21st European Young Statisticians meeting – Proceedings

Published by Faculty of Mathematics, Studentski trg 16, Belgrade

For publisher: Zoran Rakić, dean

Editors: Bojana Milošević, Marko Obradović

ISBN: 978-86-7589-137-6

21st European Young Statisticians Meeting

29 July–2 August 2019, Belgrade, Serbia

Proceedings

Eds. Bojana Milošević, Marko Obradović

Bernoulli Society
for Mathematical Statistics
and Probability

МАТЕМАТИЧКИ ФАКУЛТЕТ
У БЕОГРАДУ
1873

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 twenty seven European countries participating at the 21st 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
- [-] forty eight invited young scientists gave 20-minutes lectures.

The topics presented include, but are not limited to

- Applied statistics in biology, medicine, ...
- Bayesian inference
- Characterizations of probability distributions
- Extreme and record value theory
- Functional statistics
- Goodness-of-fit testing
- High-dimensional statistics
- Regression models
- Robust estimation

- Spatial statistics
- Stochastic processes
- Survival analysis
- Time series analysis

More information about the Conference such as 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 www.eysm2019.matf.bg.ac.rs .

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

We would like to thank 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 would also like to express our deep gratitude to our student-volunteers, in the hope that this event will be a driving force for their future academic achievements. We also appreciate very much the help of the staff of the Faculty of Mathematics. A special thanks goes to our sponsors and the Ministry of Education, Science and Technical Development of the Republic of Serbia 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.

Belgrade, September 2019 Local organizing committee

21st European Young Statisticians Meeting

Organized by

Faculty of Mathematics, University of Belgrade

Under the auspices of

Bernoulli Society for Mathematical Statistics and Probability

Ministry of Education, Science and Technological Development of the Republic of Serbia

International Organizing Committee

Apostolos Batsidis, University of Ioannina, Greece

Bettina Porvázsnyik, University of Debrecen, Hungary

Bojana Milošević, University of Belgrade, Serbia

Bruno Ebner, Karlsruher Institut für Technologie (KIT), Germany

David Preinerstorfer, Université libre de Bruxelles, Belgium

Deniz İnan, Marmara University, Turkey

Eduardo García-Portugués, Carlos III University of Madrid, Spain

Johanna Ärje, Tampere University of Technology, Finland

Juan-Juan Cai, Delft University of Technology, Netherlands

Laetitia Teixeira, University of Porto, Portugal

Måns Thulin, Uppsala University, Sweden

Marko Obradović, University of Belgrade, Serbia

Nenad Šuvak, University of Osijek, Croatia

Nina Munkholt Jakobsen, Technical University of Denmark, Denmark

Radim Navrátil, Masaryk University, Czech Republic

Riccardo De Bin, University of Oslo, Norway

Tobias Fissler, Imperial College London, United Kingdom

Wiktor Ejsmont, Mathematical Institute University of Wroclaw, Poland

Local Organizing Committee

Blagoje Ivanović, University of Belgrade, Serbia

Bojana Milošević, University of Belgrade, Serbia

Danijel Subotić, University of Belgrade, Serbia

Marija Minić, University of Belgrade, Serbia

Marko Obradović, University of Belgrade, Serbia

Keynote speakers

Ana Colubi, Justus-Liebig University of Giessen, Germany

Igor Pruñster, Bocconi University, Milan, Italy

M. Dolores Jiménez Gamero, University of Seville, Spain

Pavle Mladenović, University of Belgrade, Serbia

Peter Rousseeuw, KU Leuven, Belgium

Conference structure: keynote lectures, invited lectures.

Conference language: English

Contents

Papers

Two types of Bayesian excursion set estimates based on Gaussian process models

Dario Azzimonti[1*]

[1] *Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland*

Abstract: We consider the problem of estimating the set of points where an unknown real-valued function is above a certain threshold in the setting where only few function evaluations are available. In a Bayesian framework, we estimate the function with a Gaussian process (GP) regression model and we study the excursion set of the posterior GP distribution which is a random set. The posterior expectation of this random set provides an estimate for the unknown excursion set. We review here two random closed set expectations: the distance average expectation and the Vorob'ev expectation. We compare them empirically on a novel example that shows that Vorob'ev estimates are less informative if the measure of the set is hard to estimate. This intuition suggests a new definition of conservative estimate for excursion sets.

Keywords: Gaussian process, excursion set estimation, experimental design.

AMS subject classification: 62M30, 60G15

1 Introduction

We consider the problem of estimating an excursion set for an unknown expensive to evaluate function. More formally, our object of interest is the set

$$\Gamma^* = \{x \in \mathbb{X} : f(x) \geq t\}, \tag{1}$$

where $\mathbb{X} \subset \mathbb{R}^d$, and $f : \mathbb{X} \subset \mathbb{R}^d \to \mathbb{R}$ is an unknown expensive to evaluate function. Such problems can be found, for example, in reliability engineering [6, 4] where Γ^* represents safe configurations for a particular system and the function is often evaluated with a (possibly noisy) computer experiment.

In this paper, we estimate the function f with a Gaussian process (GP) regression model [10], and then we use the posterior distribution of the process to obtain estimates for Γ^*. In particular here we compare two types of set estimates based on two random closed set expectations: the Vorob'ev expectation [9, 7] and the distance average expectation [9, 2].

[*] Corresponding author: dario.azzimonti@idsia.ch

2 Gaussian process regression

A Gaussian process $\xi \sim GP(m, k)$ with mean function $m : \mathbb{X} \to \mathbb{R}$ and positive definite kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, is a stochastic process such that, for any $x_1, \ldots, x_\ell \in \mathbb{X}$ and any $\ell > 0$, the vector $[\xi_{x_1}, \ldots, \xi_{x_\ell}]^T$ is a multivariate Gaussian with mean $[m(x_1), \ldots, m(x_\ell)]^T$ and covariance matrix $K_\ell = [k(x_i, x_j)]_{i,j=1,\ldots,\ell}$. In a Bayesian setting, this stochastic process defines a prior over functions, and, given an observation model, we can estimate the true function by looking at the posterior distribution of the GP.

We consider n function evaluations $\mathbf{f}_n = [f(x_1), \ldots, f(x_n)] \in \mathbb{R}^n$, possibly corrupted by noise, i.e. we have $\mathbf{x}_n = [x_1, \ldots, x_n]$ and $\mathbf{y}_n = [y_1, \ldots, y_n]$, where $y_i = f(x_i) + \epsilon_i$, $i = 1, \ldots, n$, with ϵ_i independent measurement noise. We assume that the unknown function f is a realization of $\xi \sim GP(m, k)$, i.e. $p(\mathbf{f}_n) = N(m(\mathbf{x}_n), K_n)$, where $K_n = [k(x_i, x_j)]_{i=1,\ldots,n}$. The likelihood of the observations is Gaussian, i.e. $p(\mathbf{y}_n \mid \mathbf{f}_n) = N(\mathbf{f}_n, \sigma^2_{noise} I_n)$ where σ^2_{noise} is the noise variance and $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix. By exploiting Bayes theorem we can compute the posterior distribution $p(\mathbf{f}_n \mid \mathbf{y}_n) = \frac{p(\mathbf{f}_n)p(\mathbf{y}_n|\mathbf{f}_n)}{p(\mathbf{y}_n)}$. In the GP regression case the posterior has the remarkable property of being normally distributed with analytical formulae for the posterior mean m_n and covariance k_n, see, e.g. [10], chapter 1.

In practice the covariance kernel k often depends on hyper-parameters which, along with σ^2_{noise} and possibly some parameters describing the mean function m, are usually unknown. Several techniques are available in this case, see, e.g. [12], chapters 3, 4. Our focus here is on the set estimation part, so we always assume that all hyper-parameters are given and we omit conditioning on the hyper-parameters in our notation. In the experimental section we plug-in maximum likelihood estimators for the kernel hyper-parameters and use an empirical Bayes estimator for $p(\mathbf{f}_n \mid \mathbf{y}_n)$ which is still Gaussian in this case.

3 Set estimation

In order to provide estimates for Γ^* we exploit the posterior distribution of the GP. We assume that the process ξ is continuous, then $\Gamma = \{x \in \mathbb{X} : \xi(x) \geq t\}$ is a random closed set. The posterior distribution of ξ induces a posterior random set Γ and, by taking its expectation, we obtain an estimate for Γ^*. There is no unique definition for $\mathbb{E}[\Gamma]$, see [9, 1] for a more comprehensive treatment; here we compare the Vorob'ev and the distance average expectation.

Vorob'ev expectation

A key tool for computing the Vorob'ev expectation is the posterior *coverage probability*, defined, for $x \in \mathbb{X}$, as

$$p_n(x) = P(x \in \Gamma \mid \boldsymbol{\xi}_n = \mathbf{y}_n) = P_n(x \in \Gamma),$$

Since the process ξ is Gaussian, the coverage probability can be computed as $p_n(x) = \Phi\left(\frac{m_n(x)-t}{\sqrt{k_n(x,x)}}\right)$, where $\Phi(\cdot)$ is the CDF of a standard Normal random variable, m_n and k_n are the posterior mean and covariance kernel [7, 1]. By thresholding the function p_n, we define a family of set estimates $\{Q_\rho : \rho \in [0,1]\}$ called Vorob'ev quantiles, where $Q_\rho = \{x \in \mathbb{X} : p_n(x) \geq \rho\}$.
The Vorob'ev expectation chooses a specific value for ρ by exploiting the notion of a finite measure on $\mathbb{X}$, here denoted by μ. Depending on the application, the measure μ has different meanings, for example, if the input space is compact, μ is often the volume; in other applications μ is a probability distribution on $\mathbb{X}$. The Vorob'ev expectation [9, 7, 1] is then the quantile with the closest measure to the (posterior) expected measure of Γ, i.e. the quantile Q_{ρ_V} with $\rho_V \in \arg\min_{\rho\in[0,1]} \mid \mu(Q_\rho) - \mathbb{E}\left[\mu(\Gamma) \mid \boldsymbol{\xi}_n = \mathbf{y}_n\right] \mid$.

Distance average expectation

The Vorob'ev approach is based on the definition of a measure μ on $\mathbb{X}$. The distance average expectation is instead built on a notion of distance. For simplicity we consider here the Euclidean distance, however the notion can be generalized [9, 1]. Let us denote with $\delta(x, x')$ the Euclidean distance between $x, x' \in \mathbb{X}$. The *distance function* of set $A \subset \mathbb{X}$ is defined as

$$d(x, A) = \inf\{\delta(x, y) : y \in A\}, \qquad \text{for } x \in \mathbb{X}.$$

If $A = \Gamma$ is a random set, then $d(x, \Gamma)$ is a random variable for each $x \in \mathbb{X}$, we define the mean distance function as $\bar{d}(x) = \mathbb{E}\left[d(x,\Gamma)\right]$ for each $x \in \mathbb{X}$. We build a family of possible set estimates $D_u = \{x \in \mathbb{X} : \bar{d}(x) \leq u\}$, defined by thresholding $\bar{d}$. The distance average expectation is then the set D_{u^*} where

$$u^* \in \arg\min_{u>0} \left(\int_{\mathbb{X}} (d(t, D_u) - \bar{d}(x))^2 dt\right)^{1/2}.$$

See [2] for an application of this expectation to the set estimation problem described earlier. In that paper, the authors also propose a fast approximate method to generate posterior realizations of Γ that can be used to empirically compute $\bar{d}$ and the distance average expectation D_{u^*}.

Comparison

We now compare the two expectations previously reviewed on the following family of functions designed to highlight their differences. We consider

$$f(x_1, x_2; \gamma) = 5e^{-5\gamma\left(x_1^2 - x_2 - x_2^2\right)^2} \tag{2}$$

indexed by $\gamma \in (0, 1]$, where $x = (x_1, x_2) \in \mathbb{R}^2$ and the excursion set $\Gamma^* = \{f(x) \geq t\}$ with $t = 4$. This family of functions has the property that, by increasing γ, Γ^*

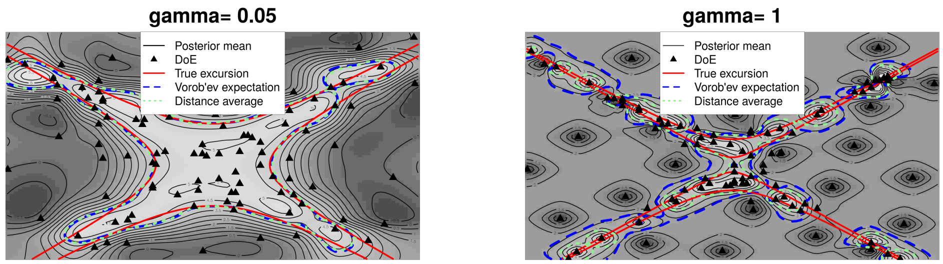


Figure 1: Analytical function in equation (2), two values for γ.

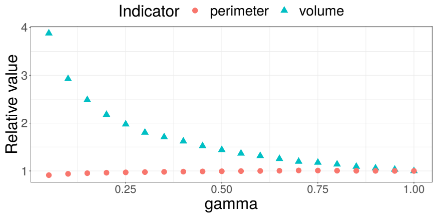


Figure 2: Indicators for Γ^*

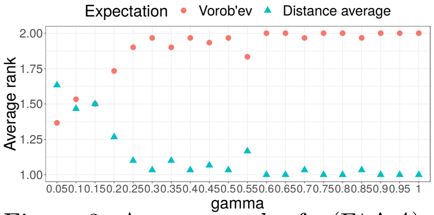


Figure 3: Average rank of $\mu(\Gamma^* \Delta A)$

becomes "thinner", i.e. its volume decreases faster than its perimeter. Figure 2 shows the volume and perimeter of Γ^* for different γ, relative to their values when $\gamma = 1$. As γ increases, the perimeter does not change significantly while the volume decreases by 4 times.

We consider a GP model built with $n = 100$ function evaluations at a design of experiments (DoE) chosen to always have points inside the excursion set. We compute the Vorob'ev expectation, Q_{ρ_V}, with the usual volume and the distance average expectation, D_{u^*}, with the Euclidean distance. Figure 1 shows the posterior GP mean computed on $n = 100$ evaluations of f, the true excursion set computed with 10000 evaluations of f on a grid and the set estimates for two values of γ. We compare the two estimates with two metrics. The first one is the true distance in measure with respect to Γ^*, i.e. $\mu(\Gamma^* \Delta A)$[1] where $A = Q_{\rho_V}$ or D_{u^*}, μ is the usual volume and Γ^* is computed from 10000 evaluations of f. The second one is the squared L^2 norm of the difference between the distance function $d(\cdot, \Gamma^*)$ and $d(\cdot, A)$, i.e. $d_{L^2}(\Gamma^*, A) := \int_{\mathbb{X}} (d(x, \Gamma^*) - d(x, A))^2 dx$, where A and Γ^* are computed as described above. We repeat the experiment 30 times with different (randomized) DoEs[2]. Figure 3 shows the average rank of $\mu(\Gamma^* \Delta A)$ for the two expectations over all repetitions as a function of γ. The metric $d_{L^2}(\Gamma^*, A)$ shows a similar behavior: table 1 reports the average value of the metric for three values of γ and the average rank in parenthesis.

In this example, both metrics seem to confirm that as the set volume decreases, thus becoming a less informative quantity for identifying the set, the distance average becomes a better estimator for Γ^*.

[1] $A \Delta B = A \setminus B \cup B \setminus A$

[2] Run in R, packages DiceKriging[11], DiceDesign[8], pGPx[2] available on CRAN

	$\gamma = 0.05$	$\gamma = 0.5$	$\gamma = 1$
Vorob'ev	$7.52 \cdot 10^{-4}$ (1.63)	$1.84 \cdot 10^{-3}$ (1.9)	$1.98 \cdot 10^{-3}$ (1.9)
Distance Avg	$7.31 \cdot 10^{-4}$ (1.37)	$1.43 \cdot 10^{-3}$ (1.1)	$1.45 \cdot 10^{-3}$ (1.1)

Table 1: Average values (rank) of $d_{L^2}(\Gamma^*, A)$ for the two expectations.

4 Conservative estimation

In some applications, such as, e.g., reliability engineering [4] or climatology [5], we would like to constrain our estimates in such a way that only regions with high probability of excursion are selected. *Conservative estimates*, as developed in [5, 3, 4], provide such probabilistic guarantee.

The idea is to select a set estimate that has probability at least $\alpha \approx 1$ to be inside Γ. Since the empty set is always included in Γ, we can obtain non trivial estimates by selecting the "largest" estimate, according to some indicator, that satisfies the probabilistic inclusion. In full generality, let us consider a family of set estimates $\mathfrak{C} = \{C_\theta, \theta \in \Theta\}$, where each set C_θ is an estimate indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^k$, then a conservative estimate for Γ at level α is a set C_{θ^*} with

$$\theta^* \in \arg\max_{\theta \in \Theta} \{I(C_\theta) : P_n(C_\theta \subset \Gamma) \geq \alpha\} \tag{3}$$

where $I : \mathfrak{C} \to \mathbb{R}_+$ associates to each set a value, such as its volume, measure or diameter. In practice, different choices for $\mathfrak{C}$ and I lead to different estimates.

Usually, [5, 3, 4], $\mathfrak{C}$ is the family of Vorob'ev quantiles $\{Q_\rho,\ \rho \in [0,1]\}$ and $I(\cdot)$ is a measure μ on $\mathbb{X}$. A conservative estimate with these choices is the largest (in terms of measure) Vorob'ev quantile such that $P(Q_\rho \subset \Gamma) \geq \alpha$. See [3] for a full implementation with a fast method to compute $P_n(Q_\rho \subset \Gamma)$ and [4] for sequential strategies to reduce the uncertainty on such estimates.

The comparison shown in the previous section, suggests that, when an informative measure μ is not available, an alternative choice based on the distance average could be useful. In this case a valid alternative could be a conservative estimate based on distance average where $\mathfrak{C}$ is the family $\{D_u : u \geq 0\}$ as defined in section 3 and $I(\cdot)$ as the diameter of the set, defined as $diam(A) = \sup\{\delta(x, x') : x, x' \in A\}$. This estimate only depends on the metric δ and can be computed by empirically estimating the probability of inclusion with the pseudo-realizations method introduced in [2].

5 Conclusion

In this paper we briefly reviewed two types of estimators for Γ^* and we showed, on a novel synthetic example, how the distance average expectation might provide better estimates when the measure of the set is not very informative. The conclusion is dependent on the metrics we used to evaluate how informative is the estimator. Nonetheless the observation above could lead to an alternative definition of conservative estimates. We proposed the first steps towards computing this estimate, however more studies are required to better analyze the advantages of distance average based conservative estimates.

Acknowledgements: The author acknowledges support from the SNSF, grant number 167199. The topics presented here are mostly the results of joint works with: D. Ginsbourger, C. Chevalier, J. Bect, and Y. Richet.

Bibliography

[1] D. Azzimonti. *Contributions to Bayesian set estimation relying on random field priors.* PhD thesis, University of Bern, 2016.

[2] D. Azzimonti, J. Bect, C. Chevalier, and D. Ginsbourger. Quantifying uncertainties on excursion sets under a Gaussian random field prior. *SIAM/ASA J. Uncertain. Quantif.*, 4(1):850–874, 2016.

[3] D. Azzimonti and D. Ginsbourger. Estimating orthant probabilities of high dimensional Gaussian vectors with an application to set estimation. *J. Comput. Graph. Statist.*, 27(2):255–267, 2018.

[4] D. Azzimonti, D. Ginsbourger, C. Chevalier, J. Bect, and Y. Richet. Adaptive Design of Experiments for Conservative Estimation of Excursion Sets. *Under revision, hal-01379642*, 2019.

[5] D. Bolin and F. Lindgren. Excursion and contour uncertainty regions for latent Gaussian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77(1):85–106, 2015.

[6] C. Chevalier, J. Bect, D. Ginsbourger, E. Vazquez, V. Picheny, and Y. Richet. Fast kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465, 2014.

[7] C. Chevalier, D. Ginsbourger, J. Bect, and I. Molchanov. Estimating and quantifying uncertainties on level sets using the Vorob'ev expectation and deviation with Gaussian process models. In D. Uciński, A. Atkinson, and C. Patan, editors, *mODa 10 – Advances in Model-Oriented Design and Analysis.* Physica-Verlag HD, 2013.

[8] J. Franco, D. Dupuy, O. Roustant, G. Damblin, and B. Iooss. *DiceDesign: Designs of Computer Experiments*, 2013. R package version 1.3.

[9] I. Molchanov. *Theory of Random Sets.* Springer London, London, 2017.

[10] C. E. Rasmussen and C. K. Williams. *Gaussian processes for machine learning.* MIT Press, 2006.

[11] O. Roustant, D. Ginsbourger, and Y. Deville. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization. *Journal of Statistical Software*, 51(1):1–55, 2012.

[12] T. J. Santner, B. J. Williams, and W. I. Notz. *The Design and Analysis of Computer Experiments.* Springer New York, New York, NY, 2018.

A new method for the estimation of distribution functions in parametric models

Steffen Betsch[1*]

[1]*Institute of Stochastics, Karlsruhe Institute of Technology*

Abstract: A new method for the estimation of the cumulative distribution function (CDF) in a parametric model of continuous probability distributions is proposed. The method is based on a simple representation of the CDF and provides continuous estimation functions which, in the given simulations, outperform the empirical CDF vastly. Moreover, the new approach is applicable to models where the CDF cannot be given explicitly, and not even the normalization constant of the density function needs to be known. The main goal of this contribution is to draw attention to this flexible method of estimation, and to stimulate further research.

Keywords: CDF estimation, Cumulative distribution function
AMS subject classification: 62F99, 62E10

1 A new class of estimators

In applications, statisticians frequently encounter the problem of estimating the CDF based on a given (real-valued) sample $X_1, \ldots, X_n$. The empirical distribution function $\widehat{F}_n(\cdot) = n^{-1} \sum_{j=1}^{n} \mathbb{1}\{X_j \leq \cdot\}$, being the most prominent estimator, is of non-parametric nature and admits strong consistency properties for both discrete and continuous probability distributions without any further assumptions. It is however not a continuous function itself even if the data stem from a continuous probability distribution, and further knowledge on the data, for instance that the observations come from a parametric model, cannot be incorporated. Of course, in parametric models where the CDF is given explicitly, the immediate alternative to the empirical CDF is to estimate the parameters of the model from the data and to plug the estimated values into the explicit formula. However, even for simpler classical models, like the normal- or Gamma distribution, no explicit form of the CDF exists and its calculation is possible by numerical integration only. In this contribution we introduce a new, flexible class of CDF-estimators for parametric classes of continuous probability distributions on the positive half axis of the real line which admit a Lebesgue density. In particular, the method addresses the situation of

[*] Corresponding author: steffen.betsch@kit.edu

non-normalized models, that is, the estimators can be used even if the normalization (integration) constant of the density functions is intractable. Such models frequently occur in machine learning and signal processing, and we refer to [4] for a little more detail about this particular situation and a method for the estimation of parameters in those models.

The construction of the new estimators is essentially based on the necessity part of the characterizations given in [2]. Since the argument is simple and rather short, we give it below, focusing on the particular situation at hand. From the results in [2] it is immediate that similar estimators can be constructed when the density function is supported by the whole real line or by a bounded interval.

As for now, assume that $\{p_\vartheta \,|\, \vartheta \in \Theta\}$, $\Theta \subset \mathbb{R}^d$, is a parametric model given through density functions defined on the interval $(0, \infty)$, which are continuously differentiable, and satisfy $p_\vartheta(x) > 0$ for each $x > 0$. Further assume that for every $\vartheta \in \Theta$,

$$\int_0^\infty x \big| p_\vartheta'(x) \big| \mathrm{d}x + \int_t^\infty \big| p_\vartheta'(x) \big| \mathrm{d}x < \infty, \quad t > 0. \tag{1}$$

Let $X, X_1, \ldots, X_n$ be independent and identically distributed (iid.) random variables with distribution given through $p_0 = p_{\vartheta_0}$, for some (unknown) $\vartheta_0 \in \Theta$, and denote by F_0 the distribution function of X. By (1) we may use the fundamental theorem of calculus and Fubini's theorem to calculate

$$\begin{aligned} F_0(t) = \int_0^t p_0(s)\,\mathrm{d}s &= -\int_0^t \int_s^\infty p_0'(x)\,\mathrm{d}x\,\mathrm{d}s \\ &= \int_0^t \mathbb{E}\left[-\frac{p_0'(X)}{p_0(X)}\,\mathbb{I}\{X > s\} \right] \mathrm{d}s \\ &= \mathbb{E}\left[-\frac{p_0'(X)}{p_0(X)}\,\min\{X, t\} \right], \quad t > 0. \end{aligned} \tag{2}$$

This simple relation leads us to suggesting the estimation of F_0 as follows: Take some consistent estimator $\hat{\vartheta}_n$ of ϑ_0 based on the sample $X_1, \ldots, X_n$ and consider

$$\widehat{T}_n(t) = -\frac{1}{n} \sum_{j=1}^n \frac{p_{\hat{\vartheta}_n}'(X_j)}{p_{\hat{\vartheta}_n}(X_j)}\,\min\{X_j, t\}, \quad t > 0.$$

If the relation $\vartheta \mapsto \frac{p_\vartheta'}{p_\vartheta}$ is sufficiently smooth, $\widehat{T}_n$ ought to be a good (pointwise!) estimator of the right-hand side of (2) and thus of F_0. Indeed, in many examples it is obvious from the form of $\frac{p_\vartheta'(x)}{p_\vartheta(x)}$ that $\widehat{T}_n(t) \to F_0(t)$ in probability or almost surely (a.s.), depending on whether $\hat{\vartheta}_n$ is weakly or strongly consistent. Under a suitable and rather weak Hölder-continuity assumption for the mapping $\vartheta \mapsto \frac{p_\vartheta'(x)}{p_\vartheta(x)}$ on compact subsets of Θ it can also be shown in a more general setting that $\widehat{T}_n(t) \to F_0(t)$ in probability. If $(x, \vartheta) \mapsto \frac{p_\vartheta'(x)}{p_\vartheta(x)}$ is measurable then the function $t \mapsto \widehat{T}_n(t)$ is easily seen to be a random element of the space of continuous functions on the positive axis. Thus, the question arises whether (minimal) sufficient conditions can be identified so that $\widehat{T}_n$ converges to F_0 in the space of continuous functions, i.e., if the quantity

$$\big\|\widehat{T}_n - F_0\big\|_\infty = \sup_{t\,>\,0} \big|\widehat{T}_n(t) - F_0(t)\big|$$

converges to zero in some stochastic mode of convergence. The study of this question is not pursued here, but to motivate further research on the topic, we provide two examples in one of which these results can be inferred explicitly, and we illustrate how the new estimation method competes against the classical estimators with respect to the uniform metric given above.

2 Example: The exponential distribution

Assume that $X, X_1, \ldots, X_n$ are iid. exponentially distributed with rate parameter $\vartheta_0 > 0$, i.e., X has density function $p_0(x) = \vartheta_0 e^{-\vartheta_0 x}$, $x > 0$. Denote by $\overline{X}_n$ the sample mean, and let $\hat{\vartheta}_n = (\overline{X}_n)^{-1}$ be the maximum likelihood estimator, which converges to ϑ_0 a.s. As classical estimators for F_0, we consider the empirical CDF $\widehat{F}_n$, and $\widetilde{F}_n(x) = 1 - e^{-\hat{\vartheta}_n x}$, the theoretical CDF with estimated parameter. Our new estimator takes the form

$$\widehat{T}_n(t) = \frac{1}{n} \sum_{j=1}^{n} (\overline{X}_n)^{-1} \min\{X_j, t\}, \quad t > 0.$$

It is readily seen that $\widehat{T}_n$ is itself a CDF (a.s.). Therefore, since $\widehat{T}_n(t) \to F_0(t)$ a.s. for each $t > 0$, the proof of the Glivenko-Cantelli theorem for the empirical CDF also yields $\|\widehat{T}_n - F_0\|_\infty \to 0$ a.s. Table 1 gives an idea on how the estimators perform in the uniform metric. We calculate $\|\widehat{F}_n - F_0\|_\infty$ by the well-known formula for the Kolmogorov-Smirnov statistic, obtain

$$\big\|\widetilde{F}_n - F_0\big\|_\infty = \left| \Big(\frac{\hat{\vartheta}_n}{\vartheta_0}\Big)^{\frac{\vartheta_0}{\vartheta_0 - \hat{\vartheta}_n}} - \Big(\frac{\hat{\vartheta}_n}{\vartheta_0}\Big)^{\frac{\hat{\vartheta}_n}{\vartheta_0 - \hat{\vartheta}_n}} \right|$$

explicitly, and approximate

$$\big\|\widehat{T}_n - F_0\big\|_\infty \approx \sup\nolimits_{j\,=\,1,\ldots,300} \big|\widehat{T}_n(t_j) - F_0(t_j)\big|,$$

where $0 < t_1 < \cdots < t_{300} \leq X_{(n)} = \max\{X_1, \ldots, X_n\}$ are equidistant points (to justify this approximation, note that $\widehat{T}_n(t) = 1$ a.s. for all $t \geq X_{(n)}$). The first quantity of course, does not depend on the underlying distribution. All simulations were performed with Python 3.7.2 (as provided by the Python Software Foundation, https://www.python.org, accessed 19 September 2019).

The results in Table 1 are not surprising, but confirm, at least in this simple example, that the new estimation method is sound: The theoretical CDF with estimated parameter fares best for it incorporates the most information about the underlying

ϑ_0	n	$\Vert\widehat{F}_n - F_0\Vert_\infty$	$\Vert\widetilde{F}_n - F_0\Vert_\infty$	$\Vert\widehat{T}_n - F_0\Vert_\infty$
0.5	25	0.168	0.0591	0.0925
	50	0.1194	0.0417	0.0651
	100	0.0852	0.0296	0.0461
1	25	0.1674	0.0586	0.092
	50	0.1203	0.0421	0.0655
	100	0.085	0.0292	0.0455
2	25	0.167	0.0589	0.0922
	50	0.1194	0.0414	0.0647
	100	0.0851	0.0292	0.0455

Table 1: (Approximated) values calculated with 10,000 exponentially distributed Monte Carlo samples for sample sizes $n = 25, 50, 100$.

parametric model. The new estimator follows behind and clearly outperforms the empirical CDF. This is insofar promising, as that the second estimator $\widetilde{F}_n$ is not available for more complex distributions that do not have an explicit CDF (see the next example).

3 Example: The Nakagami distribution

Let $X, X_1, \ldots, X_n$ be iid. Nakagami-distributed with parameters $m_0, s_0 > 0$, i.e., X has Lebesgue density

$$p_0(x) = \frac{2m_0^{m_0}}{\Gamma(m_0)\, s_0^{m_0}}\, x^{2m_0-1} \exp\Big(- \frac{m_0}{s_0}\, x^2 \Big), \quad x > 0,$$

where Γ denotes the Gamma function. By $\hat{m}_n$ and $\hat{s}_n$ we denote the maximum likelihood estimators which, in the simulation, are calculated with the 'stats.nakagami.fit'-method in the 'scipy'-module of Python (see [6]). Our estimator for the CDF takes the form

$$\widehat{T}_n(t) = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{2\hat{m}_n X_j}{\hat{s}_n} - \frac{2\hat{m}_n - 1}{X_j} \right) \min\{X_j, t\}, \quad t > 0,$$

and we compare it to the empirical CDF $\widehat{F}_n$. Since the CDF of the Nakagami distribution can only be given in terms of the regularized incomplete Gamma function, we do not use the naive estimator in this case. We have $\lim_{t \searrow 0} \widehat{T}_n(t) = 0$ and $\widehat{T}_n(t) = 1$ a.s., $t \geq X_{(n)}$ (note that $\hat{s}_n = n^{-1} \sum_{j=1}^{n} X_j^2$), so we calculate the quantities in Table 2 just like for the exponential distribution. It is however not so clear if $\widehat{T}_n$ is itself increasing and hence a CDF. Still, the simulation results indicate that a uniform consistency result should hold in this case.

(m_0, s_0)	n	$\lVert \widehat{F}_n - F_0 \rVert_\infty$	$\lVert \widehat{T}_n - F_0 \rVert_\infty$
$(1, 1)$	25	0.167	0.0995
	50	0.1205	0.0697
	100	0.0857	0.0493
$(0.25, 1.5)$	25	0.1676	0.0939
	50	0.119	0.0646
	100	0.0851	0.0452
$(2, 0.25)$	25	0.1679	0.1060
	50	0.1194	0.0738
	100	0.085	0.052
$(3, 0.75)$	25	0.1669	0.1067
	50	0.1194	0.0751
	100	0.0852	0.053

Table 2: (Approximated) values calculated with 10,000 Nakagami-distributed Monte Carlo samples for sample sizes $n = 25, 50, 100$.

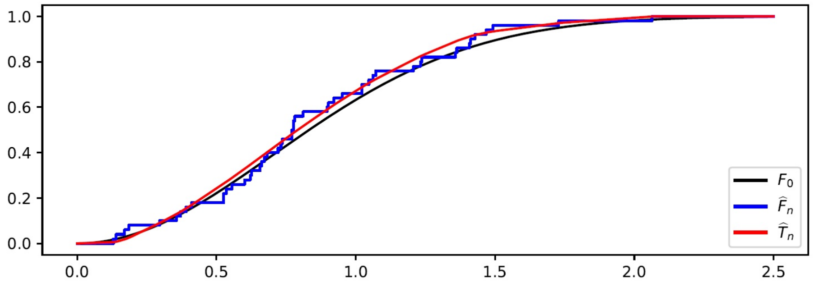


Figure 1: Plots for a Nakagami-distributed sample of size 50 with underlying parameters $m_0 = 1$, $s_0 = 1$.

It is readily seen from Table 2 that our new estimator outperforms the empirical CDF also for this slightly more complicated distribution. Note that preliminary simulations have shown that the performance depends on the choice of estimators for m_0 and s_0, as is to be expected. Figure 1 illustrates that, since $\widehat{T}_n$ is a continuous function, it also constitutes the visually more convincing estimator as compared to the empirical CDF.

4 Comments

The article at hand is to be read as a suggestion rather than a comprehensive account of the new estimation method for the CDF, and we give some thoughts on possible further research. First note that we did not discuss the integrability condition (1), but in [2] it was argued that this condition is not restrictive. The authors of [2] also prove that under few additional conditions the explicit formula for the CDF given in equation (2) completely characterizes the distribution of F_0 within a large class of probability distributions. They also give similar characterizations for distributions supported by the whole real line or by bounded intervals, so CDF-estimators of a comparable type can be constructed thereof. These characterizations are also used to construct (consistent) goodness-of-fit tests, see e.g. [1]. We have indicated that a uniform convergence result similar to that of the empirical CDF should hold. Sufficient conditions for this conjecture and its precise proof remain to be found. Additionally, such a consistency result could possibly be extended by examining whether there exists a scaling factor $C(n)$ such that $C(n)\|\widehat{T}_n - F_0\|_\infty$ converges to a non-degenerate limit distribution. A larger scale simulation study, including different metrics and other approaches (like kernel methods or approaches from [3] and [5]), would help to understand the behavior of the estimators better, in particular with respect to the impact of the parameter estimators that have to be plugged in.

Acknowledgements: The author thanks Bruno Ebner for his careful reading of the manuscript and his helpful suggestions.

Bibliography

[1] S. Betsch and B. Ebner. A new characterization of the Gamma distribution and associated goodness-of-fit tests. *Metrika*, 82(7):779-806, 2019.

[2] S. Betsch and B. Ebner. Fixed point characterizations of continuous univariate probability distributions and their applications. *ArXiv e-prints*, 1810.06226v2, 2019.

[3] B. Funke and C. Palmes. A note on estimating cumulative distribution functions by the use of convolution power kernels. *Stat. Probabil. Lett.*, 121:90–98, 2017.

[4] A. Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *J. Mach. Learn. Res.*, 6:695–709, 2005.

[5] A. Jokiel-Rokita and R. Magiera. Minimax estimation of a cumulative distribution function by converting to a parametric problem. *Metrika*, 66(1):61–73, 2007.

[6] E. Jones and T. Oliphant and P. Peterson and others. SciPy: Open Source Scientific Tools for Python. *http://www.scipy.org*, accessed 19 September 2019, 2001.

New class of supremum-type exponentiality tests based on V-empirical Laplace transforms and Puri-Rubin characterization

Marija Cuparić,[1*] **Bojana Milošević**[1] **and Marko Obradović**[1]

[1]*Faculty of Mathematics, University of Belgrade*

Abstract: We propose a new class of scale free goodness-of-fit tests for the exponential distribution based on the Puri-Rubin characterization. For the construction of test statistics we employ weighted L^∞ distance between V-empirical Laplace transforms of the random variables that appear in the characterization. We derive the asymptotic properties, and to assess the quality of the tests, we calculate the approximate Bahadur efficiency for some common close alternatives. For small sample sizes, a simulated power study is performed. The tests are shown to be very efficient and powerful in comparison to many other exponentiality tests.

Keywords: asymptotic efficiency, Laplace transform, V-statistic with estimated parameters

AMS subject classification: 62G10, 62G20

1 Introduction

Consider testing the null hypothesis that the data come from the exponential distribution, i.e. $H_0 : X \sim \mathcal{E}(\lambda)$, where $\lambda > 0$ is an unknown scale parameter. The characterization based approach to goodness-of-fit testing usually provides test statistics that do not depend on the unknown parameter, and are, therefore, suitable for testing such a composite hypothesis. Some characterization based exponentiality tests can be found in e.g. [8], [9],[10].

Here we construct a supremum-type test based on the famous Puri-Rubin characterization proposed in [13].

Characterization 1. Let X_1 and X_2 be two independent copies of a random variable X with pdf $f(x)$. Then X and $|X_1 - X_2|$ have the same distribution, if and only if for some $\lambda > 0$, $f(x) = \lambda e^{-\lambda x}$, for $x \geq 0$.

Common tools for assessing the quality of the test are its power, in case of small and moderate sample sizes, and its asymptotic efficiency. Since the asymptotic distribution of supremum-type statistics is not normal, we opt for the Bahadur efficiency.

[*] Corresponding author: marijar@matf.bg.ac.rs

2 Test statistic

Let $X_1, X_2, ..., X_n$ be independent and identically distributed (i.i.d.) non-negative random variables with an unknown absolutely continuous distribution function F. In view of the Characterization 1, we propose the following family of test statistics, depending on the tuning parameter $a > 0$:

$$L^{\mathcal{P}}_{n,a} = \sup_{t>0} \Big| \Big(\frac{1}{n} \sum_{i_1=1}^{n} e^{-tY_{i_1}} - \frac{1}{n^2} \sum_{i_1,i_2=1}^{n} e^{-t|Y_{i_1}-Y_{i_2}|} \Big) e^{-at} \Big|, \tag{1}$$

where $Y_i = \frac{X_i}{\bar{X}}$, $i = 1, 2, \dots, n$, is the scaled sample. The scaling of the sample is done to make the statistic scale invariant under the null hypothesis, and the role of the tuning parameter is to emphasize different type of differences between the null and the alternative hypothesis.

For a fixed t, the expression in the absolute parenthesis of (1) can be represented as

$$V_n(t; \hat{\lambda}_n) = \frac{1}{n^2} \sum_{i_1,i_2} \Phi(X_{i_1}, X_{i_2}, t; \hat{\lambda}_n) e^{-at}, \tag{2}$$

where Φ is a symmetric function of its arguments. Therefore $V_n(t; \hat{\lambda}_n)$ is a V-statistic with an estimated parameter. Thus the test statistic $L^{\mathcal{P}}_{n,a}$ can be represented as $\sup\limits_{t\geq 0} |V_n(t; \widehat{\lambda}_n) e^{-at}|$, where $\{V_n(t; \widehat{\lambda}_n)\}$ is a V-empirical process.

The asymptotic behaviour of $L^{\mathcal{P}}_{n,a}$ is given in the following theorem.

Theorem 2. *Let $X_1, ..., X_n$ be i.i.d. with an exponential distribution. Then*

$$\sqrt{n} L^{\mathcal{P}}_{n,a} \xrightarrow{D} \sup_{t>0} |\eta(t)|,$$

where $\eta(t)$ is centered Gaussian process with the covariance function

$$K(s,t) = \frac{e^{-a(s+t)} st(4 + 4s + 4t + 3st)}{12(1+s)(2+s)(1+t)(2+t)(1+s+t)} \tag{3}$$

Proof. In [4] was shown that, for fixed t, the statistics $\sqrt{n} V_n(t; \widehat{\lambda}_n)$ from (2) and $\sqrt{n} V_n(t; \lambda)$ are asymptotically equally distributed, and that their distribution does not depend on λ. Hence, $\sqrt{n} V_n(t; \widehat{\lambda}_n) e^{-at}$ converges in $D(0, \infty)$ to the centered Gaussian process $\{\eta(t)\}$ (see [14]) with the covariance function

$$K(s,t) = e^{-a(s+t)} \int_0^\infty \int_0^\infty \Phi(x, y, t; 1) \Phi(x, z, s; 1) e^{-x-y-z} dx dy dz$$

which after some calculation becomes (3). Therefore $L^{\mathcal{P}}_{n,a}$ converges to $\sup_{t>0} |\eta(t)|$. This completes the proof. □

3 Approximate Bahadur efficiency

Let $\mathcal{G} = \{G(x;\theta),\ \theta > 0\}$ with corresponding densities $\{g(x;\theta)\}$ be a family of alternative distribution functions with finite expectations, such that $G(x,\theta) = 1 - e^{-\lambda x}$, for some $\lambda > 0$, if and only if $\theta = 0$, and the regularity conditions for V-statistics with weakly degenerate kernels from [12, Assumptions ND] are satisfied. Within this class of alternatives, the null hypothesis can be restated as $H_0 : \theta = 0$. For two sequences of test statistics, T_n and V_n, having the same null and alternative hypothesis, the relative Bahadur efficiency is defined as the ratio of the sample sizes needed to reach the same power when the size of the tests approaches zero. For close alternatives, i.e. alternatives from $\mathcal{G}$ for which θ is close to zero, this ratio can be expressed as the limit when $\theta \to 0$ of the ratio of the Bahadur approximate slopes (see [1]):

$$e^*_{T,V} = \lim_{\theta\to 0} \frac{c^*_T(\theta)}{c^*_V(\theta)}.$$

The Bahadur approximate slopes of T_n (and V_n) can be calculated as (see [11])

$$c^*_T(\theta) = a_T b^2_T(\theta), \tag{4}$$

where a_T is the coefficient next to x^2 in the expansion of the logarithmic tail of the limiting distribution, and $b_T(\theta)$ is the limit in probability of T_n under the alternative.

The approximate local Bahadur slope of $L^{\mathcal{P}}_{n,a}(\hat\lambda_n)$, for close alternatives, is derived in the following theorem.

Theorem 3. *For the statistic $L^{\mathcal{P}}_{n,a}(\hat\lambda_n)$ and a given alternative density $g(x,\theta)$ from $\mathcal{G}$, the local Bahadur approximate slope is given by*

$$c^*_L(\theta) = \frac{1}{\sup_{t\geq 0} K(t,t)} \sup_{t\geq 0} \Big(2\int_0^\infty \tilde\varphi_1(x;t) g'_\theta(x;0)dx\Big)^2 \cdot \theta^2 + o(\theta^2), \theta \to 0,$$

where $\tilde\varphi_1(x;t) = E(\Phi(\cdot)e^{-at}|X_1 = x)$ with Φ being defined in (2).

Proof. The tail behaviour of the random variable $\sup_{t>0}|\eta_t|$ is equal to the inverse of the supremum of the covariance function, i.e. the $a_L = \frac{1}{\sup_{t>0} K(t,t)}$ (see [7]).

Since $\overline{X}_n$ converges almost surely to its expected value $\mu(\theta)$, using the Law of large numbers for V-statistics with estimated parameters (see [6]), it follows that $V_n(t;\hat\lambda)e^{-at}$ converges to

$$b_L(t;\theta) = E_\theta(\Phi(X_1, X_2, t; \mu(\theta))e^{-at}).$$

Expanding $b_L(\theta)$ in the Maclaurin series we obtain

$$b_L(t;\theta) = 2\int_0^\infty \widetilde\varphi_1(x,t) g'_\theta(x;0)dx \cdot \theta + o(\theta),$$

where $\widetilde\varphi_1(x,t) = E(\Phi(X_1, X_2, t; 1)e^{-at}|X_1 = x_1)$. According to the Glivenko-Cantelli theorem for V-statistics [5], the limit in probability under the alternative for statistics $L^{\mathcal{P}}_{n,a}$ is equal to $\sup_{t\geq 0}|b_L(t;\theta)|$. Inserting this into (4) completes the proof. □

We calculate approximate Bahadur efficiency with respect to the LRT (see [2]) for some common alternatives, namely Weibull, Gamma, linear failure rate (LFR) and mixture of exponential distributions with negative weights (EMNW(3)) distribution. Their densities can be found e.g. in [4]. The efficiencies, as functions of the tuning parameter a, are shown on Figure 1.
In comparison to the results presented in [3] we may conclude that our new tests are reasonably to highly efficient.

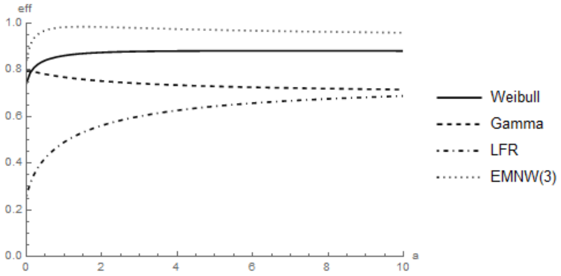


Figure 1: Local approximate relative Bahadur efficiency of $L^{(\mathcal{P})}_{n,a}$ with respect to the LRT

4 Power study

Here we present the simulated powers of new tests against alternatives whose densities can be found e.g. in [3]. The empirical powers are obtained using the Monte Carlo procedure based on 10000 replicates, for sample sizes $n = 20$ and level of significance $\alpha = 0.05$. For comparison purpose we use the same labels as in [4] and [15]. The powers vary from moderate to very high.

Table 1: Percentage of rejected hypotheses for $n = 20$

Alt.	$Exp(1)$	$W(1.4)$	$\Gamma(2)$	HN	U	$CH(0.5)$	$CH(1)$	$CH(1.5)$	$LF(2)$	$LF(4)$	$EV(1.5)$	$LN(0.8)$	$LN(1.5)$	$DL(1)$	$DL(1.5)$	$W(0.8)$	$\Gamma(0.4)$
$L^{\mathcal{P}}$	5	47	64	25	67	17	17	17	34	46	47	55	1	39	82	1	27
$L^{\mathcal{P}}_{n,0.5}$	5	50	63	29	77	24	23	22	40	53	58	47	6	34	80	2	34
$L^{\mathcal{P}}_{n,1}$	5	50	62	31	79	23	22	22	41	55	55	40	12	32	81	3	42
$L^{\mathcal{P}}_{n,2}$	5	49	62	31	80	22	25	25	41	57	58	40	27	33	78	4	47
$L^{\mathcal{P}}_{n,5}$	5	51	62	31	80	24	23	23	42	57	60	34	39	29	76	7	53
$L^{\mathcal{P}}_{n,10}$	5	47	60	32	80	22	23	23	41	57	60	33	47	28	75	8	57

Acknowledgements: The work of M. Cuparić and B. Milošević is supported by the MNTRS, Serbia under Grant No. 174012.

Bibliography

[1] R. R. Bahadur. On the asymptotic efficiency of tests and estimates. *Sankhyā: The Indian Journal of Statistics*, pages 229–252, 1960.

[2] R. R. Bahadur. Rates of convergence of estimates and test statistics. *The Annals of Mathematical Statistics*, 38(2):303–324, 1967.

[3] M. Cuparić, B. Milošević, and M. Obradović. New consistent exponentiality tests based on V-empirical Laplace transforms with comparison of efficiencies. *Preprint, arXiv:1904.00840*, 2019.

[4] M. Cuparić, B. Milošević, and M. Obradović. New l^2-type exponentiality tests. *SORT*, 43(1):25–50, 2019.

[5] R. Helmers, P. Janssen, and R. Serfling. Glivenko-Cantelli properties of some generalized empirical df's and strong convergence of generalized L-statistics. *Probability theory and related fields*, 79(1):75–93, 1988.

[6] H. Iverson and R. Randles. The effects on convergence of substituting parameter estimates into U-statistics and other families of statistics. *Probability Theory and Related Fields*, 81(3):453–471, 1989.

[7] M. B. Marcus and L. Shepp. Sample behavior of Gaussian processes. In *Proc. of the Sixth Berkeley Symposium on Math. Statist. and Prob*, volume 2, pages 423–421, 1972.

[8] B. Milošević. Asymptotic efficiency of new exponentiality tests based on a characterization. *Metrika*, 79(2):221–236, 2016.

[9] B. Milošević and M. Obradović. New class of exponentiality tests based on U-empirical Laplace transform. *Statistical Papers*, 57(4):977–990, 2016.

[10] B. Milošević and M. Obradović. Some characterization based exponentiality tests and their Bahadur efficiencies. *Publications de L'Institut Mathematique*, 100(114):107–117, 2016.

[11] Ya. Yu. Nikitin. *Asymptotic efficiency of nonparametric tests.* Cambridge University Press, New York, 1995.

[12] Ya. Yu. Nikitin and I. Peaucelle. Efficiency and local optimality of nonparametric tests based on U- and V-statistics. *Metron*, 62(2):185–200, 2004.

[13] P. S. Puri and H. Rubin. A characterization based on the absolute difference of two iid random variables. *The Annals of Mathematical Statistics*, 41(6):2113–2122, 1970.

[14] B. W. Silverman. Convergence of a class of empirical distribution functions of dependent random variables. *The Annals of Probability*, 11(3):745–751, 1983.

[15] H. Torabi, N. H. Montazeri, and A. Grané. A wide review on exponentiality tests and two competitive proposals with application on reliability. *Journal of Statistical Computation and Simulation*, 88(1):108–139, 2018.

Inference on high-dimensional graphical models via pairwise likelihood truncation

Claudia Di Caterina,[1*] **Davide Ferrari**[1,2] **and Davide La Vecchia**[3]

[1]*Free University of Bozen-Bolzano, Faculty of Economics and Management*
[2]*School of Mathematics and Statistics, University of Melbourne*
[3]*University of Geneva, Geneva School of Economics and Management*

Abstract: Undirected graphical models are popular in a number of fields due to their interpretablity and flexibility in describing complex multivariate distributions. Efficient estimation and selection of graphs, however, remain challenging when the number of connections is large relative to the sample size, even under the Gaussian distributional assumption. Within a composite likelihood framework, a novel methodology which simultaneously estimates parameters and selects edges is proposed. The procedure consists of minimizing the divergence of the pairwise composite likelihood score from the full likelihood score, subject to a constraint representing the graph sparsity. The empirical performance of such approach is assessed through data simulated from a Gaussian random field.

Keywords: Composite likelihood estimation, model selection, ℓ_1-penalization.
AMS subject classification: 62H12, 62M40, 62P10.

1 Introduction

Undirected graphical models have been extensively applied in a variety of fields, such as medicine, physics and engineering, due to their flexibility and facility of interpretation. These models describe complex multivariate distributions through the product of simpler clique-specific sub-models (e.g. pairwise models describing edges). One crucial question is how to select the structure of large graphs, i.e. how to obtain the list of edges from a large set of feasible edges. Numerous works in the literature propose to achieve such selection through solutions hinging on likelihoods with ℓ_1-type penalties [4, 5]. To address situations where the maximization of the likelihood function is impracticable, penalized approaches based on composite likelihood (CL) estimation have been suggested [1, 2, 7]. Within the CL framework, intractability of the full likelihood is avoided by taking a weighted combination of low-dimensional likelihood objects [6].

[*] Corresponding author: claudia.dicaterina@unibz.it

In this paper, a new strategy to determine the non-zero edges in high-dimensional graphs, called truncated pairwise likelihood (TPL), is introduced. In TPL, a data-driven combination of pairwise likelihood objects built on node pairs is selected by minimizing a distance between the maximum likelihood and the pairwise likelihood score functions, subject to a ℓ_1-penalty discouraging the inclusion of too many terms in the final estimating equation. The proposed criterion may be interpreted as to maximizing the statistical accuracy of the selected model for a given level of sparsity in the graph.

2 Pairwise likelihood truncation

Let $X = (X_1, \dots, X_d)$ be the random vector following a joint distribution indexed by the parameter of interest $\theta \in \Theta \subseteq \mathbb{R}^q$, $q \geq 1$. Consider the undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for the set $\mathcal{V}$ of nodes and the set $\mathcal{E}$ of edges related to single variables and variable pairs in X, respectively. Supposing that the distribution of X satisfies the Markov independence assumption, by the Hammersley-Clifford theorem, one can write the density function of X as

$$f(x_1, \dots, x_d; \theta) = \exp\left\{\sum_{s\in\mathcal{V}} \eta_s \phi_s(x_s) + \sum_{(s,t)\in\mathcal{E}} \eta_{st}(\theta)\phi_{st}(x_s, x_t) - \log a(\theta)\right\}, \qquad (1)$$

where $\{\phi_s(X_s), s \in \mathcal{V}\}$ and $\{\phi_{st}(X_s, X_t), (s,t) \in \mathcal{E}\}$ are node- and edge-specific sufficient statistics for canonical parameters $\{\eta_s\}$ and $\{\eta_{st}(\theta)\}$, respectively, and $a(\theta)$ is the normalization constant. For simplicity, the single marginal components $\{\eta_s\}$ are treated as nuisance parameters independent of θ, while the pairwise components $\{\eta_{st}(\theta)\}$ serve to describe the dependence between X_s and X_t through θ. Specifically, if $\eta_{st}(\theta) = 0$ the sth and the tth nodes are disconnected. Our focus here is on sparse graphs whose dimension $p = |\mathcal{E}| = d(d-1)/2$ is much larger than the number p_1 of non-zero edges.

Based on model (1), the marginal density for the pair (X_s, X_t) is

$$f_{st}(x_s, x_t; \theta) = \exp\{\eta_s\phi_s(x_s) + \eta_t\phi_t(x_t) + \eta_{st}(\theta)\phi_{st}(x_s, x_t) - \log a_{st}(\theta)\}, \qquad (2)$$

with $a_{st}(\theta)$ being the pairwise normalization term. For n independent observations $X^{(1)}, \dots, X^{(n)}$ on X, the maximum composite likelihood estimator $\hat{\theta}(w)$ is found by solving the q-dimensional estimating equation

$$0 = \sum_{i=1}^{n} u(\theta; w, X^{(i)}) = \sum_{(s,t)\in\mathcal{E}} w_{st} \sum_{i=1}^{n} u_{st}(\theta; X^{(i)}), \qquad (3)$$

with $u_{st}(\theta; x) = \partial \log f_{st}(x_s, x_t; \theta)/\partial\theta$ pairwise score function where the nuisance parameters are replaced by estimates $\{\hat{\eta}_s, \hat{\eta}_t\}$ obtained by separate optimization of node-specific likelihoods. The vector w of scalar coefficients $\{w_{st} \in \mathbb{R}, (s,t) \in \mathcal{E}\}$ will be hereafter referred to as composition rule. Each w_{st} represents the relative impact of $u_{st}(\theta; x)$ on the overall CL score (3), and two variables X_s and X_t are regarded as independent if $w_{st} = 0$, meaning that the associated pairwise model (2) does not contain information about θ.

A sparse composition rule $w_\lambda(\theta)$ is derived by minimizing the penalized score divergence

$$Q_\lambda(\theta, w) = \frac{1}{2} E \left\| u^{ML}(\theta; X) - u(\theta; w, X) \right\|_2^2 + \lambda \sum_{(s,t)\in\mathcal{E}} |w_{st}| \,, \tag{4}$$

where $u^{ML}(\theta, x) = \partial f(x;\theta)/\partial\theta$ is the maximum likelihood score, $\|\cdot\|_2$ denotes the ℓ_2-norm, and $\lambda \geq 0$ is a given constant. The first term in (4) is a statistical accuracy objective, while the second is a complexity penalty discouraging overly complicated graphs. The geometric properties of the ℓ_1-penalty guarantee that several elements in $w_\lambda(\theta)$, corresponding to specific edges in $\mathcal{E}$, are exactly zero for sufficiently large values of λ. This fact can be viewed as a truncation of the classical pairwise likelihood estimating equation in (3).

One issue related to the direct minimization of $Q_\lambda(\theta, w)$ is the presence of the intractable score $u^{ML}(\theta; x)$ and of the theoretical expectation based on the unknown true parameter. Let $U(\theta, x)$ be the $q \times p$ matrix with column vectors $\{u_{st}(\theta, x), (s,t) \in \mathcal{E}\}$ and define the $p \times p$ matrix $S(\theta, x) = U(\theta, x)^\top U(\theta, x)$. Exploiting the second Bartlett identity as done in [3, p. 76], formula (4) becomes

$$Q_\lambda(\theta, w) = w^\top E\left\{S(\theta; X)\right\} w - \operatorname{diag}\left[E\left\{S(\theta; X)\right\}\right] w + \lambda \sum_{(s,t)\in\mathcal{E}} |w_{st}| + c(\theta) \,,$$

being the quantity $c(\theta)$ unrelated to w. The final composition rule $\hat{w}_\lambda$ is defined as the minimizer of the data-driven criterion $\widehat{Q}_\lambda(w)$, where expectations are replaced by sample averages and evaluation at $\theta = \theta_{\text{ind}}$ with $\eta_{st}(\theta_{\text{ind}}) = 0$, $\forall (s,t) \in \mathcal{E}$, allows for the effective detection of pairwise dependencies. Then, solving equation (3) with $w = \hat{w}_\lambda$ leads to the TPL estimator $\hat{\theta}_\lambda = \hat{\theta}(\hat{w}_\lambda)$.

3 Example: Gaussian random field

Consider n independent observations on $X \sim N_d(0, \Sigma)$, where Σ is a $d \times d$ covariance matrix with diagonal elements $\sigma_1^2 = \cdots = \sigma_d^2 = 1$ and unknown off-diagonal entries $\{\theta_{st}, (s,t) \in \mathcal{E}\}$. The TPL objective $\widehat{Q}_\lambda(\theta)$ can be constructed based on pairwise scores

$$u_{st}(\theta_{st}; x_s, x_t) = \frac{\partial}{\partial\theta_{st}} \left\{ -\frac{x_s^2 + x_t^2}{2(1-\theta_{st}^2)} + \frac{\theta_{st}}{1-\theta_{st}^2} x_s x_t - \frac{1}{2}\log(1-\theta_{st}^2) \right\} \,, \tag{5}$$

since (X_s, X_t) follows a bivariate normal distribution $N_2(0, \Sigma_{st})$ with unit variances and correlation $\theta_{st} \in (-1, 1)$.

To investigate the TPL performance in terms of discrimination among edges, we generate 250 Monte Carlo samples of size $n \in \{150, 250\}$ from $N_d(0, \Sigma)$ with $d \in \{25, 50\}$ and Σ defined by setting $\theta_{st} = \theta > 0$ for $1 \leq s < t \leq 10$ and $\theta_{st} = 0$

otherwise. This choice implies a sparse graph configuration with only $p_1 = 45$ non-zero edges out of $p \in \{300, 1225\}$. Monte Carlo estimates for the false discovery proportion (FDP) and true positive proportion (TPP) of TPL are computed along the solution path for $\hat{w}_\lambda$ as

$$\text{FDP}(\lambda) = \frac{V(\lambda)}{\max\{|\{(s,t) : \hat{w}_{st}(\lambda) \neq 0\}|, 1\}} \quad \text{and} \quad \text{TPP}(\lambda) = \frac{T(\lambda)}{\max\{p_1, 1\}},$$

where $V(\lambda) = |(s,t) : \hat{w}_{st}(\lambda) \neq 0 \text{ and } \theta_{st} = 0|$ and $T(\lambda) = |(s,t) : \hat{w}_{st}(\lambda) \neq 0 \text{ and } \theta_{st} \neq 0|$ denote the number of false and true discoveries for fixed λ, respectively. Minimization of $\widehat{Q}_\lambda(w)$ is implemented by the function `sparsepca` in the R package `elasticnet` [8], over a grid of 50 values for λ. In this example, the value θ_{ind} at which the empirical criterion is calculated corresponds to $\theta_{st} = 0, \forall (s,t) \in \mathcal{E}$. Table 1 reports estimates of FDP and TPP when $d = 50$ and the average number of selected pairs across Monte Carlo samples is close to the truth. Once the graph dimension is correctly inferred by a suitable choice of λ not discussed here, the ability of TPL to detect the pairwise scores corresponding to non-zero edges increases with n and with the size of the correlation coefficient θ. Figure 1 shows the estimated TPP when the first irrelevant score enters the composition rule (left) and the estimated FDP the first time all meaningful scores are discovered (right) under the least favorable setting with $\theta = 0.3$ and $n = 150$. If d decreases from 50 to 25, the empirical distribution of the TPP shifts to the right, whilst that of the FDP moves towards 0.

Table 1: True positive proportion (TPP) and false discovery proportion (FDP) of TPL for the Gaussian random field model with $d = 50$. Estimates at each value of λ are based on 250 Monte Carlo samples of size n.

		$\theta = 0.3$		$\theta = 0.6$		$\theta = 0.9$	
n	λ	TPP	FDP	TPP	FDP	TPP	FDP
150	2.118	0.545	0.015	0.986	0.000	1.000	0.000
	1.672	0.633	0.026	0.987	0.001	1.000	0.000
	1.225	0.734	0.060	0.990	0.003	1.000	0.000
	0.779	0.844	0.149	0.993	0.020	1.000	0.001
250	2.118	0.933	0.017	0.995	0.000	1.000	0.000
	1.672	0.957	0.039	0.996	0.002	1.000	0.000
	1.225	0.978	0.092	0.998	0.009	1.000	0.000
	0.779	0.990	0.230	0.999	0.048	1.000	0.001

4 Final remarks

A novel technique for simultaneous estimation and selection of undirected graphical models within the CL framework has been illustrated. The empirical evidence presented confirms the validity of the approach when dealing with sparse Gaussian random fields. Type I error (FDP) and power (TPP) of the TPL strategy seem to

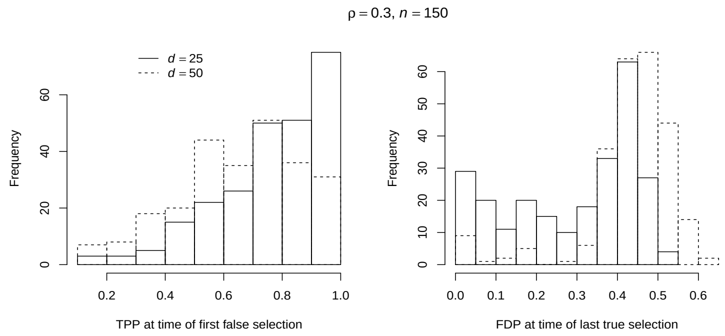


Figure 1: Empirical distributions of the true positive proportion (TPP) when the first false score enters the composition rule (left) and of the false discovery proportion (FDP) the first time all true scores are discovered (right). Results are based on 250 Monte Carlo samples.

behave particularly well when the correlation magnitude among nodes is large to moderate, even in the presence of modest samples.

For future research, it would be valuable to develop theoretical conditions on the size of λ ensuring model selection consistency of $\hat{w}_\lambda$ in a setting where both n and p diverge. The performance of the associated TPL estimator $\hat{\theta}_\lambda$ could be then compared with that of an oracle estimator based on the true non-zero edges. Derivation of empirical rules for the choice of λ and extensions to general M-estimating equations represent other avenues of further work.

Bibliography

[1] H. Höfling and R. Tibshirani. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10:883–906, 2009.

[2] J. D. Lee and T. J. Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24:230–253, 2015.

[3] B. G. Lindsay, G. Y. Yi, and J. Sun. Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21:71–105, 2011.

[4] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

[5] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38:1287–1319, 2010.

[6] C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

[7] L. Xue, H. Zou, T. Cai, et al. Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics*, 40:1403–1429, 2012.

[8] H. Zou and T. Hastie. *elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA*, 2018. R package version 1.1.1.

Filtering for stochastic heat equation with fractional noise

Vít Kubelka[1*] **and Bohdan Maslowski**[1]

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

Abstract: Linear filtering problem for infinite-dimensional Gaussian processes with finite - dimensional observation process is studied. Integral equations for the filter and for covariance of the error are derived. General results are applied to a stochastic evolution equation driven by cylindrical fractional Brownian motion observed at finitely many points on an unbounded domain.

Keywords: Kalman - Bucy filter, stochastic evolution equations, Gaussian processes, heat equation

AMS subject classification: 60H05, 60H15

Introduction

This paper deals with the linear filtering problem for infinite - dimensional Gaussian processes with finite - dimensional observation. In general, by the filtering problem we understand the case when there is a process (called signal) which is not observable and we would like to find an optimal estimate of this process. Such an estimate is based on observation of another process (called observation process) which is typically some noisy perturbation of a functional of the signal.
The aim of this paper is to give an overview of [3]. These results are further extended for a signal defined on a more general domain which enables to state more specific integral equations for the filter. For a short survey of this field see the introductory part of [3].
The paper is divided into two Sections. Section 1 contains an overview of the main results of the work [3]. In the second Section, the general results are applied to a stochastic heat equation driven by cylindrical fractional Brownian motion on an unbounded domain which slightly extends Example 3.4 in [3].
The space of bounded linear operators mapping a Banach space X to a Banach space Y is denoted as $\mathcal{L}(X,Y)$, $\mathcal{L}(X) := \mathcal{L}(X,X)$. The space of Hilbert-Schmidt operators on a Hilbert space H is denoted by $\mathcal{L}_2(H)$.

[*] Corresponding author: kubelka@karlin.mff.cuni.cz

1 Solution of the filtering problem

Consider separable Hilbert spaces H and V, where $H = (H, \langle\cdot,\cdot\rangle_H, \|\cdot\|_H)$, $V = (V, \langle\cdot,\cdot\rangle_V, \|\cdot\|_V)$, such that $V \subset H$, V is dense in H and identifying H with the dual H^* the embeddings

$$V \hookrightarrow H = H^* \hookrightarrow V^*$$

are continuous and dense. The duality pairing between V and V^* is defined by the inner product on H, that is $\langle u, v\rangle_{V,V^*} = \langle u, v\rangle_H$ for $u \in V \subset H$ and $v \in H \subset V^*$. Such construction is called rigged Hilbert space or Gelfand triple and it enables to work with pointwise observations of the signal driven by a stochastic partial differential equation. The larger space H (which is usually a Lebesgue space on a domain) is suitable for the definition of the noise term and the stochastic integral of the signal, while the signal itself lives in the smaller space V which can be contained in the space of continuous functions (for which values at given points are well defined).

Consider the signal $\theta = \{\theta_t, t \in [0,T]\}$ that is a centered Gaussian mean - square continuous measurable process in V defined on stochastic basis $(\Omega, F, P, (F_t))$. Let $\xi = \{\xi_t, t \in [0,T]\}$ denote an $\mathbb{R}^n$ - valued observation process given as

$$\xi_t = \int_0^t A(s)\theta_s \,\mathrm{d}s + W_t, \tag{1}$$

where A is a bounded strongly measurable function from $[0,T]$ to $L(V, \mathbb{R}^n)$. Here $W = \{W_t, t \in [0,T]\}$ is a standard $\mathbb{R}^n$ - valued (F_t) - Wiener process independent of the signal θ.

Further, assume that for each $t \in [0,T]$ operator $A(t)$ takes the form

$$A(t)b = (\langle b, A_1(t)\rangle_{V,V^*}, \ldots, \langle b, A_n(t)\rangle_{V,V^*})^T, \quad b \in V,$$

where $A_1(t), \ldots, A_n(t) \in V^*$. Note that the dual operator $A^*(t)$: $\mathbb{R}^n \to V^*$ takes the form $A^*(t)z = \sum_{i=1}^n z_i A_i(t)$ for all $z \in \mathbb{R}^n$.

In the paper the optimal estimate $\widehat{\theta}_t$ called filter, defined as

$$\widehat{\theta}_t = \mathbb{E}[\theta_t | F_t^\xi]$$

is studied. Here $(F_t^\xi)_{t\in[0,T]}$ is the filtration generated by the observation process ξ. It is well known that the above defined conditional expectation is the optimal estimate (in the mean-square sense) of the signal θ given the observation ξ.

For arbitrary $x, y \in V$ we define tensor product $x \circ y \in \mathcal{L}(V^*, V)$, $(x \circ y)v = x\langle y,v\rangle_{V,V^*}$, $v \in V^*$. Let $K^\theta(t,s) = \mathbb{E}[\theta_t \circ \theta_s]$, $t, s \in [0,T]$ be the covariance operator of the process $\theta_t, t \in [0,T]\}$. Notice that the mean - square continuity of the process implies that the mapping $K^\theta : [0,T]^2 \to \mathcal{L}(V^*, V)$ is strongly continuous and bounded.

The following theorem provides the tool to obtain the filter.

Theorem 1. *Let $\Lambda = \{(t,s) \in [0,T]^2; 0 \leq s \leq t \leq T\}$. The filter $\widehat{\theta}$ satisfies the stochastic integral equation*

$$\widehat{\theta}_t = \int_0^t \Phi(t,s)A^*(s)\,\mathrm{d}\xi_s - \int_0^t \Phi(t,s)A^*(s)A(s)\widehat{\theta}_s\,\mathrm{d}s, \quad t \in [0,T], \tag{2}$$

where function $\Phi\colon \Lambda \to \mathcal{L}(V^*,V)$ defined as $\Phi(t,s) = \mathbb{E}[\theta_t \circ (\theta_s - \widehat{\theta}_s)]$, $(s,t) \in \Lambda$, is strongly continuous and is the unique solution to the integral equation

$$\Phi(t,s) = K^{\theta}(t,s) - \sum_{j=1}^{n} \int_0^s (\Phi(t,r)A_j(r)) \circ (\Phi(s,r)A_j(r))\,\mathrm{d}r. \tag{3}$$

Moreover, for all $t \in [0,T]$, $\Phi(t,t)$ is the covariance of the estimation error at time $t \in [0,T]$, that is,

$$\Phi(t,t) = \mathbb{E}\left[(\theta_t - \widehat{\theta}_t) \circ (\theta_t - \widehat{\theta}_t)\right] \tag{4}$$

holds.

Proof. See Theorem 1.1 in [3]. For uniqueness see Theorem 2.1 in [3]. □

Remark 1. *Linear Stochastic Evolution Equation driven by Wiener Process*
Suppose that the signal is given as an H-valued solution of a stochastic evolution equation driven by H-valued cylindrical Wiener process and the observation $\xi = \{\xi_t, t \in [0,T]\}$ is given by the equation (1) where $V = H$.
In this case, under some additional assumptions (cf. [3], Example 3.1), the equations (2) and (3) simplify to the infinte - dimensional analogue of standard Kalman - Bucy filter as shown in Theorem 3.2 in [3].

Remark 2. Note that, in general, if the signal $\theta = \{\theta_t, t \in [0,T]\}$ takes its values in the Hilbert space H, the family $(A(s))_{s\in[0,T]}$ of observation operators $H \to \mathbb{R}^n$ can be characterised by n functionals from the dual space of H and the concept of rigged Hilbert space is not needed. In this case using adjoint operators A^* and Φ^* the equation (3) takes the form (cf. [3], Remark 2.1)

$$\Phi(t,s) = K^{\theta}(t,s) - \int_0^s \Phi(t,r)A^*(r)A(r)\Phi^*(s,r)\,\mathrm{d}r. \tag{5}$$

2 An example of stochastic heat equation

The results of previous section are now applied to the case when signal is governed by the following parabolic equation

$$\partial_t u = \Delta u + \eta \tag{6}$$

on $[0,T] \times \mathcal{D}$ with initial condition $u(0,\cdot) = 0$ and with the Dirichlet boundary condition

$$u\Big|_{[0,T] \times \partial\mathcal{D}} = 0,$$

where Δ is a Laplace operator. The domain $\mathcal{D} \subset \mathbb{R}^d$ is open and has the C^m -extension property (cf. [2]). It is well known that this property is satisfied, for instance, if $\mathcal{D} = \mathbb{R}^d$, $\mathcal{D} = (\mathbb{R}^d)_+$ or if $\mathcal{D}$ is bounded with Lipschitz boundary. The noise η is fractional in time and correlated in space. This system can be reformulated as the stochastic evolution equation

$$\mathrm{d}\theta_t = \Delta\theta_s \,\mathrm{d}s + G\,\mathrm{d}B_t, \quad \theta_0 = 0, \tag{7}$$

where $\{B_t, t \in [0,T]\}$ is a cylindrical fractional Brownian motion with Hurst parameter $h \in (0,1)$ on $\mathcal{D}$. The equation is considered in the Hilbert space $H = L^2(\mathcal{D})$. Due to the boundary condition the Laplacian generates an analytic semigroup $(S(t), t \geq 0)$ on H. The noise covariance G is supposed to be Hilbert-Schmidt on H. In virtue of [1] the equation (7) has a unique mild solution $\theta = \{\theta_t, t \in [0,T]\}$ which is continuous in time in the space $H_\delta = Dom\left((\beta - \Delta)^\delta\right)$ for a fixed β large enough and $0 \leq \delta < h$. If, moreover,

$$h > \frac{d}{4}$$

then taking $\delta \in (1/4, h)$ we have that the space $V := H_\delta$ is continuously embedded into the space of continuous functions $C(\mathcal{D})$ (cf. [2], Theorem 1.6.1), i.e.

$$V = H_\delta \hookrightarrow C(\mathcal{D}) \hookrightarrow L^2(\mathcal{D}) = H.$$

Hence it make sense to consider observation operator defined as

$$A\theta_t = (\theta_t(z_1), \ldots, \theta_t(z_n)), \tag{8}$$

where $z_1, \ldots, z_n \in \mathcal{D}$, which corresponds to pointwise observation of the signal process θ at these points. To find the filter we can use the rigged Hilbert space setting of the Theorem 1 which now reads (cf. [3], Corollary 3.5 and the note following the proof of the Corollary 3.5):

Theorem 2. *Consider the observation process $\xi = \{\xi_t, t \in [0,T]\}$ given by (1) with operator $A(t) = A$ defined by (8). Then the filter $\widehat{\theta}$ satisfies stochastic integral equation*

$$\widehat{\theta}_t = \sum_{j=1}^{n} \int_0^t \Phi_{z_j}(t,s)\,\mathrm{d}\xi_s^j - \sum_{j=1}^{n} \int_0^t \Phi_{z_j}(t,s)\widehat{\theta}_s(z_j)\,\mathrm{d}s, \quad t \in [0,T], \tag{9}$$

where $\Phi_{z_i}\colon \Lambda \to C(\mathcal{D})$ is defined as $\Phi_{z_i}(t,s) = \mathbb{E}[(\theta_s - \widehat{\theta}_s)(z_i)\theta_t]$ for all $(t,s) \in \Lambda$, $i = 1, \ldots, n$ and integral equation

$$\Phi_{z_i}(t,s) = \mathbb{E}\left[\theta_s(z_i)\theta_t\right] - \sum_{j=1}^{n} \int_0^s \Phi_{z_j}(s,r)(z_i)\Phi_{z_j}(t,r)\,\mathrm{d}r, \quad i = 1, \ldots, n \tag{10}$$

is satisfied.

Further, we can specify the covariances $\mathbb{E}\left[\theta_s(z_i)\theta_t\right]$ that appear in the equation (10). Suppose, for simplicity, that $h > 1/2$ and $n = 1$ (i.e. the process $\theta = \{\theta_t, t \in [0,T]\}$ is observed at a single point $z_1 \in \mathcal{D}$). Since the noise term G is Hilbert-Schmidt it may be expressed as

$$[G(f)](\xi) = \int_{\mathcal{D}} k(\xi,\eta) f(\eta)\, \mathrm{d}\eta, \quad f \in H = L^2(\mathcal{D}),$$

where $k \in L^2(\mathcal{D} \times \mathcal{D})$. It is also well known that the semigroup $(S(t), t \in \mathbb{R})$ may be represented by a Green function $g : [0,T] \times \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, that is,

$$[S(t)(f)](\xi) = \int_{\mathcal{D}} g(t,\xi,\eta) f(\eta)\, \mathrm{d}\eta, \quad f \in H,\ t > 0.$$

For example if $\mathcal{D} = (0,\infty)$ (i.e. Δ is the Dirichlet Laplacian in $L^2(0,\infty)$), we have

$$g(t,\xi,\eta) = \frac{1}{\sqrt{4\pi t}} \left(e^{-\frac{(\xi-\eta)^2}{4t}} - e^{-\frac{(\xi+\eta)^2}{4t}} \right), \quad \xi, \eta \geq 0,$$

and if $\mathcal{D} = \mathbb{R}$, g is the Gaussian kernel

$$g(t,\xi,\eta) = \frac{1}{\sqrt{4\pi t}} e^{-\frac{(\xi-\eta)^2}{4t}}, \quad \xi, \eta \geq 0.$$

The composition $S(t)G$ may be written as

$$[S(t)G(f)](\xi) = \int_{\mathcal{D}} \widetilde{g}(t,\xi,\eta) f(\eta)\, \mathrm{d}\eta, \quad f \in H,\ t > 0,$$

where the composition kernel $\widetilde{g}$ is given by

$$\widetilde{g}(t,\xi,\eta) = \int_{\mathcal{D}} g(t,\xi,\lambda) k(\lambda,\eta)\, \mathrm{d}\lambda.$$

Now it is standard to compute the covariance

$$\mathbb{E}\left[\theta_s(z_1)\theta_t\right](\eta) = \int_0^s \int_0^t \phi_h(\lambda,r) \int_{\mathcal{D}} \widetilde{g}(s-r,z_1,\xi)\widetilde{g}(t-\lambda,\eta,\xi)\, \mathrm{d}\xi\, \mathrm{d}\lambda\, \mathrm{d}r,$$

for $(s,t) \in \Lambda$, $\eta \in \mathcal{D}$, where $\phi_h(\lambda,r) = h(2h-1) \mid \lambda - r \mid^{2h-2}$ and $(\eta,\lambda) \in \Lambda$.

Acknowledgements: This research was partially suported by GAUK Grant no. 980218, Czech Science Foundation (GAČR) Grant no. 19-07140S and by the SVV Grant No. 260454.

Bibliography

[1] P. Čoupek, B. Maslowski, and M. Ondreját. L^P-valued stochastic convolution integral driven by Volterra noise. *Stochastics and Dynamics*, 18(06), 1850048, 2018.

[2] D. Henry. *Geometric Theory of Semilinear Parabolic Equations.* Lecture Notes in Mathematics, 840, Springer, 1981.

[3] V. Kubelka and B. Maslowski. Filtering of Gaussian processes in Hilbert spaces. http://arxiv.org/abs/1903.11464, 2019.

Selection consistency of two-step selection method for misspecified binary model

Mariusz Kubkowski[1,2*] **and Jan Mielniczuk**[2,1]

[1] *Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland*[2] *Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland*

Abstract: We prove selection consistency of two-stage selection procedure in case of misspecified binary regression model. In the discussed framework number of predictors may depend exponentially on a sample size.

Keywords: misspecification, binary regression, consistent selection, subgaussianity, Generalized Information Criterion

AMS subject classification: Primary 62J05, 62J07, secondary 90C25

1 Introduction

We consider two-stage selection method of random predictors $\mathbf{X} \in \mathbb{R}^p$ when the underlying binary regression model:

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = q(\mathbf{x}) \tag{1}$$

is misspecified. We discuss a problem of finding consistent estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ minimizes risk function:

$$R(\mathbf{b}) = \mathbb{E}\rho(\mathbf{b}^T\mathbf{X}, Y)$$

for $\mathbf{b} \in \mathbb{R}^p$ and $\rho\colon\ \mathbb{R} \times \{0, 1\} \to \mathbb{R}$ is a given convex function. We call model (1) misspecified, when $q(\mathbf{x}) = q(\boldsymbol{\beta}^T\mathbf{x})$ and corresponding minus log-likelihood is not equal ρ. In this case an aim of selection is to find $\hat{\boldsymbol{\beta}}$ the support of which recovers the support of $\boldsymbol{\beta}^*$ with high probability.
Recent advancements in data gathering allow for much larger number of observations n to be collected and to much larger number of variables p to be measured. We very often encounter a situation, where $p \gg n$ and consistent estimators of $\boldsymbol{\beta}^*$

[*] Corresponding author: m.kubkowski@mini.pw.edu.pl

do not exist, unless we impose additional model constraints. Assumptions typically concern structure of covariance matrix and sparsity constraints for $\boldsymbol{\beta}^*$.
The proposed procedure of selection consists of screening and ordering predictors by Lasso and then selecting a subset of predictors which minimizes Generalized Information Criterion on the nested family pertaining to them.
In the contribution we discuss sufficient conditions (proved in [3] on the parameters of the method and distribution of $(\mathbf{X}, Y)$ under which the above procedure is consistent.

2 Properties of Lasso estimator in misspecified binary model

We consider an i.i.d. random sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n) \overset{d}{=} (\mathbf{X}, Y) \in \mathbb{R}^{p_n} \times \{0,1\}$ where $p = p_n$. We assume that coordinates X_{ij} of $\mathbf{X}_i$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p_n$ are subgaussian $Subg(\sigma_{jn}^2)$ with $\sigma_{jn} > 0$. For future reference let $s_n = \max_j \sigma_{jn}$ and assume that $\limsup_n s_n^2 < \infty$. Empirical risk is defined as (where $\mathbf{b} \in \mathbb{R}^{p_n}$):

$$R_n(\mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \rho(\mathbf{b}^T \mathbf{X}_i, Y_i). \tag{2}$$

We are interested in properties of Lasso estimator of $\boldsymbol{\beta}^*$ defined as:

$$\hat{\boldsymbol{\beta}}_L = \underset{\mathbf{b} \in \mathbb{R}^{p_n}}{\arg\min}\, R_n(\mathbf{b}) + \lambda ||\mathbf{b}||_1, \tag{3}$$

where $\lambda > 0$. We assume that $\rho(\cdot, y)$ is convex function which is bounded from below by $m \in \mathbb{R}$. These two properties assure that $\hat{\boldsymbol{\beta}}_L$ exists (see [3]). Properties of uniqueness and sparsity of $\hat{\boldsymbol{\beta}}_L$ are discussed in [6] and [5]. Assumptions on distribution of $(\mathbf{X}, Y)$ are not needed to ensure uniqueness of $\hat{\boldsymbol{\beta}}_L$ in any of the proofs in this section. We only assume that $\boldsymbol{\beta}^*$ exists and is unique in order to obtain Lasso consistency. We are interested not only in estimation of $\boldsymbol{\beta}^*$, but also in estimation of the set of active predictors:

$$s^* = \operatorname{supp} \boldsymbol{\beta}^* = \{j \in \{1, \ldots, p_n\} : \ \beta_j^* \neq 0\}. \tag{4}$$

We introduce the following notation:

$$B_1(r) = \{\boldsymbol{\Delta} \in \mathbb{R}^{p_n} : \ ||\boldsymbol{\Delta}||_1 \leq r\}, \tag{5}$$

$$W(\mathbf{b}) = R(\mathbf{b}) - R(\boldsymbol{\beta}^*), \tag{6}$$

$$W_n(\mathbf{b}) = R_n(\mathbf{b}) - R_n(\boldsymbol{\beta}^*), \tag{7}$$

$$S(r) = \sup_{\mathbf{b} : \mathbf{b} - \boldsymbol{\beta}^* \in B_1(r)} |W(\mathbf{b}) - W_n(\mathbf{b})|, \tag{8}$$

$$\beta_{min}^* = \min_{i \in s^*} |\beta_i^*|. \tag{9}$$

The main Theorem in this section is Theorem 1 below. Idea of the proof is based on fact that if $S(r)$ defined in (8) is sufficiently small, then $\hat{\boldsymbol{\beta}}_L$ lies in a ball $\{\boldsymbol{\Delta} \in \mathbb{R}^{p_n}: \ ||\boldsymbol{\Delta} - \boldsymbol{\beta}^*||_1 \le r\}$ (see Lemma 1). In Lemma 2 we prove a tail inequality for $S(r)$, from which Theorem 1 follows.

Before we proceed to the statement of Lemma 1, we define cone $\tilde{\mathcal{C}}_\varepsilon$ and restricted minimal eigenvalue $\kappa_{\mathbf{H}}(\varepsilon)$ on that cone:

$$\tilde{\mathcal{C}}_\varepsilon = \{\boldsymbol{\Delta} \in \mathbb{R}^{p_n}: \ ||\boldsymbol{\Delta}_{s^{*c}}||_1 \le (3+\varepsilon)||\boldsymbol{\Delta}_{s^*}||_1\}, \tag{10}$$

$$\kappa_{\mathbf{H}}(\varepsilon) = \inf_{\boldsymbol{\Delta} \in \tilde{\mathcal{C}}_\varepsilon} \frac{\boldsymbol{\Delta}^T \mathbf{H} \boldsymbol{\Delta}}{\boldsymbol{\Delta}^T \boldsymbol{\Delta}}, \tag{11}$$

where $\varepsilon > 0$ and $\mathbf{H} \in \mathbb{R}^{p_n \times p_n}$ is nonnegative definite matrix. The following margin condition is also required:

(MC) There exist $\vartheta, \varepsilon, \delta > 0$ and non-negative definite matrix $\mathbf{H} \in \mathbb{R}^{p_n \times p_n}$ such that for all $\mathbf{b} \in \mathbb{R}^{p_n}$ with $\mathbf{b} - \boldsymbol{\beta}^* \in \tilde{\mathcal{C}}_\varepsilon \cap B_1(\delta)$ we have:

$$R(\mathbf{b}) - R(\boldsymbol{\beta}^*) \ge \frac{\vartheta}{2}(\mathbf{b} - \boldsymbol{\beta}^*)^T \mathbf{H} (\mathbf{b} - \boldsymbol{\beta}^*).$$

In (MC) we expect that $\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^* \in \tilde{\mathcal{C}}_\varepsilon \cap B_1(\delta)$ with high probability for some $\varepsilon, \delta > 0$ and (MC) is satisfied when ρ is linear or logistic loss (see [3]), $\mathbf{H} = D^2 R(\boldsymbol{\beta}^*)$ is hessian matrix of risk function and X_{ij} are subgaussian (in the case of linear loss) or bounded variables (in the case of logistic loss).

Lemma 1. *Let $\rho(\cdot, y)$ be convex function for all y. Assume that $\lambda > 0$. Moreover, assume margin condition (MC) with $\vartheta, \epsilon, \delta > 0$ and $p_n \times p_n$ matrix $\boldsymbol{H} \ge 0$. If for some $r \in (0, \delta]$ we have $S(r) \le \bar{C}\lambda r$ and $2|s^*|\lambda \le \kappa_{\boldsymbol{H}}(\varepsilon)\vartheta\tilde{C}r$, where $\bar{C} = \varepsilon/(8+2\varepsilon)$ and $\tilde{C} = 2/(4+\varepsilon)$, then $||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 \le r$.*

Lemma 2. *Let $\rho(\cdot, y)$ be convex function for all y and satisfy Lipschitz condition for some $L > 0$ and for all b_1, b_2, y: $|\rho(b_1, y) - \rho(b_2, y)| \le L|b_1 - b_2|$. Assume that $X_{ij} \sim Subg(\sigma_{jn}^2)$ for all i, j. Then for $r, t > 0$:*

$$\mathbb{P}(S(r) > t) \le \frac{14 L r s_n \sqrt{\log(p_n \vee 2)}}{t\sqrt{n}}.$$

In the Theorem below we consider inequality (12), from which it follows that λ and $|s^*|$ cannot be too large constants and $\kappa_{\mathbf{H}}(\varepsilon)$ and β^*_{min}. Therefore this inequality imposes sparsity, covariance structure and appropriate signal strength in this model. Remark 3 gives asymptotic conditions when the assumptions of this theorem are satisfied.

Theorem 1. *Let $\rho(\cdot, y)$ be convex function for all y and satisfy Lipschitz condition with $L > 0$. Assume that $X_{ij} \sim Subg(\sigma_{jn}^2)$ for all i, j, margin condition (MC) is satisfied for $\varepsilon, \delta, \vartheta > 0$, non-negative definite matrix $\boldsymbol{H} \in \mathbb{R}^{p_n \times p_n}$ and let*

$$\frac{2|s^*|\lambda}{\vartheta \kappa_{\boldsymbol{H}}(\varepsilon)} \le \tilde{C} \min\left\{\frac{\beta^*_{min}}{2}, \delta\right\}, \tag{12}$$

where $\tilde{C} = 2/(4+\varepsilon)$. Then:

$$\mathbb{P}\left(||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 \leq \frac{\beta^*_{min}}{2}\right) \geq 1 - \frac{14(8+2\varepsilon)Ls_n\sqrt{\log(p_n \vee 2)}}{\varepsilon\lambda\sqrt{n}}.$$

Proposition 1. *(Separation property) If assumptions of Theorem 1 are satisfied, $\log p_n = o(\lambda^2 n)$ and $\kappa_{\boldsymbol{H}}(\varepsilon) > d$ for some $d, \varepsilon > 0$ for large n, $|s^*|\lambda = o(\min\{\beta^*_{min}, 1\})$, then*

$$\mathbb{P}\left(||\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}^*||_1 \leq \frac{\beta^*_{min}}{2}\right) \to 1.$$

Moreover

$$\mathbb{P}\left(\max_{i \in s^{*c}} |\hat{\beta}_{L,i}| \leq \min_{i \in s^*} |\hat{\beta}_{L,i}|\right) \to 1.$$

3 GIC minimization

Consider an arbitrary family $\mathcal{M} \subseteq 2^{\{1,\ldots,p_n\}}$ of models (which may be data-dependent) such that $s^* \in \mathcal{M}, \forall w \in \mathcal{M}: \ |w| \leq k_n$ a.e. and $k_n \in \mathbb{N}_+$ is some sequence. We define Generalized Information Criterion (GIC) as:

$$GIC(w) = n \min_{\mathbf{b} \in \mathbb{R}^{p_n}: \ \mathbf{b}_{w^c} = \mathbf{0}_{|w^c|}} R_n(\mathbf{b}) + a_n|w|, \tag{13}$$

where $a_n > 0$ is some penalty. Moreover, we consider condition similar to (MC) for $w \subseteq \{1, \ldots, p_n\}$ and some $\epsilon, \theta > 0$:

$C_\epsilon(w)$: $R(\mathbf{b}) - R(\boldsymbol{\beta}^*) \geq \theta||\mathbf{b} - \boldsymbol{\beta}^*||_2^2$ for all $\mathbf{b} \in \mathbb{R}^{p_n}$ such that $\operatorname{supp} \mathbf{b} \subseteq w$ and $||\mathbf{b} - \boldsymbol{\beta}^*||_2 \leq \epsilon$.

Proposition below yields consistency of GIC on supersets of s^*.

Proposition 2. *Assume that $\rho(\cdot, y)$ is convex, Lipschitz function with constant $L > 0$, $X_{ij} \sim Subg(\sigma^2_{jn})$, condition $C_\epsilon(w)$ holds for some $\epsilon, \theta > 0$ and for every $w \subseteq \{1, \ldots, p_n\}$ such that $|w| \leq k_n$, $k_n \ln(p_n \vee 2) = o(n)$ and $k_n \ln(p_n \vee 2) = o(a_n)$. Then we have*

$$\mathbb{P}(\min_{w \in \mathcal{M}: s^* \subset w} GIC(w) \leq GIC(s^*)) \to 0.$$

The most restrictive condition of Proposition 2 is $k_n \ln(p_n \vee 2) = o(a_n)$. We note that in the case when $p_n \geq n$ and $k_n = d$, EBIC penalty equal $\log n + 2\gamma \log p_n$ corresponds to the borderline of this condition.
Proposition below yields consistency of GIC on subsets of s^*.

Proposition 3. *Assume that loss $\rho(\cdot, y)$ is convex, Lipschitz function with constant $L > 0$, $X_{ij} \sim Subg(\sigma^2_{jn})$, condition $C_\epsilon(s^*)$ holds for some $\epsilon, \theta > 0$ and $a_n|s^*| = o(n\min\{1, \beta^*_{min}\}^2)$, then*

$$\mathbb{P}(\min_{w \in \mathcal{M}: w \subset s^*} GIC(w) \leq GIC(s^*)) \to 0.$$

4 Selection consistency of SS procedure

SS (Screening and Selection) procedure, being a modification of SS procedure in [4], is the following:

1. Find $\hat{\boldsymbol{\beta}}_L$ for some $\lambda > 0$, then sort all nonzero coefficients of $\hat{\boldsymbol{\beta}}_L$: $|\hat{\beta}_{L,j_1}| \geq \ldots \geq |\hat{\beta}_{L,j_k}| > 0$.
2. Define $\mathcal{M}_{SS} = \{\emptyset, \{j_1\}, \{j_1, j_2\}, \ldots, \{j_1, j_2, \ldots, j_k\}\}$.
3. Find $\hat{s} = \underset{w \in \mathcal{M}_{SS}}{\arg\min}\, GIC(w)$.

Proposition 4 and Remark 3 describe the situations when SS procedure works.

Proposition 4. *Assume that $\rho(\cdot, y)$ is convex, Lipschitz function with constant $L > 0$, $X_{ij} \sim Subg(\sigma_{jn}^2)$ and $\boldsymbol{\beta}^*$ exists and is unique. If $k_n \in \mathbb{N}_+$ is some sequence, margin condition (MC) is satisfied for some $\vartheta, \delta, \varepsilon > 0$, condition $C_\epsilon(w)$ holds for some $\epsilon, \theta > 0$ and for every $w \subseteq \{1, \ldots, p_n\}$ such that $|w| \leq k_n$, $\mathcal{M}_{SS}$ is hierarchical family constructed in the step 2 of SS procedure. If $\mathbb{P}(\forall w \in \mathcal{M}_{SS} : |w| \leq k_n) \to 1$, $|s^*| \leq k_n$, $\liminf_n \kappa_{\boldsymbol{H}}(\varepsilon) > 0$ for some $\varepsilon > 0$ and $\boldsymbol{H} \geq 0 \in \mathbb{R}^{p_n \times p_n}$, $\log(p_n) = o(n\lambda^2)$, $k_n\lambda = o(\min\{\beta_{min}^*, 1\})$, $k_n \log p_n = o(n)$, $k_n \log p_n = o(a_n)$, $a_n k_n = o(n \min\{\beta_{min}^*, 1\}^2)$, then we have:*

$$\mathbb{P}(\hat{s} = s^*) \to 1.$$

Remark 3. If $p_n = O(e^{cn^\gamma})$ for some $c > 0$, $\gamma \in (0, 1/2)$, $\xi \in (0, 0.5 - \gamma)$, $u \in (0, 0.5 - \gamma - \xi)$, $k_n = O(n^\xi)$, $\lambda = C_n\sqrt{\log(p_n)/n}$, $C_n = O(n^u)$, $C_n \to +\infty$, $n^{-\frac{\gamma}{2}} = O(\beta_{min}^*)$, $a_n = dn^{\frac{1}{2}-u}$, then assumptions about asymptotic behaviour of parameters in Proposition 4 are satisfied.

Bibliography

[1] P. Bühlmann and S. van de Geer *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer Science & Business Media, 2011.

[2] M. Kubkowski and J. Mielniczuk. Selection consistency of two-step selection method for misspecified logistic regression, *in preparation*, 2018.

[3] M. Kubkowski. Misspecification of binary regression model: properties and inferential procedures. *PhD Thesis*, 2019.

[4] P. Pokarowski, A. Prochenka, M. Frej, W. Rejchel and J. Mielniczuk. Improving Lasso for Model Selection and Prediction. *Unpublished manuscript*, available from: `https://www.univie.ac.at/seam/inference2018/abstracts/contributed/rejchel.pdf`, 2018.

[5] S. Rosset, J. Zhu and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5(Aug):941–973, 2004.

[6] R. J. Tibshirani. The Lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

On random environment integer-valued autoregressive models - a survey

Petra Laketa[1*]

[1] *University of Niš, Faculty of Sciences and Mathematics, Serbia*

Abstract: A survey of random-environment INAR models is presented. There is a motivation for the introduction of the random environment into INAR models. Different models that are defined are considered, and also their mutual properties. Also, the description of problems that arise in estimation and application are given, as well as the approaches that overcome these problems.

Keywords: random environment; INAR

AMS subject classification: 62M10

1 Introduction

The integer-valued autoregressive (INAR) models can be used when we want to describe behavior of some random events which we count. These models are firstly introduced in [5] and [1], and are based on the thinning operators. There are a lot of models, which differ in the thinning operators or in the marginal distribution, which are all stationary.

All mentioned models have one important property - they are stationary. However, there are a lot of data with non-stationary character, for which they would be therefore inappropriate. One of the non-stationary models which have dealt with this are considered in [2].

From here comes the motivation for the consideration of some kind of randomness in the environment of the process, which will lead to some non-stationary characteristics. The first random environment INAR model, of order one, is introduced in [6]. Then, process of higher order is given in [7], and conditional least square estimators for it are given in [4]. Finally, generalized random environment INAR models of higher order are introduced in [3].

Here will be given some discussion about this class of models. Firstly, there is the description of the random environment as well as the formal definition. Then there are introduced random environment INAR models. Finally, there is the description of the estimation of the unknown parameters and also application of the models considered.

[*] Corresponding author: petra.laketa@pmf.edu.rs

2 Random environment

As it was mentioned in the introduction, the main idea of random environment integer-valued autoregressive models is that the environment is not permanent, but has random variations. Also, what is more important, these variations affect the main process. Precisely, they change its marginal distribution, by changing its parameters. Therefore, we have situation that conditions in which the values of the process are observed variate and cause the change of the marginal distribution, which results in non-stationarity of the process. This is very important property, because it was very frequent to describe the data with the non-stationary character by the stationary INAR models. Of course, it was not enough convenient. This class of models is attempt to overcome problem with application of the INAR models to this type of data.
In order to describe this phenomena, in [7] is defined a process that will describe conditions of the environment by the following definition

Definition 1. A sequence of random variables $\{Z_n\}$, $n \in \mathbb{N}_0$, is called the r states random environment process, for $r \in \mathbb{N}$, if it is a Markov chain, which is taking values in $E_r = \{1, 2, \ldots, r\}$. More generally, $\{Z_n\}$, $n \in \mathbb{N}_0$, is the random environment process if it is the r states random environment process, for some $r \in \mathbb{N}$.

Since the stress is not on the random environment process, we are not interested in the real description of the environment and its the values, but only in its change that affects the change of the marginal distribution of the considered process. This is why E_r contains the first r positive integers.

3 The first-order random environment INAR model

It is now possible to construct the first-order random environment INAR model, which is done in [7] in the following way

Definition 2. A non-negative integer-valued sequence of random variables $\{X_n(Z_n)\}$, $n \in \mathbb{N}_0$, is said to be the r states random environment integer-valued autoregressive process of order 1 ($RrINAR(1)$) if it is given by

$$X_n(Z_n) = \sum_{i=1}^{X_{n-1}(Z_{n-1})} U_i + \varepsilon_n(Z_{n-1}, Z_n),\ n \in \mathbb{N},$$

where

$$X_n(Z_n) = \sum_{z=1}^{r} X_n(z) I_{\{Z_n = z\}},$$

$$\varepsilon_n(Z_{n-1}, Z_n) = \sum_{z_1=1}^{r} \sum_{z_2=1}^{r} \varepsilon_n(z_1, z_2) I_{\{Z_{n-1}=z_1, Z_n=z_2\}},$$

$\{U_i\}$, $i \in \mathbb{N}$, is a counting sequence of independent and identically distributed (i.i.d.) random variables generating a thinning operator, $\{Z_n\}$ is an r states random environment process introduced by Definition 1 and $\{\varepsilon_n(i,j)\}$, $n \in \mathbb{N}_0$, $i,j \in E_r$, are sequences of i.i.d. random variables, which meet the following conditions:

$(A1)$ $\{Z_n\}$, $\{\varepsilon_n(1,1)\}$, $\{\varepsilon_n(1,2)\}$, $\dots$, $\{\varepsilon_n(r,r)\}$, are mutually independent for all $n \in \mathbb{N}_0$,

$(A2)$ Z_m and $\varepsilon_m(i,j)$ are independent of $X_n(l)$ for $n < m$ and any $i,j,l \in E_r$.

For any $n \in \mathbb{N}$, random variables $X_n(Z_n)$ and $\varepsilon_n(Z_{n-1}, Z_n)$ can be interpreted as mixtures

$$X_n(Z_n) \overset{d}{=} \begin{cases} X_n(1), \text{ w.p. } P(Z_n = 1), \\ X_n(2), \text{ w.p. } P(Z_n = 2), \\ \quad \vdots \qquad\qquad \vdots \\ X_n(r), \text{ w.p. } P(Z_n = r), \end{cases}$$

and

$$\varepsilon_n(Z_{n-1}, Z_n) \overset{d}{=} \begin{cases} \varepsilon_n(1,1), \text{ w.p. } P(Z_{n-1} = 1, Z_n = 1), \\ \varepsilon_n(1,2), \text{ w.p. } P(Z_{n-1} = 1, Z_n = 2), \\ \quad \vdots \qquad\qquad\qquad \vdots \\ \varepsilon_n(r,r), \text{ w.p. } P(Z_{n-1} = r, Z_n = r). \end{cases}$$

In practice, it is simpler to assume that realization $\{z_n\}$ of the random environment process is known in advance. This assumption makes it easier to obtain the distribution of the residuals. Actually, this does not exclude the randomness of the environment. In fact, when we apply the model, we can firstly determine sequence $\{z_n\}$, and then consider model with realized values of random environment. If we want to predict values of the process $X(Z_n)$, we can predict Z_n based on z_{n-1} and then again use model with realized values $\{z_n\}$

Having that on mind, in [7] the sequence $\{X_n(z_n)\}$ is considered based on the realized random environment process, with additional assumptions about marginal distribution and the thinning operator. Such process is named the r states random environment $INAR(1)$ process with the determined geometric marginal distribution, based on the negative binomial thinning operator $(RrNGINAR(1))$.

As an illustration, in Figure 1 is presented the simulated sequence of the random states of length 50 with initial probabilities vector $(0.5, 0.5)$ and transition matrix $\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$. The plot of the simulated $R2NGINAR$ series based on that sequence of states with parameters $\mu = (1,5)$ and $\alpha = 0.15$ is given in Figure 2. We can see from the plot that parts corresponding to the second state show greater values, which is consequence of the fact that the second state corresponds to the greater expectation $\mu_2 = 5$ than the first $\mu_1 = 1$.

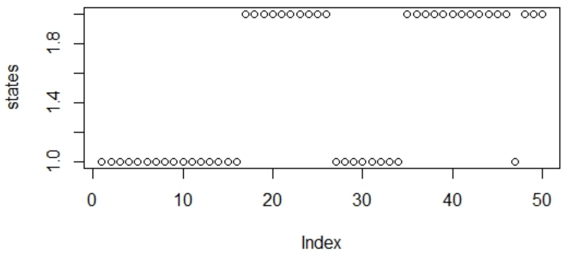


Figure 1: Plot of the random states

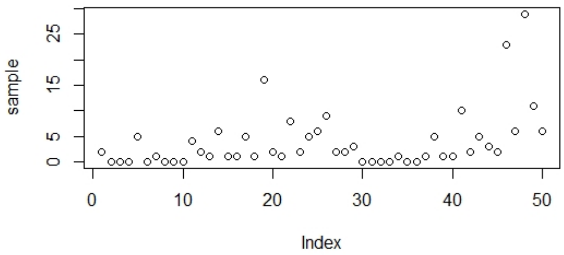


Figure 2: Plot of the simulated series

4 Generalized random environment integer-valued autoregressive model of higher order

The influence of the environment can be more significant. In the previous section it was assumed that only marginal distribution is determined by external conditions. This can be expanded also to the thinning parameter and to the order of the process. For that purpose, the following sets are introduced: $\mathcal{M} = \{\mu_1, \mu_2, ..., \mu_r\}$ with possible values of the parameters of the marginal geometric distribution, $\mathcal{A} = \{\alpha_1, \alpha_2, ..., \alpha_r\}$ that contains different values of thinning parameters and $\mathcal{P} = \{p_1, p_2, ..., p_r\}$ consisting of values of the maximal orders of the process in the appropriate state. Therefore, for every state $i \in E_r$ we have parameters μ_i, α_i and p_i that are valid for the model when that state of circumstances takes place. This is done in [3]. The model has the following form

$$X_n(z_n) = \begin{cases} \alpha_{z_n} * X_{n-1}(z_{n-1}) + \varepsilon_n(z_{n-1}, z_n), & \text{w.p. } \phi^{(z_n)}_{1,P_n}, \\ \alpha_{z_n} * X_{n-2}(z_{n-2}) + \varepsilon_n(z_{n-2}, z_n), & \text{w.p. } \phi^{(z_n)}_{2,P_n}, \\ \vdots & \vdots \\ \alpha_{z_n} * X_{n-P_n}(z_{n-P_n}) + \varepsilon_n(z_{n-P_n}, z_n), & \text{w.p. } \phi^{(z_n)}_{P_n,P_n}, \end{cases} \tag{1}$$

Here, if we use $P_n = p_{z_n}$, it would be too complex to obtain the distribution of the residuals. Therefore, two different approaches are used. For the first model, named INAR process with r-states, distribution parameters set $\mathcal{M}$, thinning parameters set $\mathcal{A}$ and maximal order set $\mathcal{P}$ (RrINARmax($\mathcal{M}, \mathcal{A}, \mathcal{P}$)), it holds $P_n = \min\{p_{z_n}, p_n^*\}$, where

$$p_n^* = \max\{i \in \{1, 2, ..., n\} : z_{n-1} = z_{n-2} = ... = z_{n-i}\}.$$

So, when the change of the state occurs, the process order becomes one, and then it becomes greater until it reaches the maximal order value for that state and remain maximal until the new state change. The alternative is that it does not grow gradually, but remains 1 until it can reach the maximal order, so in that case only possible orders for the state i are 1 and p_i. Precisely,

$$P_n = \begin{cases} p_{z_n}, & p_n^* \geq p_{z_n} \\ 1, & p_n^* < p_{z_n} \end{cases}$$

This model is named the random environment INAR process with r-states, distribution parameters set $\mathcal{M}$, thinning parameters set $\mathcal{A}$ and order set $\mathcal{P}$ (RrGINAR$_1$($\mathcal{M}, \mathcal{A}, \mathcal{P}$)).

What all random environment INAR models have in common is that they rely on the another sequence which represents states of the circumstances. This sequence determines marginal distribution at every moment and also the way that different elements of the process depend on each other. Therefore, these models have some similarities in the structure and also the approaches for estimation and application. Since they are non-stationary, it cannot be proved that Yule-Walker estimators are strongly consistent in a way it was done for other stationary INAR models. For that reason, the sample is divided into the pieces on which the process could be considered as stationary. Namely, from the moment when state i occurs to the moment when it changes to other state j, we have stationary process with fixed parameters μ_i, α_i and p_i, so estimators defined on that part of the sample would be strongly consistent and, consequently, their mean would be also strongly consistent as the linear combination of the strongly consistent estimators. This is the basic approach that is used in obtaining the Yule-Walker estimators for the random environment INAR models. Correctness of such defined estimators is confirmed on the values simulated from the appropriate models. They showed convergence to the true parameter values. In simulations, first step is construction of the sequence $\{z_n\}$ based on the given transition matrix. Then, the process $X_n(z_n)$ can be generated using the formula for the wanted random environment INAR model.

When it comes to application, the problem that arise is determination of the sequence $\{z_n\}$. The way it is solved is something new in the estimation for INAR models. Namely, the data is clustered into r different clusters for the given number r and then each cluster is assigned to the one state. After this procedure, we can take $z_n = i$ if x_n belongs to the i-th cluster. This makes sense because we already explained that values of z_n, which are from E_r are important only in service of distinction between different states. The number r should be chosen in dependence of the sample size. For the small samples, if we choose large value for r, then there will be a lot of small parts of the samples which can be treated as stationary, so estimates will be poor. But, enlarging r makes model more flexible, so it can better fit the data. Therefore, we should find optimum solution in combining these two principles. Relative to other competitive models, random environment INAR models showed best performance on the selected real-life data and therefore, they justify their introduction.

Bibliography

[1] M. A. Al-Osh and A. A. Alzaid. First-order integer-valued autoregressive (INAR(1)) process. *J. Time Ser. Anal.*, 8:261–275, 1987.

[2] N. M. Khan, R. Sunecher, and V. Jowaheer. Modelling a non-stationary BINAR(1) Poisson process. *J. Stat. Comput. Simul.*, 86(15):3106–3126, 2016.

[3] P. N. Laketa, A. S. Nastić and M. M. Ristić. Generalized Random Environment INAR Models of Higher Order. *Mediterr. J. Math.*, 15(1):9, 2018.

[4] P. N. Laketa and A. S. Nastić. Conditional least squares estimation of the parameters of higher order Random environment INAR models. *Facta Universitatis, Series Mathematics and Informatics*, to appear, 2019.

[5] E. McKenzie. Some simple models for discrete variate time series. *Water Resour. Bull.*, 21:645–650, 1985.

[6] A. S. Nastić, P. N. Laketa and M. M. Ristić. Random Environment Integer-Valued Autoregressive process. *J. Time Ser. Anal.*, 37:267–287, 2016.

[7] A. S. Nastić, P. N. Laketa and M. M. Ristić. Random environment INAR models of higher order. *Revstat Stat. J.*, 17(1):35–65, 2019.

Oscillating sequences of partial-sums discrete probability distributions

Lívia Leššová[1]*

[1] *Department of Applied Mathematics and Statistics, Faculty of Mathematics, Physics and Informatics of Comenius University in Bratislava, Mlynská dolina dolina 842 48, Bratislava*

Abstract: The limit of the repeated partial sums is defined. For these repeated partial summations, examples are presented where the limit does not exist and a sequence of probability distributions obtained by the partial sums oscillates.

Keywords: Dicrete probability distributions, partial summations, limit distribution, oscillation.

AMS subject classification: 60E05.

1 Partial sums distributions

Let $\{P_x^{(1)}\}_{x=0}^{\infty}$ and $\{P_x^*\}_{x=0}^{\infty}$ be two probability mass functions of univariate discrete distributions defined on nonnegative integers. Partial-sums discrete probability distributions are defined in general in [3] by

$$P^{(1)}(X=x) = P_x^{(1)} = c_1 \sum_{j=x}^{\infty} g(j) P_j^*, \quad x = 0, 1, 2, ...,$$

where c_1 is a normalizing constant, $\{P_x^*\}_{x=0}^{\infty}$ is a parent and $\{P_x^{(1)}\}_{x=0}^{\infty}$ is a descendant of the first generation. If we take $\{P_x^{(1)}\}_{x=0}^{\infty}$ as the parent of the new generation while function $g(j)$ remains unaltered, we obtain a descendant of the second generation

$$P_x^{(2)} = c_2 \sum_{j=x}^{\infty} g(j) P_j^{(1)}, \quad x = 0, 1, 2, ...,$$

where c_2 is a normalizing constant. Let the parent be the descendant of the $(k-1)$-st generation. Then the descendant of the k-th generation is

$$P_x^{(k)} = c_k \sum_{j=x}^{\infty} g(j) P_j^{(k-1)}, \quad x = 0, 1, 2, ...,$$

[*] Corresponding author: livia.lessova@fmph.uniba.sk

where c_k is a normalizing constant.
The existence of a limit distribution

$$P_x^{(\infty)} = \lim_{k\to\infty} P_x^{(k)} \tag{1}$$

of such repeated partial summations for some given parent $\{P_x^*\}_{x=0}^{\infty}$ and a given function $g(j) = a$ (which is called a geometric summation) was studied in [4]. In [1] it was proved that if a discrete distribution has a finite support, the limit distribution (1) can be found (with some restrictions on function g(j)) using the power method (a numerical method often used to find the dominant eigenvalue and its corresponding eigenvector of a matrix). This result was extended in [2] for bivariate partial summations.

2 Oscillating sequences of distributions

In the past only the existence of the limit (1) was proved under some conditions (see [1, 2, 4]). An example where the limit does not exist has not been provided so far. In the following, some examples of oscillating sequences of descendant distributions which arise as results of repeated partial summations will be presented.

Oscillating distributions with support size two

Let the parent distribution with the support size two be given by $P_0^* = \frac{1}{10}$, $P_1^* = 1 - P_0^* = \frac{9}{10}$. If $g(0) = -1$ and $g(1) = 1$, the first descendant is

$$\begin{aligned} P_0^{(1)} &= c_1 g(0) P_0^* + c_1 g(1) P_1^* = -c_1 \frac{1}{10} + c_1 \frac{9}{10} = c_1 \frac{8}{10}, \\ P_1^{(1)} &= c_1 g(1) P_1^* = c_1 \frac{9}{10}, \end{aligned}$$

then after determining the normalizing constant c_1 we obtain

$$P_0^{(1)} = \frac{8}{17}, \qquad P_1^{(1)} = \frac{9}{17}.$$

The descendant of the second generation is

$$\begin{aligned} P_0^{(2)} &= c_2 g(0) P_0^{(1)} + c_2 g(1) P_1^{(1)} = \frac{1}{10}, \\ P_1^{(2)} &= c_2 g(1) P_1^{(1)} = \frac{9}{10}, \end{aligned}$$

which is identical with the parental distribution. This sequence of descendant distributions determined by the partial summation with $g(0) = -1$, $g(1) = 1$ with the starting distribution $P_0^* = \frac{1}{10}$, $P_1^* = \frac{9}{10}$ oscillates with the period $k = 2$. This example can be generalized for the probability distributions with the support size two.

The aim is to identify every period and parent distribution for which the sequence of partial sums oscillates. For the oscillation with a period k we will take $g(0) = a$ and $g(1) = b$ (the normalizing constant is "hidden" in the function $g(j)$), which fully determines function g(j) (the support size is two). The k-th descendant is

$$\begin{aligned} P_0^{(k)} &= a^k P_0^* + (a^{k-1}b + a^{k-2}b^2 + ... + ab^{k-1} + b^k)P_1^*, \\ P_1^{(k)} &= b^k P_1^*. \end{aligned}$$

It is obvious that $g(1) = b = 1$ (or $g(1)$ can be -1, but this case is analogical). Consequently,

$$\begin{aligned} P_0^{(k)} &= a^k P_0^* + \left(\sum_{i=0}^{k-1} a^i\right) P_1^*, \\ P_1^{(k)} &= P_1^*, \end{aligned}$$

for $k \geq 1$. When we put $P_0^{(k)} = P_0^*$ and if $a \neq -1$, then

$$\begin{aligned} P_0^{(k)} = P_0^* &= \frac{a^{k-1} + a^{k-2} + ... + a + 1}{-a^k + a^{k-1} + a^{k-2} + ... + a + 2} = \\ &= \frac{a^{k-1} + a^{k-2} + ... + a + 1}{(2-a)(a^{k-1} + a^{k-2} + ... + a + 1)} = \frac{1}{2-a}. \end{aligned}$$

For $a \leq 1$ and $a \neq -1$ we obtain distribution $P_0^* = \frac{1}{2-a}$, $P_1^* = \frac{1-a}{2-a}$ which is invariant with respect to partial summation given by $g(0) = a$, $g(1) = 1$ (it remains unchanged after applying this partial summation):

$$\begin{aligned} P_0^{(1)} &= a\frac{1}{2-a} + \frac{1-a}{2-a} = \frac{1}{2-a}, \\ P_1^{(1)} &= \frac{1-a}{2-a}. \end{aligned}$$

For $a = -1$ we obtain

$$\begin{aligned} P_0^{(k)} &= (-1)^k P_0^* + \left[(-1)^{k-1} + (-1)^{k-2} + ... + (-1) + 1\right] P_1^* = P_0^*, \\ P_1^{(k)} &= P_1^*, \end{aligned} \tag{2}$$

from which we can easily obtain the invariant distribution if we set $k = 1$. This invariant distribution is

$$P_0^* = \frac{1}{3}, \qquad P_1^* = \frac{2}{3}.$$

For the even period $k = 2l, l = 1, 2, ...$ it follows from (2) that

$$\begin{aligned} P_0^{(k)} &= P_0^* + [(-1) + 1 + ... + (-1) + 1]\, P_1^* = P_0^*, \\ P_1^{(k)} &= P_1^*, \end{aligned}$$

and we can see that the lowest period is $k = 2$. Indeed, another even period for which two is not a period does not exist. There is also another requirement - the descendant of the first generation must be a probability distribution, which means that in our case with $a = -1$ and $k = 2$ the parent distribution must satisfy $P_0^* \leq P_1^*$, in other words $P_0^* \leq \frac{1}{2}$. The only odd period is one, i.e. the constant sequence of the same distributions (see the abovementioned invariant distribution).

Oscillating distributions with support size three

Results from the previous section can be extended to discrete distributions with support size three. For distributions with the support size three we need three values of the function $g(j)$ in $0, 1$ and 2. With the values $g(1)$ and $g(2)$ we have the similar situation as in the case with the support size two, because the value $g(0)$ does not influence the transformation of P_1^*, P_2^*:

$$\begin{aligned} P_0^{(1)} &= c_1 \left[g(0)P_0^* + g(1)P_1^* + g(2)P_2^* \right], \\ P_1^{(1)} &= c_1 \left[g(1)P_1^* + g(2)P_2^* \right], \\ P_2^{(1)} &= c_1 g(2)P_2^*. \end{aligned}$$

From the previous section we know that if $g(1) = -1$ and $g(2) = 1$ (normalizing constant is "hidden" in this function), sequences $\{P_1^{(i)}\}_{i=1}^{\infty}$, $\{P_2^{(i)}\}_{i=1}^{\infty}$ oscillate for $P_1^* \leq P_2^*$ with the period $k = 2$, and odd periods (excluding invariant distributions) are not possible.

For $g(0) = a$, $g(1) = -1$, $g(2) = 1$ and an even period $k = 2l, l = 1, 2, ...$ we obtain for $P_0^{(k)}$ ($P_1^{(k)}$, $P_2^{(k)}$ remain unchanged for an even period)

$$P_0^{(k)} = a^k P_0^* + (-a^{k-1} + a^{k-2} - \cdots + 1)P_1^* + (a^{k-1} + a^{k-3} + \cdots + a)P_2^*,$$

and after some trivial algebraic operations

$$(a-1)(a+1)P_0^* + (1-a)P_1^* + aP_2^* = 0,$$

which gives similar results as in the case that the support has size two. A sequence of descendant distributions can oscillate only with period $k = 2$ if the parent distribution is

$$\begin{aligned} P_0^* &= P_0^*, \\ P_1^* &= \frac{aP_0^* + P_0^* - a - a^2 P_0^*}{1 - 2a}, \\ P_2^* &= \frac{aP_0^* + a^2 P_0^* + 1 - 2P_0^* - a}{1 - 2a}. \end{aligned}$$

Naturally, there are some conditions which the parent distribution must satisfy so that its descendants of all generations be proper probability distributions. The conditions are presented in Table 1. For every a and P_0^* from these intervals, a sequence of descendant distributions oscillates with period $k = 2$. There is only one

a	P_0^*
$(-\infty, 0)$	$\left(0, \frac{1}{3-a}\right)$
$\left(0, \frac{1}{2}\right)$	$\left(\frac{-a}{a^2-a-1}, \frac{1}{3-2a^2}\right)$
$\left(\frac{1}{2}, 1\right)$	$\left(\frac{1}{3-2a^2}, \frac{-a}{a^2-a-1}\right)$

Table 1: Conditions for a and P_0^*.

exception, when the period is $k = 1$ and the distribution is invariant with respect to the partial summation, which is true if the parent distribution is

$$P_0^* = \frac{1}{4-3a}, \qquad P_1^* = \frac{1-a}{4-3a}, \qquad P_2^* = \frac{2(1-a)}{4-3a}.$$

The case with $g(2) = -1$ can be obtained analogically.

Acknowledgements: Supported by grants VEGA 2/0054/18 and UK/334/2019.

Bibliography

[1] M. Koščová, R. Harman and J. Mačutek. Iterated partial summations applied to finite-support discrete distributions. *Mathematica Slovaca*, 2019 (to appear).

[2] L. Leššová, J. Mačutek. On the limit behaviour of finite-support bivariate discrete probability distributions under iterated partial summations. *https://arxiv.org/abs/1903.03316*, (accessed on 21-March-2019).

[3] J. Mačutek. On two types of partial summations. *Tatra Mountains Mathematical Publications*, 26: 403–410, 2003.

[4] J. Mačutek. A limit property of the geometric distribution. *Theory of Probability and its Applications*, 50(2): 316–319, 2006.

On blind source separation under martingales: A probability theoretic perspective

Niko Lietzén[1]*

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

Abstract: In this short paper, we consider linear blind source separation for stochastic processes that have conditional dependency structures. We present a conditional version of a linear blind source separation model. The conditional dependency structure is imposed to the model via discrete time martingales. We present some theoretical foundations for solving the corresponding conditional blind source separation problem and provide discussion regarding our future work on the topic.

Keywords: blind source separation, joint diagonalization, martingales, multivariate analysis

AMS subject classification: 60G42, 62H99, 94A12

1 Introduction

The blind source separation (BSS) problem is a recurring and a widely studied topic in several fields of science. Usually, the goal in BSS is to reverse the effects of an unknown mixing system and to recover some signals of interest. There exists several applications where blind source separation is utilized, e.g., finance, biomedical applications and telecommunications, see [2] for a collection.

Linear BSS has been widely studied under assumptions of weakly stationary processes. However, there exists an increasing demand for models that allow time varying correlations, especially in applications of finance such as modeling financial returns. In this paper, we present a version of the linear BSS model called the conditional blind source separation (cBSS) model, in which we assume that the latent processes of interest are discrete time martingales. The presented cBSS model is able to model conditional dependency structures, which are not captured by traditional BSS models, and which are relevant in several financial applications. In this short paper, we present the cBSS model and provide some theoretical groundwork for solving the corresponding cBSS problem. In particular, we consider the identification of the solutions for the cBSS problem. Furthermore, the finite sample estimation procedure and the derivation of the corresponding asymptotical properties are discussed.

*Corresponding author: niko.lietzen@aalto.fi

2 Conditional blind source separation

Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space where $\mathbb{F} := (\mathcal{F}_t)_{t\in\{0\}\cup\mathbb{N}}$ such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$. We use $\mathbf{x} \in m\mathcal{F}$ to denote that $\mathbf{x}$ is $\mathcal{F}$-measurable. We next present a conditional blind source separation (cBSS) model. The presented version of the cBSS model does not include a location parameter. A constant location parameter could be straightforwardly implemented into the model, which would result in some minor modifications on the following assumptions.

Definition 1. A $\mathbb{R}^p$-valued stochastic process $\mathbf{x}_\bullet := (\mathbf{x}_t(\omega))_{t\in\{0\}\cup\mathbb{N}}$ follows a conditional blind source separation (cBSS) model if

$$\mathbf{x}_t = \mathbf{A}\mathbf{z}_t, \qquad \text{for all } t \in \{0\} \cup \mathbb{N},$$

where $\mathbf{A} \in m\mathcal{F}_0$ is a $\mathbb{R}^{p\times p}$-matrix of rank p and the $\mathbb{R}^p$-process $\mathbf{z}_\bullet := (\mathbf{z}_t(\omega))_{t\in\mathbb{N}}$ is $\mathbb{F}$-adapted and square-integrable such that,

(Z1) $\mathbb{P}\left[\{\omega \in \Omega : z_0(\omega) = \mathbf{0}_p\}\right] = 1,$

(Z2) $\mathbb{E}\left[\mathbf{z}_t \mid \mathcal{F}_{t-1}\right] = \mathbf{z}_{t-1},\ \mathbb{P}\text{-a.s.},$

(Z3) $\exists s, t \in \mathbb{N} : \mathbb{E}\left[\mathbf{z}_s\mathbf{z}_s^\top\right] = \mathbf{\Lambda}_s,\ \mathbb{E}\left[\mathbf{z}_t\mathbf{z}_t^\top\right] = \mathbf{\Lambda}_t \ \text{ and } \ \tilde{\mathbf{\Lambda}}_{st} = \mathbf{\Lambda}_s\mathbf{\Lambda}_t^{-1},$

where $\mathbf{\Lambda}_s \in m\mathcal{F}_s$, $\mathbf{\Lambda}_t \in m\mathcal{F}_t$ are $\mathbb{R}^{p\times p}$-valued diagonal matrices with positive diagonal entries and $\tilde{\mathbf{\Lambda}}_{st}$ is a diagonal matrix with distinct diagonal elements.

The latent process $\mathbf{z}_\bullet$, given in Definition 1, satisfies the properties of a $\mathbb{R}^p$-variate discrete martingale. The martingale property of $\mathbf{z}_\bullet$ together with condition (Z1) imply that the process is centered, i.e., $\mathbb{E}[\mathbf{z}_t] = \mathbf{0}_p$ for every $t \in \mathbb{N}$. Furthermore, since the process $\mathbf{z}_\bullet$ is square-integrable, the martingale property gives $\mathbb{E}[\mathbf{z}_t\mathbf{z}_{t-\tau}^\top] = \mathbb{E}[\mathbf{z}_{t-\tau}\mathbf{z}_{t-\tau}^\top]$, $\tau \in \mathbb{N}$. Since every autocovariance matrix is equal to the covariance matrix at some point of time, many widely used blind source separation procedures are unapplicable here. For example, classic versions of the algorithm for multiple unknown signals extraction (AMUSE) [7] and second order blind identification (SOBI) [1] procedures rely on assumptions of weak stationarity. Since there exists no nontrivial martingales that are weakly stationary, no nontrivial weakly stationary processes $\mathbf{z}_t$ can satisfy conditions (Z1)-(Z3). Thus, it is not surprising that AMUSE and SOBI cannot be applied here. The assumption that $\mathbf{\Lambda}_s$ and $\mathbf{\Lambda}_t$ have positive diagonal elements is natural, since they are variances. Consequently, the diagonal entries of $\tilde{\mathbf{\Lambda}}$ are also positive. We will further discuss the identification condition (Z3) after the following definition that quantifies the solutions for the cBSS problem.

Definition 2. Let $\mathbf{x}_\bullet$ be a process that follows the cBSS model, such that condition (Z3) holds for some fixed $s, t \in \mathbb{N}$. A $\mathbb{R}^{p\times p}$-valued matrix $\mathbf{\Gamma}$ is a solution to the cBSS problem, if the process $\mathbf{y}_\bullet := (\mathbf{\Gamma}\mathbf{x}_t(\omega))_{t\in\{0\}\cup\mathbb{N}}$ satisfies conditions (Z1), (Z2) and condition (Z3) holds for the fixed t and s.

Under the cBSS model, we have that the process $\mathbf{x}_\bullet$ is also a $\mathbb{F}$-adapted and square-integrable process that satisfies conditions (Z1) and (Z2). Thus, the identification of the solution $\mathbf{\Gamma}$ relies solely on condition (Z3). However, the condition (Z3) is not enough to produce unique solutions for the cBSS problem. Let $\mathbf{P}$ be a permutation matrix, $\mathbf{D}$ a diagonal matrix with positive diagonal entries and $\mathbf{J}$ a sign-change matrix, that is, $\mathbf{J}_{kk} \in \{-1, 1\}$ for every $k \in \{1, \dots, p\}$. Then, let $\tilde{\mathbf{y}}_\bullet := (\mathbf{PJD\Gamma x}_t(\omega))_{t \in \{0\} \cup \mathbb{N}}$, where $\mathbf{\Gamma}$ is a solution to the cBSS problem. The process $\tilde{\mathbf{y}}_\bullet$ is square-integrable and $\mathbb{F}$-adapted and it satisfies conditions (Z1) and (Z2). Additionally, condition (Z3) is satisfied since,

$$\mathbb{E}[\tilde{\boldsymbol{y}}_h \tilde{\boldsymbol{y}}_h^\top] = \mathbf{PJD\Lambda}_h(\mathbf{PJD})^\top = \mathbf{PD\Lambda}_h\mathbf{DP}^\top = \mathbf{L}_h,$$

where $\mathbf{L}_h$ is a diagonal matrix with positive diagonal entries at the fixed points of time $h \in \{s, t\}$. Furthermore, we have that,

$$\mathbf{L}_s\mathbf{L}_t^{-1} = \mathbf{PD\Lambda}_s\mathbf{DP}^\top\mathbf{PD}^{-1}\mathbf{\Lambda}_t^{-1}\mathbf{D}^{-1}\mathbf{P}^\top = \mathbf{P}\tilde{\mathbf{\Lambda}}_{st}\mathbf{P}^\top = \tilde{\mathbf{L}}_{st},$$

where $\tilde{\mathbf{L}}_{st}$ is a diagonal matrix that has the same distinct diagonal entries, possibly in some permuted order, as the matrix $\tilde{\mathbf{\Lambda}}_{st}$. Hereby, with the current cBSS model assumptions, we can at best hope to recover the latent process up to the matrices $\mathbf{P}, \mathbf{J}$ and $\mathbf{D}$.

Recovering the latent processes up to order and heterogeneous scaling is sufficient for many applications, in particular, for the case when the most relevant information of $\tilde{\mathbf{z}}_\bullet$ is contained in its waveform, i.e., its shape. These identifiability issues are not considered to be problematic in the signal processing literature, see e.g. [6] for further discussion.

It is hereby justified to not distinguish between solutions that solve the same cBSS problem. We say that solutions $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ are equivalent, if there exists a permutation matrix $\mathbf{P}$, a sign-change matrix $\mathbf{J}$ and a diagonal matrix (with positive diagonal elements) $\mathbf{D}$, such that $\mathbf{\Gamma}_1 = \mathbf{PJD\Gamma}_2$ and we denote the corresponding equivalence relation as $\mathbf{\Gamma}_1 \equiv \mathbf{\Gamma}_2$. Note that under the cBSS model assumptions, we have for every solution $\mathbf{\Gamma}$ that $\mathbf{\Gamma} \equiv \mathbf{A}^{-1}$, that is, $\mathbf{\Gamma A} = \mathbf{PJD}$. With this equivalence property between the solutions and the unknown mixing matrix $\mathbf{A}$, we can present a solution procedure for the cBSS problem.

Theorem 1. *Let $\boldsymbol{x}_\bullet$ be a process that satisfies Definition 1 and let $\mathbf{\Sigma}_t = \mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top]$. Then, every cBSS solution $\mathbf{\Gamma}$ satisfies the eigenvector-eigenvalue equation,*

$$\mathbf{\Sigma}_t^{-1}\mathbf{\Sigma}_s\mathbf{\Gamma}^\top = \mathbf{\Gamma}^\top\boldsymbol{L},$$

where $\boldsymbol{L}$ is a diagonal matrix with distinct diagonal elements.

Proof of Theorem 1. Since $\mathbf{x}_t$ follows the cBSS model, we have for some fixed t that $\mathbf{\Sigma}_t = \mathbf{A\Lambda}_t\mathbf{A}^\top$ and,

$$\mathbf{\Sigma}_t\mathbf{\Gamma}^\top = \mathbf{A\Lambda}_t\mathbf{A}^\top\mathbf{\Gamma}^\top = \mathbf{A\Lambda}_t\left(\mathbf{PJD}\right)^\top = \mathbf{A}\tilde{\mathbf{\Lambda}}_t\mathbf{P}^\top, \tag{1}$$

where $\tilde{\mathbf{\Lambda}}_t = \mathbf{\Lambda}_t\mathbf{DJ}$ is a diagonal matrix with nonzero diagonal entries and consequently invertible. By right-multiplying both sides of Eq. (1) with $(\tilde{\mathbf{\Lambda}}_t\mathbf{P}^\top)^{-1}$, gives that,

$$\mathbf{\Sigma}_t\mathbf{\Gamma}^\top\mathbf{P}\tilde{\mathbf{\Lambda}}_t^{-1} = \mathbf{A}. \tag{2}$$

Then, since Eq. (1) also holds for some fixed $s \neq t$, and by using Eq. (2), we get that,

$$\mathbf{\Sigma}_s \mathbf{\Gamma}^\top = \mathbf{\Sigma}_t \mathbf{\Gamma}^\top \mathbf{P}\mathbf{J}\mathbf{D}^{-1}\mathbf{\Lambda}_t^{-1}\mathbf{\Lambda}_s \mathbf{D}\mathbf{J}\mathbf{P}^\top = \mathbf{\Sigma}_t \mathbf{\Gamma}^\top \mathbf{P}\tilde{\mathbf{\Lambda}}_{st}\mathbf{P}^\top = \mathbf{\Sigma}_t \mathbf{\Gamma}^\top \mathbf{L}, \tag{3}$$

where $\mathbf{L} = \mathbf{P}\tilde{\mathbf{\Lambda}}_{st}\mathbf{P}^\top$ is a diagonal matrix that has the same distinct diagonal entries, possibly in some permuted order, as the matrix $\tilde{\mathbf{\Lambda}}_{st}$. The claim then follows by left-multiplying both sides of Eq. (3) with $\mathbf{\Sigma}_t^{-1}$. □

Thus, by Theorem 1, we get solutions for the cBSS problem by finding the eigenvectors of $\mathbf{\Sigma}_t^{-1}\mathbf{\Sigma}_s$. As with eigenvalue-eigenvector problems usually, the eigenvectors are only unique up to order and heterogeneous scaling, which is perfectly in line with our identifiability conditions. Similar versions of Theorem 1 are often utilized in the BSS estimation procedure, see e.g. [4, 5]. Note that, we can form a finite sample version of Theorem 1 by replacing $\mathbf{\Sigma}_s$ and $\mathbf{\Sigma}_t$ with some appropriate estimators. In practice, the fixed s, t can be chosen, e.g., such that the diagonal elements of $\tilde{\mathbf{\Lambda}}_{st}$ are as distinct as possible.

In the BSS literature, the model is often defined such that either $\mathbf{\Lambda}_s$ or $\mathbf{\Lambda}_t$ is e.g. the identity matrix and the other is a diagonal matrix with distinct diagonal elements. In this case, we can also identify the scaling matrix $\mathbf{D}$, and we can think of the estimation procedure as a two step process that includes finding a suitable scaling and rotation for the latent $\mathbf{x}_\bullet$. Note that the corresponding case is included in our model, although in the estimations procedure one would have to fix the scales of the eigenvectors accordingly.

3 Discussion and future work

In this short paper, we presented the conditional blind source separation (cBSS) model and provided some theoretical foundations in order to solve the corresponding cBSS problem. In future work, we will provide finite sample estimators for the cBSS problem and derive some relevant asymptotic properties for them. Note that the model assumptions we have made, in this paper, are minimal and in order to make the estimation procedure meaningful, and make the derivation of the asymptotic properties possible, we will have to require more structure from either the latent process $\mathbf{z}_\bullet$ or from the matrix $\mathbf{A}$.

The estimation procedure benefits significantly, if we considered the fully conditional version of the cBSS model. The stronger version of condition (Z3) of Definition 1 is formed by requiring that the conditional covariance matrices $\mathbb{E}[\mathbf{z}_s\mathbf{z}_s^\top \mid \mathcal{F}_{s-1}]$ and $\mathbb{E}[\mathbf{z}_t\mathbf{z}_t^\top \mid \mathcal{F}_{t-1}]$ are diagonal and that the product of the diagonal matrices has distinct diagonal entries. Clearly, these stronger conditions imply the current condition (Z3) since $\mathbb{E}[\mathbf{z}_s\mathbf{z}_s^\top] = \mathbb{E}[\mathbb{E}[\mathbf{z}_s\mathbf{z}_s^\top \mid \mathcal{F}_{s-1}]]$ and similarly for t. The assumption of conditional uncorrelated components seems to be a good middle ground between the overly weak assumption of unconditional uncorrelated components and the overly strong assumption of independent components.

If we, in addition to the conditional version of condition (Z3), replaced condition (Z2) with $\mathbb{E}[\mathbf{z}_t \mid \mathcal{F}_{t-1}] = 0$, for every $t \in \mathbb{N}$, we would get a latent martingale difference sequence $\mathbf{z}_\bullet$ that follows the so-called conditional uncorrelated components

(CUC) model presented in [3]. Note that the CUC model is a special case of the cBSS model and thus the theory presented in this paper can be directly applied to the CUC case also. The CUC model is motivated by financial applications where conditional dependency structures are present, such as generalized auto-regressive conditional heteroscedasticity (GARCH) processes, see [3] for additional details. Thus, considering GARCH processes in the context of the cBSS model could provide fruitful results from both the theoretical and the applied perspective.

Acknowledgements: Niko Lietzén gratefully acknowledges financial support from the Emil Aaltonen Foundation (grant 180144). The author would also like to thank the anonymous referee, Pauliina Ilmonen, Lauri Viitasaari and Joni Virta for providing their insights, which have helped improve the paper.

Bibliography

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on signal processing*, 45(2):434–444, 1997.

[2] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications.* Academic Press, 2010.

[3] J. Fan, M. Wang, and Q. Yao. Modelling multivariate volatilities via conditionally uncorrelated components. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 70(4):679–702, 2008.

[4] P. Ilmonen, J. Nevalainen, and H. Oja. Characteristics of multivariate distributions and the invariant coordinate system. *Statistics & Probability Letters*, 80(23-24):1844–1853, 2010.

[5] L. Parra and P. Sajda. Blind source separation via generalized eigenvalue decomposition. *Journal of Machine Learning Research*, 4(Dec):1261–1269, 2003.

[6] L. Tong, R.-W. Liu, V. C. Soon, and Y.-F. Huang. Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems*, 38(5):499–509, 1991.

[7] L. Tong, V. Soon, Y. Huang, and R. Liu. AMUSE: a new blind identification algorithm. In *IEEE International Symposium on Circuits and Systems*, pages 1784–1787. IEEE, 1990.

Statistical analysis of parameter estimators in the fractional Vasicek model

Stanislav Lohvinenko[1][*]

[1] *Taras Shevchenko National University of Kyiv*

Abstract: The fractional Vasicek model, described by the stochastic differential equation $dX_t = (\alpha - \beta X_t)\,dt + \gamma\,dB_t^H$, where B^H is a fractional Brownian motion, is studied. It is assumed that the parameters $x_0 \in \mathbb{R}$, $\gamma > 0$ and $H \in (0,1)$ are known. The problem of estimating α and β is considered. Least squares, maximum likelihood and alternative estimators are constructed, and their asymptotic properties are established.

Keywords: Fractional Brownian motion, fractional Vasicek model, parameter estimation, consistency, asymptotic normality.

AMS subject classification: 60G22, 62F10, 62F12.

1 Introduction

The standard Vasicek model was proposed and studied by O. Vasicek [6] in 1977 for the purpose of interest rate modeling. It is described by the following stochastic differential equation

$$dX_t = (\alpha - \beta X_t)\,dt + \gamma\,dW_t, \tag{1}$$

where $\alpha, \beta, \gamma \in \mathbb{R}_+$, and W is a standard Wiener process. From the financial point of view, β corresponds to the speed of recovery, the ratio α/β is the long-term average interest rate, and γ represents the stochastic volatility. Now the Vasicek model is widely used not only in finance, but also in various scientific areas such as economics, biology, physics, chemistry, medicine and environmental studies.
In our research we deal with the fractional Vasicek model of the form

$$dX_t = (\alpha - \beta X_t)\,dt + \gamma\,dB_t^H, \tag{2}$$

where the Wiener process W is replaced with B^H, a fractional Brownian motion with Hurst index $H \in (0,1)$. This generalization of the model (1) enables one to model processes with long-range dependence. Such processes appear in finance, hydrology, telecommunication, turbulence and image processing.

[*] Corresponding author: stanislav.lohvinenko@gmail.com

2 Model description

Let $(\Omega, \mathfrak{F}, \mathbf{P})$ be a complete probability space. Let $B^H = \{B_t^H,\ t \geq 0\}$ be a fractional Brownian motion on this probability space, that is, a centered Gaussian process with covariance function

$$\mathbb{E}B_t^H B_s^H = \frac{1}{2}\left(s^{2H} + t^{2H} - |t-s|^{2H}\right).$$

We consider the continuous (and even Hölder up to order H) modification of B_t^H that exists due to the Kolmogorov theorem.
We study the fractional Vasicek model, described by the stochastic differential equation

$$X_t = x_0 + \int\limits_0^t (\alpha - \beta X_s)\, ds + \gamma B_t^H, \quad t \geq 0. \tag{3}$$

We assume that the parameters $x_0 \in \mathbb{R}$, $\gamma > 0$ and $H \in (0,1)$ are known. Such assumption can be made due to existence of many methods to estimate parameters γ and H (for example, see [1] and [3, Remark 2.1]). The main goal is to estimate parameters $\alpha \in \mathbb{R}$ and $\beta > 0$ by continuous observations of a trajectory of X on the interval $[0, T]$.
Following [2], for $0 < s < t \leq T$, define

$$\kappa_H = 2H\Gamma\left(3/2 - H\right)\Gamma\left(H + 1/2\right), \qquad \lambda_H = \frac{2H\Gamma(3-2H)\Gamma(H+1/2)}{\Gamma(3/2-H)},$$
$$k_H(t,s) = \kappa_H^{-1} s^{1/2-H}(t-s)^{1/2-H}, \qquad w_t^H = \lambda_H^{-1} t^{2-2H}.$$

Define also next stochastic processes

$$S_t = \frac{1}{\gamma}\int_0^t k_H(t,s)\, dX_s, \qquad P_H(t) = \frac{1}{\gamma}\frac{d}{dw_t^H}\int_0^t k_H(t,s) X_s\, ds,$$
$$Q_H(t) = \frac{1}{\gamma}\frac{d}{dw_t^H}\int_0^t k_H(t,s)(\alpha - \beta X_s)\, ds = \frac{\alpha}{\gamma} - \beta P_H(t).$$

Process S is called fundamental semimartingale. It has the following properties [2, Theorem 1]:

1. process S is a $(\mathfrak{F}_t)$-semimartingale with the decomposition

$$S_t = \int_0^t Q_H(s)\, dw_s^H + M_t^H, \tag{4}$$

where $M_t^H = \int_0^t k_H(t,s)\, dB_s^H$ is a Gaussian martingale, whose variance function $\langle M^H \rangle = w^H$;

2. process X admits the representation

$$X_t = \int_0^t K_H(t,s)\, dS_s, \tag{5}$$

where

$$K_H(t,s) = \gamma H(2H-1)\int_s^t r^{H-1/2}(r-s)^{H-3/2}\,dr;$$

3. the natural filtrations of processes S and X coincide.

3 Main results

Let us introduce the least squares estimators of the unknown parameters:

$$\widehat{\alpha}_T^{(1)} = \frac{(X_T - X_0)\int_0^T X_t^2\,dt - \int_0^T X_t\,dX_t \int_0^T X_t\,dt}{T\int_0^T X_t^2\,dt - \left(\int_0^T X_t\,dt\right)^2}, \tag{6}$$

$$\widehat{\beta}_T^{(1)} = \frac{(X_T - X_0)\int_0^T X_t\,dt - T\int_0^T X_t\,dX_t}{T\int_0^T X_t^2\,dt - \left(\int_0^T X_t\,dt\right)^2}. \tag{7}$$

Theorem 1 ([5, Theorem 2.1])**.** *Let $H \in [\frac{1}{2}, 1)$. Then the estimators $\widehat{\alpha}_T^{(1)}$ and $\widehat{\beta}_T^{(1)}$ are strongly consistent.*

Since the discretization and simulation of $\widehat{\alpha}_T^{(1)}$ and $\widehat{\beta}_T^{(1)}$ when $H \neq 1/2$ is quite difficult, we introduce alternative estimators:

$$\widehat{\beta}_T^{(2)} = \left(\frac{1}{\gamma^2 H\Gamma(2H)\,T^2}\left(T\int_0^T X_t^2\,dt - \left(\int_0^T X_t\,dt\right)^2\right)\right)^{-\frac{1}{2H}}, \tag{8}$$

$$\widehat{\alpha}_T^{(2)} = \frac{\widehat{\beta}_T^{(2)}}{T}\int_0^T X_t\,dt. \tag{9}$$

Theorem 2 ([5, Theorem 2.2])**.** *Let $H \in (0, 1)$. Then the estimators $\widehat{\alpha}_T^{(2)}$ and $\widehat{\beta}_T^{(2)}$ are strongly consistent.*

In applications usually the observations cannot be continuous. The estimators $\widehat{\alpha}_T^{(2)}$ and $\widehat{\beta}_T^{(2)}$ can be discretized as follows.
Let $h > 0$. Assume that a trajectory of X is observed at times $t_k = kh$, $k = 0, 1, \dots, n$. Define

$$\widehat{\beta}_n^{(3)} = \left(\frac{1}{\gamma^2 H\Gamma(2H)\,n^2}\left(n\sum_{k=0}^{n-1} X_{kh}^2 - \left(\sum_{k=0}^{n-1} X_{kh}\right)^2\right)\right)^{-\frac{1}{2H}}, \tag{10}$$

$$\widehat{\alpha}_n^{(3)} = \frac{\widehat{\beta}_n^{(3)}}{n}\sum_{k=0}^{n-1} X_{kh}. \tag{11}$$

Theorem 3 ([5, Theorem 2.3])**.** *Let $H \in (0,1)$. Then the estimators $\widehat{\alpha}_n^{(3)}$ and $\widehat{\beta}_n^{(3)}$ are strongly consistent.*

Applying the analog of the Girsanov formula for a fractional Brownian motion (see [2, Theorem 3]), we obtain next likelihood ratio:

$$\begin{aligned}\Lambda_H(T) &= \exp\left\{\int_0^T Q_H(t)\,dS_t - \frac{1}{2}\int_0^T (Q_H(t))^2\,dw_t^H\right\}\\ &= \exp\left\{\frac{\alpha}{\gamma}S_T - \beta\int_0^T P_H(t)\,dS_t - \frac{\alpha^2}{2\gamma^2}w_T^H\right.\\ &\quad\left. + \frac{\alpha\beta}{\gamma}\int_0^T P_H(t)\,dw_t^H - \frac{\beta^2}{2}\int_0^T (P_H(t))^2\,dw_t^H\right\}.\end{aligned} \tag{12}$$

Now we can construct maximum likelihood estimators.

Theorem 4 ([4, Theorem 3.1])**.** *Let $H > 1/2$ and β is known. The MLE for α is*

$$\widehat{\alpha}_T^{(4)} = \frac{S_T + \beta\int_0^T P_H(t)\,dw_t^H}{w_T^H}\,\gamma. \tag{13}$$

It is unbiased, strongly consistent and normal:

$$T^{1-H}\left(\widehat{\alpha}_T^{(4)} - \alpha\right) \overset{d}{=} \mathcal{N}\left(0, \lambda_H\gamma^2\right).$$

Theorem 5 ([4, Theorem 3.2])**.** *Let $H > 1/2$ and α is known. The MLE for β is*

$$\widehat{\beta}_T^{(5)} = \frac{\frac{\alpha}{\gamma}\int_0^T P_H(t)\,dw_t^H - \int_0^T P_H(t)\,dS_t}{\int_0^T (P_H(t))^2\,dw_t^H}. \tag{14}$$

It is strongly consistent and asymptotically normal:

$$\sqrt{T}\left(\widehat{\beta}_T^{(5)} - \beta\right) \xrightarrow{d} \mathcal{N}(0, 2\beta).$$

Theorem 6 ([4, Theorem 3.4])**.** *Let $H > 1/2$. The MLEs for α and β equal*

$$\begin{aligned}\widehat{\alpha}_T^{(6)} &= \frac{\int_0^T P_H(t)\,dS_t \int_0^T P_H(t)\,dw_t^H - S_T\int_0^T (P_H(t))^2\,dw_t^H}{\left(\int_0^T P_H(t)\,dw_t^H\right)^2 - w_T^H\int_0^T (P_H(t))^2\,dw_t^H}\,\gamma,\\ \widehat{\beta}_T^{(6)} &= \frac{w_T^H\int_0^T P_H(t)\,dS_t - S_T\int_0^T P_H(t)\,dw_t^H}{\left(\int_0^T P_H(t)\,dw_t^H\right)^2 - w_T^H\int_0^T (P_H(t))^2\,dw_t^H}.\end{aligned} \tag{15}$$

They are consistent and asymptotically normal:

$$T^{1-H}\left(\widehat{\alpha}_T^{(6)} - \alpha\right) \xrightarrow{d} \mathcal{N}(0, \lambda_H\gamma^2), \qquad \sqrt{T}\left(\widehat{\beta}_T^{(6)} - \beta\right) \xrightarrow{d} \mathcal{N}(0, 2\beta).$$

It can be seen that both estimators $\widehat{\alpha}_T^{(6)}$ and $\widehat{\beta}_T^{(6)}$ depend on four elements: S_T, $\int_0^T P_H(t)\,dS_t$, $\int_0^T P_H(t)\,dw_t^H$ and $\int_0^T (P_H(t))^2\,dw_t^H$. Calculation of their joint moment generating function [3, Theorem 3.4] gives the following result.

Theorem 7 ([3, Theorem 4.2])**.** *Let $H > 1/2$. The vector maximum likelihood estimator $\left(\widehat{\alpha}_T^{(6)}, \widehat{\beta}_T^{(6)}\right)$ for vector parameter (α, β) is asymptotically normal:*

$$\begin{bmatrix} T^{1-H}\left(\widehat{\alpha}_T^{(6)} - \alpha\right) \\ \sqrt{T}\left(\widehat{\beta}_T^{(6)} - \beta\right) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda_H \gamma^2 & 0 \\ 0 & 2\beta \end{bmatrix}\right), \quad T \to \infty, \tag{16}$$

hence estimators $\widehat{\alpha}_T^{(6)}$ and $\widehat{\beta}_T^{(6)}$ are asymptotically independent.

Acknowledgements: I would like to thank Kostiantyn Ralchenko for an invaluable constant support throughout whole my research and Olga Zhuchenko for the theorem about discretized estimators.

Bibliography

[1] C. Berzin, A. Latour, J.R. León. *Inference on the Hurst Parameter and the Variance of Diffusions Driven by Fractional Brownian Motion.* Springer, 2014.

[2] M. Kleptsyna, A. Le Breton, M.-C. Roubaud. Parameter estimation and optimal filtering for fractional type stochastic systems. *Stat. Inference Stoch. Process.*, 3:173–182, 2000.

[3] S.S. Lohvinenko, K.V. Ralchenko. Asymptotic distribution of maximum likelihood estimator in fractional Vasicek model. *Teor. Imovir. Mat. Stat.*, 99:134–151, 2018 (In Ukrainian).

[4] S.S. Lohvinenko, K.V. Ralchenko. Maximum likelihood estimation in the fractional Vasicek model. *Lithuanian J. Statist.*, 56(1):77–87, 2017.

[5] S.S. Lohvinenko, K.V. Ralchenko, O.M. Zhuchenko. Asymptotic properties of parameter estimators in fractional Vasicek model. *Lithuanian J. Statist.*, 55(1):102–111, 2016.

[6] O. Vasicek. An equilibrium characterization of the term structure. *J. Finance Econ.*, 5(2):177–188, 1977.

Semi-Markov Processes in Reliability: Theory and Applications

Andreas Makrides,[1*] **Alex Karagrigoriou**[2] **and Vlad Stefan Barbu**[1]

[1]*Université de Rouen, Laboratoire de Mathématiques Raphaël Salem, UMR 6085, Avenue de l'Université, BP.12, F76801 Saint-Étienne-du-Rouvray, France*
[2]*Lab of Statistics and Data Analysis, Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Karlovasi, Samos 83200, Greece*

Abstract: This work deals with multi state systems that we model by means of semi-Markov processes. The main characteristic of this work is that the sojourn times in a given state which are seen to be independent not identically distributed random variables are assumed to belong to two different general classes of distributions. The first class of distributions is closed under maxima and contains distributions, like the Bernoulli distribution, the power function distribution and the extreme value Type I distribution, while the second is closed under minima and includes the exponential, the Weibull, the Pareto, the Rayleigh and the Erlang truncated exponential distribution [1]. Here we concentrate only on the first class of distributions and we obtain maximum likelihood estimators of the parameters of interest and investigate their asymptotic properties. Furthermore, plug-in type estimators are furnished for various reliability indices related to the system under study.

Conclusively, our main objective is the proposal of parsimonious modeling for multi-state systems under a semi-Markov framework. Thus, we introduce a useful and powerful tool with a reduced number of parameters, which can be of great importance from a practical point of view.

Keywords: multi-state system, reliability theory, survival analysis, semi-Markov processes, parameter estimation

AMS subject classification: 60J27, 60K15, 60K20, 62N02, 62N05, 90B25

[*] Corresponding author: antreas.makridis@univ-rouen.fr

1 Introduction

Let us consider a multi-state system with state space $E = \{1, 2, \ldots, N\}$, defined on a probability space $(\Omega, \mathcal{F}, P)$. As it will introduced in next Section the evolution of the system is assumed to follow a continuous time semi-Markov process.

The main feature of this work is that the sojourn times between states are assumed to belong to two different general classes of distributions. The first class of distributions is closed under maxima and contains several distributions, like the Bernoulli distribution, the power function distribution and the extreme value Type I distribution. The second class is closed under minima and includes the exponential, the Weibull, the Pareto, the Rayleigh and the Erlang truncated exponential distribution (cf. [1]). However, here we concentrated only on the first class of distributions.

The outline of the paper is as follows. Some preliminaries regarding semi-Markov processes and multi-state systems for a family of distributions closed under maxima are presented in Section 2. Section 3 provides the likelihood function and the associated maximum likelihood estimators of the parameters under investigation. Section 4 is devoted to study the Markov renewal function and the semi-Markov renewal matrix. Finally, Section 5 presents simulation results for evaluating the accuracy of the proposed methodology.

2 Semi-Markov Processes and Multi State Systems

Semi-Markov (SM) processes are typical tools for the modeling of technical systems. Such classes of stochastic processes generalize typical Markov jump processes by allowing general distributions for sojourn times [4].

Assume that the time evolution of the system is governed by a stochastic process $Z = (Z_t)_{t\in\mathbb{R}_+}$. Let $S = (S_n)_{n\in\mathbb{N}}$ be the successive time points when state changes in $(Z_t)_{t\in\mathbb{R}_+}$ occur and $J = (J_n)_{n\in\mathbb{N}}$ the successive visited states at these time points. Furthermore, $X = (X_n)_{n\in\mathbb{N}}$ are the successive sojourn times in the visited states. Therefore, $X_n = S_n - S_{n-1}$, $n \in \mathbb{N}^*$, and, by convention, we set $X_0 = S_0 = 0$.

We assume that (J, S) is a *Markov renewal process* (cf. [4]) and $Z = (Z_t)_{t\in\mathbb{R}_+}$ is a *semi-Markov (SM) process* associated to (J, S), where $Z_t := J_{N(t)}$, with $N(t) := \max\{n \in \mathbb{N} \mid S_n \leq t\}$, $t \in \mathbb{R}_+$. A SM model is characterized by its *initial distribution* $\alpha = (\alpha_1, \ldots, \alpha_N)$, $\alpha_j := \mathbb{P}(J_0 = j)$, $j \in E$, and by the *semi-Markov kernel* $Q_{ij}(t) := \mathbb{P}(J_n = j, X_n \leq t | J_{n-1} = i)$. Let us also introduce the *transition probabilities* of the embedded Markov chain $(J_n)_{n\in\mathbb{N}}$, $p_{ij} := \mathbb{P}(J_n = j | J_{n-1} = i) = \lim_{t\to\infty} Q_{ij}(t)$, and the *conditional sojourn time distribution functions*

$$W_{ij}(t) := \mathbb{P}(S_n - S_{n-1} \leq t | J_{n-1} = i, J_n = j)\mathbb{P}(X_n \leq t | J_{n-1} = i, J_n = j). \quad (1)$$

Observe that $Q_{ij}(t) = p_{ij} W_{ij}(t)$.

Let T_{ij} be a potential time spent in state i before moving (directly) to state j. We denote by $F_{ij}(t; \theta_{ij})$ its cumulative distribution function (cdf), where θ_{ij} is the m-dimensional parameter involved in the underlying distribution. We assume that the distribution of T_{ij} is absolutely continuous with respect to the Lebesgue measure; an associated density is denoted by $f_{ij}(t;\theta_{ij})$.

The dynamic of the system is as follows: the next state to be visited after state i is the one for which T_{il} is the maximum. Thus, for our semi-Markov system, the semi-Markov kernel becomes

$$Q_{ij}(t) = \mathbb{P}(\max_l T_{il} \leq t \text{ \& the max occurs for } j | J_{n-1} = i) = p_{ij} W_{i\bullet}(t),$$

where $\quad p_{ij} = \mathbb{P}(J_n = j | J_{n-1} = i) = \mathbb{P}(T_{ij} \geq T_{il}, \forall l | J_{n-1} = i) \quad$ and

$$W_{ij}(t) = \mathbb{P}(\max_l T_{il} \leq t | J_{n-1} = i) =: W_{i\bullet}(t), \text{ independent of } j,$$

is the cdf of the sojourn time in state i (unconditional to the next state to be visited). Note that $\sum_j Q_{ij}(t) = W_{i\bullet}(t)$, where $W_{i\bullet}(t)$ is absolutely continuous w.r.t. the Lebesgue measure and has a density denoted by $f_{i\bullet}(t)$.

INID RANDOM VARIABLES

As mentioned earlier, we consider the class of distributions closed under maxima, focus on independent but not necessarily identically distributed (*inid*) random variables and consider the case where the distributions $F_{ij}(\cdot;\theta_{ij})$, $i, j = 1, \ldots, N$, are of the same functional form but with different parameters. A member of this class with parameter a is assumed to verify [1]

$$F(t;a) := (F(t;1))^a \,, \tag{2}$$

$F(t;a)$ is absolutely continuous w.r.t. Lebesgue measure with density $f(t;a)$.

The above class includes the Bernoulli, the power function and the extreme value Type I distributions. A representative example comes from structural engineering where engineers are interested in stress and strain diagrams that graphically display the basic material characteristics when designing various types of constructions like bridges, highways or buildings.

The following result, shows that the maximum order statistic from an *inid* random sample from the above class has a distribution belonging to the same class. More precisely, the above class of distributions is closed under maxima.

Lemma 1. *Let $X_1, \ldots, X_N$ be inid random variables such that $X_i \sim F(x;a_i)$ which belongs to class (2), $i = 1, 2, \ldots, N$. Then, the distribution function $F^{(N)}$ of the maximum order statistic $X_{(N)}$ belongs also to (2) (cf. [1]).*

Under the above class of distributions, the following result concerning the main semi-Markov characteristics can be proved. For notational convenience, we set $F(t) := F(t;1)$, $f(t) := f(t;1)$ and $Q_{ij}\left(t;a_{im};m = 1, \ldots, N\right) := Q_{ij}(t)$.

Note that, the dependence of semi-Markov kernel to a_{im} is due to the fact that the parameter of $F(\cdot)$ included in the semi-Markov kernel is $\sum\limits_{m\in E} a_{im}$.

Proposition 1. *Under the setup of this section, the following results hold:*

$$Q_{ij}(t) = \frac{a_{ij}}{\sum\limits_{m\in E} a_{im}} \cdot [F(t)]^{\sum\limits_{m\in E} a_{im}}, \quad p_{ij} = \frac{a_{ij}}{\sum\limits_{m\in E} a_{im}}, \tag{3}$$

$$W_{i\bullet}(t) = [F(t)]^{\sum\limits_{m=1}^{N} a_{im}} \quad \text{and} \quad f_{i\bullet}(t) = \sum_{m=1}^{N} a_{im}\, [F(t)]^{\sum\limits_{m=1}^{N} a_{im}} \frac{f(t)}{F(t)}. \tag{4}$$

3 Maximum Likelihood Estimation

For estimation purposes, one sample path as well as several sample paths are considered. On each situation we investigate both the (right) censored and the uncensored cases. However, here we deal with the general case where some of the sojourn times are censored either at the beginning and/or at the end for several trajectories. Given L censored sample paths,
$\left\{ {x_0^{(l)}}^{\delta_b^{(l)}}, j_0^{(l)}, x_1^{(l)}, j_1^{(l)}, x_2^{(l)}, \ldots, j_{N^l(M)}^{(l)}, {u_M^{(l)}}^{\delta_e^{(l)}} \right\}$, $l = 1, \ldots, L$,
where $\delta_b^{(l)}, \delta_e^{(l)}$ take the values 1 or 0 if we have censoring or not, respectively. Then the associated likelihood is

$$\mathcal{L} = \left(\prod_{i\in E} \alpha_i^{N_{i,0}(L)} \right) \left(\prod_{i,j\in E} p_{ij}^{\sum\limits_{l=1}^{L} N_{ij}^{(l)}(M)} \right) \left(\prod_{l=1}^{L} \prod_{i\in E} \prod_{k=1}^{N_i^{(l)}(M)} f_{i\bullet}(x_i^{(l,k)}) \right) \times$$
$$\times \left(\prod_{i\in E} \prod_{k=1}^{\overline{N}_{i\bullet}^{b}(L)} \overline{W}_{i\bullet}(x_{i,0}^{(k)}) \right) \left(\prod_{i\in E} \prod_{k=1}^{\overline{N}_{i\bullet}^{e}(L)} \overline{W}_{i\bullet}(x_{i,M}^{(k)}) \right). \tag{5}$$

The MLE $\hat{a}_{ij}(L, M)$ can be easily obtained from the above equation.

4 Markov Renewal Function and semi-Markov Transition Matrix

The *Markov renewal function* $\Psi_{ij}(t)$, $i, j \in E$, $t \geq 0$, is defined as (cf. [4])

$$\Psi_{ij}(t) := \mathbb{E}_i[N_j(t)] = \sum_{n=0}^{\infty} \sum_{k\in E} \int_0^t Q_{ik}(ds) Q_{kj}^{(n-1)}(t-s) \tag{6}$$

The *semi-Markov transition matrix (function)* is defined as

$$P_{ij}(t) := \mathbb{P}(Z_t = j | Z_0 = i), i, j \in E. \tag{7}$$

Let $W(t) := diag\left(W_i(t); i \in E\right) = diag\left(\sum_j Q_{ij}(t); i \in E\right) = diag\left(Q \cdot \mathbf{1}_N\right)(t)$ be the diagonal matrix with the (i,i) element equal to $W_i(t) = \sum_{j \in E} Q_{ij}(t)$, where $\mathbf{1}_N = (\underbrace{1, \cdots, 1}_{N})^\top$, $()^\top$ denoting the transposed of a vector.

Then the semi-Markov transition matrix is given by [4]

$$P(t) = \left((I_N - Q)^{(-1)} \star (I_N - W)\right)(t) = \left(\Psi \star (I_N - W)\right)(t), \tag{8}$$

where $\Psi(t) = \left(\Psi_{ij}(t)\right)_{i,j \in E}$ and it is shown that $(I_N - Q)^{(-1)}(t) = \Psi(t)$.
Using the expressions given in (3), (4) and the estimators of the previous Section, the estimators of $\Psi(t)$ and $P(t)$ are easily obtained.

5 Simulations

A series of simulations in R is analyzed to evaluate the accuracy of the proposed procedure. The case of L sample paths of a semi-Markov process with censoring rate 50% at the beginning and/or at the end is considered with $M = 1000$. The sojourn times follow the Power Function distribution
$f(t; a_{ij}) = \frac{a_{ij}}{c^{a_{ij}}} t^{a_{ij}-1}, \quad 0 \leq t \leq c = 100$, with $a_{ii} = 0$, $a_{12} = 0.5$, $a_{13} = 1$, $a_{21} = 1.2$, $a_{23} = 0.9$, $a_{31} = 1.4$, $a_{32} = 1.5$, $i = 1, 2, 3$.

L S.E.	5	10	100	1000
$\hat{a}_{ij}(L,M)$	2.423613	2.98703×10^{-1}	6.50080×10^{-2}	4.48959×10^{-2}
$\hat{p}_{ij}(L,M)$	2.67957×10^{-2}	1.95020×10^{-2}	7.11892×10^{-2}	4.09628×10^{-4}

Table 1: Standard errors of $\hat{a}_{ij}(L,M)$ and $\hat{p}_{ij}(L,M)$ for various L.

t S.E.	1	25	50	100
$\hat{P}_{ij}(t;L,M)$	1.61768×10^{-7}	5.61013×10^{-5}	2.23401×10^{-3}	9.09540×10^{-4}

Table 2: Standard errors of the estimators $\hat{P}_{ij}(t;L,M)$ for various t.

The results show that the S.E. decrease as L increases which implies a better accuracy for $\hat{a}_{ij}(L,M)$ and $\hat{p}_{ij}(L,M)$. Furthermore, $\hat{P}_{ij}(t;L,M)$ is extremely good as expected, for small values of t it starts deteriorating as t increases and becomes better again as t approaches the upper limit.

Bibliography

[1] K. Balasubramanian, M.-I. Beg and R.-B. Bapat. On families of distributions closed under extrema. *Sankhya A*, 53:375–388, 1991.

[2] V.-S. Barbu, C. Brard, D. Cellier, M. Sautreuil and N. Vergne. Parametric estimation of semi-Markov chains (submitted), 2017.

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

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

On the Convergence and Robustness of Mean-vector Estimation of Inlier Distribution

Arshak Minasyan[1*]

[1] *Department of Mathematics and Mechanics Yerevan State University, YerevaNN Research Lab*

Abstract: Collier and Dalalyan [2] study the problem of common p-dimensional mean vector estimation of inliers among n independent Gaussian vectors by iteratively using soft-thresholding operator. The presented method is the approximation of the solution of a non-convex optimization problem involving the Huber function. We simplify this problem and reduce it to the Fermat-Weber location problem using Huber function instead of Euclidean distance. A modified version of the Iteratively Re-weighted Least Squares (IRLS) method is presented that minimizes the resulting objective function along with a global convergence proof given that the starting point is chosen accordingly. We illustrate the robustness of the resulting estimator through numerical experiments and examples, which are nicely consistent with theoretical results.

Keywords: Huber function, Iteratively Re-weighted Least Squares (IRLS), Fermat-Weber problem, Robustness.

AMS subject classification: 62F35.

1 Introduction

The problem of finding a point that minimizes the sum of the distances to the given points $Y_1, \dots, Y_N$ is known as the Fermat-Weber problem [1] and is formulated as follows

$$\min_{x \in \mathbf{R}^p} \sum_{i=1}^{N} \|Y_i - x\|_2. \tag{1}$$

It can be easily seen that the closed form expression cannot be found for the solution of (1). However, the problem is convex and can be efficiently solved numerically. A method that does so with provable convergence is known as Iteratively Re-weighted Least Squares (IRLS) [4]. The idea of IRLS is to approximate each summand of (1)

[*] Corresponding author: arsh.minasyan@gmail.com

with a quadratic function, which always has a higher objective value. The analysis of IRLS method shows that the method converges globally.
In this paper we introduce a modified version of the Fermat-Weber problem, in which we replace the ℓ_2 norm with a Huber function. The motivation behind this relaxation comes from a statistical viewpoint. Imagine the problem of mean value estimation in a setup where the data is corrupted by some small number of outliers. Even a very small number of outliers can cause large bias in the standard framework, even though the posed optimization problem will be solved accurately. Hence, we choose Huber function as a robust function to solve the corresponding optimization problem.

2 Fermat-Weber Location Problem with Huber Function

Consider the following optimization problem

$$x^\star = \arg\min_{x \in \mathbf{R}^p} \sum_{i=1}^{N} \rho_\delta(\|Y_i - x\|_2) \tag{2}$$

for a set of points $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_N\}$ with $Y_i \in \mathbf{R}^p$ and the Huber function ρ_δ defined as in [3]

$$\rho_\delta(x) = \begin{cases} x^2/2, & \text{if } |x| \leq \delta, \\ \delta|x| - \delta^2/2, & \text{o.w.} \end{cases} \tag{3}$$

The standard IRLS method [4] uses the following simple idea: at iteration k, approximate the objective function with a quadratic function so that the values of the function and its derivative are equal at the current point $x^{(k)}$. Further in this section we show how IRLS method can be modified in order to be applicable to (2) preserving the desirable properties of the IRLS estimator.
For $x^{(k)} \neq Y_i$, let $t_{ik} = Y_i - x^{(k)}$ and define the following function

$$g_k^i(x) = \begin{cases} \frac{1}{2}\|t_{ik}\|_2^2, \text{ if } \|t_{ik}\|_2 \leq \delta, \\ \frac{\delta}{2}\left(\frac{1}{\|t_{ik}\|_2}\|t_{ik}\|_2^2 + \|t_{ik}\|_2 - \delta\right), \text{ o.w.} \end{cases} \tag{4}$$

It is straightforward to check that at point $x^{(k)}$ the values and derivatives of $\rho_\delta(\|Y_i - x\|_2)$ and $g_k^i(x)$ coincide. Hence, instead of solving (2) directly we minimize the following function

$$\min_{x \in \mathbf{R}^p} \sum_{i=1}^{N} w_{ik}\|Y_i - x\|_2^2 + \sum_{i \in \mathcal{O}_k} \frac{\delta}{2}(\|t_{ik}\|_2 - \delta), \tag{5}$$

where $\mathcal{O}_k := \{i \in [N] : \|t_{ik}\|_2 > \delta\}$ and w_{ik} are defined as follows

$$w_{ik} = \begin{cases} \frac{1}{2}, \text{ if } \|t_{ik}\|_2 \leq \delta, \\ \frac{\delta}{2\|t_{ik}\|_2}, \text{ o.w.} \end{cases} \tag{6}$$

Notice that (5) is equivalent to

$$\min_{x\in\mathbf{R}^p}\sum_{i=1}^{N} w_{ik}\|Y_i - x\|_2^2, \tag{7}$$

the solution of which reads as

$$x_k^\star := \frac{\sum_{i=1}^{N} w_{ik}Y_i}{\sum_{i=1}^{N} w_{ik}} = \arg\min_{x\in\mathbf{R}^p}\sum_{i=1}^{N} w_{ik}\|Y_i - x\|_2^2.$$

However, we need to do something more for achieving the global convergence property. The idea is that we do not want to make too large steps, which formally translate into the following two rules:

1. if $\|t_{ik}\|_2 > \delta$ and $\|Y_i - x_k^\star\|_2 \le \delta$, instead of $x^{(k+1)} = x_k^\star$ we take $x^{(k+1)} = \tilde{x}_k$, where $\tilde{x}_k$ is defined as follows:
$$\tilde{x}_k \in \alpha x^{(k)} + (1-\alpha)x_k^\star \quad \text{for some } \alpha \in (0,1) \text{ and } \|Y_i - \tilde{x}_k\|_2 = \delta. \tag{8}$$
2. if $\|t_{ik}\|_2 \le \delta$ and $\|Y_i - x_k^\star\|_2 > \delta$, instead of $x^{(k+1)} = x_k^\star$ we take $x^{(k+1)} = \tilde{x}_k$, where $\tilde{x}_k$ is defined exactly the same way as in (8).

For the rest of the cases the update rule reads as follows:

$$x^{(k+1)} = \frac{\sum_{i=1}^{N} w_{ik}Y_i}{\sum_{i=1}^{N} w_{ik}}. \tag{9}$$

Denote

$$f_k(x) = \sum_{i=1}^{N} w_{ik}\|Y_i - x\|_2^2 + \sum_{i\in\mathcal{O}_k}\frac{\delta}{2}(\|t_{ik}\|_2 - \delta), \quad f(x) = \sum_{i=1}^{N}\rho_\delta(\|Y_i - x\|_2). \tag{10}$$

Lemma 1. *There exists $\alpha \in (0,1)$ such that for $\tilde{x}^{(k)} = \alpha x^{(k)} + (1-\alpha)x_k^\star$ it holds*

$$f(\tilde{x}^{(k)}) \le f_k(\tilde{x}^{(k)}) \le f_k(x^{(k)}) = f(x^{(k)}). \tag{11}$$

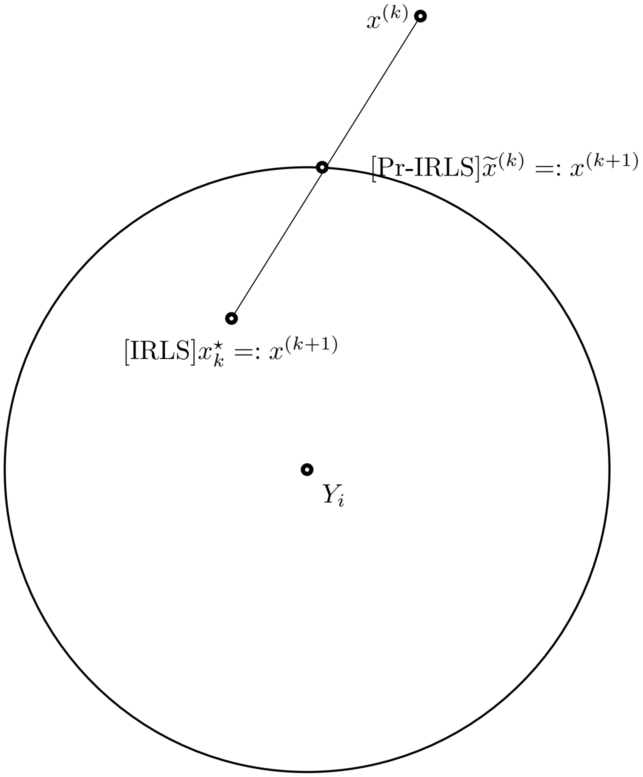
Proof. Notice that the inequality $f_k(x) \le f_k(x^{(k)})$ is true for all $x := \alpha x^{(k)} + (1-\alpha)x_k^\star$, due to the convexity of $f_k(\cdot)$, where $\alpha \in (0,1)$. Next, we will show that there is a point on the line connecting $x^{(k)}$ and $x_k^\star$ such that the first inequality holds as well. From the definition of $g_k^i(\cdot)$ it follows that for the cases when for some $i \in [n]$ either $\|t_{ik}\|_2 \le \delta$ and $\|t_{ik+1}\|_2 \le \delta$ or $\|t_{ik}\|_2 > \delta$ and $\|t_{ik+1}\|_2 > \delta$, then (11) is indeed true, since for all $i \in [n]$ it holds $\rho_\delta(\|Y_i - x^{(k+1)}\|_2) \le w_{ik}\|Y_i - x^{(k+1)}\|_2^2 + \frac{\delta}{2}(\|t_{ik}\|_2 - \delta)$ then $f(x^{(k+1)}) \le f_k(x^{(k+1)})$. In other cases when $\|t_{ik}\|_2 \le \delta$ and $\|t_{ik+1}\|_2 > \delta$ or $\|t_{ik}\|_2 > \delta$ and $\|t_{ik+1}\|_2 \le \delta$, we take the projection of $x_k^\star$ onto the ball $\|Y_i - x\|_2 \le \delta$. Now, according to (8) we will get $\alpha_1, \ldots, \alpha_p$, where p is the number of cases for which the projection is needed. Notice that $\alpha = \max\{\alpha_1, \ldots, \alpha_p\}$ we obtain $\tilde{x}_k = \alpha x^{(k)} + (1-\alpha)x_k^\star$, for which $\|t_{ik}\|_2 \le \delta$ and $\|Y_i - \tilde{x}_k\|_2 \le \delta$ or $\|t_{ik}\|_2 > \delta$ and $\|Y_i - \tilde{x}_k\|_2 > \delta$ is true for all $i \in [n]$, which completes the proof.

□

Remark 1. The resulting $\alpha = \max\{\alpha_1, \dots, \alpha_p\}$ will be projected only on the ball that is the closest to the point $x^{(k)}$, however ensuring that both $x^{(k)}$ and $\tilde{x}^{(k)}$ are either inside or outside of all balls $\|Y_i - x\|_2 \leq \delta$, where $i \in [p]$.

The following lemma can be easily verified and the formal proof is omitted from the paper.

Lemma 2. *If $x^{(k)} = x^\star$ then $x^{(k+1)} = x^\star$ and if $x^{(k)} \neq Y_i$ for all $i \in [N]$ and $x^{(k+1)} = x^{(k)}$ then $x^{(k)} = x^\star$.*


$x^{(k)}$
[Pr-IRLS]$\widetilde{x}^{(k)} =: x^{(k+1)}$
[IRLS]$x^\star_k =: x^{(k+1)}$
Y_i

Theorem 1. *For all but countable set of initial values $x^{(0)}$ and for all $i \in [N]$ if at each iteration $k \geq 1$ $x^{(k)} \neq Y_i$ then the above defined sequence $\{x^{(k)}\}_{k \geq 1}$ converges to $x^\star$.*

Proof. For all but countable set of initial values $x^{(0)}$, the sequence of $x^{(k)}$s lies inside the convex hull of $Y_1, \dots, Y_N$, which is a compact set. After applying Bolzano–Weierstrass theorem, we have $\lim_{l \to \infty} x_l^{(k)} = x$. Showing that $x \equiv x^\star$ will prove the theorem. If $x^{(k+1)} = x^{(k)}$ for some k and $x^{(k)} \neq Y_i$ for all $i \in [N]$, then by Lemma 2 $x^{(k)} = x^\star$. Otherwise, by Lemma 1 we have $f(x^{(0)}) > f(x^{(1)}) > f(x^{(2)}) > \cdots > f(x^{(k)}) > \cdots > f(x^\star)$. Hence, $\lim_{k \to \infty} f(x_l^{(k)}) - f(T(x_l^{(k)})) = 0$, where $T(\cdot)$ is the update rule described above. Then, the continuity of $T(\cdot)$ implies $\lim_{l \to \infty} T(x_l^{(k)}) = T(x)$, yielding $f(x) = f(T(x))$ and $T(x) = x$, if $x \neq Y_i$ for all

$i \in [N]$, then by Lemma 2 $x \equiv x^{\star}$. Given the assumption that no $x^{(k)}$ coincides with one of the given datapoints Y_i concludes the proof of the theorem. □

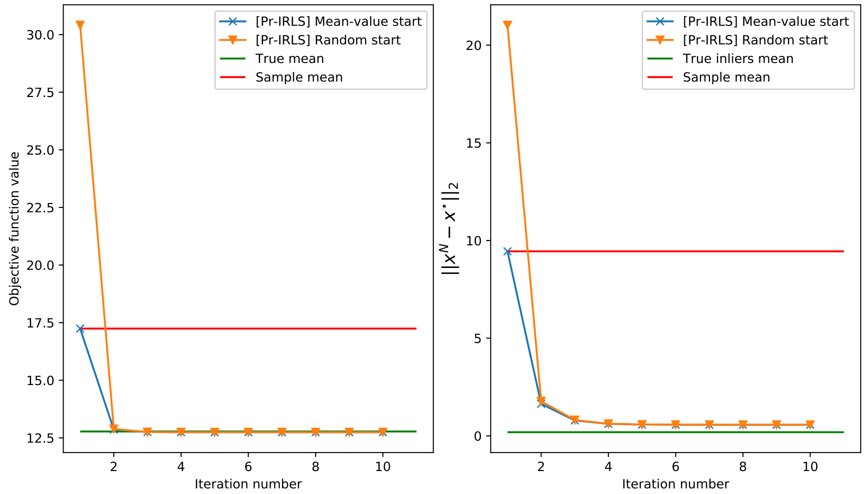


Figure 1: $N = 200, p = 30, \varepsilon = 0.2$. Inlier distribution: $\mathcal{N}(\mathbf{10}, I_p)$, Outlier distribution: $\mathcal{N}(\mathbf{10} + 5 \cdot \boldsymbol{\theta}, I_p)$ with $\boldsymbol{\theta} \in \mathbf{R}^p$ and $\|\boldsymbol{\theta}\|_0 \leq \varepsilon \cdot N$, non-zero elements of $\boldsymbol{\theta}$ are i.i.d. from $\mathcal{U}[0,1]$. Mean-value start: $x^{(0)} := N^{-1} \sum_{i=1}^{N} Y_i$, random start: $x^{(0)} \sim \mathcal{N}(\mathbf{0}, I_p)$.

The update rule for $\|t_{ik}\|_2 > \delta$ and $\|Y_i - x_k^{\star}\|_2 \leq \delta$ is provided in Figure 2. Figure 1 illustrates the convergence of the described method on a simulated dataset. In the plots we indicated the true and sample mean values to show that the obtained solution can be seen as a robust mean estimator for inliers.

Bibliography

[1] J. Brimberg. The Fermat-Weber location problem revisited. *Mathematical Programming,* 71*, 71–76*, 1995.

[2] O. Collier and A. Dalalyan. Minimax estimation of a p-dimensional linear functional in sparse Gaussian models and robust estimation of the mean. *arXiv:1712.05495v3*, 2018.

[3] P. J. Huber. Robust estimation of a location parameter. *The annals of mathematical statistics, 35(1) : 73–101*, 1964.

[4] E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnes est minimum. *Tohoku Mathematical Journal.,* 43*, 355–396*, 1937.

Shape and Order Constraints in Nonparametric Regression

Alexandre Mösching[1][*]

[1] *University of Bern*

Abstract: Imposing a nonparametric shape constraint in a statistical model has shown its benefit on several occasions, for example in circumstances where a parametric model is hard to justify but a shape constraint on the distribution is natural. We consider constraints on an unknown family of distributions $(F_x)_{x\in\mathbb{X}}$, with a fixed subset $\mathbb{X} \subset \mathbb{R}$, and discuss nonparametric estimation procedures based on a sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ such that, conditional on the X_i, the Y_i are independent random variables with distribution functions F_{X_i}.

Keywords: Likelihood ratio order, log-concavity, shape constraint.

AMS subject classification: 62G05, 62G07, 62G08.

1 Introduction

For any fixed set $\mathbb{X} \subset \mathbb{R}$, let $(F_x)_{x\in\mathbb{X}}$ be a family of conditional distribution functions of the form $F_x(y) := \mathbb{P}(Y \le y \,|\, X = x)$, for $(x, y) \in \mathbb{X} \times \mathbb{R}$. Suppose that we observe $n \ge 1$ pairs

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \;\in\; \mathbb{X} \times \mathbb{R},$$

such that, conditional on $X_1, X_2, \ldots, X_n$, the responses $Y_1, Y_2, \ldots, Y_n$ are independent with respective distribution functions $F_{X_1}, F_{X_2}, \ldots, F_{X_n}$.

In this article, we treat $(F_x)_{x\in\mathbb{X}}$ as unknown and investigate nonparametric estimation methods for the family of distribution functions, quantiles or densities under various shape constraints.

[*] Corresponding author: alexandre.moesching@stat.unibe.ch

Usual Stochastic Order. The family $(F_x)_{x\in\mathbb{X}}$ is stochastically ordered if:

$$\textit{For any fixed } y \in \mathbb{R}, F_x(y) \textit{ is decreasing in } x \in \mathbb{X}. \tag{1}$$

Such a stochastic order appears natural on several occasions. For example, an employee's income Y presumably increases with their age X. In forecasting, the measured cumulative precipitation amount Y is expected to increase with the numerical predictions X of the same quantity.

The above stochastic ordering constraint has a dual characterization in terms of the minimal and maximal β-quantiles of F_x, respectively defined by $F_x^{-1}(\beta) := \min\{y \in \mathbb{R} : F_x(y) \geq \beta\}$ and $F_x^{-1}(\beta+) := \inf\{y \in \mathbb{R} : F_x(y) > \beta\}$, for each $x \in \mathbb{X}$ and $\beta \in (0,1)$. More precisely, (1)⇔(a)⇔(b) with:

(a) $F_x^{-1}(\beta)$ is increasing in $x \in \mathbb{X}$ for any fixed $\beta \in (0,1)$.

(b) $F_x^{-1}(\beta+)$ is increasing in $x \in \mathbb{X}$ for any fixed $\beta \in (0,1)$.

Let $Q_x(\beta)$ be any β-quantile of F_x and assume that it is increasing in $x \in \mathbb{X}$.

The estimation of $(F_x)_{x\in\mathbb{X}}$ under constraint (1) can be done via nonparametric monotone least squares, as discussed in [3], with the estimator $(\hat{F}_x)_{x\in\mathbb{X}}$ typically computed with the pool-adjacent-violators algorithm [8]. On the other hand, nonparametric monotone regression quantiles gives estimators $(\hat{Q}_x)_{x\in\mathbb{X}}$ of $(Q_x)_{x\in\mathbb{X}}$, see [4].

In our manuscript [6], we give detailed descriptions of each of the aforementioned estimators and prove that the quantiles of $(\hat{F}_x)_{x\in\mathbb{X}}$ yield a large family of estimated quantile curves containing the estimators $(\hat{Q}_x)_{x\in\mathbb{X}}$, but also smoother ones. This shows that the estimators $\hat{Q}_x$ are consistent with the quantiles of $\hat{F}_x$ and that these two estimation techniques provide two equivalent ways to characterize the unknown distribution.

Likelihood Ratio Order. Suppose that F_x has density f_x w.r.t. Lebesgue measure for each $x \in \mathbb{X}$. Then, the family $(F_x)_{x\in\mathbb{X}}$ is increasing in the likelihood ratio order if:

$$\textit{For any fixed } x_1, x_2 \in \mathbb{X}, x_1 < x_2, \frac{f_{x_2}(y)}{f_{x_1}(y)} \textit{ is increasing in } y \in \mathbb{R}. \tag{2}$$

One can verify that (2) implies (1). This constraint finds its interest for example in discriminant analysis, where the confidence that an observation y comes from population F_{x_2} rather than F_{x_1} increases with $f_{x_2}(y)/f_{x_1}(y)$. In statistical testing, uniformly most powerful tests can be build when the likelihood ratio is increasing in y, see [5].

The estimation of $(f_x)_{x\in\mathbb{X}}$ under constraint (2) has been studied in the two-sample case by [10], that is, when $\#\mathbb{X} = 2$. Their estimator is nevertheless not fully nonparametric as they maximize a smoothed log-likelihood which necessitates a choice of a kernel function and its bandwidth. This method was chosen because the standard log-likelihood is unbounded otherwise, which may explain why the case of general subsets $\mathbb{X} \subset \mathbb{R}$ has not received any attention in the literature.

2 Likelihood Ratio Order with a Shape Constraint

In our ongoing work [7], we impose a log-concave shape constraint on the family of densities $(f_x)_{x\in\mathbb{X}}$. This additional qualitative constraint, combined with likelihood ratio ordering, enables us to get rid of any choice of tuning parameter. Furthermore, log-concave densities are of particular interest as they encompass a large family of parametric densities and are a subclass of unimodal densities. Log-concave density estimation is discussed in [2, 9].

In the present setting, the log-concavity constraint is set as follows: Suppose that $(F_x)_{x\in\mathbb{X}}$ is such that, for each $x \in \mathbb{X}$, F_x can be written as

$$F_x(y) \ = \ \int_{-\infty}^{y} \exp\varphi_x(s)\,\mathrm{d}s, \qquad \text{for } y \in \mathbb{R},$$

with $\varphi_x : \mathbb{R} \to [-\infty, \infty)$ a concave and upper semicontinuous (c.u.s.c.) function satisfying:

$$\textit{For any fixed } y \in \mathbb{R}, \varphi_x'(y+) \textit{ is increasing in } x \in \mathbb{X}. \tag{3}$$

The above right-sided derivative has the following meaning: Define for each $x \in \mathbb{X}$ the domain of φ_x as $\mathrm{dom}(\varphi_x) := \{y \in \mathbb{R} : \varphi_x(y) > -\infty\} \neq \emptyset$ and the right-sided derivative of φ_x at $y \in \mathbb{R}$ to be the usual right-sided derivative for $y \in \mathrm{dom}(\varphi_x)$, equal to ∞ for $y < \inf\mathrm{dom}(\varphi_x)$ and $-\infty$ for $y > \sup\mathrm{dom}(\varphi_x)$.

The equivalence between constraints (2) and (3) is immediate.

Maximum Likelihood Formulation

Without loss of generality, the set of data pairs $\mathcal{D} := \{(X_i, Y_i)\}_{1\le i\le n}$ can be restricted to fulfill a specific geometric property (see [7]) under which one shows that the normalized log-likelihood

$$l(\varphi) \ := \ \frac{1}{n}\sum_{i=1}^{n} \varphi_{X_i}(Y_i) \tag{4}$$

is bounded from above when restricted to the set

$$\Theta_1 \ := \ \big\{\varphi := (\varphi_x)_{x\in\mathbb{X}} : \varphi \text{ satisfy (3)}, \ \varphi_x \text{ is c.u.s.c. and } \textstyle\int \exp\varphi_x = 1\big\}.$$

Let us define the marginal datasets $\mathcal{X} := \{X_i\}_{1\le i\le n}$ and $\mathcal{Y} := \{Y_i\}_{1\le i\le n}$. Two issues can be identified if one were to estimate the unknown $\varphi := (\varphi_x)_{x\in\mathbb{X}}$ by a maximizer of (4) over Θ_1. First, such a maximizer is not necessarily unique, since the value of φ_x for $x \in \mathbb{X} \setminus \mathcal{X}$ is irrelevant for the computation of l, as long as it satisfies all the requirements of Θ_1. Secondly, it is not straightforward to deal with the constraint $\int \exp\varphi_x = 1$ as such. Therefore, we instead consider the modified log-likelihood

$$L(\varphi) \ := \ \sum_{r=1}^{R}\sum_{s=1}^{S} w_{r,s}\left(\varphi_{x_r}(y_s) - \int_{-\infty}^{\infty} \exp\varphi_{x_r}(y)\,\mathrm{d}y\right) \tag{5}$$

with $x_1 < \cdots < x_R$ the pairwise different elements of $\mathcal{X}$, $y_1 < \cdots < y_S$ those of $\mathcal{Y}$ and $w_{r,s}$ the relative frequency of (x_r, y_s) in the full sample. We then estimate φ by

$$\widehat{\varphi} \;:=\; \underset{\varphi\in\Theta}{\arg\max}\; L(\varphi), \tag{6}$$

with

$$\Theta \;:=\; \big\{\varphi := (\varphi_x)_{x\in\mathbb{X}} : \varphi \text{ satisfy (3) with } \mathbb{X} \text{ replaced by } \mathcal{X} \text{ and } \varphi_x \text{ is c.u.s.c.}\big\}.$$

One shows that $\widehat{\varphi}$ is indeed unique and that $L(\widehat{\varphi}) = \sup_{\varphi\in\Theta_1} l(\varphi) - 1$. Furthermore, for some $1 = s_1 \le \cdots \le s_R \le S$ and $1 \le S_1 \le \cdots \le S_R = S$ defined from the geometry of $\mathcal{D}$, one proves that for $1 \le r \le R$, $\widehat{\varphi}_{x_r}$ is linear on $[y_s, y_{s+1}]$, $s_r \le s < S_r$, and $\widehat{\varphi}_{x_r} \equiv -\infty$ outside of these intervals, see [7].

Finite Dimensional Constrained Optimization Problem

Define $\mathcal{P} := \cup_{r=1}^{R} \cup_{s=s_r}^{S_r} \{(r,s)\}$, $\mathbb{R}^{\mathcal{P}} := \{(\varphi_{r,s})_{(r,s)\in\mathcal{P}} : \varphi_{r,s} \in \mathbb{R}\}$ and $\mathbb{K}^{\mathcal{P}}$ to be the cone in $\mathbb{R}^{\mathcal{P}}$ of feasible elements $\boldsymbol{\varphi} := (\varphi_{r,s})_{(r,s)\in\mathcal{P}}$ such that

$$\frac{\varphi_{r,s+1} - \varphi_{r,s}}{y_{s+1} - y_s} \;\le\; \frac{\varphi_{r,s} - \varphi_{r,s-1}}{y_s - y_{s-1}}, \tag{7}$$

$$\frac{\varphi_{r,s+1} - \varphi_{r,s}}{y_{s+1} - y_s} \;\le\; \frac{\varphi_{r+1,s+1} - \varphi_{r+1,s}}{y_{s+1} - y_s}, \tag{8}$$

where (7) holds for $s_r < s < S_r, 1 \le r \le R$ and expresses the concavity constraint, and where (8) holds for $s_r \vee s_{r+1} \le s < S_r \wedge S_{r+1}, 1 \le r < R$, and stands for constraint (3). Then, problem (6) is identified with the finite dimensional constrained optimization problem

$$\underset{\boldsymbol{\varphi}\in\mathbb{K}^{\mathcal{P}}}{\arg\max}\; \mathcal{L}(\boldsymbol{\varphi}), \tag{9}$$

whose solution $\widehat{\boldsymbol{\varphi}}$ is unique and so that $\mathcal{L}(\widehat{\boldsymbol{\varphi}}) = L(\widehat{\varphi}) = \sup_{\varphi\in\Theta_1} l(\varphi) - 1$, with the target functional $\mathcal{L}$ being essentially a parametric version of (5).

An active set algorithm for $\widehat{\varphi}$

Active set methods are particularly well suited for optimization problems with inequality constraints on the parameter space. In [1] for example, an active set strategy was employed to estimate shape-constrained density ratios, including log-concave densities in the one-sample case. In our ongoing work [7], we extend this approach and build a specific algorithm to solve (9). The idea of the method reads as follows:

1. Start with a feasible $\boldsymbol{\varphi} \in \mathbb{K}^{\mathcal{P}}$ such that $\mathcal{L}(\boldsymbol{\varphi}) > -\infty$ and go to 2.

2. Define the set of deactivated equality constraints of $\boldsymbol{\varphi}$ as

   $$\begin{aligned} C(\boldsymbol{\varphi}) \;:=&\; \Big\{(r,s,y) \in \mathcal{P} \times \{y\} : \frac{\varphi_{r,s+1} - \varphi_{r,s}}{y_{s+1} - y_s} \neq \frac{\varphi_{r,s} - \varphi_{r,s-1}}{y_s - y_{s-1}}\Big\} \\ &\cup \Big\{(r,s,x) \in \mathcal{P} \times \{x\} : \frac{\varphi_{r,s+1} - \varphi_{r,s}}{y_{s+1} - y_s} \neq \frac{\varphi_{r+1,s+1} - \varphi_{r+1,s}}{y_{s+1} - y_s}\Big\}. \end{aligned}$$

   Check if $\boldsymbol{\varphi}$ is *locally optimal*, in the sense that

   $$\mathcal{L}(\boldsymbol{\varphi}) \;\geq\; \mathcal{L}(\boldsymbol{\psi}), \qquad \text{for all } \boldsymbol{\psi} \in \mathbb{R}^{\mathcal{P}} \text{ such that } C(\boldsymbol{\psi}) \subset C(\boldsymbol{\varphi}).$$

   To this end, define an element $\boldsymbol{\varphi}_{\mathrm{new}}$ as the result of a Newton step in a linear subspace of $\mathbb{R}^{\mathcal{P}}$ determined by $\boldsymbol{\varphi}$ and compute the directional derivative δ of $\mathcal{L}$ at $\boldsymbol{\varphi}$ along $\boldsymbol{\varphi}_{\mathrm{new}} - \boldsymbol{\varphi}$.

   If $\delta > 0$, replace $\boldsymbol{\varphi}$ by a convex combination of $\boldsymbol{\varphi}$ and $\boldsymbol{\varphi}_{\mathrm{new}}$ that makes it feasible in $\mathbb{K}^{\mathcal{P}}$ and restart with 2.

   Otherwise, if $\delta \leq 0$, then $\boldsymbol{\varphi}$ is locally optimal and one proceeds to 3.

   Each repetition of step 2. has the effect of strictly improving the likelihood score and possibly activating some constraints.

3. Suppose that $\boldsymbol{\varphi} \in \mathbb{K}^{\mathcal{P}}$ is locally optimal and check whether it is also *globally optimal*, in the sense that $\boldsymbol{\varphi} = \widehat{\boldsymbol{\varphi}}$.

   For this purpose, compute the directional derivative $\delta_{\boldsymbol{\varepsilon}}$ of $\mathcal{L}$ at $\boldsymbol{\varphi}$ in the direction of some simple elements $\boldsymbol{\varepsilon}$ which have the specificity to deactivate at least one constraint of $\boldsymbol{\varphi}$. Only finitely many of these specific elements exist and the search for $\boldsymbol{\varepsilon}$ such that $\delta_{\boldsymbol{\varepsilon}} > 0$ is done efficiently via dynamic programming.

   If such an $\boldsymbol{\varepsilon}$ exists, replace $\boldsymbol{\varphi}$ by $\boldsymbol{\varphi} + t\boldsymbol{\varepsilon}$ for some $t > 0$ which makes it feasible in $\mathbb{K}^{\mathcal{P}}$ and strictly increases the likelihood score. One then restarts with 2.

   Otherwise, if $\delta_{\boldsymbol{\varepsilon}} \leq 0$ for all $\boldsymbol{\varepsilon}$, return $\widehat{\boldsymbol{\varphi}} := \boldsymbol{\varphi}$.

In reality, the inequalities employed in the algorithm are treated up to some precision $\delta_o > 0$, so that after finitely many iterations, the procedure stops and provides an approximate solution $\widehat{\boldsymbol{\varphi}}$ to (9). Using linear interpolation, one constructs $\widehat{\varphi} \in \Theta_1$, yielding an approximate maximizer of (4).

Acknowledgements: This work was supported by Swiss National Science Foundation.

Bibliography

[1] L. Dümbgen, A. Mösching, and C. Strähl. Active set algorithms for estimating shape-constrained density ratios. Preprint available from arXiv: https://arxiv.org/abs/1808.09340, 2019.

[2] L. Dümbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009.

[3] H. El Barmi and H. Mukerjee. Inferences under a stochastic ordering constraint. *Journal of the American Statistical Association*, 100(469):252–261, 2005.

[4] R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[5] E. L. Lehmann, and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 2005.

[6] A. Mösching and L. Dümbgen. Monotone least squares and isotonic quantiles. Preprint available from arXiv: https://arxiv.org/abs/1901.02398, 2019.

[7] A. Mösching and L. Dümbgen. Strong stochastic order with a shape constraint. In preparation, 2019.

[8] T. Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*. Wiley, 1988.

[9] R. J. Samworth. Recent progress in log-concave density estimation. *Statistical Science*, 33(4):493–509, 2018.

[10] T. Yu, P. Li, and J. Qin. Density estimation in the two-sample problem with likelihood ratio ordering. *Biometrika*, 104(1):141–152, 2017.

From theory to application: a spatio-temporal modelling perspective

Michele Nguyen[1]*

[1] *Malaria Atlas Project, Oxford Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK*

Abstract: Gaussian random fields with Matérn covariances are the most popular modelling tools in Spatial Statistics. We draw on examples of recent advances in related theory as well as application. For the former, we look at ambit fields, a relatively new class of spatio-temporal random fields which were introduced for turbulence modelling. Since these are expressed as stochastic integrals over Lévy bases, they allow for non-Gaussianity. By looking at two fundamental subclasses, the spatio-temporal Ornstein-Uhlenbeck (STOU) and the mixed STOU process, we show how one can modify the spatio-temporal correlation of the general ambit field to model different phenomena. On the application side, we use a spatio-temporal random field to model malaria seasonality. The connection between the Matérn covariance and a linear fractional stochastic partial differential equation speeds up inferences and prediction.

Keywords: Spatio-temporal Statistics, random fields, statistical modelling.

AMS subject classification: 62M30, 62M40, 62M86, 62F10, 62F12, 62F15.

1 Introduction

Data are being collected at higher frequencies and spatial resolutions. To address their complexity, researchers have been developing spatio-temporal models and methodologies. These use the characteristics of each location-time observation such as the amount of rainfall as well as the correlation between the observations [4, 6]. The latter arises by virtue of close proximity in space-time and can represent factors which we do not have data for; thus mitigating the problem with omitted variables in complex phenomena.
The most popular ingredient of these models is the spatial Gaussian Matérn random field (RF).

*Corresponding author: michele.nguyen@bdi.ox.ac.uk

Definition 1 (Matérn spatial covariance)**.**
A Gaussian Matérn RF $\{Y(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^d}$ has a covariance of the form:

$$\mathrm{Cov}(Y(\mathbf{x}), Y(\mathbf{x}+\mathbf{h})) = \frac{2^{1-\nu}\phi^2}{(4\pi)^{\frac{d}{2}}\Gamma(\nu+\frac{d}{2})\kappa^{2\nu}}(\kappa||\mathbf{h}||)^{\nu}K_{\nu}(\kappa||\mathbf{h}||),$$

where $d \in \mathbb{N}$, $\mathbf{h} \in \mathbb{R}^d$ and K_ν denotes the modified Bessel function of second kind and order $\nu > 0$. The parameters $\phi > 0$, ν and $\kappa > 0$ can be interpreted as variance, smoothness and decay parameters respectively.

$Y(\mathbf{x})$ has been shown to be a solution of a linear fractional stochastic partial differential equation (SPDE) [5]:

$$(\kappa^2 - \triangle)^{\frac{\alpha}{2}} Y(\mathbf{x}) = \phi W(\mathbf{x}),$$

where $\alpha = \nu + d/2$, $\triangle = \sum_{i=1}^{d} \delta^2/\delta x_i^2$ is the Laplace operator and $W(\mathbf{x})$ is Gaussian white noise. Through the Green's function of the differential operator, $Y(\mathbf{x})$ can expressed as a stochastic integral:

$$Y(\mathbf{x}) = \int_{\mathbb{R}^d} k(\mathbf{x}, \boldsymbol{\xi}) W(\mathrm{d}\boldsymbol{\xi}),$$

$$\text{where } k(\mathbf{x}, \boldsymbol{\xi}) = \frac{2^{1-\frac{\alpha-d}{2}}\phi}{(4\pi)^{\frac{d}{2}}\Gamma(\frac{\alpha}{2})\kappa^{\alpha-d}}(\kappa||\mathbf{x}-\boldsymbol{\xi}||)^{\frac{\alpha-d}{2}} K_{\frac{\alpha-d}{2}}(\kappa||\mathbf{x}-\boldsymbol{\xi}||).$$

Recently, there has been renewed interest in viewing SPDE solutions as RFs and investigating their probabilistic as well as statistical properties. Of special mention is the work on ambit fields, a family of non-Gaussian spatio-temporal RFs. These were introduced for turbulence modelling and have stochastic integrals as their core components [2]. For further details on the connection between ambit fields and SPDEs, interested readers are referred to [1].

While research on spatio-temporal ambit fields is on the rise, most work is still concentrated on the purely temporal setting. In the second section, we hope to motivate further study on spatial and spatio-temporal ambit fields by introducing two fundamental subclasses: the spatio-temporal Ornstein-Uhlenbeck process (STOU) and the mixed spatio-temporal Ornstein-Uhlenbeck (MSTOU) process [10, 11]. By focusing on shape of the integration set and the Lévy basis, we show that these ambit fields are able to model clusters in space-time as well as bridge between short-range and long-range dependence.

In addition to paving the way to more interesting spatio-temporal models, the link between Gaussian Matérn RFs and SPDEs has been used to make inference and prediction with such models computationally feasible [8]. We illustrate this in the third section by fitting a log-linear spatio-temporal regression model to the monthly proportions of malaria cases in Madagascar. Malaria is a disease caused by the Plasmodium parasite and remains a major cause of child mortality in sub-Saharan Africa [13]. Understanding location-specific seasonal characteristics is useful for maximising the impact of interventions.

2 Spatio-temporal Ornstein-Uhlenbeck processes

Lévy-driven spatio-temporal Ornstein-Uhlenbeck (STOU) processes form one of the first few subclasses of ambit fields introduced in [3]. The classic, non-mixed process has temporal exponential correlations which are characteristic of OU processes. More generally, the construction can be seen as an extension of a stationary OU process in time where the Lévy basis allows for non-Gaussianity and the ambit set aids the modelling in space-time.

Definition 2 (MSTOU and STOU processes)**.**
Let a random field $Y = \{Y_t(\mathbf{x})\}_{x \in \mathbb{R}^d, t \in \mathbb{R}}$ be defined as follows:

$$Y_t(\mathbf{x}) = \int_0^\infty \int_{A_t(\mathbf{x})} \exp(-\lambda(t-s)) L(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}s, \mathrm{d}\lambda). \tag{1}$$

where L is a Lévy basis over the product space of space-time and the parameter space of λ. If λ is associated with a probability density other than the Dirac delta measure, we call Y a *MSTOU process*. Otherwise, Y is a *STOU process*.
The ambit set, $A_t(\mathbf{x}) \subset \mathbb{R}^d \times \mathbb{R}$, can be interpreted as a causality cone in physics and is restricted to be translation invariant as well as non-anticipative. We also require that $A_s(\mathbf{x}) \subset A_t(\mathbf{x}), \forall s < t$.

The shape of $A_t(\mathbf{x})$ determines the kind of non-separable spatio-temporal covariance we can model. Although STOU processes have exponential temporal correlation and hence short-range (SR) dependence in time, MSTOU process can exhibit long-range (LR) temporal dependence for various choices of the probability density of λ. We illustrate this through the following example:

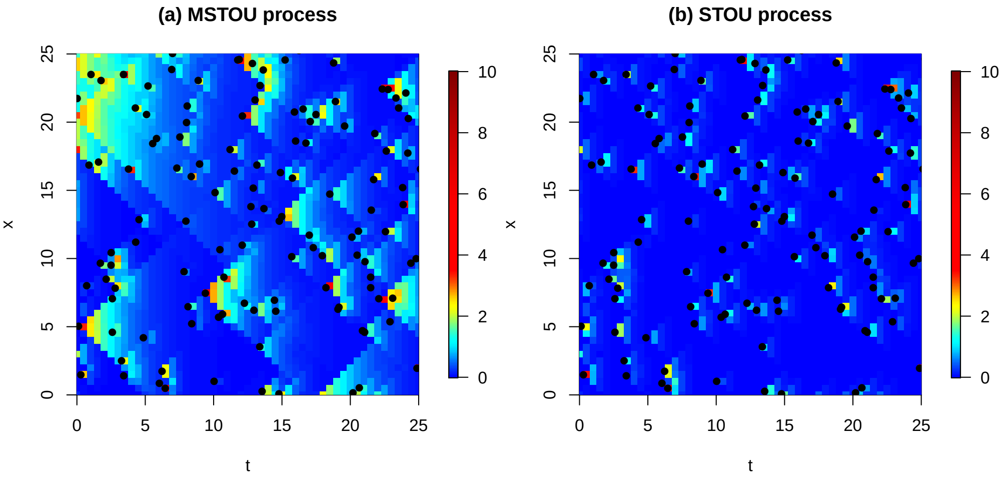
Example 1. Let E be an arbitrary bounded subset of space-time, then L is a spatio-temporal compound Poisson Lévy basis if:

$$L(E) = \sum_{k=-\infty}^{\infty} J_k \mathbf{1}_{\{(\Gamma_k, \lambda_k) \in E\}}.$$

Here, the jump sizes $J_k \overset{i.i.d.}{\sim} \text{Gamma}(\alpha_Z, \beta_Z)$ for $k \in \mathbb{N}$, $\left\{\Gamma_k = \left(\Gamma_k^{(1)}, \Gamma_k^{(2)}\right)\right\}_{k \in \mathbb{N}}$ denote the spatio-temporal jump locations of a Poisson process $N = (N_t(\mathbf{x}))_{(\mathbf{x},t) \in \mathbb{R}^d \times \mathbb{R}}$ with intensity μ, and $\{\lambda_k\}_{k \in \mathbb{N}}$ is an independent, identically distributed (i.i.d.) sequence of decay rates with probability density function f. The three components, $\{J_k\}_{k \in \mathbb{N}}$, N and $\{\lambda_k\}_{k \in \mathbb{N}}$ are independent of each other.
Suppose that $d = 1$ so that $x \in \mathbb{R}$ and $A_t(x) = \{(\xi, s) : s \leq t, |x - \xi| \leq c|t - s|\}$ for some $c > 0$. If $f(\lambda)$ is a Gamma(α, β) density where $\alpha > d + 1 = 2$ and $\beta > 0$, the spatio-temporal covariance $\text{Cov}(Y_t(x), Y_{t+d_t}(x + d_x))$ is equal to:

$$\begin{aligned} &\text{Var}(L') \int_0^\infty \int_{A_t(x) \cap A_{t+d_t}(x+d_x)} \exp(-2\lambda(t-s) - \lambda d_t) \mathrm{d}\xi \mathrm{d}s f(\lambda) \mathrm{d}\lambda, \\ &= \frac{c\beta^\alpha \,\text{Var}(L')}{2(\beta + \max(|d_t|, |d_x|/c))^{\alpha-2}(\alpha-2)(\alpha-1)}. \end{aligned}$$

Figure 1: Heat plots of: (a) the MSTOU process and (b) the corresponding STOU process with rate parameter $\int_0^\infty \lambda f(\lambda)\mathrm{d}\lambda = \alpha/\beta = 3$. The black dots denote the positions of the jumps in space-time.


(a) MSTOU process

(b) STOU process

This has the desirable property known as 'non-separability' because the temporal and spatial distances interact. In general, since the covariance depends on the ambit sets' intersections, their shapes determine the functional form.

To see that the MSTOU process can model both LR and SR temporal dependence, we integrate the temporal covariance over $\tau = d_t$:

$$\int_0^\infty \mathrm{Cov}(Y_t(x), Y_{t+\tau}(x))\mathrm{d}\tau = \frac{c\beta^3\,\mathrm{Var}(L')}{2(\alpha-2)(\alpha-1)(\alpha-3)},$$

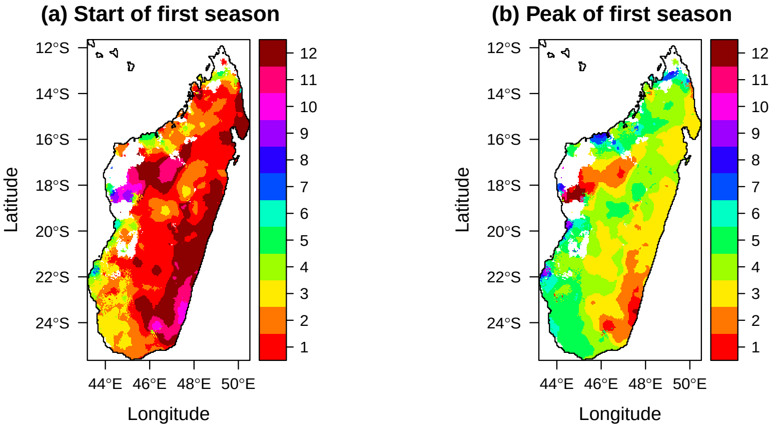
for $\alpha > 3$. So, the process has temporal LR dependence for $2 < \alpha \leq 3$ when the integral diverges but SR dependence otherwise.

In Figure 1, we show heat plots of realisations from a MSTOU and STOU process with the same ambit set and underlying Poisson process. Just as how a temporal OU process is used to model financial volatility clusters in time, the MSTOU and STOU processes can be used to model clusters in space-time. While the field values decay at the same rate for the STOU process, the values decay at varying rates for each jump in the MSTOU process. Its lower rate parameters lead to larger clusters which is consistent with its long memory.

3 Modelling malaria seasonality

While the Gaussian Matérn RF and its stochastic integral representation can be extended to create more interesting spatio-temporal models like ambit fields, the added complexity in terms of correlation structures and non-Gaussianity make statistical inference more tedious. Currently, only moment-based procedures have been

Figure 2: (a) Estimated start and (b) peak months of the first malaria transmission season in Madagascar. These are preliminary results based on median proportion estimates and the white regions on the island denote the areas where the entropy is not well-defined.



implemented to estimate the parameters of STOU and MSTOU processes [10, 11]. On the other hand, under the assumption of Gaussianity, the link between Gaussian Markov RFs and their approximate SPDE solutions has been recently used to alleviate a longstanding problem faced when conducting likelihood-based inference for spatial models: computing the inverse and log-determinants of a large covariance matrix. The method, Integrated Nested Laplace Approximation (INLA), which is set up in a Bayesian framework, has made convenient through the R package, R-INLA [4].

With the newfound computational capability, spatio-temporal modelling can be conducted for larger data sets and wider application scenarios. We illustrate this by modelling malaria seasonality in Madagascar. In particular, we apply the following model to monthly case records from 2669 health facilities between 2013 and 2016 [9]:

$$\log(p_{i,j}) = X_{ij}^T \boldsymbol{\beta} + \phi_{ij} + \epsilon_{ij}. \tag{2}$$

Here, $p_{i,j}$ represents the fraction of cases in month i at location j, X_{ij} is a m-dimensional covariate vector including an intercept, $\boldsymbol{\beta} \in \mathbb{R}^m$ is the corresponding parameter vector and $\epsilon \sim N(0, \sigma_e^2)$ denotes i.i.d. noise. The spatio-temporal Gaussian field ϕ is constructed such that:

$$\phi_{i,j} = \begin{cases} \xi_{1j} \text{ for } i = 1, \\ a\phi_{i-1,j} + \xi_{i,j} \text{ for } i = 2, \ldots, 12, \end{cases} \tag{3}$$

$|a| < 1$ and $\xi_{i,j}$ correspond to zero-mean Gaussian innovations which are temporally independent but spatially coloured with a Matérn covariance.

With INLA, backwards regression to select covariates such as lagged monthly temperature and rainfall is computationally feasible. Figure 2 shows the start and peak of the first transmission season as identified via least-squares fits of the median estimated monthly case proportions to rescaled von Mises densities. The latter is used to ensure continuity of the seasonal pattern between December and January and is also convenient for identifying seasonal peaks since these correspond to the mean parameters. The estimated peaks in Figure 2(b) tie in with current literature which state that most areas experience the peak around April while the eastern coast experiences it earlier around February [7].

By defining the transmission season to be the period when the monthly proportions exceed 1/12, we can also identify its start, end and length in a consistent manner. This allows us to make fair comparisons across the study region. In practice, the estimated start months of the transmission season, as shown in Figure 2(a), are also useful for the planning of indoor residual spraying campaigns since these typically need to be completed by then.

4 Conclusion and further work

In this paper, we used the examples of ambit fields and malaria seasonality to highlight advances in spatio-temporal modelling on both the theory and the practical fronts. The link between random fields and stochastic partial differential equations have proved useful in both cases.

Although ambit fields have been successfully applied to turbulence and finance modelling, there is need for continued research on its properties and spatio-temporal applications. In the context of likelihood-based inference, there has been recent progress for other, simpler non-Gaussian spatial models. While the associated R package, LANG, is still under development, the Monte Carlo expectation-maximization algorithm for non-Gaussian spatial Matérn fields presents an exciting way forward [5, 12].

Similarly, since the seasonality results in Section 3 are preliminary, additional analysis is being done with the realisations of the fitted model so as to obtain uncertainty measures. As researchers develop methodologies to study the increasing wealth of spatiotemporal data, it is hoped that we will vastly improve our understanding of the world we live in.

Acknowledgements: My gratitude goes to Imperial College for my PhD scholarship and the Bill & Melinda Gates Foundation for supporting my current research. I would also like to thank Almut Veraart for the fruitful PhD discussions as well as Daniel Weiss, Rosalind Howes and the rest of the Malaria Atlas Project team for their feedback on the seasonality work.

Bibliography

[1] O. E. Barndorff-Nielsen, F. E. Benth, and A. E. D. Veraart. Ambit processes and stochastic partial differential equations. In *Advanced mathematical methods for finance*, pages 35–74. Springer, 2011.

[2] O. E. Barndorff-Nielsen, F. E. Benth, and A. E. D. Veraart. *Ambit Stochastics.* Springer Nature, Switzerland, 2018.

[3] O. E. Barndorff-Nielsen and J. Schmiegel. Lévy-based tempo-spatial modelling: with applications to turbulence. *Uspekhi Mat. Nauk*, 59(1):63–90, 2004.

[4] M. Blangiardo and M. Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA.* John Wiley & Sons, New York, 2015.

[5] D. Bolin. Spatial Matérn Fields Driven by Non-Gaussian Noise. *Scandinavian Journal of Statistics*, 41(3):557–579, 2014.

[6] N. Cressie and C. K. Wikle. *Statistics for spatio-temporal data.* John Wiley & Sons, Hoboken, 2011.

[7] R. E. Howes, S. A. Mioramalala, B. Ramiranirina, T. Franchard, A. J. Rakotorahalahy, D. Bisanzio, P. W. Gething, P. A. Zimmerman, and A. Ratsimbasoa. Contemporary epidemiological overview of malaria in Madagascar: operational utility of reported routine case data for malaria control planning. *Malaria journal*, 15(1):502, 2016.

[8] F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

[9] M. Nguyen, R. E. Howes, T. C. D. Lucas, K. E. Battle, E. Cameron, H. S. Gibson, J. Rozier, S. Keddie, E. Collins, R. Arambepola, S. Y. Kang, C. Hendriks, M. A. Nambinisoa, P. W. Gething, and D. J. Weiss. Mapping malaria seasonality: a case study from Madagascar. Submitted to journal. Preprint available at arXiv:1901.10782, 2019

[10] M. Nguyen and A. E. D. Veraart. Spatio-temporal Ornstein-Uhlenbeck processes: theory, simulation and statistical inference. *Scandinavian Journal of Statistics*, 44(1):46–80, 2017.

[11] M. Nguyen and A. E. D. Veraart. Bridging between short-range and long-range dependence with mixed spatio-temporal Ornstein-Uhlenbeck processes. *Stochastics*, 90(7):1023–1052, 2018.

[12] J. Wallin and D. Bolin. Geostatistical modelling using non-Gaussian Matérn fields. *Scandinavian Journal of Statistics*, 42(3):872–890, 2015.

[13] World Health Organization. *World Malaria Report 2018.* Geneva, 2018.

Pseudo-observations and a variance inequality

Morten Overgaard[1*]

[1] *Department of Public Health, Aarhus University*

Abstract: A method based on jack-knife pseudo-observations has been used for regression analysis when outcomes may be missing. This pseudo-observation method has been shown to produce a variance of the regression parameter estimate that is not consistently estimated by standard variance estimators. The asymptotic bias of the standard variance estimators has been shown to be upwards in some specific cases. In this paper, the upwards bias is established in a somewhat broader generality.

Keywords: Regression analysis, survival analysis, pseudo-observations

1 Introduction

The pseudo-observation method is a regression method that can be used when some outcome values are missing. It was suggested by [2] and it has mainly been studied in a survival analysis context where outcomes such as survival past some time point or death of a certain cause before a certain time may be missing due to right censoring. The method is based on substituting the outcome variable for a variable of pseudo-observations, which will be the jack-knife pseudo-values, $\hat{\theta}_{n,i} = n\hat{\theta}_n - (n-1)\hat{\theta}_n^{(i)}$ of an estimator of the expectation of the outcome V, $\theta = \mathrm{E}(V)$. Here, $\hat{\theta}_n$ is the overall estimate of θ based on a sample of size n and $\hat{\theta}_n^{(i)}$ is the estimate obtained by the same estimator on the same sample with the ith observation left out. Suppose the regression model tells us that a regression parameter $\beta_0 \in \mathbb{R}^q$ exists such that $\mathrm{E}(V \mid Z) = \mu(\beta_0; Z)$ for some given function μ and covariates Z. The pseudo-observation method now involves solving an estimating equation of the type

$$\frac{1}{n}\sum_{i=1}^{n} A(\beta; Z_i)(\hat{\theta}_{n,i} - \mu(\beta; Z_i)) = 0 \tag{1}$$

to obtain a regression parameter estimate $\hat{\beta}_n$ where A is a given q-dimensional vector function. The form of (1) covers many typical estimating equations except for the

[*] Corresponding author: moov@ph.au.dk

fact that outcomes V_i have been substituted by pseudo-observations $\hat{\theta}_{n,i}$. Originally, [2] suggested basing variance estimation on the Huber–White type estimator

$$\hat{\mathrm{Var}}(\hat{\beta}_n)_{n,\mathrm{HW}} = \frac{1}{n}\hat{M}_n^{-1}\hat{\Sigma}_{n,\mathrm{HW}}(\hat{M}_n^{\mathsf{T}})^{-1}, \tag{2}$$

where $\hat{M}_n = n^{-1}\sum_{i=1}^n A(\hat{\beta}_n; Z_i)\partial/\partial\beta\mu(\hat{\beta}_n; Z_i)$ and

$$\hat{\Sigma}_{n,\mathrm{HW}} = \frac{1}{n}\sum_{i=1}^n A(\hat{\beta}_n; Z_i)A(\hat{\beta}_n; Z_i)^{\mathsf{T}}(\hat{\theta}_{n,i} - \mu(\hat{\beta}_n; Z_i))^2. \tag{3}$$

Building on results of [3], the results of [4] indicate that this approach to variance estimation is generally inappropriate. In that paper, the estimate $\hat{\theta}_n$ is seen as the product of applying a functional to the empirical distribution, $\hat{\theta}_n = \phi(F_n)$ for some functional ϕ, where F_n is, essentially, the empirical distribution of the n observations $X_1, \dots, X_n$ that the estimator is based on. The first and second order influence functions associated with the functional ϕ as an estimator are denoted $\dot{\phi}$ and $\ddot{\phi}$. The result by [4], based on various assumptions, states that the asymptotic variance of $n^{1/2}(\hat{\beta}_n - \beta_0)$ will take the form $M^{-1}\Sigma(M^{\mathsf{T}})^{-1}$. Here, $M = \mathrm{E}(A(\beta_0; Z)\partial/\partial\beta\mu(\beta_0; Z))$ is consistently estimated by $\hat{M}_n$ from before under the implied assumptions. The inner part is of the form $\Sigma = \mathrm{Var}(h_0(X, Z) + h_1(X))$ where

$$h_0(x, z) = A(\beta_0; z)(\theta + \dot{\phi}(x) - \mu(\beta_0; z)) \tag{4}$$

and

$$h_1(x) = \mathrm{E}(A(\beta_0; Z)\ddot{\phi}(X, x)). \tag{5}$$

Because the pseudo-observation $\hat{\theta}_{n,i}$ approximates $\theta + \dot{\phi}(X_i)$ in this setting, it can be seen that $\hat{\Sigma}_{n,\mathrm{HW}}$ estimates $\mathrm{Var}(h_0(X, Z))$ consistently but is not generally estimating Σ consistently.

In a survival and competing risks setting where $\hat{\theta}_n$ corresponds to the Kaplan–Meier or Aalen–Johansen estimators, the papers of [3] and [5] reveal that $\hat{\Sigma}_{n,\mathrm{HW}}$ and so $\hat{\mathrm{Var}}(\hat{\beta}_n)_{n,\mathrm{HW}}$ is upwards biased. In this paper, this result will be established in a more general setting by following a similar approach and proving similar results as in [3] and [5].

2 Main result

The setting that we will consider is a survival setting and is a special case of the setting of [6]. Specifically, we consider an underlying event time $T > 0$ and an underlying event type $D \in \{1, \dots, d\}$ and a censoring time $C > 0$. The exit time $\tilde{T} = T \wedge C$ and the exit type $\tilde{D} = D\mathbf{1}(T \leq C)$ are observed in addition to the covariates Z. We will be working under the completely independent censoring assumption $C \perp\!\!\!\perp (T, D, Z)$. Modelling n observations, $(X_1, Z_1), \dots, (X_n, Z_n)$ are independent replications of (X, Z) where $X = (\tilde{T}, \tilde{D})$. The censoring distribution is captured by the function G given by $G(s) = \mathrm{P}(C \geq s) = \prod_0^{s-}(1 - \Lambda(\mathrm{d}u))$ and this can be estimated by a Kaplan–Meier style estimate $\hat{G}_n(s) = \prod_0^{s-}(1 - \hat{\Lambda}_n(\mathrm{d}u))$ where $\hat{\Lambda}_n(s) = \int_0^s \hat{K}_n(u)^{-1}\hat{H}_{n,0}(\mathrm{d}u)$ where $\hat{K}_n(s)$ is the empirical version of $K(s) = \mathrm{P}(\tilde{T} > s) + \mathrm{P}(\tilde{T} = s, \tilde{D} = 0)$ and $\hat{H}_{n,0}(s)$ is the empirical version of $H_0(s) = \mathrm{P}(\tilde{T} \leq s, \tilde{D} = 0)$. It can be seen that $\hat{G}_n(s)$ is consistently estimating $G(s)$ under the independent censoring assumption.

We will consider a certain time point $t > 0$ with $\mathrm{P}(\tilde{T} > t) > 0$ of particular interest. The outcome V that we consider should be a real-valued random variable that is available at time t in the sense that it is a function of $(T \wedge t, D\mathbf{1}(T \leq t))$. This includes the examples of survival past t, $V = \mathbf{1}(T > t)$, death of a certain cause before time t, $V = \mathbf{1}(T \leq t, D = j)$, life time up to time t, $V = T \wedge t$, life time lost due to a certain cause up to time t, $V = (t - T \wedge t)\mathbf{1}(D = j)$, and more. Such outcomes are observed when $C \geq T \wedge t$. This means we can use the estimator

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{V\mathbf{1}(C \geq T \wedge t)}{\hat{G}_n(\tilde{T} \wedge t)} \tag{6}$$

to estimate $\theta = \mathrm{E}(V)$ consistently under the independent censoring assumption. Among other examples, this type of estimator covers the Kaplan–Meier estimator and the Aalen–Johansen estimator in a competing risks setting. Seeing this estimator as a functional ϕ, the influence function of this estimator is derived by [6] under certain assumptions and can be expressed as

$$\begin{aligned}\dot{\phi}(x) = {} & \frac{v(x)}{G(\tilde{t} \wedge t)} - \theta \\ & + \int\int_0^{t-} \frac{v(x^*)\mathbf{1}(\tilde{t}^* > u)}{G(\tilde{t}^* \wedge t)} \frac{1}{1 - \Delta\Lambda(u)} \frac{1}{K(u)} M_{x,0}(\mathrm{d}u)F(\mathrm{d}x^*)\end{aligned} \tag{7}$$

where $M_{x,0}(s) = N_{x,0}(s) - \int_0^s Y_x(u)\Lambda(\mathrm{d}u)$ with $N_{x,0}(s) = \mathbf{1}(\tilde{t} \leq s, \tilde{d} = 0)$ and $Y_x(s) = \mathbf{1}(\tilde{t} > s) + \mathbf{1}(\tilde{t} = s, \tilde{d} = 0)$ and F is the distribution of $X = (\tilde{T}, \tilde{D})$ with x and x^* denoting $(\tilde{t}, \tilde{d})$ and $(\tilde{t}^*, \tilde{d}^*)$ respectively. Splitting $K(u)$ into the product $S(u)G(u)$ where $S(u) = \mathrm{P}(T > u)$, a change in the order of integration reveals

$$\dot{\phi}(x) = \frac{v(x)}{G(\tilde{t} \wedge t)} - \theta + \int_0^{t-} \mathrm{E}(V \mid T > u)\frac{1}{G(u+)} M_{x,0}(\mathrm{d}u) \tag{8}$$

since $S(u)^{-1}\int v(x^*)\mathbf{1}(\tilde{t}^* > u)G(\tilde{t}^* \wedge t)^{-1}F(\mathrm{d}x^*) = S(u)^{-1}\,\mathrm{E}(V\mathbf{1}(C \geq T \wedge t)\mathbf{1}(\tilde{T} > u)G(\tilde{T} \wedge t)^{-1}) = \mathrm{E}(V \mid T > u)$ for $u < t$ and $G(u+) = (1 - \Delta\Lambda(u))G(u)$. We may then use the identity of $\mathbf{1}(C \geq s)G(s)^{-1} - 1 = -\int_0^{s-} G(u+)^{-1}M_C(\mathrm{d}u)$ where $M_C(s) = N_C(s) - \Lambda(C \wedge s)$ with $N_C(s) = \mathbf{1}(C \leq s)$ to obtain the expression

$$h_0(X, Z) = A(\beta_0; Z)\big(V - \mathrm{E}(V \mid Z) - \int_0^{t-} \frac{V - \mathrm{E}(V \mid T > u)}{G(u+)} M_{X,0}(\mathrm{d}u)\big). \tag{9}$$

An expression of the second order derivative of the functional ϕ is given in the supplement of [6]. Based on this, the expression

$$\mathrm{E}(\ddot{\phi}(X, x) \mid Z) = \int_0^{t-} \frac{\mathrm{E}(V \mid T > u, Z) - \mathrm{E}(V \mid T > u)}{G(u+)} \frac{S(u \mid Z)}{S(u)} M_{x,0}(\mathrm{d}u) \tag{10}$$

where $S(u \mid z) = \mathrm{P}(T > u \mid Z = z)$ can be obtained in a similar manner as laid out above. We let

$$W(u) = \mathrm{E}\left(A(\beta_0; Z)\big(\mathrm{E}(V \mid T > u, Z) - \mathrm{E}(V \mid T > u)\big)\frac{S(u \mid Z)}{S(u)}\right) \tag{11}$$

to obtain

$$h_1(x) = \int_0^{t-} W(u)\frac{1}{G(u+)}M_{x,0}(\mathrm{d}u). \tag{12}$$

The main result can be stated as follows.

Theorem 1. *We have*

$$\Sigma = \mathrm{Var}(h_0(X, Z)) - \mathrm{Var}(h_1(X)) \tag{13}$$

such that $\Sigma < \mathrm{Var}(h_0(X, Z))$ *unless* $\mathrm{Var}(h_1(X)) = 0$. *In addition,*

$$\mathrm{Var}(h_1(X)) = \int_0^{t-} W(u)W(u)^{\mathsf{T}}S(u)\frac{1}{G(u+)}\Lambda(\mathrm{d}u) \tag{14}$$

such that $\mathrm{Var}(h_1(X)) = 0$ *if and only if* $W(u) = 0$ *for* Λ*-almost all* $u \in (0, t)$.

Proof. This follows from martingale properties of M_C which hold even in the conditional distribution given (T, D, Z). In particular, $M_C^2 - [M_C]$ is a martingale where $[M_C]$ is the optional variation process of M_C which is given by $[M_C](s) = \int_0^s (1 - \Delta\Lambda(u))N_C(\mathrm{d}u) - \int_0^s \Delta\Lambda(u)M_C(\mathrm{d}u)$ and has expectation $\mathrm{E}([M_C](s) \mid T, D, Z) = \int_0^s G(u+)\Lambda(\mathrm{d}u)$. The process given by $M_{X,0}(s) = \int_0^s \mathbf{1}(T > u)M_C(\mathrm{d}u)$ is then similarly a martingale in the conditional distribution, as is also the case with the vector-valued processes given by $M_1(s) = \int_0^s W(u)G(u+)^{-1}M_{X,0}(\mathrm{d}u)$ and $M_2(s) = \int_0^s A(\beta_0; Z)(V - \mathrm{E}(V \,|\, T > u))G(u+)^{-1}M_{X,0}(\mathrm{d}u)$. Now, $M_1(s)M_1(s)^{\mathsf{T}} - [M_1](s)$ defines a martingale in the conditional distribution and in particular $M_1(t-)M_1(t-)^{\mathsf{T}}$ and $[M_1](t-)$ will have the same conditional expectation which will result in

$$\mathrm{E}(h_1(X)h_1(X)^{\mathsf{T}} \mid T, D, Z) = \int_0^{t-} W(u)W(u)^{\mathsf{T}}\frac{\mathbf{1}(T > u)}{G(u+)}\Lambda(\mathrm{d}u) \tag{15}$$

by properties of the optional variation process as found in Chapter 2 of [1]. Taking expectation reveals (14). We have $\mathrm{Cov}(h_0(X, Z), h_1(X)) = -\,\mathrm{E}(M_2(t-)M_1(t-)^{\mathsf{T}})$ since $\mathrm{E}(M_{X,0}(s) \mid T, D, Z) = 0$. By the same argument as above, we find

$$\begin{aligned}&\mathrm{E}(M_2(t-)M_1(t-)^{\mathsf{T}} \mid T, D, Z)\\ &\quad= \int_0^{t-} A(\beta_0; Z)(V - \mathrm{E}(V \mid T > u))W(u)^{\mathsf{T}}\frac{\mathbf{1}(T > u)}{G(u+)}\Lambda(\mathrm{d}u).\end{aligned} \tag{16}$$

Here, $\mathrm{E}((V - \mathrm{E}(V \mid T > u))\mathbf{1}(T > u) \mid Z) = (\mathrm{E}(V \mid T > u, Z) - \mathrm{E}(V \mid T > u))S(u \mid Z)$ such that $\mathrm{E}(M_2(t-)M_1(t-)^{\mathsf{T}}) = \int_0^{t-} W(u)W(u)^{\mathsf{T}}S(u)G(u+)^{-1}\Lambda(\mathrm{d}u)$ and this establishes $\mathrm{Cov}(h_0(X, Z), h_1(X)) = -\,\mathrm{Var}(h_1(X))$ and thus (13). □

3 Conclusion

The result presented here generalizes the expression given in [5] for the outcome $V = \mathbf{1}(T \leq t, D = 1)$ and pseudo-observations based on the Aalen–Johansen estimator. The definition of W is slightly different here in comparison to that paper.
It is of interest to find a consistent variance estimate. In [4], a suggestion on this matter can be found. This suggestion is based on a plug-in version of $\Sigma = \mathrm{Var}(h_0(X, Z) + h_1(X))$.
Looking at the expression of $\mathrm{Var}(h_1(X))$, it seems the bias will be limited unless there is a close connection between V and Z simultaneously with a large censoring hazard. For that reason, this bias may be of minor importance in many applications. Also, the consequence of using the standard Huber–White type variance estimate will be a conservative rather than an invalid analysis according to Theorem 1.

Acknowledgements: I thank Erik Thorlund Parner and Jan Pedersen for discussions on the subject and comments on an earlier draft of this paper. This research is supported by the Novo Nordisk Foundation, grant NNF17OC0028276.

Bibliography

[1] P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993.

[2] P. K. Andersen, J. P. Klein, and S. Rosthøj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

[3] M. Jacobsen and T. Martinussen. A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics*, 43(3):845–862, 2016.

[4] M. Overgaard, E. T. Parner, and J. Pedersen. Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *Ann. Statist.*, 45(5):1988–2015, 10 2017.

[5] M. Overgaard, E. T. Parner, and J. Pedersen. Estimating the variance in a pseudo-observation scheme with competing risks. *Scandinavian Journal of Statistics*, 45(4):923–940, 2018.

[6] M. Overgaard, E. T. Parner, and J. Pedersen. Pseudo-observations under covariate-dependent censoring. *Journal of Statistical Planning and Inference*, 202:112 – 122, 2019.

A Comparative Study of Change-Point Analysis Techniques for Outbreak Detection

Christina Parpoula[1*], **Emmanouil–Nektarios Kalligeris**[1], **Alex Karagrigoriou**,[1]

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

Abstract: This paper provides some guidelines for the implementation and effective use of the mechanism of change-point analysis for the detection of epidemics, and discusses some of the statistical issues involved in the evaluation and optimal selection among change-point analysis-based approaches for the very early and accurate outbreak detection. The empirical comparative study provides evidence that statistical methods based on change-point analysis have several appealing properties compared to the current practice for the detection of epidemics.

Keywords: change-point analysis, distribution-free, statistical process control, statistical modeling, outbreak detection

AMS subject classification: 62-07, 62-09, 62G05, 62G08, 62G10, 62P10

1 Introduction

In the modern world, the timely detection of epidemics has been recognized as an extremely important problem of biosurveillance (see [4] and references therein). The current approach to influenza-like illness (ILI rate) surveillance (implemented by the European Centre for Disease Prevention and Control-ECDC and Centers for Disease Control and Prevention-CDC) is based on Serfling's cyclic regression method [5] by which epidemics are detected and reported when morbidity/mortality exceeds the epidemic threshold. This typical approach suffers from several shortcomings, such as the need for non-epidemic data to model the baseline distribution, and the fact that observations are treated not only as independent but also as identically distributed. Although the second issue can be overcome by properly adjusted modeling of the time series data, the first issue is a fundamental obstacle toward the development of an automated surveillance system for influenza.

Towards this end, this paper aims at the implementation and evaluation of cutting-edge change-point analysis-based methods for detecting changes in location of univariate ILI rate data. The main tool in this methodology is detection of unusual

[*] Corresponding author: parpoula.ch@aegean.gr

trends, like the beginning of an unusual trend that marks a switch from a control state to an epidemic state. Therefore, a beginning of an epidemic trend is a change point whose timely detection will predict occurrence of a new epidemic. The rest of this paper is organized as follows. In Section 2, the statistical framework is introduced. In Section 3, an empirical comparative study is performed. Finally, in Section 4, some concluding remarks are made.

2 Phase I Distribution-Free Change-Point Analysis

Several approaches for detecting outbreaks of infectious diseases in the literature are directly inspired by, or related to, methods of Statistical Process Control (SPC). In an epidemiological surveillance problem, the underlying process distribution is not normal and usually unknown. Hence statistical properties of commonly used SPC charts could be highly affected. In this paper, we implement important aspects of univariate distribution-free Phase I change-point analysis and apply some of the recent developments in this area, in order to develop a novel SPC charting method that works best for monitoring and outbreak detection processes.

Let x_i represent the ith observation, $i = 1, \ldots, m$, collected from the distribution of a quality characteristic, either continuous or discrete, X. When the process is in-control (IC), these observations are assumed to be independent with an unknown but common cumulative distribution function (c.d.f.) $F_0(x)$, whereas the out-of-control (OC) state can be described by a multiple change-point model, that is $F_0(x)$ if $0 < i \leq \tau_1$, $F_1(x)$ if $\tau_1 < i \leq \tau_2$, $\ldots$, $F_k(x)$ if $\tau_k < i \leq m$, where $0 < \tau_1 < \tau_2 < \ldots < \tau_k < m$ denote k change points and $F_r(*)$, $r = 0, \ldots, k$, are unknown c.d.f. which, at one or several times, may shift in position. Note here that the shift times τ_i are also assumed to be unknown. This Phase I analysis procedure provides a statistical test for verifying the hypothesis system H_0: the process was IC ($k = 0$) vs. H_1: the process was OC ($k > 0$) and identifying the time of the changes when the hypothesis of an IC process is rejected. This hypothesis testing system (performed in Phase I) requires the specification of a nominal false alarm probability (FAP). Following the recursive segmentation and permutation (RS/P) approach of Capizzi and Masarotto in [1], choosing an acceptable FAP value, say α, we test the stability over time of the level parameter. The following steps need to be executed for level-changes detection, i.e., detection of single or multiple level shifts.

Let us consider the problem of testing the null hypothesis that the process was IC against the alternative hypothesis that the process mean experienced an unknown number of step shifts. In such a case, a set of test (control) statistics is needed for detecting $1, 2, \ldots, K$ step shifts with K denoting the maximum number of hypothetical change points. The mean values $\mu_0, \ldots, \mu_k$, and the change points are assumed to be unknown. Further, defining $\tau_0 = 0$ and $\tau_{k+1} = m$, it is also assumed that $\tau_r - \tau_{r-1} \geq l_{MIN}$, $r = 1, \ldots, k+1$, where l_{MIN} is a (user pre-specified) constant giving the minimum number of successive observations allowed between two change points. For a sequence of individual observations, the control statistic and the possible change points are computed using a simple forward recursive segmentation approach. The algorithm starts with $k = 0$ and then proceeds in K successive stages. At the beginning of stage k, the interval $[1, m]$ is partitioned into k subintervals, each having a length greater or equal to l_{MIN}. At stage k, one of these subintervals is split, adding a new potential change point. The new change point is selected maximizing

$$\sum_{i=1}^{k+1} (\hat{\tau}_i - \hat{\tau}_{i-1})(\bar{x}(\hat{\tau}_{i-1}, \hat{\tau}_i) - \bar{x}_{om})^2 \tag{1}$$

conditionally on the results of the previous stages. Here $\bar{x}_{om}$ represents the overall mean (om) of observations, $\bar{x}(\alpha, b) = \frac{1}{b-\alpha} \sum_{i=\alpha+1}^{b} x_i$, and $0 = \hat{\tau}_0 < \hat{\tau}_1 < \cdots < \hat{\tau}_k < \hat{\tau}_{k+1} = m$ are the boundaries of the new partition. The control statistic T_k, $k = 1, \ldots, K$ is equal to the attained maximum value of Eq. (1). Therefore, given a test statistic, its p-value can be calculated, as the proportion of permutations under which the statistic value exceeds or is equal to the statistic computed from the original sample of observations. Choosing an acceptable FAP, say α, then, for p-value$< \alpha$, the null hypothesis that the process was IC is rejected.

3 Comparative Study

This paper focuses on the study of weekly ILI rate data (provided from the Hellenic Centre for Disease Control and Prevention) for Greece, between September 29, 2014 (week40/2014) and October 2, 2016 (week39/2016), which were used for analysis purposes. Here, we perform the RS/P approach for both periods under study (1^{st} period: week40/2014-week39/2015, 2^{nd} period: week40/2015-week39/2016) executing $L = 100000$ permutations with $K = \max\left(3, \min\left(50, \left[\frac{m}{15}\right]\right)\right)$ and $l_{MIN} = 5$. Our procedure signals possible changes of the mean (p-value< 0.001 for a change in level). The extracted signaled start (sw) and end weeks (ew) of the epidemics were sw01-ew14/2015 and sw01-ew08/2016.

In our study, the ability of RS/P method to detect the true (and correct amount of) change-points is tested through benchmarking. Therefore, RS/P derived change-points are compared with those derived after executing **1. the standard CDC and ECDC flu detection algorithm (Serfling's model)** [5]

$$\textbf{M11:} \quad X(t) = \alpha_0 + \alpha_1 t + \gamma_1 \cos(\frac{2\pi t}{m}) + \delta_1 \sin(\frac{2\pi t}{m}) + \varepsilon(t), \tag{2}$$

where $X(t)$ are the observed time series values (weekly ILI rate), $\varepsilon(t)$ are centered zero-mean random variables with variance σ^2, m denotes the number of observations within one year, and model coefficients are estimated by least squares method, **2. an extended Serfling's model** presented by Parpoula et al. in [4]

$$\textbf{M23:} \quad X(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \gamma_1 \cos(\frac{2\pi t}{m}) + \delta_1 \sin(\frac{2\pi t}{m})$$

$$+\gamma_2\cos(\frac{4\pi t}{m})+\delta_2\sin(\frac{4\pi t}{m})+\gamma_3\cos(\frac{8\pi t}{m})+\delta_3\sin(\frac{8\pi t}{m})+\varepsilon(t), \qquad (3)$$

3. a mixed model with a linear trend, 12-month seasonal periodicity, Auto-Regressive Moving Average (ARMA) terms, that is ARMA(2,1), and the minimum temperature (mintemp) as a random meteorological covariate presented by Kalligeris et al. in [3]

$$\textbf{MXM11:}\quad X(t)=\alpha_0+\alpha_1 t+\gamma_1\cos(\frac{2\pi t}{m})+\delta_1\sin(\frac{2\pi t}{m})$$

$$+\phi_1 x_{t-1}+\phi_2 x_{t-2}+\epsilon_t+\lambda_1\epsilon_{t-1}+\omega_1\text{mintemp}, \qquad (4)$$

and **4. Segment Neighbourhood (SegNeigh) algorithm** which uses an optimization step that searches over all previous change-point locations and picks the one that gives the optimal segmentation up to time t, presented by Kalligeris et al. in [2].

As aforementioned, the current approach to influenza surveillance is based on Serfling's cyclic regression model (**M11**). Parpoula et al. in [4], developed extended Serfling-type periodic regressions models, and through an exhaustive search process (using ANOVA comparisons and AIC, BIC information criteria) the best fitting model **M23** was selected. The aforementioned procedure allowed Parpoula et al. in [4] to extract the signaled start and end weeks of the epidemics, i.e., sw01-ew13/2015, sw01-ew08/2016. It is worth to be noted that the signaled start and end weeks were found to be identical considering either Serfling's model (**M11**) or extended Serfling's model (**M23**). Then, the above results motivated Kalligeris et al. in [3] to incorporate ARMA terms and random meteorological covariates in the model structure, for identifying the epidemics (sw01-ew12/2015, sw05-ew08/2016). Further, Kalligeris et al. in [2] established that the change-point detection analysis (SegNeigh algorithm) in conjunction with periodic-type ARMA modeling with covariates is capable of modeling time series data with typical and non-typical parts and identifying the beginning and end of the extreme periods that occurred (sw01-ew12/2015, sw01-ew08/2016).

Therefore, we then examine the ability of the RS/P, **MXM11** and SegNeigh approaches to detect the true change-points compared to the standard and extended CDC and ECDC flu detection algorithm (models **M11** & **M23**). The diagnostic performance of a test to discriminate between two groups (here, epidemic from non-epidemic) is typically evaluated using Receiver Operating Characteristic (ROC) curve analysis, and its related statistics/metrics (Accuracy-ACC, Sensitivity-SENS, Specificity-SPEC, Area Under the ROC curve-AUC). Hence we estimated these metrics along with their 95% Confidence Interval (CI) (exact Clopper-Pearson CIs for ACC, SENS and SPEC, exact binomial CI for each derived AUC) for each method (as shown in Table 1). Table 1 indicates that RS/P and SegNeigh approaches (higher ACC, SENS and AUC values) outperform **MXM11**, and seem to detect successfully the true change-points compared to the standard approach to influenza surveillance.

Table 1: Metrics for RS/P, MXM11 and SegNeigh approaches

Metric	RS/P	MXM11	SegNeigh
ACC (95% CI)	99.05% (94.81% to 99.98%)	95.24% (89.24% to 98.44%)	99.05% (94.81% to 99.98%)
SENS (95% CI)	100,0% (83,89% to 100,0%)	76,19% (52,83% to 91,78%)	95,24% (76,18% to 99,88%)
SPEC (95% CI)	98,81% (93,54% to 99,97%)	100,0% (95,71% to 100,0%)	100,0% (95,71% to 100,0%)
AUC (95% CI)	0,988 (0,944 to 0,999)	0,881 (0,803 to 0,936)	0,976 (0,926 to 0,996)

4 Concluding Remarks

In this paper, we implemented and evaluated cutting-edge change-point analysis-based methods for detecting changes in location of univariate ILI rate data. The empirical comparative study provides evidence that statistical methods based on change-point analysis have several appealing properties compared to the current practice for the detection of epidemics. In particular, RS/P and SegNeigh approaches, both succeeded in early and accurate outbreak detection. Both RS/P and SegNeigh approaches are advantageous since they can be applied to historical data without the need for distinguishing between epidemic and non-epidemic periods in the data, and single or multiple mean shifts can be detected. Further, RS/P Phase I distribution-free change-point analysis method is able to guarantee a prescribed false alarm probability without any knowledge about the (in-control) underlying distribution, whereas SegNeigh algorithm in conjunction with periodic-type ARMA modeling with covariates is capable of modeling time series data with typical and non-typical parts.

Bibliography

[1] G. Capizzi and G. Masarotto. Phase I Distribution-Free Analysis of Univariate Data. *Journal of Quality Technology*, 45:273–284, 2013.

[2] E.-N. Kalligeris, A. Karagrigoriou and C. Parpoula. Periodic-Type Auto-Regressive Moving Average Modeling with Covariates for Time-Series Incidence Data via Changepoint Detection, *Statistical Methods in Medical Research*, 2019, DOI: 10.1177/0962280219871587

[3] E.-N. Kalligeris, A. Karagrigoriou and C. Parpoula. On Mixed PARMA Modeling of Epidemiological Time Series Data. *Communications in Statistics-Case Studies, Data Analysis and Applications*, 2019, DOI: 10.1080/23737484.2019.1644253.

[4] C. Parpoula, A. Karagrigoriou and A. Lambrou. *Epidemic Intelligence Statistical Modelling for Biosurveillance.* In J. Blömer et al. (Eds.): MACIS 2017, Lecture Notes in Computer Science (LNCS), 10693, pp. 349-363, Springer International Publishing AG, 2017.

[5] R. Serfling. Methods for Current Statistical Analysis of Excess Pneumonia-Influenza Deaths. *Public Health Reports*, 78:494–506, 1963.

Goodness-of-fit Tests for Logistic Distribution Based on Some Characterization

Ilia Ragozin[1,2*]

[1] *Department of Mathematics and Mechanics of Saint Petersburg State University, Universitetsky pr. 28, Stary Peterhof 198504, Russia*

[2] *National Research University - Higher School of Economics, Souza Pechatnikov, 16, St.Petersburg 190008, Russia*

Abstract: We build two location-free goodness-of-fit tests for the logistic distribution based on a recent characterization by Hu and Lin [2]. The test statistics are based on suitable functionals of U-empirical distribution functions. One of them has the integral structure, the second one is of Kolmogorov type. For every test statistic we describe the large deviation asymptotics under the null-hypothesis. Then we calculate the local Bahadur efficiency for certain alternatives. Conditions of local optimality in Bahadur sense are also studied.

Keywords: Bahadur efficiency, logistic distribution, large deviations, goodness-of-fit tests, U-statistics.

AMS subject classification: 62G10, 62G20, 62G30, 62E10.

1 Introduction

This paper is dedicated to certain statistical tests based on characterizations. The idea of construction of such tests goes back to Yu. V. Linnik's paper [6]. Here we consider goodness-of-fit criteria based on the characterization of the logistic distribution. This characterization belongs to Hu and Lin [2] and can be formulated as follows.

Theorem. *Let X and Y be independent identically distributed random variables (iid rv's) with a continuous distribution function (df) L and let E be a standard exponential rv independent of X and Y. Then X and $min(X, Y) + E$ are identically distributed iff L belongs to the logistic family of df's with arbitrary shift parameter having the density*

$$l(x+\theta) = \frac{e^{x+\theta}}{(1+e^{x+\theta})^2}, \, \theta \in \mathbb{R}.$$

[*] Corresponding author: Ragza@yandex.ru

Let $X_1, ..., X_n$ be iid rv's with continuous d.f. $G(x)$. Using this characterization, we can test a composite null hypothesis H_0: G is a logistic d.f. with the density $l(\cdot + \theta)$ against the alternative H_1 under which H_0 is wrong. The papers on goodness-of-fit problem for logistic family are rather sparse, we can mention only [7] and [13].
Let $F_n(t), t \in \mathbb{R}$ be the usual empirical df and build the U-statistical empirical df using the convolution with the exponential rv

$$U_n(t) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \left(1 - e^{(\min(X_i, X_j) - t)}\right) I\left\{\min(X_i, X_j) < t\right\}.$$

Thus, the tests for the null-hypothesis H_0 can be based on the following statistics:

$$LU_n = \int_{-\infty}^{\infty} \left(F_n(t) - U_n(t)\right) dF_n(t) \quad \text{and} \quad KU_n = \sup_t |F_n(t) - U_n(t)|.$$

One of the purposes of this paper is the asymptotic comparison of the sequences statistic which are not asymptotically normal. So we will compare the constructed statistics using the Bahadur efficiency concept, which is described in detail in [1] and [8]. The quality of test statistics is measured by the so-called *exact slope.*
Let formulate the Bahadur fundamental theorem. Assume that the distribution of the sequence of observations $\mathbb{P}_\theta$ is determined by the parameter $\theta \in \Theta$, where Θ is a parametric set, and we test the null hypothesis $H_0 : \theta \in \Theta_0 \subset \Theta$ against alternative hypothesis $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$.

Bahadur Theorem. *Suppose that the sequence of statistics T_n satisfies the following conditions:*

1. $T_n \xrightarrow{\mathbb{P}_\theta} b(\theta)$, *$\theta \in \Theta_1$, where $-\infty < b(\theta) < \infty$, and*
2. $\lim_{n\to\infty} n^{-1} \ln \mathbf{P}_\theta(T_n \ge z) = -k(z)$ *for each $\theta \in \Theta_0$ and any z from an open interval I, on which function k is continuous and $\{b(\theta), \theta \in \Theta_1\} \subset I$.*

Then for all $\theta \in \Theta_1$, the exact slope $c_T(\theta)$ exists and can be calculated as

$$c_T(\theta) = 2k(b(\theta)).$$

We denote the Kullback-Leibler "distance" [1] between the alternative and the null-hypothesis H_0 by $K(\theta)$. In our case H_0 is composite, hence for any alternative density $f(x, \theta)$ one has

$$K(\theta) = \inf_{v \in \mathbb{R}} \int_{-\infty}^{\infty} \ln \frac{f(x,\theta)}{l(x+v)} f(x,\theta) dx. \tag{1}$$

The exact slopes always satisfy the inequality $c_T(\theta) \le 2K(\theta)$, so the local Bahadur efficiency of the sequence of statistics T_n is defined as

$$eff_T = \lim_{\theta \to 0} \frac{c_T(\theta)}{2K(\theta)}. \tag{2}$$

Now we present the alternatives $f_i(x,\theta), x \in \mathbb{R}, i = 1, ..., 3$, which we consider in this paper:

- 1. Scale alternative:

$$f_1(x,\theta) = \frac{e^{\theta + xe^{\theta}}}{(1+e^{xe^{\theta}})^2}.$$

- 2. Hyperbolic cosine alternative:

$$f_2(x,\theta) = \frac{\Gamma(\theta+2)}{\Gamma^2(\frac{\theta}{2}+1)} \frac{e^{\left(x+\frac{\theta x}{2}\right)}}{(1+e^x)^{\theta+2}}$$

- 3. Sine-alternative from [5] with the density for small θ:

$$f_3(x,\theta) = l(x) - 2\pi\theta\cos(2\pi L(x))l(x).$$

We present the main parts of the Taylor series expansion of Kullback-Leibler information as $\theta \to 0$ for our alternatives in Table 1.

alternative	f_1	f_2	f_3
$K(\theta)$	$0.715 \cdot \theta^2$	$0.08877 \cdot \theta^2$	$9.8696 \cdot \theta^2$

Table 1: Kullback-Leibler information as $\theta \to 0$

2 The statistic LU_n

Consider the auxiliary function

$$g(x,y,z) = \left(1 - e^{(\min(x,y)-z)}\right) I\left\{\min(x,y) < z\right\},\ x,y,z \in \mathbb{R}.$$

The statistic LU_n is asymptotically equivalent to the U-statistic of degree 3 with the centered kernel:

$$\Phi(x,y,z) = \frac{1}{2} - \frac{1}{3}\left(g(x,y,z) + g(y,z,x) + g(x,z,y)\right). \tag{3}$$

The projection of this kernel $\Psi(t) = \mathbb{E}\left(\Phi(X,Y,Z)|Z=t\right)$ is

$$\Psi(t) = -\frac{2}{3}\left(Li_2(-e^t) + t\ln(e^t+1) - \frac{1}{2}\ln^2(e^t+1) + \frac{7e^t+1}{4(e^t+1)}\right),$$

where $Li_2(z) = -\int_0^z \frac{\ln(1-t)}{t}dt,\ z \in \mathbb{C}$.
The variance of this projection is

$$\Delta^2 = \mathbb{E}\Psi^2(X) = 0.00195,$$

therefore, the kernel $\Phi(X,Y,Z)$ is non-degenerate [4]. Using Hoeffding's Theorem [4], [3] we get the following result.

Theorem 1.
Under the hypothesis H_0, the statistic LU_n satisfies the following limit relation as $n\to\infty$:

$$\sqrt{n}LU_n \xrightarrow{d} \mathcal{N}\left(0, 9\Delta^2\right).$$

Also the kernel Φ is centered, non-degenerate and bounded, so we apply the results on large deviations of non-degenerate U–statistics from [9] and obtain the following theorem:

Theorem 2.

$$\lim_{n\to\infty} n^{-1}\ln\mathbb{P}(LU_n > t) = h(t),$$

where the function $h(t)$ is continuous for small t and $h(t)\sim -\frac{t^2}{18\Delta^2}$, $t\to 0$.
Using Bahadur Theorem and Theorem 2, see also [11], we can calculate the local Bahadur exact slope of statistics LU_n : where $f(x,\theta)$ is the alternative density.

$$c_{LU}(\theta) \sim \frac{\left(\int\limits_{-\infty}^{\infty}\Psi(x)f'_\theta(x,0)dx\right)^2\theta^2}{\Delta^2}, \quad \text{as} \quad \theta\to 0, \tag{4}$$

where $f(x,\theta)$ is the alternative density.
We present the exact slopes and the values of local Bahadur efficiency of the statistic LU_n in the table below.

Alternative	f_1	f_2	f_3
$c_{LU}(\theta)$	$1.1970\cdot\theta^2$	$0.1371\cdot\theta^2$	$14.978\cdot\theta^2$
eff_{LU}	0.837	0.77	0.759

Table 2: Local Bahadur efficiencies of the statistic LU_n

3 The statistic KU_n

In this section we consider the Kolmogorov type statistic KU_n. For a fixed $t\in\mathbb{R}$ the expression $F_n(t)-U_n(t)$ is the U-statistic with the following kernel depending on t:

$$\Phi_1(X,Y;t) = \left(1 - e^{\min(X,Y)-t}\right)I\{\min(X,Y)<t\} - \frac{1}{2}\left(I\{X<t\}+I\{Y<t\}\right).$$

The projection of the family of kernels $\Phi_1(X,Y;t)$ is equal to

$$\begin{aligned}\Psi_1(s,t) = \mathbb{E}\left(\Phi_1(X,Y;t)|Y=s\right) &= \frac{e^{\min(s,t)}(1+e^{-t})}{1+e^{\min(s,t)}} - e^{-t}\ln(e^{\min(s,t)}+1)\\ &+ \frac{I\{s<t\}(1-e^s(1+2e^{-t}))}{2(1+e^s)} - \frac{e^t}{2(1+e^t)}.\end{aligned} \tag{5}$$

Now we calculate the variance of this projection under H_0:

$$\Delta_1^2(t) := \mathbb{E}_X \Psi_1^2(X,t) = \frac{e^{3t} + 8e^{2t} + 8e^t - 4(e^t+1)(e^t+2)\ln(e^t+1)}{4e^{2t}(e^t+1)^2},$$

hence, the family of kernels $\Phi_1(X,Y;t)$ is non-degenerate [10] and besides

$$\Delta_1^2 = \sup_{t\in\mathbb{R}} \Delta_1^2(t) \approx 0.02322\ldots.$$

The limit distribution of the statistics KU_n is unknown. Using the methods of ([12]), one can show that the $U-$empirical process

$$\eta_n(t) = \sqrt{n}\left(F_n(t) - U_n(t)\right), \qquad t \in \mathbb{R};$$

weakly converges as $n \to \infty$ to a certain Gaussian process $\eta(t)$ with complicated covariance. Then the sequence of statistics $\sqrt{n}KU_n$ converges in distribution to $\sup_{t\in\mathbb{R}} |\eta(t)|$, but we are not able to find this distribution. Hence it is reasonable to define the critical values for KU_n by simulation.
The family of kernels $\Phi_1(X,Y;t)$ is centered and bounded in the sense described in [10]. Applying the large deviation theorem for the supremum of the family of non-degenerate U–statistics from [10], we get the following result.

Theorem 3. For $z > 0$

$$\lim_{n\to\infty} n^{-1} \ln \mathbb{P}\left\{KU_n > z\right\} = w(z) \sim -\frac{z^2}{8\Delta_1^2},$$

where the function $w(z)$ is continuous for sufficiently small $z > 0$.
Using Theorem 3, we can obtain the following expression for the exact local slope $c_{KU}(\theta)$:

$$c_{KU}(\theta) = \frac{\sup\limits_{t\in\mathbb{R}} \left(\int\limits_{-\infty}^{\infty} \Psi_1(x;t) f'_\theta(x,0)dx \right)^2}{\Delta_1^2} \cdot \theta^2. \tag{6}$$

We calculate the exact slopes of statistic KU and the values of local Bahadur's efficiency and collect them in the table below .

alternative	f_1	f_2	f_3
$c_{KU}(\theta)$	$0.5033 \cdot \theta^2$	$0.0509 \cdot \theta^2$	$15.801 \cdot \theta^2$
eff_{KU}	0.352	0.287	0.800

Table 3: local Bahadur efficiencies of the statistic KU_n

Acknowledgements: I would like to express my gratitude to my scientific advisor Prof. Ya. Yu. Nikitin for interesting problem statement and very useful advice. I acknowledge the support of the grant SPbGU-DFG 6.65.37.2017.

Bibliography

[1] R. R. Bahadur. *Some limit theorems in statistics.* SIAM, Philadelphia, 1971.

[2] C. -Y. Hu and G. D. Lin. Characterizations of the logistic and related distributions. *Journ. of Mathem. Anal. and Appl.*, 463(1):79–92, 2018.

[3] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19:293–325, 1948.

[4] V. S. Korolyuk and Y. V. Borovskich. *Theory of U-statistics.* Springer Science & Business Media, 2013.

[5] C. Ley and D. Paindaveine. Le Cam optimal tests for symmetry against Ferreira and Steel's general skewed distributions. *J. Nonparam. Stat.*, 21(8):943–967, 2009.

[6] Yu. V. Linnik. Linear forms and statistical criteria I, II. *Ukrain. Mathem. J.*, 5:207–243; 247–290, 1953 (in Russian). *Engl. transl. in Selected Transl. in Mathem. Stat. and Probab.*, 3:1–90, Amer. Math. Soc., Providence, RI, 1963.

[7] B. Milošević and M. Obradović. Two-dimensional Kolmogorov-type goodness-of-fit tests based on characterizations and their asymptotic efficiencies. *J. Nonparam. Stat.*, 28(2):413–427, 2016.

[8] Ya. Yu. Nikitin. *Asymptotic Efficiency of Nonparametric Tests.* Cambridge University Press, NY, 1995.

[9] Ya. Yu. Nikitin and E. V. Ponikarov. Rough large deviation asymptotics of Chernoff type for von Mises functionals and U-statistics. *Proc. of St.Petersburg Math. Soc.*, 7:124–167, 1999. *Engl. transl. in AMS Transl.*, ser.2, 203:107–146, 2001.

[10] Ya. Yu. Nikitin. Large deviations of U-empirical Kolmogorov-Smirnov tests and their efficiency. J. Nonparam. Stat. 22:649–668, 2010.

[11] Ya. Yu. Nikitin and I Peaucelle. Efficiency and local optimality of distribution-free tests based on U- and V- statistics. *Metron*, LXII:185–200, 2004.

[12] B. W. Silverman. Convergence of a class of empirical distribution functions of dependent random variables. *Ann. Probab.*, 11:745–751, 1983.

[13] M. A. Stephens. Tests of fit for the logistic distribution based on the empirical distribution function. *Biometrika*, 66(3):591–595, 1979.

nn

Overview of the limit properties of the MAP partitions in the Dirichlet Process Normal-Normal Mixture Model

Łukasz Rajkowski[1*]

[1]*Faculty of Mathematics, Informatics and Mechanics, University of Warsaw*

Abstract: The increasing popularity of the Dirichlet Process Mixture Models (DPMMs) and other nonparametric Bayesian mixture models can be explained by the fact that they do not require a specification of the number of clusters in advance. This short paper presents an overview of the large sample results concerning the maximum a posteriori (the MAP) partition in the DPMM where the cluster means have Gaussian distribution and, for each cluster, the observations within the cluster have Gaussian distribution with a priori specified covariance matrix. The detailed description of these findings is contained in [2].

Keywords: Dirichlet Process Mixture Models, Chinese Restaurant Process
AMS subject classification: 62F15

1 The Model

The model we discuss here is based on *the Dirichlet Process*, which is a probability distribution on the space of probability measures. It allows to construct a model in which the distribution of observations is a random mixture from a parametrized family of distributions; this model is called Dirichlet Process Mixture Model (DPMM). For clustering purposes it is more convenient to formulate this model starting from a probability distribution on possible partitions of the data. This leads to *the Chinese Restaurant Process*. Before formulating the definition, it is good to mention the following, culinary metaphor: imagine that the observations' indices, say $1, 2, \dots, n$, are the customers that enter a restaurant. Customer 1 chooses any table he/she wants, but every customer that follows joins a table which is already occupied with a probability proportional to the number of customers sitting there and he chooses an empty table with probability proportional to some predefined parameter α. In this way the probability that 7 customers are partitioned as $\{1, 3\}, \{2, 4, 6, 7\}, \{5\}$ is given by

$$\frac{\alpha}{\alpha} \cdot \frac{\alpha}{1+\alpha} \cdot \frac{1}{2+\alpha} \cdot \frac{1}{3+\alpha} \cdot \frac{\alpha}{4+\alpha} \cdot \frac{2}{5+\alpha} \cdot \frac{3}{6+\alpha} = \frac{\alpha^3 \cdot 1! \cdot 3! \cdot 0!}{\alpha(\alpha+1)\dots(\alpha+6)}. \tag{1}$$

[*] Corresponding author: l.rajkowski@mimuw.edu.pl

Definition 1. *The Chinese Restaurant Process* with parameter α is a Markov chain of random partitions $(\mathcal{J}_n)_{n\in\mathbb{N}}$, where $\mathcal{J}_n$ is a partition of $[n] = \{1, 2, \ldots, n\}$, that satisfies

$$\mathcal{J}_{n+1} \mid \mathcal{J}_n = \{J_1, \ldots, J_k\} = \begin{cases} \{J_1, \ldots, J_i \cup \{n+1\}, \ldots, J_k\} & \text{with prob. } \frac{|J_i|}{n+\alpha} \\ \{J_1, \ldots, J_k, \{n+1\}\} & \text{with prob. } \frac{\alpha}{n+\alpha} \end{cases} . \tag{2}$$

We write $\mathcal{J}_n \sim \mathrm{CRP}(\alpha)_n$. Clearly, the probability of a given partition $\mathcal{J}$ of $[n]$ is given by $\frac{\alpha^{|\mathcal{J}|}}{(\alpha)^{\uparrow|\mathcal{J}|}} \prod_{J\in\mathcal{J}}(|J|-1)!$, where $(\alpha)^{\uparrow k} = \alpha(\alpha+1)\ldots(\alpha+k-1)$.

Suppose we want to model clustered multivariate data, which within every cluster have the Gaussian distribution with known covariance matrix Σ. In this case we may sample the partition of the data from the Chinese Restaurant Process and then independently for every cluster sample the cluster mean from some predefined Gaussian distribution and sample the observations belonging to that cluster from the Gaussian distribution with cluster's mean and covariance matrix Σ. This model is formally stated as

$$\begin{aligned} \mathcal{J} &\sim \mathrm{CRP}(\alpha)_n \\ \boldsymbol{\theta} = (\theta_J)_{J\in\mathcal{J}} \mid \mathcal{J} &\overset{\text{iid}}{\sim} \mathcal{N}(\mu, \mathrm{T}) \\ \mathbf{x}_J = (x_j)_{j\in J} \mid \mathcal{J}, \boldsymbol{\theta} &\overset{\text{iid}}{\sim} \mathcal{N}(\theta, \Sigma) \quad \text{for } J \in \mathcal{J}. \end{aligned} \tag{3}$$

where $\mu \in \mathbb{R}^d$ and $\mathrm{T}, \Sigma \in \mathbb{R}^{d,d}$ are the hyperparameters of the model. Formulation (3) is equivalent to the Dirichlet Process Mixture Model with normal distribution as the base and component measures.

Naturally, the goal is not to simulate data, but to make an inference about the clustering structure of given data. To this end, we apply the Bayesian approach: upon receiving the observations we may compute, using Bayes formula, the conditional probability (*the posterior*) on the space of the partitions of observations. It is impossible to compute this value *exactly* since the norming constant is intractable, but there are MCMC methods of sampling from the posterior. In the article we investigate the properties of *the MAP estimator*, i.e. the partition that maximises the posterior probability. Note that, by Bayes formula, the posterior is proportional to the joint probability on partitions and observations, which is easy to compute and in turn is proportional to

$$C^{|\mathcal{J}|} \prod_{J\in\mathcal{J}} \frac{|J|!}{|J|^{(d+2)/2} \det R_{|J|}} \cdot \exp\left\{\frac{1}{2} \sum_{J\in\mathcal{J}} |J| \cdot \left\| R_{|J|}^{-1} R^2 \overline{\mathbf{x}}_J \right\|^2 \right\} =: Q_{\mathbf{x}}(\mathcal{J}) \tag{4}$$

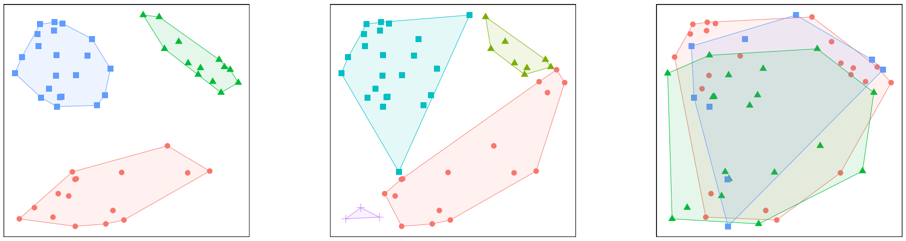
where $C = \alpha/\sqrt{\det T}$, $R = \Sigma^{-1/2}$, $R_m = (\Sigma^{-1} + T^{-1}/m)^{1/2}$ for $m \in \mathbb{N}$, $\|\cdot\|$ is the standard Euclidean norm in $\mathbb{R}^d$ and $\overline{\mathbf{x}}_J = \frac{1}{|J|}\sum_{j\in J} x_j$ is the mean of observations in the cluster J.

Definition 2. The *maximal a posteriori* (MAP) partition of $[n]$ with observed $\mathbf{x} = (x_i)_{i=1}^n$ is any partition of $[n]$ that maximises $Q_{\mathbf{x}}(\cdot)$ (or, equivalently, the posterior probability). We denote the maximiser by $\hat{\mathcal{J}}(\mathbf{x})$.

2 Results and Examples

Our first result concerns the geometry of the MAP partition. It states that the convex hulls of the clusters are 'almost disjoint' (they may have at most one point in common, which must be a data point). When the data points are distinct, it clearly implies that the convex hulls are disjoint (cf. Figure 1).

Proposition 1. *For every $n \in \mathbb{N}$ if $J_1, J_2 \in \hat{\mathcal{J}}(x_1, \ldots, x_n)$, $J_1 \neq J_2$ and A_k is the convex hull of the set $\{x_i \colon i \in J_k\}$ for $k = 1, 2$ then $A_1 \cap A_2$ is an empty set or a singleton $\{x_i\}$ for some $i \leq n$.*



(a) This is a convex partition.

(b) This is a convex partition which is not ideal.

(c) This partition is not convex.

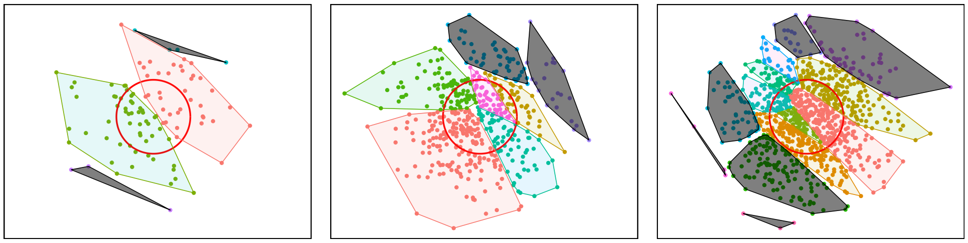
Figure 1: Illustration of the convexity property of a partition of the data.

Proposition 2 states that the clusters are of reasonable size; if the sequence of means of squared Euclidean norms of observations is bounded, then the size of clusters that intersect any fixed ball is comparable with the number of observations. As a result, the number of clusters that intersect any fixed ball remains bounded as the number of observations increases (cf. Figure 2).

Proposition 2. *If $\sup_n \frac{1}{n}\sum_{i=1}^n \|x_n\|^2 < \infty$ then*

$$\liminf_{n\to\infty} \min\{|J| \colon J \in \hat{\mathcal{J}}(x_1, \ldots, x_n), \exists_{j\in J} \|x_j\| < r\}/n > 0$$

for every $r > 0$.



(a) $n = 100$ (b) $n = 500$ (c) $n = 1000$

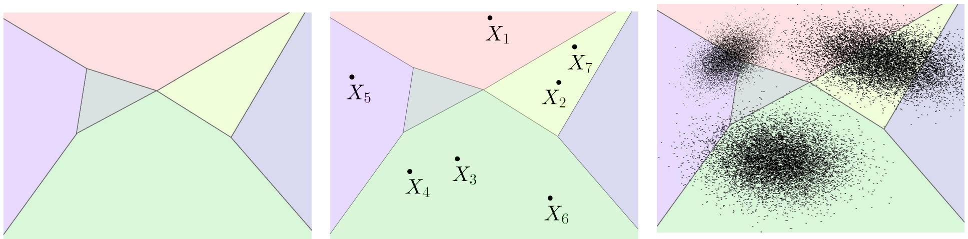
Figure 2: Illustration of Proposition 2. The red circle is arbitrarily fixed and the clusters it intersects are coloured. The number of observations in each coloured cluster is proportional to n and the number of these clusters remains bounded as $n \to \infty$.

Now consider a slightly different setting. Let P be a probability distribution on $\mathbb{R}^d$ and let $X_1, X_2, \ldots \sim P$. Assume that $\mathcal{A}$ is a finite partition of the observation space $\mathbb{R}^d$ into sets with positive P measure. This *induces* a random partition of indices in a natural way: two indices are in the same cluster if the respective observations are in the same element of $\mathcal{A}$. Formally, we consider a partition of $[n]$ given by $\mathcal{J}_n^{\mathcal{A}} = \big\{\{i \leq n \colon x_i \in A\} \colon A \in \mathcal{A}\big\}$ (cf. Figure 3). Therefore we can compute the *posterior score* of this random partition (computed with respect to our Normal DPMM model). If we take the n-th root of this value, it converges.

Lemma 1. *Let $\mathcal{A}$ be a finite partition of $\mathbb{R}^d$ consisting of Borel sets with positive P measure. Then almost surely $\sqrt[n]{Q_{\mathbf{X}_{1:n}}(\mathcal{J}^{\mathcal{A}})} \approx \frac{n}{e} \exp\left\{\Delta(\mathcal{A})\right\}$, where for any finite family $\mathcal{G}$ of measurable sets:*

$$\Delta(\mathcal{G}) = \frac{1}{2} \sum_{G \in \mathcal{G}} P(G) \left\| \Sigma^{-1/2} \mathbb{E}(X \mid X \in G) \right\|^2 + \sum_{G \in \mathcal{G}} P(G) \ln P(G). \tag{5}$$

Hence, in some sense, the Δ function measures how well the partition $\mathcal{A}$ fits the probability measure P (according to our DPMM model).



(a) Partition $\mathcal{A}$ of the observation space $\mathbb{R}^2$.

(b) $\mathcal{J}_7^{\mathcal{A}}$ is equal to $\{\{1\},\{2,7\},\{3,4,6\},\{7\}\}$

(c) We can approximate $\sqrt[10^5]{Q_{X_{1:10^5}}(\mathcal{J}^{\mathcal{A}}_{10^5})}$

Figure 3: Illustration of the induced partitions and Lemma 1.

We already know that the convex hulls of the clusters in the MAP partition are pairwise disjoint. It implies that in a way the clusters of the MAP are *induced* by their convex hulls in the sense given above. It poses a natural question: does the analogue of Lemma 1 hold for the posterior score of the MAP partition? It is possible to justify the positive answer when P has a bounded support and the family of convex sets $\mathcal{C}$ is a *Glivenko-Cantelli class* with respect to P, i.e. $\sup_{C\in\mathcal{C}}\big|P_n(C)-P(C)\big|\overset{a.s.}{\to}0$ $(*)$.

Lemma 2. *Assume that P has a bounded support and satisfies $(*)$. Let $\hat{\mathcal{A}}_n=\big\{\mathrm{conv}\{\mathbf{X}_j\colon j\in J\}\colon J\in\hat{\mathcal{J}}\big\}$. Then $\sqrt[n]{Q_{\mathbf{X}_{1:n}}(\hat{\mathcal{J}}_n)}\overset{a.s.}{\approx}\frac{n}{e}\exp\{\Delta(\hat{\mathcal{A}}_n)\}$.*

It should be pointed out that the family of probabilities P that satisfy $(*)$ is relatively large. For example, in [1] it is proved that if for every $C\in\mathcal{C}$ the boundary ∂C can be covered by countably many hyperplanes plus a set of P-measure zero, then $(*)$ holds for P (so it is true e.g. for distributions continuous with respect to the Lebesgue measure on $\mathbb{R}^d$).

Lemma 2 allow us to infer about the limits of the MAP partitions for large samples sizes. In order to formulate this result, we need to consider the *symmetric distance metric* over P-measurable sets, which is defined by $d_P(A,B)=P\big((A\setminus B)\cup(B\setminus A)\big)$. This can be easily extended to a metric $\overline{d}_P$ over finite families of measurable subsets of $\mathbb{R}^d$. Let $\boldsymbol{M}_\Delta$ denote the set of finite partitions that maximise the function Δ.

Proposition 3. *Assume that P has bounded support and is continuous with respect to Lebesgue measure. Then $\boldsymbol{M}_\Delta\neq\emptyset$ and $\inf_{\mathcal{M}\in\boldsymbol{M}_\Delta}\overline{d}_P(\hat{\mathcal{A}}_n,\mathcal{M})\overset{a.s.}{\to}0$.*

As indicated by the following Proposition 4 and the examples, this may lead to an inconsistent behaviour of the MAP partition.

Proposition 4. *Assume that P has bounded support and is continuous with respect to Lebesgue measure. Then for every $K\in\mathbb{N}$ there exists an $\varepsilon>0$ such that if $\|\Sigma\|:=\sup_{v\in\mathbb{R}^d}\|\Sigma v\|/\|v\|<\varepsilon$ then $|\hat{\mathcal{J}}_n|>K$ for sufficiently large n.*

Example 1. Let P be the uniform distribution on $[-1,1]$. It can be computed (see [2], Supplement A) that the unique optimiser of the Δ function is the partition of $[-1,1]$ into K segments of equal lengths, where $K\approx\Sigma^{-1/2}/\sqrt{3}$. It is worth noting that the variance of the data within a segment of length $2\Sigma^{-1/2}/\sqrt{3}$ is equal to $\Sigma^{-1/2}$, so in this case the MAP clustering splits the data to adjust the empirical within-group covariance to the model assumptions.

Example 2. Let P be a mixture of two normals: $P=\frac{1}{2}(\nu_{-1.01}+\nu_{1.01})$, where ν_m is the normal distribution with mean m and variance 1. Choose the model parameters consistent with the input distribution, i.e. $d=\alpha=\Sigma=\mathrm{T}=1$. It can be computed numerically that $\Delta(\{(-\infty,0],(0,\infty)\})<0=\Delta(\{\mathbb{R}\})$. In this case a partition of the data into positive and negative is intuitive and for sufficiently large data input the posterior score for the two clusters partition is smaller than for a single cluster. This may be taken as an indication of inconsistency of the MAP estimator in this setting.

3 Discussion

It is clear that the setting of our considerations is rather limited. Firstly, the object of our analysis is a very specific DPM model. It is natural to investigate if these result hold for different models, for example when we allow the covariance matrices between clusters to vary. Secondly, the limiting results contained here are proved in the case where the support of the input distribution is bounded. In this case the model is clearly misspecified. Extending these results to the case of unbounded input distribution P would greatly improve the applicability of this work.

Bibliography

[1] J. Elker, D. Pollard, and W. Stute. Glivenko-Cantelli theorems for classes of convex sets. *Advances in Applied Probability*, 11(4):820–833, 1979.

[2] Ł. Rajkowski. Analysis of the maximal a posteriori partition in the Gaussian Dirichlet Process Mixture Model. *Bayesian Anal.*, 14(2):477–494, 06 2019.

Application of generalized random forests for survival analysis

Helene Charlotte Rytgaard*

[1]*Section of Biostatistics, University of Copenhagen*

Abstract: We are interested in estimating treatment effects on the absolute risk of an event in a survival analysis setting. The particular approach taken in this paper is based on the generalized random forest (GRF) [2] methodology that we adapt to right-censored data. We formulate the estimation problem in terms of counterfactual outcomes where both treatment and censoring act as a coarsening on the underlying survival time, and define our target parameter as the solution to an inverse probability weighted estimating equation. To grow the forest, we use a partitioning scheme (splitting criteria) based on the influence function for our target parameter. The result is a nonparametric estimator for the treatment effect on survival.

Keywords: Survival analysis, random forests, causal inference, treatment effects, censored data.

1 Introduction

Estimation of average treatment effects by means of machine learning methods has applications in fields such as biostatistics and econometrics and is a popular alternative to parametric and semiparametric methods. This article is concerned with the adaption of the generalized random forest (GRF) [2] framework, a recent extension of the original random forest [4] based on subsampling and honesty, to estimation of treatment effects based on right-censored data. The GRF methodology is formulated in terms of estimating equations of the form,

$$\mathbb{E}\big[\psi_{\theta(x),\nu(x)}(O)\,|\,X=x\big]=0, \tag{1}$$

for estimation of a parameter $\theta(x)$ based on data $(X_i, O_i) \in \mathcal{X} \times \mathcal{O}$, $\mathcal{X} \subseteq \mathbb{R}^p$ where $\psi_{\theta(x),\nu(x)}(\cdot)$ is a scoring function and $\nu(x)$ is an optional nuisance parameter. GRFs have been applied to estimation of heterogeneous treatment effects [12, 2] but not for censored time-to-event outcomes.

*Corresponding author: hely@sund.ku.dk

A random forest consists of trees where each tree recursively splits subsamples of data using a specific partitioning scheme. Central to the GRF algorithm is that the partitioning scheme targets specifically the estimation of $\theta(x)$. The idea is to label subjects with the influence function of a local estimator for the target parameter. Then a split is implemented such as to maximize heterogeneity in the labeled subjects. By averaging over the neighborhoods defined by each tree, the forest outputs a weighting function that can be used to find solutions to the estimating equation (1).
To adapt the GRF methodology to the survival analysis setting we consider a specific estimation equation that involves a Kaplan-Meier integral for which we derive the influence function. The forest weights define a kernel function based on which we construct an estimator that solves the estimating equation of interest. That way, we obtain a nonparametric estimator, allowing for covariate-dependent censoring, that is targeted directly towards the treatment effect on survival.

2 Setting and notation

Suppose we make $n \in \mathbb{N}$ independent and identically distributed observations of,

$$X \in \mathbb{R}^p, \quad O = (A, \tilde{T}, \Delta) \in \{0,1\} \times \mathbb{R}_+ \times \{0,1\},$$

where $\tilde{T}$ is a continuous time-to-event outcome observed under right-censoring, $\Delta \in \{0,1\}$ is an indicator of event, $X \in \mathcal{X} \subseteq \mathbb{R}^p$ is a vector of baseline covariate values, and $A \in \{0,1\}$ is a binary treatment assigned at baseline. We represent the observed data $(X, A, \tilde{T}, \Delta)$ as a many-to-one mapping on the full data structure (X, T^0, T^1) induced by a coarsening by (A, C) [10, 9]. Here, C is the censoring time and T^a is the uncensored counterfactual event time that would result if treatment had been set to $A = a$. The observed survival outcome variables are then given as $\tilde{T} = T^A \wedge C$ and $\Delta = \mathbb{1}\{T^A \leq C\}$.
Our interest is in the counterfactual distributions $F^a(t \,|\, x) = P(T^a \leq t \,|\, X = x)$ for $a = 0,1$. We further use the notation $F(t, a \,|\, x) = P(T \leq t, A = a \,|\, x)$, $G(t, a \,|\, x) = P(C > t, A = a \,|\, x)$, $H^\delta(t, a \,|\, x) = P(\tilde{T} \leq t, \Delta = \delta, A = a \,|\, X = x)$ for $\delta = 0,1$ and $H(t, a \,|\, x) = P(\tilde{T} \geq t, A = a \,|\, X = x)$. For a fixed timepoint $t_0 > 0$, we assume coarsening at random (CAR) [10, 6], which implies $T^a \perp\!\!\!\perp (C, A) \,|\, X$, $a = 0,1$. We also assume positivity, $P(C > t_0, A = a \,|\, X) > \eta > 0$, a.s. for $a = 0,1$. We note that, under these assumptions, the conditional density of an observation O (with respect to an appropriate dominating measure) can be expressed as,

$$\begin{aligned} &P(\tilde{T} \in dt, \Delta = 1, A = a \,|\, X = x) \\ &\qquad = P(T^a \in dt \,|\, X = x) P(C > t, A = a \,|\, X = x), \end{aligned}$$

and we have the following relations,

$$F^a(t \,|\, x) = \frac{H^1(t, a \,|\, x)}{G(t, a \,|\, x)}, \qquad G(t, a \,|\, x) = \prod_{s \in (0,t]} \left(1 - \frac{H^0(ds, a \,|\, x)}{H(s, a \,|\, x)}\right), \tag{2}$$

where $\prod$ denotes the product integral [1].

3 Kernel estimation

We are concerned with estimation of $\theta(x) = \theta_1(x) - \theta_0(x)$, where,

$$\theta_a(x) = \int_0^\infty \mathbb{1}\{t > t_0\}\, dF^a(t \,|\, x), \quad a = 0, 1. \tag{3}$$

The dependence on the timepoint of interest, t_0, is implicit in the notation for $\theta_a(x)$. We note that $\theta_a(x)$ is defined as a functional of the distribution F^a of the unobservable T^a. By CAR and positivity, we can rewrite (3) using (2) as,

$$\theta_a(x) = \int_0^\infty \mathbb{1}\{t > t_0\}\, \frac{H^1(dt, a \,|\, x)}{G(t, a \,|\, x)}, \quad a = 0, 1. \tag{4}$$

This corresponds to an inverse probability weighted estimating equation of the form,

$$\mathbb{E}\Bigg[\sum_{a \in \{0,1\}} (2a - 1) \bigg(\int_0^\infty \mathbb{1}\{t > t_0\}\, \frac{H^1(dt, a \,|\, x)}{G(t, a \,|\, x)} \bigg) - \theta(x) \,\Bigg|\, X = x \Bigg] = 0,$$

with nuisance parameters (H^1, G). We consider the following estimators, for a kernel weighting function $K(x, x') \geq 0$,

$$\begin{aligned} \hat{H}^\delta_K(t, a \,|\, x) &= \sum_{i=1}^n K(x, x_i)\, \mathbb{1}\{\tilde{T}_i \leq t, \Delta_i = \delta, A_i = a\}, \quad \text{for } \delta = 0, 1, \\ \hat{H}_K(t, a \,|\, x) &= \sum_{i=1}^n K(x, x_i)\, \mathbb{1}\{\tilde{T}_i > t, A_i = a\}. \end{aligned}$$

The kernel function $K(x, x')$ is used to place more weight on observations in the covariate space $\mathcal{X}$ that are close to x. We define the estimators,

$$\hat{\theta}_{K,a}(x) = \int_0^\infty \mathbb{1}\{t > t_0\}\, \frac{\hat{H}^1_K(dt, a \,|\, x)}{\hat{G}_K(t, a \,|\, x)}, \;\; \hat{G}_K(t, a \,|\, x) = \prod_{s \leq t} \left(1 - \frac{\hat{H}^0_K(ds, a \,|\, x)}{\hat{H}_K(s, a \,|\, x)} \right).$$

In the GRF framework we replace the kernel weighting function $K(x, x')$ by forest-based weights as we will show in the following.

4 GRF for survival analysis

A random forest consists of a set of $B \in \mathbb{N}$ trees that each provides a partitioning of the covariate space. The following outlines the tree building process for the b^{th} tree in the GRF framework.

1. *Subsampling.* An index set $\mathcal{J}_b$ of size $s_n < n$ is sampled randomly from $\{1, \dots, n\}$ without replacement.

2. *Honesty.* The index set $\mathcal{J}_b$ is divided randomly into $\mathcal{J}_b^1 \cup \mathcal{J}_b^2$ of sizes $\lfloor s_n/2 \rfloor$ and $\lceil s_n/2 \rceil$.
3. *Splitting.* The tree is grown by recursively implementing binary axis-aligned splits of the covariates space based on the samples $\{i : i \in \mathcal{J}_b^1\}$. We describe the particular splitting rule used for estimation of our target parameter below.

The randomness induced by subsampling together with randomly selecting a smaller set of variables as candidates for a split ensures diversity of the different trees of the forest.

Splitting is central to the tree building scheme. In the GRF framework, splitting rules are targeted specifically towards the target parameter $\theta(x)$. Particularly, the idea is to implement splits of a mother node $M \subseteq \mathcal{X}$ into daughters $D_1 \cup D_2 = M$ so as to maximize,

$$\mathcal{L}(D_1, D_2) \equiv \sum_{j=1}^{2} P(X \in D_j \mid X \in M) \, \mathbb{E}\big[\big(\hat{\theta}_{D_j} - \theta(X)\big)^2 \mid X \in D_j\big]. \tag{5}$$

Here, $\hat{\theta}_{D_j}$ is the estimate of the target parameter in the j^{th} daughter node, corresponding to the kernel weight $K_{D_j}(x, x') = \mathbb{1}\{x' \in D_j\}$. As proposed in [2], we will approximate the splitting criterion in (5) in the following way. Let $\hat{\theta}_M$ be the estimator for the target parameter, corresponding to the kernel weight $K_M(x, x') = \mathbb{1}\{x' \in M\}$. Define,

$$\Psi(H^1, G) = \sum_{a \in \{0,1\}} (2a - 1) \int_0^\infty \mathbb{1}\{t > t_0\} \, \frac{H^1(dt, a)}{G(t, a)}.$$

We derive the influence function of the estimator $\hat{\theta}_M = \Psi(\hat{H}^1_M, \hat{G}_M)$ as the Gâteaux derivative of the functional $\Psi(H^1, G) = \Psi_1(H^1, G) - \Psi_0(H^1, G)$ in direction of δ_{O_i} [5, 11]. The influence function is given as $\mathrm{IF}(H^1, G) = \mathrm{IF}_1(H^1, G) - \mathrm{IF}_0(H^1, G)$, where, for $a = 0, 1$,

$$\begin{aligned}
\mathrm{IF}_a(H^1, G)(O_i) = \Bigg(& \frac{\mathbb{1}\{\tilde{T}_i > t_0, \Delta_i = 1, A_i = a\}}{G(\tilde{T}_i, a)} + \mathbb{1}\{A_i = a\} \times \\
& \Bigg(\frac{1 - \Delta_i}{H(\tilde{T}_i, a)} \int_{\tilde{T}_i}^\infty \mathbb{1}\{t > t_0\} \frac{H^1(dt, a)}{G(t, a)} \\
& - \int_0^\infty \mathbb{1}\{t > t_0\} \frac{H^1(dt, a)}{G(t, a)} \Bigg(\int_0^{t \wedge \tilde{T}_i} \frac{H^0(ds, a)}{\big(H(s, a)\big)^2} \Bigg) \Bigg) \Bigg) \\
& - \Psi_a(H^1, G).
\end{aligned}$$

Now, we can approximate the splitting criterion defined in (5) by,

$$\tilde{\mathcal{L}}(D_1, D_2) \equiv \sum_{j=1,2} \frac{1}{\sum_{i=1}^n \mathbb{1}\{X_i \in D_j\}} \Bigg(\sum_{\{i : X_i \in D_j\}} \mathrm{IF}(\hat{H}^1_M, \hat{G}_M)(O_i) \Bigg)^2. \tag{6}$$

The estimated influence function $\text{IF}(\hat{H}^1_M, \hat{G}_M)(O_i)$ represents the rate of change in $\hat{\theta}_M$ in direction of $O_i \in D_j$, and the criterion defined by (6) seeks to separate samples in a way such that the estimates in the daughter nodes, $\hat{\theta}_{D_j}$, $j = 1, 2$, differ as much as possible from the estimate in the mother node, $\hat{\theta}_M$.
Node M specific estimation and the approximation by (6) is defined locally for $X \in M$. When the splitting process is repeated iteratively, we move through smaller and smaller neighborhoods defined by each current mother node. We let $L_b(x) \subseteq \mathcal{X}$ denote the terminal node of the b^{th} tree that contains $x \in \mathcal{X}$. Forest weights are obtained by averaging over the neighborhoods $L_b(x)$, $b = 1, \dots, B$,

$$\alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{b,i}(x), \text{ where, } \alpha_{b,i}(x) = \frac{\mathbb{1}\{X_i \in L_b(x), i \in \mathcal{J}_b^2\}}{\sum_{k=1}^{n} \mathbb{1}\{X_k \in L_b(x), k \in \mathcal{J}_b^2\}}. \tag{7}$$

Our forest estimator for $\theta(x)$ is defined as,

$$\hat{\theta}_\alpha(x) = \sum_{a\in\{0,1\}} (2a-1) \int_0^\infty \mathbb{1}\{t > t_0\} \frac{d\hat{H}^1_\alpha(t,a)}{\hat{G}_\alpha(t,a)},$$

using the kernel function defined by the forest $K(x, x_i) = \alpha_i(x)$. The terminal nodes shrinking around x for $n \to \infty$ implies that $K(x, x_i) = \alpha_i(x) \to \delta_x$ for $n \to \infty$.

5 Discussion

In this paper we have demonstrated how the GRF methodology can be adapted to right-censored data. We have proposed a forest-based kernel weighted estimator of the treatment effect on the absolute risk and derived the influence curve to be used for the recursive splitting scheme. That way, estimation is targeted directly towards the treatment effect and optimized for the timepoint of interest.
We note that this stands in contrast to the existing random forest algorithms for survival analysis, see for instance [7, 8]. For these, splitting rules are typically based on two-sample tests for right-censored data focusing on survival estimation over the whole time range. Our approach will be useful in the application of average treatment effects as a variable importance measure. Another extension of interest deals with competing risks analysis. Here our methods could be used to rank a list of treatments in terms of their effect on hospitalization with depression or bipolar disorder in presence of the competing risk of death.

Bibliography

[1] P. K. Andersen, O. Borgan, R. D. Gill and N. Keiding. *Statistical models based on counting processes.* Springer Science & Business Media, 1993.

[2] S. Athey, J. Tibshirani and S. Wager. Generalized random forests. *The Annals of Statistics.* 47(2):1148–1178, 2019.

[3] P. J. Bickel, C. A. J. Klaassen, Y. Ritov and J. A. Wellner. *Efficient and adaptive inference in semiparametric models*, Johns Hopkins University Press, Baltimore, 1993.

[4] L. Breiman. Random forests. *Machine learning* 45(1):5–32, 2001.

[5] R. D. Gill. *Lectures on survival analysis.* Springer, Berlin, Heidelberg, 1994.

[6] R. D. Gill, M. J. van der Laan, and J. M. Robins. Coarsening at random: Characterizations, conjectures, counter-examples. *Proceedings of the First Seattle Symposium in Biostatistics*, 255–294, 1997.

[7] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer and others. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

[8] H. C. Rytgaard and T. A. Gerds. Random Forests for Survival Analysis. *Wiley StatsRef: Statistics Reference Online*, 1–8, 2018.

[9] A. Tsiatis. *Semiparametric theory and missing data.* Springer Science & Business Media, 2007.

[10] M. J. van der Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality.* Springer Science & Business Media, 2003.

[11] A. W. van der Vaart. *Asymptotic statistics.* Cambridge university press, 2000.

[12] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association.*, 113(523):1228-1242, 2018.

Bayesian model selection for a family of discrete valued time series models

Panagiota Tsamtsakiri[1*] **and Dimitris Karlis**[1]

[1] *Athens University of Economics and Business*

Abstract: Models for univariate count time series can be split into two main categories. The first one is known as parameter driven models where the time autocorrelation comes from an underlying latent process for the mean of the discrete process and the second category is the so called observation driven models where the current observations are related to the past observations in some way. This allows for easier construction and estimation of the model but due to the discreteness and the positivity of the counts special treatment is needed.

In this paper we consider the INARCH model, proposed by [3] and detailed in [4] and [5]. The models have a feedback mechanism for the mean process which is related deterministically with its past values together with past observations. Integer Autoregressive Heteroskedastic (INARCH) models belong to the class of observation driven models. While INARCH type models have gained interest, the problem of selecting the order of the terms in the specification of the models is not developed. We aim at contributing to this direction by proposing a Bayesian model selection approach for INARCH models.

In our study we consider that the conditional distribution of Y_t given the past values is a Poisson distribution and mean linked linearly or log-linearly with past values and past observations. We propose a Bayesian approach based on a transdimensional MCMC approach. At the same time we describe Bayesian estimation for INARCH models which has not been attempted so far. A real data application will be given. Simulation evidence to support the usage of the approach is also provided.

Keywords: discrete valued time series; transdimensional MCMC; model selection;

AMS subject classification: 62M10

1 Introduction to INGARCH models

We consider that

$$Y_t \mid \mathcal{F}_{t-1}^{Y} \sim Poisson(\lambda_t), \tag{1}$$

[*] Corresponding author: ptsamtsak@aueb.gr

where $\mathcal{F}^{Y}_{t-1}$ is the σ field generated by $\{Y_s : s \leq t\}$. In the linear case parameter λ_t given by

$$\lambda_t = b_0 + \sum_{i=1}^{p} \alpha_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Y_{t-j} \tag{2}$$

Due to construction of 2 we presume that b_0, α_i and b_j are positive and initial values Y_0 and λ_0 are fixed. In addition considering Poisson process, conditional mean $E[Y_t \mid \mathcal{F}_{t-1}]$ and conditional variance $Var[Y_t \mid \mathcal{F}_{t-1}]$ are equal to the parameter λ_t. This model was proposed by [4] based on GLM theory. Conditions of stationarity in two models are very important because of their use in estimation of parameters via bayesian methods. A necessary and sufficient condition to be stationary with the combination of positivity is

$$0 < \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} b_j < 1 \tag{3}$$

Estimation via classical approaches has been discussed by [3, 4] A modification of the linear INGARCH model is the log-linear INGARCH model introduced by [5]

$$Y_t \mid \mathcal{F}_{t-1} \sim Poisson(\lambda_t = \exp(\nu_t)), \tag{4}$$

where

$$\lambda_t = e^{b_0} \prod_{i=1}^{p} \lambda_{t-i}^{\alpha_i} \prod_{j=1}^{q} y_{t-j}^{b_j} \tag{5}$$

In this model parameters α_i,b_j and b_0 take values in $\mathbb{R}$ and both negative and positive correlation take into account. Equation above is inefficient in computational methods when zeros are presented as observations in logarithmic scale. So a different representation was proposed by [5] and it maps zeros of Y_{t-j} into zeros of $\log(Y_{t-j} + 1)$

$$\lambda_t = e^{b_0} \prod_{i=1}^{p} \lambda_{t-i}^{\alpha_i} \prod_{j=1}^{q} (y_{t-j} + 1)^{b_j} \tag{6}$$

This model is more computationally demanding because λ_t increase or dicrease fast accordingly to the parameters' values α_i and b_j. In this case conditions of stationarity, ergodic properties and estimation via classical approaches have been examined by [5].

2 Trans-dimensional MCMC

Due to intractability of integrals in Bayes factor, MCMC methods for bayesian model selection have been discussed by [1, 2, 6]. At all of those methods the crucial problem is the matching of dimensions in the parameter space. An alternative method for the construction of the parameter space both in nested and non-nested models has been proposed by [7]. We consider a countable set of models M, a model indexed by m $\in \mathcal{M}$ and $\theta_m \in \Theta_m$ a vector of unknown parameters. We combine all parameter vectors in one mixture parameter vector which take values of the cartesian product of models' parameter spaces $\mathbf{\Theta_{m_1}} \times \mathbf{\Theta_{m_2}} \times \cdots \times \mathbf{\Theta_{m_{10}}}$. Consequently if a model is disconnected from the likelihood then is generated from pseudopriors and posterior model probabilities are estimated by the following

$$\hat{p}(M \mid \mathbf{y}) = \frac{\sum_{i=1}^{b} I(M_i = M)}{B} \tag{7}$$

where B is the total number of iterations and M_i dentes the model we are in the i-th iteration.

3 Simulation study

In this section we present results from a small simulation experiment aiming at examining whether our approach can identify the correct structure of the time series that generated the data. We make use of sample size n=200 close to the one used later for the application and consider 10 competing models from the linear and log-linear INARCH family. For examining if this method is appropriate for model selection in our case where we have nested and non-nested models we consider two criteria.

CRITERION 1: Each conditional prior must be proper (integrating to one) and cannot be arbitrarily vague in the sense of almost all of its mass being outside any believable compact set.

CRITERION 2: (Model selection consistency) If data $\mathbf{y}$ have been generated from model M_i then posterior model probability of model M_i should converge to 1 as the sample size n $\longrightarrow \infty$.

In our case for the accomplishment of criterion 1 we suggest as pseudopriors normal densities obtaining mean and variance after "pilot runs" of MCMC for each model and considering that those pseudopriors are proper. More specifically we generate 100 datasets of size n=200 for each model and we run a trans-dimensional Markov chain with length 10000. According to criterion 2 posterior model probability of model from which we generate the data, must be close to 1 while the probabilities must be small for the other models. We have compared 10 models, 5 models of the linear family and five from the log-linear. Looking the results from figure 1 we can see that even for small sample size we can identify the correct structure with great success for most of the models. As expected in most cases the preferred model is one close (in the sense of the parameters setup) which is reasonable.

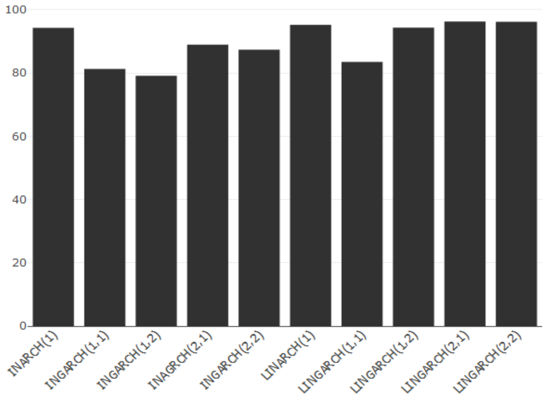


Figure 1: Averages of posterior model probabilities after 100 samples of data from each model

Predictions in time varying prediction volatility in linear and log-linear INGARCH models can be calculated. A 1-step ahead predictive pmf is defined as

$$P(\lambda_{t+1} \mid y_1, y_2, \dots y_t) = \sum_{m \in \mathcal{M}} P(\lambda_{t+1} \mid y_1, \dots y_t, m) p(m \mid y_1, \dots y_t) \tag{8}$$

which is an average of the posterior predictive distribution under each model weighted by their posterior model probabilities. For each parameter vector we calculate at each iteration λ_{t+1}. Each sampled point should be taken with probability $p(m \mid y_1, \dots, y_t)$. Then we obtain the sample of $p(\lambda_{t+1} \mid m, y)$ by weighting all samples of $p(\lambda_{t+1} \mid y)$ by the corresponding $p(m \mid y_1, \dots, y_t)$.

4 Application

We illustrate our approach by an application to estimate the parameters for each of five linear and five log-linear INGARCH models and to compare those models. The data consist of monthly counts of poliomyelitis cases in the United States from 1970 to 1983 (168 observations) reported by the Centres for Disease Control and discussed in [8] among others. In transdimensional method we concentrate jumps between five linear and five log-linear INGARCH models when both parameters α_i and b_j are positive and less than 1. For the ten linear and log-linear INGARCH models where parameters α_i and b_j are all positive and $\sum_i \alpha_i + \sum_j b_j < 1$, we apply trans-dimensional MCMC method of [7] and posterior probabilities are presented in table 1. The INGARCH(1,1) model is the one mostly visited which indicates that this is the selected model. Note that the 4 best models are of the linear type while the log-linear models have much smaller posterior probabilities.

Model	Posterior probability	Bayes Factor
INGARCH(1,1)	0.4564	3245.066
INGARCH(0,1)	0.3741	2659.801
INGARCH(1,2)	0.0641	455.838
INGARCH(2,1)	0.0588	418.188
LINGARCH(0,1)	0.0341	242.644
INGARCH(2,2)	0.0085	60.462
LINGARCH(1,2)	0.0023	16.199
LINGARCH(2,1)	0.0016	11.527
LINGARCH(2,2)	0.0001	1.000

Table 1: Posterior probabilities and Bayes factor for 5 INGARCH and 5 Log-INGARCH models

In our full work we make predictions in time varying prediction volatility in linear and log-linear INGARCH models.

Acknowledgements: This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme ≪ Human Resources Development, Education and Lifelong Learning ≫ in the context of the project " Strengthening Human Resources Research Potential via Doctorate Research " (MIS-5000432), implemented by the State Scholarships Foundation (IKY).

Bibliography

[1] B. P. Carlin and S. Chib. Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):473–484, 1995.

[2] P. Dellaportas, J. J. Forster, and I. Ntzoufras. On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12(1):27–36, 2002.

[3] R. Ferland, A. Latour, and D. Oraichi. Integer-valued garch process. *Journal of Time Series Analysis*, 27(6):923–942, 2006.

[4] K. Fokianos, A. Rahbek, and D. Tjøstheim. Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439, 2009.

[5] K. Fokianos and D. Tjøstheim. Log-linear poisson autoregression. *Journal of Multivariate Analysis*, 102(3):563–578, 2011.

[6] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[7] T. Lodewyckx, W. Kim, M. D. Lee, F. Tuerlinckx, P. Kuppens, and E.-J. Wagenmakers. A tutorial on bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, 55(5):331–347, 2011.

[8] S. L. Zeger. A regression model for time series of counts. *Biometrika*, 75(4):621–629, 1988.

On decomposable multi-type Bellman-Harris branching process for modeling cancer cell populations with mutations

Kaloyan Vitanov[1*] **and Maroussia Slavtchova-Bojkova**[2]

[1]*Sofia University "St. Kliment Ohridski"*
[2]*Sofia University "St. Kliment Ohridski"; Institute of Mathematics and Informatics, Bulgarian Academy of Sciences*

Abstract: Metastasis, the spread of cancer cells from a primary tumour to secondary location(s) in the human organism, is the ultimate cause of death for the majority of cancer patients. That is why, it is crucial to understand metastases and their evolution in order to successfully combat the disease.

We consider a metastasized cancer cell population after some medical treatment (e.g. chemotherapy). Arriving in a different environment the cancer cells may change their characteristics concerning lifespan and reproduction, thus they may differentiate into different types. Even if the treatment is effective (resulting in subcritical reproduction of all cancer cell types), however, it is possible during cell division for mutations to occur. These mutations can produce a new cancer cell type that is adapted to the treatment (having supercritical reproduction). Cancer cells from this new type may lead to a non-extinction process.

As a continuation of [3] we model the above scenario with a decomposable multi-type Bellman-Harris branching process. Expanding [2] and [4] we investigate relevant quantities such as the probability of extinction of the process until time t and as $t \to \infty$, the number of occurred supercritical mutants until time t and as $t \to \infty$ and the time until the first occurrence of a mutant starting a non-extinction process. We also propose numerical schemes for performing calculations.

Keywords: Decomposable multi–type branching processes, Probability of extinction, Mutations, Waiting time to escape mutant

AMS subject classification: Primary 60J80, Secondary 62P10

1 Notations and model description

We now state the constructive definition of our multi–type Bellman-Harris branching process (BHBP).

Definition 1. Define a multi–type BHBP with $n + 1, n \geq 1$, types of cells, as follows:

[*] Corresponding author: kvitanov@uni-sofia.bg

1. There are $n+1$, $n \geq 1$, different types of cells;

2. All cells from all types reproduce independently. Each cell type i, $i = 0, \dots, n$, has a (possibly) distinctive (continuous) distribution $G_i(t) = \mathbb{P}(\tau_i \leq t)$, $G_i(0^+) = 0$, of the lifespan τ_i and a (possibly) distinctive (discrete) distribution $\{p_{ik}\}_{k=0}^{\infty}$, $\sum_{k=0}^{\infty} p_{ik} = 1$, of the number of cells in the offspring ν_i. We denote $f_i(s) = \sum_{k=0}^{\infty} p_{ik} s^k, |s| \leq 1$ to be the probability generating function (p.g.f.) of the offspring ν_i;

3. Each descendant of a type-i cell, $i = 1, \dots, n$ can mutate at birth, independently of other cells, to any other type, with probabilities u_{ik}, $0 \leq u_{ik} \leq 1, k = 0, \dots, n$, $\sum_{k=0}^{n} u_{ik} = 1$. Descendants of the mutant type 0 cannot mutate to another type, i. e. $u_{00} = 1$, meaning also that there is no backward mutation. The process is decomposable;

4. Formally $\left\{\mathbf{Z}(t) = \left(Z^0(t), Z^1(t), \dots, Z^n(t)\right)\right\}_{t \geq 0}$, where $\{Z^i(t)\}_{t \geq 0}$ stands for the number of cells of type $i, i = 0, \dots, n$ at time t respectively.

A representation of the relationships between cell types is given in Figure 1. The interested reader is referred to [1] for related models in discrete time.

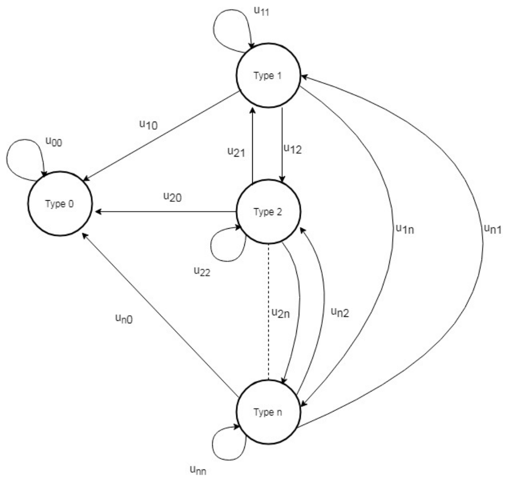


Figure 1: Flow diagram of transitions between types leading to the decomposable multi-type BHBP.

2 Number of mutations to type 0

Let us denote by $I_i(t)$ and I_i, $i = 1, \dots, n$, the random variables (r.v.) being the number of mutations to type 0 that have so far occurred until time t and the number of mutations to type 0 during the whole multi-type BHBP respectively, when the process starts with a single cell of type i. The p.g.f. of $I_i(t)$ and I_i, $i = 1, \dots, n$, will be denoted by

$$h_{I_i(t)}(s) = \mathbb{E}(s^{I_i(t)}), |s| \le 1$$

and

$$h_{I_i}(s) = \mathbb{E}(s^{I_i}), |s| \le 1.$$

Using the assumption of independence in cell reproduction, in [2] we identify the recurrence relationships:

Theorem 1. *The following integral equations hold:*

$$h_{I_i(t)}(s) = 1 - G_i(t) + \int_0^t f_i\Big(u_{i0}s + \sum_{r=1}^n u_{ir} h_{I_r(t-y)}(s)\Big) dG_i(y),$$

$$h_{I_i}(s) = f_i\Big(u_{i0}s + \sum_{r=1}^n u_{ir} h_{I_r}(s)\Big).$$

From the definitions of the r.v. $I_i(t)$ and I_i it is clear that $\lim_{t\to\infty} I_i(t) = I_i$ almost surely (a.s.). Considering that there is a one-to-one correspondence between r.v.s and p.g.f.s, it follows that $\lim_{t\to\infty} h_{I_i(t)}(s) = h_{I_i}(s)$, $|s| \le 1$.

3 Probabilities of extinction

We define the probabilities of extinction of the multi–type BHBP before a fixed time t, as follows:

$$q_i(t) = \mathbb{P}\Big(\sum_{m=0}^n Z^m(t) = 0 | Z^i(0) = 1, Z^j(0) = 0, j \ne i\Big), i = 0, \dots, n.$$

Again, due to the assumption of independence in cell reproduction, as can be seen in [5], we are able to derive recurrence relationships between $q_i(t)$, namely

Theorem 2. *The following integral equations hold:*

$$q_0(t) = \int_0^t f_0\Big(q_0(t-y)\Big) dG_0(y),$$

$$q_i(t) = \int_0^t f_i\Big(\sum_{r=0}^n u_{ir} q_r(t-y)\Big) dG_i(y), i = 1, \dots, n.$$

Further, we have

Theorem 3. *There exist* $\lim_{t\to\infty} q_i(t) = q_i$*, such that* $q_i(t) \le q_i$*,* $\forall t \ge 0$*,* $i = 0, \dots, n$*. Moreover, if the types* $i = 1, \dots, n$ *are subcritical, then the probabilities* q_i *satisfy the following equations:*

$$q_0 = f_0(q_0),$$

$$q_i = h_{I_i}(q_0) = f_i\Big(u_{i0}q_0 + \sum_{r=1}^n u_{ir} h_{I_r}(q_0)\Big), i = 1, \dots, n.$$

4 Time until occurrence of a mutant, starting a non – extincting multi–type BHBP

We introduce r.v.s T_i, $i = 1, \ldots, n$, denoting the time it takes for the occurrence of the first mutant, initiating a non–extincting multi–type BHBP, provided that the process starts with one cell of type i. Such a mutant, leading to a non–extincting processes, is called "successful" and the fact that it starts such a process is often paraphrased as "the process escapes extinction". We define $T_i = \infty$ as the event that no "successful" mutant has occurred during a process beginning with one cell of type i. That way $T_i \in (0, \infty]$.

Theorem 4. *Assume that types i, $i = 1, \ldots, n$ are subcritical. Let the process start with 1 cell type i, $i = 1, \ldots, n$. Then the distribution of r.v. T_i has the following properties:*

(i)

$$\mathbb{P}(T_i > t) \equiv Q_{i,t} = h_{I_i(t)}(q_0);$$

$$Q_{i,t} = 1 - G_i(t) + \int_0^t f_i(u_{i0}q_0 + \sum_{r=1}^n u_{ir}Q_{r,t-y})dG_i(y), \quad Q_{i,0} = 1;$$

(ii)

$$\mathbb{P}(T_i = \infty) = q_i = h_{I_i}(q_0);$$

In addition if type 0 is supercritical, then

(iii)

$$\mathbb{E}(T_i | T_i < \infty) = \frac{1}{1 - q_i} \int_0^\infty \big[h_{I_i(t)}(q_0) - h_{I_i}(q_0)\big] dt,$$

The proof of an extended version of Theorem 4 can be found in [5].

5 Calculation schemes

In this section we will briefly sketch the numerical schemes we use to calculate the derived quantities. Note that due to the established limit behaviour of $h_{I_i(t)}(s)$ and $q_i(t)$ we can obtain $h_{I_i}(s)$ and q_i by calculating for t sufficiently large.

I. Calculation scheme for $h_{I_i(t)}(s)$:

1. Let $t = 0$. For every $i = 1, \ldots, n$

$$h_{I_i(0)}(s) = 1.$$

2. Let $t = kh, k = 1, 2, \ldots$. For every $i = 1, \ldots, n$

$$h_{I_i(kh)}(s) \approx 1 - G_i(kh) + \\ + \sum_{j=1}^k f_i\Big(u_{i0}s + \sum_{r=1}^n u_{ir}h_{I_r\big((k-j)h\big)}(s)\Big) \times \Big(G_i\big(jh\big) - G_i\big((j-1)h\big)\Big).$$

Derivation:

1. Let $t = 0$. It is clear that for every $i = 1, \dots, n$
$$h_{I_i(0)}(s) = 1 - G_i(0) = 1.$$

2. Let $t = kh, k = 1, 2, \dots$ Note that for every $i = 1, \dots, n$ we can write
$$\int_0^{kh} f_i\Big(u_{i0}s + \sum_{r=1}^{n} u_{ir} h_{I_r(kh-y)}(s)\Big) dG_i(y) =$$
$$= \sum_{j=1}^{k} \int_{(j-1)h}^{jh} f_i\Big(u_{i0}s + \sum_{r=1}^{n} u_{ir} h_{I_r(kh-y)}(s)\Big) dG_i(y).$$

   Approximating the integrals in the sum through the right rectangle rule, we arrive at:
$$h_{I_i(kh)}(s) \approx 1 - G_i(kh) +$$
$$+ \sum_{j=1}^{k} f_i\Big(u_{i0}s + \sum_{r=1}^{n} u_{ir} h_{I_r\big((k-j)h\big)}(s)\Big) \times \Big(G_i\big(jh\big) - G_i\big((j-1)h\big)\Big).$$

□

II. Calculation scheme for $q_i(t)$:

1. Let $t = 0$. For every $i = 0, \dots .n$:
$$q_i(0) = \mathbb{P}\Big(\sum_{m=0}^{n} Z^m(0) = 0 | Z^i(0) = 1, Z^j(0) = 0, j \neq i\Big) = 0.$$

2. Let $t = kh, k = 1, 2, \dots,\ i = 1 \dots, n$
$$q_0(kh) \approx \sum_{j=1}^{k} f_0\Big(q_0\big((k-j)h\big)\Big) \times \Big(G_0\big(jh\big) - G_0\big((j-1)h\big)\Big),$$
$$q_i(kh) \approx \sum_{j=1}^{k} f_i\Big(\sum_{r=0}^{n} u_{ir} q_r\big((k-j)h\big)\Big) \times \Big(G_i\big(jh\big) - G_i\big((j-1)h\big)\Big).$$

Derivation: The derivation is analogous to the derivation of the scheme for calculating $h_{I_i(t)}(s)$. □

Acknowledgements: This research was supported, in part, by the Bulgarian National Science Fund at the Ministry of Science and Education, Grant No. KP-06-H22/3.

Bibliography

[1] M. Serra and P. Haccou. Dynamics of escape mutants. *Theor. Popul. Biol.*, 72:167–178, 2007.

[2] K. Vitanov and M. Slavtchova-Bojkova. Multitype branching processes in continuous time as models of cancer. *Annuaire de l'Universite de Sofia "St. Kl. Ohridski", Fac. Math and Inf.* 104:193–200, 2017.

[3] M. Slavtchova-Bojkova, P. Trayanov and S. Dimitrov. Branching processes in continuous time as models of mutations: Computational approaches and algorithms. *Computational Statistics and Data Analysis*, 113:111–124, 2017.

[4] M. Slavtchova-Bojkova and K. Vitanov. Modelling cancer evolution by multitype age-dependent branching processes. *Compt. rend. de l'acad. bulgare des Sci.*, 71:1297–1305, 2018.

[5] M. Slavtchova-Bojkova and K. Vitanov. Multi-type age-dependent branching processes as models of metastasis evolution. *Stochastic Models*, 35(3):284–299, 2019.

Author index

Diamond sponsors

Gold sponsors

Silver sponsors