Statistično društvo Slovenije
Statistical Society of Slovenia

Bernoulli Society
for Mathematical Statistics
and Probability

REPUBLIC OF SLOVENIA
STATISTICAL OFFICE

23rd European Young Statisticians Meeting

11.–15. September 2023, Ljubljana, Slovenia

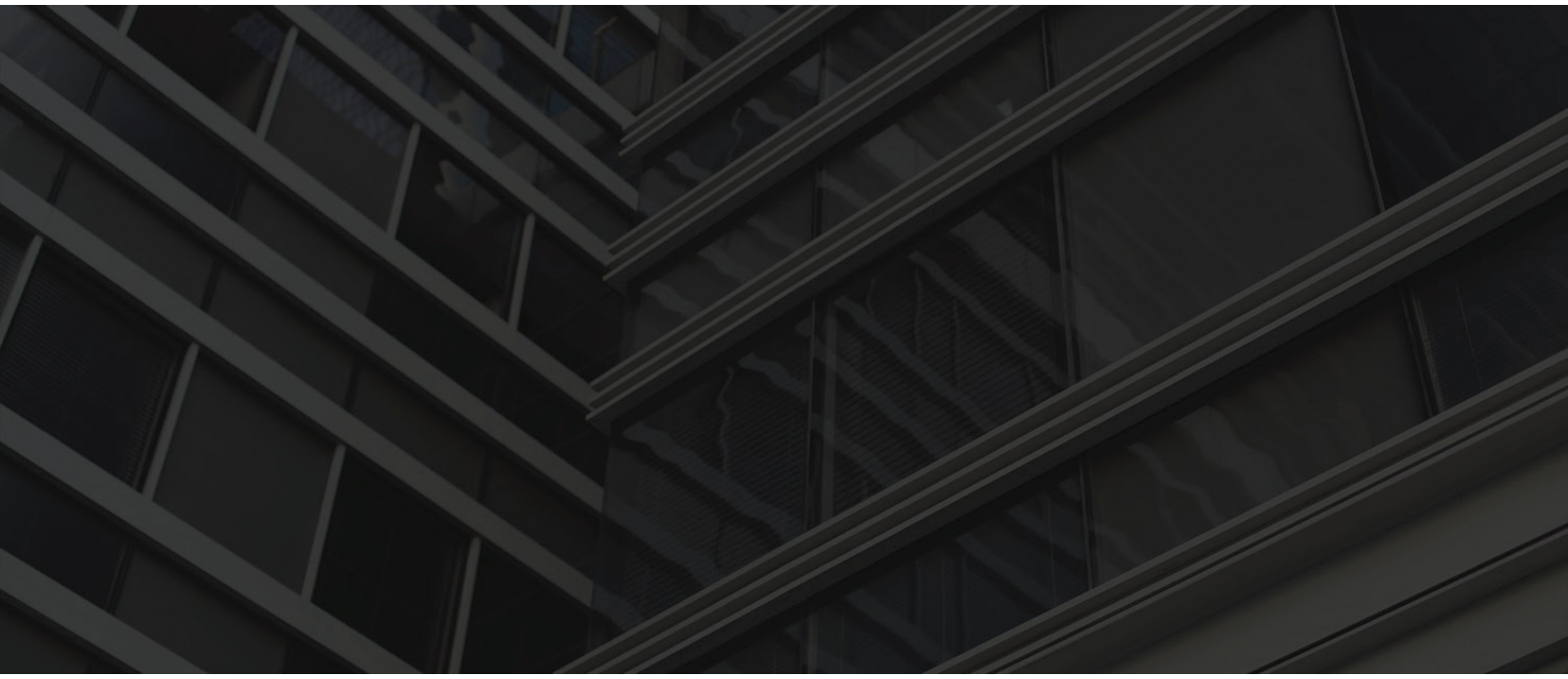
Book of abstracts

23rd European Young Statisticians Meeting – Book of Abstracts

Published by: Statistical Society of Slovenia, Litostrojska cesta 54, 1000 Ljubljana, Slovenia
organizer Statistical Society of Slovenia
Editor: Andrej Srakar
Place of publication: Ljubljana, Slovenia
Year of publication: 2025
ISBN: 978-961-94283-5-1 (PDF)

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID 212510979
ISBN 978-961-94283-5-1 (PDF)

23. Evropsko srečanje mladih statistikov je bilo izvedeno prek spleta in na Inštitutu za ekonomska raziskovanja (IER) v Ljubljani. Inštitutu se zahvaljujemo za podporo pri izvedbi dogodka.



Preface

This booklet contains basic information about the 23rd European Young Statisticians Meeting (23nd EYSM) which was held virtually, in the organization of Statistical Society of Slovenia from Monday, September 11th to Friday, September 15th, 2023.

There were twenty five European countries participating at the 23rd EYSM. The International Organizing Committee (IOC) was responsible for invitation of at most two young scientists from each country whose research interests were from pure probability theory to applied statistics. Here the term »young scientist« refers to scientists who are less than thirty years of age or have two to eight years of research experience.

The scientific part of the Conference was organized as follows:

- six eminent scientists from the field of mathematical statistics and probability held 60-minutes keynote lectures
- fourty seven invited young scientists held 20-minutes lectures.

Following the tradition of European Young Statisticians Meetings, there were no parallel sessions. The lectures of invited young scientists were divided into thirteen sessions (two or three sessions per day), nine of them having four lectures, and four of them three lectures each. Sessions were set in such manner that lectures inside one session have belong roughly to the same research area, or at least have something in common. However, the sessions were not named since in some cases it was unlikely to find a title that would suit all the lectures within the session.

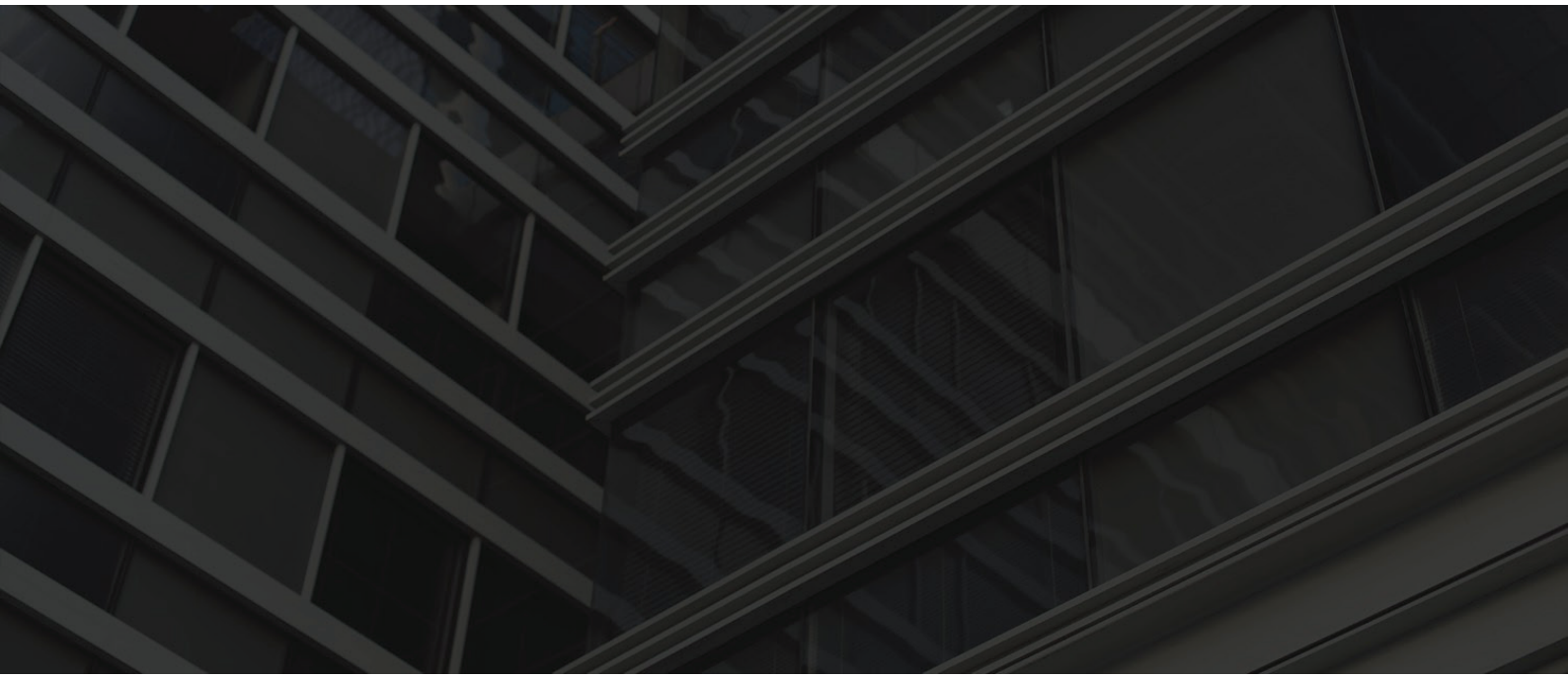
The topics presented included, but were not limited to:

- Applied statistics in economics, finance, biology, etc.
- Applied stochastic models
- Bayesian statistics and computation
- Causal inference
- Central limit theorem and asymptotics
- Dependent data
- Design of experiments
- Directional data analysis
- Distributional topics
- Econometrics
- Functional data analysis
- High-dimensional statistics
- Nonparametric statistics
- Random matrix theory
- Robust statistics
- Schramm-Loewner evolution
- Spatial data
- Statistical learning
- Stochastic differential equations
- Stochastic processes

All invited young scientists had an opportunity to publish a short paper, i.e., extended abstract of their lectures, in the Proceedings of the 23rd EYSM. The review process for short papers was organized by the IOC, in the way that the IOC representatives personally acted as referees or proposed reviewers for papers of participants they invited.

This booklet, beside all important information about the Conference, contains the scientific program, abstracts of all lectures that were given at the 23rd EYSM and the list of participants together with their affiliations and contact information. Abstracts of all contributors are given in order following the schedule of lectures from the scientific program.

More details about the **23rd European Young Statisticians Meeting** could be found at the Conference website https://sites.google.com/view/eysm2023.



Organizers

23rd European Young Statisticians Meeting

Co-Organized by
Statistical Society of Slovenia
Statistical Office of the Republic of Slovenia (SORS)

Under the auspices of
Bernoulli Society for Mathematical Statistics and Probability

International Organizing Committee
Alexandre Jacquemain, Belgium
Hrvoje Planinić, Croatia
Jan Vávra, Czechia
Niko Lietzen, Finland
Anne van Delft, Germany/United States
Christina Parpoula, Greece
Federico Camerlenghi, Italy
Aliaksandr Hubin, Norway
Łukasz Rajkowski, Poland
Marija Cuparić, Serbia
Andrej Srakar, Slovenia
Jakob Peterlin, Slovenia
Elena Castilla González, Spain
Dario Azzimonti, Switzerland
Chengchun Shi, UK (England)
Laura Boyle, UK (Northern Ireland)

Local Organizing Committee

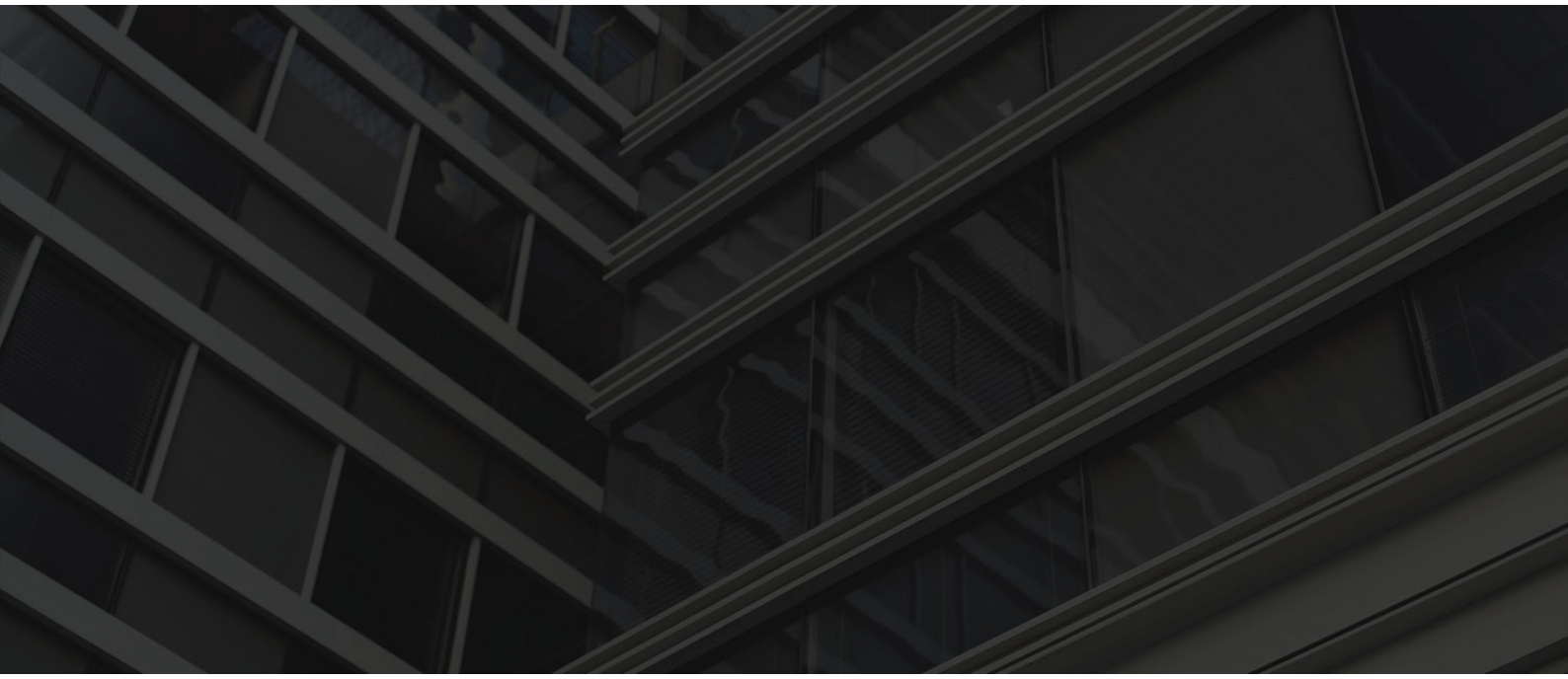
Andrej Srakar, Chair, Institute for Economic Research and University of Ljubljana
Damjan Manevski, Institute for Biostatistics and Medical Informatics, University of Ljubljana, Faculty of Medicine
Jakob Peterlin, Institute for Biostatistics and Medical Informatics, University of Ljubljana, Faculty of Medicine

Keynote Speakers

Nina Holden, New York University
Daniela Witten, University of Washington
Davy Paindaveine, Université Libre de Bruxelles
Aad van der Vaart, Delft University of Technology
Mihael Perman, University of Ljubljana and University of Primorska
Vladimir Batagelj, University of Ljubljana

Conference Structure: keynote lectures, invited lectures.

Conference Language: English



Scientic Program

Monday, September 11

10:30 - 11:00 OPENING CEREMONY

11:00 - 12:40 **Session 1** – Chairman: *Federico Camerlenghi*

11:00 - 11:25 *Bálint Horváth*
Constructing Nonparametric Simultaneous Confidence Bands for Band-Limited Functions

11:25 - 11:50 *Kimon Ntotsis*
Exploring Nonlinear Associations between Features: A Tool for Capturing Complex Relationships

11:50 - 12:15 *Mathieu Sauvenier*
Multivariate Multiscale model for Locally Stationary Processes

12:15 - 12:40 *Mario Beraha*
Frequency and cardinality recovery from sketches: a statistical viewpoint

12:40 - 14:00 Mid-day Break

14:00 - 15:15 **Session 2** – Chairman: *Aliaksandr Hubin*

14:00 - 14:25 *Matej Benko*
Bayesian Estimation of the Battery State of Charge Model Parameters

14:25 - 14:50 *Jon Lachmann*
Subsampling MCMC for Bayesian Variable Selection and Model Averaging in Bayesian Generalized Nonlinear Models

14:50 - 15:15 *Francesca Romana Crucinio*
Divide-and-Conquer sequential Monte Carlo with applications to high dimensional filtering

15:15 - 15:40 Short Break 1

15:40 - 16:40 **Keynote Lecture 1**, Chairman: *Federico Camerlenghi*

Aad VAN DER VAART
Delft University of Technology, The Netherlands
Nonparametric Bayesian uncertainty quantification: a review and some open problems

16:40 - 17:00 Short Break 2

17:00 - 18:40 **Session 3** – Chairman: *Hrvoje Planinić*

17:00 - 17:25 *Predrag Pilipovic*
Parameter Estimation in First and Second-Order Nonlinear Stochastic Differential Equations using the Strang Splitting Scheme

17:25 - 17:50 *Petra Lazić*
Subgeometric ergodicity of regime-switching diffusion processes

17:50 - 18:15 *Miha Brešar*
Lower bounds for convergence rates of ergodic strong Markov processes

18:15 - 18:40 *Zhongwei Zhang*
Extremal Dependence of Stochastic Processes Driven by Exponential-Tailed Lévy Noise

Tuesday, September 12

11:00 - 12:40 **Session 1** – Chairman: *Jan Vávra*

11:00 - 11:25 *Lídia André*
Joint modelling of the bulk and tail of bivariate data

11:25 - 11:50 *Aoife Hurley*
Joint Species Spatial Modelling of Deer Count Data

11:50 - 12:15 *Tiia-Maria Pasanen*
A Bayesian Spatio-Temporal Analysis of Markets During the Finnish 1860s Famine

12:15 - 12:40 *Ana Martins*
On the theory of space-time models for counts

12:15 - 13:40 Mid-day Break

13:40 - 15:20 **Session 2** – Chairman: *Chengchun Shi*

13:40 - 14:05 *Sarah Pirenne*
Approximate Post-Selection Inference For Variables Selected by Adaptive and Group Lasso

14:05 - 14:30 *Marouane Il Idrissi*
Cooperative game theory and importance quantification

14:30 - 14:55 *Vlad Raul Constantinescu*
Interpolation property of shallow neural networks

14:55 - 15:20 *Lars Henry Berge Olsen*
Estimating Conditional Shapley Values in Model Explanation

15:20 - 15:40 Short Break 1

15:40 - 16:40 **Keynote Lecture 2**, Chairman: *Andrej Srakar*

Mihael PERMAN
University of Ljubljana and University of Primorska, Slovenia:
Kolmogorov's statistic and point processes

16:40 - 17:00 Short Break 2

17:00 - 18:40 **Session 3** – Chairman: *Andrej Srakar*

17:00 - 17:25 *Katarina Halaj*
On diagnostic tests for circular regression models

17:25 - 17:50 *Qin Fang*
On the Modelling and Prediction of High-Dimensional Functional Time Series

17:50 - 18:15 *Jaakko Pere*
On extreme behavior of multivariate and infinite dimensional observations

18:15 - 18:40 *Alessia Caponera*
Functional estimation of anisotropic covariance and autocovariance operators on the sphere

Wednesday, September 13

11:00 - 12:15 **Session 1** – Chairman: *Dario Azzimonti*

11:00 - 11:25 *Alexandre Lecestre*
Robust estimation in finite state space hidden Markov models

11:25 - 11:50 *Carlos de la Calle-Arroyo*
Design augmentation: a tool for experimenters

11:50 - 12:15 *Henrik Imberg*
Optimal Subsampling in Measurement-Constrained Experiments

12:15 - 13:40 Mid-day Break

13:40 - 15:20 **Session 2** – Chairman: *Jakob Peterlin*

13:40 - 14:05 *Erika Lettrichová*
Stochastic modeling of age-specific fertility rates

14:05 - 14:30 *Ana Zalokar*
Optimal switching among hedging strategies in equity-linked products with guarantees

14:30 - 14:55 *Razvan-Cornel Sfetcu*
Ordering Awad-Tsallis Quantile Entropy and Applications to Some Stochastic Models

14:55 - 15:20 *Žikica Lukić*
On a Novel Two-Sample Test for Matrix Distributions and Application in Finance

15:20 - 15:40 Short Break 1

15:40 - 17:20 **Session 3** – Chairman: *Alexandre Jacquemain*

15:40 - 16:05 *Sophia Loizidou*
A data-adaptive algorithm for multiple change-point detection

16:05 - 16:30 *Victor Gallego*
Fast Adaptation with Bradley-Terry Preference Models in Text-To-Image Retrieval and Generation

16:30 - 16:55 *Jonathan Henderson*
Using Spatial Point Process Models to Understand Cell Loss in Glaucoma

16:55 - 17:20 *Tomasz Skalski*
Pattern Recovery by SLOPE

17:20 - 17:40 Short Break 2

17:40 - 18:40 **Keynote Lecture 3**, Chairman: *Anne van Delft*

Daniela M. WITTEN
University of Washington, United States
Data thinning and its applications

Thursday, September 14

11:00 - 12:00 **Keynote Lecture 4**, Chairman: *Dario Azzimonti*
Vladimir BATAGELJ
University of Ljubljana, Slovenia
Analysis of bibliographic networks

12:00 - 13:40 Mid-day Break

13:40 - 14:55 **Session 1** – Chairman: *Hrvoje Planinić*
13:40 - 14:05 *Diana Avetisian*
Parameter estimation in stochastic heat equations driven by fractional and mixed fractional Brownian motion
14:05 - 14:30 *Tomislav Kralj*
CLT for the perimeter of the convex hull spanned by two independent random walks
14:30 - 14:55 *Tobias Overgaard*
Adding unit clauses to the random 2-SAT problem

14:55 - 15:20 Short Break 1

15:20 - 16:20 **Keynote Lecture 5**, Chairman: *Andrej Srakar*
Nina HOLDEN
Courant Institute of Mathematical Sciences, New York University, United States
Random curves and surfaces

16:20 - 16:40 Short Break 2

16:40 - 18:20 **Session 2** – Chairman: *Marija Cuparić*
16:40 - 17:05 *Arianna Casanova Flores*
Desirability: some developments and applications
17:05 - 17:30 *Małgorzata Łazęcka*
Resampling methods in conditional independence testing
17:30 - 17:55 *Marianna Lakatos-Szabó*
Parametric post-processing of dual-resolution precipitation forecasts
17:55 - 18:20 *Saifuddin Syed*
Non-reversible parallel tempering on optimized paths

Friday, September 15

11:00 - 12:15 **Session 1** – Chairman: *Christina Parpoula*

11:00 - 11:25 *Spiros Dafnis*
Weak Runs in Multistate Trials and Applications in Statistical Process Monitoring

11:25 - 11:50 *Shauna Mooney*
Using service statistics to inform modern contraceptive prevalence estimates for all women of reproductive age

11:50 - 12:15 *Maria Nareklishvili*
Feature Selection for Personalised Policy Analysis

12:15 - 13:40 Mid-day Break

13:40 - 14:55 **Session 2** – Chairman: *Andrej Srakar*

13:40 - 14:05 *Mykyta Yakovliev*
Asymptotic Properties of Parameter Estimators for Mixed Fractional Brownian Motion with Trend

14:05 - 14:30 *Martina Petráková*
Some notes about Poisson–Laguerre tessellation with unbounded weights

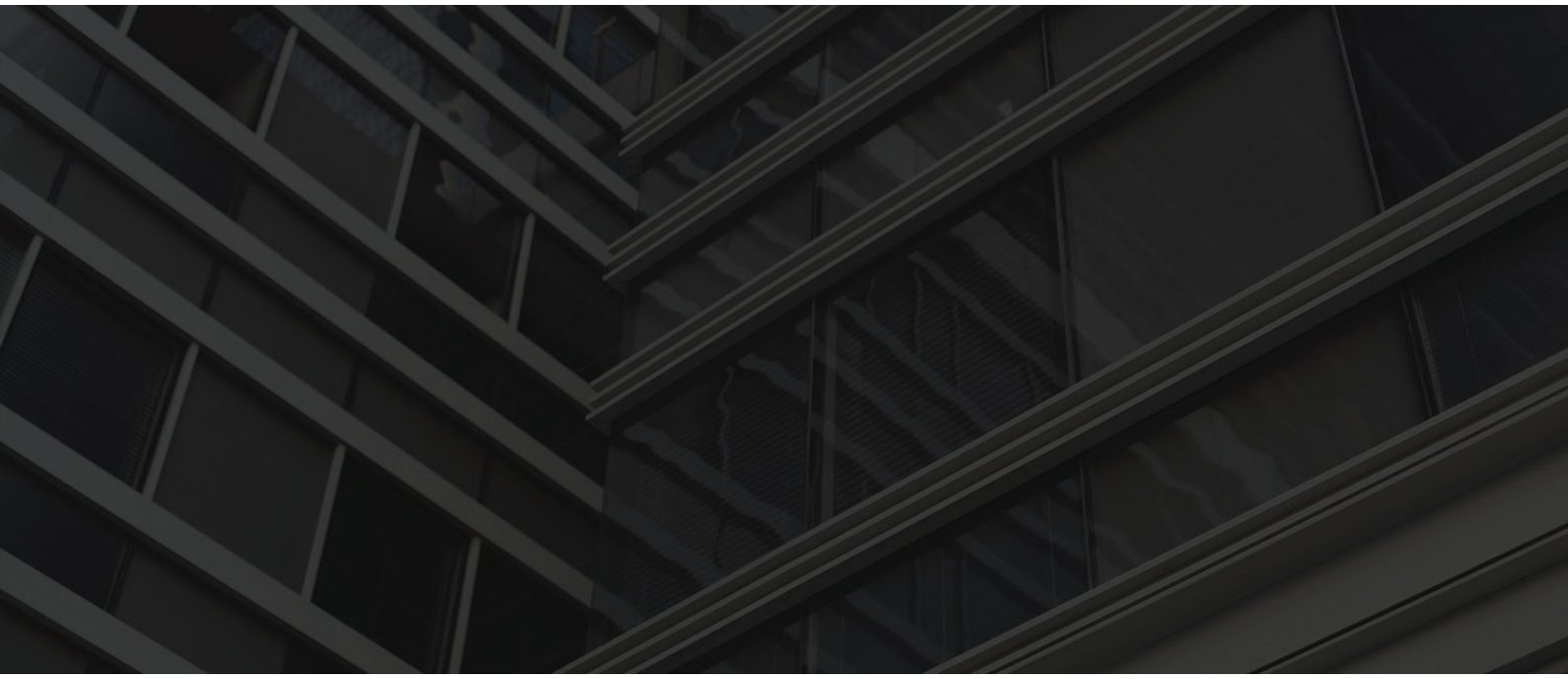
14:30 - 14:55 *Jana Reker*
Multi-Point Functional Central Limit Theorem for Wigner Matrices

14:55 - 15:20 Short Break

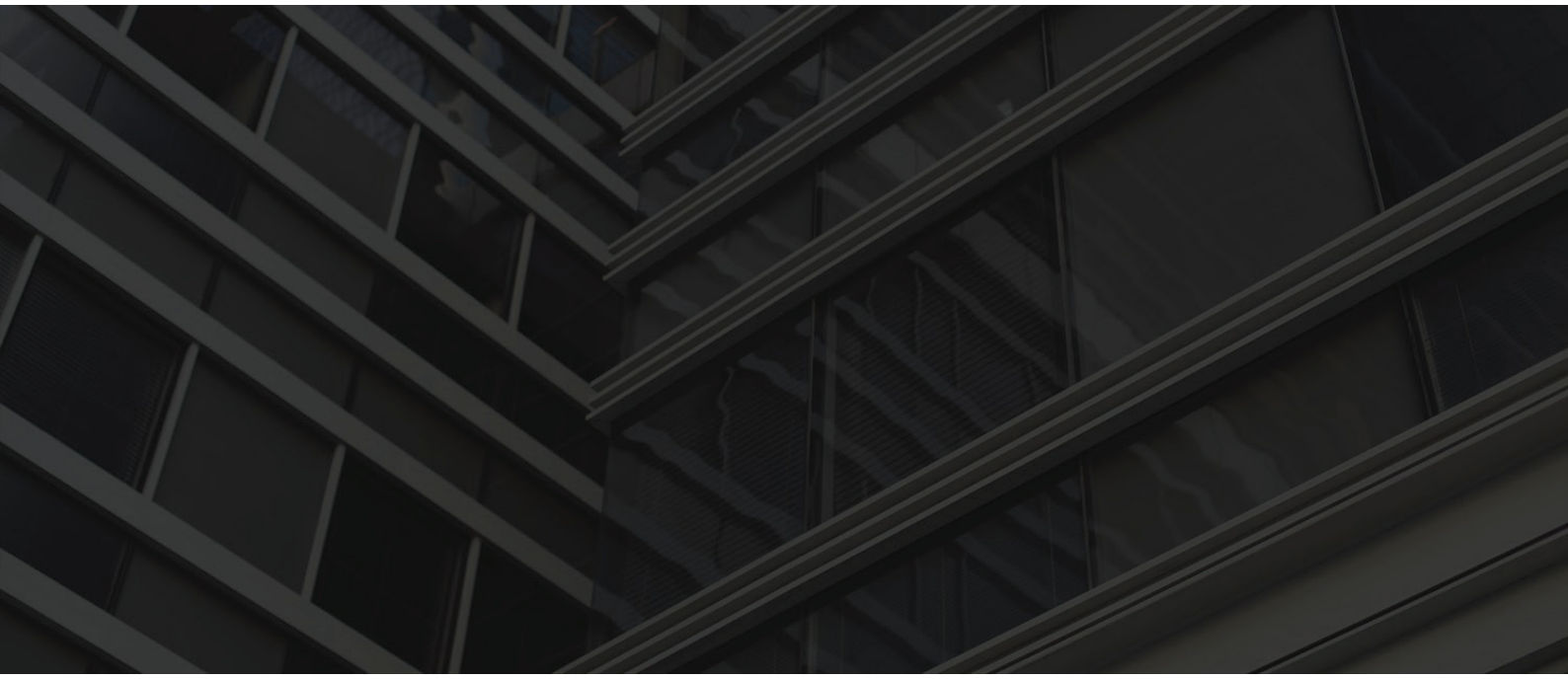
15:20 - 16:20 **Keynote Lecture 6**, Chairman: *Alexandre Jacquemain*

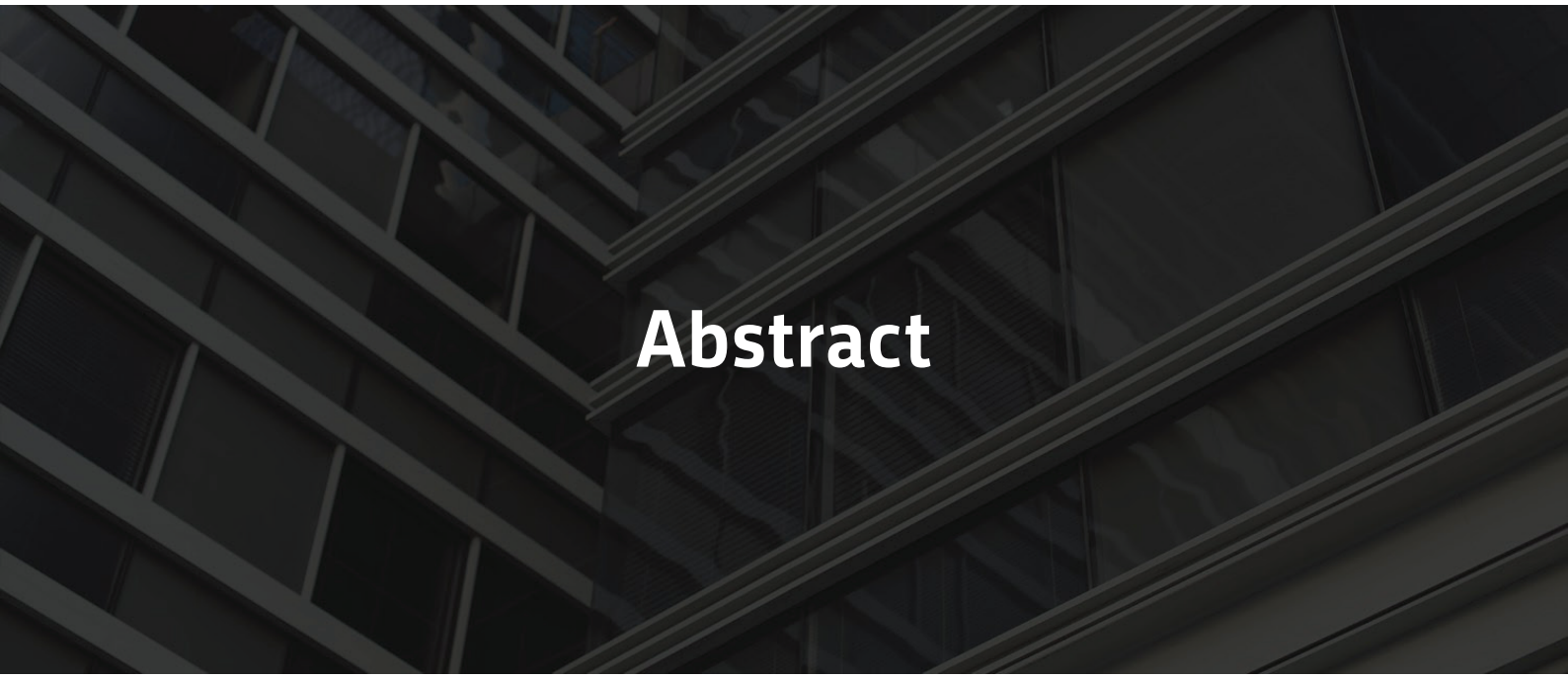
Davy PAINDAVEINE
Université Libre de Bruxelles, Belgium
Inference in principal component analysis under weak identifiability

16:20 - 17:00 CLOSING CEREMONY



Contents





Abstract

Functional estimation of anisotropic covariance and autocovariance operators on the sphere

Alessia Caponera[1]

[1]*Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca*

Abstract

In this talk we present nonparametric estimators for the second-order central moments of possibly anisotropic spherical random fields, within a functional data analysis context. We consider a measurement framework where each random field among an identically distributed collection of spherical random fields is sampled at a few random directions, possibly subject to measurement error. The collection of random fields could be i.i.d. or serially dependent. Though similar setups have already been explored for random functions defined on the unit interval, the nonparametric estimators proposed in the literature often rely on local polynomials, which do not readily extend to the (product) spherical setting. We therefore formulate our estimation procedure as a variational problem involving a generalized Tikhonov regularization term. Using the machinery of reproducing kernel Hilbert spaces, we establish representer theorems that fully characterize the form of our estimators. We determine their uniform rates of convergence as the number of random fields diverges, both for the dense (increasing number of spatial samples) and sparse (bounded number of spatial samples) regimes. A preliminary exploration of climate data will also be discussed.

Based on joint works with Victor M. Panaretos, Julien Fageot, Matthieu Simeoni, and Almond Stöcker.

Keywords: functional data analysis, measurement error, representer theorem, sparse sampling, spherical random field.

AMS subject classifications: Primary 62G08; secondary 62M.

Robust estimation in finite state space hidden Markov models

Alexandre Lecestre[1]

[1] *University of Luxembourg*

Abstract

We consider stationary hidden Markov models (HMMs) with finite state space and aim at estimating the different parameters in a robust way. It is possible to deduce them from the stationary law P_L of L consecutive observations, for L large enough. Therefore we build a robust estimator of P_L. This means we do not assume the observations to be generated by a hidden Markov model and consider possible contamination or outliers. The ρ-estimators developed by Baraud et al. [1] and Baraud & Birgé [2] are quite general and are proven to be robust in the independent setting. Our strategy is to extend their framework to handle dependent observations without losing the robustness properties. We select a subset of the observations that should be almost independent. This is possible because under weak assumptions HMMs are mixing with exponential decay rate. We prove a non-asymptotic risk bound on the squared Hellinger distance between our estimator and the target distribution. This general result is free of any assumption on the data. We can derive convergence rates with a posteriori assumptions. This method can be applied more generally to processes with Markovian and similar mixing properties. We briefly consider the estimation of the invariant distribution of a diffusion process.

Keywords: robust estimation, dependent observations, hidden Markov models, diffusion process.

AMS subject classifications: 62M05

Acknowledgements: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 811017.

References

[1] Baraud, Y., Birgé, L. and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Inventiones mathematicae.* 207, 425–517.

[2] Baraud, Y. and Birgé, L. (2018). Rho-estimators revisited: General theory and applications. *The Annals of Statistics.* 46, 3767–3804.

On the theory of space-time models for counts

Ana Martins[1], Manuel G. Scotto[2], Christian H. Weiß[3] and Sónia Gouveia[1,4]

[1]*Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and Department of Electronics, Telecommunications and Informatics (DETI), University of Aveiro, Aveiro, Portugal.*

[2]*Center for Computational and Stochastic Mathematics (CEMAT), Department of Mathematics, IST, University of Lisbon, Lisbon, Portugal.*

[3]*Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany.*

[4]*Intelligent Systems Associate Laboratory (LASI), Portugal.*

Abstract

This work introduces a new class of statistical models for the analysis of space-time series of counts. The new class is addressed as space-time ARMA (STINARMA), and hold clear connections with existing classes of models based on the ARMA framework, since it constitutes an integer counterpart of the STARMA class of Pfeifer and Deutsch, 1980 [1] based on the binomial thinning operator (BTO hereafter) of Steutel and Van Harn, 1979 [2]. Moreover, it can be seen as a spatio-temporal extension (STARMA-type) of the popular integer-valued ARMA (INARMA) class.

The probabilistic and the statistical properties of the STINARMA model are discussed in detail. Specifically, the theoretical first- and second- order moments, as well as the space-time autocovariance and autocorrelation functions are derived. Furthermore, estimation approaches based on method of moments, conditional least squares and conditional maximum likelihood are proposed, with their finite-sample performance being evaluated through a simulation study. Finally, the STINARMA models were applied to the analysis of space-time count data consisting of the daily number of hospital admissions over time in different Portuguese locations.

Keywords: spatio-temporal models; STINARMA; STARMA; integer-valued models; INARMA; binomial thinning operator.

AMS subject classifications: 62M10, 62H11, 62H12.

Acknowledgements: AM acknowledges a PhD grant from the FCT (SFRH/BD/143973/2019). This work was partially funded by the Foundation for Science and Technology, FCT, through national (MEC) and European structural (FEDER) funds, in the scope of the research projects UIDB/04621/2020 (CEMAT/UL) and UIDB/00127/2020 (IEETA/UA).

References

[1] Pfeifer P. E., Deutsch S. J. (1980). A Three-Stage Iterative Procedure for Space-Time Modeling, *Technometrics*, 22 (1), 35-47.

[2] Steutel, F. W., Van Harn, K. (1979). Discrete Analogues of Self-Decomposability and Stability, *The Annals of Probability*, 7 (5), 893-899.

Optimal switching among hedging strategies in equity-linked products with guarantees

Ana Zalokar[1] **and Mihael Perman**[1 2]

[1] *University of Primorska, Slovenia*

[2] *University of Ljubljana, Slovenia*

Abstract

Equity-linked insurance policies are one of the most widespread insurance products. In many cases such contracts have guarantees like a minimum return over the lifetime of the policy [1]. Liabilities arising from such guarantees must be hedged by suitable investments [2,3]. There are restrictions on hedging strategies in many jurisdictions but with the more flexible regulatory framework of Solvency 2 there are alternative ways to hedge certain guaranteed products using derivative securities [4]. In this talk we investigate when it is optimal to switch to hedging liabilities with derivative securities in the framework of the Cox-Ross-Rubinstein model. This leads to optimal stopping problems that can be solved explicitly. Mortality is also incorporated in the model. The results may indicate the level of reserves necessary to meet obligations with the desired level of confidence. In particular the strategy may be applicable in adverse market conditions.

Keywords: equity-linked products, hedging, optimal stopping

AMS subject classifications: 60G40, 91G30, 62P05.

Acknowledgements: This work is supported in part by the Slovenian Research Agency (research program P3-0384 and research project J1-4414).

References

[1] Bacinello, A. R., Millossovich, P., Olivieri, A. and Pitacco, E. (2011). Variable annuities: a unifying valuation approach. *Insurance Math. Econom.* 49, 285–297.

[2] Møller, T. (2001). Hedging equity-linked life insurance contracts. *N. Am. Actuar. J.* 5 , 79–95.

[3] Nonnenmacher, D. J. F. (1998) Guaranteed equity-linked products. *Proceedings of the 8th International AFIR Colloquium*, 413–428.

[4] Perman, M. and Zalokar, A. (2019), Some extensions of optimal stopping with financial applications. *Ars mathematica contemporanea*, Vol. 16, no. 2, 465–472.

[5] Shiryayev (1978). *Optimal stopping rules*, Springer-Verlag, New York.

Joint Species Spatial Modelling of Deer Count Data

Aoife K Hurley[1], Ruth F Carden[2] and James Sweeney[1]

[1]*Department of Mathematics and Statistics, University of Limerick, Ireland*

[2]*School of Archaeology, University College Dublin, Ireland*

Abstract

Accurately predicting the populations and spatial distribution of wild species is important from both a management and conservation perspective. Data on wild species can be collected in various ways including; tracking devices, motion activated cameras as well as conducting site surveys. Site survey data is generally collected at a fine spatial resolution (along a predefined path or within a 10KM square) as collection is limited by human sense (vision and hearing) and the costs of collecting data itself. In conjunction with the survey data, other informative data, such as land cover covariates or cull data from a hunting season may be available, though at a different spatial resolution, creating issues of spatial misalignment in the data sources.

In the Republic of Ireland, there are four main species of deer in the wild, with the main three being Fallow, Red and Sika. Previous studies have looked at the distribution and abundance of these species in the Republic of Ireland [1, 2], but neither account for the correlation between these species nor the effect of different land coverings on their spatial distribution. We aim to jointly model the three species of deer using both observational and cull data, allowing the borrowing or pooling of information across the species, to estimate the impacts of some land coverings, and create species distribution maps across the Republic of Ireland.

Keywords: Joint Species Modelling, Spatial Statistics, Spatial Modelling

AMS subject classifications: 62H11

Acknowledgements: This work has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number 18/CRT/6049.

References

[1] Carden, R.F., Carlin, C.M., Marnell, F., Mcelholm, D., Hetherington, J., Gammell, M.P. (2011). Distribution and range expansion of deer in Ireland. *Mammal Review* 41(4), 313–325.

[2] Morera-Pujol, V., Mostert, P.S., Murphy, K.J., Burkitt, T., Coad, B., McMahon, B.J., Nieuwenhuis, M., Morelle, K., Ward, A.I., Ciuti, S. (2023). Bayesian species distribution models integrate presence-only and presence–absence data to predict deer distribution and relative abundance. *Ecography* 2023(2).

Desirability: some developements and applications

Arianna Casanova Flores[1]

[1]*IDSIA USI-SUPSI Istituto Dalle Molle di studi sull'intelligenza artificiale, Lugano, Switzerland*
arianna.casanova@idsia.ch

Abstract

Desirability is an extension of de Finetti's Bayesian theory [4] specifically developed to handle imprecision in probabilities. This formalism offers several advantages. Firstly, it encompasses, but it is at the same time easier to work with, most of the generalisations of probability proposed in the literature, including lower and upper probabilities and previsions, convex sets of distributions, belief functions, possibility measures, and more [7]. In particular, it can deal more effectively than other models with the problem of sets of measure zero [6]. Additionally, it permits to simultaneously deal with considerations of beliefs and values and to consider any domain and possibility space [8].

The objective of this presentation is to further enhance the modeling capabilities of desirability by exploring additional developments and applications of this formalism.

We begin by investigating its applicability as a general formalism for aggregating opinions [3]. Subsequently, we analyse its connections with *information algebras* [5, 2], general algebraic structures introduced to manage information and inference at an abstract level. Finally, we discuss relaxations of some of its basic assumptions [1].

Keywords: imprecise probabilities, coherence, opinions aggregation, information algebras, monetary scale

AMS subject classifications: 60A99.

Acknowledgements: I would thank my PhD supervisor Prof. M. Zaffalon and all my co-authors: Prof. E. Miranda, Prof. J. Kohlas and Prof. A. Benavoli.

References

[1] Casanova, A., Benavoli, A., and Zaffalon, M. (2023). Nonlinear desirability as a linear classification problem. *International Journal of Approximate Reasoning*, 152:1–32.

[2] Casanova, A., Kohlas, J., and Zaffalon, M. (2022). Information algebras in the theory of imprecise probabilities. *International Journal of Approximate Reasoning*, 142:383–416.

[3] Casanova, A., Miranda, E., and Zaffalon, M. (2021). Joint desirability foundations of social choice and opinion pooling. *Annals of Mathematics and Artificial Intelligence*, 89:965–1011.

[4] De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'Institut Henri Poincaré*, volume 7, pages 1–68.

[5] Kohlas, J. (2003). *Information Algebras: Generic Structures for Inference*. Springer-Verlag.

[6] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.

[7] Walley, P. (2000). Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2-3):125–148.

[8] Zaffalon, M. and Miranda, E. (2021). Desirability foundations of robust rational decision making. *Synthese*, 198(27):6529–6570.

Constructing Nonparametric Simultaneous Confidence Bands for Band-Limited Functions

Balázs Csanád Csáji[1,2] **Bálint Horváth**[1,3]

[1]*Institute for Computer Science and Control (SZTAKI), Eötvös Loránd Research Network (ELKH), 13-17 Kende utca, 1111, Budapest, Hungary*

[2]*Institute of Mathematics, Eötvös Loránd University (ELTE), Budapest, Hungary*

[3]*Institute of Mathematics, Budapest University of Technology and Economics (BME), Budapest, Hungary*

Abstract

In this work, we study the task of constructing simultaneous and nonparametric confidence bands for band-limited functions. Our method is based on Paley-Wiener kernels [1] and uses a finite sample of independent and identically distributed (i.i.d.) input-output pairs. It is distribution-free in the sense that only mild statistical assumptions are required for the measurement noises: we only assume that their joint distribution is invariant with respect to a group operation [2]. Since the data is i.i.d., the permutation group is a suitable choice. Here, for simplicity, we also assume that the distribution of the inputs is apriori known.

First, we study a noise-free variant, when there are no measurement noises affecting the outputs, only the inputs are random. This allows us to provide an intuitive explanation for the proposed construction, which leads to solving convex optimization problems. Then, the problem is generalized to the case of measurement noises, where the solution also builds on gradient perturbation methods [3], which can construct confidence intervals at observed inputs. In both cases, our algorithms provide nonasymptotic guarantees for the inclusion of the true regression function in the confidence band, simultaneously for all possible inputs.

Finally, we show that our approach is strongly uniformly consistent, as well, where the case of having no measurement noises is discussed in detail. A related result for the case of measurement noises is also stated, and numerical examples are shown for both cases, demonstrating the proposed constructions.

Keywords: confidence bands, convex optimization, reproducing kernel Hilbert spaces, nonlinear system identification, Paley-Wiener spaces, band-limited functions, randomized algorithms

AMS subject classifications: 46E22, 62G15, 68Q32

Acknowledgements: This research was supported by the European Union within the framework of the Artificial Intelligence National Laboratory, RRF-2.3.1-21-2022-00004; and by the TKP2021-NKTA-01 grant of the National Research, Development and Innovation Office, Hungary.

References

[1] Csáji, B. Cs., Horváth, B. (2022). Nonparametric, Nonasymptotic Confidence Bands With Paley-Wiener Kernels for Band-Limited Functions, *IEEE Control Systems Letters*, 6, 3355-3360.

[2] Csáji, B. Cs., Horváth, B. (2023) Improving Kernel-Based Nonasymptotic Simultaneous Confidence Bands, *22nd IFAC World Congress*, Yokohama, Japan, 9-14 July.

[3] Csáji, B. Cs., Kis, K. B. (2019). Distribution-Free Uncertainty Quantification for Kernel Methods by Gradient Perturbations. *Machine Learning*, 108(8-9), 1677-1699.

Design augmentation: a tool for experimenters

Carlos de la Calle-Arroyo[1], Mariano Amo-Salas[2], Jesús López-Fidalgo[1] and Licesio J. Rodríguez-Aragón[2]

[1]*Instituto de Ciencia de los Datos e Inteligencia Artificial (DATAI), Universidad de Navarra*

[2]*Instituto de Matemática Aplicada (IMACI), Universidad de Castilla-La Mancha*

Abstract

Experimentation has been the reliable path for scientific discovery. However, statistical bases for it were set fairly late by Fisher [3]. In this work, model dependent design of experiments, called optimal experimental designs, are considered. In particular, designs viewed as a probability measure, approximate designs [4].

Optimal experimental designs usually have too few points and tend to have some very extremal observations, with points lying on the boundary of the design space. Often times the number of different points is the same as the number of parameters to be estimated for models, specially for models with a single explanatory variable. In this scenary, proper model adequacy testing is not possible. Due to this and other experimenter concerns, such as constraints, particular domain practice or need for robust estimation, optimal designs are not directly used in experimentation. They; therefore, are often used as a reference to measure how efficient are the designs used in practice.

In this work a methodology for design augmentation [1] is proposed as an alternative to simply rely on optimal as a benchmark. Based on the General Equivalence Theorem [5], the procedure allows to control the efficiency when adding points to a given design. The experimenter can then, starting from the optimal design, or a regulated experimental plan, add points controlling the efficiency in order to enhance the initial design to its liking. The full procedure is presented for D-optimality, while an analogous solution for D_S- and L-optimality is tentatively provided. Its software implementation is in the R package *optedr* [2].

Keywords: Optimal Design, D-optimality, Augmented Designs, General Equivalence Theorem.

AMS subject classifications: 62K05.

Acknowledgements: This work was sponsored by Ministerio de Ciencia e Innovación PID2020-113443RB-C21 and by Consejería de Educación, Cultura y Deportes of Junta de Comunidades de Castilla-La Mancha and Fondo Europeo de Desarrollo Regional SBPLY/17/180501/000380.

References

[1] de la Calle-Arroyo, C., Amo-Salas, M., López-Fidalgo, J., Rodríguez-Aragón, L. J. and Wong, W. K. (2023). A methodology to D-augment experimental designs. *Chemometrics and Intelligent Laboratory Systems.* 237.

[2] de la Calle-Arroyo, C., López-Fidalgo, J. and Rodríguez-Aragón, L. J. (2021). *optedr: Calculating Optimal and D-Augmented Designs.* https://cran.r-project.org/package=optedr.

[3] Fisher, R. A. (1935). *The Design of Experiments*, Macmillan.

[4] Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society Serie B.* 21, 272–319.

[5] Kiefer, J. and Wolfowitz, J. (1959). The equivalence of two extremum problems. *Canadian Journal of Mathematics.* 12, 363–366.

Parameter estimation in stochastic heat equations driven by fractional and mixed fractional Brownian motion

Diana Avetisian and Kostiantyn Ralchenko

Department of Probability Theory, Statistics and Actuarial Mathematics,
Taras Shevchenko National University of Kyiv, Ukraine

Abstract

We investigate a stochastic heat equation with two types of noises: fractional Brownian noise and a mixed fractional Brownian noise.

On the first step, we consider a one-dimensional stochastic heat equation driven by a space-only fractional Brownian noise. We prove that a solution to this equation is a stationary and ergodic Gaussian process. These results enable us to construct a strongly consistent estimator of the diffusion parameter assuming that Hurst parameter H is known. We prove asymptotic normality of suggested estimator.

We construct strongly consistent estimators of two unknown parameters, namely, the diffusion parameter σ and the Hurst parameter $H \in (0, 1)$, based on the discrete-time observations of a solution and prove joint asymptotic normality of the estimators in the case $H \in (0, \frac{3}{4})$.

On the second step, we consider a stochastic heat equation with a mixed fractional Brownian noise. We investigate the covariance structure, stationarity, upper bounds and asymptotic behavior of the solution. Based on its discrete-time observations, we construct a strongly consistent estimator for the Hurst index H and prove the asymptotic normality for $H < 3/4$. Then assuming the parameter H to be known, we deal with joint estimation of the coefficients at Wiener process and at fractional Brownian motion. We show that constructed joint estimator is strongly consistent and prove asymptotic normality for $H \in (0, \frac{1}{2}) \cup (\frac{1}{2}, \frac{3}{4})$.

The quality of all estimators is illustrated by simulation experiments.

Keywords: Stochastic partial differential equation, mixed fractional Brownian motion, Hurst index estimation, strong consistency, asymptotic normality.

AMS subject classifications: 60G22, 60H15, 62F10, 62F12.

References

[1] D. Avetisian, K. Ralchenko, *Ergodic properties of the solution to a fractional stochastic heat equation, with an application to diffusion parameter estimation*, Modern Stochastics: Theory and Applications, 2020, Vol. 7, No. 3, 339 – 356.

[2] D. A. Avetisian, K. V. Ralchenko, *Estimation of the Hurst and diffusion parameters in fractional stochastic heat equation*, Theory of Probability and Mathematical Statistics, 2021, Vol. 104, 61 - 76.

[3] D. Avetisian, K. Ralchenko, *Parameter estimation in mixed fractional stochastic heat equation*,Modern Stochastics: Theory and Applications, 2023, Vol. 10, No. 2, 175 – 195.

Stochastic modeling of age-specific fertility rates

Erika Lettrichová[1]

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

Abstract

Demographic process of fertility is one of the main natural mechanisms that influences the change in human populations. There are different demographics indicators connected with fertility; in this study, we focus specifically on age-specific fertility rate. Age-specific fertility rate represents the number of live births to women of certain age to the average number of women in that certain age. The goal of modeling fertility is to determine the main factors affecting fertility rates and reveal the relationships and connections between them. Out of several ways in which fertility can be modeled, we've decided to use stochastic models based on Lee's model [1], improved with various combinations of age, period and cohort parameters (see [2]). The aim of our work is to present several stochastic models for modeling age-specific fertility rates in selected populations and evaluating the best-fitting model for each of the selected populations. For quantitative comparison of models, we use the values of the likelihood function, the number of estimated parameters and the value of the modified Bayesian information criterion. We evaluate the quality of the models using estimated model residuals.

Keywords: stochastic models, age-specific fertility rate, Lee's model, probability predictions

AMS subject classifications: 62M20, 62P25, 91B70, 91D20

Acknowledgements: This research was supported by the Grant for PhD students of the Comenius University Bratislava No. UK/228/2023.

References

[1] Lee, R. D. (1993). *Modeling and forecasting the time series of US fertility: Age distribution, range, and ultimate level*, Journal of the American Statistical Association, vol. 87, no. 419, 1993, pp. 659–671.

[2] Lettrichová, E. (2022). *Stochastic modeling of age-specific fertility rates in selected countries*. Master's thesis. Faculty of Mathematics, Physics and Informatics, Comenius University Bratislava, Slovakia.

Divide-and-Conquer sequential Monte Carlo with applications to high dimensional filtering

Francesca Romana Crucinio[1]

[1] *CREST, ENSAE, Institut Polytechnique de Paris*

Abstract

In this talk I will describe the Divide-and-Conquer sequential Monte Carlo (DaC-SMC) algorithm introduced in [1], an extension of standard SMC which exploits auxiliary tree-structured decompositions of the target distribution to turn the overall inferential task into a collection of smaller sub-problems. The tree structure embedded in DaC-SMC makes it more easy to parallelise and distribute than standard SMC, I will present some recent results on the convergence properties of these algorithms and describe an application of DaC-SMC to high dimensional filtering for spatial state space models [2].

This talk is based on joint work with Adam M. Johansen and Juan Kuntz.

Keywords: distributed computing, Bayesian inference, particle filtering, state-space models

AMS subject classifications: 65C05

Acknowledgements: FRC acknowledges support from the EPSRC (grant # EP/R034710/1).

References

[1] Lindsten, F., Johansen, A.M., Næsseth, C.A., Kirkpatrick, B., Schön, T.B., Aston, J.A.D. and Bouchard-Côté, A (2017). Divide-and-Conquer with sequential Monte Carlo. *Journal of Computational and Graphical Statistics*, 26(2):445–458.

[2] Crucinio, F.R. and Johansen, A.M. (2023). A divide and conquer sequential Monte Carlo approach to high-dimensional filtering, *Statistica Sinica* In press.

Optimal Subsampling in Measurement-Constrained Experiments

Henrik Imberg

Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Gothenburg, Sweden

Abstract

We consider the problem of optimal subsample selection in an experiment setting where observing, or utilising, the full dataset for statistical analysis is practically unfeasible. This may be due to, e.g., computational or economic cost-constraints. As a result, statistical analyses must be restricted to a subset of data. Choosing this subset in a manner that captures as much information as possible is essential. Existing subsampling methods are often limited in scope and use optimality criteria (e.g., A-optimality) with well-known deficiencies, such as lack of invariance to the measurement-scale of the data and parameterisation of the model.

We present a theory of optimal design for general data subsampling problems, including finite population inference, parametric density estimation, and regression modelling using generalised linear models or quasi-likelihood methods. Our theory encompasses and generalises most existing methods in the field of optimal subdata selection based on unequal probability sampling and inverse probability weighting. We derive optimality conditions and optimal sampling schemes for a general class of optimality criteria under Poisson and multinomial sampling designs. We also study optimal design from an expected-distance-minimising perspective. This naturally leads us to a novel class of linear optimality criteria with good theoretical and practical properties, including computational tractability and invariance under non-singular affine transformations of the data and under a re-parameterisation of the model. We discuss the use of sequential optimal design for the implementation of optimal subsampling methods in practice. An active sampling strategy is proposed that iterates between estimation and data collection with optimal subsamples, guided by machine learning predictions on yet unseen data. The methodology is illustrated on an application in the vehicle safety domain.

Keywords: active sampling, inverse probability weighting, M-estimation, optimal design, unequal probability sampling.

AMS subject classifications: 62B15, 62D05, 62K05, 62L05, 68T09.

Acknowledgements: This paper is a summary of the work contained in [1, 2, 3], which is joint work with Marina Axelson-Fisk, Johan Jonasson, Xiaomi Yang, Jonas Bärgman, and Carol Flannagan, Chalmers University of Technology and University of Michigan Transportation Research Institute. The author would also like to thank Malin Svärd and Simon Lundell at Volvo Car Corporation for sharing data.

References

[1] H. Imberg. *Optimal Subsampling Designs Under Measurement Constraints.* PhD thesis, Chalmers University of Technology, Gothenburg, 2023.

[2] H. Imberg, M. Axelson-Fisk, and J. Jonasson. Optimal subsampling designs. `arXiv:2304.03019 [math.ST]`, 2023.

[3] H. Imberg, X. Yang, C. Flannagan, and J. Bärgman. Active sampling: A machine-learning-assisted framework for finite population inference with optimal subsamples. `arXiv:2212.10024 [stat.ME]`, 2022.

On extreme behavior of multivariate and infinite dimensional observations

Jaakko Pere

Aalto University School of Science, Department of Mathematics and Systems Analysis, Espoo, Finland

Abstract

There is no canonical ordering between multivariate or infinite dimensional observations, and consequently, it is not straightforward to extend extreme value theory (see [1] for a review) to multidimensional setting. However, often some univariate reductions of the original observations such as averages, supremums or norms reveal much or all of the tail behavior of the original high dimensional data. For example, under regularly varying elliptical distributions, Mahalanobis distances encode all the information about the tails of the original multivariate observations [2]. Then the analysis is simplified significantly since only sufficient regularity for the tails of the one dimensional mapped observations is required, instead of more complicated regularity assumptions, such as regular variation of the original high dimensional observations [3].

This work considers extreme behavior of multivariate and infinite dimensional observations through their one dimensional reductions, with particular interest in elliptical models and functional data. Asymptotic results for estimators of extreme value index or extreme quantiles of specific univariate mapped values corresponding to the original high dimensional observations are derived. Approximation errors that arise naturally in practical applications are taken into account. For example, under elliptical distributions one rarely knows the true location or scatter matrices, or in the context of functional data analysis possibly only discretizations of the sample paths are observed.

Keywords: Extreme value theory, Multivariate statistics, Functional data analysis

AMS subject classifications: 62G32, 60G70

Acknowledgements: This work was supported by the Vilho, Yrjö and Kalle Väisälä Foundation.

References

[1] L. De Haan and A. Ferreira. *Extreme Value Theory: An Introduction.* Springer, New York, 2007.

[2] H. Hult and F. Lindskog. Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied Probability,* 34(3):587-608, 2002.

[3] H. Hult and F. Lindskog. Regular variation for measures on metric spaces. *Publications de l'Institut Mathématique,* 80(94):121-140, 2006.

Multi-Point Functional Central Limit Theorem for Wigner Matrices

Jana Reker[1]

[1]*IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria. E-Mail: jana.reker@ist.ac.at.*

Abstract

Consider the random variable $X := \mathrm{Tr}(f_1(W)A_1 \dots f_k(W)A_k)$ where W is a Hermitian Wigner matrix, $k \in \mathbb{N}$, and we choose regular functions $f_1, \dots, f_k$ as well as bounded deterministic matrices $A_1, \dots, A_k$. With the expectation of X recently identified by Cipolloni, Erdős, and Schröder [1, 2], we study the fluctuations around $\mathbb{E}X$ and give a functional central limit theorem on macroscopic and mesoscopic scales. Analyzing the underlying combinatorics further leads to explicit formulas for the variance of X as well as the covariance of X and $Y := \mathrm{Tr}(f_{k+1}(W)A_{k+1} \dots f_{k+\ell}(W)A_{k+\ell})$ of similar build. The results match the structure of formulas in second-order free probability obtained by Male, Mingo, Peché, and Speicher [3], which were previously only available for f_j being polynomials.

Keywords: Wigner matrix, functional central limit theorem, fluctuation moments, resolvent

AMS subject classifications: 60B20, 15B52, 46L54.

Acknowledgements: I am very grateful to László Erdős for suggesting the topic and the many discussions during my work on this project. Partially supported by ERC Advanced Grant "RMTBeyond" No. 101020331.

References

[1] Cipolloni, G., Erdős, L., and Schröder, D. (2022) Optimal multi-resolvent local laws for Wigner matrices. *Electron. J. Probab.* 27, 1–38.

[2] Cipolloni, G., Erdős, L., and Schröder, D. (2023) Thermalization for Wigner matrices. *J. Funct. Anal.* 282(8), 109394.

[3] Male, C., Mingo, J. A., Peché, S. and Speicher, R. (2022) Joint global fluctuations of complex Wigner and deterministic matrices. *Random Matrices: Theory Appl.* 11(2), 2250015.

Using Spatial Point Process Models to Understand Cell Loss in Glaucoma

Jonathan Henderson[1] **Benjamin M Davis**[2] **and Hannah J Mitchell**[1]

[1]*Mathematical Sciences Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast, Northern Ireland*

[2]*CLF OCTOPUS Imaging Facility, STFC Rutherford Appleton Laboratories, oxfordshire, UK*

Abstract

Glaucoma represents the leading cause of irreversible blindness globally. It is characterised by progressive degeneration of retinal ganglion cells (RGCs). Current therapeutic treatments to prevent the disease from progressing can be highly effective provided that they are administered early in the disease process, however, there are cases where cell loss is still observed even after patients begin treatment [1]. Therefore, an understanding of the neuro-degenerative process particularly the behaviour of the RGC cells as the disease progresses will maximise information to enable clinicians and researchers to develop treatment plans for this disease. The dataset for this study contains the location and sizes of RGCs within retinas of rats which have been subject to different models to replicate the disease process. In particular, the models used were the partial optic nerve transection (pONT), and ocular hypertension (OHT) to represent severe and milder cases of the disease respectively [2]. The locations of the RGCs within the retina can be treated as a point pattern enabling spatial point process techniques to be applied to the data. Through the application of summary functions such as Ripley's K function [3], it was found that the pattern of the RGCs are aggregated. This study will look at how to model this behaviour using the log Gaussian Cox process through the use of the integrated nested Laplace approximation [4] and will look to understand if marks such as the cell size have an effect on the areas of the retina which are most likely to experience cell loss.

Keywords: Point pattern analysis, log Gaussian Cox process, Neuro-degenerative diseases, integrated nested Laplace approximation.

AMS subject classifications: 62H11

Acknowledgements: Thanks to Benjamin Davis for the providing the dataset for the study as well as help and guidance with understanding the biology and also thanks to Hannah Mitchell for her guidance during the project.

References

[1] Davis, B. M., Crawley, L., Pahlitzsch, M., Javaid, F., and Cordeiro, M. F. (2016). Glaucoma: the retina and beyond. *Acta neuropathologica.* 132, 807-826.

[2] Davis, B. M., Salinas-Navarro, M., Cordeiro, M. F., Moons, L., and De Groef, L. (2017). Characterizing microglia activation: a spatial statistics approach to maximize information extraction. *Scientific reports, 7*(1), 1576.

[3] Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of applied probability, 13*(2), 255-266.

[4] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology), 71*(2), 319-392.

Subsampling MCMC for Bayesian Variable Selection and Model Averaging in Bayesian Generalized Nonlinear Models

Jon Lachmann[1] **and Aliaksandr Hubin**[2]

[1]*Stockholm University, Indicio Technologies AB*

[2]*NMBU, UiO*

Abstract Bayesian Generalized Nonlinear Models (BGNLM) [1] offer a flexible nonlinear alternative to GLM while still providing better interpretability than machine learning techniques such as neural networks, which are still a special case of BGNLM. In BGNLM, the methods of Bayesian Variable Selection and Model Averaging are applied in an extended GLM setting. Models are fitted to data using MCMC within a genetic framework by an algorithm called GMJMCMC. In this paper, we combine GMJMCMC with a novel algorithm called S-IRLS-SGD for estimating the marginal likelihoods in BGLM/BGNLM by subsampling from the data. This allows to apply GMJMCMC to tall data. Furthermore, we present a new implementation of the algorithm as a package in the programming language R. We then present a reproducibility study where our implementation of GMJMCMC is shown to recover the third Kepler's law in an analytic form. We carry out a simulation study demonstrating the performance of the contributed algorithm as compared to a less scalable full sample approach. Finally, we provide a prediction experiment.

Keywords: BMA, MJMCMC, Subsampling, Laplace Approximation

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

References

[1] Hubin, A. (2021). *Flexible Bayesian Nonlinear Model Configuration*, Journal of Artificial Intelligence Research.

On diagnostic tests for circular regression models

Bruno Ebner[1] **and Katarina Halaj**[2] **and Bojana Milošević**[3] **and Mirjana Veljović**[3]

[1] *Institute of Stochastics, Karlsruhe Institute of Technology (KIT)*

[2] *Faculty of Transport and Traffic Engineering, University of Belgrade*

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

Abstract

We present an overview of goodness-of-fit tests for circular data and apply them to residuals of the circular-circular multiplicative regression model proposed in [1]. In this model, the regression curve is expressed as a Möbius transformation and the angular error is assumed to be distributed as a wrapped Cauchy distribution. We examine some theoretical properties of test statistics considered in this setting. Furthermore, we conduct a comprehensive Monte Carlo study to compare the performance of the presented tests. The application is illustrated on several real data sets.

Keywords: circular data, circular-circular multiplicative regression, model specification testing, wrapped Cauchy distribution, Möbius transformation.

AMS subject classifications: 62H11

Acknowledgements: The research was supported by the bilateral cooperation project "Modeling complex data - Selection and Specification" between the Federal Republic of Germany (funded by the Federal Ministry of Education and Research (BMBF) and the German Academic Exchange Service (DAAD), project-ID: 57656429) and the Republic of Serbia (funded by the Ministry of Science, Technological Development and Innovations, project-ID: 337-00-19/2023-01/6).

References

[1] Kato, S., Shimizu, K. and Shieh, G.S. (2008). *A circular-circular regression model*, Statistica Sinica, 18(2), 633–645.

[2] Rao Jammalamadaka S., Jimenez-Gamero, M.D., Meintanis, S.G. (2019). *A class of goodness-of-fit tests for circular distributions based on trigonometric moments,* SORT-Statistics and Operations Research Transactions, 43, 317-336.

[3] Mardia K. V. , Jupp P. E. (2009). *Directional Statistics,* John Wiley & Sons.

Exploring Nonlinear Associations between Features: A Tool for Capturing Complex Relationships

Kimon Ntotsis[1], Andreas Artemiou[2], and Alex Karagrigoriou[3]

[1] *NIHR Leicester Biomedical Research Centre-Respiratory, Leicester, UK & Department of Respiratory Sciences, University of Leicester, UK*

[2] *Department of Mathematics, Cardiff University, United Kingdom & School of Technology and Innovation, University of Limassol, Cyprus*

[3] *Laboratory of Statistics and Data Analysis, University of the Aegean, Greece & Graphic Era Deemed to be University, India*

Abstract

When evaluating associations between features, many techniques traditionally rely on linear associations. However, real-life data often exhibits associations that deviate from linearity, necessitating the development of new approaches to assess these non-linear associations. This study introduces an association coefficient as a novel tool for detecting associations between features, utilizing kernel functions. The methodology is investigated in comparison with already existing and frequently utilized coefficients through simulations and real-world case studies. The analysis results revealed that the proposed coefficient consistently outperforms existing coefficients in terms of accuracy and performance in many cases. The methodology demonstrated optimal or satisfactory accuracy rates regardless of the underlying data distribution and sample size, without favouring linear or non-linear associations. These findings underscore the potential utility of the coefficient in various applications and emphasize its significance and impact on the modelling process. Overall, this study contributes to the advancement of statistical modelling by highlighting the effectiveness of the proposed coefficient and its potential for further research in any field where the capturing of complex relationships are vital for inferential, descriptive or predictive purposes. It underscores the importance of considering non-linear associations and provides evidence supporting the adoption of such coefficients in practical applications.

Keywords: Association, kernel function, measures of association.

AMS subject classifications: 62H20, 62R07, 46E22.

Estimating Conditional Shapley Values in Model Explanation

Lars Henry Berge Olsen[1]

[1]*Department of Mathematics, University of Oslo, Norway,* `lholsen@math.uio.no`

Abstract

Complex machine learning (ML) models often obtain accurate predictions for supervised learning problems in numerous fields but at the cost of interpretability. Not understanding the input's influence on the ML model's output is a significant drawback; hence, the field of explainable artificial intelligence has proposed several types of post hoc explanation frameworks [4]. One of the most commonly used explanation frameworks is *Shapely values*, a promising model-agnostic local explanation methodology with a solid mathematical foundation and unique theoretical properties from cooperative game theory [3, 7]. The Shapely value explanation framework explains predictions $f(\boldsymbol{x}^*)$ made by any predictive model f by providing variable importance scores that add up to the individual predictions. The precision of the explanations depends on the characteristics of the observation $\boldsymbol{x}^*$, and practitioners should consider this when applying Shapley value explanations. There are several types of Shapley value explanations [2], but we focus on conditional Shapley values for predictive models fitted to tabular data. Estimating precise conditional Shapley values is difficult as they require the estimation of non-trivial conditional expectations for all variable combinations, which grows at an exponential rate [1, 5]. We can either directly estimate the conditional expectation by regression or model the input variables' dependencies combined with Monte Carlo integration. In [6], we extend and develop new estimation methods and systematize them into classes for comparison and evaluation. In higher dimensions, we can restrict the Shapley value computations to a reduced subset of the variable combinations to simplify the computations using only the most influential combinations. This can speed up the computation time without significantly impacting the precision of the Shapley value estimates.

Keywords: Shapley values, explainable artificial intelligence, prediction explanation, feature dependence.

AMS subject classifications: 62D10, 62E17, 62G05, 62G07, 62P99, 68T01, 91A12

References

[1] Aas K., Jullum M., Løland A., *Explaining individual predictions when features are dependent: More accurate approximations to Shapley values*, Artificial Intelligence, 298 (2021), 103502.

[2] Chen H., Covert I., and Lundberg S., Lee S., *Algorithms to estimate Shapley value feature attributions*, arXiv preprint arXiv:2207.07605 (2022).

[3] Lundberg S., Lee S., *A unified approach to interpreting model predictions*, Advances in neural information processing systems (2017), 4765-4774.

[4] Molnar C., *A Guide for Making Black Box Models Explainable* (2nd ed.), `christophm.github.io/interpretable-ml-book`, 2022.

[5] Olsen L. H. B., Glad I. K., Jullum M., Aas K., *Using Shapley Values and Variational Autoencoders to Explain Predictive Models with Dependent Mixed Features*, Journal of Machine Learning Research, 23:213 (2022), 1-51.

[6] Olsen L. H. B., Glad I. K., Jullum M., Aas K., *A Comparative Study of Methods for Estimating Conditional Shapley Values and When to Use Them*, arXiv preprint arXiv:2305.09536 (2023).

[7] Shapley L. S., *A value for n-person games*, Contributions to the Theory of Games, 2:28 (1953), 307-317.

Joint modelling of the bulk and tail of bivariate data

Lídia André[1], **Jennifer Wadsworth**[1] **and Adrian O'Hagan**[2]

[1] *Lancaster University, UK*

[2] *University College of Dublin, Ireland*

Abstract

Modelling the full data set accurately is important in situations where the interest lies in both extreme and non-extreme data. In a univariate setting, there is an increasingly rich literature which concern the fit of the bulk and tail of a distribution simultaneously. However, when more than one variable is involved, methods that perform inference and allow for a smooth transition between the two regions are scarce in the literature, motivating novel approaches.

We propose a dependence model which blends two copulas with different characteristics over the whole range of the support. One copula is tailored to the body and the other to the tail, with a dynamic weighting function employed to transition smoothly between them. We investigate extremal dependence properties, and show that the weighted model is flexible enough to capture a wide variety of structures.

We then apply our model to study the dependence between temperature and ozone concentration at two sites in the UK. Since the concentration of pollutants in the air might be so high that harmful levels exist in the bulk of the data, we not only care about the probability of exceeding extreme pollutant levels, but also about the probability of exceeding harmful yet locally moderate levels, from a public health perspective. By comparing with single copula fits, we show that the blended model is not only able to provide a better fit in both the bulk and tail of the data set, but is also able to capture different dependence structures within the same data. This presentation is based on the work in [1].

Keywords: Copulas, Dependence, Extremal Dependence

AMS subject classifications: 62H05, 62H10

Acknowledgements: This work is based on work completed while Lídia André was part of the EPSRC funded STOR-i centre for doctoral training (EP/S022252/1).

References

[1] André, L. M., Wadsworth, J. L., O'Hagan, A, (2022). Joint modelling of the body and tail of bivariate data. *(preprint)*, `https://arxiv.org/abs/2209.05795`

Resampling methods in conditional independence testing

Małgorzata Łazęcka[1,3], **Bartosz Kołodziejek**[2], **Jan Mielniczuk**[2,3]

[1] *University of Warsaw*

[2] *Warsaw University of Technology*

[3] *Polish Academy of Sciences*

Abstract

This presentation focuses on the testing of conditional independence between two discrete random variables, X and Y, given a third discrete variable Z, using an information-theoretic measure called Conditional Mutual Information

$$CMI(X,Y|Z) = \sum_z p(z) \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}.$$

We examine two resampling methods in this context: Conditional Randomisation [1] and Conditional Permutation [3] schemes. In Conditional Randomisation, we copy the observations $(Y_i, Z_i)_{i=1}^n$ from the sample $(X_i, Y_i, Z_i)_{i=1}^n$ and then resample X_i from the distribution $P_{X|Z_i}$ for $i \in 1, 2, \ldots, n$ independently, resulting in new observations X_i^*. In Conditional Permutation, we also copy $(Y_i, Z_i)_{i=1}^n$, but the vector $(X_i^*)_{i=1}^n$ consists of the values of $(X_i)_{i=1}^n$ permuted separately for each layer $\{i : Z_i = z\}$ of Z for $z \in \mathcal{Z}$, where $\mathcal{Z}$ denotes the support of Z. Throughout the presentation, we will analyze the properties of these resampling scenarios.

We investigate the asymptotic behaviour of estimates of a vector of probabilities for the two resampling scenarios, establish their asymptotic normality and ordering between asymptotic covariance matrices. Using these results, we derive asymptotic distributions of the empirical Conditional Mutual Information based on resampling schemes. In addition, we will also provide a non-asymptotic viewpoint.

We conduct simulations to examine the behavior of tests based on both asymptotic and non-asymptotic approaches. During the presentation, we will show the most conclusive results.

The presented results come from [2].

Keywords: conditional independence, conditional mutual information, resampling, asymptotic distribution, permutation

AMS subject classifications: 62E20, 62F40, 62H15

References

[1] Candès E., Fan Y., Janson L., Lv J.:, *On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection*, Panning for gold: model-X knockoffs for high dimensional controlled variable selection. J. R. Stat. Soc., B: Stat. Methodol. 80 (2018), 551–577

[2] Łazecka M., Kołodziejek B., Mielniczuk J., *Analysis of Conditional Randomisation and Permutation schemes with application to conditional independence testing*, in revision.

[3] Tsamardinos I., Borboudakis G., *Testing improves Bayesian network learning*, Lecture Notes in Computer Science vol. 6323 LNAI, 2010, pp. 322–337

Feature Selection for Personalised Policy Analysis

Maria Nareklishvili[1], Nicholas Polson[2] and Vadim Sokolov[3]

[1] *University of Oslo*

[2] *University of Chicago*

[3] *George Mason University*

Abstract

In this paper, we propose Forest-PLS, a feature selection method for analyzing policy effect heterogeneity in a more flexible and comprehensive manner than is typically available with conventional methods. In particular, our method is able to capture policy effect heterogeneity both within and across subgroups of the population defined by observable characteristics. To achieve this, we employ partial least squares to identify target components of the population and causal forests to estimate personalized policy effects across these components. We show that the method is consistent and leads to asymptotically normally distributed policy effects. To demonstrate the efficacy of our approach, we apply it to the data from the Pennsylvania Reemployment Bonus Experiments, which were conducted in 1988-1989. The analysis reveals that financial incentives can motivate some young non-white individuals to enter the labor market. However, these incentives may also provide a temporary financial cushion for others, dissuading them from actively seeking employment. Our findings highlight the need for targeted, personalized measures for young non-white male participants.

Keywords: causal forests, feature selection, policy analysis.

AMS subject classifications: 62Pxx, 62Gxx

Acknowledgements: We acknowledge the funding by the Research Council of Norway.

Parametric post-processing of dual-resolution precipitation forecasts

Marianna Szabó[1,2], Estíbaliz Gascón[3] and Sándor Baran[1]

[1]*Faculty of Informatics, University of Debrecen, Hungary*

[2]*Doctoral School of Informatics, University of Debrecen, Hungary*

[3] *European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*

Abstract

Recently, all major weather centres issue ensemble forecasts which - even covering the same domain - differ both in the ensemble size and spatial resolution. These parameters highly determine both the forecast skill of the prediction and the computation cost. In the last few years, the plans of upgrading the configuration of the Integrated Forecast System of the European Centre for Medium-Range Weather Forecasts (ECMWF) from a single forecast with 9 km resolution and a 51-member ensemble with 18 km resolution induced an extensive study of the forecast skill of both raw and post-processed dual-resolution predictions comprising ensemble members of different horizontal resolutions.
We investigate the predictive performance of the censored shifted gamma (CSG) [3] ensemble model output statistic (EMOS) [2] approach for statistical post-processing with the help of dual-resolution 24h precipitation accumulation ensemble forecasts over Europe with various forecast horizons. As high-resolution, the operational 50-member ECMWF ensemble is considered, which is extended with a 200-member low-resolution (29-km grid) experimental forecast. The investigated dual-resolution combinations consist of subsets of these two forecast ensembles with equal computational cost, being equivalent to the cost of the operational ensemble. Our case study verifies that, compared with the raw ensemble combinations, EMOS post-processing results in a significant improvement in forecast skill and the differences between the various dual-resolution combinations are reduced to a non-significant level. Moreover, the semi-locally trained CSG EMOS is fully able to catch up with the state-of-the-art quantile mapping [1] and provides an efficient alternative without requiring additional historical data essential in determining the quantile maps.

Keywords: ensemble forecast, post-processing, EMOS, dual-resolution, precipitation

AMS subject classifications: 62P12

Acknowledgements: Sándor Baran and Marianna Szabó were supported by the Hungarian National Research, Development and Innovation Office under Grant No. K142849. They are also grateful to the ECMWF for supporting their research stay in Reading. Finally, the authors are indebted to Martin Leutbecher for his help in connection with the new ECMWF IFS configuration.

References

[1] Gascón, E., Lavers, D., Hamill, T. M. , Richardson, D. S., Ben Bouallègue, Z., Leutbecher, M. and Pappenberger, F. (2019). Statistical postprocessing of dual-resolution ensemble precipitation forecasts across Europe. *Quart. J. Roy. Meteor. Soc.* 145, 3218–3235.

[2] Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman,T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.* 133, 1098–1118.

[3] Baran, S. and Nemoda, D. (2016). Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics* 27, 280–292.

Frequency and cardinality recovery from sketches: a statistical viewpoint

Mario Beraha[1], **Stefano Favaro**[1,2], **and Matteo Sesia**[3]

[1]*Department of Economics and Statistics, University of Torino*

[2]*Collegio Carlo Alberto*

[3]*Department of Data Sciences and Operations, University of Southern California*

Abstract

Given a lossy-compressed representation, or sketch, of data with values in a set of symbols, the frequency recovery problem considers the estimation of the empirical frequency of a new data point [1]. This is a classical problem in computer science, with applications in diverse fields such as statistical machine learning, cybersecurity, web and social network data analysis, natural language processing, sequencing analysis in biological sciences, and privacy-protecting data analysis.

We present a statistical framework for frequency recovery, whereby data $X_1, \ldots, X_n$ are not observable and we have access only to a sketch obtained by hashing data with a random function. We assume that observations $X_1, \ldots, X_n$ are i.i.d. from an unknown distribution $P = \sum p_s \delta_s$ on a universe of symbols $\mathbb{S}$. We characterize the conditional distribution of $f_{X_{n+1}} = \sum_{i=1}^{n} I(X_i = X_{n+1})$ given the sketch, from which point estimates are derived. To deal with the unknown P we consider first a worst-case approach, from which we recover the original count-min sketch algorithm [2] and then propose a class of *smoothed* estimators obtained by averaging the frequentist estimators over a "prior distribution" for P. Previously, fully Bayesian nonparametric approaches have been proposed, but proven to be computationally challenging, see [1] and the references therein. We then extend our framework to the problem of recovering the number of distinct symbols in the (unobserved) sample given the sketch. Contrary to algorithmic approaches, whereby a different sketch of the data is needed for the distinct recovery, we show that within our model-based framework, it is possible to answer to both frequency and distinct recovery problems using only one sketch.

When considering a sketch obtained from multiple hash functions, we show how the conditional distribution is $f_{X_{n+1}}$ is intractable outside trivial cases. Taking inspiration from *multimodal* learning, we propose to aggregate the distributions obtained by considering independently each hash function, and show that this yields tractable estimators whose empirical performances we evaluate on simulated and real data.

Keywords: frequency recovery, sketch, random hashing, Bayesian nonparametrics

AMS subject classifications: 62F15,62G05,68P30

Acknowledgements: Mario Beraha and Stefano Favaro received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257.

References

1 [Beraha and Favaro(2023)] Beraha, M., and Favaro, S. (2023). Random measure priors in Bayesian frequency recovery from sketches. *Preprint arXiv:2303.15029*.

2 [Cormode and Muthukrishnan(2005)] Cormode, G. and Muthukrishnan, S. (2005). An improved data stream summary: the count-min sketch and its applications. Journal of Algorithms 55, 58–75.

Cooperative game theory and importance quantification

Marouane Il Idrissi[123], Nicolas Bousquet[234], Fabrice Gamboa[1], Bertrand Iooss[123], Jean-Michel Loubes[1]

[1]*Institut de Mathématiques de Toulouse, Université Toulouse III.*

[2]*EDF R&D, Lab Chatou.*

[3]*SINCLAIR AI Laboratory.*

[4]*LPSM, Sorbonne Université.*

Abstract

Being able to quantify the importance of random inputs of an input-output black-box model is at the cornerstone of the fields of sensitivity analysis (SA) and explainable artificial intelligence (XAI). To perform this task, methods such as *Shapley effects* and *SHAP* have received a lot of attention. The former offers a solution for output variance decomposition with non-independent inputs, and the latter proposes a way to decompose predictions of predictive models. Both of these methods are based upon the *Shapley values*, an allocation mechanism from the *cooperative game theory*.

This presentation aims a shedding light on the underlying mechanism behind the paradigm of cooperative games for input importance quantification. To that extent, a link is drawn an extension of the *Möbius inversion formula* to boolean lattices leading to coalitional decompositions [1]. Allocations can be seen as aggregations of such decomposition, leading to a more general view of the importance quantification problem.

This generalization is leveraged in order to solve a problem in the context of global SA with dependent inputs. The Shapley effects are known not to be able to detect exogenous inputs (i.e., variables not in the model). Using a different allocation, namely the *proportional values*, lead to interpretable importance indices with the ability to identify such inputs [2].

These indices are illustrated on a classical problem of surrogate modeling of a costly numerical model: the transmittance performance of an optical filter. It allows for clear and interpretable decision rules for feature selection and dimension reduction.

Keywords: Uncertainty quantification ; Sensitivity analysis ; Interpretability ; Importance quantification

AMS subject classifications: Primary: 62P30 ; Secondary: 91A12, 62H99

References

[1] Il Idrissi, M., Bousquet, N., Gamboa, F., Iooss, B., and Loubes, J. M. (2023). *On the coalitional decomposition of parameters of interest*. arXiv preprint arXiv:2301.02539.

[2] Herin, M., Il Idrissi, M., Chabridon, V. and Iooss, B. (2022). *Proportional marginal effects for global sensitivity analysis*. arXiv preprint arXiv:2210.13065.

Some notes about Poisson–Laguerre tessellation with unbounded weights

Martina Petráková[1] and Zbyněk Pawlas[1]

[1] *Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University*

Abstract

The object of our research is the Laguerre tessellation (a generalisation of the well-known Voronoi tessellation) and particularly its random counterpart, called the Poisson–Laguerre tessellation. We are interested in the behaviour of certain characteristics of this random tessellation – e.g. the perimeter of the cells or the ratio of volumes of the neighbouring cells. As the exact distribution of these functionals is hard to derive, except for trivial cases, the primary concern is to examine their asymptotic properties. Several central limit theorems were derived using the stabilisation method (see [1]), which forced the assumption of bounded weights. In this talk we will present some properties of the Poisson–Laguerre tessellation, derived using the concept of tempered configurations (see [3]). Our goal is to use these properties together with the bound on the Wasserstein distance developed in [2] to derive central limit theorems for functionals of the Poisson–Laguerre tessellation with unbounded weights.

Keywords: Poisson–Laguerre tessellation, tempered configurations, central limit theorem

AMS subject classifications: 60G55, 60D05

Acknowledgements: This work was supported by the Czech Science Foundation, project no.22-15763S.

References

[1] Flimmel, D., Pawlas, Z. and Yukich, J. (2020). Limit theory for unbiased and consistent estimators of statistics of random tessellations. *Journal of Applied Probability* 57(2), 679–702.

[2] Last, G., Peccati, G. and Schulte, M. (2016). Normal approximation on Poisson spaces: Mehler's formula, second order Poincaré inequalities and stabilization. *Probability Theory and Related Fields* 165, 667—723.

[3] Roelly, S. and Zass, A. (2020). Marked Gibbs point processes with unbounded interaction: an existence result. *Journal of Statistical Physics* 179, 4, 972—996.

Bayesian Estimation of the Battery State of Charge Model Parameters

Matej Benko[1] **and Libor Žák**[1]

[1] *Institute of Mathematics, Faculty of Mechanical Engineering, Brno University of Technology, Technická 2896/2, 616 69 Brno, Czech Republic*

Abstract

To estimate hidden state from noisy observations, the Kalman filter [2] has become a commonly used approach in many technological applications. Such examples of them are navigation and control of vehicles from, e.g., Global Positioning System or radar measurements and battery state of charge estimation from current and terminal voltage on the battery. The estimation problem has to be described by a dynamic state-space model. This model is a hidden Markov model. For the linear state-space model, the Kalman filter provides the minimum-variance unbiased estimates of the hidden states. If the linearity assumption of the model is not satisfied, a generalization of the Kalman filter has to be used.

The performance of mentioned algorithms is significantly influenced by the defined state-space dynamic model. The model is usually determined by physical laws depended on some unknown parameters (temperature, humidity, etc.) and statistical parameters: process noise (the quantity describing how well the model fits the reality) and measurement noise (expected error of the sensors). These parameters can be empirically guessed by some expert or statistically estimated [1] from the behavior of the model.

In this contribution, the battery state of charge estimation is considered. This problem has a nonlinear state-space model, so the mentioned generalizations have to be incorporated. The Bayesian approach to the physical parameters (temperature, battery aging, and current losses) is taken over together with process noise estimation. The identifiability of the particular parameters is discussed, and the modification of the model is considered. The examples are provided in a simulated data set.

Keywords: Hidden Markov model, Kalman filtering, Particle filtering, Bayesian estimation, Parameter identifiability

AMS subject classifications: 62M05

Acknowledgements: This topic is part of the Doctoral Study of Matej Benko under the supervision of Libor Žák at Brno University of Technology.

References

[1] E. Mazor, A. Averbuch, Y. Bar-Shalom and J. Dayan. *Interacting multiple model methods in target tracking: a survey*, IEEE Transactions on Aerospace and Electronic Systems, **34**(1), 103-123, 1998, doi: 10.1109/7.640267.

[2] CHUI, C. K. and G. CHEN. *Kalman filtering with real-time applications.* Fifth edition. Cham, Switzerland: Springer, 2017. ISBN 978-3-319-47612-4.

Multivariate Multiscale model for Locally Stationary Processes

Mathieu Sauvenier[1] **and Sébastien Van Bellegem**[1]

[1]*Center for Operations Research and Econometrics, Université catholique de Louvain*

Abstract

The prevalent models for analyzing multivariate zero mean time series typically hinge on the assumption of covariance stationarity, signifying that the second order structure of the vector time series remains constant over time (*e.g.* [1]). This assumption greatly facilitates the mathematical analysis of covariance estimators between two multivariate time series. Nevertheless, many time series in applied sciences do not adhere to covariance stationarity and instead exhibit a time-varying second-order structure, where variance and covariance evolve over time. The focal point of the paper is a model of locally stationary wavelet processes. Here, time series are conceptualized as a linear combination of certain easily tractable functions, known as wavelets, and random increments. These types of models, founded on the principle of local stationarity [2], have been previously explored for univariate non-stationary time series (*e.g.* [3] [5]). They enable the modeling and estimation of the time-varying autocovariance.

The paper presents a Multivariate Multiscale model for Locally Stationary Processes. The model allows correlation between increments across different time series and scales, a novel achievement in the field. Identification and estimation theories are obtained thanks to the new concept of Cross-Correlation Wavelet Functions (CCWF), serving as a generalization of the concept of Autocorrelation Wavelet Functions (ACWF) [4], to measure redundancy levels among sets of non-decimated discrete wavelet functions. Owing to the linear independence of the CCWF, a unique asymptotic representation of the Evolutionary Wavelet Spectrum (EWS) is obtained in the CCWF domain for the proposed model. A consistent estimator for the EWS, called the corrected and smoothed wavelet periodogram, is studied. Simulations and an econometric application demonstrate the practical utility of the method.

Keywords: Non-stationary time series, multivariate wavelets processes, wavelet spectrum, local stationarity

AMS subject classifications: Primary 62M10, Secondary 62G05, 62M15

Acknowledgements: The authors gratefully acknowledge Daniel Koch for seminal discussions about the model presented in this paper.

References

[1] Brillinger, D. (2001). *Time series: data analysis and theory*, Society for Industrial and Applied Mathematics.

[2] Dahlhaus, R (1997). Fitting time series models to nonstationary processes. *The annals of Statistics*, 25(1), 1–37.

[3] Nason, G., Von Sachs, R., and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2), 271–292.

[4] Saito, N., and Beylkin, G. (1993). Multiresolution representations using the autocorrelation functions of compactly supported wavelets. *IEEE Transactions on Signal Processing*, 41(12), 3584-3590.

[5] Van Bellegem, S. and and Von Sachs, R. (2008). Locally adaptive estimation of evolutionary wavelet spectra. *Annals of Statistics*, 36(4), 1879-1924.

SUBEXPONENTIAL LOWER BOUNDS FOR $f-$ERGODIC MARKOV PROCESSES

Miha Brešar[1] **and Aleksandar Mijatović**[1]

[1] *University of Warwick*

Abstract We provide a criterion for establishing lower bounds on the rate of convergence in f-variation of a continuous-time ergodic Markov process to its invariant measure. The criterion consists of novel super- and submartingale conditions for certain functionals of the Markov process. It provides a general approach for proving lower bounds on the tails of the invariant measure and the rate of convergence in f-variation of a Markov process, analogous to the widely used Lyapunov drift conditions for upper bounds. Our key technical innovation produces lower bounds on the tails of the heights and durations of the excursions from bounded sets of a continuous-time Markov process using path-wise arguments.

We apply our theory to elliptic diffusions and Lévy-driven stochastic differential equations with known polynomial/stretched exponential upper bounds on their rates of convergence. Our lower bounds match asymptotically the known upper bounds for these classes of models, thus establishing their rate of convergence to stationarity. The generality of the approach suggests that, analogous to the Lyapunov drift conditions for upper bounds, our methods can be expected to find applications in many other settings.

Keywords: subgeometric ergodicity, lower bounds, rate of convergence in f-variation and total variation, invariant measure,

AMS subject classifications: 60J25

Acknowledgements: MB and AM are supported by EPSRC grant EP/V009478/1. AM is also supported by EPSRC grant EP/W006227/1.

References

[1] Brešar, M. and Mijatović, A. (2024). Subexponential lower bounds for f-ergodic Markov processes. *Probability Theory and Related Fields*, 58 pages, to appear. Available at arXiv:2403.14826 [math.PR], https://link.springer.com/content/pdf/10.1007/s00440-024-01306-z.pdf.

Asymptotic Properties of Parameter Estimators for Mixed Fractional Brownian Motion with Trend

Mykyta Yakovliev[1] **and Kostiantyn Ralchenko**[1]

[1] *Taras Shevchenko National University of Kyiv*

Abstract

We consider the following mixed fractional Brownian motion with a linear trend:

$$X_t = \theta t + \sigma W_t + \kappa B_t^H, \quad t \geq 0, \tag{1}$$

where W is a Wiener process, B^H is a fractional Brownian motion with Hurst index H, B^H is independent of W. We aim to estimate unknown parameters $(\theta, H, \sigma, \kappa)$ based on observed $\{X_{kh}, k = 0, 1, 2, \dots\}$, $h > 0$.

Following [1], we introduce the next four statistics:

$$\begin{aligned}
&\phi_N := \frac{X_{Nh}}{N} = \frac{1}{N}\sum_{k=0}^{N-1}\left(X_{(k+1)h} - X_{kh}\right), \quad \xi_N := \frac{1}{N}\sum_{k=0}^{N-1}\left(X_{(k+1)h} - X_{kh}\right)^2, \\
&\eta_N := \frac{1}{N}\sum_{k=0}^{N-1}\left(X_{(k+1)h} - X_{kh}\right)\left(X_{(k+2)h} - X_{(k+1)h}\right), \\
&\zeta_N := \frac{1}{N}\sum_{k=0}^{N-1}\left(X_{(k+2)h} - X_{kh}\right)\left(X_{(k+4)h} - X_{(k+2)h}\right),
\end{aligned} \tag{2}$$

and consider the following estimators for the parameters $(\theta, H, \sigma^2, \kappa^2)$:

$$\begin{aligned}
&\hat{\theta}_N = \frac{\phi_N}{h}, && \widehat{H}_N = \frac{1}{2}\log_{2+}\frac{\zeta_N - 4\phi_N^2}{\eta_N - \phi_N^2}, \\
&\hat{\kappa}_N^2 = \frac{\eta_N - \phi_N^2}{h^{2\widehat{H}_N}(2^{2\widehat{H}_N - 1} - 1)}, && \hat{\sigma}_N^2 = \frac{1}{h}\left(\xi_N - \phi_N^2 - \hat{\kappa}_N^2 h^{2\widehat{H}_N}\right).
\end{aligned} \tag{3}$$

We prove the asymptotic normality of statistics (2) and evaluate their asymptotic covariance matrix. Thereafter, we obtain the main result on asymptotic normality of (3) by applying the delta-method. Additionally, we investigate the behaviour of the estimators numerically, using Monte Carlo simulations.

Keywords: Brownian motion, Wiener process, mixed model, asymptotic distribution.

AMS subject classifications: 60G22, 62F12.

References

[1] A. Kukush, S. Lohvinenko, Y. Mishura, and K. Ralchenko, Two approaches to consistent estimation of parameters of mixed fractional Brownian motion with trend, Statistical Inference for Stochastic Processes 25 (2021), no. 1, 159–187.

Subgeometric ergodicity of regime-switching diffusion processes

Petra Lazić[1] **and Nikola Sandrić**[2]

[1]*Department of Mathematics, University of Zagreb, Zagreb, Croatia*

[2]*Department of Mathematics, University of Zagreb, Zagreb, Croatia*

Abstract

I will discuss subgeometric ergodicity of a class of regime-switching diffusion processes. These are processes that, beside the continuous, diffusive one, have a second, discrete component which changes the behaviour of the process at random times. The theory about them is quite interesting as it shows that in many ways they exhibit different characteristics than classical diffusion processs. In this talk, following the results from [1], I will derive conditions on the drift and diffusion coefficients which result in subgeometric ergodicity of the corresponding semigroup, that is, which allow us to find explicit bounds on the rate of the convergence with respect to two distance functions: the total variation distance and the class of Wasserstein distances.

Keywords: Regime-switching diffusion process, Subgeometric ergodicity, Total variation distance, Wasserstein distance

AMS subject classifications: 60J25, 60J27, 60J60, 60J75

References

[1] Lazic, P. and Sandric, N. (2022). On subgeometric ergodicity of regime-switching diffusion processes *Nonlinear Analysis: Hybrid Systems*, 46: 101262.

Parameter Estimation in First and Second-Order Nonlinear Stochastic Differential Equations using the Strang Splitting Scheme

Predrag Pilipovic[1], **Adeline Samson**[2], **Susanne Ditlevsen**[1]

[1]*Department of Mathematical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark*

[2]*Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*

Abstract

Stochastic differential equations (SDEs) are powerful tools for modeling complex dynamical systems. In this work, we propose a parameter estimation framework based on the Strang splitting scheme for both first and second-order nonlinear SDEs. The Strang splitting scheme is a numerical approximation of the solution to SDE that allows us to construct a pseudo-likelihood function for maximum likelihood estimation (MLE) of the parameters. For first-order SDEs, we develop consistent and efficient estimators, outperforming traditional approaches like Euler-Maruyama discretization or Local linearization in terms of accuracy and speed [1]. In the case of second-order SDEs, we transform them into a system of first-order SDEs by introducing an auxiliary velocity variable. We address the challenges of hypoellipticity and partial observation. If we assume that we observe the auxiliary velocity, the resulting Strang estimator is both consistent and efficient, unlike the Euler-Maruyama-based estimator which does not exist due to hypoellipticity. However, in practical scenarios where the velocity variable is unobserved, we approximate it using finite difference methods. Although this approximation leads to a slight loss of efficiency, resulting in a slightly larger asymptotic variance for the Strang estimator. Our proposed methodology provides a practical and reliable solution for parameter estimation in both first and second-order nonlinear SDEs using the Strang splitting scheme. The integration of pseudo-likelihood offers a consistent MLE approach, enhancing the accuracy and efficiency of parameter inference in SDE models.

Keywords: Strang Splitting Scheme, Stochastic Differential Equations, Parameter Estimation, Hypoellipticity, Partial Observation

AMS subject classifications: 37M15, 60G65, 62F12, 62H12, 62M99

Acknowledgements: This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956107, "Economic Policy in Complex Environments (EPOC)"; Novo Nordisk Foundation NNF20OC0062958; and Independent Research Fund Denmark — Natural Sciences 9040-00215B.

References

[1] Pilipovic, P., Samson, A. and Ditlevsen, S. (2022) *Parameter Estimation in Nonlinear Multivariate Stochastic Differential Equations based on Splitting Schemes*, arXiv preprint arXiv:2211.11884

On the Modelling and Prediction of High-Dimensional Functional Time Series

Jinyuan Chang[1], Qin Fang[2], Xinghao Qiao[2] and Qiwei Yao[2]

[1]*Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China*

[2]*Department of Statistics, London School of Economics, London, WC2A 2AE, UK*

Abstract

We propose a two-step procedure to model and predict high-dimensional functional time series, where the number of function-valued time series p is large in relation to the length of time series n. Our first step performs an eigenanalysis of a positive definite matrix, which leads to a one-to-one linear transformation for the original high-dimensional functional time series, and the transformed curve series can be segmented into several groups such that any two subseries from any two different groups are uncorrelated both contemporaneously and serially. Consequently in our second step those groups are handled separately without the information loss on the overall linear dynamic structure. The second step is devoted to establishing a finite-dimensional dynamical structure for all the transformed functional time series within each group. Furthermore the finite-dimensional structure is represented by that of a vector time series. Modelling and forecasting for the original high-dimensional functional time series are realized via those for the vector time series in all the groups. We investigate the theoretical properties of our proposed methods, and illustrate the finite-sample performance through both extensive simulation and three real datasets.

Keywords: Dimension reduction, Eigenanalysis, Functional thresholding, Hilbert–Schmidt norm, Segmentation transformation.

AMS subject classifications: 62R10, 15A18, 62M10

Ordering Awad-Tsallis Quantile Entropy and Applications to Some Stochastic Models

Răzvan-Cornel Sfetcu[1]

[1] *University of Bucharest, Faculty of Mathematics and Computer Science, Str. Academiei 14, 010014, Bucharest, Romania*

Abstract

We define a stochastic order for Awad-Tsallis residual entropy and prove the preservation of this order in some stochastic models, such as the proportional hazard rate model, proportional reversed hazard rate model and proportional odds model.

Keywords: Awad-Tsallis entropy, proportional hazard rate model, proportional reversed hazard rate model, proportional odds model

AMS subject classifications: 60E15, 60K10, 62N05, 90B25, 94A17

References

[1] Nanda, A.K.; Sankaran, P.G. and Sunoj, S.M. (2014). Rényi's Residual Entropy: A Quantile Approach. *Statist. Probab. Lett.* 85, 114–121.

[2] Navarro, J.; del Aguila, Y. and Asadi, M. (2010). Some New Results on the Cumulative Residual Entropy. *J. Statist. Plann. Inference* 140, 310–322.

[3] Shaked, M. and Shanthikumar, J.G. (2007). *Stochastic Orders*, Springer Science Business Media LLC.

[4] Sunoj, S.M. and Sankaran, P.G. (2012). Quantile Based Entropy Function. *Statist. Probab. Lett.* 82, 1049–1053.

[5] Yan, L. and Kang, D.-t. (2016). Some New Results on the Rényi Quantile Entropy Ordering. *Statistical Methodology* 33, 55–70.

Non-reversible parallel tempering on optimized paths

Saifuddin Syed

Department of Statistics, University of Oxford

Abstract

Markov chain Monte Carlo (MCMC) methods are the most widely used tools in Bayesian statistics for making inferences from complex posterior distributions. For challenging problems where the posterior is high-dimensional with well-separated modes, MCMC algorithms can get trapped exploring local regions of high probability. Parallel tempering (PT) tackles this problem by delegating the task of global exploration to a tractable reference distribution (e.g. prior) which communicates to the target (e.g. posterior) through a sequence of parallel MCMC algorithms targeting distributions of increasing complexity to the target.

The classical approach to designing PT algorithms relied on a reversibility assumption, making PT challenging to tune and even deteriorating performance when introducing too many parallel chains. This talk will introduce a new non-reversible paradigm for PT that dominates its reversible counterpart while avoiding the performance collapse endemic to reversible PT methods. We will then establish near-optimal tuning guidelines, an efficient black-box methodology scalable to GPUs.

Keywords: Monte Carlo Methods, Markov Chain Monte Carlo, Bayesian Inference, Parallel Computing, Computational Statistics

AMS subject classifications: 60J22

Approximate Post-Selection Inference For Variables Selected by Adaptive and Group Lasso

Sarah Pirenne and Gerda Claeskens

ORStat and Leuven Statistics Research Center, KU Leuven

Abstract

Performing model selection on data causes selection uncertainty. This invalidates classical inference which assumes that the model is fixed before the analysis and is correctly specified. Instead, the model is chosen based on the data and is thus random [1]. Selective inference deals with this selection uncertainty by conditioning confidence intervals and p-values on the subspace of the data which leads to the same model selection as the observed data. However, characterizing this selection event is a challenging task. Much literature in selective inference is applicable to model selection events which can be characterized as a polyhedron, i.e. as a set of linear inequalities in the data variables [2]. Examples are forward stepwise regression and regularized regression using the lasso, the fused lasso or the elastic net. For more complicated selection events, which cannot be expressed as polyhedral constraints on the data space, methods for selective inference are lacking. We develop an algorithm for approximate selective inference for parameters after model selection events which do not necessarily have to be characterizable as a polyhedron. We apply this algorithm for conducting inference post-selection by the adaptive lasso and the group lasso. We demonstrate through experiments on simulated data that our algorithm effectively controls the false discovery rate and is computationally efficient. We also illustrate our proposed method to a real-world dataset. See also [3].

Keywords: post-selection inference, selective inference, group lasso, adaptive lasso, parametric programming

AMS subject classifications: 62J07, 62F03

Acknowledgements: G.Claeskens acknowledges the support of the Research Foundation Flanders and KU Leuven Research Fund C1-project C16/20/002.

References

[1] W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. `https://doi.org/10.48550/arXiv.1410.2597`, 2014.

[2] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of statistics*, 44(3):907–927, 2016.

[3] Pirenne, S., & Claeskens, G. (2024). Parametric programming-based approximate selective inference for adaptive lasso, adaptive elastic net and group lasso. Journal of Statistical Computation and Simulation, 94(11):2412–2435.

Using service statistics to inform modern contraceptive prevalence estimates for all women of reproductive age

Shauna Mooney[1] **and Dr Niamh Cahill**[1]

[1]*Department of Mathematics and Statistics, Maynooth University, Ireland*

Abstract

Family Planning 2030 (FP2030) is a global movement committed to improving safe access to reproductive health services worldwide. Monitoring key family planning indicators in FP2030 focus countries is crucial for advancing FP2030's goals. This has been effectively established for women of reproductive age who are married or in a union (MWRA)[1], however progress-tracking for women who are not married or in a union (UWRA) can be associated with higher levels of uncertainty due to various factors, including limited data availability[2]. One of the key family planning indicators used to monitor progress is modern contraceptive prevalence (mCPR). FP2030 recognises access to family planning services as a human right, regardless of marital status. For comprehensive monitoring, FP2030 emphasizes monitoring mCPR for all women (WRA), i.e. the combination of MWRA and UWRA, to highlight family planning needs of all women populations.

The Family Planning Estimation Model (FPEM) combines a Bayesian hierarchical model with country-specific time trends to estimate and project country-level family planning indicators, including mCPR, for MWRA, UWRA, and WRA. However, FPEM relies primarily on survey-based data, which can result in data gaps due to large-scale surveys being conducted on average every 3-5 years. This can lead to estimates and projections that are not data-driven.

Service statistics are a readily available data source, that are routinely collected in conjunction with family-planning service delivery. Service statistics are used to derive Estimated Modern Use (EMU), a family-planning indicator. However, EMUs provide a biased estimate of the level of mCPR. To address this bias, the relationship between changes in EMUs and changes in mCPR can be used to inform mCPR estimates where recent survey data are absent. In the current version of FPEM, EMUs have been used as supplementary data to directly inform estimates of mCPR for MWRA[3]. However, service statistics and hence EMUs are representative of all women, this work therefore aims to enhance FPEM by using EMU data to directly inform mCPR for all women. Such updates can ultimately help to increase confidence in estimates and projections of mCPR for both married women and unmarried women, helping countries to more closely monitor progress towards their family planning goals.

Keywords: Family Planning Estimation Model (FPEM), Bayesian, biased data, modern contraceptive use

AMS subject classifications: 62P07

Acknowledgements: This work has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI 18/CRT/6049.

References

[1] Cahill, N. et al. (2018). Modern contraceptive use, unmet need, and demand satisfied among women of reproductive age who are married or in a union in the focus countries of the Family Planning 2020 initiative: a systematic analysis using the Family Planning Estimation Tool. *The Lancet*, 391, 870–882.

[2] Kantorová, V. et al. (2020). Estimating progress towards meeting women's contraceptive needs in 185 countries: A Bayesian hierarchical modelling study. *PLoS medicine*, 17(2), e1003026.

[3] Cahill, N. et al. (2021). Using family planning service statistics to inform model-based estimates of modern contraceptive prevalence. *PLoS One*, 16, e0258304.

A data-adaptive algorithm for multiple change-point detection

Andreas Anastasiou[1] **and Sophia Loizidou**[2]

[1] *University of Cyprus*

[2] *University of Luxembourg*

Abstract

In this talk, a new data-adaptive method, called DAIS (Data Adaptive ISolation), is introduced for the estimation of the number and the location of change-points in a given data sequence. The proposed method can detect changes in various different signal structures; we focus on the examples of piecewise-constant and continuous, piecewise-linear signals. We highlight, however, that our algorithm can be extended to other frameworks, such as piecewise-quadratic signals. The data-adaptivity of our methodology lies in the fact that, at each step, and for the data under consideration, we search for the most prominent change-point in a targeted neighbourhood of the data sequence that contains this change-point with high probability. Using a suitably chosen contrast function, the change-point will then get detected after being isolated in an interval. The isolation feature enhances estimation accuracy, while the data-adaptive nature of DAIS is advantageous regarding, mainly, computational complexity but also accuracy. The simulation results presented indicate that DAIS is at least as accurate as state-of-the-art competitors.

Keywords: Change-point detection, data-adaptivity, isolation, thresholding criterion

AMS subject classifications: 62G05

Weak Runs in Multistate Trials and Applications in Statistical Process Monitoring

Spiros D. Dafnis[1], Theodoros Perdikis[2] and Markos V. Koutras[1]

[1]*Department of Statistics and Insurance Science, University of Piraeus, Piraeus, Greece*

[2]*Department of Statistics, Athens University of Economics and Business, Athens, Greece*

Abstract

Dafnis and Makri [1] have recently introduced and defined the concept of weak runs in binary trials (see, also, [2]). The authors have studied binomial-type distributions related to such generalized runs and they have highlighted that the new theoretical results can be applied to various fields, such as Agriculture, Finance and Reliability Engineering. Apart from the applications in these fields, it is also highlighted in the literature that weak runs can be profitably applied to statistical process monitoring. More specifically, Dafnis et al. [3] have introduced and studied an improved chi-square control chart based on weak runs in binary trials. Presently, we shall present new classes of Shewhart-type control charts based on an extension of the concept of weak runs under a multistate framework. The study of the control charts' characteristics (run length distribution, average run length) and the numerical experimentation provide useful hints for the implementation of the new control charts by a practitioner and reveal that the proposed charts significantly improve the detection power (as compared to other existing charts) especially for small shifts in the process mean.

Keywords: Waiting time distribution, Markov chain, Shewhart-type control charts, Runs rules, Average run length.

AMS subject classifications: 60J10.

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 2014-2020" in the context of the project "Innovative Titanium Nanoparticles for Development of Autocleaning and Auto antibacterial Application" (MIS 5131364, TiClean).

References

[1] Dafnis, S. D. and Makri, F. S. (2022). Weak Runs in Sequences of Binary Trials. *Metrika* 85(5), 573–603.

[2] Dafnis, S. D. and Makri, F. S. (2023). Distributions Related to Weak Runs With a Minimum and a Maximum Number of Successes: A Unified Approach. *Methodol. Comput. Appl. Probab.* 25(1), 24.

[3] Dafnis, S. D. Perdikis T. and Papadopoulos, G. K. (2023). Improved Chi-Square Control Charts with Weak-Run Rules. *Commun. Stat. - Theory Methods*, https://doi.org/10.1080/03610926.2023.2215357.

A Bayesian Spatio-Temporal Analysis of Markets During the Finnish 1860s Famine

Tiia-Maria Pasanen[1] **and Miikka Voutilainen**[2] **and Jouni Helske**[1] **and Harri Högmander**[1]

[1]*Department of Mathematics and Statistics, University of Jyväskylä, Finland*

[2]*Department of History and Ethnology, University of Jyväskylä, Finland*

Abstract

A Bayesian spatio-temporal model is developed to study pre-industrial grain market integration during the Finnish famine of the 1860s [1]. The data analysed consists of 80 regional time series covering nine years of monthly rye prices in Finland. The model takes into account several problematic features often present when analysing such spatially interdependent time series. For example, compared with the error correction methodology commonly applied in econometrics, this approach allows simultaneous modelling of multiple interdependent time series avoiding cumbersome statistical testing needed to predetermine the often artificial market leader as a point of reference. Furthermore, introducing a flexible spatio-temporal structure enables analysing detailed regional and temporal dynamics of the market mechanisms, for example the asymmetric neighbour dependencies.

Keywords: Bayesian statistics, error correction model, Finnish famine, market integration, spatio-temporal model

AMS subject classifications: 62P20

Acknowledgements: The study was supported by the Finnish Cultural Foundation and the Academy of Finland grants 331817, 311877 and 308975. The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

[1] Pasanen, T.-M., Voutilainen, M., Helske, J., Högmander, H. (2022). Bayesian Spatio-Temporal Analysis of Markets During the Finnish 1860s Famine, *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71 (5), 1282—1302.

Adding unit clauses to the random 2-SAT problem

Tobias L. Overgaard[1]

[1]*Department of Mathematics, Aarhus University*

Abstract

In this talk we will characterize the asymptotics of the random mixed 1- and 2-SAT problem. To elaborate, the Boolean Satisfiability (SAT) Problem refers to the problem of determining whether a given set of m Boolean constraints over n variables can be simultaneously satisfied, i.e. all evaluate to 1 under some interpretation of the variables in $\{0, 1\}$. If we choose the m constraints i.i.d. uniformly at random among the set of disjunctive clauses of length 2, then the problem is known as the random 2-SAT problem, and it is well known that this problem undergoes a structural phase transition; taking $m = \alpha n$ for $\alpha > 0$, the probability of there existing a satisfying assignment tends in the large n limit to 1 if $\alpha < 1$, and to 0 if $\alpha > 1$. At the turn of the millennium it was noted by Dmitris Achlioptas that one can add any polynomial in $\log(n)$ random clauses of length 1 – so-called unit clauses – to the problem without changing these asymptotic probabilities. We investigate how far one can take this procedure of adding unit clauses.

Keywords: Satisfiability, random formulas, unit clauses, phase transition.

AMS subject classifications: 60K35, 68R99.

Acknowledgements: This talk presents joint work with Andreas Basse-O'Connor and Mette Skjøtt.

Pattern Recovery by SLOPE

Małgorzata Bogdan[3], Xavier Dupuis[5], Piotr Graczyk[2], Bartosz Kołodziejek[4], Tomasz Skalski[1,2], Patrick Tardivel[5], Maciej Wilczyński[1]

[1] *Wrocław University of Science and Technology*

[2] *University of Angers, France*

[3] *University of Wrocław*

[4] *Warsaw University of Technology*

[5] *University of Burgundy, France*

Abstract

Sorted L-One Penalized Estimator (SLOPE), a generalization of the LASSO estimator, was introduced by Bogdan, van den Berg, Sabatti, Su and Candès in 2015. It is a convex regularization method for fitting high-dimensional regression models. While LASSO can eliminate redundant predictors by setting the corresponding regression coefficients to zero, SLOPE can also identify clusters of variables with the same absolute values of regression coefficients.

In this talk I will discuss sufficient and necessary conditions for the proper identification of the SLOPE pattern, i.e. of the proper sign and of the proper ranking of the absolute values of individual regression coefficients, including a proper clustering. I will also mention the strong consistency of pattern recovery by SLOPE in an asymptotic case when the number of columns in the design matrix is fixed, but the sample size diverges to infinity.

Keywords: linear regression, SLOPE, pattern recovery, irrepresentability condition

AMS subject classifications: 62J05

Acknowledgements: This research was supported by a French Government Scholarship.

References

[1] M. Bogdan, X. Dupuis, P. Graczyk, B. Kołodziejek, T. Skalski, P. Tardivel, M. Wilczyński. Pattern Recovery by SLOPE. ArXiv 2203.12086.

[2] P. Graczyk, U. Schneider, T. Skalski, P. Tardivel. Pattern Recovery in Penalized and Thresholded Estimation and its Geometry. 2023. hal-03262087.

[3] T. Skalski, P. Graczyk, B. Kołodziejek, M. Wilczyński. Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal. Probability and Mathematical Statistics 42(2):283-302, 2022

CLT for the perimeter of the convex hull spanned by two independent random walks

Tomislav Kralj[1] **and Daniela Ivanković**[1]

[1] *University of Zagreb, Faculty of Science, Department of Mathematics*

Abstract

In this research, we limit ourselves to $\mathbb{R}^2$ and consider two independent random walks with linearly independent drift vectors. If we denote the perimeter of the obtained convex hull up to time n with L_n, we show that $1/\sqrt{n}\,(L_n - \mathbb{E}\,(L_n))$ converges in distribution to $N\left(0, \sigma^2\right)$, where the constant $\sigma^2 > 0$ can be explicitly expressed. Ultimately, the idea is to expand this concept to more walks in multiple dimensions and for other intrinsic volumes. This research extends the argument made in the papers [2], [3], and [4].

Keywords: Central Limit Theorem, Random Walks, Convex Hulls, Stochastic Geometry

AMS subject classifications: 60D05

Acknowledgements: I would like to express my profound gratitude to my PhD advisors, Professors Nikola Sandrić and Stjepan Šebek. Their enduring patience and invaluable assistance were the pillars that supported this endeavor. Moreover, this work serves as a generalization of the problem initially examined by Professor Andrew Wade in collaboration with his doctoral students. I am deeply indebted to him for his insightful advice and constructive feedback, which significantly enhanced the quality of this work.

References

[1] Durrett, R. (2019). *Probability: Theory and Examples*. Cambridge University Press, 49.

[2] McRedmond, J. and Wade, A. (2018). The convex hull of a planar random walk: perimeter, diameter, and shape.

[3] Wade, A. and Xu, C. (2015). Convex hulls of planar random walks with drift. *Proceedings of the American Mathematical Society*, 143(1), 433–445.

[4] Wade, A. and Xu, C. (2015). Convex hulls of random walks and their scaling limits. *Stochastic Processes and their Applications*, 125(11), 4300–4320. Elsevier.

Fast Adaptation with Bradley-Terry Preference Models in Text-To-Image Retrieval and Generation

Víctor Gallego[1]

[1]*Komorebi AI Technologies*

Abstract

Recently, large multimodal models, such as CLIP [2] and Stable Diffusion [3] have experimented tremendous successes in both foundations and applications. However, as these models increase in parameter size and computational requirements, it becomes more challenging for users to personalize them for specific tasks or preferences. In this work, we address the problem of adapting the previous models towards sets of particular human preferences, aligning the retrieved or generated images with the preferences of the user. We leverage the Bradley-Terry preference model [1] to develop a fast adaptation method that efficiently fine-tunes the original model, with few examples and with minimal computing resources. Extensive evidence of the capabilities of this framework is provided through experiments in different domains related to multimodal text and image understanding, including preference prediction as a reward model, image retrieval, and generation tasks.

Keywords: preference models, multimodal, text-to-image, retrieval, image generation

AMS subject classifications: 62H30, 62H35

References

[1] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022.

Interpolation property of shallow neural networks

Vlad Raul Constantinescu[1,2] **and Ionel Popescu**[1,3]

[1] *University of Bucharest, Faculty of Mathematics and Computer Science, 14 Academiei str., 70109, Bucharest, Romania*

[2] *"Gheorghe Mihoc – Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, 13 Calea 13 Septembrie, 050711 Bucharest, Romania*

[3] *Simion Stoilow Institute of Mathematics of the Romanian Academy, P.O. Box 1-764, RO-014700 Bucharest, Romania*

Abstract

In this work, we prove that in the overparametrized regime, a shallow neural network can interpolate any data set, i.e. the loss function has a global minimum value equal to zero as long as the activation function is not a polynomial of small degree. Additionally, if such a global minimum exists, then the locus of global minima is a manifold. Furthermore, we give a characterization of the Hessian of the loss function evaluated at the global minima, and we provide a practical probabilistic method of finding the interpolation point.

Keywords: neural networks, interpolation, optimization, global minima

AMS subject classifications: 60

Acknowledgements: I am grateful to my Ph.D supervisor, Ionel Popescu, for proposing this research topic for my doctoral research and for his guidance through this work. Also, this work has been supported from UEFISCDI PN-III-P4-ID-PCE-2020-2498.

References

[1] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal (2019). *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proceedings of the National Academy of Sciences, 116(32):15849– 15854

[2] Sebastien Bubeck (2015). *Convex optimization: Algorithms and complexity*, Foundations and Trends in Machine Learning, 8(3-4):231–357

[3] Yaim Cooper (2021). *Global minima of overparameterized neural networks*, SIAM Journal on Mathematics of Data Science, 3(2):676–691

[4] William F Donoghue (2014). *Distributions and Fourier transforms*, Academic Press

[5] Chaoyue Liu, Libin Zhu, and Mikhail Belkin (2022). *Loss landscapes and optimization in over-parameterized non- linear systems and neural networks*, Applied and Computational Harmonic Analysis, 59:85–116

[6] Samet Oymak and Mahdi Soltanolkotabi (2020). *Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks*, IEEE Journal on Selected Areas in Information Theory, 1(1):84– 105

[7] Allan Pinkus (1999). *Approximation theory of the mlp model in neural networks*, Acta numerica, 8:143–195

Extremal Dependence of Stochastic Processes Driven by Exponential-Tailed Lévy Noise

Zhongwei Zhang[1], David Bolin[2], Sebastian Engelke[1], and Raphaël Huser[2]

[1] *Research Center for Statistics, University of Geneva, 1205 Geneva, Switzerland*

[2] *Statistics Program, CEMSE Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia*

Abstract

Stochastic processes driven by exponential-tailed Lévy noise constitute important extensions of their Gaussian counterparts in order to capture deviations from Gaussianity, more flexible dependence structures, and sample paths with jumps. Popular examples include non-Gaussian Ornstein–Uhlenbeck (OU) processes and type G Matérn stochastic partial differential equation (SPDE) random fields. This paper is concerned with the open problem of determining the extremal dependence structure induced by these processes. Both process types admit approximations on grids or triangulations that are used in practice for efficient simulations and inference. We first show that these approximations can be expressed as special cases of a class of linear transformations of independent, exponential-tailed random variables, that bridge asymptotic dependence and independence in a novel, tractable way. This result is of independent interest since models that can capture both extremal dependence regimes are scarce and the construction of such flexible models is an active area of research. Based on this fundamental result, we show that the exponential-tailed non-Gaussian OU process is asymptotically independent, but with a different residual tail dependence function than its Gaussian counterpart. Furthermore, we show that the integral approximation of the type G Matérn SPDE fields is asymptotically independent, as well, provided that the mesh is fine enough, and for certain smoothness parameters, when the mesh size tends to zero, the limiting residual tail dependence function of the approximation model is derived. Our results are illustrated through simulation studies.

Keywords: exponential tails, extremal dependence, non-Gaussian OU process, type G Matérn SPDE fields

AMS subject classifications: 60G70, 60G10, 62H20

On a Novel Two-Sample Test for Matrix Distributions and Application in Finance

Žikica Lukić[1] **and Bojana Milošević**[2]

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

[2]*Faculty of Mathematics University of Belgrade, Serbia*

Abstract

The complexity of data gathered in recent years has increased the popularity of matrix methods. The primary focus of this presentation will be to discuss recent developments in testing statistical hypotheses in the field of matrix distributions. Specifically, we will explore a newly proposed test for evaluating the equality of two symmetric positive definite matrix distributions. This test is constructed by integrating the squared difference of two empirical Hankel transforms using the Wishart measure.

Particular attention will be given to the application of the aforementioned test in finance, specifically in relation to major cryptocurrencies like Bitcoin and Ethereum, as well as the analysis of Bitcoin-related events [1]. Moreover, recent Bitcoin-related events will also be considered. Lastly, the presentation will include the application of novel tests to datasets from major S&P 500 companies [2].

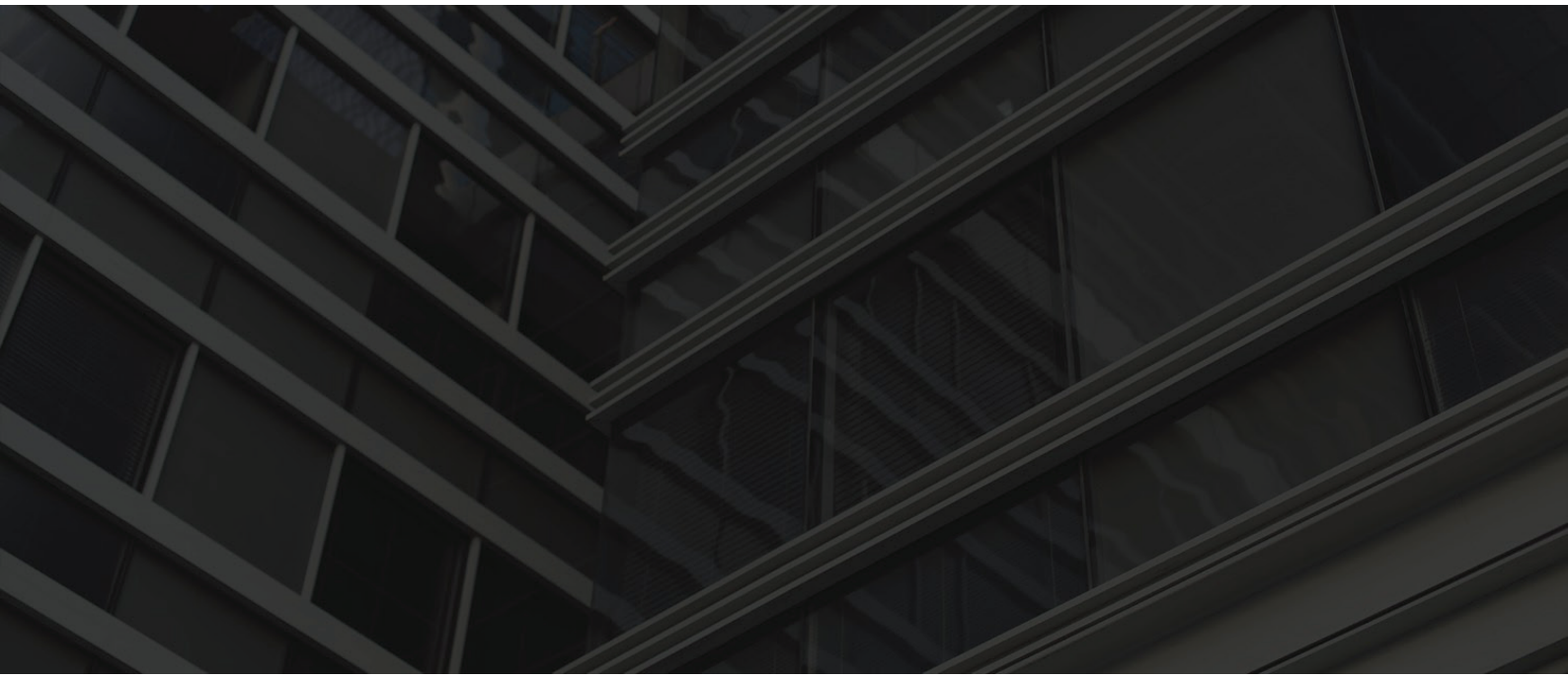
Keywords: Hankel transform, Wishart distribution, stability of cryptomarkets, stock data

AMS subject classifications: Primary: 62H15; Secondary: 62P05, 62E20

References

[1] Fruehwirt, W., Hochfilzer, L., Weydemann, L. and Roberts, S. (2021). Cumulation, crash, coherency: A cryptocurrency bubble wavelet analysis. *Finance Research Letters*, 40, 101668.

[2] Péli, T. (2023). Assessing the Financial Performance of the Companies that Shape the S&P 500 Index. *Acta Polytechnica Hungarica*, 20(3).



Author index