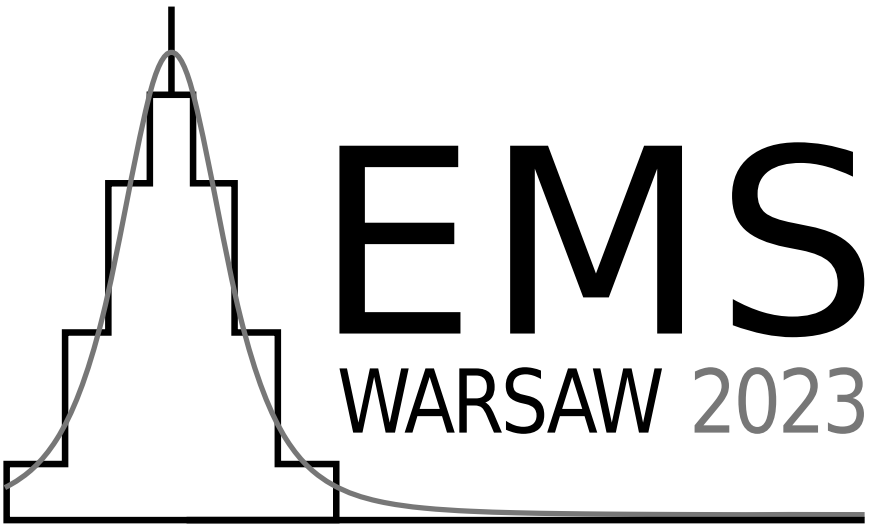
European Meeting of Statisticians 2023

Warsaw 3–7 July, 2023

Book of Abstracts



The European Meeting of Statisticians (EMS), sponsored by the European Regional Committee of the Bernoulli Society, is the main conference in statistics and probability in Europe. EMS is a conference where statisticians from all regions meet to exchange ideas and talk about the newest developments in the broad field of statistics and probability theory. The very first EMS meeting was held in Dublin in 1962, and the 34th European Meeting of Statisticians (EMS 2023) – will take place in Warsaw.

The meeting is organized by the University of Warsaw, the Warsaw University of Technology, and the Polish Mathematical Society under the auspices of the Committee of Mathematics of the Polish Academy of Sciences. The conference will be held at the campus of the University of Warsaw and will start on Monday 3.7.2023 and it will end on Friday 7.7.2023. The program consists of invited and contributed lectures, and posters, addressing a full range of subjects in statistics and its various applications.

EXCELLENCE INITIATIVE
RESEARCH
UNIVERSITY

UNIVERSITY OF WARSAW
MIM
FACULTY
OF MATHEMATICS, INFORMATICS
AND MECHANICS

rok zał. 1919
ptm

Faculty of Mathematics
and Information Science
WARSAW UNIVERSITY OF TECHNOLOGY

Strategic Sponsor:

RESURGO
MUTATA
EADEM
Bernoulli Society
for Mathematical Statistics
and Probability

PZU

Program Committee

Judith Rousseau (chair)
Denis Belometsny
Małgorzata Bogdan
Arnoldo Frigessi
Eric Moulines
Sonia Petrone
Stijn Vansteelandt

Organizing Committee

Przemysław Grzegorzewski
Krystyna Jaworska
Katarzyna Kmita
Anna Kozak
Bartosz Kołodziejek
Adrian Łydka
Błażej Miasojedow
Jan Mielniczuk
Jacek Miękisz
Agnieszka Piliszek
Łukasz Rajkowski
Wojciech Rejchel
Krzysztof Rudaś
Paweł Teisseyre

Keynotes

OPENING LECTURE – **Angelika Rohde** (University of Freiburg)
Angelika Rohde received her Ph.D. at Universität Heidelberg and Universität Bern for the dissertation on New Multiscale Approaches to Nonparametric Statistical Inference. Her professional career includes a postdoctoral fellowship at CREST/HEC/Ecole Polytechnique, Paris. Next, she was a junior professor of Stochastics at Universität Hamburg, associate professor for Mathematical Statistics at Ruhr-Universität Bochum and since 2016 she is a full professor of Mathematical Stochastics at Albert-Ludwigs-Universität Freiburg. Her scientific interests concern mainly mathematical statistics (adaptive inference under constraints, nonparametric statistics of stochastic processes, transfer learning) and probability theory (esp. stochastic analysis).

FORUM LECTURE – **Gérard Biau** (Sorbonne University, Paris)
Gérard Biau is a full professor at the Probability, Statistics, and Modeling Laboratory (LPSM) of Sorbonne University, Paris. His research is mainly focused on developing new methodologies and rigorous mathematical theory in statistical learning and artificial intelligence while trying to find connections between statistics and algorithms. He was a member of the Institut Universitaire de France from 2012 to 2017 and served from 2015 to 2018 as the president of the French Statistical Society. In 2018, the French Academy of Sciences awarded him the Michel Monpetit-Inria Prize. He is currently the director of the Sorbonne Center for Artificial Intelligence (SCAI).

EUROPEAN MATHEMATICAL SOCIETY COOPERATION LECTURE – **Arnaud Doucet** (University of Oxford)
Arnaud Doucet obtained his Ph.D. from the University of Paris-XI in 1997. Ever since he has held faculty positions at the University of Cambridge, Melbourne University, the University of British Columbia, and the Institute of Statistical Mathematics (Tokyo). He joined Oxford University in 2011 where he is currently a Statutory Professor (Oxford speak for chair) in the Department of Statistics. Since 2019, he is also a Senior Research Scientist at Google DeepMind. He was an IMS Medallion lecturer in 2016, was elected an IMS Fellow in 2017, and was awarded the Guy Silver medal of the Royal Statistical Society in 2020.

SPECIAL INVITED LECTURE – **Po-Ling Loh** (University of Cambridge)
She grew up in Madison, Wisconsin. After graduating from the California Institute of Technology with a BS in math and minor in English, she moved to UC Berkeley, where she subsequently earned an MS in computer science and Ph.D. in statistics. From 2014-2016, she was an assistant professor in the Department of Statistics at the Wharton School of the University of Pennsylvania. From 2016-2020, she was an assistant and then associate professor in the Department of Electrical and Computer Engineering and the Department of Statistics at the University of Wisconsin-Madison. She was a visiting associate professor in the Department of Statistics at Columbia University from 2019-2020. In 2021, she began my appointment at the University of Cambridge.

SPECIAL INVITED LECTURE – **Jonas Peters** (ETH Zurich)
Jonas is a professor of statistics at the Department of Mathematics at ETH Zurich. Previously, he has held positions at the Department of Mathematical Sciences at the University of Copenhagen and the Max-Planck-Institute for Intelligent Systems in Tuebingen. He studied Mathematics at the University of Heidelberg and the University of Cambridge and obtained his Ph.D. jointly from MPI and ETH. He is interested in inferring causal relationships from different types of data and in building statistical methods that are robust with respect to distributional shifts. In his research, Jonas seeks to combine theory, methodology, and applications. His work relates to areas such as computational statistics, causal inference, graphical models, and testing independence.

CLOSING LECTURE – **Marta Blangiardo** (Imperial College, London)
Marta Blangiardo is a professor of Biostatistics in the Department of Epidemiology and Biostatistics at Imperial College London and leads the Biostatistics and Data Science theme of the MRC Centre for Environment and Health. She is one of the PIs of the Turing-RSS Health Data lab and has an honorary academic contract with the UK Health Security Agency. Her main interests are related to the methodological aspects of environmental exposure estimation and spatial and spatiotemporal models for disease mapping and for risk assessment.

Invited sessions

IS01. Recent advances in change point analysis
Organizer: Haeran Cho (University of Bristol)
Speakers:
Andreas Anastasiou (University of Cyprus),
Florian Pein (Lancaster University),
Lorenzo Trapani (University of Leicester)

IS02. Network inference and causality
Organizer: Veronica Vinciotti (University of Trento), Statistical Network Science Committee
Speakers:
Sara Magliacane (University of Amsterdam, MIT-IBM Watson AI Lab),
Francesco Stingo (University of Florence),
Ernst Wit (Universita della Svizzera italiana)

IS03. Theory for Bayesian nonparametrics
Organizer: Stéphanie van der Pas (Amsterdam UMC)
Speakers:
Natalia Bochkina (University of Edinburgh),
Sonia Petrone (Bocconi University, Milan),
Lasse Vuursteen (TU Delft)

IS04. Spatial statistics: testing issues
Organizer: Aila Särkkä (Chalmers University of Technology and University of Gothenburg)
Speakers:
Chiara Fend (RPTU Kaiserslautern-Landau, Germany),
Tomáš Mrkvička (University of South Bohemia),
Jonas Wallin (Lund University)

IS05. Computational statistics
Organizer: Anthony Lee (University of Bristol)
Speakers:
Yuansi Chen (Duke University),
Anthony Lee (University of Bristol),
Sven Wang (Massachusetts Institute of Technology)

IS06. Efficient and robust estimation in causal inference
Organizer: Alex Luedtke (University of Washington)
Speakers:
Antoine Chambaz (Université Paris Cité),
Nathan Kallus (Cornell University),
Alex Luedtke (University of Washington)

IS07. Nonparametric statistics for stochastic processes
Organizer: Mathias Trabs (Karlsruhe Institute of Technology)
Speakers:
Randolf Altmeyer (University of Cambridge),
Markus Bibinger (University of Würzburg),
Mariucci Ester (Paris-Saclay (UVSQ))

IS08. Fairness
Organizer: Nicolas Schreuder (Università di Genova)
Speakers:
Evgenii Chzhen (CRNS, LMO, Université Paris-Saclay),
Patrick Loiseau (Inria),
Nicolas Schreuder (Università di Genova, MaLGa)

IS09. Inverse problems
Organizer: Alexander Goldenshluger (University of Haifa)
Speakers:
Alexander Goldenshluger (University of Haifa, Israel),
Anatoli Juditsky (LJK, University Grenoble-Alpes),
Mathias Trabs (Karlsruhe Institute of Technology)

IS10. Complex network methods for complex real data
Organizer: Clelia di Serio (Università della Svizzera Italiana, Lugano and Università Vita Salute San Raffaele, Milano)
Speakers:
Spyros Balafas (Università Vita-Salute San Raffaele),
Lorenzo Giammei (University of Milano-Bicocca),
Rūta Juozaitienė (Vytautas Magnus University)

IS11. Robustness sensitivity and uncertainty of statistical estimation
Organizer: Fanny Yang (ETH Zurich)
Speakers:
Carlos Cinelli (University of Washington),
Nicola Gnecco (University of Copenhagen),
Yingzhen Li (Imperial College London)

IS12. Advances in functional depth measures
Organizer: Laura Maria Sangalli (Politecnico di Milano)
Speakers:
Sara Lopez-Pintado (Northeastern University),
Pavlo Mozharovskyi (LTCI, Telecom Paris, Institut Polytechnique de Paris),
Stanislav Nagy (Charles University)

IS13. Recent advances in high-dimensional inference
Organizer: Eduard Belitser (VU Amsterdam)
Speakers:
Eduard Belitser (VU),
Subhashis Ghoshal (North Carolina State University),
Deborah Sulem (Barcelona School of Economics-Universitat Pompeu Fabra)

IS14. Federated Learning
Organizer: Kerrie Mengersen (Queensland University of Technology)
Speakers:
Conor Hassan (Queensland University of Technology),
Samuel Kaski (Aalto University),
Marco Lorenzi (Inria Sophia Antipolis and Université Côte d'Azur)

IS15. New developments in learning algorithms
Organizers: Alain Durmus (Ecole Polytechnique) and Błażej Miasojedow (University of Warsaw)
Speakers:
Gersende Fort (CNRS & Institut de Mathématiques de Toulouse),
Aurelien Lucchi (University of Basel),
Sean Meyn (University of Florida)

IS16. Centennials of important events in the Probability Theory
Organizer: Adam Jakubowski (Nicolaus Copernicus University in Torun), President of the Bernoulli Society
Speakers:
Adam Jakubowski (Nicolaus Copernicus University),
Rene Schilling (TU Dresden),
Aleksander Weron (Wroclaw University of Science and Tehnology)

IS17. Statistical inference for stochastic processes
Organizer: Mark Podolskij (Aarhus University)
Speakers:
Denis Belomestny (Duisburg-Essen University),
Mark Podolskij (University of Luxembourg),
Claudia Strauch (Aarhus University)

IS18. Latent variable models
Organizer: Elisabeth Gassiat (Université Paris-Saclay)
Speakers:
Aurore Delaigle (University of Melbourne),
Elisabeth Gassiat (Université Paris-Saclay),
Luc Lehéricy (Université Côte d'Azur, CNRS)

Contents

Kumaraswamy-Normal Control Chart for Process Monitoring

Wednesday, 5.07
12:40–13:00
CS 16
Room 2180

Saheed A. Afolabi

Department of Mathematics,
College of Computing and Mathematics,
King Fahd University of Petroleum and Minerals,
Kingdom of Saudi Arabia
afolabisa63@gmail.com

Abstract

Over the years, different probability distributions have been in existence for modelling in different areas of research. One such family of distribution is named Kumaraswamy (Kw) distributions. Many researchers have worked on the combination of Kw and Beta (B) distributions using different approaches based on existing literatures. This study presents new control limits based on Kumaraswamy-Normal distribution from Kumaraswamy-Generalized (Kw-G) family of distributions that are more flexible in controlling the skewness of data. Moreover, an algorithm is developed in R software for generating the random numbers from Kumaraswamy Normal Distribution which was used to evaluate and compare the performance of some basic control charts under Kw-Normal environment. The performance of the proposed chart is evaluated through an extensive Monte Carlo simulation study. The numerical results demonstrated that the proposed control chart outperforms the Kumaraswamy control chart in terms of run length analysis. Finally, applying Kw-N control chart in modeling real life data showing the practical applicability.

Keywords

Kumaraswamy distribution, Kumaraswamy-Normal, Control chart, Statistical process control.

Bibliography

[1] Abbas N., (2012). Memory-type Control Charts in Statistical Process Control. Research Gate.

[2] Cordeiro, G. M. and de Castro, M. J., (2011). A New Family of Generalized Distributions. *Journal of Statistical Computation and Simulation.* Volume 00, No. 00, August 2009, 1–17. 81 (2011), 883–893.

[3] de Araujo Lima-Filho, L. M., & Mariano Bayer, F. (2021). Kumaraswamy control chart for monitoring double bounded environmental data. *Communications in Statistics: Simulation and Computation*, 50(9), 2513–2528. https://doi.org/10.1080/03610918.2019.1635159.

Causal effect identification under preferential selection: extended graphical criteria for regression adjustment via mediation analysis

Tuesday, 4.07
16:00–18:00
Poster

Johan de Aguas[1,2], **Johan Pensar**[2] **and Guido Biele**[2]

[1] Department of Mathematics, University of Oslo;
{johanmd,johanpen}@math.uio.no
[2] Department of Child Health and Development, Norwegian Institute of Public Health;
guido.biele@fhi.no

Abstract

Selection bias is driven by the preferential exclusion of units from the data. It impedes valid statistical and causal inference, as collected samples are not entirely representative of the target population. Besides, it induces extraneous flows of non-causal association between treatment and outcome under analysis [1]. Generalizations to the *back-door criterion* have been developed to recover the *interventional distribution* from observational data under confounding and selection bias, by employing *adjustment sets* containing no *causal node* (mediators and their descendants) [2]. We extend available graphical criteria to recover the conditional average treatment effect (CATE) when a causal node is required to *d*-separate the outcome from the selection mechanism. By leveraging external unbiased data, we define a *blocking set* as a set of causal nodes such that: *i)* the direct and indirect effects flowing through the blocking mechanism are recoverable, and *ii)* such effects correspond to a counterfactual decomposition of the targeted CATE. When identifiability criteria are met, the proposed approach shows superior performance in simulations against alternative methods based on inverse probability weighting (IPW). Under partial identification, it still obtains a better bias-variance trade-off than IPW. We apply it to real-world data, to estimate the effects of ADHD medication on children's school performance in Norway.

Bibliography

[1] Bareinboim E., Pearl J. *Controlling Selection Bias in Causal Inference*. In N. D. Lawrence and M. Girolami, editors, Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, volume 22 of Proceedings of Machine Learning Research, pages 100-108, La Palma, Canary Islands, 21-23 Apr 2012. PMLR. URL https://proceedings.mlr.press/v22/bareinboim12.html.

[2] Joe H., *Generalized adjustment under confounding and selection biases*. Proceedings of the AAAI Conference on Ar- tificial Intelligence, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.12125. URL https://ojs.aaai.org/index.php/AAAI/article/view/12125

Automatic model training under restrictive time constraints

Monday, 3.07
15:45–16:05
CS 01
Room 1.01

Lukas Cironis[1], **Jan Palczewski**[2], **Georgios Aivaliotis**[2]

[1] 4-Extra Ltd (United Kingdom)
[2] School of Mathematics, University of Leeds (United Kingdom)

Abstract

We develop a hyperparameter optimisation algorithm, Automated Budget Constrained Training, which balances the quality of a model with the computational cost required to tune it. The relationship between hyperparameters, model quality and computational cost must be learnt and this learning is incorporated directly into the optimisation problem. At each training epoch, the algorithm decides whether to terminate or continue training, and, in the latter case, what values of hyperparameters to use. This decision weighs optimally potential improvements in the quality with the additional training time and the uncertainty about the learnt quantities. The performance of our algorithm is verified on a number of machine learning problems encompassing random forests and neural networks. Our approach is rooted in the theory of Markov decision processes with partial information and we develop a numerical method to compute the value function and an optimal strategy.

Optimal parameter estimation for linear SPDEs

Tuesday, 4.07
11:00–11:30
IS 07
Room 0.06

Randolf Altmeyer[1], **Anton Tiepner**[2], **Martin Wahl**[3]

[1] University of Cambridge (United Kingdom)
[2] Aarhus University (Denmark)
[3] Bielefeld University (Germany)

Abstract

The problem of parameter estimation in a general second order linear stochastic partial differential equation (SPDE) is considered. One trajectory of the solution to the SPDE is observed continuously in time and averaged in space over a small window at multiple locations.

Extending on results in [1], estimators for the diffusivity, transport and reaction coefficients are constructed and analysed. These estimators are shown to be minimax rate optimal by proving an explicit lower bound in the asymptotic regime where the spatial window shrinks to zero and with a growing number of observations. The rate of convergence depends on the differential order in which the respective coefficient appears, with the fastest rate achieved for the diffusivity and the slowest rate for reaction terms.

The proof for the lower bound relies on an explicit analysis of the reproducing kernel Hilbert space for the observed Gaussian process and the underlying SPDE, which we derive here, and which may be of independent interest.

Bibliography

[1] Altmeyer R., Reiß M. *Nonparametric estimation for linear SPDEs from local measurements*, Annals of Applied Probability, 31 (1) 1-38, 2021.

[2] Altmeyer R., Tiepner A., Wahl M., *Optimal parameter estimation for linear SPDEs from multiple measurements*, arXiv preprint arXiv:2211.02496.

Generalized multiple change-point detection in the structure of multivariate, possibly high-dimensional, data sequences

Monday, 3.07
15:45–16:15
IS 01
Room 0.03

Andreas Anastasiou, Aggelos Papanastasiou

University of Cyprus

Abstract

The extensive emergence of big data techniques has led to an increasing interest in the development of change-point detection algorithms that can perform well in a multivariate, possibly high-dimensional setting. In this talk, we propose a new method for the consistent estimation of the number and location of multiple generalized change-points in multivariate, possibly high-dimensional, noisy data sequences. The number of change-points is allowed to increase with the sample size and the dimensionality of the given data sequence. Having a number of univariate signals, which constitute the unknown multivariate signal, our algorithm can deal with general structural changes; we focus on changes in the mean vector of a multivariate piecewise-constant signal, as well as changes in the linear trend of any of the univariate component signals. Our proposed algorithm, labelled Multivariate Isolate-Detect (MID) allows for consistent change-point detection in the presence of frequent changes of possibly small magnitudes in a computationally fast way.

Nonparametric regression with clustered observations

Tuesday, 4.07
11:00–11:20
CS 08
Room 5440

Stanislav Anatolyev

CERGE-EI (Czech Republic)

Abstract

We consider a nonparametric mean regression for clustered samples, where observations are independent across clusters, but may exhibit within-cluster dependence and be accompanied with conditional heteroskedasticity. The clusters may have different size, and their average size may be small, moderately large, or seriously large. We focus on the Nadaraya-Watson (NW) mean regression estimator, and derive its asymptotic distribution under corresponding scenarios of a fixed or slowly growing, and rapidly growing average cluster size, which interplays with the rate of bandwidth shrinkage as the total sample size increases, and yields different rates of convergence. We also discuss optimal bandwidth selection; the rule turns out to be uniform across the cases. The form of asymptotic variance of the NW estimator depends on the growth rate of the average cluster size via the dominance or balance between within-cluster error variances and error covariances. We propose a number of asymptotic variance estimates, suitable in different situations, and prove their consistency. One of these estimators is robust to the growth rate of the average cluster size, and leads to robust confidence intervals. Finally, we provide an illustration of the developed inferrential tools in a nonparametric regression of log wage on age in a one-year slice of the CPS dataset.

Sample Splitting techniques for high dimensional Multivariate Survival data

Tuesday, 4.07
15:00–15:20
CS 11
Room 2180

Beryl Ang'iro[1,2], **Roel Braekers**[1]

[1] Hasselt University (Belgium)
[2] Karatina University(Kenya)

Abstract

Large data sets come from either a large number of repeated measurements per outcome for some clusters or because of hierarchies in the data with large numbers of within-unit replication. An extraordinarily large sample size may pose computational challenges to the data analyst making conventional inferential methods, such as maximum likelihood, prohibitive. Furthermore, accounting for correlation in clustered data remains a challenge especially when the total number of clusters are large and/or there are large number of replications per cluster. Clustered multivariate failure times are frequently analysed using frailty and copula models. Here, the different response event times are grouped into clusters hence the dependence between the different event times in a cluster is of interest. In this talk, we introduce a copula model and propose a sample splitting method to analyze large-scale multivariate failure time data.

This study is motivated by the Kenya Demographic health Survey (KDHS, 2014) data in which there are a large number of households (14,000) with 1 to 6 children per household. Due to this, the optimization in both the frailty and copula models tend to be numerically unstable. As a solution to this problem we use sample splitting methods based on the work of [1] and [2]. The sample splitting methodology refers to a three step process: first randomly splitting the data into sub-samples: then analyse each sub-sample separately: and finally combining the results of these analyses using appropriate weighted averages. The combined results provides a statistical inference which is similar to the one from analyzing the entire data set all at once.

Keywords

High dimensional data, Multivariate Survival data, Sample Splitting, Copula model.

Bibliography

[1] Molenberghs., *Pseudo-likelihood methodology for partitioned large and complex samples*, Statistics & probability letters, 81, no. 7 (2011): 892–901.

[2] Wang, Wei, Shou-En Lu, Jerry Q. Cheng, Minge Xie, and John B. Kostis, *Multivariate survival analysis in big data: A divide-and-combine approach*, Biometrics 78, no. 3 (2022): 852–866.

Assumption-Lean Quantile Regression

Thursday, 6.07
16:40–17:00
CS 27
Room 1.40

Georgi Baklicharov[1], **Stijn Vansteelandt**[1], **Christophe Ley**[2]

[1] Ghent University (Belgium)
[2] University of Luxembourg (Luxembourg)

Abstract

Quantile regression is a powerful tool for detecting varying associations across different parts of the dependent variable's distribution. However, when using quantile regression to parameterize the conditional association between an exposure and an outcome, given covariates, two potential issues are often ignored. Firstly, the exposure coefficient estimator may not converge to a meaningful quantity when the model is misspecified, and secondly, variable selection methods may induce excess uncertainty, rendering inferences overly optimistic. In this paper, we address these issues by introducing a nonparametric main effect estimand that still captures the (conditional) association of interest, even when the quantile model is misspecified. This estimand is estimated using the efficient influence function under the nonparametric model, allowing for the incorporation of data-adaptive procedures such as variable selection and machine learning. Our approach provides a flexible and reliable method for detecting associations that are robust to model misspecification and excess uncertainty induced by variable selection methods.

Comparing Communication Strategies with Networks Estimated From Hierarchical EEG Data

Thursday, 6.07
11:00–11:30
IS 10
Room 0.06

Spyros Balafas[1], **Clelia Di Serio**[4], **Riccardo Lolatto**[5], **Marco Mandolfo**[5], **Anna Maria Bianchi**[5], **Ernst Wit**[3], **Chiara Brombin**[1]

[1] University Center for Statistics in the Biomedical Sciences (CUSSB), Universita Vita–Salute San Raffaele, Milan
[2] Department of Electronics, Information Science and Bioengineering, Politecnico di Milano, Milan
[3] Institute of Computing (CI), UniversitÃ della Svizzera italiana, Lugano
[4] Faculty of Biomedical Sciences, UniversitÃ della Svizzera italiana, Lugano
[5] Department of Management, Economics, and Industrial Engineering, Politecnico di Milano, Milan

Abstract

Developing a methodological framework and advanced statistical tools to measure the effectiveness of fundraising campaigns is fundamental to enhance communication strategies. Nowadays this is particular crucial for healthcare campaigns aimed at raising public awareness towards extremely sensitive aspects of health that suffer from lack of visibility while needing financial support for advancing research. Healthcare fundraising involves many challenges including unique setting and relationships between potential donors, patients and healthcare providers. Thus, finding evidence for measuring the emotional impact of these campaigns is a major challenge.

In this study, we focus on brain activation as a proxy measure of video effectiveness and we compare within a multivariate setting physiological responses induced by four fundraising campaigns realized using different communication strategies. We model electroencephalographic (EEG) signals using methodological tools belonging to graphical models to estimate partial correlation networks corresponding to different advertisements. The structure of estimated EEG networks is then compared in post-hoc analysis using resampling procedures.

Results show statistically significant structural differences in the networks, that may reflect different cognitive and emotional activation across video advertisements. The proposed approach is quite flexible and appealing as it allows to analyse induced physiological response in detail, both at a local and a global level, accounting for interrelationships among collected EEG data and participantsâ€™ heterogeneity, without the need of deriving composite scores as commonly done in neuromarketing research areas. Moreover, the approach is very general and may be successfully applied in psychological and neuroscientific research fields whenever brain response to affective stimuli is of interest.

On preformance the randomized Euler algorithm under inexact information

Tuesday, 4.07
11:20–11:40
CS 07
Room 2180

Marcin Baranek

Abstract

In this presentation we investigate error of the randomized Euler algorithm in the case when we have available only noisy information about the coefficients of the underlying SDE and about the driving Wiener process. We consider two classes of disturbed Wiener processes: The first contains disruptive functions of the $\mathcal{C}^{1,2}$ smoothness. The second class contains α, β hölder continuous corrupting functions. Our results shows upper bounds on the error of the randomized Euler algorithm under noisy information and for the both classes of disturbed Wiener processes. Finally, results of numerical experiments will also be reported.

Nearly Unstable Integer-Valued ARCH Process and Unit Root Testing

Thursday, 6.07
11:20–11:40
CS 22
Room 5440

Wagner Barreto-Souza[1], **Ngai Hang Chan**[2]

[1] University College Dublin (Ireland)
[2] The Chinese University of Hong Kong (Hong Kong)

Abstract

This paper introduces a Nearly Unstable INteger-valued AutoRegressive Conditional Heteroscedastic (NU-INARCH) process for dealing with count time series data. It is proved that a proper normalization of the NU-INARCH process weakly converges to a Cox-Ingersoll-Ross diffusion in the Skorohod topology. The asymptotic distribution of the conditional least squares estimator of the correlation parameter is established as a functional of certain stochastic integrals. Numerical experiments based on Monte Carlo simulations are provided to verify the behavior of the asymptotic distribution under finite samples. These simulations reveal that the nearly unstable approach provides satisfactory and better results than those based on the stationarity assumption even when the true process is not that close to non-stationarity. A unit root test is proposed and its Type-I error and power are examined via Monte Carlo simulations. As an illustration, the proposed methodology is applied to the daily number of deaths due to COVID-19 in the United Kingdom. ArXiv version: `https://arxiv.org/abs/2107.07963`

Inference for multivariate branching Ornstein–Uhlenbeck processes

Friday, 7.07
14:20–14:40
CS 35
Room 0.06

Krzysztof Bartoszek[1], John Tredgett Clarke[2], Jesualdo Fuentes-González[3], Venelin Mitov[4], Jason Pienaar[3], Marcin Piwczyński[5], Radosław Puchałka[5], Krzysztof Spalik[6], Kjetil Lysne Voje[7]

[1] Linköping University (Sweden)
[2] Ludwig-Maximilians-Universität München (Germany),
[3] Florida International University (USA)
[4] IntiQuan GmbH (Switzerland)
[5] Nicolaus Copernicus University in Toruń (Poland)
[6] University of Warsaw (Poland)
[7] University of Oslo (Norway)

Abstract

The recent development of fast computational algorithms for phylogenetic comparative methods finally allows for considering multiple hypotheses concerning the co-adaptation of traits and also designing complex studies on the properties of inference methods. The **mvSLOUCH** [2] `R` package for maximum likelihood estimation of multivariate Ornstein–Uhlenbeck process evolving on phylogenies has been redesigned to take advantage of the computational engine provided by **PCMBase** [3], conducting likelihood calculations for Gaussian process on trees. We used the new **mvSLOUCH** to investigate whether competing multivariate Ornstein–Uhlenbeck models can be distinguished, analyzing both simulated and empirical data [1]. We looked into relationships between traits in ungulates, in the genus *Ferula* (Apiaceae), in angiosperms and carnivorans. We find which parameters, and compound statistics of the multivariate Ornstein–Uhlenbeck process are easier to estimate, and which are more difficult. However, we also surprisingly find that forcing the sign of the diagonal of the drift matrix can seriously affect model identifiability. We also demonstrate that measurement error can make it very difficult to distinguish between models. Properties of the underlying phylogeny can make likelihood evaluation, and hence estimation, impossible. Contrary to some studies, we do not find any bias against Brownian motion processes. Interestingly, when an Ornstein–Uhlenbeck process is chosen over (the simulated under) Brownian motion, the drift is estimated to be very close to 0.

Bibliography

[1] Bartoszek K., Fuentes-González J., Mitov V., Pienaar J., Piwczyński M., Puchałka R., Spalik K., Voje K. L., *Model Selection Performance in Phylogenetic Comparative Methods under multivariate Ornstein–Uhlenbeck Models of Trait Evolution*, Systematic Biology, (2022), in press.

[2] Bartoszek K., Pienaar J., Mostad P., Andersson A., Hansen T. F., *A phylogenetic comparative method for studying multivariate adaptation*, Journal of Theoretical Biology, 314 (2012), 204–215.

[3] Mitov V., Bartoszek K., Asimomitis G., Stadler T., *Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts*, Theoretical Population Biology, 131 (2020), 66–78.

Size biased samples when modeling extreme phenomena: statistical inference and applications

Wednesday, 5.07
11:00–11:20
CS 16
Room 2180

Apostolos Batsidis[1], **Polychronis Economou**[2] **and George Tzavelas**[3]

[1] University of Ioannina (Greece)
[2] University of Patras (Greece)
[3] University of Piraeus (Greece)

Abstract

The present talk deals with two possible sources of bias that arise naturally from the selection procedure when modeling extreme phenomena, usually observed in meteorology, environmental and biomedical studies. More specifically, the first type of bias arises when the maximum values from a series of biased samples are observed, while the second one arises when a biased sample from a set of maximum values is selected. Since the ignorance of a bias leads to spurious findings in many fields, aspects of adjusting these findings have been considered by many authors. Motivated by [2] and [3], the concept of weighted distributions is used not only as a method of describing both cases but also as an adjustment methodology applicable to many situations in which the recorded observations cannot be considered as a random sample from the parent distribution (see, for example, [1]). In this context, the properties of the maximum likelihood estimators are studied, while the differences between the two types of bias and the impact of ignoring the biasness are revealed with the use of numerical examples.

Acknowledgements

Apostolos Batsidis acknowledges support of this work by the project "Establishment of capacity building infrastructures in Biomedical Research (BIOMED-20)" (MIS 5047236) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme Competitiveness, Entrepreneurship and Innovation (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

Bibliography

[1] Patil, G.P., Rao, C.R., *Weighted Distributions and Size-Biased Sampling with Applications to Wildlife Populations and Human Families*, Biometrics, 34 (1978), 179–189.
[2] Fisher, R.A., *The effect of methods of ascertainment upon the estimation of frequencies*, Annals of Eugenics, 6 (1934), 13–25.
[3] Rao, C.R., *On Discrete Distributions Arising out of Methods of Ascertainment*, Sankhya: The Indian Journal of Statistics, Series A, 27 (1965), 311–324.

Almost Omnibus Nonparametric Inference for Two Independent Samples

Thursday, 6.07
11:20–11:40
CS 20
Room 1.40

Jonas Beck[1], **Patrick B. Langthaler**[2], **Arne C. Bathke**

[1] University of Salzburg
[2] Paracelsus Medical University

Abstract

Different statistical functionals are omnipresent in nonparametric statistics. Functionals like the Mann-Whitney functional (relative effect) $P(X < Y)$ measure a location or stochastic tendency effect for two independent samples. We extend this by additionally incorporating an overlap index (nonparametric dispersion measure). We develop the joint asymptotic distribution of their rank-based estimators and construct confidence regions based on a resampling approach.

Extending the two-sample rank sum test (Wilcoxon, Mann and Whitney), we propose a new test based on these functionals for the hypothesis of distribution equality. As we simultaneously test different functionals we get a much larger consistency region as in the classical test and in most cases a substantial improvement in the power to the Kolmogorow Smirnov Test, as our simulation shows. We additionally show the usability of our approach by applying the test to different real data sets.

Machines don't go for lunch: A new diurnal adjustment for trade durations

Thursday, 6.07
16:20–16:40
CS 29
Room 4420

Markus Belfrage

Hanken School of Economics (Finland)

Abstract

A new diurnal adjustment method is proposed for stock trade durations. Trade durations are often modelled by the class of Autoregressive Conditional Duration (ACD) models, where it is assumed that the seasonality factor acts multiplicatively on all durations. This assumption is shown to be violated for ultra-high precision trade data when a large portion of trades are executed by computer algorithms. A two-component mixture model that varies nonparametrically over time of the day is developed as a response to the heterogeneity in the diurnal seasonality caused by the mix of machines and regular traders. An estimation algorithm much in the flavor of expectation maximization is proposed and applied to a large set of Apple (AAPL) trades using data from Nasdaq Historical TotalView ITCH at the nanosecond precision.

Bayesian UQ and structure detection for multiple change-points models

Thursday, 6.07
11:00–11:30
IS 13
Room 0.03

Eduard Belitser[1] **and Subhashis Ghosal**[2]

[1] VU Amsterdam (The Netherlands)
[2] North Carolina State University (USA)

Abstract

Data observed over a long period of time may contain several change-points, where the distribution of a variable changes but remains the same over the blocks in between. This useful qualitative structure allows precise estimation and uncertainty quantification for a long vector of parameters. Detecting these change-points is another important objective. We derive a concentration inequality for an empirical Bayes procedure, obtain the frequentist coverage of a suitable confidence ball of the optimal size constructed from the posterior distribution and study the problem of change-points detection. We adopt an oracle approach to quantify the estimation error locally and show that the estimation error of the proposed procedure matches with the oracle rate, thus automatically implying minimax optimality, adaptively over all change-points structures. Under a condition on the minimum magnitude of the changes, we show that precisely all change-points are detected with high probability, and accompany this with a lower bound result asserting the minimality of that condition. Our results are non-asymptotic and robust in that normality is used only as a working model in the procedure, but the true distribution may not be normal. We discuss important extensions of our results to Hilbert space-valued parameters to address the multiple change-points problem for multivariate and functional data.

Statistical inference for McKean-Vlasov SDEs

Thursday, 6.07
16:00–16:30
IS 17
Room 0.03

Denis Belomestny[1], **Vytaute Pilipauskaite**[2], **Mark Podolskij**[2]

[1] Department of Mathematics, Duisburg-Essen University
[2] Department of Mathematics, University of Luxembourg

Abstract

In this paper we study the problem of semiparametric estimation for a class of McKean-Vlasov stochastic differential equations. Our aim is to estimate the drift coefficient of a MV-SDE based on observations of the corresponding particle system. We propose a semiparametric estimation procedure and derive the rates of convergence for the resulting estimator. We further prove that the obtained rates are essentially optimal in the minimax sense.

Bibliography

[1] Belomestny, D., Pilipauskaite, V., & Podolskij, M. (2023, February). Semiparametric estimation of McKean-Vlasov SDEs. In Annales de l'Institut Henri Poincare (B) Probabilites et statistiques (Vol. 59, No. 1, pp. 79-96). Institut Henri Poincare.

Risk estimation of late radio-induced cardiac disease in childhood cancer survivors

Wednesday, 5.07
11:40–12:00
CS 18
Room 4420

Mahmoud Bentriou[1], **Véronique Letort-Le Chevalier**[1], **Sarah Lemler**[1], **Rodrigue Allodji**[2]

[1] Université Paris-Saclay, CentraleSupélec, MICS laboratory, 91190, Gif-sur-Yvette, France
[2] INSERM, CESP, Cancer and Radiation Team, F-94805 Villejuif, France

Abstract

Cardiac disease (CD) is a primary long-term diagnosed pathology among childhood cancer survivors, which may occur as a side effect of radiotherapy. Our work is motivated by the long-term follow-up of 7670 5-year survivor patients from the French Childhood Cancer Survivors Study (FCCSS), a cohort study set up to better understand and anticipate the risks induced by cancer treatments. Whole-body 3D dose distributions delivered during radiotherapy treatments are reconstructed for the cancer survivors. Standard models for predicting the radio-induced risk induced rely on aggregated dose data, such as dose-volume indicators, that do not include spatial information. In this study, texture-based features called dosiomics [2] are extracted from these 3D dose distributions and integrated into a survival analysis performed by Cox Proportional Hazard and Random Survival Forests with regularization screening methods in a context of high right censorship and dimension. We show how these texture-based models catch additional information by focusing on the time-dependent statistical estimators of the bounded inverse-probability-of-censoring weights C-index [1] and the Brier score metrics used for performance evaluation.

Bibliography

[1] Uno H., Cai T., Pencina M. J., D'Agostino R. B., Wei L. J., *On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data*, Statistics in Medicine, 2011.

[2] Zhang X., Zhang Y., Zhang G., Qiu X., Tan W., Yin X., Liao L., *Deep Learning With Radiomics for Disease Diagnosis and Treatment: Challenges and Potential*, Frontiers in Oncology, 2022.

Selecting interaction effects in regression using I-priors

Wednesday, 5.07
11:00–11:20
CS 15
Room 1.40

Wicher Bergsma[1], **Haziq Jamil**[2]

[1] London School of Economics and Political Science (UK)
[2] Universiti Brunei Darussalam (Brunei)

Abstract

Additive models with interactions have been considered extensively in the literature, using estimation methods such as splines or Gaussian process regression. At least two difficulties have hampered their application: (i) estimating the models can be difficult due to potentially many tuning (or smoothing) parameters, and (ii) model selection may be difficult due to a lack of adequate. In this paper we attempt to address these issues with a novel approach to estimating and selecting additive models with and without interaction effects.

Firstly, we extend the I-prior methodology [1] to multiple covariates, each of which may be multidimensional. For this purpose, we define a class of hierarchical interaction models assuming the regression function lies in a *reproducing kernel Krein space* (RKKS), and derive the (possibly indefinite) reproducing kernel for the models. The I-prior is an objective prior for a statistical parameter based on its Fisher information. In the present case, the I-prior for the regression function is Gaussian with covariance kernel proportional to its (positive definite) Fisher information, with support a subset of the assumed RKKS, i.e., the I-prior is proper. The I-prior methodology has some theoretical and practical advantages over competing methods such as Gaussian process regression and Tikhonov regularization. A practical (computational) advantage is that it permits an EM algorithm with simple E and M steps to find the maximum marginal likelihood estimators of the *scale* parameters (also known as tuning parameters), making their estimation easier than for competing methods.

A second and very simple innovation we introduce is a parsimonious specification of models with interactions. That is, each covariate is assigned a single scale parameter, regardless of the number of interactions present in the model. This approach is mathematically justified and has two key advantages: (i) estimation of models with interactions is simplified due to the presence of fewer scale parameters, and (ii) model selection (among models with different interactions present) is simplified, in that simply the model with the highest marginal likelihood can chosen.

An R-package implementing our methodology is available [2].

Bibliography

[1] Bergsma, W. P., *Regression with I-priors*, Econometrics and Statistics, 14 (2020), 89–111.

[2] Jamil, H. and W. P. Bergsma, *iprior: An R package for regression modelling using I-priors*, `https://arxiv.org/abs/1912.01376`.

Two-sided Matrix Regression

Monday, 3.07
16:45–17:05
CS 02
Room 1.40

Nayel Bettache, Cristina Butucea

CREST, ENSAE, Institut Polytechnique de Paris (France)

Abstract

The two-sided matrix regression model $Y = A^*XB^* + E$ aims at predicting Y by taking into account both linear links between columns targets of Y and the column features of X, via the unknown matrix B^*, and also among the row targets of Y and the row features of X, via the unknown matrix A^*. We assume that the rank of the signal is smaller than the dimensions of the observation Y (rank sparsity). This high dimensional problem is highly non-identifiable but the prediction risk benefits from the rank constraint as it attains faster rates under this sparsity assumption.

We propose low-rank predictors in this high-dimensional matrix regression model via rank-penalized least squares. The criterion is non jointly convex but we propose explicit predictors based on the Singular Value Decomposition of the target matrix Y and the feature matrix X and show optimal prediction bounds. We also give sufficient conditions for consistent rank selector and propose a fully data-driven rank-adaptive procedure. We assume independent and identically distributed centered gaussian entries $\mathcal{N}(0, \sigma^2)$ for the noise matrix E and consider both cases of known and unknown variance σ^2.

In the oracle situation where $r^* = \text{rank}(A^*XB^*)$ and σ^2 are known, we build explicit predictors $(\hat{A}_{r^*}, \hat{B}_{r^*})$ solutions to the non-convex minimization problem

$$\min_{\substack{A,B:\\ \text{rank } A \wedge \text{rank } B \leq r^*}} \|Y - AXB\|_F^2.$$

In this case, we give an explicit upper bound on the Frobenius prediction risk $\|A^*XB^* - \hat{A}_{r^*}X\hat{B}_{r^*}\|_F^2$ in high probability.

In the unknown r^* - known σ^2 situation, we propose rank-adaptive predictors $(\hat{A}_{\hat{r}}, \hat{B}_{\hat{r}})$, selected from the family $\{(\hat{A}_r, \hat{B}_r) : r \in \{1, \dots, \min(n, p, \text{rank}(X))\}\}$ by a model selection procedure analogous to that of [1].

For a given $\lambda > 0$, we consider $\hat{r} := \arg\min_{r \in [n \wedge p \wedge r_X]} \left\{\|Y - \hat{A}_r X \hat{B}_r\|_F^2 + \lambda r\right\}$. We then build the predictors $(\hat{A}_{\hat{r}}, \hat{B}_{\hat{r}})$ and provide an upper bound on the Frobenius prediction risk $\|A^*XB^* - \hat{A}_{\hat{r}}X\hat{B}_{\hat{r}}\|_F^2$ in high probability given a σ^2-dependent lower bound on λ. The bounds attained by our procedure are analogous to those for the low-rank matrix regression models in [2] and [3]. We also study the consistency of the rank selector $\hat{r}$ and see when it recovers the true rank r^* with high probability.

Finally we investigate the unknown variance case and provide a data-adaptive rank estimator in this situation. We extend the previously obtained upper bounds to this situation and provide a computationally efficient algorithm for λ selection.

Bibliography

[1] Bunea F., She Y., Wegkamp M.H., *Optimal selection of reduced-rank estimators of high-dimensional matrices*, The Annals of Statistics, 39 (2011), 1282–1309.

[2] Rohde A., Tsybakov A.B., *Estimation of high-dimensional low-rank matrices*, The Annals of Statistics, 39 (2011), 887–930.

[3] Giraud C., *Low rank multivariate regression*, Electronic Journal of Statistics, 5 (2011), 775–799.

On Modeling Bivariate Lifetime Data in the Presence of Inliers

Tuesday, 4.07
16:00–18:00
Poster

Sumangal Bhattacharya[1], **Ishapathik Das**[1] and **Muralidharan Kunnummal**[2]

[1] Department of Mathematics and Statistics, Indian Institute of Technology Tirupati, Tirupati, India, 517506

[2] Department of Statistics, Faculty of Science, The Maharaja Sayajirao University of Baroda, Vadodara, India, 390002

Abstract

In reliability, the lifetime data is usually modeled using one or two parametric distributions, such as Weibull, gamma, log-normal, Pareto, etc., which are unimodal by nature. Sometimes, the data may contain many zeros or close to zero data points, defined as inliers (instantaneous or early failure observations) in the literature. The usual modeling approach using the unimodal parametric distributions may not provide expected results for such data in the presence of inliers. Furthermore, correlated bivariate observations with inliers frequently occur in reliability; here, we propose a method of modeling bivariate lifetime data with instantaneous and early failure observations. We construct a new bivariate distribution function by combining bivariate uniform and Weibull distributions. The bivariate Weibull distribution is obtained using a 2-dimensional copula, assuming the marginal distributions as two parametric Weibull distributions. We derive some properties of that modified bivariate Weibull distribution, mainly the joint probability density function, the survival (reliability) function, and the hazard (failure rate) function. The model's unknown parameters are estimated using the Maximum Likelihood Estimation (MLE) technique combined with a machine learning clustering algorithm. Numerical examples are provided using simulated data to illustrate and test the performance of the proposed methodologies. The method is also applied to real data and compared with existing methods in the literature.

Recurrent and Residual Neural Networks as differential equations

Wednesday, 5.07
09:30–10:30,
14:00–15:00
Plenary Lecture
Room 0.03

Gérard Biau[1], **Adeline Fermanian**[1], **Pierre Marion**[1], **Jean-Philippe Vert**[2]

[1] Sorbonne University (France)
[2] Owkin (France)

Abstract

Deep learning has become a prominent approach for many applications, such as computer vision or neural language processing. The mathematical understanding of these methods is still incomplete. A recent approach is to consider neural networks as discretized versions of differential equations. I will first give an overview of this emerging field and then discuss new results on Recurrent and Residual Neural Networks, which are common types of deep networks.

Testing for jumps in processes with integral fractional part and jump-robust inference on the Hurst exponent

Tuesday, 4.07
11:30–12:00
IS 07
Room 0.06

Markus Bibinger, Michael Sonntag

Julius-Maximilians-Universität Würzburg (Germany)

Abstract

We present a test for jumps based on high-frequency observations of a fractional process with an additive jump component. The Hurst exponent of the fractional process is unknown. For this reason, truncation methods which are classical for semimartingales are not feasible. It is shown that our test statistic, the maximal absolute, second-order increment, normalised with a scaled jump-robust spot volatility estimate, converges to a Gumbel distribution under the null hypothesis of no jumps. The asymptotic theory under infill asymptotics builds upon extreme value theory for weakly dependent, stationary time series and extends techniques for the semimartingale case from the literature. We prove consistency under the alternative hypothesis when there are jumps. Moreover, we establish convergence rates for local alternatives. We compare our results to existing methods for the semimartingale setting. We demonstrate sound finite-sample properties in a simulation study. In the process, we show that inference on the Hurst exponent of a rough fractional process can be performed robustly with respect to jumps. This provides an important insight for the growing literature on rough volatility.

L-statistics as quantile estimators

Friday, 7.07
11:00–11:20
CS 31
Room 1.40

Mariusz Bieniek, Luiza Pańczyk

Maria Curie Skłodowska University (Poland)

Abstract

We consider the problem of quantile estimation by suitably chosen L-statistics. For a given sample size n we determine optimal L-statistics as estimators of the quantile of a given order $p \in (0, 1)$. We use a new criterion of optimality, introduced in our paper [1], based on sharp bounds on the bias of the estimation. First we study the most popular cases of single order statistics and linear combinations of a pair of successive order statistics. Next, we study the mean square error of derived estimators and we compare them with well known L-statistics such as Kaigh-Lachenbrugh or Harell-Davis estimators.

Bibliography

[1] Bieniek, M., Pańczyk, L. On the choice of the optimal single order statistic in quantile estimation. Annals of the Institute of Statistical Mathematics 75 (2023), 303–333.

Supply chain contracts with returns handling under price-sensitive stochastic demand

Friday, 7.07
14:20–14:40
CS 34
Room 1.01

Milena Bieniek

Institute of Management and Quality Sciences, Faculty of Economics,
Maria Curie-Sklodowska University in Lublin (Poland)

Abstract

E-commerce constitutes a system for the online purchase and sale of services and commodities. The flood of returns constitutes a vital problem of e-commerce. It is definitely more serious than in the case of traditional sales. Research shows that online orders are returned by shoppers more than three times as frequently as when the same goods are purchased offline.

We investigate the manufacturer handling strategy which occurs in online shopping, in a centralized or decentralized channel under the wholesale price contract. The retailer's optimal order quantity, price, and the manufacturer's wholesale price are derived assuming additive uncertainty in demand. The possibility of negative demand realizations, which may occur in adverse market circumstances, is verified in the investigated models. It was proved that the imposition of the non-negativity prerequisite on demand is vital to obtain complete results. The non-negativity constraint in this study incorporates consumer's returns handling costs which is different than the previously used constraint. As an extension of the additive case, the model with iso-elastic demand is discussed. The theoretical analysis is enriched with numerical examples.

Our work extends the model of [1] for the manufacturer returns handling by adding price as a endogenous variable. In [1], the retail price is given in advance and the consumer demand is stochastic and price-free. Compared to that work, in this article the demand function is not only stochastic but also linearly price-sensitive. The study shows that the non-negativity limitation is indispensable to avoid suboptimal solutions and to ensure the completeness of the discussion.

Bibliography

[1] Liu, Jian and Xiao, Tiaojun and Tian, Chen and Wang, Huimin, *Ordering and returns handling decisions and coordination in a supply chain with demand uncertainty*, International Transactions in Operational Research, 27 (2020), 1033-1057, doi: 10.1111/itor.12542.

Spatio-temporal Bayesian models for environmental epidemiology: methods and examples

Friday, 7.07
16:00–17:00
Plenary Lecture
Room 0.03

Marta Blangiardo[1,2]

[1] MRC Centre for Environment and Health, Imperial College London (UK)
[2] Turing-RSS Health Data Lab (UK)

Abstract

In this talk I will give an overview of the statistical modelling approach used for analysing data characterised by spatial and temporal dependence and will focus on applications in the field of environmental epidemiology. I will start introducing the general statistical framework used in environmental epidemiology when dealing with spatial and temporal data and then present how these methods have been used on two different real-world applications:

1. **Air pollution sources and their health impact** – Particulate matter (PM) is a complex mix of organic and inorganic compounds of distinct sources, with a range of physical and chemical properties, which might have a different harmful effects to health. Disentangling total ambient PM concentration into its sources is key for developing strategies to reduce PM through targeted actions. I will present a Bayesian model for (i) apportioning airborne particles into sources using a nonparametric approach and (ii) assessing the impact of each source on health outcomes. To illustrate the proposed model framework, I will use particle size data measured at an urban background site in London (UK) and cardio-respiratory hospital admission.
2. **Wastewater based epidemiology in the COVID pandemic** – The utility of wastewater-based epidemiology as an early warning tool has been explored widely across the globe during the current COVID-19 pandemic. However, no attempt has been made to develop a model that predicts wastewater viral concentration at fine spatio-temporal resolutions covering an entire country, a necessary step towards using wastewater monitoring for the early detection of local outbreaks. I will first show how we can model the relationship between weekly viral concentration in wastewater at specific locations and a collection of covariates covering socio-demographics, land cover and virus-associated genomic characteristics. I will then discuss the potential for joint modelling of wastewater data and COVID-19 prevalence on a space-time resolved domain to improve the performance of public health surveillance systems.

Approximate subgraph count: asymptotic normality and jackknife estimate of variance

Friday, 7.07
11:00–11:20
CS 30
Room 1.01

Mindaugas Bloznelis

Vilnius University (Lithuania)

Abstract

We are interested in the number N_F of copies of a small graph F in a large graph G. Given s we sample a random subset S of size $|S| = s$ of vertices of G and count the number t_S of copies of F incident to S. The weighted count $t_S/(1-(v_G-s)_{v_F}/(v_G)_{v_F})$, where v_F and v_G stand for the number of vertices of F and G, is an unbiased estimator of N_F. Assuming that G is a sparse random community affiliation graph with clustering [1, 2] and that F is balanced and 2-connected we show that the estimator is asymptotically normal and establish the consistency of its jackknife variance estimator.

Bibliography

[1] Bloznelis M., Leskelä, L., Clustering and percolation on superpositions of Bernoulli random graphs, *Random Structures and Algorithms*, DOI: 10.1002/rsa.21140.

[2] Bloznelis M., Karjalainen J., Leskelä L., Normal and stable approximation to subgraph counts in superpositions of Bernoulli random graphs, arxiv:2107.02683.

Semi-parametric Bernstein-von Mises theorem for linear models with one-sided error

Wednesday, 5.07
11:00–11:30
IS 03
Room 0.06

Natalia Bochkina[1], **Judith Rousseau**[2], **J.-B. Salomond**[3], **Johan van der Molen Moris**[4]

[1] University of Edinburgh and Maxwell Institute, UK
[2] University of Oxford, UK
[3] University of Universite Paris-Est Creteil, France
[4] University of Cambridge, UK

Abstract

We consider the problem of linear regression with one-sided iid errors where the density of the errors is unknown, from a Bayesian perspective. This is an extension of the problem of estimating an unknown density on $[\theta, \infty)$ where θ is also unknown, to the case θ being dependent on some covariates. For the case of known density, this problem was considered by Chernozhukov and Hong (2004) in the context of procurement auctions.

We state general sufficient conditions for the local concentration of the marginal posterior of the parameters in the linear regression model (Bernstein - von Mises type theorem) which have a faster 1/n contraction rate and a constrained multivariate exponential distribution with constraint dependent on the observed data, under an adaptive prior of the unknown smooth density. To estimate the unknown density, we consider a mixture of Gamma densities convolved with a Uniform distribution which ensures that a priori the unknown density is decreasing and smooth. The additional challenge in estimating the unknown density here is that it is crucial to estimate consistently the value of the unknown density of errors at zero, as it is the scale parameter of the limiting constrained exponential distribution. In particular, to ensure that the error density is asymptotically pointwise consistent in a neighbourhood of zero, instead of a usual Dirichlet mixture weights, we consider a mixture based on a non-homogeneous Completely Random Measure. We illustrate performance of this approach on simulated data, and apply it to model distribution of bids in procurement auctions.

Maximum likelihood estimation for discrete exponential families and random graphs

Thursday, 6.07
14:20–14:40
CS 25
Room 2180

Krzysztof Bogdan[1], **Michał Bosy**[2], **Tomasz Skalski**[1]

[1] Wrocław University of Science and Technology
[2] University College London

Abstract

We shall discuss a criterion for the existence of the maximum likelihood estimator for discrete exponential families, based on the paper [1]. Our criterion is simple to use, as we show in various settings, e.g., for exponential models of random graphs. As an application, we specify the size of independent identically distributed samples for which the maximum likelihood estimator exists with high probability.

Bibliography

[1] K. Bogdan, M. Bosy, T. Skalski. Maximum likelihood estimation for discrete exponential families and random graphs. *ALEA Lat. Am. J. Probab. Math. Stat. 19 (2022), no. 1, 1045â€“1070.* `https://alea.impa.br/articles/v19/19-43.pdf`

Sparse Graphical Modelling via the sorted L_1 – Norm

Thursday, 6.07
15:00–15:20
CS 26
Room 5440

M. Bogdan[1,2], **Giovanni Bonaccolto**[3], **I. Hejny**[2], **Philipp J. Kremer**[4], **S. Paterlini**[5], **R. Ricobello**[5], **P. Sobczyk**[6] **and J. Wallin**[2]

[1] U. Wroclaw
[2] U. Lund
[3] U. Enna
[4] EBS, Wiesbaden
[5] U. Trento
[6] UAM, Poznan

Abstract

Sparse graphical modelling has attained widespread attention across various academic fields. We propose two new graphical model approaches, Gslope and Tslope, which provide sparse estimates of the precision matrix by penalizing its sorted L_1-norm, and relying on Gaussian and T-student data, respectively. Gslope and Tslope reduce the dimension by identifying zeros in the precision matrix as well as by identifying clusters of parameters which are equal to each other. We present theoretical results concerning this low dimensional pattern recovery and provide the selection of the tuning parameters which allows for control of the number of false edges between the disjoint graph components. Extensive simulations and real world analysis show the advantages of the new methods as compared to other state-of-the-art sparse graphical modelling approaches.

Robust Bayesian estimation and prediction in gamma-gamma model of claim reserves

Tuesday, 4.07
15:00–15:20
CS 09
Room 1.01

Agata Boratyńska

Warsaw School of Economics SGH (Poland)

Abstract

The problem of robust Bayesian estimation of chain ladder (development) factors and Bayesian prediction of claim reserves is considered. Two different classes of priors (classes of gamma distributions with bounded parameters and a nonparametric class based on distortion of a prior cummulative distribution function) are presented. The oscillation (as a measure of robustness) of Bayes estimators and predictors of reserves, when priors are in the considered class, is calculated and the posterior regret Γ-minimax (PRGM) rules as optimal procedures are obtained. The numerical example compares different methods of the estimation of development factors and the calculation of claim reserves. The chain ladder estimators and predictors, exact Bayes estimators and predictors, PRGM estimators and predictors for aforementioned classes of priors and empirical credibility estimators and predictors are considered. It is shown that the variability of the expected value parameter of a prior has a greater impact on the oscillation of the Bayes estimators of the development factors and Bayes predictors of reserves than the variability of the shape parameter. The Bayes predictors are more robust with respect to the distortion of a prior c.d.f. than the fluctuation of the expected value parameter of a Gamma prior. A distortion of the shape of the prior c.d.f. has also a smaller impact on the value of PRGM estimators and predictors than the variability of the parameters of Gamma distribution. The difference between the Bayesian and chain ladder estimators (and predictors) depends mainly on the difference between the expected value parameter of the prior and the chain ladder estimator of development factor. The presented results are published in the paper A. Boratyńska, Z. Zielińska-Kolasińska *Robust Bayesian estimation and prediction in gamma-gamma model of claim reserves* in Insurance: Mathematics and Economics 105 (2022), 194–202.

A modified dividing local Gaussian processes algorithm for theoretical particle physics applications

Tuesday, 4.07
16:00–18:00
Poster

Timo C. Braun, Anders Kvellestad, Riccardo De Bin

University of Oslo (Norway)

Abstract

In the search of new types of fundamental particles and new interactions that go beyond the Standard Model of particle physics, physicists face severe issues due to the many time-consuming physics calculations and simulations needed to evaluate the joint likelihood function. Since the set of relevant experimental results that enter the likelihood function is usually updated between every new fit, there is the need for an algorithm that provides an approximate value in a reasonable time. To be able to tackle the issue and provide an estimate of the likelihood value *on-the-fly* during the execution of the fit, here we propose an approach based on the DLGP algorithm of Lederer et al (2020) [1]. The parametric space is divided in overlapping regions organized within a tree structure, where each subregion is modelled by a mixture of Gaussian processes. In particular, motivated by our physics problem, we extend Lederer et al (2020)'s approach to improve its prediction accuracy, both by better estimating the parameters of the Gaussian Processes and by implementing suitable forms of the covariance function. Further modifications include new ways to split the parameter space in subregions and to characterize the overlap between them. Our approach is illustrated on data from the GAMBIT project [2]: To mimic the real case scenario, we pass the data points to our algorithm one by one, as if they were sampled by the differential evolution scanner used in the physics experiment. A dedicated R package, `GPTreeO`, has been created to facilitate the use of our modified dividing local Gaussian processes algorithm in practice.

Bibliography

[1] Lederer, A., Conejo, A. J. O., Maier, K., Xiao, W., Umlauft, J., & Hirche, S. (2020). Real-time regression with dividing local Gaussian processes. *arXiv preprint arXiv:2006.09446*.

[2] GAMBIT Collaboration: Athron, P., Balázs, C. et al (2019). Combined collider constraints on neutralinos and charginos. *The European Physical Journal C, 79: 1–52.*

A varying-coefficient beta-binomial model for the analysis of DNA-methylation sequencing data

Wednesday, 5.07
11:20–11:40
CS 16
Room 2180

Katarzyna Górczak[1,2], **Jürgen Claesen**[1,3], **Tomasz Burzykowski**[1,4]

[1] Hasselt University (Belgium)
[2] Open Analytics (Belgium)
[3] Amsterdam University Medical Center (The Netherlands)
[4] Medical University of Bialystok (Poland)

Abstract

DNA methylation is an important epigenetic modification involved in gene regulation. Advances in the next generation sequencing technology have enabled the retrieval of DNA-methylation information at single-base-resolution. However, due to the sequencing process and the limited amount of isolated DNA, DNA-methylation data are often noisy and sparse, which complicates the identification of differentially methylated regions (DMRs), especially when few replicates are available. We present a varying-coefficient model for detecting DMRs by using single-base-resolved methylation information. The model simultaneously smooths the methylation profiles and allows detection of DMRs, while accounting for additional covariates. The proposed model takes into account possible overdispersion by using a beta-binomial distribution. The overdispersion itself can be modeled as a function of the genomic region and explanatory variables. We illustrate the properties of the proposed model by applying it to a real-life case study.

Variables selection in high dimension in a joint model of survival times and longitudinal outcomes with random effects

Monday, 3.07
15:45–16:05
CS 02
Room 1.40

Antoine Caillebotte[1,2], **Estelle Kuhn**[2], **Judith Legrand**[1], **Sarah Lemler**[3], **Elodie Marchadier**[1]

Paris-Saclay University, INRAE,
[1] UMR GQE-Moulon,
[2] MaIAGE, France,
[3] Paris-Saclay University, Laboratoire MICS, France

Abstract

A current challenge in many fields is understanding the interactions between dependent dynamic phenomena. For example, in plant science, the dynamics of plant development and the spread of an epidemic disease or pests in this field. Mathematical modeling is a powerful tool for capturing interactions between phenomena. Indeed, the joint modeling of several phenomena has shown its efficiency in several areas, notably in medicine, pharmacology, and biology (see [1]).

In this work, we consider a joint model of survival time and repeated measurements, longitudinal data in our setting, of a population of individuals. More precisely, we model the evolution of a dynamic variable of interest over time using a nonlinear mixed-effects model (see [2]). This model is combined with a Cox model (see [3]) in which we introduce classical covariates characterizing the individual considered and also a link function based on the mixed-effects model function of the related dynamic (see [4]). Moreover, we consider a high-dimensional setting where the number of covariates in the Cox model is larger than the sample size. It is necessary to adapt the statistical and numerical approaches for inference to deal with this high-dimensional setting since classical tools failed. Therefore we propose a new method for parameter inference and variable selection. We consider the penalized maximum likelihood estimate using a LASSO penalty to account for the high dimensionality of covariates. We implement a preconditioned proximal stochastic gradient algorithm to compute this estimate, dealing both with unobserved random effects of the mixed-effects model and the penalty term. We fixed the penalty coefficient using cross-validation. Finally, we highlight the performance of our method through an intensive simulation study and apply it to a real data set in plant sciences.

Bibliography

[1] Kerioui, Marion et al., *Modelling the association between biomarkers and clinical outcome: An introduction to nonlinear joint models*, British Journal of Clinical Pharmacology 88.4, 2022, pp. 1452-1463.

[2] Davidian, M. and D.M. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1995.

[3] Cox, D. R., *Regression Models and Life-Tables*, Journal of the Royal Statistical Society. Series B (Methodological) 34.2, 1972, pp. 187-220

[4] Rizopoulos, Dimitris, *Joint models for longitudinal and time-to-event data: With applications in R.*, CRC press, 2012.

Learning, evaluating and analysing a recommendation rule for early blood transfer in the ICU

Tuesday, 4.07
11:00–11:30
IS 06
Room 0.03

Antoine Chambaz[1], **Nicolas Gatulle**[2], **Julie Josse**[3], **Pan Zhao**[3] **and the Traumabase Group**

[1] Université Paris Cité, CNRS, MAP5, F-75006 Paris (France)
[2] Sorbonne University, GRC 29, AP-HP, DMU DREAM, Department of Anaesthesiology and Critical Care, Pitié-Salpêtrière Hospital, Paris (France)
[3] PreMeDICaL team, Inria, Desbrest Institute of Epidemiology and Public Health, University of Montpellier (France)

Abstract

Severely injured patients experiencing haemorrhagic shock often require massive transfusion. Early transfusion of blood products (plasma, platelets and red blood cells) is common and associated with improved outcomes in the hospital. However, determining a right amount of blood products is still a matter of scientific debate.

The speaker will present and discuss a methodology to learn, evaluate and analyse a recommendation rule for early blood transfer in the ICU. The study uses data from the French Traumabase, a French observatory for Major Trauma.

When does Metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm?

Friday, 7.07
14:00–14:30
IS 05
Room 0.03

Yuansi Chen[1], **Khashayar Gatmiry**[2]

[1] Duke University
[2] MIT

Abstract

We analyze the mixing time of Metropolized Hamiltonian Monte Carlo (HMC) with the leapfrog integrator to sample from a distribution on $\mathbb{R}^d$ whose log-density is smooth, has Lipschitz Hessian in Frobenius norm and satisfies isoperimetry. We bound the gradient complexity to reach ϵ error in total variation distance from a warm start by $\tilde{O}(d^{1/4}\text{polylog}(1/\epsilon))$ and demonstrate the benefit of choosing the number of leapfrog steps to be larger than 1. To surpass previous analysis on Metropolis-adjusted Langevin algorithm (MALA) that has $\tilde{O}(d^{1/2}\text{polylog}(1/\epsilon))$ dimension dependency in Wu et al. (2021), we reveal a key feature in our proof that the joint distribution of the location and velocity variables of the discretization of the continuous HMC dynamics stays approximately invariant. This key feature, when shown via induction over the number of leapfrog steps, enables us to obtain estimates on moments of various quantities that appear in the acceptance rate control of Metropolized HMC.

A Shared Frailty Competing Risk Model with Time-varying Covariates: an Application to Hereditary Breast and Ovarian Cancer Families

Tuesday, 4.07
16:00–18:00
Poster

Yun-Hee Choi[1], **Seungwoo Lee**[1], **Laurent Briollais**[2,3]

[1] Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, The University of Western Ontario (Canada)
[2] Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital (Canada)
[3] Biostatistics, Dalla Lana School of Public Health, University of Toronto (Canada)

Abstract

Families with hereditary breast and ovarian cancer (HBOC) syndrome are at increased risk of developing both breast and ovarian cancers, mostly due to BRCA 1/2 gene mutation, and are recommended to undergo frequent screening or prophylactic surgery for prevention or early detection. Time-to-event data arising from HBOC families are often complicated by multiple correlated outcomes, time-varying interventions, competing risks, and familial correlation.

We developed a shared frailty competing risks model with time-varying covariates to evaluate time-varying effects of interventions on the risk of breast cancer in the presence of competing events such as ovarian cancer or death. Flexible modelling of time-varying covariate effects was implemented using either a permanent exposure or exponential decays model, or a non-parametric smoothing approach by B-splines. The proposed model addresses several statistical challenges, including correlated competing risks, time-varying covariate, and time-varying effect and familial correlation, and provides age-specific cancer risk estimates adjusted for individual's intervention time, mutation status, and familial correlation. The performance of the proposed models was evaluated through simulation studies under different scenarios of time-dependent effect models.

We also applied our proposed models to 498 HBOC families habouring BRCA1 mutated genes recruited through the Breast Cancer Family Registry to evaluate the time-dependent effects of mammographic screening and risk-reducing salpingo oophorectomy on breast cancer risk in the presence of competing risks. We provided the cause-specific penetrance estimates of breast cancer for the individuals in HBOC families, using three time-dependent models—permanent exposure, exponential decay, and B-splines.Their performance of those models was compared using time-dependent C-index and Brier scores with HBOC family data.

Uncovering Data Symmetries: Estimating Covariance Matrix in High-Dimensional Setting With 'gips' R Package

Tuesday, 4.07
16:00–18:00
Poster

Adam Przemysław Chojecki, Paweł Morgen

Warsaw University of Technology (Poland)

Abstract

In high-dimensional settings, where the number of variables exceeds the number of observations, accurately estimating the covariance matrix poses a significant challenge. This poster presents a novel approach that leverages the identification of symmetries within the data to improve covariance matrix estimation. In the 'gips' R package [2], we implement the Bayesian model selection procedure within Gaussian vectors (invariant under the permutation group) introduced in [1].

Our method aims to capture the underlying low-dimensional structure by exploring the permutation symmetries within the data. Identifying symmetries enables us to interpret relationships in data in a new and natural way. The 'gips' package provides a comprehensive set of functions that facilitate identifying and utilizing symmetries, making it a valuable resource for researchers working with high-dimensional data.

We demonstrate the effectiveness of our approach through simulations and real-world data examples. Our results show that incorporating data symmetries leads to more reliable covariance matrix estimates, enabling better inference and decision-making.

The presented novel approach contributes to the growing field of statistical methods for $p > n$, offering promising avenues for future research and practical applications.

Bibliography

[1] P. Graczyk, H. Ishi, B. Kołodziejek and H. Massam., *Model selection in the space of Gaussian models invariant by symmetry*, The Annals of Statistics, 2022

[2] A. Chojecki, P. Morgen, B. Kołodziejek, *gips: Gaussian Model Invariant by Permutation Symmetry*, CRAN, 2022, `https://CRAN.R-project.org/package=gips`, `https://przechoj.github.io/gips/`

A symmetry-based simulation method for Bayesian nonparametric models with binary response data

Monday, 3.07
16:45–17:05
CS 01
Room 1.01

Dennis Christensen

University of Oslo (Norway)
Norwegian Defence Research Establishment (FFI) (Norway)

Abstract

Binary responses are a central object of study in statistics, and arise naturally in applications such as bioassay, current-status data and sensitivity testing [1, 2]. There has been a large interest in applying Bayesian nonparametric models to such data since the early 1970s [3, 4]. For models based on Dirichlet processes, posterior inference is possible via Markov chain Monte Carlo (MCMC) methods [5]. However, for many modern model choices, such MCMC based methods fail [6, 7, 8]. Here we present a new importance sampling algorithm for nonparametric models given exchangeable binary response data [9]. Unlike approximate methods, this algorithm converges to the true posterior distribution, and also yields a consistent estimator for the marginal likelihood. It can be applied to any model from which samples can be generated, or even approximately generated. The main idea behind the algorithm is to exploit the symmetries introduced by exchangeability of the data, and then to correct for this exploitation by multiplying by an appropriate importance weight. Calculating the weights turns out to be equivalent to evaluating the permanents of a certain class of 0-1 matrices, which we prove can be done in polynomial time by deriving an explicit algorithm. We apply the new technique to both synthetic and real data.

Bibliography

[1] Keiding N., Begtrup K., Scheike T. H., Hasibeder G., *Estimation from current-status data in continuous time*, Lifetime Data Analysis, 2 (1996), 119–129.

[2] Dixon W. J., Mood A. M., *A method for obtaining and analyzing sensitivity data*, Journal of the American Statistical Association, 43:241 (1948), 109–126.

[3] Antoniak C. E., *Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems*, The Annals of Statistics, 2:6 (1974), 1152–1174.

[4] Bhattacharya P. K., *Posterior distribution of a Dirichlet process from quantal response data*, The Annals of Statistics, 9:4 (1981), 803–811.

[5] Escobar M. D., West M., *Bayesian density estimation and inference using mixtures*, Journal of the American Statistical Association, 90:430 (1995), 577–588.

[6] Hjort N. L., Walker S. G., *Quantile pyramids for Bayesian nonparametrics*, The Annals of Statistics, 37:1 (2009), 105–131.

[7] Lijoi A., Prünster I., Rigon T., *The Pitman-Yor multinomial process for mixture modelling* Biometrika, 107:4 (2020), 891–906.

[8] Lijoi A., Prünster I., Rigon T., *Finite-dimensional discrete random structures and Bayesian clustering*, Journal of the American Statistical Association, advance online publication (2023).

[9] Christensen D., *Inference for Bayesian nonparametric models with binary response data via permutation counting*, Bayesian Analysis, advance online publication (2022).

Small Total-Cost Constraints in Contextual Bandits with Knapsacks, with Application to Fairness

Tuesday, 4.07
15:00–15:30
IS 08
Room 0.06

Evgenii Chzhen[1], **Christophe Giraud**[1], **Zhen Li**[2], **and Gilles Stoltz**[1]

[1] Université Paris-Saclay, LMO, CNRS
[2] BNP Paribas

Abstract

We consider contextual bandit problems with knapsacks [CBwK], a problem where at each round, a scalar reward is obtained and vector-valued costs are suffered. The learner aims to maximize the cumulative rewards while ensuring that the cumulative costs are lower than some predetermined cost constraints. We assume that contexts come from a continuous set, that costs can be signed, and that the expected reward and cost functions, while unknown, may be uniformly estimated – a typical assumption in the literature. In this setting, total cost constraints had so far to be at least of order $T^{3/4}$, where T is the number of rounds, and were even typically assumed to depend linearly on T. We are however motivated to use CBwK to impose a fairness constraint of equalized average costs between groups: the budget associated with the corresponding cost constraints should be as close as possible to the natural deviations, of order $\sqrt{T}$. To that end, we introduce a dual strategy based on projected-gradient-descent updates, that is able to deal with total-cost constraints of the order of $\sqrt{T}$ up to poly-logarithmic terms. This strategy is more direct and simpler than existing strategies in the literature. It relies on a careful, adaptive, tuning of the step size.

Long Story Short: Omitted Variable Bias in Causal Machine Learning

Friday, 7.07
11:00–11:30
IS 11
Room 0.03a

Victor Chernozhukov[1], **Carlos Cinelli**[2], **Whitney Newey**[1], **Vasilis Syrgkanis**[3]

[1] MIT (USA)
[2] University of Washington, Seattle (USA)
[3] Microsoft Research (USA)
[4] Stanford (USA)

Abstract

We derive general, yet simple, sharp bounds on the size of the omitted variable bias for a broad class of causal parameters that can be identified as linear functionals of the conditional expectation function of the outcome. Such functionals encompass many of the traditional targets of investigation in causal inference studies, such as, for example, (weighted) average of potential outcomes, average treatment effects (including subgroup effects, such as the effect on the treated), (weighted) average derivatives, and policy effects from shifts in covariate distribution—all for general, nonparametric causal models. Our construction relies on the linear (Riesz) representer of the target functional. Specifically, we show how the bound on the bias depends only on the additional variation that the latent variables create both in the outcome and in the representer for the parameter of interest. In leading examples, this bound is further shown to depend on easily interpretable and familiar quantities, such as the gains in variation explained in the outcome and the treatment (or other interpretable quantities) created by latent confounders. Therefore, simple plausibility judgments on the maximum explanatory power of latent confounders are sufficient to place overall bounds on the size of the bias. Finally, we develop new flexible and efficient statistical inference methods on the learnable components of the bounds, based on debiased machine learning. We emphasize that all of the above results are new for all of the leading special and general cases, including average causal effects in nonparametric models with a binary treatment, average causal derivatives in nonparametric models with a continuous treatment, and average causal effects in partially linear models with an arbitrary treatment.

Estimation of the complier causal hazard ratio under dependent censoring

Thursday, 6.07
17:00–17:20
CS 27
Room 1.40

Gilles Crommen[1], **Jad Beyhum**[2], **Ingrid Van Keilegom**[1]

[1] ORSTAT, KU Leuven, Naamsestraat 69, Leuven, 3000, Belgium
[2] CREST, ENSAI, 51 Rue Blaise Pascal, Bruz, 35170, France

Abstract

In this paper, we are interested in studying the causal effect of an endogenous binary treatment on a dependently censored duration outcome. By dependent censoring, it is meant that the duration time (T) and censoring time (C) are not statistically independent of each other, even after conditioning on the measured covariates. The endogeneity issue is handled by making use of a binary instrumental variable for the treatment. To deal with the dependent censoring problem, it is assumed that on the stratum of compliers: (i) T follows a semiparametric proportional hazards model; (ii) C follows a fully parametric model; and (iii) the relation between T and C is modeled by a parametric copula, such that the association parameter can be left unspecified. In this framework, the treatment effect of interest is the complier causal hazard ratio (CCHR). We devise an estimation procedure that is based on a weighted maximum likelihood approach, where the weights are the probability of an observation being a complier. The weights are estimated nonparametrically in a first stage, followed by the estimation of the CCHR. Conditions under which the model is identifiable are given, a two-step estimation procedure is proposed, and some important asymptotic properties are established. Simulations are used to assess the validity and finite-sample performance of the estimation procedure.

Accurate bias estimation with applications to focused model selection

Tuesday, 4.07
12:20–12:40
CS 06
Room 1.40

Ingrid Dæhlen, Nils Lid Hjort and Ingrid Hobæk Haff

University of Oslo (Norway)

Abstract

The mean squared error (MSE) of an estimator decomposes into the sum of its variance and squared bias. For many of commonly used estimators, the former is of order $O(1/n)$. Hence, if we want to approximate the mean squared error, we need to make sure that the estimators have errors of smaller orders.

Lately, such approximations have been sought and developed in the literature concerning the focused information criterion, see e.g. [1, 2, 3]. Our work corrects the estimators of [3], which can be shown to have errors of order $O(1/n)$, the same order as the variance of the estimator whose MSE we are trying to approximate. This is achieved by deriving precise estimators of the bias and squared bias. Our results hold for a large class of estimators, and in particular, we derive an expression for and estimator of the bias of maximum likelihood estimators in (possibly) misspecified parametric models. This result can be seen as a novel and model-agnostic extension of the results in [4].

Lastly, we define a new and precise focused information criterion and illustrate the methods on a data set containing the number of battle deaths in the 95 most recent and concluded inter-state wars. The example illustrates the potentially large impact of using a less precise estimator of the MSE.

Bibliography

[1] Claeskens, G. and Hjort, N. The focused information criterion, *Journal of the American Statistical Association*, 98, (2003) 900-916.

[2] Claeskens, G., Cunen, C. and Hjort, N. Model selection via focused information criteria for complex data in ecology and evolution, *Frontiers in Ecology and Evolution*, 7, (2019) 415.

[3] Jullum, M. and Hjort, N.L. Parametric or nonparametric: The FIC approach, *Statistica Sinica*, 27, (2017) 951-981.

[4] Cox, D. R. and Snell, E. J. A general definition of residuals, *Journal of the Royal Statistical Society. Series B (Methodological)*, 30, (1968) 248-275.

Statistical Inference for Rough and Persistent Volatility

Friday, 7.07
14:00–14:20
CS 35
Room 0.06

Camilla Damian[1], **Rüdiger Frey**[2], **Stefan Voigt**[3]

[1] TU Wien (Austria)
[2] WU Vienna University of Economics and Business (Austria)
[3] University of Copenhagen (Denmark)

Abstract

In previous work, we considered a filtering and parameter estimation strategy in a rough volatility model for high-frequency data based on a Markovian representation of fractional (resp. Liouville) Brownian motion. However, frameworks based on processes of this type are not suitable for capturing other stylized facts associated with volatility (such as long memory): they are not stationary and, because of the self-similarity property, they cannot account for both roughness and long-term persistence.

In this work, in the spirit of [1], we devise an alternative model that is rich enough to accommodate both effects, while preserving a structure suitable for similar filtering and estimation strategies. We consider a setup where observations are modeled as a doubly-stochastic Poisson process, while the state process can be approximated with a finite-dimensional superposition of Ornstein-Uhlenbeck processes and is constructed in such a way that sample paths exhibit roughness whereas the autocorrelation function decays polynomially.

This framework allows for a more flexible parametrization and provides a foundation for a satisfactory and thorough empirical analysis of real tick-by-tick data. In particular, given their interpretation and role in the model, we provide different estimation strategies for different parameters: while those governing the long-term behavior of signal and observations can be inferred via moment methods, particle-based methods are more suitable for those controlling the roughness of sample paths (essentially as they are very sensitive to the unoberservability of the volatility process).

Bibliography

[1] Bennedsen M., Lunde A., Pakkanen M., *Decoupling the Short- and Long-Term Behavior of Stochastic Volatility*, Journal of Financial Econometrics, 20 (2022), 961–1006.

Doubly-Inflated Poisson INGARCH Models for Count Time Series

Thursday, 6.07
12:00–12:20
CS 22
Room 5440

Sumen Sen[1], **Ishapathik Das**[2], **Fathima Ayoob**[2]

[1] Department of Mathematics and Statistics, Austin Peay State University, Clarksville, TN (USA)
[2] Department of Mathematics and Statistics, Indian Institute of Technology Tirupati, Tirupati, AP (India)

Abstract

Researchers usually model count time series data assuming the responses follow the Poisson distribution. The Poisson distribution may be adequate to fit the count time series data when there is no inflated point in the data, i.e., some observation is more likely to occur than it should happen for Poisson models. The zero-inflated Poisson models are well studied in literature [1, 2] when the number of zeros is unexpectedly higher than the other observations in the data. Many count time-series observations may contain more than one inflated point in the data [3], which zero-inflated Poisson models may not nicely fit. Here we propose doubly-inflated Poisson INGARCH models to accommodate the count time-series observations having two inflation points in the data. The stationarity conditions and the autocorrelation functions are derived for the corresponding models. The model's unknown parameters are estimated using the maximum likelihood estimation (MLE) method and the expectation-maximization (EM) algorithm. The simplified steps for estimating the model's parameters using both methods are provided. The proposed methodologies are illustrated using simulated data and applied to real data.

Bibliography

[1] Zhu, F., *Zero-inflated Poisson and negative binomial integer-valued GARCH models*, ournal of Statistical Planning and Inference, 142 (2012), 826-839.

[2] Lee, S., Kim, D. and Seok, S., *Modeling and inference for counts time series based on zero-inflated exponential family INGARCH models*, Journal of Statistical Computation and Simulation, 91 (2021), 2227-2248.

[3] Das, I., Sen, S., Chaganty, N.R. and Sengupta, P., *Regression for doubly inflated multivariate Poisson distributions*, Journal of Statistical Computation and Simulation, 89 (2019), 2549-2561.

Latent variables in measurement error problems with excess zeros

Tuesday, 4.07
14:00–14:30
IS 18
Room 0.03

Aurore Delaigle

University of Melbourne (Australia)

Abstract

In the classical measurement error problem, we observe a contaminated version W of a unobservable latent variable of interest X, where W and X are related through $W = X+U$, with X and U independent and U an unobserved measurement error which is independent of X. In this talk we are interested in the case where the latent variable of interest represents the long-term average of a quantity that changes every day such as activity level, nutrient intake or exposure to a toxic substance. There, we observe repeated contaminated versions of X, that is, we observe $W_{i,j}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$, where n is the number of individuals in the sample J is the number of replicates per individual. In cases where the quantity of interest corresponds to an intermittent phenomenon, $W_{i,j}$ is equal to zero if the jth replicate is collected on a day where the phenomenon is not present. After transformation ignored here for simplicity, these data are often modelled by a two-part model of the type

$$W_{i,j} = X_i + U_{i,j} \quad \text{if } W_{i,j} > 0,$$
$$P(W_{i,j} > 0|X_i) = H(X_i),$$

where H is an increasing function, and the latent variable of interest is defined by $T_i = E(W_{i,j}|X_i)$.

I will discuss semiparametric and nonparametric approaches to estimating the density f_T of T and illustrate the method on EATS food consumption data. This is joint work with Felix Camirand Lemyre and Raymond Carroll.

Universal densities for stationary processes

Wednesday, 5.07
11:20–11:40
CS 17
Room 5440

Łukasz Dębowski

Polish Academy of Sciences (Poland)

Abstract

A prefix-free code is called universal if its length divided by the sample length is a strongly consistent estimator of the entropy rate in a given class of stationary ergodic sources. It is known that universal codes exist for sources over a finite alphabet but not over a countably infinite one. We generalize the problem of universal coding as the problem of universal densities with respect to a given reference measure on a countably generated measurable space. By definition, universal densities consistently estimate the differential entropy rate of any stationary ergodic source with a finite differential entropy rate.

We show constructively that universal densities exist if the reference measure is finite. To exhibit a universal density, we combine the prediction by partial matching (PPM) code with the recently proposed non-parametric differential (NPD) entropy rate estimator. The PPM code approximates a finite-alphabet source as a Markov chain of a high order, whereas the NPD estimator approximates a non-discrete source as a finite-alphabet one. To enforce universality of this construction, we put a prior over both countably many Markov orders and countably many quantization levels. The proof of universality applies Barron's asymptotic equipartition for densities and continuity of f-divergences.

As an application, we show that any universal density induces a strongly consistent Cesàro mean estimator of the conditional density given an infinite past. In turn, this yields a universal predictor with the $0-1$ loss for a countable alphabet. We also exhibit a strongly consistent entropy rate estimator with respect to the Lebesgue reference measure in the class of stationary ergodic Gaussian processes.

The full paper is available as preprint [1].

Bibliography

[1] Ł. Dębowski. Universal densities exist for every finite reference measure. Preprint `https://arxiv.org/abs/2209.11981`, under review in IEEE Transactions on Information Theory, 2022.

On Lasso and Slope drift estimators for Lévy-driven Ornstein–Uhlenbeck processes

Monday, 3.07
15:45–16:05
CS 03
Room 2180

Niklas Dexheimer[1], **Claudia Strauch**[1]

[1] Aarhus University (Denmark)

Abstract

We investigate the problem of estimating the drift parameter of a high-dimensional Lévy-driven Ornstein–Uhlenbeck process under sparsity constraints. It is shown that both Lasso and Slope estimators achieve the minimax optimal rate of convergence (up to numerical constants), for tuning parameters chosen independently of the confidence level, which improves the previously obtained results for standard Ornstein–Uhlenbeck processes.

Copula based quantile modelling under dependent censoring

Wednesday, 5.07
12:20–12:40
CS 18
Room 4420

Myrthe D'Haen[1,2], **Ingrid Van Keilegom**[2], **Anneleen Verhasselt**[1]

[1] CenStat, DSI, Hasselt University (Belgium)
[2] ORSTAT, KU Leuven (Belgium)

Abstract

In survival analysis, random right censoring may cause part of the observations corresponding to censoring times C instead of survival times T. Whereas the censoring is often assumed to be independent, this assumption is not always realistic in practice. Therefore, to obtain more accurate inference for the survival time, we propose a quantile model for T that is part of a bivariate model for (T, C), capturing the possible dependency using a copula function.

For the copula model, we use recent results of Czado and Van Keilegom [1], showing that complete specification of the copula (including its association parameter) is not necessary to guarantee identifiability of the model, as opposed to many copula models in current literature. This comes at the cost of parametric marginal distributions. However, this flexibility reduction on the level of the marginals we overcome by using the family of so-called *enriched asymmetric Laplace* (EAL) distributions. Starting from an asymmetric Laplace distribution, orthonormal Laguerre polynomials are included. While maintaining the parametric nature of the marginals, the EAL family also yields sufficient flexibility. Inspired by the proofs in Czado and Van Keilegom [1], we can prove identifiability of the parameters in our model. As this in particular comprises all parameters for the survival time, inference is actually done not only on T's quantiles, but on its entire distribution.

Bibliography

[1] Czado, C., and Van Keilegom, I., *Dependent censoring based on parametric copulas*, Biometrika (2022), https://doi.org/10.1093/biomet/asac067.

Mapper algorithms – how to visualize your data, even if they are high dimensional

Friday, 7.07
11:00–11:20
CS 33
Room 5440

Paweł Dłotko

Abstract

It is well known that various statistics are often not sufficient to tell apart various samples. Two dimensional examples like Anscombe's quartet and the recent Datasaurus dataset showed samples of samples with the same summary statistics yet very different 'shape'. In this talk I will present tools of topological data analysis, most notably mapper type algorithms and cluster graph techniques that allow encapsulating the shape of high dimensional samples. Their construction and sample used-cases will be discussed.

From Denoising Diffusion Models to Dynamic Transport Models – Generative Modeling and Inference

Friday, 7.07
09:30–10:30
Plenary Lecture
Room 0.03

Arnaud Doucet

Google DeepMind

Abstract

Denoising diffusion models are a novel powerful class of techniques for generative modeling and inference. These models have superseded generative adversarial networks over the past two years as they are flexible, easy to train and provide state-of-the-art results in numerous application domains such as image synthesis and protein design. In this talk, we will review these methods, illustrate them on a variety of applications and discuss their limitations. We will then show how recent alternative techniques based on dynamic mass transport ideas can resolve some of these limitations. In particular, we will focus on Schrödinger bridges, an entropy-regularized version of the dynamic optimal transport, and present a novel simple method to approximate them numerically.

Propensity Weighting in the Estimation of Direct Effects

Tuesday, 4.07
16:00–18:00
Poster

Christiana Drake[1], **Julie Smith-Gagen**[2]

[1] University of California, Davis (USA)
[2] University of Nevada, Reno (USA)

Abstract

We investigate propensity weighting in assessing direct effects in a model where treatment may be mediated by another risk factor. In particular, we compare two approaches to estimating direct effects, counterfactual approaches as discussed by Vanderweele [1] and principal stratification as suggested by Frangakis and Rubin [2], [3]. We demonstrate the ideas via simulation studies and apply the method to a study of Cancer of Unknown Primary (CUP). The exposure variable is Cancer of Unknown Primary (CUP) and the mediator is treatment. A CUP is con- firmed after a set of recommended tests are performed but a primary cancer is not found. A CUP is unconfirmed if the tests are not done. The outcome is survival beyond a fixed time point. A direct effect of a CUP diagnosis would suggest that CUP is directly affecting survival, irrespective of treatment. An indirect effect would be survival by modified by the treatment.

Bibliography

[1] Vanderweele., *Explanation in Causal Inference: Methods for Mediation and Interaction*, Oxford University Press, 2015.
[2] Frangakis CE and Rubin DB, *Principal Stratification in Causal Inference*, Biometrics, 58 (2002), 21-29.
[3] Rubin DB, *Direct and Indirect Causal Effects via Potential Outcomes*, Scandinavian Journal of Statistics, 31 (2004), 161-170

Affine-equivariant inference for multivariate location under L_p loss functions

Thursday, 6.07
12:00–12:20
CS 20
Room 1.40

Alexander Dürre[1], **Davy Paindaveine**[2]

[1] Mathematical Institute, Leiden University (Netherlands)
[2] ECARES and Department of Mathematics, Université libre de Bruxelles (Belgium)

Abstract

We consider the fundamental problem of estimating the location of a d-variate probability measure under an L_p loss function. The naive estimator, that minimizes the usual empirical L_p risk, has a known asymptotic behavior but suffers from several deficiencies for general p, the most important one being the lack of equivariance under affine transformations. We propose a collection of L_p location estimators that minimize the size of suitable ℓ-dimensional data-based simplices. For $\ell = 1$, these estimators reduce to the naive ones, whereas, for $\ell = d$, they are equivariant under affine transformations. Irrespective of ℓ, these estimators reduce to the sample mean for $p = 2$, whereas for $p = 1$, the estimators provide the well-known spatial median and Oja median for $\ell = 1$ and $\ell = d$, respectively.

Under very mild assumptions, we derive an explicit Bahadur representation and establish asymptotic normality for the new estimators. To allow for large sample size n and/or large dimension d, we introduce a version of our estimators relying on incomplete U-statistics. We also define related location tests and derive explicit expressions for the asymptotic power under contiguous local alternatives.

Moment equations for homographic difference equations with random structure

Monday, 3.07
17:05–17:25
CS 04
Room 5440

Irada Dzhalladova

Kyiv National Economic University (Ukraine)

Abstract

Homographic difference equation with coefficients dependent on Markov process is studied. To investigate the behaviour of its solutions, equations of appropriate moments of the first order are derived and analysed. The set of states is finite. Using stochastic approach, one can investigate a number of aspects relating to a variety of phenomena in finance and economics.The paper constructs mathematical models of nonlinearity and perturbable. On the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we consider the first order difference equation with coefficients dependent on Markov process in the form

$$x_{n+1} = \frac{B(\xi)}{x_n}, \quad n = 1, 2, \dots, \tag{1}$$

ξ is the Markov chain of the two states θ_1, θ_2 with the probabilities $p_k = P\{\xi = \theta_k\}$, $k = 1, 2$, that satisfy the system of difference equations

$$\begin{aligned} p_1(n+1) &= (1-\lambda)p_1(n) + \nu p_2(n), \\ p_2(n+1) &= \lambda p_1(n) + (1-\nu)p_2(n), \end{aligned} \tag{2}$$

where $1-\lambda, \nu, \lambda, 1-\nu \in (0,1)$ are elements of transition matrix π_{ks} in Markov chain. The state x_n, $n = 1, 2, \dots$ is called a solution of system (4) within the meaning of a strong solution if it satisfies (4) [2].

Definition 1. The function

$$\begin{aligned} E^{(1)}\{b_n\} &= \sum_{k=1}^{2} E_k^{(1)}\{b_n\} \\ E_k^{(1)}\{b_n\} &= \int_{\mathbb{R}} z f_k(n, z)\, dz, \qquad k = 1, 2, \end{aligned} \tag{3}$$

is called moment of the first order for a solution b_n, $n = 1, 2, \dots$ of (4). The values $E_k^{(1)}\{b_n\}$, $k = 1, 2$ are called particular moments of the first order[2].

Theorem. *System of moment equations of the first order for a solution z_n, $n = 1, 2, \dots$ of (4) is of the form*

$$E_k^{(1)}\{b_{n+1}\} = \sum_{s=1}^{2} \pi_{ks} p_s(n) A_k E_s^{(1)}\{b_n\} \tag{4}$$

where k, s - numbers of states of Markov chain ξ_n.

Bibliography

[1] I. A. Dzhalladova: *Optimization of Stochastic System*, KNEU, Kiev, 2005, Ukraine.
[2] M. Ružičková, I. A. Dzhalladova: The optimization of Solutions of the Dynamic Systems with Random Structure, *Abstr. Appl. Anal.* **2011**, Article ID 486714, 18 pages.

Plugin Machine Learning for Treatment Effect Estimation on Networks

Friday, 7.07
11:20–11:40
CS 30
Room 1.01

Corinne Emmenegger[1], **Meta-Lina Spohn**[1], **Timon Elmer**[2], **Peter Bühlmann**[1]

[1] Seminar for Statistics, ETH Zurich (Switzerland)
[2] Department of Humanities, Social and Political Sciences, ETH Zurich (Switzerland)

Abstract

Causal inference methods for treatment effect estimation usually assume independent units. However, this assumption is often questionable because units may interact, resulting in spillover effects between units. We develop augmented inverse probability weighting (AIPW) for estimation and inference of the direct effect of the treatment with observational data from a single (social) network with spillover effects. We use plugin machine learning and sample splitting to obtain a semiparametric treatment effect estimator that converges at the parametric rate and asymptotically follows a Gaussian distribution. We apply our AIPW method to the Swiss StudentLife Study data to investigate the effect of hours spent studying on exam performance accounting for the students' social network.

Nonparametric density estimation of small jumps of Lévy processes

Tuesday, 4.07
12:00–12:30
IS 07
Room 0.06

Duval Céline[1], **Jalal Taher**[2], **Mariucci Ester**[2]

[1] Université de Lille (France)
[2] Université Paris-Saclay (UVSQ) (France)

Abstract

We consider the problem of estimating the density of the process associated with the small jumps of a pure jump Lévy process, from discrete observations of one trajectory. We discuss results both from low and high frequency observations, for Lévy processes possibly of infinite variation.

In a low frequency setting, we propose an estimator obtained via a spectral approach which achieves the minimax rate with respect to the L_2-loss.

In a high frequency setting we propose two different estimators, a kernel estimator and one based on a spectral approach, which achieve the same rate of convergence. This rate depends on the sampling scheme and on the Blumenthal-Getoor index of the process. Finally, we discuss the optimality of these results.

Goodness-of-Fit Tests for Spatial Point Processes

Monday, 3.07
15:45–16:15
IS 04
Room 0.06

Chiara Fend, Claudia Redenbach

RPTU Kaiserslautern-Landau (Germany)

Abstract

In spatial point process statistics most research focuses on defining new parametric models, studying their properties and establishing estimation methods for their parameters. Nowadays many fitting procedures are available that can be used by everyone. This makes fitting arbitrary models to some observed point pattern quite easy. As a consequence, the task of statistical model validation of a fitted point process model becomes more and more important.

The aspect of model validation that we investigate here is goodness-of-fit testing. In recent years several new approaches for goodness-of-fit tests in the setting of spatial point processes have been introduced. The latest contributions propose either new functional summary statistics, e.g. originating from topological data analysis [1], or new test procedures such as global envelope tests [3].

Furthermore, scoring rules such as the Continuous Ranked Probability Score have been introduced for spatial point processes [2] for quantifying the fit of a proposed model. These scoring rules can then also be incorporated into the goodness-of-fit test setting.

In this talk we consider a single realization x_0 of a point process $X \sim P$ on $\mathbb{R}^d, d > 1$. We review both classical and recent test approaches for a simple hypothesis of a fully specified model P_0 and for composite hypotheses of the form

$$H_0 : P \in \mathcal{P}_\Theta = \{P_\theta \mid \theta \in \Theta \subset \mathbb{R}^p, p \geq 1\} \quad \text{vs.} \quad H_1 : P \notin \mathcal{P}_\Theta.$$

We discuss novel ways of how one can combine multiple aspects from different test statistics into combined goodness-of-fit tests.

A power study for planar stationary and isotropic point processes compares the robustness of the tests with respect to parameters such as the subset of the domain of the functional summary statistic that is taken into account, the size of the observation window or the number of simulations under the null hypothesis in case of Monte Carlo tests.

Bibliography

[1] Biscio, C.A.N., Chenavier, N., Hirsch, C., Svane, A.M., *Testing goodness of fit for point processes via topological data analysis*, Electronic Journal of Statistics, 14(1) (2020), 1024–1074.

[2] Heinrich-Mertsching, C., Thorarinsdottir, T.L., Guttorp, P., Schneider, M., *Validation of point process predictions with proper scoring rules*, arXiv preprint arXiv:2110.11803, 2021.

[3] Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., Hahn, U., *Global envelope tests for spatial processes*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(2) (2017), 381–404.

Sparse M-estimators in semi-parametric copula models

Wednesday, 5.07
11:20–11:40
CS 15
Room 1.40

Jean-David Fermanian[1], **Benjamin Poignard**[2]

[1] Ensae-Crest, Palaiseau, (France)
[2] Osaka University (Japan)

Abstract

We study the large sample properties of sparse M-estimators in the presence of pseudo-observations. Our framework covers a broad class of semi-parametric copula models, for which the marginal distributions are unknown and replaced by their empirical counterparts. It is well known that the latter modification significantly alters the limiting laws compared to usual M-estimation. We establish the consistency and the asymptotic normality of our sparse penalized M-estimator and we prove the asymptotic oracle property with pseudo-observations, possibly in the case when the number of parameters is diverging. Our framework allows to manage copula-based loss functions that are potentially unbounded. Additionally, we state the weak limit of multivariate rank statistics for an arbitrary dimension and the weak convergence of empirical copula processes indexed by maps. We apply our inference method to Canonical Maximum Likelihood losses with Gaussian copulas, mixtures of copulas or conditional copulas. The theoretical results are illustrated by two numerical experiments.

Bibliography

[1] Berghaus B., Bücher A., Volgushev S., *Weak convergence of the empirical copula process with respect to weighted metrics*, Bernoulli, 23 (2017), 743-772.
[2] Fan J., Li R., *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American statistical Association, 96 (2001), 1348-1360.
[3] Fan J., Peng H., *Nonconcave penalized likelihood with a diverging number of parameters,* The Annals of Statistics, 32 (2004), 928-961.
[4] Radulović D., Wegkamp M., Zhao Y., *Weak convergence of empirical copula processes indexed by functions,* Bernoulli, 23 (2017), 3346-3384.

Iterative regularization methods for ill-posed generalized linear models

Tuesday, 4.07
14:40–15:00
CS 11
Room 2180

Gianluca Finocchio, Tatyana Krivobokova

University of Vienna (Austria)

Abstract

We study the problem of expected-log-likelihood optimisation in ill-posed generalised linear models (GLMs) where only some projections of the covariates are relevant for the response. The relevant features identify a possibly low-dimensional relevant subspace for the response. It is assumed that the source of ill-posedness is a joint low-dimensionality of the response and the relevant covariates in the sense of a latent factor GLM. We propose a novel iteratively-reweighted-partial-least-squares (IRPLS) algorithm and show that it is better than any other regularisation algorithm, based on projection or penalisation, in estimating the relevant subspace. Under regularity assumptions on the latent factor GLM, we provide convergence rates of the IRPLS estimator with high probability and show that, in may special cases, they nearly match those of the maximum likelihood estimator in the latent factor GLM, which is an oracle achieving an optimal parametric rate. Our findings are confirmed by numerical studies.

Time series emulation of carbon cycle model simulations

Tuesday, 4.07
16:00–18:00
Poster

Nina Fischer[1], **Amy Wilson**[1], **Chris Dent**[1], **Mathew Williams**[2]

[1] University of Edinburgh, School of Mathematics (UK)
[2] University of Edinburgh, School of Geosciences (UK)

Abstract

Simulators, complex physical models implemented in computer code, are a fundamental tool to assess climate change and terrestrial carbon dynamics. Coupled with observational data from satellites, models can infer unobserved ecosystem properties. However, their computational cost typically restricts the number of possible system runs. This limits our understanding of landscapes' complex physical processes and associated uncertainties, hindering mitigation and adaptation strategies essential to meet climate change mitigation targets.

To facilitate quick model calibration and inference, this talk presents statistical methodology to overcome computational constraints by building emulators, or statistical approximations of the simulators. We combined these emulators with history matching to find the set of input combinations for which the simulation gives acceptable matches to observed data. In this work, we applied this to the carbon cycle model DALEC Crop, which models the effect of Nitrogen fertilisation on wheat yield in a field over a crop growing season. We applied dimension reduction techniques to emulate Leaf Area Index (LAI) time series and then history matched predictions to Sentinel-2 LAI observations. We then used the input space found in a DALEC Crop forward run to produce plausible yield predictions, skipping DALEC's computationally costly model calibration process, to demonstrate the potential to reduce computational costs by using suitable statistical methodology.

Prediction-based uncertainty quantification in learning models

Tuesday, 4.07
14:00–14:20
CS 09
Room 1.01

Sandra Fortini, Sonia Petrone

Bocconi University (Italy)

Abstract

Prediction plays a central role in the foundations of Bayesian statistics, and it has become the main focus in many areas of machine learning, as opposed to the more traditional focus on inference. In this talk, we discuss how uncertainty expressed by the posterior distribution and credible intervals can be understood in terms of prediction within a Bayesian framework. The posterior law of the unknown distribution is centered on the predictive distribution, and we prove that it is marginally asymptotically Gaussian with the variance depending on the predictive updates, that is on how the predictive rule incorporates information as new data become available. This approach allows to obtain asymptotic credible intervals only based on the predictive rule (without having to specify the model and the prior law), sheds light on frequentist coverage as related to the predictive learning rule, and, we believe, opens a new perspective towards a notion of *predictive efficiency* that seems to call for further research. A recent reference is [1].

Bibliography

[1] Fortini, S., Petrone, S. *Prediction-based uncertainty quantification for exchangeable sequences*, Phil. Trans. R. Soc. A, Volume 381, Issue 2247 (2023).

GLARMA model and Bootstrap approaches: An application to respiratory diseases and air pollutants

Tuesday, 4.07
16:00–18:00
Poster

Ana Júlia A. Câmara[1], Valdério A. Reisen[2], Glaura Franco[1], and Pascal Bondon[3]

[1] Universidade Federal de Minas Gerais (Brazil)
[2] Universidade Federal da Bahia (Brazil)
[3] Université Paris-Saclay (France)

Abstract

The generalized linear autoregressive moving average (GLARMA) model has been used in epidemiological studies to evaluate the impact of air pollutants in human health, as frequently the response variable is a nonnegative integer-valued time series. These health effects are quantified through the relative risk (RR) measure. Inference for the RR can be based on the asymptotic properties of the maximum likelihood estimator for the GLARMA parameters, which can be troublesome if the series sizes are not very large. Thus, this work proposes to study different types of bootstrap confidence intervals (CI) for the RR calculated from the GLARMA model. An extensive simulation study was performed, revealing that the model parameter which accounts for the autocorrelation present in the data can influence the coverage of the intervals. The numerical simulations also showed that problems can arise if the covariates present an autocorrelation structure since the CI coverage is impacted even for stationary time series. To solve this problem, the use of the vector autoregressive filter (VAR) in the covariates is suggested.

A non-parametric construction and learning method of a graphical model

Thursday, 6.07
11:20–11:40
CS 19
Room 1.01

Konrad Furmańczyk

Institute of Information Technology, Warsaw University of Life Sciences (Poland)

Abstract

We present a non-parametric construction of a graphical model. Our construction contains an undirected graph that represents conditional independence for general random variables defined by the conditional dependence coefficient ([1]). In this construction the set of edges are defined as $E = \{(i,j) : R_{i,j} \neq 0\}$, where $R_{i,j}$ is a the conditional dependence coefficient for X_i and X_j given $(X_1, \ldots, X_p) \backslash \{X_i, X_j\}$.

Graph structure learning is proposed by two steps selection procedure:

a) first, we compute the matrix of sample version of the conditional dependence coefficient $\widehat{R_{i,j}}$;
b) next, for some prespecificated threshold $\lambda > 0$ we choose an edge $\{i,j\}$ if

$$max\left\{\left|\widehat{R_{i,j}}\right|, \left|\widehat{R_{j,i}}\right|\right\} \geq \lambda.$$

Under some mild assumptions we present the result of consistency model selection.

Proposed graph recovery structure has been evaluated on artificial and real datasets. We also applied our graph recovery procedure for learning partial correlation graphs for the elliptical distribution.

Bibliography

[1] Azadkia M., Chatterjee S. *A simple measure of conditional dependence*, The Annals of Statistics 49 (2021), 3070-3102.
[2] Furmańczyk K. *A construction of a graphical model*, 2022 Preprint.
[3] Lauritzen, S. *Graphical Models*, Clarendon Press, 1996.
[4] Rossell D., Zwiernik P. *Dependence in Elliptical Partial Correlation Graphs*, Electron. J. Statist. 15 (2021), 4236-4263.

On the role of phase type distributions in population genetics

Monday, 3.07
16:45–17:05
CS 03
Room 2180

Asger Hobolth[1], **Iker Rivas-González**[2], **Mogens Bladt**[3], **Andreas Futschik**[4]

[1] Department of Mathematics, Aarhus University (Denmark)
[2] Bioinformatics Research Center, Aarhus University (Denmark)
[3] Department of Mathematical Sciences, University of Copenhagen (Denmark)
[4] Institute of Applied Statistics, Johannes Kepler University(Austria)

Abstract

A phase-type distribution describes the time to absorption in a continuous- or discrete-time Markov chain. Phase-type distributions have been successfully used in different fields of applied probability. Such applications, as well as an introduction to the theory of discrete and continuous phase type distributions can be found for instance in [1]. This talk intends to investigate how results from phase type theory can be applied to derive properties of the Kingman-coalescent process which is used to model the genealogy of samples of DNA sequences in population genetics. [3] observed that properties of the coalescent process can be derived in an alternative and elegant way using phase type theory. We will both review existing results and provide new ones. It turns out that features such as the time to the most recent common ancestor and the total branch length are phase-type distributed. Reward transformations lead to easy calculation of covariances and correlation coefficients between e.g. tree height, tree length, external branch length and internal branch length. Phase type calculations also provide an alternative approach to derive results from classical first step analysis ([2]).

A further result is that the site frequency spectrum follows a multivariate discrete phase-type distribution and the joint distribution of total branch lengths in the two-locus coalescent-with-recombination model is multivariate phase-type distributed. In general, phase-type distributions provide a powerful mathematical framework for coalescent theory because phase-type distributions are analytically tractable using matrix manipulations. Together with properly specified rewards, phase type distributions can also be used to derive likelihood functions for statistical inference. Providing an alternative to previous work based on the Laplace transform, we derive likelihoods for small size coalescent trees based on phase-type theory. Their application towards statistical inference is also explored.

Bibliography

[1] Bladt M., Nielsen B.F., Matrix-Exponential Distributions in Applied Probability. Springer, New York, 2017.
[2] Wakeley, J. Coalescent Theory: An Introduction. Roberts & Company Publishers, 2008.
[3] Hobolth A., Siri-Jegousse A., and Bladt M. Phase-type distributions in population genetics. Theoretical population biology, 127:16â€“32, 2019.

Applying Functional Data Analysis to periodically correlated time series

Friday, 7.07
12:00–12:20
CS 32
Room 2180

Elżbieta Gajecka-Mirek

Cracow University of Technology (Poland)

Abstract

The functional approach is becoming popular in modern statistics. It assumes that observations are functions, not random values. If we are dealing with very large datasets, we can treat the time series as a set of random functions. In practice, often, we deal with very numerous data that shows periodic behavior. In the presentation, the use of a functional autoregressive model for modeling periodic time series will be shown. More over the method of identifying the period of the data set will be presented. The periodic functional autoregressive model will be applied to periodic signals observed with high frequency which are bearing data.

Bibliography

[1] Horvath, L., Kokoszka, P., *Inference for Functional Data with Applications*, Springer Series in Statistics, 2012.

[2] Kidziński, L., Kokoszka, P., Neda, M.J., *Principal components analysis of periodically correlated time series*, Journal of Time Series Analysis, vol. 39, no 4 (2018), pp. 502-522.

[3] Kokoszka, P., Reimherr, M., *Asymptotic normality of the principal components of functional time series*, Stochastic Processes and their Applications, vol. 123 (2013), pp. 1546-1562.

[4] Leśkow, J., Skupień, M., *An application of functional data analysis to local damage detection*, Statistics in Transition, vol. 20, no 1 (2019), pp. 131-151.

Exact non-parametric inference for spatial point processes; with applications from Epidemiology to Ecology

Monday, 3.07
16:25–16:45
CS 01
Room 1.01

Dani Gamerman

Universidade Federal do Rio de Janeiro (Brazil)

Abstract

Point processes are one of the most commonly encountered observation processes in Spatial Statistics. Model-based inference for them depends on the likelihood function. In standard settings, the likelihood depends on the entirely unknown intensity function, and can not be computed analytically. A number of approximating techniques have been proposed to handle this difficulty. In this talk, we present recent work on exact solutions that solve this infinite-dimensional problem without resorting to approximations. This approach also accomodates for unobserved heterogeneity and regression components. The solution is based on model specifications with Gaussian processes that impose smoothness constraints on the intensity function. Details are provided in [7] and outlined in the presentation.

Applications to cardiological epidemiology and archaeoecology illustrate the results. In the first application, our approach is able to quantify the relevance of personal characteristics of an individual, his/her residential location and their interaction in their likelihood of death by a cardiovascular disease. In the second application, our approach provides and exact, model-based solution for the analysis of presence-only data in Ecology [2]. This approach allows for the prediction of the number of pre-Columbian earthworks still hidden in Amazonia and their possible relation with the current composition of the tropical forest (see [3]). Our approach was also extended in [4] to propose an exact model solution for preferential sampling in Geostatistics.

Extensions towards non-stationary specifications beyond Gaussian processes to handle more general intensity functions are under development. This setting is particularly useful to describe jumps and discontinuities, typically encountered in many applications. Time permitting, this topic will be briefly outlined.

This talk is based on joint work with Flavio Gonçalves, Guido Moreira, Izabel Nolau, Jony A. Pinto Jr and Douglas Silva.

Bibliography

[1] Gonçalves, F. B. and Gamerman, D., *Exact Bayesian inference in spatiotemporal Cox processes driven by multivariate Gaussian processes*, Journal of the Royal Statistical Society, Series B, 80 (2018), 157–175.

[2] Moreira, G. A. and Gamerman, D., *Analysis of presence-only data via exact Bayes, with model and effects identification*, Annals of Applied Statistics, 16 (2022), 1848–1867.

[3] Peripato, V., Levis, C., Moreira, G. A., Gamerman, D. et al., *Over 10,000 Pre-Columbian earthworks are still hidden throughout Amazonia*, submitted (2023).

[4] Silva, D. M. and Gamerman, D., *Exact Bayesian Geostatistics under Preferential Sampling*, submitted (2023).

On birth-death circuit chains in fixed environments: analysis of a generalized sample path case

Friday, 7.07
14:00–14:20
CS —
Room 1.40

Chrysoula Ganatsiou

University of Thessaly (Greece)

Abstract

In recent years a systematic research has been developed (Kalpazidou [1], MacQueen [10], Qian Minping and Qian Min [11] and others) in order to investigate representations of the finite-dimensional distributions of Markov processes (discrete or continuous parameter) having an invariant measure, as decompositions in terms of the *circuit passage functions*

$$J_c(i,j) = \begin{cases} 1, & \text{if i,j are consecutive states of c,} \\ 0, & \text{otherwise,} \end{cases}$$

for any directed sequence $c = (i_1, i_2, \ldots, i_\nu, i_1)$ of states called a *circuit*, $\nu > 1$ of the corresponding Markov process. This research has stimulated a motivation towards the representation of Markov processes through directed circuits and weights in terms of circuit passage functions in fixed or random environments as well as the study of specific problems associated with Markov processes in a different way. The representations are called *circuit representations* while the corresponding discrete or continuous parameter Markov chains generated by directed weighted circuits are called *circuit chains*.

By using the circuit representation theory of Markov processes we investigate the unique representations by directed circuits and weights of birth-death chains in fixed environments for a generalized sample path case ([3], [5]). Loosely speaking this is a process which combines the property of a random walk with reflection at zero and the discrete or continuous parameter nature of the transitions times ([2], [4], [6, 7, 8, 9]). This will give us the possibility to study suitable criteria regarding the properties of transience and recurrence for the corresponding *"adjoint"* Markov chains (*circuit chains*) describing uniquely the birth-death chains by directed weighted circuits in fixed environments for a generalized sample path case as well as important stochastic properties such as reversibility and entropy production ([12]).

Bibliography

[1] Kalpazidou S., *Cycle Represantations of Markov Processes*, Spriner, 1995.

[2] Ganatsiou Ch., *On cycle representations of random walks in fixed, random environments*, in Proceedings of the 58th World Congress of the International Statistical Institute, Dublin, 2011, pp. 4628–4631.

[3] Ganatsiou Ch., *On cycle representations of discrete-time birth-death processes*, in Proceedings of the 29th European Meeting of Statisticians, Budapest, 2013.

[4] Ganatsiou Ch., *On the study of transience and recurrence of the Markov chain defined by directed weighted circuits associated with a random walk in fixed environment*, J.Prob. Stat., 2013, Article ID 424601.

[5] Ganatsiou Ch. et al., *On discrete-time birth-death processes defined by directed weighted circuits*, JP. J. Biostatist. 11(2)(2014), 103–116.

[6] Ganatsiou Ch., *On the study of circuit chains associated with a random walk with jumps in fixed, random environments: criteria of recurrence and transience*, In: T. Rassias, P. Pardalos (eds.), Mathematical Analysis and Applications, Springer Optimization and Its Applications, vol. 154, Springer, 2019, pp. 185–203.

[7] Ganatsiou Ch., *On irreducible denumerable continuous parameter circuit chains: analysis of a generalized sample path case*, Quaestiones Mathematicae, 45(1) (2022), pp. 41–53.

[8] Ganatsiou Ch., *On the Approximation of Extinction Time for the Discrete-Time Birth-Death Circuit Chains in Random Environments*, In: Rassias T., Daras N. J. (eds), Approximation and Computation in Science and Engineering, Springer, 2022, pp. 333–348.

[9] Ganatsiou Ch., *"On the Expected Extinction Time for the Adjoint Circuit Chains associated with a Random Walk with Jumps in Random Environments"*, In: A. Nikeghbali, P. Pardalos, A. Raigorodskii, M. Th. Rassias (eds.), High Dimensional Optimization and Probability: With a view towards Data Sciences, Springer, 2022, pp. 219–239.

[10] MacQueen J. , *Circuit processes*, Ann. Probability, 9(4)(1981), 604–610.

[11] Qian Minping, Qian Min, *Circulation for recurrent Markov chains*, Z. Wahrsch. Verw. Gebiete, 59(2) (1982), 203–210.

[12] Qian Minping, Qian Min, *The reversibility and the entropy production of Markov processes*, Contemp. Math., 118(2) (1991), 255–261.

Scalable Hierarchical Bayesian Model for Multiple Sclerosis Progression using Brain Imaging and Clinical Features

Wednesday, 5.07
11:40–12:00
CS 16
Room 2180

Habib Ganjgahi[1,2], **Chris Holmes**[1]

[1] Statistics Department, University of Oxford (United Kingdom)
[2] Big Data Institute, University of Oxford (United Kingdom)

Abstract

Multiple sclerosis (MS) is a chronic and ultimately debilitating disease of the central nervous system that affects approximately 2.5 million individuals worldwide. MS subtypes have been defined only based on the clinical disease course of patients (relapse and physical discability), they reveal limitations in prognosticating long-term outcome, or in providing homogeneous treatment targets. However, there is growing evidence that MS patients suffer from neurodegeneration, inflation measured by different magnetic resonance imaging (MRI) modalities and cognitive impairment.

We have developed a scalable multivariate hierarchical Bayesian model (probabilistic latent variable followed by hidden Markov model (HMM)) that can handle different data modalities (binary, count, ordinal and continuous variables) and structured missingness to characterize MS progression in a data driven way using both brain imaging and clinical variables. The proposed model 1) exploits shared information between different brain imaging modalities and clinical variables to find key MS dimensions and corresponding composite scores 2) discover and describe homogeneous states of MS based on composite scores change over time by putting HMM prior on the latent variables.

The method was applied to the NO.MS dataset which is currently the largest and most comprehensive clinical trial dataset in MS with longitudinal data on more than 8000 MS patients that contains both clinical assessments (relapse occurrence, cognitive performance, physical disability status, walking speed and hand coordination) and brain imaging markers (brain atrophy, T2 lesion volume, T1 Gd lesion). We discovered four key-dimensions of MS: a) physical disability; b) subclinical disease burden/associated cognitive deficits; ongoing inflammation either as c) MRI lesions and/or d) clinical relapses. We identified 9 states which can be grouped into 4 clinical Meta-states solely based on transition probability matrix clustering that allow a complete description: i) Early MS (clinically stable patients), ii) acute relapse, iii) transition state and iv) evolved MS.

Choice of input parameters in robust clustering

Thursday, 6.07
16:00–16:20
CS 28
Room 2180

Luis A. García-Escudero[1], **Christian Hennig**[2], **Agustín Mayo-Iscar**[1], **Gianluca Morelli**[3], **Marco Riani**[3]

[1] University of Valladolid (Spain)
[2] University of Bologna (Italy)
[3] University of Parma (Italy)

Abstract

Outliers are known to be potentially detrimental to many statistical procedures and therefore the use of robust alternatives is recommended. With this idea in mind, robust clustering methods have been designed to better resist a certain fraction of outlying observations in Cluster Analysis. The TCLUST method introduced in [1] is a robust clustering procedure based on "impartial" trimming that requires the specification of the number of clusters k and the trimming fraction α. The term impartial means that is the data set itself that indicates what fraction α of observations to be trimmed.

A graphical procedure to select sensible values for k and α was proposed in [2] that is based on the visual inspection of the so-called "classification trimmed likelihood" curves and some heuristic arguments were proposed for justifying its application. We will now provide some theoretical background for this graphical tool and also to better understand the elements involved in it. In addition, a parametric bootstrap method will be presented to find a reduced list of sensible selections for these k and α parameters. This reduced list is returned in an fully automated way, without the need for visual inspection of the curves and avoiding subjective decisions related to the sample variability. The user can then use this reduced list to choose the robust cluster partition that better suits their ultimate clustering purposes using standard cluster validation tools.

Bibliography

[1] García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar. A., *A general trimming approach to robust cluster analysis*, Annals of Statistics, 36 (2008), 1324-1345.

[2] García-Escudero L.A., Gordaliza A., Matrán C., Mayo-Iscar. A., *Exploring the number of groups in robust model-based clustering*, Statistics and Computing, 21 (2011), 585-599.

More powerful universal post-selection inference

Tuesday, 4.07
14:20–14:40
CS 11
Room 2180

Daniel Garcia Rasines, G. Alastair Young

Imperial College London (United Kingdom)

Abstract

In contemporary statistical applications, selection of the formal inferential problem is typically done after some level of interaction with the data. Usually, an initial exploratory analysis is used to identify interesting aspects of the population under study, and then the same dataset is used to learn about them. It is well known that such "data snooping" invalidates classical inferential procedures, and many approaches have been proposed in recent years to correct it.

In this talk, we present a simple alternative to data splitting based on randomization, which can be used to provide valid inference after selection irrespective of the selection mechanism and allows for higher selection and inferential power than the former. We provide theoretical and empirical comparisons of both methods, as well as a Central Limit Theorem for the randomization approach in misspecified models.

Bibliography

[1] Garcia Rasines D., Young G. A, *Splitting strategies for post-selection inference*, Biometrika (2022).

Clustering and non parametric hidden Markov models

Tuesday, 4.07
14:30–15:00
IS 18
Room 0.03

Elisabeth Gassiat, Ibrahim Kaddouri, Zacharie Naulet

Université Paris-Saclay (France)

Abstract

Clustering is a statistical approach which seeks to infer interesting hidden structure in data. We focus here on model-based clustering, see [5] for a recent survey of the subject. The observations are assumed to be distributed as a mixture model, and a clustering procedure seeks to retrieve the sequence of clusters (also called hidden states) associated to the observations. Since parametric mixtures may be unable to handle the complex nature of real-world data, we are interested in non parametric mixture modeling.

In a hidden Markov model (HMM), to each observation is associated a non-observed random variable (thus hidden) which determines the class the observation belongs to. These hidden variables are assumed to form a Markov chain and the observations are independent conditionally to the sequence of hidden states. It has been understood in the recent years that HMMs are mixture models for which non parametric identifiability is possible. Several nonparametric estimation procedures have been proposed in the literature [2, 3, 1] and the limits of learning HMMs have been recently investigated [1].

I shall present recent results from [4] in which we provide strong theoretical garanties for the clustering risk when using HMMs. By the way, I will also provide upper and lower bounds linking the minimal clustering risk with the clustering risk of the Bayes classifier, correcting earlier results in the literature.

Bibliography

[1] Abraham K., Gassiat E., Naulet Z., *Frontiers to the learning of nonparametric hidden Markov models*, submitted, 2023.

[2] de Castro Y., Gassiat E., Lacour C., *Minimax adaptive estimation of non-parametric hidden Markov models*, Journal of Machine Learning Research, 17, 11, 2016.

[3] de Castro Y., Gassiat E., Le Corff S., *Consistent estimation of the filtering and marginal smoothing distributions in non-parametric hidden Markov models*, IEEE Trans. Info. th, 63 (8), 4758-4777, 2017.

[4] Gassiat E., Kaddouri I., Naulet, Z., *Clustering of nonparametric Hidden Markov Models observations*, submitted, 2023.

[5] McNicholas P. *Model-Based Clustering*, Journal of Classification, 33, 331-373, 2016.

Unified convergence theory for Bayesian sparse linear regression with nuisance parameters

Thursday, 6.07
11:30–12:00
IS 13
Room 0.03

Seonghyun Jeongr[1], **Subhashis Ghosal**[2]

[1] Yonsei University (Republic of Korea)
[2] North Carolina State University (United States of America)

Abstract

We study frequentist asymptotic properties of Bayesian procedures for high-dimensional Gaussian sparse regression in the presence of nuisance parameters that can be finite-, high-, or infinite-dimensional. A spike-and-slab prior is used for the sparse regression coefficients. Conditions on the nuisance parameters that retain the optimal posterior contraction under sparsity are examined and discussed. A Bernstein-von Mises-type theorem for sparse regression coefficients is also obtained, that is subsequently used to establish model selection consistency under certain beta-min conditions and compatibility conditions. The results are illustrated with a variety of examples for models involving measurement error, missing data, heteroscedasticity, graphical structure, partial linear structure, and others.

Bibliography

[1] Jeong, S. and Ghosal, S., *Unified Bayesian asymptotic theory for sparse linear regression*. Electronic Journal of Statistics, 15 (2021), 3040–3111.

Integrating causal Bayesian networks and potential outcomes to evaluate the effect of home-based working on Italian firms

Thursday, 6.07
11:30–12:00
IS 10
Room 0.06

Lorenzo Giammei

University of Milan-Bicocca (Italy)

Abstract

Covid-19 generated an unprecedented shock on the Italian economy, which severely affected firm performance. This work focuses on estimating the causal effect of implementing home-based working (HBW) after the pandemic outbreak on manufacturing firms' expected revenues. The analysis uses a unique firm-level dataset [1], which captures a rich set of features before and after the spread of the virus.

Causal effect estimation is performed implementing an integrated approach that merges Causal Graphs [2] and Potential Outcomes frameworks [3]. At first, the dataset is used to learn a causal diagram that encodes theory-based assumptions and causal information contained in the data through a structural learning algortihm [4]. An adjustment set is then selected by applying the back-door criterion on the obtained graph. Bootstrap-based robustness checks are performed in order to assess the validity of the adjustment set. Lastly, causal estimates are computed with matching [5], using the chosen adjustment set to ensure unconfoundedness. The adopted integrated approach permits the assessment of whether the assumptions required by matching are satisfied, relying on the structure of the learned causal graph. This increases the analysis clarity and the reliability of the obtained results.

The estimates confirm the presence of a positive effect of the implementation of HBW on expected revenues. The treatment seems to be particularly effective in providing revenue stability and mitigating of losses. The findings are consistent with the fact that HBW equips firms with greater flexibility and helps contain productivity decreases in Covid times.

Bibliography

[1] Balduzzi P., Brancati E., Brianti M., Schiantarelli F., *The Economic Effects of COVID-19 and Credit Constraints: Evidence from Italian Firms' Expectations and Plans*, IZA Discussion Paper, 13629 (2020).

[2] Pearl J., *Causal diagrams for empirical research*, Biometrika, 82 (1995), 669-688.

[3] Rubin D.B., *Causal inference using potential outcomes: Design, modeling, decisions*, Journal of the American Statistical Association, 100 (2005), 322-331.

[4] Spirtes P., Glymour C., Scheines R., *Causation, Prediction, and Search* , MIT press, 2000.

[5] Stuart E. A., *Matching methods for causal inference: A review and a look forward*, Statistical science: a review journal of the Institute of Mathematical Statistics, 25 (2010), 1-21.

Distribution generalization in semi-parametric models: A control function approach

Friday, 7.07
11:30–12:00
IS 11
Room 0.03a

Nicola Gnecco[1], **Sebastian Engelke**[2], **Jonas Peters**[3], **and Niklas Pfister**[1]

[1] University of Copenhagen (Denmark)
[2] University of Geneva (Switzerland)
[3] ETH Zurich, Zurich (Switzerland)

Abstract

In this talk, we present a framework to study the problem of distribution generalization from a causal perspective. We consider an anchor regression setting [1] with a nonparametric causal function, where interventions on exogenous variables (the anchors) induce distributional shifts in the predictors and the response. In such a setup, one is usually interested in the minimax function, i.e., minimizing the worst-case mean square error over all distributions induced by interventions on the anchors. In the nonlinear setting, however, the existing literature does not characterize explicitly or identify the minimax function. In this work, we characterize the minimax function explicitly and study its identification properties under certain assumptions on the noise terms. Furthermore, we propose a simple adaptation to the random forests algorithm [2] to learn the minimax function from data.

Bibliography

[1] Rothenhaeusler, D., Meinshausen, N., Buehlmann, P., & Peters, J. (2021). Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society Series B: Statistical Methodology, 83(2), 215-246.
[2] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Evaluation of Some Ridge Parameter Estimators in Logistic Regression Based on MSE and Skewness Criteria

Thursday, 6.07
16:40–17:00
CS 28
Room 2180

Atila Göktaş, Selman Mermi, Özge Akkuş

Muğla Sıtkı Koçman University, Faculty of Science, Department of Statistics (TURKEY)

Abstract

The logistic regression is a powerful statistical method used to analyze the linear relationship between a dependent variable and one or more predictor variables. In this model, dependent variable contains two (binary) or more than two (multinomial) categories, unlike the linear regression model where the dependent variable is numerical. Multicollinearity is a common issue that can arise in logistic regression models when predictor variables are higly correlated with each other. It can lead to unstable and unreliable estimates of the model coefficients, just like in the linear regression model. Furthermore, it may result in inflated standard errors, which can lead to incorrect hypothesis testing and confidence interval estimation for the regression coefficients. The ridge regression method, developed by Hoerl and Kennard (1970) to overcome multicollinearity, was first adapted to the logistic regression model by Schaefer et al. (1984). Many ridge parameter estimators have been proposed for the multiple linear ridge regression model in the literature. The goal of this study is to adapt some popular of those estimators to the binary logistic ridge regression model and compare their efficiencies among themselves as well as with the maximum likelihood (ML) approach. Outlier detection and normality criteria were used to evaluate the performances of ridge parameter estimators, as well as the classical mean Square Error (MSE) criterion. For this purpose, a Monte Carlo simulation study with 10000 replication was conducted, taking into account simulation parameters on small, medium, and large scales in terms of the number of regressors (p), sample size (n), and correlation level (ρ).

According to the simulation results, all ridge parameter estimators surpass the ML method in terms of the MSE criterion in all cases. Moreover, the ridge parameter estimators with the best and worst performance in terms of these three criteria were determined and presented to the attention of researchers.

Bibliography

[1] Hoerl, A. E., Kennard, R. W., *Ridge Regression: Biased Estimaton to Nonorthogonal Problems*, Technometrics, 12(1), (1970), 69–82, https://doi.org/10.1080/00401706.1970.10488635.

[2] Schaefer, R. L., Roi, L. D., Wolfe, R. A., *A ridge logistic estimator*, Communications in Statistics – Theory and Methods, (1984), 13(1), https://doi.org/10.1080/03610928408828664.

Statistical problems for Smoluchowski processes

Thursday, 6.07
15:00–15:30
IS 09
Room 0.03

A. Goldenshluger, R. Jacobovic

University of Haifa (Israel)

Abstract

Suppose that particles are randomly distributed in $\mathbb{R}^d$, and they are subject to identical stochastic motion independently of each other. The Smoluchowski process describes fluctuations of the number of particles in an observation region over time. In this talk we discuss probabilistic properties of the Smoluchowski processes and consider related statistical problems. In particular, we consider two different models of the particle displacement process: the undeviated uniform motion (when a particle moves with random constant velocity along a straight line) and the Brownian motion displacement. In the setting of the undeviated uniform motion we study the problems of estimating the mean speed and the speed distribution, while for the Brownian displacement model the problem of estimating the diffusion coefficient is considered. In all these settings we develop estimators with provable accuracy guarantees.

Causal Spatial Quantile Regression

Thursday, 6.07
12:00–12:20
CS 21
Room 2180

Yan Gong[1,3], **Reetam Majumder**[2], **Brian J. Reich**[2], **Raphaël Huser**[3]

[1] Imperial College London (UK)
[2] North Carolina State University (USA)
[3] King Abdullah University of Science and Technology (Saudi Arabia)

Abstract

In this work, we propose a novel spatial quantile regression-based framework, for estimating heterogeneous treatment effects which are non-linear, spatially-varying, and quantile-dependent. While our main goal is to estimate quantile treatment effects at extreme quantile levels, we develop a general methodology that can be used to target any quantile of interest (from low to high quantiles). We extend the semi-parametric neural-network-based quantile regression framework proposed by [1] by further integrating spatial features in the model and then exploit this flexible model for estimating spatial quantile treatment effects. The methodology is validated by simulation and we apply the method in quantifying spatial quantile treatment effects of maternal smoking on extreme low birth weight of newborns in North Carolina, United States.

Bibliography

[1] Xu, S. G., Reich, B. J., *Bayesian nonparametric quantile process regression and estimation of marginal quantile effects*, Biometrics, 2023, 79(1), 151-164.

Simplicial depths in the fuzzy setting

Thursday, 6.07
12:20–12:40
CS 20
Room 1.40

Luis González-De La Fuente[1], **Alicia Nieto-Reyes**[1], **Pedro Terán**[2]

[1] Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria (Spain)
[2] Departamento de Estadística e Investigación Operativa y Didáctica de las Matemáteicas, Universidad de Oviedo (Spain)

Abstract

Simplicial depth is one of the most well-known and studied statistical depth funtions in the multivariate setting and it is based on the notion of multivariate simplex. The simplicial depth of a point $x \in \mathbb{R}^n$ with respect to a distribution F is just the probability that x belongs to the simplex generated by $n+1$ independent and identical distributed random variables with distribution F. Firstly, we will propose a natural generalization of the notion of simplex for fuzzy sets, namely the pseudosimplex. Basically, a fuzzy set belongs to the pseudosimplex generated by a colletion of fuzzy sets if its support function are in between the minimum and the maximum of the support functions of the collection of fuzzy sets for every α-level and every direction $u \in \mathbb{S}^{n-1}$. Secondly, we will introduce three plausible generalizations of simplicial depth for the fuzzy setting, based on the multivariate and functional cases. The first notion is called the *naive simplicial depth* and its definition is analogous as in the multivariate case, but with the notion of pseudosimplex. The other simplicial depths proposed are the *fuzzy simplicial depth* and the *modified fuzzy simplicial depth*. They try to measure which *proportion* of a fuzzy set is contained in the pseudosimplex generated by a collection of independent observations from a fuzzy random variable. This *proportion* is measured with respect to the Lebesgue measure in $[0,1]$ and the normalized Haar measure in $\mathbb{S}^{n-1}$. Thirdly, we will study whether those simplicial depths fulfill the newly defined concept of statistical depth for the fuzzy setting. We will illustrate the behavior of our proposals through the use of simulated and real data.

Controlled Sequential Monte Carlo Methods for Continuous-Time Diffusion Models

Friday, 7.07
14:40–15:00
CS 35
Room 0.06

Andras Fulop[1], **Luca Gonzato**[2], **Jeremy Heng**[1], **Fabio Trojani**[3]

[1] ESSEC Business School (Paris-Singapore)
[2] University of Vienna (Austria)
[3] University of Geneva (Switzerland)

Abstract

In this paper, we propose a general econometric approach to the estimation of continuous-time diffusion models. The combined usage of data augmentation and controlled Sequential Monte Carlo (SMC) methods [2] allows to control discretization bias and provides efficient likelihood estimators, necessary for Bayesian inference. We test our methodology by considering a multifactor option pricing model where the latent volatility is a continuous-time Wishart process as defined in [1]. To deal with the highly informative nature of options data, we propose a likelihood tempering procedure to gradually introduce information from observations and construct optimal proposal distributions that deliver zero variance likelihood estimators. Numerical experiments on simulated data illustrate that our approach yields excellent tracking of the latent states and outperforms standard particle filtering techniques in terms of marginal likelihood estimation. The proposed SMC method is then embedded in a pseudo-marginal SMC^2 algorithm to perform parameter estimation. We illustrate the strengths of our proposed methodology by estimating a three-factor Wishart stochastic volatility model using time series of option prices on the S&P500 market.

Bibliography

[1] Gruber, P., Tebaldi, C., Trojani, F., *The price of the smile and variance risk premia*, Management Science, 2020

[2] Heng, J., Bishop, A. N., Deligiannidis, G., Doucet, A., *Controlled Sequential Monte Carlo*, Annals of Statistics, 2020

Bounds on expectations of spacings from DRFR distributions

Tuesday, 4.07
14:40–15:00
CS 13
Room 4420

Agnieszka Goroncy

Nicolaus Copernicus University (Poland)

Abstract

We consider the sequence of independent and identically distributed random variables that have the decreasing reversed failure rate. For such setting we present the solution to the problem of determining the upper positive mean-variance bounds on the expectations of the spacings, i.e. the differences of consecutive order statistics. We also provide the equality conditions. The results are obtained with use of the projection method.

Bibliography

[1] Danielak, K. (2003). Sharp upper bounds on trimmed means from restricted families. *Statistics* **37**, 305–324.

[2] Goroncy, A., Rychlik, T. (2022). Bounds on the expectations of order statistics for the monotone reversed failure rate distributions, submitted.

[3] Goroncy, A., Rychlik, T. (2016). Evaluations of expectations of order statistics and spacings based on IFR distributions. *Metrika* **79**, 635–657.

[4] Goroncy, A., Rychlik, T. (2015). Optimal bounds on expectations of order statistics and spacings from nonparametric families of distributions generated by convex transform order. *Metrika* **78**, 175–204.

Estimating periodicity in disease dynamics

Tuesday, 4.07
16:00–18:00
Poster

Natalia Bochkina, Niamh Graham

University of Edinburgh

Abstract

Periodic processes are ubiquitous in the context of Biology and Medicine. Periodicity in disease progression has not been widely explored, although there has been investigation into the periodicity of CRP, a blood plasma protein which is a biomarker for inflammation [2] and [3]. Here we aim to model periodicity in the disease progression of an inflammatory disease, acute pancreatitis (AP).

Principal component analysis (PCA) is used to reduce high dimensional medical data to two most relevant unobserved factors. These two factors, termed protective and damaging factors, characterise the disease progression and are modelled by a complex valued Gaussian process with Markovian structure. The covariance function of this process coincides with the damped oscillating solution of the system of differential equations describing the interaction of these disease driving factors. Calculation of the sample covariance of the Gaussian process enables the inference of the parameters of the function describing disease progression. This approach was proposed in [1]. This approach also allows the estimation of critical time points in the disease course. These critical time points are hypothesised to correspond to points during the disease where medical intervention is most effective.

We develop appropriate inference for this approach, namely we derive appropriate asymptotic confidence regions for parameters based on the MLEs that lead to consistent and efficient inference.

We apply this method to estimate periodicity in the disease progression of acute pancreatitis. As the considered pancreatitis data has missing values, we apply the approach of Josse et al. [4] to perform imputation (under the assumption of missing at random) and demonstrate how this improves accuracy of parameter estimation. We show that the first two principal components for the pancreatitis data have the interpretation of damaging and protecting factors conjectured by [1] which have a clear medical interpretation, and identify periodicity of disease progression, as well as optimal medical intervention times and their uncertainty.

Bibliography

[1] Bart, A. G. et al. *Modeling disease dynamics and survivor functions by sanogenesis curves*, Journal of statistical planning and inference, 2005, pp. 33-51.

[2] Coventry, Brendon J., et al., *CRP identifies homeostatic immune oscillations in cancer patients: a potential treatment targeting tool?* Journal of translational medicine, 2009, pp. 1–8.

[3] Dorraki, Mohsen, et al., *On detection of periodicity in C-reactive protein (CRP) levels*, Scientific reports, 2018, pp. 1–7.

[4] Josse, Julie and Husson, François . *missMDA: a package for handling missing values in multivariate data analysis*, Journal of Statistical Software, 2016, pp. 1–31.

Nonparametric tests for imprecise data analysis

Thursday, 6.07
14:40–15:00
CS 25
Room 2180

Przemyslaw Grzegorzewski[1,2], **Milena Zacharczuk**[1]

[1] Warsaw University of Technology (Poland)
[2] Polish Academy of Sciences (Poland)

Abstract

When faced with imprecise or vague observations, so often in real-life data, especially human ratings based on opinions or perceptions, nonparametric statistical methods are usually used. Random fuzzy numbers (also known as fuzzy random variables) constitute a convenient model which allows grasping both randomness, associated with the data generation mechanism, and fuzziness, connected with data imprecision. However, the derivation of nonparametric tests for fuzzy data cannot be done by directly generalizing nonparametric tests for real-valued data to a fuzzy environment. Indeed, in analyzing fuzzy data from the statistical perspective we immediately come upon some key obstacles, like the nonlinearity associated with the fuzzy number arithmetic, the lack of suitable probability distribution models or no limit theorems for random mechanisms producing fuzzy data which could be directly applied in statistical inference.

To overcome these drawbacks one usually applies an appropriate metric between fuzzy data and a bootstrapped central limit theorem for general space-valued random elements [1] or permutation tests [2, 3, 4]. In this contribution, we consider a novel approach for test construction that combines permutation tests with the credibility index [5] measuring the degree of the dominance for each pair of random variables.

Bibliography

[1] Blanco-Fernández A., Casals M.R., Colubi A., Corral N., García-Bárzana M., Gil M.A., González-Rodríguez M., López M.T., Lubiano M.A., Montenegro M., Ramos-Guajardo A.B., de la Rosa de Sáa S., Sinova, B., *A distance-based statistic analysis of fuzzy number-valued data* (with Rejoinder), International Journal of Approximate Reasoning, 55 (2014), 1487–1501, 1601–1605.

[2] Grzegorzewski P., *Permutation k-sample goodness-of-fit test for fuzzy data*, In: 2020 IEEE International Conference on Fuzzy Systems (2020), 1–8.

[3] Grzegorzewski P., *Two-sample test for comparing ambiguity in fuzzy data*, In: 2022 IEEE International Conference on Fuzzy Systems (2022), 1–8.

[4] Grzegorzewski P., Gadomska O., *Nearest neighbor tests for fuzzy data*, In: 2021 IEEE International Conference on Fuzzy Systems (2021), pp. 1–6.

[5] Liu B., Uncertainty Theory: An Introduction to its Axiomatic Foundations, Springer, 2004.

Imputomics: comprehensive missing data imputation for metabolomics data

Tuesday, 4.07
16:00–18:00
Poster

Krystyna Grzesiak[1], **Jarosław Chilimoniuk**[1,2], **Dominik Nowakowski**[2], **Jakub Kała**[3], **Adam Krętowski**[2], **Małgorzata Bogdan**[1], **Michał Ciborowski**[2], **Michał Burdukiewicz**[2,4]

,

[1] University of Wrocław (Poland)

[2] Medical University of Białystok (Poland)

[3] Why R? Foundation (Poland)

[4] Autonomous University of Barcelona, Cerdanyola del Valles (Spain)

Abstract

Metabolomics is a rapidly growing field that aims to identify and quantify small molecules, or metabolites, in biological samples. These metabolites are key indicators of the biochemical state of cells, tissues, and organisms. Their analysis can provide valuable insights into biological processes, disease mechanisms, and drug effects [1]. However, metabolomics data often suffer from missing values due to technical limitations, biological variability, and experimental design. The presence of missing data can hinder statistical analysis, leading to biased or incomplete results and limiting the scope and accuracy of metabolomics research, highlighting the need for appropriate imputation methods[2].

In this research project, we aim to provide a comprehensive overview of the existing methods for imputing missing values in metabolomics datasets, considering different patterns of missingness: completely at random (MCAR), missing at random (MAR), non-random missingness (MNAR) and mixtures [3].

We systematically reviewed the literature on imputation methods for metabolomics data, identifying 68 different imputation methods from 21 articles. These methods included simpler approaches based on the constant imputation (as the mean imputation), but also more advanced techniques based on K-nearest neighbors, random forests or principal component analysis.

To make these methods more accessible to researchers, we developed an R package called Imputomics, which includes implementations of the identified methods. Additionally, we created a user-friendly webserver that enables researchers to upload their data and select the appropriate imputation method from the available options. We will also develop a benchmarking framework to evaluate the performance of these methods on simulated and real-world datasets. The benchmarking framework will include imputation accuracy, bias, and computational efficiency metrics.

Bibliography

[1] Villas-Boas, S. G. et al. (2005). Mass spectrometry in metabolome analysis. Mass Spectrometry Reviews, 24(5), 613–646.

[2] Armitage, E. G. et al. (2015). Missing value imputation strategies for metabolomics data. ELECTROPHORESIS, 36(24), 3050–3060.

[3] Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581–592.

Bootstrap test for variance components in nonlinear mixed effects models for small sample size in presence of nuisance parameters and singular Fisher Information Matrix

Tuesday, 4.07
16:00–18:00
Poster

Tom Guédon[1], **Charlotte Baey**[2], **Estelle Kuhn**[1]

[1] MaIAGE, INRAE, Université Paris-Saclay, Domaine de Vilvert, 78352 Jouy-en-Josas, France
[2] Laboratoire Paul Painlevé, Université de Lille, 59655 Villeneuve d'Asq, France

Abstract

We address the problem of variance components testing in nonlinear mixed effects models in non regular settings. We consider N individuals each measured J times. We denote by y_{ij} the jth observation of the ith individual for $i = 1, \dots, N$ and $j = 1, \dots, J$. We consider the following nonlinear mixed effects model:

$$y_i = g(x_i, \beta, \Lambda\xi_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_J), \quad \xi_i \sim \mathcal{N}(0, I_p)$$

where g is a known nonlinear function, x_i gathers all the covariates of the ith individual, β is the unknown vector of fixed effects, Λ is an upper triangular matrix which is a scaling parameter for the random effect ξ_i, and σ^2 is the positive noise variance. The variables (ε_i) and (ξ_i) are assumed independent and mutually independent.

We consider the likelihood ratio statistic in order to test that some components of the covariance matrix of the scaled random effect $\Lambda\xi_i$ are zero. Regarding non regular settings, two main issues may arise. On the one hand, under the null hypothesis the true parameter is not an interior point of the parameter space which prevent from using the usual asymptotic theory of maximum likelihood estimation. This has been considered in the literature (see [1], [3] for example). On the other hand, in the specific context of nonlinear mixed-effects models, the Fisher information matrix is singular when considering a zero scaling parameter Λ. This is in particular the case when unknown untested variances, also called nuisance parameters, are also zero. Those two issues make the asymptotic distribution of the likelihood ratio statistic difficult or even impossible to identify.

We propose a parametric bootstrap procedure to perform this test. The benefits of the bootstrap in our context is twofold: first it performs well when dealing with small sample sizes, second it enables to take into account the presence of nuisance parameters. However the two issues cited above are sources of inconsistency of the bootstrap. First the boundary issue can be solved using a modified bootstrap parameter that shrinks toward zero the nuisance parameters as developed in [2]. We show that the singularity issue is also solved thanks to this shrinkage, as long as the shrinking parameter converges fast enough toward zero. The main contribution of this work is a procedure that is straightforward to apply even to nonlinear models, and deals with the issue of nuisance parameters that is not, to our knowledge, considered in the mixed effects models literature. We illustrate the performance of our procedure on simulated datasets and real data.

Bibliography

[1] Donald WK Andrews. “Estimation when a parameter is on a boundary”. In: Econometrica 67.6 (1999), pp. 1341–1383.
[2] Giuseppe Cavaliere et al. “Bootstrap inference on the boundary of the parameter space, with application to conditional volatility models”. In: *Journal of Econometrics* (2020).
[3] Steven G Self and Kung-Yee Liang. “Asymptotic properties of maximum likelihood estima- tors and likelihood ratio tests under nonstandard conditions”. In: *Journal of the American Statistical Association* 82.398 (1987), pp. 605–610.

Biclustering Multivariate Longitudinal Data with Application to a Diffusion Tensor Imaging Study

Wednesday, 5.07
11:40–12:00
CS 15
Room 1.40

Jaroslaw Harezlak[1], **Caleb Weaver**[2], **Yu-Chien Wu**[3], **Qiuting Wen**[3], **Luo Xiao**[2]

[1] Indiana University (USA)
[2] North Carolina State University (USA)
[3] Indiana University School of Medicine (USA)

Abstract

Biclustering [1, 2] is the task of simultaneously clustering the samples and features of a data set. In doing so, subsets of samples that exhibit similar behaviors across subsets of features can be identified. Motivated by a longitudinal diffusion tensor imaging study of sport-related concussion (SRC) [3], we present the problem of biclustering multivariate longitudinal data in which subjects and features are grouped simultaneously based on longitudinal patterns rather than magnitude. We propose a penalized regression based method for solving this problem by exploiting the heterogeneity in the longitudinal patterns within subjects and features. We evaluate the performance of the proposed methods via a simulation study and perform an analysis of the motivating data set. In this analysis, we reveal subgroups of SRC cases that exhibit heterogeneous patterns of white-matter abnormalities.

Bibliography

[1] Madeira SC, Oliveira AL. *Biclustering algorithms for biological data analysis: a survey*, IEEE/ACM transactions on computational biology and bioinformatics 2004; 1(1): 24–45.
[2] Chi EC, Allen GI, Baraniuk RG. *Convex biclustering*, Biometrics 2017; 73(1): 10–19.
[3] Mustafi SM, Harezlak J, Koch KM, et al. Acute white-matter abnormalities in sports-related concussion: a diffusion tensor imaging study from the NCAA-DoD CARE Consortium. Journal of neurotrauma 2018; 35(22): 2653–2664.

Reverse Information Projections and Optimal E-Statistics

Thursday, 6.07
14:00–14:20
CS 25
Room 2180

Peter Harremoës[1], **Tyron Lardy**[2,3], **Peter Grünwald**[2,3]

[1] Copenhagen Business College
[2] Leiden University
[3] Centrum Wiskunde & Informatica

Abstract

Information projections appear in a multitude of problems in probability and statistics alike. In particular, reverse information projections have recently been shown to lead to optimal e-statistics for certain hypothesis testing problems. On a measurable space $(\Omega, \mathcal{F})$, an e-statistic for testing a hypothesis $\mathcal{C}$ is simply a non-negative statistic $E : \Omega \to \mathbb{R}_{>0}$ with expected value bounded by one under $\mathcal{C}$, i.e. $\sup_{Q\in\mathcal{C}} \int_\Omega E(\omega)\,\mathrm{d}Q \leq 1$ [1]. An e-statistic evaluated on data is called an e-value, and the larger the e-value, the larger the evidence against the null. Rejecting $\mathcal{C}$ whenever the e-value is larger than $1/\alpha$ leads to a test of $\mathcal{C}$ with type-I error guarantee α by a simple application of Markov's inequality: $Q(E(\omega) > 1/\alpha) \leq \alpha \int_\Omega E(\omega)\,\mathrm{d}Q \leq \alpha$ for all $Q \in \mathcal{C}$. Furthermore, e-values computed for consecutive tests of the same hypothesis can simply be multiplied to measure the total evidence [2]. This property has lead to the popularization of e-statistics in the field of anytime-valid testing [3], where the goal is to have a {`reject`, `do not reject`}-decision rule for each sample size that preserves the type-I error uniformly over time. In experiments with accumulating data, such stopping rules allow for continuous monitoring and analysis of the data without compromising the validity of the consequent inference.

Furthermore, as an analogue to statistical power, $\int_\Omega \log E(X)\,\mathrm{d}P$ gives an inherent notion of optimality for e-statistics whenever the alternative is given by a single distribution P. The former quantity is called the growth-rate and the e-statistic with the highest growth-rate is called Growth-Rate Optimal (GRO) relative to P. In [1] it is shown that whenever the Kullback-Leibler (KL) divergence between P and $\mathcal{C}$ is finite, the GRO e-statistic is uniquely given by a likelihood ratio between P and its reverse information projection (RIPr) on $\mathcal{C}$. That is, there exists a unique measure $\hat{Q}$ such that $D(P\|\hat{Q}) = \inf_{Q\in\mathcal{C}} D(P\|Q) < \infty$, and $\mathrm{d}P/\mathrm{d}\hat{Q}$ is the GRO e-statistic. This result is only applicable if the KL is finite, or equivalently when there are no e-statistics with infinite growth-rate. In this work we show that under much weaker assumptions, there exists a measure that generalizes the RIPr and gives rise to an e-statistic in the same way as the RIPr does. This e-statistic is optimal in the sense that its growth-rate relative to any other e-statistic is positive. Finally, we discuss conditions under which the measure we arrive at is a strict sub-probability distribution, which also gives new insights into ordinary reverse information projections.

Bibliography

[1] Grünwald P., de Heide R., Koolen W., *Safe Testing*, arXiv preprint arXiv:1906.07801, 2023, Accepted for JRSS, Series B.

[2] Vovk V., Ruodu W., *E-values: Calibration, combination and applications.* The Annals of Statistics, 2021,49.3: 1736-1754
[3] Ramdas A., Grünwald G., Vovk V., Shafer G., *Game-theoretic statistics and safe anytime-valid inference*, arXiv preprint arXiv:2210.01948, 2022

Warped metrics for optimisation and sampling in probabilistic models

Thursday, 6.07
16:20–16:40
CS 28
Room 2180

Marcelo Hartmann, Bernardo Williams, Hanlin Yu & Arto Klami

University of Helsinki (Finland)

Abstract

In Bayesian statistics, Markov Chain Monte Carlo methods (MCMC) and the maximum a posterior estimate (MAP) are predominant inference schemes as they allow Bayesian inference to be operational in practice. Since they offer summaries of interest for the parameters in the posterior distribution, they have been subject of much research since its inception, see for example [1]. However, for more complex models, both classical sampling and optimisation are harder as the geometry of the target distribution plays a clear role in the performance of the computational algorithms, see [2] and references therein. To circumvent this problem the Fisher information (FI) matrix has been the ubiquitous choice for geometric MCMC methods and natural gradients (Fisher score) for the MAP estimate. The FI, though, is in practice hard to use as usually does not have closed-form arithmetics. Recently, alternatives to the FI have been proposed in the literature, see for example [10, 4]. These choices are computationally attractive, however they require extra parameters to be tuned so that the sampler or the optimisation procedure would have better performance.

In this work the present a metric-tensor obtained from the theory of warped product spaces [5] which provide us another way to define a metric that can be computational friendly and can also ease the problem of selecting tuning parameters. We introduce a function's graph on an ambient space equipped with a warped metric, this operation induces a metric on the tangent space that has desired computational properties such as closed-form inverse matrix and determinants. We further make a specification of the metric that can be used for both sampling and optimisation schemes, and empirically show that it can improve both tasks with minimal extra computational costs leading to $\mathcal{O}(D)$ memory load and $\mathcal{O}(D^2)$ arithmetic operations. We apply the proposed metric in the context of manifold metropolis-adjusted Langevin Dynamics (MMALA) and Riemannian optimisation schemes, comparing with methods on Euclidean spaces such as classical Langevin dynamics and conjugate gradient optimisation counterparts.

Bibliography

[1] W. K. Hastings., *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*, Biometrika , 57, 1, (1970).

[2] M. Girolami & B. Calderhead. , *Riemann manifold Langevin and Hamiltonian Monte Carlo methods.* Journal of Royal Statistical Society Series B, (2011).

[3] T. S.Kleppe. *Log-density gradient covariance and automatic metric tensors for Riemann manifold Monte Carlo methods.* ArXiV (2022)

[4] M. Hartmann, M. Girolami & A. Klami. *Lagrangian Manifold Monte Carlo on Monge Patches.* AISTATS (2022).

[5] B. O'Neill. *Semi-Riemannian Geometry with applications to relativity.* Elsevier (1983).

Federated Learning of Structured Probabilistic Models

Thursday, 6.07
16:00–16:30
IS 14
Room 1.01

Conor Hassan, Robert Salomone, Kerrie Mengersen

Queensland University of Technology (Australia)

Abstract

In this work, we address the challenges of federated learning for structured probabilistic models, aiming to expand the class of models and types of inference applicable in the Bayesian federated learning setting. We present Structured Federated Variational Inference (SFVI) [1], a novel algorithm designed to perform inference on distributed data sources without violating data privacy. Furthermore, we introduce a communication-efficient variant of SFVI, called SFVI-Avg, inspired by the FedAvg algorithm [2].

Our approach leverages structured variational inference techniques, widely used in Bayesian machine learning, to adapt to the federated learning context. Notably, our methods show promise in modelling heterogeneity between different data sources, making them particularly suitable for real-world applications. We demonstrate the effectiveness of our proposed algorithms by examining their performance across various applications, such as Bayesian multinomial regression, hierarchical Bayesian neural networks, and topic models. This work contributes to the development of more versatile and efficient federated learning methods, showing promise for real-world applications that demand secure, privacy-preserving models.

Bibliography

[1] Hassan C., Salomone R., Mengersen K., *Federated Variational Inference Methods for Structured Latent Variable Models*, arXiv preprint arXiv:2302.03314 (2023). Under review.

[2] McMahan, B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B.A., 2017, April. *Communication-efficient Learning of Deep Networks from Decentralized Data* In Artificial Intelligence and Statistics (pp. 1273–1282). PMLR.

Weak pattern convergence for SLOPE and its robust versions

Thursday, 6.07
14:40–15:00
CS 26
Room 5440

Ivan Hejny[1], **Jonas Wallin**[1], **Malgorzata Bogdan**[1,2]

[1] Department of Statistics, Lund University (Sweden)
[2] Institute of Mathematics, University of Wroclaw (Poland)

Abstract

The Sorted L-One Estimator (SLOPE) is a popular regularization method in regression, which induces clustering of the estimated coefficients. That is, the estimator can have coefficients of identical magnitude. In this paper, we derive an asymptotic distribution of SLOPE for the ordinary least squares, Huber, and Quantile loss functions, and use it to study the clustering behavior in the limit. This requires a stronger type of convergence since clustering properties do not follow merely from the classical weak convergence. We establish asymptotic control of the false discovery rate for the asymptotic orthogonal design of the regressor. We also show how to extend the framework to a broader class of regularizers other than SLOPE.

Evaluation of Different Sequential Testing Strategies applied to Dielectric Breakdown Requirements of a Batch of High-Voltage Devices

Tuesday, 4.07
16:00–18:00
Poster

Kai Hencken[1], **Patrick Rumpelt**[2], **Michael Weuffel**[2], **Thomas Schmölzer**[2]

[1] ABB Switzerland Ltd, Corporate Research (Switzerland)
[2] ABB Germany, Ratingen (Germany)

Abstract

Sequential testing is a common approach to decide between different hypotheses especially for quality testing. In the case of composite hypotheses there exist different approaches to it. These are evaluated here for a practical application: testing the dielectric breakdown performance of a batch of high-voltage equipment. They have to fulfill this with a high level of confidence. Each batch is therefore tested individually using some well-defined voltage pulses, see [1] for details. An assessment of the overall probability of a randomly selected device to hold the required high voltage depends on the variability on different levels: There is scattering for each individual device, between different devices, and finally between batches. A number of tests on each device is performed to acquire the individual breakdown probability distribution. The value of a (low) quantile of this distribution is then used as the quality criteria and the distribution of these quantiles is then further analyzed. The probability of this quantile value to be below a critical value, serves as the quality criteria, whether the batch is accepted or not.

As dielectric testing is expensive and time-consuming, and in addition can lead to the desctruction or deterioriation of the device, methods to minimize the number of test or test devices are of interest [2]. In this work it is therefore investigated how differential sequential hypothesis testing approaches perform in this case, focusing especially on the case of two composite hypothesis and the required number of test devices. The sequential probability ratio test (SPRT) by Wald [3] is a commonly used approach. In its simplest form it is based on a binary variable, which accounts only, whether the quantile value is below or above the critical one. Such an approach is clearly sub-optimal, as it does not take into account, how far it is from the boundary. But it can be phrased as a test of two simple hypotheses. The generalization to composite approaches, with hypotheses defined over parameter ranges instead of specific values, is required. One approach is maxSPRT [4], where the maximum likelihood value within this range is used in the ratio. By assuming a log-normal distribution of the quantile value, some simplifications can be done analytically. More recently the sequential Bayes Factor hypothesis testing (SBFT) has been proposed in [5] as an alternative. Instead of an optimization a marginalization over the range with a suitable prior distribution is done.

A comparison of the performance of the different methods is done using both simulated and real data. Critical levels for the stopping of the sequence, are validated against the theoretical error rates. A focus is also on the performance in terms of the distribution of the number of tests required for the different cases. In the case of real data, bootstrap

sampling is applied to confirm the performance of the method, as only one full dataset was available. As the advantage of this method has been demonstrated, further improvements are discussed in the end.

Bibliography

[1] P. Rumpelt, M. Weuffel, T. Schmölzer. Acquisition and Evaluation of the Breakdown Voltage as a Result of the Layout and the Statistical Spread of Vacuum Gaps, 27th International Conference on Electricity Distribution, Paper nÂ°10533, 2023.

[2] P. Rumpelt, M. Weuffel, T. Schmölzer, K. Hencken. Methods for statistical analysis of the breakdown voltage of vacuum gaps. 30th International Symposium on Discharges and Electrical Insulation in Vacuum. to be published (2023).

[3] A. Wald. Sequential Tests of Statistical Hypotheses. Annals of Mathematical Statistics. 16 (2): 117 186, 1954.

[4] M. Kulldorff, R. L. Davis, M. Kolczak, E. Lewis, T. Lieu, and R. Platt. A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance. Sequential Analysis 30 (2011): 58-78.

[5] F. D. Schönbrodt, E. J. Wagenmakers, M. Zehetleitner, and M. Perugini. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. Psychological methods, 22(2), 322, 2017.

Parameter estimation based on differential equations of empirical transforms

Tuesday, 4.07
16:00–18:00
Poster

Tomasz Hołyński

Vienna University of Technology (Austria)

Abstract

A new class of robust parameter estimators is proposed based on integral transforms of probability distributions, such as the characteristic function or the Laplace transform. In the past, estimators of this type were mostly constructed as minimizers of squared distances between the model transform (of the assumed distribution) and the empirical transform (computed from the sample), e.g. [1], [2] or [3]

Here, instead of using the above distances, we construct estimators by minimizing the L2 norm of 'empirical' differential equations satisfied by the transforms. Such functionals have been already used in literature but for goodness-of-fit testing, see e.g. [4], [5]. The motivation for applying the equations in the estimation context is that it often leads to computationally attractive and reliable estimators, having closed-form expressions and being robust to outliers.

We address issues of consistency and asymptotic normality and derive influence functions of the estimators. We show that a good trade-off between robustness and efficiency requires data-driven weight functions in the L2 statistical functionals. In simulations, we study finite sample performance in pure and contaminated models, making comparisons with other kinds of estimators.

Bibliography

[1] Heathcote C.R., *The integrated squared error estimation of parameters*, Biometrika, 64 (1977), 255-264.

[2] Jiménez-Gamero M.D., Alba-Fernández, V., Batsidis, A., *Fourier methods for model selection*, Annals of the Institute of Statistical Mathematics, 68 (2016), 1-29.

[3] Jiménez-Gamero M.D., Batsidis, A., *Minimum distance estimators for count data based on the probability generating function with applications*, Metrika, 80 (2017), 503-545.

[4] Henze, N., Meintanis, S.G., Ebner, B., *Goodness-of-fit tests for the gamma distribution based on the empirical Laplace transform*, Communications in Statistics - Theory and Methods, 41 (2012), 1543-1556.

[5] Goffard, P.-O., Rao Jammalamadaka, S., Meintanis S.G., *Goodness-of-fit procedures for compound distributions with an application to insurance*, Journal of Statistical Theory and Practice, 16 (2022), article nr 52.

A novel functional model for temporal images

Tuesday, 4.07
14:00–14:20
CS 10
Room 1.40

Sonia Dembowksa[1], **Jeanine Houwing-Duistermaat**[1,2], **Haiyan Liu**[1]

[1] University of Leeds (United Kingdom)
[2] Radboud University Nijmegen (Netherlands)

Abstract

Temporal and longitudinal studies collect information repeatedly over time and are aimed to provide insight in joint changes of variables over time. Nowadays, the collected information per time point is often high dimensional, for example images, multiple omics datasets. Dimension reduction methods such as partial least squares for multiple omics or functional principal component analysis for images are available for one time point. However, application of these methods to repeatedly measured high-dimensional datasets is too computational intensive. For images, we propose a novel functional model which includes cross-sectional functional components for the high-dimensional data and components for the temporal or longitudinal dependency and an estimation method which is computationally fast.

Our model is obtained by applying Mercer's theorem to the marginal covariance function of the continuous processes underrlying the discrete observed images. Here, we integrate over time ([1]). The time component is modelled by individual specific temporal score functions. For estimation of the model, we propose a method which comprises a singular value decomposition of the inner products of the underlying processes for each time point seperately and a regression component to estimate the temporal score functions ([2]). The performance of the method is studied via simulations and a data application. Here, we focus on dense temporal data. The simulation study evaluates the ability to reconstruct the data, to estimate the score functions and principal components, and to identify the correct number of components. The dataset consists of brain fMRI images captured on 15 subjects whilst performing decision-making tasks ([2]).

The simulation study showed good performance for obtaining the correct number of components and estimation of the score functions and components when the fitted model agrees with the simulated one. When the fitted and simulated model do not agree, the approach appears to be able to reconstruct the images. Concerning the data application, our model was also able to identify known relevant brain regions. The model can also be used for sparse temporal data by using splines or longitudinal mixed models for the scores.

Bibliography

[1] Park S.Y., Staicu A.M., *Longitudinal functional data analysis*, Stat, 4, (2015), 212-226.
[2] Li Y., Huang C., Härdle W., *Spatial functional principal component analysis with applications to brain image data*, Journal of Multivariate Analysis, 170, (2019), 263-274.

Causal change point detection and localization

Tuesday, 4.07
16:00–18:00
Poster

Shimeng Huang, Rikke Soendergaard, Jonas Peters, Niklas Pfister

University of Copenhagen (Denmark)

Abstract

Testing and estimating change points in sequential data is of interest in many areas. Various notions of change points have been proposed, such as changes in mean, variance, or linear regression coefficients. In this work, we consider settings in which a response variable Y and a set of covariates $X = (X^1, \dots, X^d)$ are observed and aim to find changes in the causal mechanism of how Y depends on X. More specifically, we assume Y depends linearly on a subset of the covariates and aim to determine at what time points either the dependency on the subset or the subset itself changes. We call these time points causal change points (CCPs) which do not necessarily correspond to changes in linear regression coefficients. We propose general methodology to both detect and localize CCPs. Although motivated by causality, we define CCPs without referencing an underlying causal model. The proposed methods exploit a notion of causal invariance which allows to distinguish CCPs from other types of change points. For localization, we propose a loss function that can be combined with existing multiple change point algorithms to localize multiple CCPs and prove consistency of the CCP estimators. We evaluate and illustrate our methods on simulated and real-world datasets.

A maximum likelihood method for the estimation of population growth rate – a case study of Saudi Arabia

Tuesday, 4.07
16:00–18:00
Poster

M. Hafidz Omar, Sulyman Iyanda

Department of Mathematics and Statistics,
King Fahd University of Petroleum and Minerals,
Dhahran, (Saudi Arabia).

Abstract

Nearly everything in our environment changes and grows. Growth and reduction in population size cannot only be found in humans but also in organisms, crystals, animals, economic activity, the quantity of medicine in the blood, simple and compound interest, etc. Growth and decay processes are being described using mathematical models, be it physically, chemically, biologically, or sociologically. However, while the models are not exact representations of reality, they can nonetheless be used to comprehend it and make predictions about it. Exponential growth is one of the most frequent ways that a quantity can change. In this work, we estimate the parameters of an exponential growth model using the maximum likelihood estimator(MLE). In [2], the method of ordinary least squares was used, but it is not as general as MLE and may not be feasible in a case where a probabilistic feature about the random variables is required. Moreover, MLE is not only consistent asymptotically but also efficient asymptotically.

Bibliography

[1] Wachter, Kenneth W., *Essential demographic methods*, Harvard University Press, 2014.

Convergence in law of stochastic processes

Friday, 7.07
12:00–12:30
IS 16
Room 0.03

Adam Jakubowski

Nicolaus Copernicus University, Toruń, Poland
adjakubo@mat.umk.pl

Abstract

In 1953 Yu. Prohorov published a paper on weak convergence of probability measures on metric spaces, bringing a new, extended context to the Invariance Principle proved by Donsker two years earlier. Prohorov's formalism, publicised in books by K.R. Parthasarathy and P. Billinsley, established the equivalence of the notion of convergence in law of stochastic processes and the weak convergence of their distributions. This point of view is completely justified in metric spaces, especially in Polish spaces.

It is, however, much less satisfactory in non-metric spaces, as was shown by examples due to X. Fernique, given long time ago.

We show that in a large class of *submetric spaces* there exists a stronger mode of convergence, coinciding with the weak convergence on metric spaces, and much more suitable for needs of contemporary theory of stochastic partial differential equations.

A submetric space is a topological space $(\mathcal{X}, \tau)$ admitting a continuous metrics d that in turn determines a metric topology $\tau_d \subset \tau$ (where this inclusion is in general *strict*). As a standard (and the simplest) example may serve a separable Hilbert space equipped with the weak topology.

The number of failed components in a coherent system and its application to optimal design

Friday, 7.07
11:40–12:00
CS 31
Room 1.40

Krzysztof Jasiński

Nicolaus Copernicus University (POLAND)

Abstract

We consider a coherent system consisting of components of different types. That is, the system is composed of components having nonidentical failure time distributions. We are interested in determining the number of failed components of each type in a failed (see [1]) or in a operating coherent system (see [2]). We extend the results which are well-known in the literature for k-out-of-n systems. The obtained results are next used in the classical age replacement policy when the corrective replacement of the system is applied if a system has failed, while the preventive replacement is done before failure. We formulate the optimization problem. We wish to find the optimal replacement time.

Bibliography

[1] Jasiński K., *A study on the number of failed components in a failed coherent system consisting of different types of components*, Journal of Computational and Applied Mathematics, (2022), 1–9, https://doi.org/10.1016/j.cam.2022.114839.

[2] Jasiński K., *On the number of failed components in a coherent system consisting of multiple types of components*, Journal of Computational and Applied Mathematics, 410 (2022), 1–8.

Using the Probability of Improved Prediction for Model Selection in the Presence of Outliers

Tuesday, 4.07
11:00–11:20
CS 06
Room 1.40

Stijn Jaspers[1], **Olivier Thas**[1,2,3]

[1] Data Science Institute, I-BioStat, Hasselt University, 3590 Diepenbeek, Belgium
[2] Department of Applied Mathematics, Computer Science and Statistics, Faculty of Sciences, Ghent University, Krijgslaan 281, 9000 Ghent, Belgium[3] National Institute of Applied Statistics Research Australia (NIASRA), University of Wollongong, Northfield Ave, Wollongong, NSW 2522, Australia

Abstract

Outlying observations are known to have a large influence on the **model selection** process. Rabbi et al. (2022) [2] make a comparison of four robust linear model selection criteria that are all based on two key components that are derived from Müller and Welsh (2005) [1], i.e. a good model has:

i) the capability to fit the sample data y and $\mathbf{X}$ reasonably well, and
ii) the ability to predict future observations with great accuracy.

All four methods accommodate for the presence of outliers by considering a bounded robust loss function. The methods differ in the way they quantify the requirements above, with the most advanced method employing an out-of-bag error estimate based on a stratified bootstrap procedure. The values of the respective criteria are computed for a set of candidate models and the optimal model is selected based on the minimum value.

As an alternative, we propose the **Probability of Improved Prediction (PIP)**, which measures how more often a model gives better predictions than another model, where better is determined based on a user-defined loss function. In contrast to the four methods above, the PIP is less sensitive towards the choice of loss function as it directly compares between two models (m^0 and m^1) on an individual level by computing a score between zero and one that reflects the percentage of times a new observation y^* is predicted more accurately by model m^1 as compared to the prediction by model m^0.

A simulation study and data application show the performance of our new concept as compared to the existing methodology. Moreover, although the original four methods could probably be modified to also fit in the setting of more complicated **machine learning models**, we show that the PIP can directly be applied. In this perspective, some results with respect to gradient boosting machines are presented as well.

Bibliography

[1] Müller, S. and Welsh, A. (2005). *Outlier robust model selection in linear regression.* Journal of the American Statistical Association 100(472), 1297-1310 .

[2] Rabbi, F., Khalil, A., Khan, I., Almuqrin, M.A., Khalil, U. and Andualem, M. (2022). *Robust model selection using the out-of-bag bootstrap in linear regression*, Scientific Reports, 12, 1-10.

Prediction for the Weibull-power law process

Wednesday, 5.07
12:20–12:40
CS 17
Room 5440

Alicja Jokiel-Rokita, Ryszard Magiera

Wrocław University of Science and Technology (Poland)

Abstract

The Weibull-power law process (WPLP) is a special case of the trend-renewal process which is used in modelling of the recurrent event data. The research concerns the problem of prediction of the future time of the WPLP event. On the basis of a fairly large set of models of the WPLP, the accuracy of predictors based on maximum likelihood (ML) estimators is confronted with the accuracy of predictors based on Bayesian approach. In the case of the small size of samples of the models concerned, it is shown that the Bayesian approach applied with a non-informative prior distribution leads, in the vast majority of the models concerned, to more accurate predictors than the ML ones.

Time series GLM's by convex programming

Thursday, 6.07
14:00–14:30
IS 09
Room 0.03

Anatoli Juditsky,[1] **Arkadi Nemirovski**[2], **Yao Xie**[2], **Chen Xu**[2]

[1] University Grenoble Alpes (France)
[2] Georgia Institute of Technology (USA)

Abstract

We introduce a new computational framework for estimating parameters of generalized linear models. The proposed approach relies upon a monotone operator-based variational inequalities framework to overcome non-convexity of the loss functions of the parameter estimation problem and leads to non-asymptotic guarantees for parameter recovery. Our focus is on a class of spatio-temporal models which can be seen as a large-scale generalization of Wiener and Hammerstein-Wiener nonlinear autoregressive models [1, 2] with monotone nonlinearity, and also an extension of the popular generalized linear model (GLM) class [3] to account for dependencies among observations in spatio-temporal data.

Proposed estimates are accompanied with online instance-based accuracy bounds which use observations. Such bounds rely upon new "computation counterpart" of classical concentration inequalities for martingales [4] .

Finally, to illustrate the performance of the proposed estimation routines, we discuss results of a preliminary numerical study of Poisson spatio-temporal model based on simulated and real data.

Bibliography

[1] Schetzen, M., *The Volterra and Wiener theories of nonlinear systems*, Wiley, 1980.
[2] Billings, S., Fakhouri, S., Identification of systems containing linear dynamic and static nonlinear elements. *Automatica*, 18(1):15–26, 1982.
[3] Nelder, J.-A., Wedderburn, R., Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.
[4] Bercu, B., Delyon, B., Rio, E., *Concentration inequalities for sums and martingales*. Springer International Publishing, 2015.

A framework for relational event modelling

Thursday, 6.07
12:00–12:30
IS 10
Room 0.06

Rūta Juozaitienė[1], **Ernst Wit**[2]

[1] Vytautas Magnus University (Lithuania)
[2] Università della Svizzera italiana (Switzerland)

Abstract

Temporal network data often encode time-stamped interaction events between senders and receivers, such as co-authoring a scientific article or sending an email. A number of relational event frameworks have been proposed to address specific issues raised by modelling time-stamped data with complex temporal and spatial dependencies. These models attempt to quantify how individuals' behaviour, external factors and interaction with other individuals change the network structure over time. It is often of interest to determine whether changes in the network can be attributed to endogenous mechanisms reflecting natural relational tendencies, such as reciprocity or triadic effects.

We propose to model dynamic structure of reciprocal and triadic effects in relational event networks via stratified baseline hazards. This avoids having to define arbitrary temporal windows for what can be considered reciprocity or triadic closure. A two-step estimation framework is used, based on the stratified Cox proportional hazards model, followed by non-parametric estimation of the stratified cumulative hazard functions.

Moreover, the propensity to form (or receive) ties can also be related to individual actors' attributes. Nodal heterogeneity in the network is often modelled by including actor-specific or dyadic covariates, such as age, gender, shared neighbourhood, etc. However, capturing personality traits such as popularity or expansiveness is difficult, if not impossible. A failure to account for unobserved heterogeneity may confound the substantive effect of key variables of interest. We show how node level popularity in terms of sender and receiver effects may mask ghost triadic effects. These results suggest that unobserved heterogeneity plays a substantial role in REM estimation procedure and influences the conclusions drawn from real-world networks.

Inadmissibility of classical sharp minimax tests in high dimensional change point analysis

Wednesday, 5.07
11:00–11:20
CS 14
Room 1.01

Fabio Kalix[1], **Moritz Jirak**[2]

[1] University of Vienna (Austria)
[2] University of Vienna (Austria)

Abstract

Consider the problem of detecting changes in the mean in a high-dimensional time series. There is, by now, a well-established theory regarding this problem, in particular from an information theoretic point of view. Motivated by a cancellation effect, we design a method that, in a certain way, renders the classical (sharp) minimax test in the sparse regime inadmissible. In extreme cases, the (sharp) minimax optimal test may have zero power with respect to certain components, whereas our method has a convincingly performance. This phenomenon is by no means restricted to change-point problems, but may be equally observed in classical high-dimensional testing problems. Our framework is quite general and contains prominent dynamical systems such as random walks on the regular group, functionals of iterated random systems, functionals of (augmented) Garch models of any order, functionals of (Banach space valued) linear processes or possibly infinite memory Markov chains, dynamical systems arising from SDEs and many more.

Near-Optimal Non-Parametric Sequential Tests and Confidence Sequences with Possibly Dependent Observations

Tuesday, 4.07
11:30–12:00
IS 06
Room 0.03

Aurelien Bibaut, Nathan Kallus[1,2], Michael Lindon[2]

[1] Cornell University
[2] Netflix

Abstract

Sequential testing, always-valid p-values, and confidence sequences promise flexible statistical inference and on-the-fly decision making. However, unlike fixed-n inference based on asymptotic normality, existing sequential tests either make parametric assumptions and end up under-covering/over-rejecting when these fail or use non-parametric but conservative concentration inequalities and end up over-covering/under-rejecting. To circumvent these issues, we sidestep exact at-least-α coverage and focus on asymptotically exact coverage and asymptotic optimality. That is, we seek sequential tests whose probability of ever rejecting a true hypothesis asymptotically approaches α and whose expected time to reject a false hypothesis approaches a lower bound on all tests with asymptotic coverage at least α, both under an appropriate asymptotic regime. We permit observations to be both non-parametric and dependent and focus on testing whether the observations form a martingale difference sequence. We propose the universal sequential probability ratio test (uSPRT), a slight modification to the normal-mixture sequential probability ratio test, where we add a burn-in period and adjust thresholds accordingly. We show that even in this very general setting, the uSPRT is asymptotically optimal under mild generic conditions. We apply the results to stabilized estimating equations to test means, treatment effects, etc. Our results also provide corresponding guarantees for the implied confidence sequences. Numerical simulations verify our guarantees and the benefits of the uSPRT over alternatives.

From embarrassingly parallel learning to minimally coordinated distributed Bayesian data analysis

Thursday, 6.07
16:30–17:00
IS 14
Room 1.01

Samuel Kaski

Aalto Univerity (Finland) and University of Manchester (UK)

Abstract

Embarrassingly parallel Markov Chain Monte Carlo (MCMC) is an effective and straightforward way of distributing Bayesian inference, by simply distributing different MCMC chains to different processors, and then pooling the results. I will start by demonstrating that embarrassingly parallel MCMC can also fail embarrassingly, and introduce ways for fixing the problems. I will proceed to parallelizing also gradient-based inference, and then consider how can we federate even further. I will discuss ways to minimize the need to coordinate the processing, ultimately each party producing and releasing their results independently. I will introduce two approaches: meta-analysis of Bayesian analyses, and privacy-preserving release of twin data sets.

A semiparametric approach for interactive fixed effects panel data models

Thursday, 6.07
11:20–11:40
CS 21
Room 2180

Georg Keilbar[1], **Juan M. Rodriguez-Poo**[2], **Alexandra Soberon**[2], **Weining Wang**[3]

[1] University of Vienna (Austria)
[2] University of Cantabria (Spain)
[3] University of York (UK)

Abstract

This paper presents a new approach for the estimation and inference of the regression parameters in a panel data model with interactive fixed effects. It relies on the assumption that the factor loadings can be expressed as an unknown smooth function of the time average of covariates plus an idiosyncratic error term. Compared to existing approaches, our estimator has a simple partial least squares form and does neither require iterative procedures nor the previous estimation of factors. We derive its asymptotic properties by finding out that the limiting distribution has a discontinuity, depending on the explanatory power of our basis functions which is expressed by the variance of the error of the factor loadings. As a result, the usual “plug-in" methods based on estimates of the asymptotic covariance are only valid pointwise and may produce either over- or under-coverage probabilities. We show that uniformly valid inference can be achieved by using the cross-sectional bootstrap. A Monte Carlo study indicates good performance in terms of mean squared error. We apply our methodology to analyze the determinants of growth rates in OECD countries.

Efficient nonparametric estimation of Toeplitz covariance matrices

Friday, 7.07
11:40–12:00
CS 32
Room 2180

Karolina Klockmann, Tatyana Krivobokova

University of Vienna (Austria)

Abstract

A new nonparametric estimator for Toeplitz covariance matrices based on a periodic smoothing spline estimator of the log-spectral density function is proposed. This estimator is positive definite by construction, fully data-driven and computationally very fast. Moreover, the estimator is shown to be minimax optimal under the spectral norm for a large class of Toeplitz matrices. These results are readily extended to inverses of Toeplitz covariance matrices. Also, an alternative version of the Whittle likelihood for the spectral density based on the Discrete Cosine Transform is proposed.

Bayesian model selection in the space of Gaussian models invariant by permutation symmetry

Thursday, 6.07
11:40–12:00
CS 19
Room 1.01

Piotr Graczyk[1], **Hideyuki Ishi**[2], **Bartosz Kołodziejek**[3]

[1] Université d'Angers (France)
[2] Osaka Metropolitan University (Japan)
[3] Warsaw University of Technology (Poland)

Abstract

Colored Graphical Gaussian models (Graphical Gaussian Models with Edge and Vertex Symmetries) are especially useful when parsimony is needed. Apart from the conditional independence structure, symmetry restrictions are imposed on the concentration or partial correlation matrices. Such symmetries can be represented by a colored graph. Addition of symmetry to the conditional independence restrictions, reduces the number of parameters to estimate. Three types of such models (RCOP among them) were introduced in [3] to describe situations where some entries of concentration or partial correlation matrices are approximately equal. The RCOP models are simplest and most readily interpretable, partly due to their justification through permutation symmetries among the covariates and partly due to additional algebraic structure.

In the talk we will describe Bayesian model selection procedure which applies to the case when conditional dependency graph G is known and one looks for a graph coloring. The talk is based on [1] and [2].

Bibliography

[1] Graczyk, P., Ishi, H., Kołodziejek, B. and Massam, H. (2022) *Model selection in the space of Gaussian models invariant by symmetry.* Ann. Statist. 50, no. 3, pp. 1747–1774.

[2] Graczyk, P., Ishi, H. and Kołodziejek, B. (2022) Graphical *Gaussian models associated to a homogeneous graph with permutation symmetries*, Physical Sciences Forum, 5(1), 20, pp. 1–9.

[3] Højsgaard, S. and Lauritzen, S. L. (2008) *Graphical Gaussian models with edge and vertex symmetries.* J. R. Stat. Soc. Ser. B Stat. Methodol. 70, no. 5, pp. 1005–1027.

Invariant causal prediction for non-additive noise models

Thursday, 6.07
16:00–16:20
CS 27
Room 1.40

Lucas Kook, Sorawit Saengkyongam, Jonas Peters

University of Copenhagen (Denmark)

Abstract

Discovering causal relationships from observational data is a fundamental yet challenging task. For some applications, it may suffice to learn the causal drivers of a given response variable instead of the entire causal graph. Invariant causal prediction (ICP) is a method for causal feature selection which requires data from heterogeneous settings. ICP assumes that the mechanism of the response is the same in all settings and exploits invariance of the conditional distribution of the response given its parents across those settings. The original formulation of ICP for linear models has been extended to general independent additive noise models and to nonparameteric settings using conditional independence testing. However, additive noise models are not suitable for applications in which the response is not measured on a continuous scale, but rather reflects categories or counts, while nonparametric conditional independence testing often suffers from low power. To bridge this gap, we develop ICP for continuous, categorical, count-type, and uninformatively censored responses in parametric transformation models. We propose procedures for testing invariance based on score residuals, establish coverage guarantees and empirically show gains in power over nonparametric alternatives when the model is correctly specified. Our proposed method is implemented in the R package `tramicp`.

Weighted distributions, random truncation, and partial extremes

Wednesday, 5.07
12:00–12:20
CS 17
Room 5440

Charles Amponsah[1], **Tomasz J. Kozubowski**[2]

[1] University of Nevada (USA)
[2] University of Nevada (USA)

Abstract

Many generalized distributions in the literature are weighted distributions with their PDFs proportional to $w(x)f(x)$, where $f(x)$ is a genuine PDF and w(x) is a weight function. We are concerned with the case where the latter is a genuine cumulative distribution function or a survival function, in which case we provide stochastic representations of these constructions related to random truncation and maxima and minima of independent random variables. In connection with this, we introduce the new concept of "partial extremes", and illustrate the latter with multiple examples. In particular, we obtain new interesting results in this setting involving exponential distribution.

Bibliography

[1] Amponsah, C.., Kozubowski, T.J., *Weighted distributions, random truncation, and partial extremes*, Working paper (2023).

Testing For Global Covariate Effects in Dynamic Interaction Event Networks

Tuesday, 4.07
14:20–14:40
CS 12
Room 5440

Alexander Kreiss[1], **Enno Mammen**[2], **Wolfgang Polonik**[3]

[1] Leipzig University (Germany)
[2] Heidelberg University (Germany)
[3] University of California at Davis (United States)

Abstract

In statistical network analysis it is common to observe so called interaction data. In such data the vertices of a network are interpreted as actors who can interact along the edges of the network. We allow that the edges are randomly formed and dissolved over the observation horizon. Generally speaking, we suppose to observe a counting process for each pair of actors. The jump times of the counting processes correspond to the interactions between the respective actors. We understand interactions as events that happen instantaneously at a single time point, e.g., the moment of sending a text message. In addition, covariates are observed and the goal is to model the impact of the covariates on the interactions. We distinguish two types of covariates: global, system-wide covariates (i.e. covariates taking the same value for all individuals, such as seasonality) and local, dyadic covariates describing two actors in the network. It is then natural to formulate a model for the intensity functions of the counting processes in which the global covariates contribute to the baseline, while the local covariates contribute to the pair specific part of the intensity.

Existing continuous time network models are extended to allow for comparing a completely parametric model and a model that is parametric only in the local covariates but has a global non-parametric time component. This allows, for instance, to test whether global time dynamics can be explained by simple global covariates like weather, seasonality etc. Such results, in turn, are useful for predictions. The procedure is applied to a bike-sharing network by using weather and weekdays as global covariates and distances between the bike stations as local covariates.

The talk is based on the pre-print [1]

Bibliography

[1] Kreiss A., Mammen E., Polonik W., *Testing For Global Covariate Effects in Dynamic Interaction Event Networks*, arXiv 2103.14668, 2022.

Bootstrap-based goodness-of-fit test for parametric generalized linear models under random censorship

Thursday, 6.07
14:00–14:20
CS 23
Room 1.01

Gitte Kremling[1,2], **Gerhard Dikta**[1], **Richard Stockbridge**[2]

[1] Fachhochschule Aachen (Germany)
[2] University of Wisconsin-Milwaukee (USA)

Abstract

Many scientific studies are concerned with the duration of time until a predefined terminal event occurs. Take, for example, the time to death after cancer treatment. Due to the nature of such studies, it is possible not to observe the event for all of the individuals during the study, leading to randomly right-censored data. Typically, the interest lies in the distribution of the durations of time dependent on some given covariates such as, for example, the age of the patient or the treatment dose.

In this talk, we consider a parametric generalized linear model for the conditional distribution. In order to check whether the model fits to given data, we propose a bootstrap-based goodness-of-fit test using a modified Kaplan-Meier type empirical process with estimated parameters. Since we are considering a parametric model, the model parameters can be estimated via a maximum likelihood method. As a first step towards justifying the proposed bootstrap method, we establish a functional limit theorem for the modified Kaplan-Meier process. In particular, we show that the process converges to a centered Gaussian process with known covariance structure. For that, we make use of results for the classical Kaplan-Meier type empirical process and its bootstrap version [1, 2]. A bootstrap-based test for nonparametric censored regression is introduced in [3]. Finally, we present a simulation study indicating the validity of the goodness-of-fit test.

Bibliography

[1] Breslow, N. and Crowley, J., *A large sample study of the life table and product limit estimates under random censorship*, The Annals of Statistics, 2.3 (1974), 437-453.

[2] Lo, S.H. and Singh, K., *The Product-Limit Estimator and the Bootstrap: Some Asymptotic Representations*, Probability Theory and Related Fields, 71.3 (1986), 455-465.

[3] Stute, W., González Manteiga, W. and Sánchez Sellero, C., *Nonparametric model checks in censored regression*, Communications in Statistics-theory and Methods, 29.7 (2000), 1611-1625.

Finite sample rates for logistic regression with small noise or few samples

Tuesday, 4.07
11:20–11:40
CS 05
Room 1.01

Felix Kuchelmeister, Sara van de Geer

ETH Zurich (Switzerland)

Abstract

The logistic regression estimator is known to inflate the magnitude of its coefficients if the sample size n is small, the dimension p is (moderately) large or the signal-to-noise ratio $1/\sigma$ is large (probabilities of observing a label are close to 0 or 1). With this in mind, we study the logistic regression estimator, assuming Gaussian covariates and labels generated by the Gaussian link function, with a mild optimization constraint on the estimator's length to ensure existence. We provide finite sample guarantees for its direction, which serves as a classifier, and its Euclidean norm, which is an estimator for the signal-to-noise ratio.
We distinguish between two regimes. In the low-noise/small-sample regime ($n\sigma \lesssim p \log n$), we show that the estimator's direction (and consequentially the classification error) achieve the rate $p \log n/n$, as if the problem was noiseless. In this case, the norm of the estimator is at least of order $n/(p \log n)$, so the signal-to-noise ratio $1/\sigma$ is possibly overestimated. If instead $n\sigma \gtrsim p \log n$, the estimator's direction achieves the rate $\sqrt{\sigma p \log n/n}$, whereas its norm converges to the true norm at the rate $\sqrt{p \log n/(n\sigma^3)}$. As a corollary, the data is not linearly separable with high probability in this regime. We conclude that inference for logistic regression is possible in the regime $n\sigma \gtrsim p \log n$. In either case, it provides a competitive classifier.

Regularity of laws and statistical inference for Markov processes: from LAN property to LAD estimators

Monday, 3.07
16:25–16:45
CS 03
Room 2180

Oleksii Kulyk

Wrocław University of Science and Technology (Poland)

Abstract

The talk is focused on the applications in statistics of stochastic processes of the theoretical results about the small time distributional properties of diffusion processes, Lévy processes, and solutions to SDEs with jump noise. We will discuss statistical models based of high frequency discrete observations of a Markov process whose transition probability density is not available in an explicit form, which requires proper approximation tools for treating the likelikhood function of the model. In that general framework, statistical applications appear to bring new challenges to the classical theory of the heat kernels constructions and estimates. We will discuss several examples of such mutually beneficial links between statistics and general theory of stochastic processes, including the Local Asymptotic Normality (LAN) property or diffusion models and the Least Absolute Deviations (LAD) parameter estimators for Lévy driven SDEs, which motivate new forms of the classical parametrix method or constructing and estimating heat kernels of Markov processes.

The talk is based on [1],[2],[3] and ongoing joint research with Hiroki Masuda (Tokyo), Arturo Kohatsu-Higa (Kyoto) and Dmytro ivanenko (Kyiv).

Bibliography

[1] D. Ivanenko, A.Kulik, H.Masuda, Uniform LAN property of locally stable Lévy process observed at high frequency, *ALEA, Lat. Am. J. Probab. Math. Stat.* 12, (2015), 835 - 862.

[2] A. Kulik, Approximation in law of locally α-stable Lévy-type processes by non-linear regressions. *Electronic Journal of Probability* (2019) vol. 24, art. 83, 1-45.

[3] D. Ivanenko, A. Kohatsu-Higa, A. Kulik, Small time chaos approximations for heat kernels of multidimensional diffusions. *BIT Numerical Mathematics* (2023) 63:16.

Bayesian estimation of causal effects from observational discrete data

Tuesday, 4.07
16:00–18:00
Poster

Vera Kvisgaard[1], **Johan Pensar**[2]

[1] University of Oslo, Oslo, verahk@uio.no
[2] University of Oslo, Oslo, johanpen@uio.no

Abstract

Understanding the effect of interventions in a system of variables is central to many scientific studies. We address the problem of estimating such effects and identifying the strongest pairwise causal relationships among a set of discrete variables based on observational data only. We assume that the data has been generated by an unknown causal Bayesian network, for which the causal structure is represented by a directed acyclic graph (DAG), and that there are no latent confounders. Most of the existing procedures developed for this setting assume a linear Gaussian model, under which causal effects can be estimated by linear regression [1][2][3][4]. In this work we extend the Bayesian IDA (BIDA) [2] approach to categorical variables, making it a relevant tool for a wider set of problems. We use the Jensen-Shannon-Divergence of the multinomial intervention distributions to quantify and rank causal effects between discrete variables. The algorithm builds a posterior distribution of every pairwise causal effect by combining local Bayesian estimation of causal effects, given a specific DAG, and Bayesian model averaging over DAGs. In addition to the exact algorithm, we present an approximate version where the DAG posterior is estimated by means of Markov Chain Monte Carlo (MCMC) sampling. This makes the procedure applicable to larger networks and facilitates identification of different adjustment sets, including the o-set [4].

Bibliography

[1] Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

[2] Johan Pensar, Topi Talvitie, Antti Hyttinen, and Mikko Koivisto. A Bayesian Approach for Estimating Causal Effects from Observational Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5395–5402, 2020.

[3] Jussi Viinikka, Antti Hyttinen, Johan Pensar, and Mikko Koivisto. Towards scalable Bayesian learning of causal DAGs. *Advances in Neural Information Processing Systems*, 33:6584–6594, 2020.

[4] Janine Witte, Leonard Henckel, Marloes H Maathuis, and Vanessa Didelez. On Efficient Adjustment in Causal Graphs. *Journal of Machine Learning Research*, 21:1–45, 2020.

Bayesian ex Post Evaluation of Recursive Multi-Step-Ahead Density Prediction

Monday, 3.07
16:05–16:25
CS 01
Room 1.01

Anna Pajor[1,2], **Jacek Osiewalski**[1], **Justyna Wróblewska**[1], **Łukasz Kwiatkowski**[1]

[1] Krakow University of Economics (Poland)
[2] Jagiellonian Univesrity (Poland)

Abstract

This research is focused on developing a formal Bayesian method of recursive multi-step-ahead density prediction and its ex post evaluation (see [1] for details). Our approach remains within the framework of the standard classical Bayesian paradigm based on the Bayes factor and the likelihood-based update. We propose a new decomposition of the predictive Bayes factor into the product of partial Bayes factors, for both a finite number of consecutive k-step-ahead forecasts and the recursive updates of the posterior odds ratios based on updated data sets. The first factor in the decomposition is related to the relative k-step-ahead forecasting ability of models, while the second one measures the updating effect. To illustrate the usefulness of the proposed measures, we apply the new decomposed predictive Bayes factors to compare forecasting ability of models when the true data generating process (DGP) is known, using simulated data sets. Taking into account the effect of updating the posterior odds ratios leads to the conclusion that the best model coincides with the true DGP. However, the highest k-step-ahead forecasting ability (considered alone, without including the updating effect) can be achieved by some other, less adequate models. Finally, we investigate the predictive ability of different Vector Error Correction (VEC) models with conditional heteroscedasticity, combining three macroeconomic variables: unemployment, inflation and interest rates, separately for the US and Polish economies. The results show that the inference about the models' predictive performance depends on the forecast horizon as well as on taking into account the updating effect.

Bibliography

[1] Pajor A., Osiewalski J., Wróblewska J., Kwiatkowski Ł., *Bayesian ex Post Evaluation of Recursive Multi-Step-Ahead Density Prediction*, Bayesian Analysis, Bayesian Anal. Advance Publication (2023), 1-33.

Smooth $k-$sample tests under left truncation

Wednesday, 5.07
11:20–11:40
CS 18
Room 4420

Adrián Lago[1], **Ingrid Van Keilegom**[2], **Juan Carlos Pardo-Fernández**[1], **Jacobo de Uña-Álvarez**[1]

[1] Universidade de Vigo (Spain)
[2] KU Leuven (Belgium)

Abstract

Left truncation arises in many different applied fields due to the impossibility of observation of every individual that experiments the event of interest, frequently as a result of the way a study is designed or limitations on the measurement instruments. Truncation causes an observational bias which also induces bias in the estimators of different population quantities, such as the density function. A proper estimation of it can be derived from the Lynden-Bell estimator of the survival function for left-truncated data, proposed in [1]. Let us now consider k independent populations. It is a common applied problem to determine whether the target variables are the same in every population. To do that, a test based on an integral distance between the estimator of the density function in each population and the one of the pooled sample is proposed. Its asymptotic distribution will be studied and, due to the difficulty of its application in practice, a bootstrap resampling plan will be proposed to approximate the distribution of the test statistic. The appropriateness of the method will be studied via Monte Carlo simulations. As the test is based on density estimators, the choice of the bandwidth plays an important role on its performance, thus it will be carefully studied. Moreover, a bandwidth to maximize the power under the alternative hypothesis will be proposed, based on a double-bootstrap algorithm as the one of [2]. Lastly, the proposed test will be compared to other tests in the literature for left-truncated data, such as the Kolmogorov-Smirnov and the log-rank, under different simulation scenarios. The performance of the test will be exemplified with real data regarding pregnancy times.

Bibliography

[1] Lynden-Bell, D., *A Method of Allowing for Known Observational Selection in Small Samples Applied to 3CR Quasars*, Monthly Notices of the Royal Astronomical Society, 155 (1971), 95–118.

[2] Cao, R., Van Keilegom, I., *Empirical Likelihood Tests for Two-Sample Problems via Nonparametric Density Estimation*, The Canadian Journal of Statistics, 34 (2006), 61–77.

Resampling methods in conditional independence testing

Thursday, 6.07
14:40–15:00
CS 23
Room 1.01

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

[1] University of Warsaw (Poland)
[2] Warsaw University of Technology (Poland)
[3] Polish Academy of Sciences (Poland)

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 [3] and Conditional Permutation [2] 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. Surprisingly, despite differences in the asymptotic distributions of the probability vector estimates, the distributions coincide for the two scenarios.

In addition to showing asymptotic behaviour of the estimates, we will also provide a non-asymptotic viewpoint, specifically regarding the validity of permutation p-values for the Conditional Permutation scheme.

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 [1].

Bibliography

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

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

[3] 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 knock-offs for high dimensional controlled variable selection. J. R. Stat. Soc., B: Stat. Methodol. 80 (2018), 551–577.

Explicit convergence bounds for Metropolis Markov chains

Friday, 7.07
14:30–15:00
IS 05
Room 0.03

Christophe Andrieu[1], **Anthony Lee**[1], **Sam Power**[1], **Andi Q. Wang**[2]

[1] University of Bristol (UK)
[2] University of Warwick (UK)

Abstract

One of the simplest and enduringly popular general-purpose Monte Carlo Markov chains evolving on R^d is the random walk Metropolis (RWM) Markov chain. Despite its relative simplicity, explicit convergence bounds that scale suitably with dimension have proved elusive for decades. In recent years, progress has been made to show that for distributions with strongly convex and gradient-Lipschitz potentials there exists a specific proposal variance giving an explicit bound on the L^2-mixing time. We refine these results and obtain explicit spectral gap and L^2-mixing time bounds for RWM with arbitrary proposal variances in any dimension, demonstrating the robustness of the algorithm. We obtain the correct scaling with dimension of the spectral gap for sufficiently regular target distributions, and the mixing time bounds are of reasonable (not astronomical) order. Our positive results are quite generally applicable in principle. Essentially the same analysis can be performed for the preconditioned Crank–Nicolson Markov chain, obtaining dimension-independent bounds under suitable assumptions.

Deconvolution with unknown noise distribution and support inference

Tuesday, 4.07
15:00–15:30
IS 18
Room 0.03

Jérémie Capitao-Miniconi[1], **Élisabeth Gassiat**[1], **Sylvain Le Corff**[2], **Luc Lehéricy**[3]

[1] Université Paris-Saclay, CNRS (France)
[2] Sorbonne Université, CNRS (France)
[3] Université Côte d'Azur, CNRS (France)

Abstract

The deconvolution problem consists in recovering the distribution of a signal X based on observations $X + \varepsilon$ perturbed by an independent additive noise ε. Most existing results require the noise distribution to be known–either exactly or through a sample of the noise– in order to find a solution, which can be restrictive in practice.

We show that under very mild assumptions, it is possible to solve the deconvolution problem without any information on the noise distribution. When the signal distribution is absolutely continuous with respect to the Lebesgue measure, we propose an estimator of its density and ascertain its rates of convergence, which are minimax when the signal has compact support.

A direct application of this result is the estimation of the support of the signal. We present simple geometrical conditions on the support that ensure that our identifiability theorem applies and construct an estimator of the support that has minimax adaptive rates of convergence up to a log log factor in Hausdorff distance. These results still work when the distribution of the signal is singular.

This presentation is based on the collaborations [1] and [2].

Bibliography

[1] Gassiat É., Le Corff S., Lehéricy L., *Deconvolution with unknown noise distribution is possible for multivariate signals*, Annals of Statistics, 50.1 (2022), pp. 303–323.

[2] Capitao-Miniconi J., Gassiat É., Lehéricy L., *Support and distribution inference from noisy data*, submitted.

Classification with pairwise Markov models

Wednesday, 5.07
11:40–12:00
CS 17
Room 5440

Jüri Lember

University of Tartu, Estonia

Abstract

We consider a bivariate Markov chain $\{Z_k\}_{k\geq 1} = \{(X_k, Y_k)\}_{k\geq 1}$ taking values on $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is possibly uncountable space and $\mathcal{Y}$ is a finite set, typically referred to as the *state-space*. Process $X = \{X_k\}_{k\geq 1}$ is seen as the observed sequence and $Y = \{Y_k\}_{k\geq 1}$ is seen as hidden or latent variable sequence, often referred to as *the signal process*. The process $Z = \{Z_k\}$ is called *pairwise Markov model* (PMM) and it covers many latent variable models used in practice, such as hidden Markov models (HMM) and autoregressive regime-switching models. Generally, neither Y nor X is a Markov chain, although in many practical models like HMM, Y remains to be a Markov chain. PMM-s is a rich class of models allowing to model: dependent Markov chains; an inhomogenous Markov chain as a marginal process of homogenous Markov chain; a semi-Markov chain as a marginal process of a homogenous PMM [1].

The *decoding or segmentation problem* consists of estimating the unobserved realization of signal process $Y_{1:n}$ given the realization $x_{1:n}$ of the observation process $X_{1:n}$. Probably the most popular estimate is the path with maximum posterior probability. This path will be denoted with $v_{1:n}$:

$$v_{1:n} := \arg\max_{y_{1:n}\in\mathcal{Y}^n} P\big(Y_{t:n} = y_{1:n} | X_{1:n} = x_{1:n}\big).$$

Typically $v_{1:n}$ is called *Viterbi* or *MAP (maximum aposteriori) path*. As it is well known, Viterbi path is not (in general) the most accurate, i.e. the one that minimizes the expected number of errors. The path $w_{1:n}$ that minimizes the expected number of errors – so called *PMAP (pointwise maximum aposteriori*) path – can be obtained pointwise

$$w_t = \arg\max_{y:t\in Y} P\big(Y_t = y | x_{1:n}\big), \quad t = 1, \ldots, n.$$

As it is well known, w might have zero posterior probability. A *hybrid path* operates between the PMAP and Viterbi path and is the solution to the following problem

$$\max_{y_{1:n}\in\mathcal{Y}^n} \big[\sum_{t=1}^{n} \ln P(Y_t = y_t | X_{1:n} = x_{1:n}) + C \ln P\big(Y_{t:n} = y_{1:n} | X_{1:n} = x_{1:n}\big)\big].$$

Clearly $C = 0$ corresponds to PMAP path and when C is big enough, then the output is Viterbi path. However for any $C > 0$, the hybrid path has positive posterior probability.

We define the concept of *m-local Viterbi property* and argue that there exists constants C^m so that whenever $C > C^m$, then the hybrid path is m-locally Viterbi [2]. Then the different hybrid paths tend to differ from each other in larger and larger intervals when C grows.

Bibliography

[1] K. Kuljus, J. Lember, Hybrid classifiers of pairwise Markov models, *Methodology and computing in Applied Probability* (to appear), 2023.
[2] J. Lember, Local Viterbi property in decoding, *Information and Inference* (to appear), DOI: 10.1093/imaiai/iaad004, 2023.

A stationary non-linear time series model with multiple frequencies

Friday, 7.07
11:00–11:20
CS 32
Room 2180

Łukasz Lenart, Łukasz Kwiatkowski, Justyna Wróblewska

Krakow University of Economics (Poland)

Abstract

Modelling data that display cyclical patterns of multiple frequencies has been of substantial interest in many areas. Depending on the phenomenon at hand, the model framework allows for cyclicality in the first, second or even higher-order moments, with either known (fixed) or unknown (estimated) frequencies related to either deterministic or stochastic cyclical pattern.

In this paper, we design a non-linear weak stationary, multi-frequency stochastic cycle model drawing upon the idea of the non-linear innovations state space framework, combined with a direct movement of a sine function with a time-variable amplitude and phase shift driven by autoregression processes. Our specification belongs to the family of exponential smoothing models with a Gaussian error term, and is shown to feature some salient theoretical properties, including weak stationarity (of any order) and a pseudo-cyclical autocovariance function that vanishes similarly to a damped almost periodic function with relevant frequencies. Finally, specification of our model enables a direct decomposition of overall cyclic fluctuations into separate cyclic components corresponding to different frequencies.

Nonparametric Modeling of Event Processes with Applications to Conditional Local Independence Testing

Monday, 3.07
16:45–17:05
CS 04
Room 5440

Myrto Limnios, Niels R. Hansen

Department of Mathematical Sciences, University of Copenhagen,
Universitetsparken 5, 2100 Copenhagen Ø (Denmark)

Abstract

In the context of disease progression analysis, estimating the causal effect of a time-continuous treatment assigned to a given population is an important problem. This is particularly relevant for understanding (the causal) underlying phenomena in many applications gathering massive high-dimensional and time-dependent data structures, ranging from biomedicine to financial markets. In practice, existing learning algorithms inferring the underlying causal graph consider the progression of the recorded markers as time-continuous event processes, for which it is required to specify a model, see e.g., algorithms for directed mixed graphs [4], and the Causal Analysis algorithm [5].

We propose, in this work, a nonparametric model for testing if, a process directly influences another when conditioned on the history of others. Known as the (asymmetric) conditional local independence test, we use a version of the Local Covariance Measure from [1], where we model both the unknown intensity process and the test statistic using their finite order Volterra expansion. This results in a linear combination of tensor products of kernel functions composed of stochastic integrals w.r.t. the event processes observed up to that time. Under some assumptions of sparse decomposition and adequate regularity, the optimal parameters involved in the Volterra expansions are solution of a LASSO penalized method. Finite-sample concentration bounds for the estimation and prediction errors are derived, yielding data-driven optimal weights for the LASSO penalty term, that allow for sparse and heterodastic expansions. These results are obtained by investigating nonasymptotic probabilistic Bernstein bounds for time-dependent martingales, following the works of [2, 3]. Once the optimal parameters are estimated for a particular node of the graph, a practical implementation of the statistical test based on cross-validation and sample splitting is proposed and adapted from [1], that can be plugged in the aforementioned learning graphs algorithms.

Bibliography

[1] Christgau A. M., Petersen L., Hansen N. R., *Nonparametric Conditional Local Independence Testing*, arxiv.2203.13559, 2022.

[2] Bacry E., Bompaire M., Gaiffas S., Muzy J.-F., *Sparse and low-rank multivariate Hawkes processes*, In: Journal of Machine Learning Research, 2020, 21(50), pp. 1–32.

[3] Hansen N. R., Reynaud-Bouret P., Rivoirard V., *Lasso and probabilistic inequalities for multivariate point processes*, In: Bernoulli, 2015, 21(1), pp. 83–143.

[4] Mogensen S. W., Malinsky D., Hansen N. R., *Causal learning for partially observed stochastic dynamical systems*, In: Proceedings of the 34 th conference on Uncertainty in Artificial Intelligence, 2018, pp. 350–360.

[5] Meek C., *Toward learning graphical and causal process models*, In: Proceedings of the UAI 2014 workshop causal inference: Learning and prediction, 2014, pp. 43–48.

Pre-conditioning in Markov chain Monte Carlo

Tuesday, 4.07
11:00–11:20
CS 07
Room 2180

Max Hird, Samuel Livingstone

University College London (UK)

Abstract

Numerous tricks are used to improve the practical performance of a Markov chain Monte Carlo algorithm. One of these, known as pre-conditioning, involves transforming the state variable of interest in such a way that the target distribution for the transformed variable is easier to sample from, and then mapping back afterwards to generate samples from the originally desired distribution.

A common choice is a linear transformation of the state variable (linear preconditioning), which often drastically improves the practical performance of a Markov chain Monte Carlo algorithm, and is implemented in popular software packages (e.g. [1]). Despite this, however, quantifying the benefits of linear preconditioning is not well-studied theoretically, and rigorous guidelines for choosing pre-conditioners are not always readily available. Mixing time bounds for various samplers have been produced in recent works for the class of strongly log-concave target distributions with M-smooth potentials (e.g. [2, 3, 4]). These bounds depend strongly on a quantity known as the condition number.

We study linear preconditioning for this class of distributions, and under appropriate assumptions we provide bounds on the condition number after using a given linear preconditioner. We then provide conditions under which linear pre-conditioning provably increases the spectral gap for certain algorithms. We also find situations in which different linear preconditioning strategies can be ineffective or even detrimental. If time permits we will also briefly discuss non-linear pre-conditioners, which can be connected to Riemannian algorithms (e.g. [5]).

Bibliography

[1] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

[2] Dalalyan, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 651–676.

[3] Dwivedi, R., Chen, Y., Wainwright, M. J., & Yu, B. (2019). Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20, 1–42.

[4] Andrieu, C., Lee, A., Power, S., & Wang, A. Q. (2022). Explicit convergence bounds for Metropolis Markov chains: isoperimetry, spectral gaps and profiles. arXiv preprint *arXiv:2211.08959*.

[5] Girolami, M., & Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2), 123–214.

Bayesian Deep Learning via Function-Space Posterior Inference

Friday, 7.07
12:00–12:30
IS 11
Room 0.03a

Yingzhen Li

Imperial College London, UK

Abstract

The Bayesian deep learning community concerns about Bayesian methods on quantifying uncertainty for deep neural networks. To date, approximate inference methods (incl. variational Bayes [1]) have been researched extensively for neural network weight-space posterior inference [2, 3, 4]. However, ultimately we want to measure uncertainty for neural network outputs, i.e., function-space uncertainty. In this talk, I will review existing approaches applied to function-space posterior inference for neural networks (e.g., [5, 6]). Many of these approaches have connections to Gaussian processes (GPs, [7]), and I will also discuss our latest work on quantifying uncertainty for Transformer models with sparse GP inspired techniques [8].

Bibliography

[1] Beal M J, *Variational algorithms for approximate Bayesian inference*. Ph. D. thesis. University College London, 2003.

[2] Blundell C, Cornebise J, Kavukcuoglu K, et al., *Weight uncertainty in neural networks*. In Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37. 2015: 1613-1622.

[3] Hernández-Lobato J M, Adams R, *Probabilistic backpropagation for scalable learning of bayesian neural networks*. International conference on machine learning. PMLR, 2015: 1861-1869.

[4] Gal Y, Ghahramani Z, *Dropout as a bayesian approximation: Representing model uncertainty in deep learning*. international conference on machine learning. PMLR, 2016: 1050-1059.

[5] Ma C, Li Y, Hernández-Lobato J M, *Variational implicit processes*. In International Conference on Machine Learning. PMLR, 2019: 4222-4233.

[6] Sun S, Zhang G, Shi J, et al., *Functional Variational Bayesian Neural Networks*. International Conference on Learning Representations, 2019.

[7] Snelson E, Ghahramani Z, *Sparse Gaussian processes using pseudo-inputs*. Advances in neural information processing systems, 2005, 18.

[8] Chen W, Li Y, *Calibrating Transformers via Sparse Gaussian Processes*. International Conference on Learning Representations, 2023.

Robust regression revisited

Tuesday, 4.07
09:30–10:30
Plenary Lecture
Room 0.03

Po-Ling Loh

University of Cambridge (United Kingdom)

Abstract

This talk will discuss two projects on robust estimation in the presence of contaminated data, bringing new ideas beyond the framework of traditional M-estimation.

In the first part of the talk, we study the problem of linear regression in a setting where both the covariates and responses may be heavy-tailed and/or adversarially contaminated. We show how to modify the Huber regression estimator by first applying an appropriate "filtering" procedure to the data based on the covariates, and show that in low-dimensional settings, the filtered Huber regression estimator achieves near-optimal error rates. We further show that the commonly used least trimmed squares and least absolute deviation estimators may similarly be made robust to contaminated covariates via the same covariate filtering step.

In the second part of the talk, we study a variant of Newton's method for robust empirical risk minimization, where at each iteration of the optimization algorithm, we replace the gradient and Hessian of the objective function by robust estimators taken from literature on robust mean estimation for multivariate data. After proving a general theorem about the convergence of successive iterates to a small ball around the population-level minimizer, we study consequences of our theory in generalized linear models, when data are generated from Huber's epsilon-contamination model and/or heavy-tailed distributions. We also propose an algorithm for obtaining robust Newton directions based on the conjugate gradient method, which may be more appropriate for high-dimensional settings, and provide conjectures about the convergence of the resulting algorithm. Our algorithm enjoys the fast rates of convergence for successive iterates often achieved by second-order algorithms for convex problems, i.e., quadratic convergence in a neighborhood of the optimum, with a stepsize which may be chosen adaptively via backtracking linesearch.

This is based on joint work with Ankit Pensia (UW-Madison), Varun Jog (Cambridge), Eirini Ioannou (Edinburgh), and Muni Sreenivas Pydi (Paris Dauphine – PSL).

Statistical Discrimination in Stable Matching

Tuesday, 4.07
14:00–14:30
IS 08
Room 0.06

Rémi Castera[1], **Patrick Loiseau**[2], **and Bary S.R. Pradelski**[1]

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, Grenoble, France
[2] Inria, FairPlay Team, Palaiseau, France

Abstract

We study statistical discrimination in matching, where multiple decision-makers are simultaneously facing selection problems from the same pool of candidates. We propose a model where decision-makers observe different, but correlated estimates of each candidate's quality. The candidate population consists of several groups that represent gender, ethnicity, or other attributes. The correlation differs across groups and may, for example, result from noisy estimates of candidates' latent qualities, a weighting of common and decision-maker specific evaluations, or different admission criteria of each decision maker. We show that lower correlation (e.g., resulting from higher estimation noise) for one of the groups worsens the outcome for all groups, thus leading to efficiency loss. Further, the probability that a candidate is assigned to their first choice is independent of their group. In contrast, the probability that a candidate is assigned at all depends on their group, and — against common intuition — the group that is subjected to lower correlation is better off. The resulting inequality reveals a novel source of statistical discrimination.

On data depth for object data

Thursday, 6.07
17:00–17:30
IS 12
Room 0.06

Xiongtao Dai[1], **Sara Lopez-Pintado**[2]

[1] University of California, Berkeley (USA)
[2] Northeastern University (USA)

Abstract

Data depth is a powerful non-parametric tool for analyzing multivariate and functional data. We recently developed a novel exploratory tool for non-Euclidean object data based on data depth, extending Tukey's depth for Euclidean multivariate data [1]. The proposed metric halfspace depth, applicable to data objects in a general metric space, assigns to data points depth values that characterize the centrality of these points with respect to the distribution and provides an interpretable center-outward ranking. Desirable theoretical properties that generalize standard depth properties postulated for Euclidean data are established for the metric halfspace depth. The depth median, defined as the deepest point, is shown to have high robustness as a location descriptor both in theory and in simulation. We propose an efficient algorithm to approximate the metric halfspace depth and illustrate its ability to adapt to the intrinsic data geometry. The metric halfspace depth was applied to an Alzheimer's disease study, revealing group differences in the brain connectivity, modeled as covariance matrices, for subjects in different stages of dementia.

Bibliography

[1] Dai X., Lopez-Pintado S., *Tukey's depth for object data*, Journal of the American Statistical Association, DOI:10.1080/01621459.2021.2011298

Federated learning in medical applications: from theory to practice

Thursday, 6.07
17:00–17:30
IS 14
Room 1.01

Marco Lorenzi

Inria, France

Abstract

This talk will illustrate current advances in federated learning (FL) for sensitive applications, from a both theoretical and practical perspectives. In spite of the wide interest in the federated learning paradigm, current applications to sensitive domains, such as healthcare, are still challenging due to the complexity in dealing with heterogeneous and complex data hosted in different hospitals, as well as to the practical difficulty of deploying federated architectures in the real world.

We will first cover a novel theory for modeling the impact of clients heterogeneity on the convergence guarantees of federated learning [1, 2]. In particular, we will study the robustness and variability of federated learning to heterogeneous conditions, by introducing the notion of stochastic aggregation weights. The proposed framework allows to derive novel federated optimization schemes to maximise the representativity and minimize the variability of clients contributions across federated optimization rounds. We will also introduce a novel perspective to federated unlearning (FU), a novel FL problem aiming at providing theoretical guarantees on the removal of the contribution of a given client from a federated training procedure. Upon unlearning request from a given client, FU is based on the definition of criterion to identifies the optimal FL iteration from which FL has to be reinitialized, along with randomized perturbation mechanism to provide unlearning guarantees [3].

Finally, from the practical standpoint, the talk will introduce Fed-BioMed, a development initiative aiming at translating federated learning to healthcare applications [4]. Fed-BioMed tackles the challenges required to meet real-world translation, concerning FL security, scalability and interoperability. We will provide an illustration of the interplay between methodological development and translational effort that characterise the development of the Fed-BioMed FL platform, and discuss our current effort in delivering FL in hospitals networks.

Bibliography

[1] Fraboni Y., Vidal R., Kameni L. and Lorenzi M. *Clustered sampling: Low-variance and improved representativity for clients selection in federated learning*, In International Conference on Machine Learning (pp. 3407-3416). PMLR.

[2] Fraboni Y., Vidal R., Kameni L. and Lorenzi M. *A General Theory for Federated Optimization with Asynchronous and Heterogeneous Clients Updates*, Journal of Machine Learning Research (110):1-43, 2023.

[3] Fraboni Y., Vidal R., Kameni L. and Lorenzi M. *Sequential Informed Federated Unlearning: Efficient and Provable Client Unlearning in Federated Optimization*, arXiv preprint arXiv:2211.11656.

[4] Cremonesi F., et al.Fed-BioMed: Open, *Transparent and Trusted Federated Learning for Real-world Healthcare Applications*, White paper, arXiv:2304.12012.

On the Properties of Noise Injection in Stochastic Optimization

Thursday, 6.07
14:30–15:00
IS 15
Room 0.06

Aurelien Lucchi[1], **Antonio Orvieto**[2], **Frank Proske**[3], **Francis Bach**[4], **Hans Kersting**[5]

[1] Department of Mathematics & Computer Science, University of Basel, Basel, Switzerland.
[2] Department of Computer Science, ETH Zürich, Zürich, Switzerland
[3] Department of Mathematics, University of Oslo, Oslo, Norway
[4] Inria, Ecole Normale Supérieure, PSL Research University, Paris, France
[5] Yahoo Research, Germany

Abstract

Optimizing machine learning models is almost exclusively done using stochastic gradient descent (SGD) or some variant of it. The randomness of these algorithms can actually be beneficial, as it helps to prevent the algorithm from getting stuck in local optima and can lead to better generalization of the model to new data. This has led to a common procedure where artificial noise is explicitely injected into gradient descent (GD) to improve the performance of machine learning models. Usually, uncorrelated noise is used in such *perturbed gradient descent* (PGD) methods. It is, however, not known if this is optimal or whether other types of noise could provide better generalization performance. In this talk, I will zoom in on the problem of *correlating* the perturbations of consecutive PGD steps. We consider a variety of objective functions for which we find that GD with *anticorrelated* perturbations ("Anti-PGD") generalizes significantly better than GD and standard (uncorrelated) PGD. To support these experimental findings, we also derive a theoretical analysis that demonstrates that Anti-PGD moves to wider minima, while GD and PGD remain stuck in suboptimal regions or even diverge. This new connection between anticorrelated noise and generalization opens the field to novel ways to exploit noise for training machine learning models.

Finally, I will discuss connections between the correlation of pertburbation and a non-Markovian stochastic process known as fractional Brownian motion (fBM) that allows for the increments of the process to be (anti-)correlated. This generalizes processes based on Brownian motion, such as the Ornstein-Uhlenbeck process. We demonstrate how to discretize such processes which gives rise to the new algorithm "PGD". This method is a generalization of the known PGD and Anti-PGD algorithms. We study the properties of fPGD both theoretically and empirically, demonstrating that it possesses exploration abilities that, in some cases, are favorable over PGD and Anti-PGD.

Bibliography

[1] Lucchi, Aurelien, Frank Proske, Antonio Orvieto, Francis Bach, and Hans Kersting. "On the Theoretical Properties of Noise Correlation in Stochastic Optimization". In Advances in neural information processing systems, 2022.

[2] Orvieto, Antonio, Hans Kersting, Frank Proske, Francis Bach, and Aurelien Lucchi. "Anticorrelated noise injection for improved generalization." In International Conference on Machine Learning, 2022.

One-Step Estimation of Differentiable Hilbert-Valued Parameters

Tuesday, 4.07
12:00–12:30
IS 06
Room 0.03

Alex Luedtke, Incheoul Chung

University of Washington (United States)

Abstract

This talk presents estimators for smooth Hilbert-valued parameters, where smoothness is characterized by a pathwise differentiability condition. We show that, when the parameter space is a reproducing kernel Hilbert space, efficient, root-n rate estimators and corresponding confidence sets can be obtained under reasonable conditions. These estimators correspond to generalizations of cross-fitted one-step estimators based on Hilbert-valued efficient influence functions. We provide theoretical guarantees even when arbitrary estimators of nuisance functions are used, including those based on machine learning techniques. In Hilbert spaces without a reproducing kernel, we show that an efficient influence function does not necessarily exist, even though the parameter is pathwise differentiable. To handle these cases, we propose a regularized one-step estimator and corresponding confidence sets. We also show that pathwise differentiability, which is a central requirement of our approach, holds in many cases. In particular, we provide multiple examples of pathwise differentiable parameters and develop novel estimators for function-valued parameters that are of current interest to the causal inference community.

Invited talk: Causality-inspired ML: what can causality do for ML?

Wednesday, 5.07
12:00–12:30
IS 02
Room 0.03

Sara Magliacane[1,2]

[1] University of Amsterdam (Netherlands)
[2] MIT-IBM Watson AI Lab (US)

Abstract

Applying machine learning to real-world cases often requires methods that are robust w.r.t. heterogeneity, missing not at random or corrupt data, selection bias, non i.i.d. data etc. and that can generalize across different domains. Moreover, many tasks are inherently trying to answer causal questions and gather actionable insights, a task for which correlations are usually not enough.

Several of these issues are addressed in the rich causal inference literature. On the other hand, often classical causal inference methods require either a complete knowledge of a causal graph or enough experimental data (interventions) to estimate it accurately. Recently, a new line of research has focused on causality-inspired machine learning, i.e. on the application of ideas from causal inference to machine learning methods without necessarily knowing or even trying to estimate the complete causal graph.

In this talk, I will present an example of this line of research in the unsupervised domain adaptation case [1], in which we have labelled data in a set of source domains and unlabelled data in a target domain ("zero-shot"), for which we want to predict the labels. In particular, given certain assumptions, our approach is able to select a set of provably "stable" features (a separating set), for which the generalization error can be bound, even in case of arbitrarily large distribution shifts. As opposed to other works, it also exploits the information in the unlabelled target data, allowing for some unseen shifts w.r.t. to the source domains. While using ideas from causal inference, our method never aims at reconstructing the causal graph or even the Markov equivalence class, showing that causal inference ideas can help machine learning even in this more relaxed setting.

I will also show two examples of other causality-inspired works in the context of reinforcement learning. In both works, we can only guarantee that we learn the correct causal relations under very specific conditions, but even in the more general conditions we can show that modelling changes in the domain [2] or nonstationarity [3] through a causal lens improves the performance of the algorithms.

Bibliography

[1] Magliacane, S., Van Ommen, T., Claassen, T., Bongers, S., Versteeg, P. and Mooij, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. Advances in neural information processing systems, 31.

[2] Huang, B., Feng, F., Lu, C., Magliacane, S., and Zhang, K. (2022) AdaRL: What, Where, and How to Adapt in Transfer Reinforcement Learning. In International Conference on Learning Representations.

[3] Feng, F., Huang, B., Zhang, K., and Magliacane, S. (2022) Factored Adaptation for Non-Stationary Reinforcement Learning. In Advances in Neural Information Processing Systems.

A network autoregressive model for networks with time-varying edge weights

Tuesday, 4.07
15:00–15:20
CS 12
Room 5440

Anastasia Mantziou[1], **Gesine Reinert**[1,2], **Mihai Cucuringu**[1,2], **Victor Meirinhos**[3], **Francois Lafond**[4]

[1] The Alan Turing Institute
[2] University of Oxford, Department of Statistics
[3] Office for National Statistics
[4] University of Oxford, Institute for New Economic Thinking

Abstract

In economic and financial applications, there is often the need for analysing multivariate time series, comprising of time series for a range of quantities. Particularly, in some applications such complex systems can be associated with some underlying network describing pairwise relationships among the quantities. Accounting for the underlying network structure for the analysis of this type of multivariate time series can be particularly informative for forecasting. Our work is motivated by a data set consisting of time series of firm-to-firm transactions. In this example, pairwise relationships between SIC codes can be represented using a network, while the observed time series for each pair of SIC codes can be regarded as time-varying edge weights. Inspired by [1], we introduce the GNAR-edge model which allows modelling of multiple time series utilising the network structure, assuming that each edge weight depends not only on its past values but also also on past values of its neighbouring edges, for a range of neighbourhood stages. Results from the implementation of the GNAR-edge model on the real firm-to-firm data show good fitting and good predictive performance of the model. In addition, the method is validated through simulations.

Bibliography

[1] Knight M., Leeming K., Nason G., Nunes M., *Generalised network autoregressive processes and the GNAR package.*, arXiv preprint arXiv:1912.04758. 2019 Dec 10.

Space-time moving average models for time series of counts

Friday, 7.07
11:20–11:40
CS 32
Room 2180

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

[1] Institute of Electronics and Informatics Engineering of Aveiro (IEETA) and Department of Electronics, Telecommunications and Informatics (DETI), University of Aveiro, Aveiro, Portugal.
[2] Intelligent Systems Associate Laboratory (LASI), University of Aveiro, Portugal.
[3] Center for Computational and Stochastic Mathematics (CEMAT), Department of Mathematics, IST, University of Lisbon, Lisbon, Portugal.
[4] Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany.

Abstract

This work introduces a new class of models addressed as Space-Time Integer AutoRegressive and Moving Average (STINARMA). These models cope with time series data also exhibiting a spatial dimension, which is quite common in many applied areas. STINARMA models are developed as the integer counterpart of the continuous STARMA models [1], where the spatial component is taken into consideration through a $\boldsymbol{W}^{(\ell)}$ matrix that provides weighted information on the spatial neighbours of order ℓ. The integer nature of the data is conserved by replacing the multiplication for the (multivariate) random binomial thinning operator [2, 3] and the multivariate Gaussian distributed innovation process for a discrete valued one with independent components.

This work focuses the subclass of moving average STINMA($q_{m_1,\ldots,m_q}$) models. First- and second-order moments as well as the space-time autocorrelation (ST-ACF) function are derived to characterise theoretically the STINMA process. Furthermore, we define the theoretical region enclosing the possible ST-ACF values for the STIMA process which can be used as a pratical tool to decide if a given set of data can be properly modelled with the STINMA approach. Due to its high potential in real-data applications, the Poisson STINMA(1_1) is further analysed and several strategies aiming the estimation of the STINMA coefficients given a set of data are derived, namely based on method of moments (MM), conditional least squares (CLS) and conditional maximum likelihood (CLM). The performance of the estimation strategies in a finite sized sample is evaluated via simulation. Finally, the Poisson STINMA(1_1) model is applied to analyse real count data consisting of the daily number of hospital admissions, over time, in different Portuguese locations.

Bibliography

[1] Pfeifer P. E., Deutsch S. J., *A Three-Stage Iterative Procedure for Space-Time Modeling*, Technometrics, 22 (1), (1980), 35-47.
[2] Steutel, F. W., Van Harn, K., *Discrete Analogues of Self-Decomposability and Stability*, The Annals of Probability, 7 (5), (1979), 893-899.
[3] Franke, J. Subba Rao, T. *Multivariate First-Order Integer-Valued Autoregressions*, Technical Report, University of Kaiserslaute, (1993).

Covariance-based soft clustering of functional data based on the Wasserstein-Procrustes metric

Tuesday, 4.07
14:20–14:40
CS 10
Room 1.40

V. Masarotto[1], **G. Masarotto**[2]

[1] Leiden University (NL)
[2] Universitá di Padova (IT)

Abstract

In this work, we address the problem of clustering functional data based on their covariance structure. Previous research in this direction has primarily focused on comparing covariance operators by embedding them in a much larger, linear Hilbert–Schmidt space. However, covariances can essentially be seen as squares of Hilbert–Schmidt operators [1], indicating their intrinsic non-linearity. To tackle this, we propose a framework centered on the Wasserstein distance of Optimal Transport. We contribute a soft clustering methodology based on the Wasserstein-Procrustes distance, where the variability between clusters is penalised by a term proportional to the entropy of the partition matrix. More precisely, assume to have observed N independent samples of functional data, each of size n_i, $\{X_{1,j}\}_{j=1}^{n_1}, \ldots, \{X_{N,j}\}_{j=1}^{n_N}$. The objective is to cluster the corresponding (observed) covariances, denoted by $\widehat{\Sigma}_i$, $i = 1, \ldots, N$, in K groups. Simultaneously, we aim to estimate an $N \times K$ (*soft*) partition matrix $P = [\pi_{i,j}]$ such that $\pi_{i,j} \geq 0$ and $\sum_{j=1}^{K} \pi_{i,j} = 1$, where each element $\pi_{i,j}$ describes the degree of confidence with which the observed covariance $\widehat{\Sigma}_i$ can be assigned to the jth group. Such soft classification allows for clusters to overlap, and accommodates situations where the separation between all or some of the clusters is not well-defined. The proposed soft clustering method characterizes each group using a centroid covariance operator $\overline{\Sigma}_1, \ldots, \overline{\Sigma}_K$ and simultaneously estimates $\overline{\Sigma}_1, \ldots, \overline{\Sigma}_K$ and P minimizing

$$\sum_{i=1}^{N} \sum_{j=1}^{K} \pi_{i,j} (n_i - 1) \Pi^2 (\widehat{\Sigma}_i, \overline{\Sigma}_j)$$

subject to appropriate constraints. This formulation enables each covariance operator to be partially classified into more than one group. Additionally, we discuss techniques for estimating the number of groups, reducing the computational complexity through finite-dimensional low-resolution approximations, and testing for the presence of any cluster structure. The algorithm is illustrated using simulated and real data and comes with an `R` implementation available [2].

Bibliography

[1] Masarotto, V., Panaretos, V. M., Zemel, Y. *Procrustes metrics on covariance operators and optimal transportation of Gaussian processes*, Sankhya A, 81, 172–213 (2019).

[2] Masarotto, V., Masarotto G., *Covariance-based soft clustering of functional data based on the Wasserstein-Procrustes metric.* arXiv preprint arXiv:2212.13287 (2022).

Model-based clustering in simple hypergraphs through a stochastic blockmodel

Tuesday, 4.07
14:00–14:20
CS 12
Room 5440

Luca Brusa[1], **Catherine Matias**[2]

[1] University of Milano Bicocca (Italy)
[2] Sorbonne Université, Université de Paris Cité, Centre National de la Recherche Scientifique (France)

Abstract

We present a new hypergraph stochastic blockmodel and an associated inference procedure for model-based clustering of the nodes in simple hypergraphs. Simple hypergraphs, where a node may not appear several times in a same hyperedge, have been overlooked in the literature, though they appropriately model some high-order interactions (such as co-authorship). The model assumes latent groups for the nodes and conditional independence of the hyperedges given the latent groups. We establish the first proof of generic identifiability of the parameters in such a model. We develop a variational approximation Expectation-Maximization algorithm for parameter inference and node clustering, and derive an integrated classification likelihood criterion for model selection. We illustrate the performance of our algorithm on synthetic data and analyse a real dataset of co-authorship. Our method called `HyperSBM` is implemented in `C++` for efficiency and available as an `R` package at `https://github.com/LB1304/HyperSBM`.

Optimizing the mean-squared error of nested Monte Carlo estimators with application to sensitivity analysis

Thursday, 6.07
15:00–15:20
CS 23
Room 1.01

Henri Mermoz Kouye[1], **Gildas Mazo**[2]

[1] Inria Saclay Île-de-France, EPI Lifeware, Palaiseau (France)
[2] Univ. Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas (France)

Abstract

In simulation experiments, the estimation of an expectation of a function of a conditional expectation requires nested Monte Carlo techniques. Nested Monte Carlo estimators are computationally expensive and, often, biased. Optimizing the number of runs in the inner and outer loops, denoted by m and n, respectively, is therefore crucial. The estimation of sensitivity indices for stochastic models falls within this scope [1].

Let $\hat{\theta}_{n,m}$ be a nested Monte Carlo estimator of some unknown quantity θ. Suppose that $\hat{\theta}_{n,m}$ and θ are d-dimensional vectors and denote by $\Sigma_{n,m}$ and B_m^2 the variance-covariance matrix and the squared bias of $\hat{\theta}_{n,m}$, respectively. For instance, it is common to have $\Sigma_{n,m} = \Sigma'_m/n$ for some variance-covariance matrix Σ'_m. The mean squared error (MSE) of $\hat{\theta}_{n,m}$ is given by $\mathrm{MSE}(\hat{\theta}_{n,m}) = \mathrm{E}\,||\hat{\theta}_{n,m} - \theta||_2^2 = \mathrm{Trace}(\Sigma_{n,m}) + \mathrm{B}_m^2$. If the form of $\hat{\theta}_{n,m}$ is simple enough, $\Sigma_{n,m}$ and B_m^2 can be calculated explicitly, and hence $\mathrm{MSE}(\hat{\theta}_{n,m})$ can be optimized with respect to n and m, improving the rate of convergence of the estimator. Typically, one has $\mathrm{MSE}(\hat{\theta}_{n,m}) = O(1/n + \mathrm{B}_m^2)$.

In more complex situtations, however, the quantity of interest is of the form $g(\theta)$, and hence the estimator becomes $g(\hat{\theta}_{n,m})$, of which the MSE is given by

$$\mathrm{MSE}(g(\hat{\theta}_{n,m})) = \mathrm{E}\,||g(\hat{\theta}_{n,m}) - g(\theta)||_2^2 = \mathrm{Trace}(\Sigma_{n,m}^{(g)}) + \mathrm{B}_m^{(g)2}, \tag{1}$$

where $\Sigma_{n,m}^{(g)}$ and $\mathrm{B}_m^{(g)2}$ are the variance-covariance matrix and the squared bias of $g(\hat{\theta}_{n,m})$, respectively. In general, neither $\Sigma_{n,m}^{(g)}$ nor $\mathrm{B}_m^{(g)2}$ can be calculated, and hence it is difficult to optimize the MSE (1). If the function g has good properties (e.g., Lipschitz), then bounds can be calculated and least convergence rates can be obtained [2]. For more general functions, such as two-times differentiable functions, the problem remains open.

A quantity that bounds from above the MSE in (1) is found for the class of two-times differentiable functions that satisfy a certain regularity condition. The bound, explicit, yields that the MSE converges to zero as $n, m \to \infty$. It is argued that the bound can be interpreted as a bias-variance tradeoff, and a method to optimize it with respect to n and m is given. An application to the estimation of sensitivity indices in stochastic models is presented.

Bibliography

[1] H. M. Kouye and G. Mazo, *Asymptotic control of the mean-squared error for Monte Carlo sensitivity index estimators in stochastic models*, submitted.

[2] T. Rainforth, R. Cornish, H. Yang, A. Warrington, and F. Wood, *On nesting Monte Carlo estimators*, Proceedings of the 35th International Conference on Machine Learning, PMLR 80:4267-4276, 2018.

Pattern Boosting, a statistical learning method for chemistry applications

Tuesday, 4.07
16:00–18:00
Poster

Claudio Meggio[1] **Johan Pensar**[2], **David Balcells**[3] **and Riccardo De Bin**[4]

[1] Department of Mathematics, University of Oslo, Norway, claudm@uio.no
[2] Department of Mathematics, University of Oslo, Norway, johanpen@math.uio.no
[3] Department of Chemistry, University of Oslo, Norway, david.balcells@kjemi.uio.no
[4] Department of Mathematics, University of Oslo, Norway, debin@math.uio.no

Abstract

In many applications, for example when dealing with chemical data, observations in a dataset do not come in the standard tabular format but are represented as complex graphs structures. In such cases, a crucial part of the information is contained in the structure itself. The aim of this work is to develop a model that takes as input graphs of different dimensions and explore them in a efficient way by detecting paths that are the most informative in terms of predicting the properties related to the structure of the graph. Taking advantage of the iterative nature of the boosting approach, we propose an algorithm that at each step expands only the most significant path selected in the previous step so to keep the search space limited, avoiding evaluating complicated structure when not necessary. Every time the decision on which path should be chosen is made in a gradient boosting fashion, that means following the gradient of a loss function. Emphasis is put on the explainability of the predictions, as we are interested in identifying the paths in the graphs that are the most relevant for the prediction task. We apply our novel algorithm to a chemical dataset known as â€ťtmQMgâ€ť[1], which contains 60k molecular graphs, In this dataset the molecules are represented as graphs with atoms as nodes and bonds as edges. The goal is to predict quantum properties of the molecules, such as the HOMO/LUMO gap. These properties usually require heavy computational power to be computed, while our model can to provide comparable results using much less resources.

Bibliography

[1] Hannes Kneiding, Ruslan Lukin, Lucas Lang, Simen Reine, Thomas Bondo Pedersen, Riccardo De Bin, and David Balcells, Digital Discovery, 2023, Advance Article DOI: 10.1039/D2DD00129B

Minimax estimation in the functional regression model with a functional output

Tuesday, 4.07
11:20–11:40
CS 08
Room 5440

Gaëlle Chagny[1], **Anouar Meynaoui**[2] **and Angelina Roche**[3]

[1] University of Rouen Normandie (France)
[2] University Rennes 2 (France)
[3] University Paris Dauphine (France)

Abstract

We consider a functional linear regression model, where both the covariate and the response variable are functional random variables. We address the problem of optimal nonparametric estimation of the conditional expectation operator in this model. A collection of projection estimators over finite dimensional subspaces is first introduce. We provide a non-asymptotic bias-variance decomposition for the Mean Square Prediction error in the case where these subspaces are generated by the (empirical) PCA functional basis. The automatic trade-off is realized thanks to a model selection device which selects the best projection dimensions: the penalized contrast estimator satisfies an oracle-type inequality and is thus optimal in an adaptive point of view. These upper-bounds allow us to derive convergence rates over ellipsoidal smoothness spaces. The rates are shown to be optimal in the minimax sense: they match with a lower bound of the minimax risk, which is also proved. Finally, we conduct a numerical study, over simulated data and over two real-data sets. More details on this work can be found in the preprint [1].

Bibliography

[1] Chagny, G., Meynaoui, A., and Roche, A. (2022). Adaptive nonparametric estimation in the functional linear model with functional output. arXiv preprint arXiv:2203.00518.

Curse of Markovian Memory in Recursive Algorithms

Thursday, 6.07
15:00–15:30
IS 15
Room 0.06

Sean Meyn, Caio Kalil Lauand

Abstract

Theory and application of stochastic approximation (SA) has grown within the control systems community since the earliest days of adaptive control. This paper takes a new look at the topic, motivated by recent results establishing remarkable performance of SA with (sufficiently small) constant step-size $\alpha > 0$. If averaging is implemented to obtain the final parameter estimate, then the estimates are asymptotically unbiased with nearly optimal asymptotic covariance. These results have been obtained for random linear SA recursions with i.i.d. coefficients.

This paper obtains very different conclusions in the more common case of geometrically ergodic Markovian disturbance:

(i) The target bias is identified, even in the case of non-linear SA, and is in general non-zero.

The remaining results are established for linear SA recursions:

(ii) the bivariate parameter-disturbance process is geometrically ergodic in a topological sense;
(iii) the representation for bias has a simpler form in this case, and cannot be expected to be zero if there is multiplicative noise;
(iv) the asymptotic covariance of the averaged parameters is within $O(\alpha)$ of optimal. The error term is identified, and may be massive if mean dynamics are not well conditioned.

The theory is illustrated with application to TD-learning.

Estimation and classification for Positive Unlabelled data under selection bias

Tuesday, 4.07
11:40–12:00
CS 05
Room 1.01

Konrad Furmańczyk[1], **Jan Mielniczuk**[2], **Wojciech Reichel**[3], **Paweł Teisseyre**[2]

[1] Warsaw University of Life Sciences (Poland)
[2] Warsaw University of Technology (Poland)
[3] Nicolaus Copernicus University (Poland)

Abstract

A classification model for partial observability scenario is considered, in which instead of observing a random sample from distribution $P_{X,Y}$ supported on $R^p \times \{0,1\}$, where X is a vector of predictors and Y a class indicator, one observes a random sample from $P_{X,S}$ where S is a binary, censored version of Y in the sense that $P(S = 1|Y = 0, X = x) = 0$ and $P(S = 1|Y = 1, X = x) = e(x)$ with propensity score $e(x)$ being some unknown, usually non-constant, function of predictors' values. Recent advances in estimation of posterior probability $y(x) = P(Y = 1|X = x)$ for such observation scheme with non-constant $e(x)$(i.e. allowing for selection bias, cf [3, 2]) and ensuing classification rules will be compared with a novel proposal. Namely, double logistic model proposed in [1] assumes that both $y(x)$ and $e(x)$ are parametric functions

$$y(x) = \sigma(\beta^T x) \quad e(x) = \sigma(\gamma^T x),$$

where $\sigma(\cdot)$ is a logistic function and $\beta, \gamma \in R^p$. We show that up to the swap, parameters β and γ are identifiable and discuss properties of *non-concave* expected log-likelihood and pertaining ML estimators under the considered scenario. Moreover, double logistic model is considered from the broader approach of model fitting under misspecification (cf [4], [5]).

Bibliography

[1] Furmanczyk K., Mielniczuk J., Reichel W., Teisseyre P. *Joint estimation of posterior probability and propensity score function for positive and unlabelled data*, submitted.

[2] Gong C, Wang Q, Liu, T, Han, B., You, J., Yang, J., *Imstance dependent positive and unlabeled learning with labeling bias estimation*, 2021, IEEE Trans Pattern Anal Mach Intell, pp.1–16.

[3] Bekker, J., Roddrechts, P., Davis, J., *Beyond SCAR assumption for learning from Positive and unlabelled data*In: Úbeda Flores M. et al. (eds.), Proceedings of ECML-PKDD, 2019, pp. 71–85.

[4] Li K, Duan N., *Regression analysis under link violation*, Annals of Statistics, 17 1989, 1009-1052.

[5] Teisseyre, P, Mielniczuk, J, Łazecka M. *Different strategies ofr fitting logistic regression for positve nan unlabelled data*, Proceedings of ICCS 2020,

Random walks with asymmetric time delays

Monday, 3.07
16:05–16:25
CS 04
Room 5440

Jacek Miękisz

University of Warsaw (Poland)

Abstract

Time delays are ubiquitous in biological and socioeconomic processes. They usually lead to periodic oscillations and orbits. Here we discuss other possible effects of time delays, not present in previously studied systems. Random walks with asymmetric time delays (the so-called higher-order Markov chains) are constructed [1]. In such models, the probability of a system going to the right and to the left depends on the difference between two fitness functions in states of the system at two different times. We performed stochastic simulations and made analytical calculations for simplified models (with stepwise fitness functions). We observed new behaviors. Namely, the average position of a random walker (the expected value of the stationary probability distribution) depends on time delays. This is the combined effect of the system's stochasticity and time delays. In the deterministic version of the model, the walker moves along the cycle with an average value equal to the stationary point of the deterministic system without time delays. In the stochastic model with no time delays, the expected value of the stationary probability distribution is also equal to the stationary point of the deterministic system without time delays. We have also observed that by an appropriate symmetric shifting of fitness functions, we can reverse the effect of time delays.

Bibliography

[1] Łopuszański K., Miękisz J., *Random walks with asymmetric time delays*, Phys. Rev. E, 105 (2022), 064131.

Parameter estimation in diffusion and fractional diffusion models

Monday, 3.07
17:05–17:25
CS 03
Room 2180

Yuliya Mishura[1,2]

[1] Taras Shevchenko National University of Kyiv (Ukraine)
[2] Malardalen University (Sweden)

Abstract

This talk is the result of our common research with Kostiantyn Ralchenko and our PhD students. We consider both diffusion models involving standard Brownian motion, such as Cox-Ingersoll-Ross processes and so called Chan-Karolyi-Longstaff-Sanders processes, and the models involving fractional Brownian motion, including models with the mixture of fBms with different Hurst indices. All these models involve several (from 2 to 5) different parameters to be estimated. For example, for CIR model we provide the following parameter estimation: we obtain the rate of convergence in probability of the maximum likelihood drift parameter estimators based on the continuous-time estimators. Then we introduce the discrete versions of these estimators and investigate their asymptotic behavior. In particular, we establish the conditions for weak and strong consistency, asymptotic normality, and get the rate of convergence in probability. Also, we investigate the mixed fractional Brownian motion with trend of the form $X_t = \theta t + \sigma W_t + \kappa B_t^H$, driven by a standard Brownian motion W and a fractional Brownian motion B^H with Hurst parameter H. We develop and compare two approaches to estimation of four unknown parameters θ, σ, κ and H by discrete observations. The first algorithm is more traditional: we estimate σ, κ and H using the quadratic variations, while the estimator of θ is obtained as a discretization of a continuous-time estimator of maximum likelihood type. This approach has several limitations, in particular, it assumes that $H < \frac{3}{4}$, moreover, some estimators have too low rate of convergence. Therefore, we propose a new method for simultaneous estimation of all four parameters, which is based on the ergodic theorem. Finally, we compare two approaches by Monte Carlo simulations. The results of all this work are presented in the papers [1]–[4].

Bibliography

[1] Mishura, Y., Ralchenko, K., Dehtiar, O. *Parameter estimation in CKLS model by continuous observations*, Statistics and Probability Letters, 184, (2022), 109391.

[2] Kukush, A., Lohvinenko, S., Mishura, Y., Ralchenko, K. *Two approaches to consistent estimation of parameters of mixed fractional Brownian motion with trend*, Statistical Inference for Stochastic Processes, 25(1), (2022), pp. 159–187.

[3] Mishura, Y., Ralchenko, K., Zhelezniak, H. *Numerical approach to the drift parameter estimation in the model with two fractional Brownian motions*, Communications in Statistics: Simulation and Computation, (2022) DOI 10.1080/03610918.2022.2099556.

[4] Dehtiar, O., Mishura, Y., Ralchenko, K. *Two methods of estimation of the drift parameters of the Cox-Ingersoll-Ross process: Continuous observations*, Communications in Statistics: Theory and Methods, 51(19), (2022), pp. 6818–6833.

High-dimensional change point estimation using thresholded CUSUM statistics

Wednesday, 5.07
11:40–12:00
CS 14
Room 1.01

Per August Moen[1], **Ingrid K. Glad**[1], **Martin Tveten**[2]

[1] University of Oslo (Norway)
[2] Norwegian Computing Center (Norway)

Abstract

We consider detection and estimation of an unknown number of changes in the mean-vector sequence of high-dimensional isotropic Gaussian vectors. We propose a novel change point estimation algorithm, which we call **E**fficient **S**parsity **A**daptive **C**hange point estimator (ESAC). To detect and estimate change points, ESAC thresholds CUSUM values in a similar fashion as the test statistic proposed by [1], making it adaptive to the sparsity of each change point. That is, ESAC handles changes that occur in one, some, or all of the components of the mean vector, without the sparsity being known to the user beforehand. Furthermore, the spacing between consecutive change points, as well as the affected coordinates at each change point, are allowed to differ, making ESAC multiscale. We prove finite sample theoretical results, both for the single and multiple change point case, which guarantee that ESAC successfully estimates change points with a given error rate. To search for multiple change points, ESAC uses Seeded Binary Segmentation [2]. ESAC is therefore computationally efficient, obtaining a computational complexity that is linear up to logarithmic factors. Through extensive numerical studies we demonstrate that ESAC is highly competitive in terms of statistical accuracy and computational complexity.

Bibliography

[1] Liu, H, Gao, C, Samworth, R., *Minimax rates in sparse, high-dimensional change point detection* (2021), The Annals of Statistics, 5(2), 1081-1112.

[2] Kovacs, S., Buhlmann, P., *Seeded binary segmentation: a general methodology for fast and optimal changepoint detection* (2022), Biometrika, 110(1), 249-256.

GAMLSS models to explain and predict the length of hospital stays caused by climate change and its effects on health insurance sector

Thursday, 6.07
17:00–17:20
CS 29
Room 4420

Roberto Morales-Arsenal[1], **Ángela Rabadán-Navarrete**[2]

[1] CUNEF Universidad (Spain)
[2] Análistas Financieros Internacionales (Spain)

Abstract

Climate change is producing adverse effects that are being ignored by the insurance industry. The threat posed by the manifestation of climate change, which is called *physical risk*, refers to the modifying effect on the severity and frequency of a wide range of underwritten risks[1] of the insurance sector. There is little literature regarding the potential impact that climate change will have on the insurance sector, and especially in the health branch [1]. Most are limited to the field of health and/or environmental sciences. This study develops a methodology that evaluates the effect of $\mathsf{CO_2}$ concentrations on the duration of hospital stays and, therefore, on hospital spending in the insurance sector in Spain. For this purpose, a semi-parametric regression analysis has been applied using the generalized additive model for position, scale and shape with Box-Cox power exponential distribution (BCPE) whose main advantage is its flexibility. The following model is estimated:

$$\mathsf{CO_2} \sim \mathsf{BCPE}(\mathsf{GDP}, \mathsf{CAR.SIZE}|\mu, \sigma, \upsilon, \tau) + \varepsilon \tag{1}$$

The R package gamlss is used. This package uses two algorithms (*backfitting algorithms*) to fit the model, the CG and RS algorithms [3]. The model explains 94% of the variability of $\mathsf{CO_2}$ levels, which in turn determine the inter-annual variability of stays by 85%. Additionally, a novel $\mathsf{CO_2}$ prediction method has been developed whose predictive capacity was satisfactory. The result of the forecast exercise indicates that the annual days of hospital stays for 2019 were going to increase by 7% over the actual days recorded in 2018, for the total of the Spanish territory. The actual increase recorded during 2019 was 3%, which leaves an operating margin of 4%; this differential constitutes a conservative approach in the estimation. Annual data from 2005 to 2019 disaggregated by Autonomous Communities have been used. It is concluded that the model is useful for assessing the impact of $\mathsf{CO_2}$ pollution on future claims, which would be useful for surcharging premiums and reducing underwriting risk.

Bibliography

[1] Zhao, H. Y., Jianxiong, Zhenwei, T., Tao, L., Weilin, Z., Xing, L., Caiyan, H., Shengyong, W., Zhao, W., Wenjun, M.: *Ambient carbon monoxide and increased risk of daily hospital outpatient visits for respiratory diseases in dongguan, china.* Science of The Total Environment 668 (2019), 254–260

[2] Rigby, R.A., Stasinopoulos, D.M., *A semi-parametric additive model for variance heterogeneity.* Statistical Computing 6 (1996), 57–165

[1] Underwriting risk is the risk of loss or adverse change in the value of the commitments arising from the insurance activity, due to the inadequacy of the pricing and provisioning assumptions

On a General Class of Orthogonal Learners for the Estimation of Heterogeneous Treatment Effects

Thursday, 6.07
15:00–15:20
CS 24
Room 1.40

Paweł Morzywołek[1], **Johan Decruyenaere**[1,2], **Stijn Vansteelandt**[1]

[1] Ghent University (Belgium)
[2] Ghent University Hospital (Belgium)

Abstract

Motivated by applications in personalized medicine and individualized policy making, there is a growing interest in techniques for quantifying treatment effect heterogeneity in terms of the conditional average treatment effect (CATE). Some of the most prominent methods for CATE estimation developed in recent years are T-Learner [1], DR-Learner [2] and R-Learner [3]. The latter two were designed to improve on the former by being Neyman-orthogonal [4, 5]. However, the relations between them remain unclear, and likewise does the literature remain vague on whether these learners converge to a useful quantity or (functional) estimand when the underlying optimization procedure is restricted to a class of functions that does not include the CATE. In this work, we provide insight into these questions by discussing DR-learner and R-learner as special cases of a general class of Neyman-orthogonal learners for the CATE, for which we moreover derive oracle bounds. Our results shed light on how one may construct Neyman-orthogonal learners with desirable properties, on when DR-learner may be preferred over R-learner (and vice versa), and on novel learners that may sometimes be preferable to either of these. Theoretical findings are confirmed using results from simulation studies on synthetic data, as well as an application in critical care medicine.

Bibliography

[1] Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B., *Metalearners for estimating heterogeneous treatment effects using machine learning*, Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(10):4156–4165.

[2] Kennedy, E. H., *Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects*, arXiv, 2020, 1–37.

[3] Nie, X. and Wager, S., *Quasi-oracle estimation of heterogeneous treatment effects*, Biometrika, 2021, 108(2):299–319.

[4] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., *Double/debiased machine learning for treatment and structural parameters*, Econometrics Journal, 2018, 21(1):C1–C68.

[5] Foster, D. J., Syrgkanis, V., *Orthogonal statistical learning*, arXiv, 2019, 1–86.

On polynomial-time algorithms for data depths

Thursday, 6.07
16:00–16:30
IS 12
Room 0.06

Jérémy Guerin[1], **Pavlo Mozharovskyi**[1]

[1] LTCI, Telecom Paris, Institut Polytechnique de Paris (France)

Abstract

Data depth is a statistical function [6, 7], which, by introducing centrality-based ordering, generalizes concepts of median, quantiles, and ranks to higher dimensions [1], and eventually for more complex data. Being data-driven, it is non-parametric and possesses attractive properties of robustness and affine invariance. In its variety of depth notions and types of treated data, throughout decades data depth has become a comprehensive methodology providing a universal ordering and scoring of data. Although data depth has become increasingly important in applications, practitioners are limited by computational burden of algorithms often necessitating computational time and resources growing exponentially with space dimension (for exact calculation) [2]. To deal with non-polynomial complexity, approximating techniques [3] can be often sufficient.

Restricting to the multivariate setting, to cope with computational intractability, we introduce a novel class of depth functions: depths that can be formulated as polynomial optimization problems. This class is sufficiently large to contain (many of) the most commonly used depth notions, including those of local nature describing data with non-convex support. On the other hand, it is unified enough to suit the framework of the moment-sum-of-squares hierarchy suggested by [5] (see also [4] for a statistical application of the sum-of-squares methodology), which formulates into a sequence of relaxations of semi-definite programming problems. This sequence forms a continuum of approximations ending (in a verifiable way) into the exact computation. While it is too early to conclude about resolving the computational-time issue, the suggested methodology allows for a new stream of optimization techniques to compute (exactly and approximately) data depths.

Bibliography

[1] Chernozhukov, V., Galichon, A., Hallin, M., Henry, M., *Monge–Kantorovich depth, quantiles, ranks and signs*, The Annals of Statistics 45 (2017), 223–256.

[2] Dyckerhoff, R., Mozharovskyi P., *Exact computation of the halfspace depth*, Computational Statistics and Data Analysis, 98 (2016), 19–30.

[3] Dyckerhoff, R., Mozharovskyi, P., Nagy, S., *Approximate computation of projection depths*, Computational Statistics and Data Analysis, 157 (2021), 107166.

[4] Hopkins S. B., *Mean estimation with sub-Gaussian rates in polynomial time*, The Annals of Statistics, 48 (2018), 1193–1213.

[5] Lasserre, J. B., *The Moment-SoS hierarchy*, In: Sirakov, B. et al. (eds.), Proceedings of the International Congress of Mathematicians, Rio de Janeiro, 2018, 3773–3794.

[6] Tukey, J. W., *Mathematics and the picturing of data*, In: James, R. (ed.), Proceedings of the International Congress of Mathematicians, Vancouver, 1975, 523–531.

[7] Zuo, Y., Serfling, R., *General notions of statistical depth function*, The Annals of Statistics, 28 (2000), 461–482.

Advantages of nonparametric testing in spatial statistics

Monday, 3.07
16:15–16:45
IS 04
Room 0.06

Tomáš Mrkvička

University of South Bohemia (Czech Republic)

Abstract

Parametric methods in spatial statistics are well-developed, but these methods often lead to problems when the model is wrongly specified. The pitfall of the wrong specification may come from misspecification of the autocorrelation structure of random fields, misspecification of the form of points interaction, misspecification of the model of dependence of one spatial object to other spatial objects, or misspecification of the dependence of data on the sample locations, etc. Such misspecifications may lead to the liberality of the tests higher than 0,25 when 0,05 is the target nominal level. In addition, the parametrical methods are usually based on asymptotic approximations, which can be for real data far from reality due to the long range of dependencies. These issues can lead to slight liberality, which can be higher than 0,08 according to our simulation studies if all the assumptions of the parametric models are met.

On the other hand, the nonparametric methods are free of these assumptions, and they are designed to work correctly even in non-asymptotic situations due to the resampling strategies. Therefore, we built the R package NTSS, which collects Nonparametric Tests in Spatial Statistics. The package contains tools for selecting the relevant spatial covariates influencing the point pattern based on the independence test of the point pattern and a covariate with the presence of the nuisance covariates. Further, it contains tools for disentangling the dependence between points, marks, and covariate, the test of independence between two point patterns and two random fields.

Bibliography

[1] Dvořák J., Mrkvička T. *Nonparametric testing of the covariate significance for spatial point patterns under the presence of nuisance covariates*, `https://arxiv.org/abs/2210.05424` (2022).

[2] Dvořák J., Mrkvička T., Mateu J., Gonzales J. *Nonparametric testing of the dependence structure among points-marks-covariates in spatial point patterns*, International Statistical Review , 90/3 (2022).

[3] Mrkvička T., Dvořák J., Gonzales J., Mateu J. *Revisiting the random shift approach for testing in spatial statistics*, Spatial statistics , 42, 100430 (2020).

Statistical depth in machine learning: Kernel mean embeddings in functional data analysis

Thursday, 6.07
16:30–17:00
IS 12
Room 0.06

George Wynne[1], **Stanislav Nagy**[2]

[1] University of Bristol (United Kingdom)
[2] Charles University (Czech Republic)

Abstract

Statistical depth is the act of gauging how representative a point is compared to a reference probability measure. The depth allows introducing rankings and orderings to data living in multivariate, or function spaces.

This contribution highlights how the common h-depth [1] and related statistical depths for functional data can be viewed as a kernel mean embedding, a technique used widely in statistical machine learning. This connection facilitates answers to open questions regarding statistical properties of functional depths, as well as it provides a link between the depth and empirical characteristic function-based procedures for functional data. In particular, we obtain tractable conditions under which h-depths characterise all probability distributions on an infinite-dimensional space, the first result of this type in the literature on (functional) depth.

Bibliography

[1] Cuevas, A., Febrero, M., Fraiman, R., *On the use of the bootstrap for estimating functions with functional data*, Computational Statistics & Data Analysis, 51 (2006), 1063–1074.

Are the order statistics ordered? (revisited)

Tuesday, 4.07
14:00–14:20
CS 13
Room 4420

Jorge Navarro*

Universidad de Murcia (Spain)

Abstract

If $X_1, \ldots, X_n$ are the random variables representing some sample values, the associated (increasing) ordered values $X_{1:n} \leq \cdots \leq X_{n:n}$ are know as *order statistics*. Several properties for them in the case of IID (independent and identically distributed) samples can be seen in [1].

If $X_1, \ldots, X_n$ represent the component lifetimes of a system, then the ordered values $X_{1:n} \leq \cdots \leq X_{n:n}$ represent the lifetimes of k-out-of-n systems (systems that works when at least k components work). In this case, the Xs can be dependent and they can be heterogeneous, that is, they are not ID.

In both cases, to get stochastic comparisons for them is a relevant topic. The main stochastic orders are the stochastic (st), hazard rate (hr) and likelihood ratio (lr) orders. The main properties for them can be seen in [11].

In some case the order statistics are ordered (as expected). However, this is not always the case. In this talk we will present conditions to get stochastic comparisons in these orders for the order statistics. The conditions will depend on the copula C that models the dependence between $X_1, \ldots, X_n$. The conditions will be connected with dependence properties of C. The talk is based on the results published in [2, 3, 4].

Bibliography

[1] Arnold B.C., Balakrishnan N., Nagaraja, H.N. *A First Course in Order Statistics.* SIAM, 2008.

[2] Navarro J. *Introduction to System Reliability Theory.* Springer, 2022.

[3] Navarro J., Durante F., Fernández-Sánchez J. *Connecting copula properties with reliability properties of coherent systems.* Applied Stochastic Models in Business and Industry 37 (2021), 496–512.

[4] J. Navarro, N. Torrado, Y. del Águila (2018). Comparisons between largest order statistics from multiple-outlier models with dependence. Methodology and Computing in Applied Probability 20, 411–433.

[5] Shaked M., Shanthikumar, J.G. *Stochastic Orders.* Springer, New York, 2007.

*JN thanks the partial support of Ministerio de Ciencia e Innovación of Spain under grant PID2019-103971GB-I00/AEI/10.13039/501100011033 and grant TED2021-129813A-I00 funded by MCIN/AEI/10.13039/501100011033 and the European Union 'NextGenerationEU'/PRTR.

High-dimensional variable selection in nonlinear mixed effects models using a stochastic EM spike-and-slab

Monday, 3.07
16:25–16:45
CS 02
Room 1.40

Marion Naveau, Guillaume Kon Kam King, Renaud Rincent, Laure Sansonnet, Maud Delattre

Université Paris-Saclay (France)

Abstract

High-dimensional variable selection is widely documented in standard regression models, but there are still few tools to address it in nonlinear mixed-effects models, where data is collected repeatedly on several individuals. For all $1 \leq i \leq n$ and $1 \leq j \leq n_i$, y_{ij} the response of individual i at time t_{ij} is modelled as follows:

$$y_{ij} = g(\varphi_i, t_{ij}) + \varepsilon_{ij},\ \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$
$$\varphi_i = \mu + \boldsymbol{\beta}^\top V_i + \xi_i,\ \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(0, \Gamma),$$

where g is non-linear with respect to an individual parameter φ_i which is q-dimensional. Identifying the relevant covariates for all individual parameters amounts to selecting the support of $\boldsymbol{\beta} \in \mathcal{M}_{p\times q}$, defined by $S^* = \left\{ (\ell, m) \in \{1, \dots, p\} \times \{1, \dots, q\} \middle| \beta^*_{\ell m} \neq 0 \right\}$, where $\boldsymbol{\beta}^*$ is the true fixed effects matrix, and p is the number of covariates. To solve this problem in a high-dimensional context, that is when $p >> n$, the assumption of sparsity is made, that is each row of $\boldsymbol{\beta}^*$ is sparse. The main difficulty here is that variable selection concerns latent variables of the model.

In this work, variable selection is approached from a Bayesian perspective and a selection procedure is proposed, combining the use of a spike-and-slab prior [3] and the SAEM algorithm [2]. The first step of this procedure is to reduce the number of candidate models: similarly to Lasso regression, a grid of values for the spike parameter is explored to obtain a collection of promising sub-models $(\widehat{S}_{\nu_0})_{\nu_0 \in \Delta}$. Then, an information criterion can be used to choose the final model: a pragmatic and effective choice is to use the eBIC (extended Bayesian Information Criterion, [1]) which is tailored to the high-dimensional setting.

This approach is much faster than a classical MCMC algorithm and shows very good selection performances on simulated data. The efficiency of the proposed method is illustrated on a problem of genetic markers identification, relevant for genomic assisted selection in plant breeding. The current aim is to achieve consistency in model selection for this problem, which is a work in progress.

Bibliography

[1] Chen J., Chen Z., *Extended Bayesian information criteria for model selection with large model spaces*, Biometrika, 95 (2008), 759-771.

[2] Delyon B., Lavielle M., Moulines E., *Convergence of a stochastic approximation version of the EM algorithm*, Annals of statistics (1999), 94–128.

[3] George, E. I., McCulloch R. E., *Approaches for Bayesian variable selection*, Statistica sinica, 7 (1997), 339–373.

Adaptive testing for multiplicative convolution

Thursday, 6.07
11:00–11:20
CS 20
Room 1.40

Sergio Brenner Miguel, Jan Johannes, Bianca Neubert

University of Heidelberg (Germany)

Abstract

Given observations from a positive random variable contaminated by a multiplicative measurement error, we consider the task of non-parametric goodness-of-fit testing of its unknown density f against a given density f_o in a non-asymptotic framework. More precisely, we consider a framework analogous to [1], who considered an additive measurement error. Estimation in the multiplicative measurement error framework has already been considered for multiple cases, e.g. in [2] for the global estimation of the density f. We propose a direct and indirect testing procedure (c.f. [3] in circular convolution) based on a comparison of the Mellin transform of the density f_o and a truncated version of the empirical Mellin transform based on the estimation strategy proposed in [4]. We derive non-asymptotic radii of testing over Mellin-Sobolev spaces naturally characterising regularity and ill-posedness in this model, for which parameters are unknown in practice. Therefore, adaptive testing strategies are investigated. Considering multiple testing procedures, we obtain assumption-free procedures and analyse their performance. Compared with the non-adaptive tests, their radii of testing face a deterioration by a log-factor.

Bibliography

[1] Butucea C., *Goodness-of-fit testing and quadratic functional estimation from indirect observations*, The Annals of Statistics, 2007.

[2] Brunel E., Comte F., Genon-Catalot V., *Nonparametric density and survival function estimation in the multipl. censoring model*, TEST, 2016.

[3] Schluttenhofer S., Johannes J., *Adaptive minimax testing for circular convolution*, Mathematical Methods of Statistic, 2020.

[4] Brenner Miguel S., Comte F., Johannes J., *Spectral cut-off regularisation for density estimation under multiplicative measurement errors*, Electronic Journal of Statistics, 2021.

False discovery rate control using debiased graphical Lasso in Gaussian graphical models

Thursday, 6.07
12:20–12:40
CS 19
Room 1.01

Ensiyeh Nezakati, Eugen Pircalabelu

Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain (Belgium)

Abstract

Gaussian graphical models (GGMs), which represent the dependence structure among a set of random variables, have many applications in the real world. For example, the conditional dependence relationships between gene expressions in genetic studies can be modeled by GGMs. Inference on the structure of GGMs is challenging when the dimension is larger than the sample size. Many classical methods do not work anymore.

One of the methods of GGM estimation is based on regularized optimization, for instance, the well-known method of graphical Lasso could be used in such a case (see [1]). Due to high-dimensionality and small sample sizes in many modern applications, graphical Lasso estimation often has large uncertainty and low power in detecting the presence of edges between nodes in the graph. Our purpose in this presentation is to propose a multiple testing procedure, where we can control the false discovery rate (FDR), defined as the expected proportion of false positives among all rejections. Our procedure relies on simultaneous tests of the form

$$H_{0,ab} : \mathbf{\Theta}_{ab} = 0 \quad \text{vs} \quad H_{1,ab} : \mathbf{\Theta}_{ab} \neq 0,$$

for $1 \leq a < b \leq p$, where $\mathbf{\Theta}_{ab}$ is the (a, b)-th element of $\mathbf{\Theta}$, the inverse covariance matrix of a multivariate Gaussian distribution. In [2], the author proposed a test statistic based on a bias correction version of the sample covariance coefficients of residuals which controls the FDR. Recently, [3] proposed a procedure using debiased lasso estimator to control directional FDR in high-dimensional linear regression models. In this presentation, we introduce a multiple testing procedure which is suitable for high-dimensional settings and its test statistic is constructed based on the debiased graphical Lasso estimator. As it is shown in [2], the debiased graphical Lasso estimator is asymptotically normal, under some sparsity conditions on $\mathbf{\Theta}$. We show that this asymptotic normality holds uniformly in $1 \leq a < b \leq p$, and the test statistic based on this estimator controls the FDR below a pre-specified level $q \in [0, 1]$. Numerical performance of this procedure is investigated via a simulation study and the results show that the procedure performs well in controlling the FDR.

Bibliography

[1] Friedman J., Hastie T., Tibshirani R., *Sparse inverse covariance estimation with the graphical Lasso*, Biostatistics, 9 (2008), 432–441.

[2] Jankova J., van de Geer S., *Confidence intervals for highdimensional inverse covariance estimation*, Electronic Journal of Statistics, 9 (2015), 1205–1229.

[3] Javanmard A., Javadi H., *False discovery rate control via debiased lasso*, Electronic Journal of Statistics, 13 (2019), 1212–1253.

[4] Liu W., *Gaussian graphical model estimation with false discovery rate control*, The Annals of Statistics, 41 (2013), 2948–2978.

Conditional Independence for Continuous Time Bayesian Networks

Thursday, 6.07
14:40–15:00
CS 24
Room 1.40

Wojciech Niemiro[1,2], **Łukasz Rajkowski**[1]

[1] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw
[2] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University in Toruń

Abstract

Continuous Time Bayesian Networks (CTBNs) are multivariate Markov processes on finite state spaces. They serve as simple and flexible models which explicitly represent temporal dynamics in probabilistic causal reasoning. They were introduced (at least under this name and in this context) by Nodelman et al. (2003). The causal dependences among components of a CTBN are described by a directed graph with possible cycles. The aim of our work is to express some independence relations of the subprocesses that compose a CTBN in terms of separability conditions of the underlying graph. We continue the line of research initiated by V. Didelez (2007 and other papers). We consider both the usual probablistic notion of conditional independence and the notion of "causal independence" which is defined in terms of conditional-by-intervention probability distributions, see Ay & Polani (2008). The results are largely parallel to those obtained earlier for discrete time multivariate processes. Some of the proofs go through without much change, some other are quite different. As a by-product of this research, we also spotted and partly filled in several gaps in the existing literature. In particular, we clarify the role played by the initial distribution of the process.

Bibliography

[1] V. Didelez, Graphical models for composable finite Markov processes. *Scandinavian Journal of Statistics* 34, pages 169–185 (2007).

[2] U. Nodelman, C.R. Shelton and D. Koller, Continuous time Bayesian networks. In Proceedings of the *18th Annual Conference on Uncertainty and Artificial Intelligence* (UAI-2002), pages 378–387 (2003).

[3] N. Ay and D. Polani, Information flows in causal networks, *Advances of Complex Systems* 11, 1 pages 17–41 (2008).

Mendelian randomisation: why do we need to talk about pleiotropy?

Thursday, 6.07
17:20–17:40
CS 27
Room 1.40

Christopher Oldnall

The University of Edinburgh (United Kingdom)

Abstract

The days of classical association are coming to the near end in the world of genomics and biology, with many turning to a Causal Inference framework in an attempt to pursue true causal relationships. One such example is the well utilised framework of **Mendelian randomisation** [2]. Through the use of very large data sets, referred to casually as 'omics' data, we can look to form a causal directed acyclic graph (DAG) which determines the true causal relationship between an exposure (typically a transcriptomic or proteomic feature) and outcome (such as some form of cancer) by introducing the genomic data as an **instrumental variable**.

Sadly however as time has gone on, it has been discovered that the conditions of the Mendelian framework are often broken by complex biological situations, such as the situation of **pleiotropy**. As such we now require some extended form of the Mendelian randomisation framework to ensure we are discovering the true causal relationship and not simply introducing the genomic data as an additional confounder of the exposure and outcome. For this, the work of Sun et al [1] proves to be a good, yet mathematically complex, solution.

This talk will briefly cover the concept of the Mendelian randomisation framework and the way in which the situation of pleiotropy breaches the **exclusion criteria** before discussing a solution to the problem we face. This will be done through the introduction of a newly introduced estimator which is asymptotically efficient, proven through simulations. Following this, the talk will mainly focus on exploring a couple of new and novel applications and results will be presented through the use of large biobank data from the UK. These results will show that we must address the issue of pleiotropy in a real world setting. The results will be compared to traditional Mendelian randomisation methods so the need for a pleiotropic robust estimator can be transparent.

Bibliography

[1] Sun B., Liu Z., Tchetgen Tchetgen E., *Semiparametric Efficient G-estimation with Invalid Instrumental Variables*, Online, 2021 [https://arxiv.org/abs/2110.10615, accessed: 13-03-23].

[2] Sanderson, E et. al, *Mendelian randomization*, Nature Reviews Methods Primers, 2022, Volume 2.

Techniques for Estimating Conditional Shapley Values and When to Use Them

Thursday, 6.07
17:00–17:20
CS 28
Room 2180

Lars Henry Berge Olsen[1], **Ingrid Kristine Glad**[1], **Martin Jullum**[2], **Kjersti Aas**[2]

[1] University of Oslo (Norway)
[2] Norwegian Computing Center (Norway)

Abstract

Complex machine learning (ML) models are extensively applied to solve supervised learning problems in numerous fields, and they (often) obtain accurate predictions, but at the cost of interpretability. The lack of understanding of how the input features of the ML model influence the model's output is a major drawback, hence, the field of explainable artificial intelligence (XAI) has become an active research field proposing several types of explanation frameworks [1]. The most commonly used framework to explain ML-predictions is *Shapely values*, which is a promising model-agnostic explanation methodology with a strong mathematical foundation and unique theoretical properties from the field of cooperative game theory [2, 3, 4].

There are several algorithmic approaches for computing different types of Shapley value explanations; here 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, i.e., we need to model the input variables' dependencies [4]. [5] explicitly states that an important future research direction for conditional Shapley values is the development of new approaches and a systematic evaluation of existing approaches. In our work, we develop new methods, extend earlier proposed approaches, and systematize them into different classes for comparison and evaluation. We conduct a series of extensive simulation studies to evaluate how precisely the different techniques estimate the conditional expectations, and thereby the conditional Shapley values, for different setups. Finally, we apply the techniques on several real-world data examples and provide recommendations for when to use the different approaches and techniques.

Bibliography

[1] Molnar C., *A Guide for Making Black Box Models Explainable* (2nd ed.), christophm.github.io/interpretable-ml-book, 2022.
[2] Shapley L. S., *A value for n-person games*, Contributions to the Theory of Games, 2:28 (1953), 307–317.
[3] Lundberg S., Lee S., *A unified approach to interpreting model predictions*, Advances in neural information processing systems (2017), 4765–4774.
[4] 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.
[5] Chen H., Covert I., and Lundberg S., Lee S., *Algorithms to estimate Shapley value feature attributions*, arXiv preprint arXiv:2207.07605 (2022).

Quantifying the stability and vulnerability of ecological networks

Tuesday, 4.07
14:40–15:00
CS 12
Room 5440

Karol Opara

Systems Research Institute, Polish Academy of Sciences (Poland)

Abstract

Network analysis is a growing branch of ecology. Feeding relationships in ecosystems are described by directed, weighted graphs representing biomass transfer from primary producers through herbivores and carnivores to detritus (dead organic matter). Having assembled the world's largest database of empirical, weighted foodwebs, we investigate relationships between their stability and vulnerability. Properties of our networks, such as hierarchical structure with link strengths differing by orders of magnitude or non-uniformity in taxonomic resolution, necessitate careful choice of vulnerability indicators. We also use several classical and modern methods to ensure the interpretability of our models and assess their robustness. In the talk, we will focus on the statistical aspects of this collaborative study.

Bibliography

[1] Landi, Pietro, et al. *Complexity and stability of ecological networks: a review of the theory.* Population Ecology 60 (2018): 319–345.

[2] Kazanci, C. and Ma, Q. *Chapter 3 – System-wide measures in ecological network analysis.* In Park, Y.-S., Lek, S., Baehr, C. and Jørgensen, S. E. (eds.) Advanced Modelling Techniques Studying Global Changes in Environmental Sciences, vol. 27 of Developments in Environmental Modelling, 45-68 (Elsevier, 2015).

Inequalities for martingale transforms with unbounded coefficients

Wednesday, 5.07
11:00–11:20
CS 17
Room 5440

Tomasz Gałazka[1], **Adam Osękowski**[2]

[1] University of Warsaw (Poland)
[2] University of Warsaw (Poland)

Abstract

Suppose that $f = (f_n)_{n\geq 0}$ is a real-valued martingale and $v = (v_n)_{n\geq 0}$ is a predictable sequence. Then the transform of f by v is the martingale $g = (g_n)_{n\geq 0}$ given by

$$g_n = g_0 v_0 + \sum_{k=1}^{n} v_k(g_k - g_{k-1}), \qquad n = 0,\, 1,\, 2,\, \ldots.$$

This is the discrete-time version of the Itô stochastic integral. A celebrated result of Burkholder [1] states that for each $1 < p < \infty$ there is a finite constant C_p such that

$$\|g\|_{L^p} \leq C_p \|v\|_{L^\infty} \|f\|_{L^p}.$$

This important estimate has turned out to play an important role in many areas of mathematics, including probability theory, stochastic analysis, harmonic analysis, geometry of Banach spaces and operator theory, and has been extended in numerous directions. In particular, the question about the optimal (i.e., the least) value of the constant C_p has significant implications. In the later paper [2], Burkholder introduced a general approach which can be used to study sharp versions of such estimates, and applied it to identify the optimal (i.e., the least) value of C_p allowed in the above L^p bound. It turns out that the best choice is $\max\{p-1, (p-1)^{-1}\}$.

The purpose of the talk will be to discuss the further development of this direction. We will study the context in which the L^∞-boundedness of the predictable sequence v is replaced by the corresponding L^r-boundedness. Precisely, we will be interested in sharp versions of the estimates

$$\|g\|_{L^p} \leq C_{p,q,r} \|v\|_{L^r} \|f\|_{L^q}$$

and its weak-type counterparts.

Bibliography

[1] Burkholder, D. L., *Martingale transforms*, Annals of Mathematical Statistics 37 (1966), 1494–1504.

[2] Burkholder, D. L., *Boundary value problems and sharp inequalities for martingale transforms*, Annals of Probability 12 (1984), 647–702.

Functional data analysis for time-varying networks

Tuesday, 4.07
16:00–18:00
Poster

Puchong Paophan, Georgios Aivaliotis, Leonid Bogachev, Haiyan Liu

University of Leeds (UK)

Abstract

Statistical analysis of network data is challenging due to lack of Euclidean structure in data spaces (see [2] for an exposition). A straightforward vectorisation suffers from masking the spatial information about the nodes and edges. Analysis of time-varying networks, with time-dependent values at their nodes and/or edges, is even more challenging. However, time-varying networks are increasingly used to represent data in various fields, such as social interactions, brain connectivity, transportation systems, etc. (e.g., see [1]). A suitable methodology for this kind of data may be based on functional data analysis, with particular models of interest including functional principal component analysis (FPCA) as well as clustering and vector linear regression. Recently, a general framework for functional data analysis for time varying objects was proposed in [1], based on endowing the data metric spaces with a linear structure by introducing quasi-covariance.

In the present work, we focus on time-varying weighted undirected simple networks which can be represented as graph Laplacian matrices. We transform graph Laplacian matrices (point-wise with respect to time) into a Euclidean (tangent) space and define variances of the transformed networks. Then we perform FPCA and obtain the Karhunen-Loéve representation of the data [3]. Finally, we project this representation back to the original space. Asymptotic properties for our estimators are shown. We demonstrate the effectiveness of this methodology by simulations and by implementing it on the New York City taxi data [4]. For the latter, we can cluster and interpret the patterns of taxi usage during a week through our FPCA results. Our findings agree with the ones in [1].

Bibliography

[1] Dubey P., Müller H.-G., *Functional models for time-varying random objects*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82 (2020), 275–327.

[2] Kolaczyk, E.D., *Statistical analysis of network data: Methods and models*, Springer, 2009.

[3] Ramsay J. O. and Silverman B. W., *Functional data analysis*, Springer, 2005.

[4] *TLC Trip Record Data* (Yellow Taxi, Jan–Dec 2016), NYC Taxi & Limousine Commission, `https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page`.

Introducing a Moving Aggregate to Copula-based Time Series Models to allow for Infinite Autoregressive Order

Thursday, 6.07
11:40–12:00
CS 22
Room 5440

Sven Pappert

TU Dortmund University (Germany)

Abstract

Copula-based time series models have become popular tools in modeling the temporal dependence of stationary time series. With these models, a wide range of (non-linear) autoregressive dependence structures are attainable. However, the models can only incorporate a finite autoregressive order. In a recent work, it was shown that a possible way to allow for infinite partial dependence is to reparametrize the model in terms of classical ARMA parameters [1]. This approach, however, only works with one-parametric copulas. In this work we allow for infinite partial dependence by introducing a new class of copula-based time series models, where a moving aggregate (MAG) part is incorporated in the model updating equation. Working in a D-vine framework (as in e.g. [3], [2], [1]), we derive the non-linear AR(∞) and MAG(∞) representations and investigate the distributional properties of the process. In an application to quarterly US inflation data, we find that our model is more suitable with regards to information criteria compared to the similar model from [1]. Our model is able to incorporate copulas with an arbitrary amount of parameters (particularly two-parametric copulas). It is (almost) as easy to implement and estimate as classical ARMA models and offers a wide variety for modeling temporal dependence structures.

Bibliography

[1] McNeil, A. J., and Bladt, M., *Time series with infinite-order partial copula dependence*, Dependence Modeling (2022).

[2] Nagler, T., Krüger, D., and Min, A., *Stationary vine copula models for multivariate time series*, Journal of Econometrics 227.2 (2022): 305–324.

[3] Smith, M., Min, A., Almeida, C., & Czado, C., *Modeling longitudinal data using a pair-copula decomposition of serial dependence*, Journal of the American Statistical Association 105.492 (2010): 1467–1479.

Estimating the shape functions

Tuesday, 4.07
12:00–12:20
CS 08
Room 5440

María Dolores Jiménez-Gamero[1], **Juan Carlos Pardo-Fernández**[2]

[1] Universidad de Sevilla (Spain)
[2] Universidade de Vigo (Spain)

Abstract

Arriaza *et al.* (2019) (see [1]) introduced the *left shape* function and the *right shape* function, which, in some stochastic sense, synthesize the form of the distribution and can be employed to study the behavior of the tails and the symmetry of a random variable. Specifically, let X be a continuous random variable with probability density function f and distribution function F. For each $u \in (0,1)$, let $x_u = F^{-1}(u) = \inf\{x : F(x) \geq u\}$. The left shape function and the right shape function of X are defined as

$$L_X(u) = \mathbb{E}\{(X - x_u)^- f(X)\}, \quad u \in (0,1),$$

and

$$R_X(u) = \mathbb{E}\{(X - x_u)^+ f(X)\}, \quad u \in (0,1),$$

respectively, provided that the expectations exist, where $x^- = \max\{0, -x\}$ and $x^+ = \max\{0, x\}$, $\forall x \in \mathbb{R}$.

In this talk, we will propose nonparametric estimators of those functions. The estimators involve nonparametric estimation of the quantile and density functions. Pointwise and uniform consistency are proved under general regularity assumptions, as well as the limit in law. Simulations will be shown to study the practical performance of the proposed estimators. Finally, the analysis of a real data set will illustrate the proposed methodology.

Bibliography

[1] Arriaza A., Di Crescenzo A., Sordo M.A., Suárez-Llorens A., *Shape measures based on the convex transform order*, Metrika, 82 (2019), 99–124.

Likelihood-based Inference for Skewed Responses in a Crossover Trial Setup

Tuesday, 4.07
16:00–18:00
Poster

Savita Pareek, Kalyan Das and Siuli Mukhopadhyay

Indian Institute of Technology Bombay (India)

Abstract

This work proposes a statistical model for crossover trials with multiple skewed responses measured in each period. A 3×3 crossover trial data where different doses of a drug were administered to subjects with a history of seasonal asthma rhinitis to grass pollen is used for motivation. In each period, gene expression values for ten genes were measured from each subject. It considers a linear mixed effect model with skew normally distributed random effect or random error term to model the asymmetric responses in the crossover trials. The article examines cases (i) when a random effect follows a skew-normal distribution, as well as (ii) when a random error follows a skew-normal distribution. The EM algorithm is used in both cases to compute maximum likelihood estimates of parameters. Simulations and crossover data from the gene expression study illustrate the proposed approach.

Keywords

Crossover design, Mixed effect models, Skew-normal distribution, EM algorithm.

Towards a Measure-Theoretic Axiomatisation of Causality

Tuesday, 4.07
16:00–18:00
Poster

Junhyung Park[1], **Simon Buchholz**[1], **Bernhard Schölkopf**[1], **Krikamol Muandet**[2]

[1] Max Planck Institute for Intelligent Systems, Tübingen (Germany)
[2] CISPA – Helmholtz Center for Information Security, Saarbrücken (Germany)

Abstract

Causality is a major topic of research in many scientific domains, and in recent years, there has been a steep rise of interest from researchers in statistics and machine learning. There are many frameworks that mathematise the concept of "causality", most prominently, structural causal models (SCMs) [1], [2] and potential outcomes [3]. However, we argue that there is no universally-accepted, axiomatic framework of causality. This stands in stark contrast to *probability theory*, whose goal is to mathematise the concept of "randomness", and has a near-universally accepted axiomatisation based on measure theory [4], which is written in textbooks and taught in classrooms as *the* theory of probability.

Of course, the need for the theory of causality arose from the fact that probability theory cannot encode causal relationships. However, we argue that causality should be viewed as an extension of probability theory, and that the basic building blocks of probability theory, namely the set of outcomes, σ-algebra of events and probability measures, can, and should, be used in the quest to axiomatise the concept of causality. We also take the view that causality is in essence a study of *what happens when one intervenes on the system*, a view shared by many existing frameworks of causality and philosophers [5]. To that end, we propose *causal spaces*, which consists of a probability space along with a collection of transition probability kernels, called *causal kernels*, that encode the causal information of the space, and we also define *interventions* on causal spaces.

Our proposed framework is not only rigorously grounded in measure theory, like the axiomatisation of probability theory, but it also overcomes many limitations of SCMs in a natural way, for example cycles, latent variables, marginalisation and continuous time stochastic processes. It can be shown that systems that can be represented by existing frameworks are instances of causal spaces. We stress that our goal is not to replace existing frameworks. We recognise their value in areas that our framework falls pitifully short, in particular in interpretability and identification, and we foresee that researchers and practitioners will continue to use existing frameworks. Rather, our goal is to provide some possible first steps towards an axiomatisation of causality in the most general sense.

Bibliography

[1] Pearl J., *Causality*, Cambridge University Press, 2009.
[2] Peters J., Janzing D., Schölkopf B., *Elements of Causal Inference*, MIT Press, 2017.
[3] Imbens G., Rubin D., *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*, Cambridge University Press, 2015.

[4] Kolmogorov, A., *Foundations of the Theory of Probability*, NY: Chelsea Publishing Co, 1933.
[5] Woodward, J., *Making Things Happen: A Theory of Causal Explanation*, Oxford University Press, 2005.

Nonparametric Inference from Level Crossings for Time Warping Deformation Models

Thursday, 6.07
11:00–11:20
CS 22
Room 5440

Miroslaw Pawlak

AGH University of Science and Technology (Poland) and University of Manitoba (Canada)

Abstract

We study the problem of nonparametric inference for the time warping deformation model of stationary stochastic processes [1]. The time warping model plays an important role in a number of real-world applications and also is an inherent part of nonstationary signal processing where it is interpreted as the local signal bandwidth [3]. We examine the local bandwidth recovery problem based on samples obtained from level crossings of the output process. Utilizing the celebrated Rice theory [2] for the average number of level-crossings of stochastic processes we establish the link between the local bandwidth and the process local intensity function. A kernel estimate of the local bandwidth is proposed and its asymptotic properties are established. The positivity and often bandlimitness of the local bandwidth requires the proper correction of the kernel estimate. This goal is achieved by the choice of bandlimited and positive kernel functions. The asymptotic theory is based on the local martingale characterization of the level crossing counting process.

Bibliography

[1] Clerc M., Mallat S., *Estimating deformations of stationary processes*, The Annals of Statistics, 31 (2003), 1772–1821.

[2] Lindgren G., *Gaussian integrals and Rice series in crossing distributions – to compute the distribution of maxima and other features of Gaussian processes*, Statistical Science, 34 (2019), 100–128.

[3] Rzepka D., Pawlak M., Koscielnik D., Miskowicz M., *Bandwidth estimation from multiple level-crossings of stochastic signals*, IEEE Transactions on Signal Processing, 65 (2017), 2488–2502.

High-dimensional change-point regression with structured information

Monday, 3.07
16:15–16:45
IS 01
Room 0.03

Florian Pein[1], **Hyeyoung Maeng**[2], **Paul Fearnhead**[1], **Idris Eckley**[1]

[1] Lancaster University (UK)
[2] Durham University (UK)

Abstract

Observing a large number of time series of related processes at the same time occurs in more and more applications. Consequently, recently many works have considered the detection of change-points that occur in multiple of the time series simultaneously. Sharing the information of the change-point location among different time series increases detection power heavily. In this talk we will consider the situation where we additionally know that all changes have the same sign, i.e. all changes at a certain location are either upwards or downwards. We will propose methodology that have a larger detection power if this situation is given. We will discuss application where this setting occurs, show the increase of detection power in a minimax theory and in numerical simulations.

Invariance in practice: dynamical systems and falsifiability

Thursday, 6.07
09:30–10:30
Plenary Lecture
Room 0.03

Jonas Peters

ETH Zurich (Switzerland)

Abstract

Assume that we observe data from a response Y and a vector X of covariates under different experimental conditions (or environments). Rather than focusing only on predictability, it has been suggested to consider the invariance of a model. This can help us to infer causal effects or find models that generalize better. We consider these ideas in the light of two applications and discuss how to adapt existing methodology.

Frequentist coverage of Bayesian credible intervals: a predictive matter?

Wednesday, 5.07
11:30–12:00
IS 03
Room 0.06

Sonia Petrone

Bocconi University, Milan

Abstract

The talk moves from fundamental properties of the predictive distributions for exchangeable sequences, that will be briefly reviewed. We will then illustrate some overlooked implications in Bayesian learning. In particular, we will focus on inference on the unknown distribution $F(x)$ and give a Gaussian asymptotic approximation of the posterior distribution which suggests a predictive interpretation of (lack of) frequentist coverage of Bayesian credible intervals. This is joint work with Sandra Fortini.

Estimating location parameters of two exponential distributions with ordered scale parameters

Thursday, 6.07
15:00–15:20
CS 25
Room 2180

Lakshmi Kanta Patra[1], **Constantinos Petropoulos**[2], **Shrajal Bajpai**[1]

[1] Department of Mathematics, Indian Institute of Technology Bhilai (India)
[2] Department of Mathematics, University of Patras, Rio, Patras (Greece)

Abstract

In the present work, component-wise estimation of location parameters of two exponential distributions with ordered scale parameters under a bowl-shaped location-scale invariant loss function. We have studied the invariant estimation problem. It is shown that the best affine equivariant estimator (BAEE) is inadmissible. For this purpose, we propose stein-type non-smooth dominating estimator. Consequently, we have demonstrated the inadmissibility of UMVUE and MLE by presenting improved estimators. Using Kubokawa's IERD approach, a class of dominating estimators is obtained. It is noted the boundary estimator of the dominating class is the Brewster-Zidek estimator. We have derived the improved estimator for squared error loss and linex loss function for application. Finally a simulation has been carried out for implementation purpose.

Clustered Mallows Model

Wednesday, 5.07
12:00–12:20
CS 16
Room 2180

Luiza Piancastelli, Nial Friel

[1] University College Dublin (Ireland)

Abstract

Rankings are a type of preference elicitation that arise from experiments where judges are asked to arrange objects in decreasing order of utility. Orderings of a set $\{1, \ldots, n\}$ denoted π reflect strict preferences amongst the objects, where $\pi \in \mathcal{P}_n$ the space of permutations of n items. For a variety of reasons, strict relations can be unrealistic assumptions in practical situations. One example is **I:** the case that alternatives share common traits and henceforth could easily be indistinguishable. With moderate or large n, it becomes likely that evaluators are indifferent in some of their choices. Another possibility (**II**) is that, depending on the experiment, there can be a different importance attribution to the choices within π. For example, judges could be mostly concerned with demonstrating their top and disfavored alternatives. In this type of *top/bottom* elicitation, middle-rank items could be a close to uniform placement of those remaining.

Our work extends the famous Mallows ([1]) model to accommodate indistinguishbilities in the form of **(I)** and **(II)**. The underlying groupings of items/choices within these scenarios motivates the naming Clustered Mallows Model (CMM). In addition to providing the flexibility to mix strict and indifference preferences, the CMM can also serve as a simplified representation of ranking data under large n. The proposed probability model has a non analytical normalisation, which places CMM Bayesian inference in the class of doubly-intractable problems. We overcome this challenge the specification of a rejection sampler and the exchange algorithm ([2]). Information criteria are addressed to select between alternative groupings and real data analysis illustrate the CMM usefulness in practical situations.

Bibliography

[1] Mallows, C. L., *Non-Null Ranking Models. I*, Biometrika, 44 (1957), 114-130.

[2] Murray, I., Ghahramani, Z., MacKay, D., *MCMC for doubly-intractable distributions*, Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence, AUAI Press (2006).

On Lasso estimator for the drift function in diffusion models

Thursday, 6.07
16:30–17:00
IS 17
Room 0.03

Gabriela Ciolek[1], **Dmytro Marushkevych**[2], **Mark Podolskij**[1]

[1] University of Luxembourg (Luxembourg)
[2] University of Copenhagen (Denmark)

Abstract

In this paper we study the properties of the Lasso estimator of the drift component in the diffusion setting. More specifically, we consider a multivariate parametric diffusion model X observed continuously over the interval $[0, T]$ and investigate drift estimation under sparsity constraints. We allow the dimensions of the model and the parameter space to be large. We obtain an oracle inequality for the Lasso estimator and derive an error bound for the L^2-distance using concentration inequalities for linear functionals of diffusion processes. The probabilistic part is based upon elements of empirical processes theory and, in particular, on the chaining method.

Worst-case optimality of the Risk Inflation Criterion for prediction and selection in linear models

Tuesday, 4.07
11:40–12:00
CS 06
Room 1.40

Piotr Pokarowski

Faculty of Mathematics, Informatics and Mechanics Iniversity of Warsaw, *pokar@mimuw.edu.pl*

Abstract

Risk Inflation Criterion (RIC) selects a subset of linear model predictors $J \in \{1, \ldots, p\}$ minimizing $RSS(J) + \lambda|J|$, where $\lambda = 2log(p)\sigma^2$ and σ^2 is a known value or an upper bound on the variance of Gaussian errors. The lambda penalty formula was derived by minimizing the upper bound on the predictive MSE in the worst case (Foster and George, Ann. Statist. 1994). The aim of the presentation will be to show a generalization of RIC for sub-Gaussian errors and to derive a similar formula as a penalty leading to consistent selection with minimal identifiability of the model.

A Central Limit Theorem for Centered Purely Random Forests using U-Statistic Theory

Thursday, 6.07
15:20–15:40
CS 23
Room 1.01

Jan Rabe[1], **Natalie Neumeyer**[1], **Mathias Trabs**[2]

[1] Universität Hamburg (Germany)
[2] Karlsruhe Institute of Technology (Germany)

Abstract

Random forests are supervised learning methods that can be applied to regression and classification problems. In a regression setting a random forest averages the results of several randomized regression trees that are constructed on different subsamples of the dataset. In practice random forests appear to be successful, even when applied to high dimensional data. The statistical analysis of random forests is challenging, especially due to the dependence of the partition on the data. Even for the simpler purely random forests which omit this dependence the majority of the literature focuses on consistency results. In our work we prove an explicit central limit theorem for centered purely random forests. The proof uses results from [1] which are based on an interpretation of random forests as generalized U-Statistics. The CLT allows us to construct pointwise confidence intervals for the regression function. Moreover we discuss the construction of confidence bands.

Bibliography

[1] Peng W., Coleman T., Mentch L., *Rates of convergence for random forests via generalized u-statistics*, Electronic Journal of Statistics 16(1) (2022), pp. 232–292.

A Novel Way for Inflation Forecasting? The Case for the Close End Formulas

Monday, 3.07
16:25–16:45
CS 04
Room 5440

Yedidya Rabinovitz

University of Warsaw (Poland)

Abstract

This study introduces a novel class of close-end Inflation Forward Rate Formulas. The fundamental idea behind this formula class is that the inflation process possesses continuous (gBm) and mean reversion characteristics, as they emerge from the economic nature of supply and demand from all segments of the economy. The key to derive this forward measure is the proof that the Instantaneous Mean Reversion formula is a martingale shown via the Markov Property Theorem. Consequently, the Feynman-Kac theorem is satisfied, and the Inflation Forward Measure is derived. Two more derivations follow the Forward Diffusion Equation (Fokker-Plank), and Backward Diffusion Equation (Kolmogorov) transformations. To complete the inflation forward measure class, an additional Diminish Diffusion Equation is presented. A comparison of alternative prominent benchmarks and this formula class are analyzed and compared under two distinct monetary regimes, the gold and fiat standards, and five major economies. The time interval forecast is annual. The analysis demonstrates the usefulness of this class to capture inflation expectations and underlying dynamics over the benchmarks.

Bibliography

[1] Fergusson, K. (2020). *Forecasting inflation using univariate continuous-time stochastic models*. Journal of Forecasting, 39(1), pp. 37–46.

[2] Feynman, R. P. (1948). *Space-time approach to non-relativistic quantum mechanics*. Reviews of Modern Physics, 20(2), 367–387.

[3] Kac, M. (1951). *On some connections between probability theory and differential and integral equations*. Proceedings of the Second Berkeley Symposium, 189–215.

[4] Rabinovitz, Y. (2017). *A new S.D.E. and instantaneous mean reversion rate formula (presented via a numerical empirical model comparison)*. International Journal of Financial Engineering, 04(02n03), 1750029.

[5] Ross, Sheldon M. (1996). *Stochastic processes*. Second edition New York: Wiley.

[6] Stewart, J (1991). *Calculus*, 2nd edn. Brooks/Cole Publishing Company.

[7] Shreve, S.E. (2004). *Stochastic calculus for finance II: Continuous-time models*. Springer Science & Business Media.

[8] Vasicek, O. (1977). *An equilibrium characterization of the term structure*. Journal of financial economics, 5(2), pp. 177–188.

Local Dependence Graphs for Discrete Time Processes

Tuesday, 4.07
16:00–18:00
Poster

Wojciech Niemiro[1,2], **Łukasz Rajkowski**[1]

[1] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw
[2] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University in Toruń

Abstract

Local dependence graphs for discrete time processes encapsulate information concerning the dependence relationships between the past of the multidimensional process and its present state and as such can represent feedback loops. Even in the discrete time setting some natural questions relating the conditional (in)dependence statements in the stochastic process to separation properties of the underlying local dependence graph are scattered throughout the literature. We provide an unifying view and fill in certain gaps. In this paper we examine graphical characteristics for two kinds of conditional independences: those occurring in Markov chains under the stationary regime and independences between the past of one subprocess and the future of another given the past of the third subprocess.

Bibliography

[1] Michael Eichler and Vanessa Didelez. Causal reasoning in graphical time series models. In *23rd Annual Conference on Uncertainty in Artifical Intelligence*, 2007.

Deep Horseshoe Gaussian processes

Tuesday, 4.07
16:00–18:00
Poster

Ismael Castillo[1], **Thibault Randrianarisoa**[2]

[1] Sorbonne University (France)
[2] Bocconi University (Italy)

Abstract

This work is concerned with the study of theoretical properties of deep Gaussian processes, which have recently been proposed as natural objects to fit, similarly to deep neural networks, possibly complex features present in modern data samples, such as compositional structures. Adopting a Bayesian nonparametric approach, it is natural to use deep Gaussian processes as prior distributions, and to use the corresponding posterior distributions for statistical inference. We introduce the deep Horseshoe Gaussian process, Deep-HGP, a new prior based on Gaussian processes with squared-exponential kernel, that in particular enables data-driven choices of the key lengthscale parameters. For nonparametric regression with random design, we show that the associated tempered posterior distributions recovers the unknown true regression curve optimally in terms of quadratic loss, up to a logarithmic factor. At the same time, Deep-HGP are conceptually quite simple to construct. One main idea is that the horseshoe prior enables simultaneous adaptation to both smoothness and structure. In addition, these results also stand in the case of a high-dimensional sample space, with its dimension growing polynomially in the sample size.

Bibliography

[1] Gianluca Finocchio and Johannes Schmidt-Hieber, *Posterior contraction for deep Gaussian process priors*, Journal of Machine Learning Research, 24(66):1–49, 2023.

[2] Andreas Damianou and Neil D. Lawrence, *Deep Gaussian processes*, Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, PMLR 31:207–215, 2013.

[3] Anirban Bhattacharya, Debdeep Pati, and David Dunson, *Anisotropic function estimation using multi-bandwidth gaussian processes*, The Annals of Statistics, 42(1):352-381, 2014.

[4] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott, *The horseshoe estimator for sparse signals*, Biometrika, 97(2):465–480, 2010.

Confidence bands for the covariance kernel of Banach space valued functional data

Tuesday, 4.07
14:40–15:00
CS 10
Room 1.40

C. Reihl

University of Bayreuth (Germany)

Abstract

The use and study of functional data is a common practice in modern statistics by now. Many classical problems have been extended from real valued to functional data, such as inference for the mean and the covariance structure. Naturally, most of the proofs cannot directly be transferred from real valued spaces to function spaces. The presentation deals with one of the most common questions: Estimating the trend and the covariance function, and constructing confidence bands for banach space valued functional data.

Throughout the presentation, each function is observed at discrete points, including measurement errors. For the estimation of the trend μ and covariance kernel R, two estimators, mostly based on an idea of [1], are discussed, which use the concept of smoothing via weight functions. It will be shown, that, for large sample size and observation points, the estimators smooth out the unintended departures and are also asymptotically gaussian, regarding the supremum norm, i.e.:

Theorem 1.

$$\sqrt{n}\big(\hat{\mu}(x) - \mu(x)\big) \xrightarrow{w} \mathcal{G}\big(0, R\big) \text{ in } \big(\mathcal{C}([0,1]), ||\cdot||_\infty\big)$$
$$\sqrt{n}\big(\hat{R}(s,t) - R(s,t)\big) \xrightarrow{w} \mathcal{G}\big(0, \mathcal{R}\big) \text{ in } \big(\mathcal{C}([0,1]^2), ||\cdot||_\infty\big)$$

where $\mathcal{G}$ are gaussian processes of dimension one or two.

For that result, a special version of the functional central limit theorem from [2], based on metric entropies, is used. The asymptotic normality is then used for constructing uniform confidence bands by:

Theorem 2.

$$\left[\hat{\mu}(x) - \frac{\hat{c}_{\gamma,\mu}}{\sqrt{n}},\ \hat{\mu}(x) + \frac{\hat{c}_{\gamma,\mu}}{\sqrt{n}}\right], \text{ with } \mathbb{P}\left(||\mathcal{G}(0,\hat{R})||_\infty > \hat{c}_{\gamma,\mu} \mid Y\right) = \gamma$$
$$\left[\hat{R}(s,t) - \frac{\hat{c}_{\gamma,R}}{\sqrt{n}},\ \hat{R}(s,t) + \frac{\hat{c}_{\gamma,R}}{\sqrt{n}}\right], \text{ with } \mathbb{P}\left(||\mathcal{G}(0,\hat{\mathcal{R}})||_\infty > \hat{c}_{\gamma,R} \mid Y\right) = \gamma$$

are simultaneous $(1-\gamma)\%$ confidence bands for μ or R, where Y is the random sample.

Furthermore, a naive bootstrap method is introduced, and its consistency for the confidence bands is proven. Finally, the validity of the confidence bands for the covariance, both constructed asymptotically and via bootstrap, is checked by a simulation study.

Bibliography

[1] Degras D., *Simultaneous Confidence Bands For Nonparametric Regression With Functional Data*, Statistica Sinica 21 (2011), 1735–1765.

[2] Pollard D., *Empirical Processes: Theory and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, 1990, Vol. 2.

Improving Group Lasso for high-dimensional categorical data

Monday, 3.07
16:05–16:25
CS 02
Room 1.40

Szymon Nowakowski[1], **Piotr Pokarowski**[1], **Wojciech Rejchel**[2], **Agnieszka Sołtys**[1]

[1] University of Warsaw (Poland)
[2] Nicolaus Copernicus University in Toruń (Poland)

Abstract

Sparse modeling or model selection with categorical data is challenging even for a moderate number of variables, because roughly one parameter is needed to encode one category or level. The Group Lasso [1] is a well known efficient algorithm for selection of continuous or categorical variables, but all estimates related to a selected factor usually differ. Therefore, a fitted model may not be sparse, which makes the model interpretation difficult. To obtain a sparse solution of the Group Lasso, we propose the two-step procedure in [2]. In the first step we reduce dimensionality of a problem using the Group Lasso. To choose the final model, we use an information criterion on a small family of models prepared by clustering levels of individual factors.

We investigate theoretical and practical properties of the algorithm in a sparse high-dimensional scenario, where a sample size might be significantly smaller than a number of all variables, but it is much larger than a number of active variables.

Bibliography

[1] Yuan, M., Lin, Y., *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society, Series B 68 (2006), 49–67.

[2] Nowakowski, S., Pokarowski, P., Rejchel, W., Sołtys, A., *Improving Group Lasso for high-dimensional categorical data*, arXiv (2022).

From Classification Accuracy to Proper Scoring Rules: Elicitability of Probabilistic Top List Predictions

Tuesday, 4.07
12:20–12:40
CS 05
Room 1.01

Johannes Resin[1,2]

[1] Heidelberg University (Germany)
[2] Heidelberg Institute for Theoretical Studies (Germany)

Abstract

In the face of uncertainty, the need for probabilistic assessments has long been recognized in the literature on forecasting. In classification, however, comparative evaluation of classifiers often focuses on predictions specifying a single class through the use of simple accuracy measures, which disregard any probabilistic uncertainty quantification. I propose probabilistic top lists as a novel type of prediction in classification, which bridges the gap between single-class predictions and predictive distributions. The probabilistic top list functional is elicitable through the use of strictly consistent evaluation metrics. The proposed evaluation metrics are based on symmetric proper scoring rules and admit comparison of various types of predictions ranging from single-class point predictions to fully specified predictive distributions. The Brier score yields a metric that is particularly well suited for this kind of comparison. A preprint is available [1].

Bibliography

[1] Resin, J., *From Classification Accuracy to Proper Scoring Rules: Elicitability of Probabilistic Top List Predictions*, Preprint, `https://arxiv.org/abs/2301.11797`.

Quantile estimation for maxima of stationary sequence by the block quantile method

Thursday, 6.07
14:20–14:40
CS 23
Room 1.01

Adam Jakubowski[1], **Igor Rodionov**[2]

[1] Nicolaus Copernicus University, Toruń (Poland)
[2] École Polytechnique Fédérale de Lausanne (Switzerland)

Abstract

Efficient estimation of high quantiles of partial maxima of a time series is one of the main components in statistical analysis of extremal phenomena occurring in climatology, demography, economics and other areas of the contemporary science. One of the most popular methods in analysis of extremes of stationary time series is the so-called extremal index. Non-zero value of the extremal index allows reducing the problem to the independent case with marginals being similar in nature to the originals.

In the present research we apply another, more general notion – the so-called phantom distribution function. It is known that the existence of a phantom distribution function is quite common and that a time series can admit a continuous phantom distribution function, while its extremal index is zero (i.e. is much less informative).

We construct several estimators for quantiles of maxima of a stationary sequence under the assumption that the phantom distribution function exists. For this purpose, we develop a novel block quantiles method that combines the ideas of the classical Gumbel block maxima method, adaptive selection of the proper tail model and some regression techniques. The quality of the estimators has been examined by simulations of several classes of stationary time series.

The conclusion is that in cases when the estimators can be compared with those obtained by the classical methods, our method performs in a similar way to the classical ones for time series with regularly varying marginals and outperforms them if tails of marginal distributions are lighter. Another advantage is that the obtained estimators can be applied also to processes for which the standard tools do not work (like the Markov chain corresponding to the Metropolis-Hastings algorithm) and they still exhibit stability and quite good performance.

Continuous capture process based on advection-diffusion-reaction models within Joint species movement modelling

Tuesday, 4.07
16:00–18:00
Poster

Luisa Fernanda Rodriguez[1], **Otso Ovaskainen**[2]

[1] University of Helsinki (Finland)
[2] Jyväskylä University (Finland)

Abstract

The Joint species movement modelling framework (JSMM)[1] enables inferring species and community level movement parameters from multi-species data. Within this approach, movement parameters are modelled as function of species traits, phylogenetic relationships, and spatio-temporal covariates. I introduce a continuous capture process (CCP) that extends the JSMM framework for diffusion-advection-reaction models parametrized with low resolution capture-mark-recapture data. The JSMM originally assumed a capture process mimicking a researcher visiting a particular site and attempting to capture marked individuals during an instantaneous time. The novel CCP method is aimed for ecologists that conduct experiments in which traps while open capture individuals in a continuous fashion over a given time period. As application example, I analyzed data corresponding to six species of dung beetles over three study sites to examine how the species traversed riparian buffers connected to a continuous forest within an oil palm plantation in Sabah, Malaysian Borneo. The CCP is one of the new features of the JSMM framework implementation for diffusion-advection models. The user-friendly software aims to be a useful tool for ecologists interested in studying causes and consequences of variation in movement behavior.

Bibliography

[1] Ovaskainen, O., Ramos, D. L., Slade, E. M., Merckx, T., Tikhonov, G., Pennanen, J., Pizo, M. A., Ribeiro, M. C., and Morales, J. M., *Joint species movement modeling: how do traits influence movements?*, Ecology 100(4): e02622. 10.1002/ecy.2622.

Improving Model Choice in Classification: An Approach Based on Clustering of Covariance Matrices.

Tuesday, 4.07
12:00–12:20
CS 05
Room 1.01

David Rodríguez-Vítores, Carlos Matrán

Department of Statistics and Operational Research and IMUVA,
University of Valladolid, Spain.

Abstract

This work introduces a refinement of the Parsimonious Model for fitting a Gaussian Mixture. The improvement is based on the consideration of groupings of the covariance matrices according to a criterion, such as sharing Principal Directions. This and other similarity criteria that arise from the spectral decomposition of a matrix are the bases of the Parsimonious Model. The classification can be achieved with simple modifications of the CEM (Classification Expectation Maximization) algorithm, using in the M step suitable estimation methods known for parsimonious models. This approach leads to propose Gaussian Mixture Models for model-based clustering and discriminant analysis, in which covariance matrices are clustered according to a parsimonious criterion, creating intermediate steps between the fourteen widely known parsimonious models. The added versatility not only allows us to obtain models with fewer parameters for fitting the data, but also provides greater interpretability. We show its usefulness for model-based clustering and discriminant analysis, providing algorithms to find approximate solutions verifying suitable size, shape and orientation constraints, and applying them to both simulation and real data examples.

Bibliography

[1] Celeux G., Govaert G., *Gaussian Parsimonious Clustering Models*, Pattern Recognition, 28:5 (1995), 781–793.

[2] Celeux G., Govaert G., *A Classification EM Algorithm for Clustering and Two Stochastic Versions*, Computational Statistics and Data Analysis, 14:3 (1992), 315–332.

[3] Flury B., *Common Principal Components in k Groups*, Journal of the American Statistical Association, 79:388 (1984), 892–898.

[4] Fraley C., Raftery A. E. , *Model-Based Clustering, Discriminant Analysis, and Density Estimation*, Journal of the American Statistical Association, 97:458 (2002), 611–631.

[5] García-Escudero L., Mayo A., Riani M. , *Model-Based Clustering With Determinant-and-Shape Constraint*, Statistics and Computing, 30 (2020), 1363–1380.

Bootstrapping high-dimensional sample covariance matrices

Monday, 3.07
14:15–15:15
Plenary Lecture
Room 0.03

Angelika Rohde

University of Freiburg

Abstract

Bootstrapping is the classical approach for distributional approximation of estimators and test statistics when an asymptotic distribution contains unknown quantities or provides a poor approximation quality. For the analysis of massive data, however, the bootstrap is computationally intractable in its basic sampling-with-replacement version. Moreover, it is even not valid in some important high-dimensional applications. Combining subsampling of observations with suitable selection of their coordinates, we introduce a new "$(m, mp/n)$ out of (n, p)"-sampling with replacement bootstrap for eigenvalue statistics of high-dimensional sample covariance matrices based on n independent p-dimensional random vectors. In the high-dimensional scenario $p/n \to c \in [0, \infty)$, this fully nonparametric bootstrap is shown to consistently reproduce the underlying spectral measure if $m/n \to 0$. If $m^2/n \to 0$, it approximates correctly the distribution of linear spectral statistics. The crucial component is a suitably defined representative subpopulation condition which is shown to be verified in a large variety of situations. The proofs incorporate several delicate technical results which may be of independent interest.

Modeling ultra-high frequency trade durations with ACD type of models

Thursday, 6.07
16:00–16:20
CS 29
Room 4420

Markus Belfrage, Gunnar Rosenqvist

Hanken School of Economics, Helsinki, Finland

Abstract

Beginning with Engle and Russell [3], a large number of models of the autoregressive conditional duration (ACD) type have been developed for modeling of financial time series of durations, in particular, trade durations. For reviews see e.g. Pacurar [6], Hautsch [4] and Bhogal and Variyam [2]. Since most of these models were introduced, stock markets have undergone a rapid transformation (see e.g. [5]), and trade duration data are now vastly different. In this paper we explore the possibilities to model contemporary ultra-high frequency trade durations with ACD type models by analyzing a set of such data with nanosecond precision. We fit a range of ACD type models, evaluate them and explore their ability to cope with various characteristics found in data. Our methods are largely implemented in the R package ACDm.

Bibliography

[1] Belfrage, M., *Tools for Autoregressive Conditional Duration Models*, R package version 1.0.4.2 (2022).

[2] Bhogal S.K., Variyam R.T., *Conditional Duration Models for High-Frequency Data: A Review of Recent Developments*, Journal of Economic Surveys 33 (2018), 252–273.

[3] Engle R.F., Russell J.R., *Autoregressive Conditional Duration: A New Approach for Irregularly Spaced Transaction Data*, Econometrica, 66 (1998), 1127–1162.

[4] Hautsch N., *Econometrics of Financial High-Frequency Data*, Springer Science & Business Media, 2012.

[5] OHara M., *High frequency market microstructure*, Journal of Financial Economics, 116 (2015), 257–270.

[6] Pacurar M., *Autoregressive Conditional Duration Models in Finance: a Survey of the Theoretical and Empirical Literature*, Journal of Economic Surveys, 22 (2008), 711–751.

Logistic regression for uplift modeling

Thursday, 6.07
16:20–16:40
CS 27
Room 1.40

Madalena Szypulska[2], **Krzysztof Rudaś**[1,2]

[1] Institute of Computer Science, Polish Academy of Sciences (Poland)
[2] Warsaw University of Technology (Poland)

Abstract

To evaluate efficiency of an action (e.g. medical treatment or a marketing campaign) we divide our population randomly into two groups: treatment (on which action is taken) and control (on which action is not taken). The predicted true effect of the action on a given individual is modeled as the difference between responses in both groups. In our work we concentrate on classification uplift problem, where we try to predict scoring describing if given observation should be treated or not. In our work we present new regularization method for uplift logistic models, which gives us natural interpolation between two basic approaches (double and uplift method) and also regularize vector of coefficients used to obtain appropriate scoring.

Bibliography

[1] Imbens G.I., Rubin D.B., *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press (2015).
[2] Rudaś K., Jaroszewicz S., *Linear regression for uplift modeling*, Data Mining and Knowledge Discovery, 32 (2018) 1275–1305.
[3] Rzepakowski P., Jaroszewicz S., *Decision trees for uplift modeling with single and multiple treatments*, Knowledge and Information Systems, 32 (2012) 303–327.

Necessary and sufficient conditions for existing moments of generalized order statistics

Tuesday, 4.07
14:20–14:40
CS 13
Room 4420

Mariusz Bieniek[1], **Tomasz Rychlik**[2]

[1] University of Maria Curie Skłodowska (Poland)
[2] Institute of Mathematics, Polish Academy of Sciences (Poland)

Abstract

Vectors and sequences of generalized order statistics provide a general semi-parametric model of ordered random variables, including, among others, standard order statistics, record values, kth record values, various censoring schemes, and reliability models. The distribution of rth generalized order statistic $X_{*,r}$ depends on a vector $(\gamma_1, \ldots, \gamma_r)$ of positive parameters and a one-dimensional distribution function F.

For arbitrarily fixed parameter vector $(\gamma_1, \ldots, \gamma_r)$, we determine the necessary and sufficient conditions on the moment order $p > 0$ so that $\mathbb{E}X_{*,r}^p$ is finite for all baseline distribution functions F supported on the nonnegative half-axis with a positive and finite expectation. The conditions depend on a relation between the moment order and the minimal parameter of the generalized order statistic and its multiplicity. The number r of the generalized order statistic does not matter here except for the fact that for multiple minimal parameter we need $r \geq 2$.

Moreover, for every given $(\gamma_1, \ldots, \gamma_r)$ under the necessary and sufficient conditions for pth moment finiteness, we determine the sharp upper bounds on $\mathbb{E}X_{*,r}^p$ evaluated in the scale units being the pth powers of the mean of the model distribution function F.

High-dimensional sparse vine copula regression with application to genomic prediction

Wednesday, 5.07
12:00–12:20
CS 15
Room 1.40

Özge Sahin[1], **Claudia Czado**[1,2]

[1] Technical University of Munich (Germany)
[2] Munich Data Science Institute (Germany)

Abstract

High-dimensional data sets are often available in genome-enabled predictions. Such data sets include nonlinear relationships with complex dependence structures. For such situations, vine copula based (quantile) regression is an important tool. However, the current vine copula based regression approaches do not scale up to high and ultra-high dimensions. To perform high-dimensional sparse vine copula based regression, we propose two methods. First, we show their superiority regarding computational complexity over the existing methods. Second, we define relevant, irrelevant, and redundant explanatory variables for quantile regression. Then we show our method's power in selecting relevant variables and prediction accuracy in high-dimensional sparse data sets via simulation studies. Next, we apply the proposed methods to the high-dimensional real data, aiming at the genomic prediction of maize traits. Some data-processing and feature extraction steps for the real data are further discussed. Finally, we show the advantage of our methods over linear models and quantile regression forests in simulation studies and real data applications.

Celebrating 100 Years of Wiener Space

Friday, 7.07
11:30–12:00
IS 16
Room 0.03

René Schilling

Technische Universität Dresden (Germany)

Abstract

Norbert Wiener published in 1923 his groundbreaking and hugely influential paper on "Differential Space" which was about the first rigorous mathematical construction of Brownian motion aka the Wiener process. In this talk we discuss some historical aspects surrounding Wiener's paper

Finite Mixture Models for an underlying Beta distribution with an application to COVID-19 data

Tuesday, 4.07
11:00–11:20
CS 05
Room 1.01

Cédric Noel[1], **Jang Schiltz**[2]

[1] University of Lorraine (France)
[2] University of Luxembourg (Luxembourg)

Abstract

Finite Mixture Models in the sense of Nagin ([2]) are fuzzy logic cluster analysis models for time series. Starting from a sample of trajectories, the aim is to detect a number of subgroups of the sample, so that subjects in the same group exhibit quite similar data trajectories, whereas two subjects from two different groups have trajectories that differ in some sense. These models have been generalized by Schiltz ([4]) and are part of a larger strand of models that analyze latent evolutions in longitudinal data.
We introduce an extension of Nagin's finite mixture model to underlying Beta distributions and present our **R** package ([3]) **trajeR** which allows to calibrate the model. Then, we test the model and illustrate some of the possibilities of **trajeR** by means of an example with simulated data.
In a second part of the paper, we use this model to analyze COVID-19 related data ([1]) during the first part of the pandemic. We identify a classification of the world into five groups of countries with respect to the evolution of the contamination rate and show that the median population age is the main predictor of group membership. We do however not see any sign of efficiency of the sanitary measures taken by the different countries against the propagation of the virus.

Bibliography

[1] Hasell J. et al., *A cross-country database of COVID-19 testing*, Scientific Data, 7 (2020), pp. 345.
[2] Nagin D.S., *Group-Based Modeling of Development*, Harvard University Press, 2005.
[3] Noel C., Schiltz J., *trajeR, an R package for cluster analysis of time series*, Working Paper, University of Luxembourg, Luxembourg, 2022.
[4] Schiltz J., *A generalization of Nagin's finite mixture model*, In: Stemmler M., Von Eye A. & Wiedermann W. (eds.), Dependent Data in Social Sciences Research, Springer. 2015, pp. 107–126.

Pattern Recovery in Penalized Estimation and its Geometry

Thursday, 6.07
14:00–14:20
CS 26
Room 5440

Piotr Graczyk[1], **Ulrike Schneider**[2], **Tomasz Skalski**[3], **Patrick Tardivel**[4]

[1] University of Angers (France)
[2] TU Vienna (Austria)
[3] Wrocław University of Science and Technology (Poland)
[4] University of Burgundy (France)

Abstract

For many penalized estimators such as LASSO, SLOPE, OSCAR, PACS, fused, clustered and generalized LASSO, the penalty term is a real-valued polyhedral gauge. We focus on pattern recovery at β with respect to such a penalty term, where β is the unknown parameter of regression coefficients. For LASSO, the pattern of β only depends on the sign of β and sign recovery by LASSO is a well known topic in the literature. We introduce the notion of patterns and pattern recovery in the broad framework of gauge-penalized least-squares estimation and illustrate the patterns different polyhedral gauges. We also provide theoretical guarantees for pattern recovery, in particular the "noiseless recovery condition" is necessary for a probability of recovery larger than 1/2 and can be viewed as a generalization of the LASSO's irrepresentability condition. This condition may be relaxed using thresholded penalized least squares estimators, a class of estimators generalizing the thresholded LASSO. Indeed, we show that the "accessibility condition", a weaker condition than the "noiseless recovery condition", is necessary and asymptotically sufficient for pattern recovery in thresholded penalized estimation. We also provide a geometric interpretation of our approach to pattern recovery and the accessibility condition.

Bibliography

[1] P. Graczyk, U. Schneider, T. Skalski and P. Tardivel, *Pattern Recovery by Penalized and Thresholded Estimation and its Geometry*, preprint on hal:03262087.

Goodness-of-fit tests for non-Gaussian linear causal models

Thursday, 6.07
15:20–15:40
CS 24
Room 1.40

Daniela Schkoda, Mathias Drton

Technical University of Munich (Germany)

Abstract

Inferring causal relationships between variables solely from observational data is a central question in many scientific fields. Various algorithms have been developed to tackle this problem by leveraging different types of a priori assumptions. One prominent example is the assumption that the joint distribution of the observed variables follows a linear non-Gaussian structural equation model. In this talk, we will show that this linearity assumption may be characterized in terms of a rank constraint on a matrix formed from second and higher moments. Testing this rank constraint thus offers a new approach to testing the hypothesis that the data-generating distribution belongs to a linear non-Gaussian structural equation model. For a practical implementation of the test, we consider a multiplier bootstrap method that uses incomplete U-statistics to estimate subdeterminants as well as tests that employ asymptotic approximations of the null distribution of singular values. The methods are illustrated, in particular, for the Tuebingen collection of benchmark data sets on cause-effect pairs.

A minimax framework for quantifying risk-fairness trade-off in regression

Tuesday, 4.07
14:30–15:00
IS 08
Room 0.06

Evgenii Chzhen[1], **Nicolas Schreuder**[2]

[1] Université Paris-Saclay, CNRS, Laboratoire de mathématiques dâ€™Orsay (France)
[2] MaLGa, DIBRIS, UniversitÃ di Genova (Italy)

Abstract

In various domains, statistical algorithms trained on personal data take pivotal decisions which influence our lives on a daily basis. Recent studies show that a naive use of these algorithms in sensitive domains may lead to unfair and discriminating decisions, often inheriting or even amplifying biases present in data [2]. In this talk, I will begin by introducing the question of fairness from a statistical learning perspective. I will then present a minimax statistical framework to characterize the performance – both in terms of risk and fairness – of general estimators. I will give examples of upper and lower bounds that can be derived in this framework. In particular, I will describe a minimax optimal estimator under the linear Gaussian model with group-dependent bias.

My talk will be based on a joint work with Evgenii Chzhen [1].

Bibliography

[1] Chzhen E., Schreuder N., *A minimax framework for quantifying risk-fairness trade-off in regression*, Ann. Statist. 50(4): 2416-2442 (Aug. 2022). DOI: 10.1214/22-AOS2198

[2] Barocas S, Hardt M., Arvind N., *Fairness and Machine Learning: Limitations and Opportunities*, `http://www.fairmlbook.org`, 2019.

Classification of materials using Topological Data Analysis

Friday, 7.07
11:20–11:40
CS 33
Room 5440

Pawel Dlotko[1], **Jan F. Senge**[1,2]

[1] Dioscuri Centre for TDA, Warsaw (Poland)
[2] University of Bremen (Germany)

Abstract

Topological Data Analysis provides new approaches for interpretation of (high dimensional) data exploiting underlying structures and relationships to be used as proxies for understanding qualitative similarities and differences. Suitable vectorization of proxies like the persistence diagram can enhance Machine Learning (ML) algorithms with an intermediate layer improving performance as well as giving (other) interpretations of the data. In this talk we will highlight some of the design choices and tools of TDA when introducing pipelines for persistent homology to understand the 2-dimensional structure of processed surfaces [1] as well as 3-dimensional structures of pores [2] and how these compare and relate to standard statistics and ML.

Bibliography

[1] Senge, J.F., Astaraee, A.H., Dłotko, P., Bagherifard, S., Bosbach, W.A., *Extending conventional surface roughness ISO parameters using topological data analysis for shot peened surfaces*, Scientific Reports, 12, 5538 (2022).

[2] Buccino, F., Aiazzi, I., Casto, A., Liu, B., Sbarra, M.C., Ziarelli, G., Banfi, G., Vergani, L.M., *The synergy of synchrotron imaging and convolutional neural networks towards the detection of human micro-scale bone architecture and damage*, Journal of the Mechanical Behavior of Biomedical Materials, 137 (2023)

Nonparametric Estimation of Non-Crossing Quantile Regression Process with Deep ReQU Neural Networks

Tuesday, 4.07
11:40–12:00
CS 08
Room 5440

Guohao Shen[1] , **Yuling Jiao**[2], **Yuanyuan Lin**[3], **Joel L. Horowitz**[4], **Jian Huang**[1]

[1] The Hong Kong Polytechnic University
[2] Wuhan University
[3] The Chinese University of Hong Kong
[4] Northwestern University

Abstract

We propose a penalized nonparametric approach to estimating the quantile regression process (QRP) in a nonseparable model using rectifier quadratic unit (ReQU) activated deep neural networks and introduce a novel penalty function to enforce non-crossing of quantile regression curves. We establish the non-asymptotic excess risk bounds for the estimated QRP and derive the mean integrated squared error for the estimated QRP under mild smoothness and regularity conditions. To establish these non-asymptotic risk and estimation error bounds, we also develop a new error bound for approximating C^s smooth functions with $s > 0$ and their derivatives using ReQU activated neural networks. This is a new approximation result for ReQU networks and is of independent interest and may be useful in other problems. Our numerical experiments demonstrate that the proposed method is competitive with or outperforms two existing methods, including methods using reproducing kernels and random forests for nonparametric quantile regression.

Estimation of quantile versions of the Lorenz curve and the Gini index for the generalized Pareto distribution

Tuesday, 4.07
16:00–18:00
Poster

Alicja Jokiel-Rokita[1], **Agnieszka Siedlaczek**[2]

[1] Wroclaw University of Science and Technology (Poland)
[2] University of Opole (Poland)

Abstract

The Lorenz curve is a commonly used tool in the study of income inequality. For this purpose, it was initially defined in 1905 by Lorenz [1], whereas today it is widely used in many fields of study. However, it is only defined for distributions with a finite expected value. Therefore, Prendegast and Staudte in 2016 [2] proposed quantile versions of the Lorenz curve that do not require this assumption.
We discuss estimation of the quantile versions of the Lorenz curve in case of the generalized Pareto distribution. Variety of methods of estimation parameters of generalized Pareto distribution is reviewed. The described estimators are applied in estimation of quantile versions of the Lorenz curve. Asymptotic normality of one of the plug-in estimators of the quantile versions of the Lorenz curve in examined model is proved. Accuracy of the proposed estimators is studied in simulations. Application of the described methods in the real data analysis is showed.

Bibliography

[1] Lorenz M. O., *Methods of Measuring the Concentration of Wealth*, In: Publications of the American Statistical Association 9(70), 1905, pp.209-219.
[2] Prendergast L. A., Staudte R. G., *Quantile versions of the Lorenz curve*, In: Electronic Journal of Statistics 10(2), 2016, pp.1896-1926.

Handling many longitudinal predictors in survival analysis: a penalized regression calibration approach

Wednesday, 5.07
12:00–12:20
CS 18
Room 4420

Mirko Signorelli

Leiden University (The Netherlands)

Abstract

Longitudinal and high-dimensional measurements have become increasingly common in biomedical research. Repeated measurements data can carry important information about ageing and disease progression, which can be used to improve the accuracy of predictions of survival outcomes. Over the last decades, several methods have been developed to predict survival outcomes either from a handful of longitudinal covariates, or from a high-dimensional set of cross-sectional covariates. However, until recently methods to predict survival outcomes using covariates that are both longitudinal and high-dimensional were missing.

In this presentation I will introduce Penalized Regression Calibration (PRC, [1]), a statistical method that can be employed to predict survival using a large (potentially high-dimensional) number of longitudinally-measured covariates as predictors. PRC comprises three modelling steps: first, the trajectories described by the longitudinal predictors are flexibly modelled through the specification of univariate or multivariate mixed-effects models. Second, subject-specific summaries of the longitudinal trajectories are derived from the fitted mixed-effects models. Third, the relationship between the survival outcome and the longitudinal covariates is modelled using the subject-specific summaries as covariates in a penalized Cox model. To ensure a proper internal validation of the accuracy of predictions of survival based on PRC, we furthermore developed a cluster bootstrap optimism correction procedure (CBOCP) that allows to correct for the optimistic bias of apparent measures of predictiveness. Both PRC and the CBOCP are implemented in the `R` package `pencal` [2], available from CRAN.

After summarizing the results of simulation studies designed to assess the behaviour of PRC across different scenarios, I will conclude by illustrating an application of PRC to the problem of predicting time to dementia in a population of elderly subjects.

Bibliography

[1] M. Signorelli, P. Spitali, C. Al-Khalili Sgyziarto, The Mark-MD Consortium, and R. Tsonaka. Penalized regression calibration: a method for the prediction of survival outcomes using complex longitudinal and high-dimensional data. *Statistics in Medicine*, 40(27):6178–6196, 2021.

[2] M. Signorelli, P. Spitali, and R. Tsonaka. pencal: Penalized Regression Calibration (PRC). R package version 1.3.2. URL: https://cran.r-project.org/package=pencal.

Optimal designs for testing pairwise differences: a graph based game theoretic approach

Friday, 7.07
14:40–15:00
CS 34
Room 1.01

Arpan Singh[1], **Satya Prakash Singh**[2], **Ori Davidov**[3]

[1] Indian Institute of Technology, Hyderabad, (India)
[2] Indian Institute of Technology, Kanpur (India)
[3] University of Haifa, (Israel)

Abstract

In a variety of experimental setting there is an interest in comparing a subset of pairs–of–treatments. Such experiments usually address one of the following two scientific questions: (1) is there a difference within *any* of the selected pairs of treatments? or, (2) is there a difference within *all* of the selected pairs of treatments? In this article we propose max–min optimal designs for testing the above hypotheses using a graph based game theoretic approach. Some of the max–min designs obtained are well known, but not recognized as optimal, others are novel and provide a substantial improvement over naive designs.

Pattern recovery by SLOPE

Thursday, 6.07
14:20–14:40
CS 26
Room 5440

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

[1] University of Wrocław
[2] University of Burgundy, France
[3] University of Angers, France
[4] Warsaw University of Technology
[5] Wrocław University of Science and Technology

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.

This talk is based on the joint research with Małgorzata Bogdan, Xavier Dupuis, Piotr Graczyk, Bartosz Kołodziejek, Patrick Tardivel and Maciej Wilczyński. The research was supported by a French Government Scholarship.

Bibliography

[1] Bogdan M., Dupuis X., Graczyk P., Kołodziejek B., Skalski T., Tardivel P., Wilczyński M., *Pattern Recovery by SLOPE*, ArXiv 2203.12086 (2023).

[2] Skalski T., Graczyk P., Kołodziejek B., Wilczyński M., *Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal*, Probability and Mathematical Statistics, 42.2 (2022), 283-302.

[3] Tardivel P., Skalski T., Graczyk P., Schneider U., *Pattern Recovery in Penalized and Thresholded Estimation and its Geometry*, hal-03262087v2 (2023).

Forecast combination through dynamic linear models

Friday, 7.07
12:20–12:40
CS 32
Room 2180

Maddie Smith[1], **Nicos Pavlidis**[2], **Adam Sykulski**[3]

[1] Lancaster University (United Kingdom)
[2] Lancaster University (United Kingdom)
[3] Imperial College London (United Kingdom)

Abstract

Dynamic Linear Models (DLMs) provide a modelling framework within the state space formulation of time series, wherein both the observation equation and the state equation are linear. This linear formulation enables a closed-form Bayesian analysis, allowing relevant distributions of interest to be computed sequentially as new time series data become available [1].

We propose an application of DLMs which facilitates the combination of multiple point forecasts for the current time series of interest (similar to [2]). By working within a regression framework, we enable forecasts to be combined using dynamic weights. Allowing the forecast weights to evolve in this way is desirable, since they can be learned as new data become available. Moreover, dynamic weights are able to adapt to changes in the quality of the constituent forecasters, allowing for a changing 'best' forecast throughout time. The rate at which the weights change is controlled by the introduction of a discount factor.

This method accommodates for biased forecasters, and also correlations between individual forecasts. Furthermore, we show that this method deals effectively with missing data in the time series of interest, and can outperform a selection of benchmark forecast combination methods when applied to simulated and real data. In this talk, we will discuss generalisations of the problem to wider settings, including adaptive discounting, missing forecasters and modelling multivariate time series of interest (see [3], [4]). To the best of our knowledge, such extensions have not been considered in this context before.

Bibliography

[1] West M., Harrison J., *Bayesian Forecasting and Dynamic Models (2nd Ed.)*, Springer, 1997.

[2] Terui N., van Dijk H. K, *Combined forecasts from linear and nonlinear time series models*, International Journal of Forecasting, 18 (2002), 421-438.

[3] Zhao Z. Y., Xie M., West M., *Dynamic dependence networks: Financial time series forecasting and portfolio decisions*, Applied Stochastic Models in Business and Industry, 32 (2016), 311-332.

[4] Yusupova A., Pavlidis N. G., Pavlidis E. G. (in press), *Dynamic linear models with adaptive discounting*, International Journal of Forecasting, (2022).

On multilevel Monte Carlo algorithm for SDEs driven by countably dimensional Wiener process and Poisson random measure

Tuesday, 4.07
12:00–12:20
CS 07
Room 2180

Michał Sobieraj

Faculty of Applied Mathematics, AGH UST in Cracow (Poland)

Abstract

In this presentation we refer recent results on a multilevel Monte Carlo algorithm for weak approximation of SDEs of the following form

$$\begin{cases} \mathrm{d}X(t) = a(t, X(t))\mathrm{d}t + b(t, X(t))\mathrm{d}W(t) + \int\limits_{\mathcal{E}} c(t, X(t-), y)N(\mathrm{d}y, \mathrm{d}t), \ t \in [0, T], \\ X(0) = \eta, \end{cases} \tag{1}$$

where $T > 0$, $\mathcal{E} = \mathbf{R}^{d'} \setminus \{0\}$, $d' \in \mathbf{N}$, $W = [W_1, W_2, \ldots]^T$ is a countably dimensional Wiener process, and $N(\mathrm{d}y, \mathrm{d}t)$ is a Poisson random measure with an intensity measure $\nu(\mathrm{d}y)\mathrm{d}t$ (see [1], [2]). We assume that $\nu(\mathrm{d}y)$ is a finite Lévy measure on $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$.

Hereafter, in a certain class of coefficients $a : [0, T] \times \mathbf{R}^d \to \mathbf{R}^d$, $b : [0, T] \times \mathbf{R}^d \to l^2(\mathbf{R}^d)$, $c : [0, T] \times \mathbf{R}^d \to \mathbf{R}^d$, we investigate the upper complexity bound for the multilevel Monte Carlo algorithm which utilizes truncated dimension randomized Euler scheme $X^{RE}_{M,n}$ for pointwise approximation of $X(T)$. The upper error bound for the scheme is characterized by the following theorem.

Theorem ([3]). *There exists a positive constant K, depending only on the parameters of the class $\mathcal{F}$ of admissible functions a, b, c and input conditions η, such that for every $(a, b, c, \eta) \in \mathcal{F}$ and $M, n \in \mathbf{N}$ it holds*

$$\|X(T) - X^{RE}_{M,n}\|_{L^2} \leq K\Big(n^{-1/2} + \delta(M)\Big).$$

The theorem has a direct impact on the upper complexity bound of the multilevel Monte Carlo method which turns out to highly depend on the function δ. Therefore, the upper complexity bound of the multilevel method is investigated on certain subclasses for the inverse of δ. At the end we present results of numerical experiments performed on GPUs, where we used a suitable implementation with CUDA.

Bibliography

[1] S. N. Cohen, R. J. Elliott. *Stochastic Calculus and Applications, 2nd. ed.*, Springer, 2015.

[2] I. Gyöngy, N. V. Krylov. *On stochastic equations with respect to semimartingales I.* Stochastics, Volume 4, 1980,1–21.

[3] P. Przybyłowicz, M. Sobieraj, Ł. Stępień. *Efficient approximation of SDEs driven by countably dimensional Wiener process and Poisson random measure.* SIAM Journal on Numerical Analysis 60 (2022), 824–855.

Spectral Calibration of Time-Inhomogeneous Exponential Lévy Models

Thursday, 6.07
16:40–17:00
CS 29
Room 4420

Loek Koorevaar[1], **Jakob Söhl**[2], **Stan Tendijck**[3]

[1] IMC Trading (The Netherlands)
[2] Delft University of Technology (The Netherlands)
[3] Shell (The Netherlands)

Abstract

Model calibration is essential to pricing and hedging of financial products. A key question when calibrating a model based on prices of European call and put options is how all information contained in the option prices can be merged into one model. On the one hand, the model needs to be large enough to be sufficiently flexible and to be able to integrate all the available information. On the other hand, the model needs to be identifiable from the options traded on the market. The frequently observed volatility smile or skew is an indication that the Black–Scholes model is not flexible enough to account for the prices of options with different strike prices. Exponential Lévy models are flexible enough to model the volatility smile or skew and can therefore incorporate the information of options with different strike prices. However, empirical evidence shows that calibrating exponential Lévy models by options with different maturities leads to conflicting information. In other words, the stationarity implicitly assumed in the exponential Lévy model is not satisfied. We propose an identifiable time-inhomogeneous Lévy model that does not assume stationarity and that can integrate option prices from different maturities and different strike prices without leading to conflicting information.

In the time-inhomogeneous Lévy model, we derive the convergence rates and show confidence intervals for the estimators of the volatility, the drift, the intensity, and the Lévy density. Previously confidence intervals have been constructed for time-homogeneous Lévy models in an idealized Gaussian white noise model by Söhl [2]. In the idealized Gaussian white noise model, it is assumed that the observations are Gaussian and given continuously across the strike prices. This simplifies the analysis significantly. Here we construct the confidence intervals in a discrete observation setting for time-inhomogeneous Lévy models and the only assumption on the errors is that they are sub-Gaussian, in particular, all bounded errors with arbitrary distributions are covered. Our additional results on the convergence rates extend the paper by Belomestny and Reiß [1] from time-homogeneous to time-inhomogeneous Lévy models.

Bibliography

[1] Belomestny D., Reiß M., *Spectral calibration of exponential Lévy models*, Finance and Stochastics, 10(4) (2006), 449-474.
[2] Söhl J., *Confidence sets in nonparametric calibration of exponential Lévy models*, Finance and Stochastics, 18(3) (2014), 617-649.

Joint estimation of heterogeneous non-Gaussian functional graphical models with fully and partially observed curves

Thursday, 6.07
12:00–12:20
CS 19
Room 1.01

Eftychia Solea

CREST and ENSAI (France)

Abstract

We introduce a new methodology for estimating undirected graphical models for heterogeneous non-Gaussian multivariate functional data, such as brain activities collected by functional magnetic resonance imaging from a sample of subjects with different subtypes of a neurological disease. The goal of the new model is to estimate robustly a collection of functional graphical models, corresponding to several subpopulations that share some common dependence structure. The model is fitted via a joint estimation method employed with the hierarchical penalty that encourages a common graph structure and individual sparsity. To relax the Gaussian assumption, we consider the functional Gaussian copula graphical model proposed by [1], and propose the rank-based Kendall's tau correlation operator that extends the Kendall's tau correlation coefficient to the functional setting. We establish the concentration inequalities of the estimates and the graph selection consistency for both completely and partially observed data, while allowing the number of functions to diverge to infinity with the sample size. We demonstrate the efficiency of our method through both simulations and an analysis of the fMRI ADHD-200 data set of subjects with inattentive and combined subtypes of ADHD, and control subjects

Bibliography

[1] Solea E., and Li B., (2022), *Copula Gaussian Graphical Models for Functional Data*, Journal of the American Statistical Association, 117:538, 781–793.

Approximate Bayesian algorithm for tensor robust PCA using relative entropy

Tuesday, 4.07
14:20–14:40
CS 09
Room 1.01

Andrej Srakar

[1] Institute for Economic Research (Slovenia)
[2] University of Ljubljana (Slovenia)

Abstract

Matrix and tensor completion methods are gaining interest. They allow to study un- and semistructured datasets of some contemporary endeavours such as citizen science initiatives ([4]). Recently proposed Tensor Robust Principal Component Analysis (TRPCA, [3]) aims to exactly recover the low-rank and sparse components from their sum. We extend an own Bayesian approximate inference algorithm for TRPCA ([5]), based on regression adjustment methods ([1]) and compare it to earlier studies using variational Bayes inference. As the estimation is set in a high-dimensional context this leads to known bottlenecks which we solve using a novel proposal to use functional Bregman divergence between posterior distributions as a measure of the posterior surrogate loss.

Definition (Functional Bregman Divergence; [2]). Let $\phi : L^p(\nu) \to \mathbb{R}$ be a strictly convex, twice-continuously Fréchet-differentiable functional. The Bregman divergence $d_\phi : \mathcal{A} \times \mathcal{A} \to [0, \infty)$ is defined for all $f, g \in \mathcal{A}$ as $d_\phi[f, g] = \phi[f] - \phi[g] - \delta\phi[g; f - g]$, where $\delta\phi[g; \bullet]$ is the Fréchet derivative of ϕ at g.

Namely, we use relative entropy as divergence measure combined with more general regression adjustment perspective. We provide proofs of the asymptotic properties of the approach as well as a simulation study. In a short application, we study two datasets of citizen science initiatives, Great Greenhouse Gas Grass Off initiative where individuals collected grass samples to infer the atmospheric fossil fuel CO2 mole fraction; and COVID-19 Sledilnik/Tracker with data on pandemic variables throughout the COVID-19 pandemic.

Bibliography

[1] Blum, M. G., *Approximate Bayesian computation: a nonparametric perspective*, Journal of the American Statistical Association, 105, 491, (2010).

[2] Frigyik, B.A., Srivastava, S., Gupta, M.R., *Functional Bregman Divergence and Bayesian Estimation of Distributions*, in IEEE Transactions on Information Theory, vol. 54, no. 11, pp. 5130–5139, (2008).

[3] Lu, C., Feng, J., Chen, Y., Liu, W., Lin, Z., Yan, S., *Tensor Robust PrincipalComponent Analysis: Exact Recovery of Corrupted Low-Rank Tensors via Convex Optimization*, Available as arXiv:1708.04181v3, (2019).

[4] Robin, G. Klopp, O., Josse, J., Moulines, É., *Main Effects and Interactions in Mixed and Incomplete Data Frames*, Journal of the American Statistical Association, 115:531, 1292–1303, (2020).

[5] Srakar, A., *Approximate Bayesian Algorithm for Tensor Robust Principal Component Analysis*, In: Argiento, R., Camerlenghi, F., Paganin, S. (eds) New Frontiers in Bayesian Statistics. BAYSM 2021. Springer Proceedings in Mathematics and Statistics, vol 405. Springer, Cham, (2022).

Statistical methods for the analysis of mass spectrometry data with multiple membership

Tuesday, 4.07
16:00–18:00
Poster

Mateusz Staniak

University of Wrocław (Poland)

Abstract

Mass spectrometry (MS) is a core technology for proteomics. It allows for the identification and quantification of proteins in biological samples. In a mass spectrometry experiment, peptides â€“ smaller fragments of proteins - are ionized, separated based on their mass and charge, and quantified. Resulting data are complex and high-dimensional: they include up to thousands of proteins and tens of thousands of peptides. This type of data is important in drug discovery and other stages of drug development, and in various fields of proteomics research. Typically, mass spectrometry data are used to estimate relative abundance of proteins in different biological conditions.

One of the major challenges in the analysis of mass spectrometry is the protein inference problem: deriving a list of proteins that are present in the sample based on identified peptides. It is complicated due to three factors: false identifications of peptides based on mass spectra, presence of peptide sequences that can be attributed to multiple proteins (shared peptides) and one-hit wonders - proteins identified only by a single peptide. Similarly, protein quantification - estimating the relative abundance of proteins in different conditions - is difficult in the presence of shared peptides, as it is not clear how to distribute peptide abundance among their respective proteins. From statistical perspective, inclusion of shared peptides in models that are used to estimate protein abundances from peptide-level data introduces the multiple membership structure, in which observations (peptide intensities) may belong to multiple groups defined by proteins. Typically, shared peptides are removed from analysis of MS data, which leads to loss of peptide-level information and lack of ability to estimate abundances of proteins that are identified only by shared peptides or by a single unique peptide. Our goal is to propose a statistical methodology capable of including shared peptides in downstream MS data analysis to increase the number of proteins that can be identified and quantified reliably, and improve the power of statistical analysis.

In this talk, we will present two classes of non-linear models that can be used to describe labeled and label-free mass spectrometry experiments with shared peptides: models with peptide-specific weights and non-weighted models. In labeled experiments, multiple biological conditions or subjects may be measured jointly. In this case, peptides have natural quantitative profiles, which we use to estimate the degree of their protein membership (weights). We use these weights to estimate protein-level summaries of peptide data, which are then used for comparisons of biological conditions. We will illustrate proposed models with biological data, and provide analytical and simulation study-based results on their statistical properties.

Research presented in this talk was done in collaboration with Genentech company, Northeastern University (USA) and Hasselt University (Belgium), and was financially supported by the National Science Center grant 2020/37/N/ST6/04070 (Poland).

Early stopping for L^2-boosting in high-dimensional linear models

Thursday, 6.07
11:00–11:20
CS 21
Room 2180

Bernhard Stankewitz

Bocconi University (Italy)

Abstract

Increasingly high-dimensional data sets require that estimation methods do not only satisfy statistical guarantees but also remain computationally feasible. In this context, we consider L^2-boosting via orthogonal matching pursuit in a high-dimensional linear model and analyze a data-driven early stopping time τ of the algorithm, which is sequential in the sense that its computation is based on the first τ iterations only. This approach is much less costly than established model selection criteria, that require the computation of the full boosting path. We prove that sequential early stopping preserves statistical optimality in this setting in terms of a fully general oracle inequality for the empirical risk and recently established optimal convergence rates for the population risk. Finally, an extensive simulation study shows that at an immensely reduced computational cost, the performance of these type of methods is on par with other state of the art algorithms such as the cross-validated Lasso or model selection via a high dimensional Akaike criterion based on the full boosting path.

Bibliography

[1] Stankewitz B., *Early stopping for L2-boosting in high-dimensional linear models*, arXiv: 2210.07850 [math.ST], 2022.

Causal effects of intervening variables in settings with unmeasured confounding

Thursday, 6.07
14:20–14:40
CS 24
Room 1.40

Mats Julius Stensrud[1], **Aaron Sarvet**[1], **Lan Wen**[2]

[1] Ecole Polytechnique Federale de Lausanne (Switzerland)
[2] University of Waterloo (Canada)

Abstract

We present new results on average causal effects in settings with unmeasured exposure-outcome confounding. Our results are motivated by a class of estimands, frequently of interest in medicine and public health, that are currently not targeted by standard approaches for average causal effects. We recognize these estimands as queries about the average causal effect of an *intervening* variable. We anchor our introduction of these estimands in an investigation of the role of chronic pain and opioid prescription patterns, and illustrate how conventional approaches will lead to unreplicable estimates with ambiguous policy implications. We argue that effects of intervening variables are replicable and have clear policy implications, and furthermore are non-parametrically identified by the classical frontdoor formula. As an independent contribution, we derive a new semiparametric efficient estimator of the frontdoor formula with a uniform sample boundedness guarantee. This property is unique among previously-described estimators in its class, and we demonstrate superior performance in finite-sample settings. The theoretical results are applied to data from the National Health and Nutrition Examination Survey.

Bayesian Covariate-Dependent Gaussian Graphical Models with Varying Structure

Wednesday, 5.07
11:00–11:30
IS 02
Room 0.03

Francesco C. Stingo

University of Florence (Italy)

Abstract

We introduce Bayesian Gaussian graphical models with covariates (GGMx), a class of multivariate Gaussian distributions with covariate-dependent sparse precision matrix. We propose a general construction of a functional mapping from the covariate space to the cone of sparse positive definite matrices, that encompasses many existing graphical models for heterogeneous settings. Our methodology is based on a novel mixture prior for precision matrices with a non-local component that admits attractive theoretical and empirical properties. The flexible formulation of GGMx allows both the strength and the sparsity pattern of the precision matrix (hence the graph structure) change with the covariates. Posterior inference is carried out with a carefully designed Markov chain Monte Carlo algorithm which ensures the positive definiteness of sparse precision matrices at any given covariatesâ€™ values. Extensive simulations and a case study in cancer genomics demonstrate the utility of the proposed model. Joint work with Yang Ni and Veerabhadran Baladandayuthapani.

Bibliography

[1] Ni, Y., Stingo, F.C. and Baladandayuthapani, V. *Bayesian Covariate-Dependent Gaussian Graphical Models with Varying Structure*, Journal of Machine Learning Research, 23(242) (2022), 1-29.

Change point estimation for a stochastic heat equation

Thursday, 6.07
17:00–17:30
IS 17
Room 0.03

Markus Reiß[1], **Claudia Strauch**[2], **Lukas Trottner**[2]

[1] Humboldt-Universität zu Berlin (Germany)
[2] Aarhus University (Denmark)

Abstract

We study a change point model based on a stochastic partial differential equation corresponding to the heat equation governed by the weighted Laplacian $\Delta_\vartheta = \nabla\vartheta\nabla$ on the domain $(0,1)$ with Dirichlet boundary conditions. Based on local measurements of the solution in space with resolution δ over a finite time horizon, we develop a simultaneous M-estimator for the diffusivity parameters $\vartheta_\pm$ and the change point τ, characterizing the piecewise constant diffusivity ϑ according to

$$\vartheta(x) = \vartheta_- \mathbf{1}_{(0,\tau)}(x) + \vartheta_+ \mathbf{1}_{[\tau,1)}(x), \quad x \in (0,1).$$

We work in the general setting where the parameters $\vartheta_\pm$ are allowed to vary with the resolution δ, and we establish rates of convergence, which are identical to those obtained in classical change point problems based on independent observations. Moreover, when the diffusivity parameters $\vartheta_\pm$ are known and the jump height $\eta = \vartheta_+ - \vartheta_-$ vanishes with resolution $\delta \to 0$, we derive a limit theorem for the change point estimator and identify the limiting distribution as one familiar from the change point literature. Our investigation is self-contained and fundamentally based on a careful analysis of concentration properties of the empirical Fisher information.

Testing Single-index model under censoring

Wednesday, 5.07
11:00–11:20
CS 18
Room 4420

Ewa Strzalkowska-Kominiak

Universidad Carlos III de Madrid (Spain)

Abstract

The single-index model is a useful tool to incorporate a d-dimensional vector of covariates X into a regression model avoiding the so called "curse of dimensionality". By assuming that there exists a vector of parameters θ_0 so that the response variable depends only on the projection $\theta_0' X$, one avoids a multivariate regression. This assumption is a reasonable compromise between a fully parametric and a fully nonparametric model. Additionally, in medical or economic studies, the large number of explanatory variables is not the only problem. Very often the response variable, $Z \sim F$, is only partly observed and so censored from the right. The efficient estimators for the single-index in question were proposed by [1] and [2]. The goal of the present work is to test the single-index model assumption. Our test statistics is based on the fact that,

$$h(x,t) \equiv E((1_{\{Z \leq t\}} - F_{\theta_0}(t|\theta_0' X))1_{\{X \leq x\}})$$

equals zero under the single-index model, where $F_{\theta_0}(t|u)$ is a conditional CDF of Z given $\theta_0' X = u$. Similarly as in [3] for complete data, we propose a Cramér-von Mises type statistic as follow

$$T_n = \int \hat{h}^2(x,t) F_n(dx, dt).$$

Bibliography

[1] Strzalkowska-Kominiak, E., Cao, R. *Maximum likelihood estimation for conditional distribution single-index models under censoring.* Journal of Multivariate Analysis, 114 (2013), pp. 74–98.

[2] Strzalkowska-Kominiak, E., Cao, R. *Beran-based approach for single-index models under censoring.* Computational Statistics, 29 (2014), pp. 1243–1261.

[3] Xia, Y., Li, W.K., Tong, H., Zhang, D. (2004). *A goodness-of-fit test for single-index models.* Statistica Sinica 14 (2004), pp. 1–39.

Scalable variational Bayes methods for multivariate temporal point processes

Thursday, 6.07
12:00–12:30
IS 13
Room 0.03

Déborah Sulem[1], **Vincent Rivoirard**[2], **Judith Rousseau**[3]

[1] Barcelona School of Economics, Universitat Pompeu Fabra (Spain)
[2] Université Paris-Dauphine, PSL University (France)
[3] University of Oxford (United Kingdom)

Abstract

Multivariate Hawkes processes are temporal point processes extensively applied to model event data with dependence and interaction phenomena, e.g., neuronal spike trains, online messages, and financial transactions. In the nonparametric setting, learning the temporal dependence structure of Hawkes processes is often a computationally expensive task, all the more with Bayesian estimation methods. In the generalised nonlinear Hawkes model, the posterior distribution is non-conjugate and doubly intractable, and existing Monte-Carlo Markov Chain methods are often slow and not scalable to high-dimensional processes in practice. Recently, efficient algorithms targeting a mean-field variational approximation of the posterior distribution have been proposed. In this work, we unify existing variational Bayes inference approaches under a general framework, that we theoretically analyse under easily verifiable conditions on the prior, the variational class, and the model. Then, in the context of the popular sigmoid Hawkes model, we design adaptive and sparsity-inducing mean-field variational methods. In particular, we propose a two-step algorithm based on a thresholding heuristic to select the connectivity graph parameter of the Hawkes model. Through an extensive set of numerical simulations, we demonstrate that our approach enjoys several benefits: it is computationally efficient, can reduce the dimensionality of the problem by selecting the graph parameter, and is able to adapt to the smoothness of the underlying parameter

Bibliography

[1] Feng Zhou, Quyu Kong, Zhijie Deng, Jichao Kan, Yixuan Zhang, Cheng Feng, and Jun Zhu (2022), *Efficient Inference for Dynamic Flexible Interactions of Neural Populations*, Journal of Machine Learning Research, 23(211), 1-49.

[2] Deborah Sulem, Vincent Rivoirard, and Judith Rousseau (2022). *Scalable Variational Bayes methods for Hawkes processes.* arXiv preprint arXiv:2212.00293.

Some properties of generalized aging and reversed aging intensity functions

Friday, 7.07
11:20–11:40
CS 31
Room 1.40

Francesco Buono[1], **Maria Longobardi**[1], **Magdalena Szymkowiak**[2]

[1] Università di Napoli Federico II (Italy)
[2] Poznan University of Technology (Poland)

Abstract

The classic aging intensity and reversed aging intensity functions are defined as the ratio of the instantaneous hazard rate to its average and as the ratio of the instantaneous reversed hazard rate [2] to its baseline value, respectively. They analyzes the aging property quantitatively, the higher the aging intensity, the stronger the tendency of aging, and on the contrary, the higher the reversed aging intensity, the weaker the tendency of aging. We consider families of generalized aging intensity [3] and reversed aging intensity [1] functions. We prove that in some cases the generalized aging and reversed aging intensity functions characterize a unique distribution and in the other situations, determine a family of distributions. Moreover, we introduce generalized aging and reversed aging intensity orders and present their relations with some well known stochastic orders.

Bibliography

[1] Buono, F., Longobardi, M., Szymkowiak, M., *On Generalized Reversed Aging Intensity Functions*, Ricerche di matematica, 71 (2022), 85–108.
[2] Finkelstein, M. S., *On the Reversed Hazard Rate*, Reliability Engineering & System Safety, 78 (2002), 71–75.
[3] Szymkowiak, M., *Generalized Aging Intensity Functions*, Reliability Engineering & System Safety, 178 (2018), 198–208.

Integration by parts formulas for marked Hawkes processes

Monday, 3.07
15:45–16:05
CS 04
Room 5440

Atsushi Takeuchi

Tokyo Woman's Christian University (Japan)

Abstract

On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $\{\tau_k\,;\,k \geq 1\}$ be a simple point process and $\{Z_k\,;\,k \geq 1\}$ a sequence of independent and identically distributed random variables valued in $\mathbb{R}_0 := \mathbb{R}\backslash\{0\}$. The family $\{(\tau_k, Z_k)\,;\,k \geq 1\}$ is called the marked point process. We shall denote the corresponding counting process by $\{L(t,A)\,;\,t \geq 0,\, A \in \mathcal{B}(\mathbb{R}_0)\}$, and the conditional intensity by $\{\lambda(t,A)\,;\,t \geq 0,\, A \in \mathcal{B}(\mathbb{R}_0)\}$. Let $\pi(dz)$ be a finite measure over $\mathbb{R}_0$, and $h : [0,+\infty) \to [0,+\infty)$ be bounded and Borel measurable. Suppose that

(i) the random variable Z_1 has a smooth Lebesgue density f such that

$$\lim_{|z|\nearrow+\infty} f(z) = 0, \tag{1}$$

(ii) the $\mathbb{L}^1([0,+\infty))$-norm of the function h is strictly smaller than 1.

Then, the marked Hawkes process is defined as the marked counting one $\{L(t,A)\,;\,t \geq 0,\, A \in \mathcal{B}(\mathbb{R}_0)\}$ with the conditional intensity $\{\lambda(t,A)\,;\,t \geq 0,\, A \in \mathcal{B}(\mathbb{R}_0)\}$ such that

$$\lambda(t,A) = \pi(A) + \int_{(0,t)\times A} h(t-s)\, L(ds,dz) \tag{2}$$

for $t > 0$ and $A \in \mathcal{B}(\mathbb{R}_0)$, and $\lambda(0,A) = 0$ (cf. [1] and [2]).

Now, let us consider the processes $\{L_t\,;\,t \geq 0\}$ and $\{\lambda_t\,;\,t \geq 0\}$ defined by

$$L_t = \int_0^t \int_{\mathbb{R}_0} \ell(z)\, L(ds,dz), \quad \lambda_t = \int_{\mathbb{R}_0} \ell(z)\, \lambda(t,dz), \tag{3}$$

where $\ell \in C^\infty(\mathbb{R}_0\,;\,[0,+\infty))$ such that all derivatives of any orders more than 1 are bounded, and that the function ℓ is $\mathbb{L}^p(\mathbb{R}_0)$-integrable for $p \geq 2$. In the presentation, we shall focus on the following topics:

- the integration by parts formulas on the processes L and λ,
- the study on the (conditional) densities of $\mathbb{P}[L_t \in dy | N_t \geq 1]$ and $\mathbb{P}[\lambda_t \in dy | N_t > 1]$,
- the sensitivity on $\mathbb{E}\big[\varphi(L_t)\,\mathbb{I}_{(N_t \geq 1)}\big]$ and $\mathbb{E}\big[\varphi(\lambda_t)\,\mathbb{I}_{(N_t > 1)}\big]$.

The details on the presentation can be seen in [3] and [4].

Bibliography

[1] Brémaud P., Nappo G., Torrisi G. L., *Rate of convergence to equilibrium of marked Hawkes processes*, Journal of Applied Probability, 39 (2002), 123–136.

[2] Hawkes A. G., *Spectra of some self-exciting and mutually exciting point processes*, Biometrika, 58 (1971), 83–90.

[3] Takeuchi A., *The Bismut-Elworthy-Li type formulae for stochastic differential equations with jumps*, Journal of Theoretical Probability, 23 (2010), 576–604.

[4] Takeuchi A, *Integration by parts formulas for marked Hawkes processes*, Statistics and Probability Letters, 145 (2019), 229–237.

Treatment Effect in a Proportional Hazards Model with Instrumental Variables

Thursday, 6.07
11:40–12:00
CS 21
Room 2180

Lorenzo Tedesco, Jad Beyhum, Ingrid Van Keilegom

[1] KU LEUVEN, Belgium

Abstract

In this paper we consider the Cox's proportional hazards model, proposing a novel approach for the identification and estimation of the causal hazard ratio in the presence of unmeasured confounding factors and random right censoring. The approach is based on a discrete instrumental variable and allows to include a discrete endogenous covariate and any type of exogenous covariate. A non-parametric instrumental regression for right censored duration outcomes is used to estimate the parameter of interest of the underlying semiparametric proportional hazards model. Asymptotic properties of the estimator are derived. The approach is illustrated via simulation studies and a data application.

Cost-constrained multi-label variable selection using information theory

Tuesday, 4.07
11:20–11:40
CS 06
Room 1.40

Paweł Teisseyre[1,2], **Tomasz Klonecki**[2], **Jaesung Lee**[3]

[1] Warsaw University of Technology (Poland)
[2] Polish Academy of Sciences (Poland)
[3] Chung-Ang University (Seoul)

Abstract

We consider a cost-constrained variable selection task [1] in which the goal is to select a subset of variables to predict the target vector $Y = (Y_1, \ldots, Y_q)$ while satisfying a user-specific maximal admissible budget. Unlike in traditional feature selection problems, we assume that there are costs $c_1, \ldots, c_p$ assigned to variables $X = (X_1, \ldots, X_p)$. The cost of the variable is often associated with obtaining its value; for example, in medical applications, it can be the value extracted by a diagnostic test. The cost-constrained variable selection can be stated using the information-theoretic framework

$$S_{opt} = \arg \max_{S: C(S) \leq T} I(Y, X_S), \tag{1}$$

where $I(Y, X_S)$ is mutual information, X_S denotes a subvector of X corresponding $S \subseteq \{1, \ldots, p\}$, T is a budget, and $C(S) = \sum_{j \in S} c_j$ indicates the cost associated S. Solving task (1) directly is infeasible because the number of possible subsets S in (1) may increase exponentially. Therefore, we propose a method based on a sequential forward selection procedure that adds a new candidate feature X_k that maximizes

$$\underbrace{I(Y, X_{S\cup\{k\}}) - I(Y, X_S)}_{\text{relevance term}} - \underbrace{\lambda \cdot c_k}_{\text{penalty for the cost}} \quad \propto \quad I(Y, X_{S\cup\{k\}}) - \lambda \cdot c_k$$

where $\lambda > 0$ is the cost factor controlling the trade-off between relevance and cost. The first term measures the informativeness of the candidate feature in the context of already-selected features, and the second term is a penalty for its cost. Since, estimation of $I(X_{S\cup\{k\}})$ is challenging for large subsets S, in our method we use the lower bound of $I(X_{S\cup\{k\}})$ as proposed in [2]. The choice of the optimal value of λ is essential in the proposed method. The optimal λ should depend on the available budget and the number of relevant variables. Therefore, we propose a data-driven method of finding the optimal value of λ based on calculating the cumulative sum of the relevance terms for selected features. The experimental results on simulated, publicly available, and real-world datasets revealed that the proposed method consistently outperforms conventional methods when the available budget is low.

Bibliography

[1] Teisseyre, P. and Klonecki, T., *Controlling Costs in Feature Selection: Information Theoretic Approach*, Proceedings of the International Conference on Computational Science ICCS'21, 2021.

[2] Teisseyre, P. and Lee, J., *Multilabel all-relevant feature selection using lower bounds of conditional mutual information*, Expert Systems with Applications, 216 (2023), 1–19.

Visibility graph analysis of the reservoir-triggered seismicity at Pertusillo lake, Southern Italy

Tuesday, 4.07
16:00–18:00
Poster

Luciano Telesca, Tony A. Stabile

Institute of Methodologies for Environmental Analysis, National Research Council (Italy)

Abstract

The visibility graph (VG) has become a statistical method widely employed to characterize the dynamical properties of time series. Developed by Lacasa et al. [1] the VG converts time series into graphs, whose nodes represent the series values linked between each other by their reciprocal "visibility". Recently, the VG has been used for the statistical investigation of seismicity ([2], [3], [4], [5]), due to its capability to disclose complex features of seismic processes. One of the most interesting features that was found is the relationship between the seismic parameter b-value of the Gutenberg-Richter law and the topological VG parameter k-M slope (where M is the magnitude of the earthquake and k its connectivity degree) that seems to be characterized by universality. In this paper we applied the VG to the reservoir-triggered seismicity at Pertusillo lake (southern Italy) to analyse the relationship between the b-value and k-M. Furthermore, we applied the Multiplex Visibility Graph (MVG) [6] to study the correlation between the seismicity and the water level and volume of the reservoir. Our results strengthen the univeral character of the relationship between seismic and topological parameters, and suggest the usefulness of the VG to uncover new insights into the joint time dynamics of the seismo-hydrological process at Pertusillo reservoir.

Bibliography

[1] Lacasa L., Luque B., Ballesteros F., Luque J., Nuno J.C.*From time series to complex networks: The visibility graph*, Proceedings of the National Academy of Sciences, 105 (2008), 4972–4975.

[2] Telesca L., Lovallo M., *Analysis of seismic sequences by using the method of visibility graph*, Europhysics Letters, 97 (2012), 50002.

[3] Telesca L., Lovallo M., Ramirez-Rojas A., Flores-Marquez L. *Investigating the time dynamics of seismicity by using the visibility graph approach: Application to seismicity of Mexican subduction zone*, Physica A, 392 (2013), 6571–6577.

[4] Azizzadeh-Roodpish S., Cramer C. H. *Visibility graph analysis of Alaska crustal and Aleutian subduction zone seismicity: an investigation of the correlation between b value and k-M slope*, Pure and Applied Geophysics, 175 (2018), 4241–4252.

[5] Telesca L., Thai A. T. , Lovallo M., Cao D. T. *Visibility Graph Analysis of Reservoir-Triggered Seismicity: The Case of Song Tranh 2 Hydropower, Vietnam*, Entropy, 24 (2022), 1620.

[6] Lacasa L., Nicosia V., Latora V. *Network structure of multivariate time series*, Scientific Reports, 5 (2015), 15508.

Gaussian copula estimation for heterogeneous data

Tuesday, 4.07
16:00–18:00
Poster

Ekaterina Tomilina[1,2], **Gildas Mazo**[1], **Florence Jaffrezic**[2]

[1] Universite Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France
[2] Universite Paris-Saclay, AgroParisTech, INRAE, GABI, 78350 Jouy-en-Josas, France

Abstract

Copulas can be a good strategy to build models when working with multivariate data. For instance, the Gaussian copula links the marginal cumulative distribution functions (CDFs) of the variables together in the following way:

$$F(x_1, ..., x_d) = C_\Sigma(F_1(x_1), ..., F_d(x_d)) \equiv \Phi_\Sigma(\Phi^{-1}(F_1(x_1)), ..., \Phi^{-1}(F_d(x_d))) \quad (1)$$

where F_j denotes the marginal CDF of variable X_j, Φ_Σ denotes the Gaussian CDF of mean 0 and correlation matrix Σ and Φ^{-1} denotes the inverse standard Gaussian CDF. Semiparametric inference in model (1) has been addressed in [1, 2], where asymptotic normality is studied. However, this approach remains limited to the case of continuous variables. We propose a semi-parametric inference method for the correlation matrix of the Gaussian copula described in equation (1) in the case of heterogeneous data. To infer the correlation coefficients of Σ, the pairwise likelihood [3], given by:

$$L_n(\Sigma) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{j'=j+1}^{d} \log f_{jj'}(X_{ij}, X_{ij'}; \widehat{F}_j, \widehat{F}_{j'}, \Sigma)$$

is used, where $\widehat{F}_j$ denotes the empirical CDF or a truncated version of it, $f_{jj'}$ corresponds to the density of the pair $(X_{ij}, X_{ij'})$ and can take three different forms depending on the nature of the variables. For instance, if (X_1, X_2) is a pair of heterogeneous variables where X_1 is continuous and X_2 is discrete, it can be expressed as [4]:

$$f(x_1, x_2) = f_1(x_1)\int_{F_2(x_2-)}^{F_2(x_2)} c_\rho(F_1(x_1), v)dv$$

where c_ρ corresponds to the density of the Gaussian copula. Our main result is to show consistency of the pairwise maximum likelihood estimator in the mixed case (in the continuous case, it can be deduced from Liu [5]) and study asymptotic normality.

Bibliography

[1] Genest C., Ghoudi K., Rivest L-P., *A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions*, Biometrika, Vol. 82 No. 3 (1995), 543–552.

[2] Tsukahara H., *Semiparametric estimation in copula models*, The Canadian Journal of Statistics, Vol. 33 No. 3 (2005), 357–375.

[3] Mazo G., Karlis D., Rau A., *A randomized pairwise likelihood method for complex statistical inferences*, hal-03126620, 2022.

[4] Song P., *Correlated data analysis*, Springer, 2010

[5] Liu H., Lafferty J., Wasserman L., *The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs*, Journal of Machine Learning Research, 10 (2009), 2295–2328.

Mixed-effects models and their applications

Tuesday, 4.07
16:00–18:00
Poster

Andrzej Tomski

University of Silesia in Katowice

Abstract

Mixed-effects models are an extension of classical fixed-effects models to allow both fixed and random effects, and they are particularly used when there is non independence in the data, such as arises from a hierarchical structure. Wide range of practical applications can be found, including mathematical statistics through experimental biologists research and even they can be used for commercial research, especially in pharmacy branch. This talk presents basic theoretical facts together with applications in biology and medicine as well as some challenges faced by these models.

Bibliography

[1] Jiang, J. *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, 2007.

[2] Nowakowska J., Gazda A., Tomski A., Szwagrzyk J. *Drainage ditches enhance forest succession in a raised bog but do not affect the spatial pattern of tree encroachment,* PLoS One 16(3) (2021).

[3] Wojtasz, I., Cofta, S., Czudaj, P., Jaracz, K., Kazmierski, R. *Effect of Face Masks on Blood Saturation, Heart Rate, and Well-Being Indicators in Health Care Providers Working in Specialized COVID-19 Center*, International Journal of Environmental Research and Public Health, Vol. 19 (2022).

Multivariate goodness of fit tests based on topological data summaries

Friday, 7.07
11:40–12:00
CS 33
Room 5440

Rafał Topolnicki, Niklas Hellmer, Paweł Dłotko, Łukasz Stettner

Dioscuri Centre in Topological Data Analysis Mathematical Institute of the Polish Academy of Sciences, Warsaw, Poland

Abstract

In the talk a novel application of topology to statistical hypothesis testing will be presented [1]. A new approach, driven by concepts originating in Topological Data Analysis, to multivariate one- and two-sample goodness-of-fit tests will be discussed. The presented tests work for samples in arbitrary dimensions, having comparable power to the state-of-the-art tests in the one-dimensional case. It is demonstrated that the type I error can be controlled and their type II error vanishes with increasing sample size. Derived theoretical results will be discussed and the finite-sample performance will be addressed by extensive numerical conducted to demonstrate their finite sample power.

Bibliography

[1] `https://arxiv.org/abs/2210.14965`

Dispersal density estimation across scales

Thursday, 6.07
14:30–15:00
IS 09
Room 0.03

Marc Hoffmann[1], **Mathias Trabs**[2]

[1] Université Paris-Dauphine (France)
[2] Karlsruhe Insistute of Technology (Germany)

Abstract

We build a family of statistical experiments across spatial scales that exhibit nontrivial behaviours at certain critical levels and for which different estimation procedures with different rates of convergence enter into competition as the scale varies. To this end, we study a representative model that may serve in several applications, ranging from queuing theory to particle diffusions or biological dispersion of species and even molecular genetics.

We consider a space structured population model generated by two point clouds: a homogeneous Poisson process M with intensity $n \to \infty$ as a model for a parent generation together with a Cox point process N as offspring generation, with conditional intensity given by the convolution of M with a scaled dispersal density $\sigma^{-1} f(\cdot/\sigma)$. Based on a realisation of M and N, we study the nonparametric estimation of f and the estimation of the physical scale parameter $\sigma > 0$ simultaneously for all regimes $\sigma = \sigma_n$.

We establish that the optimal rates of convergence do not depend monotonously on the scale and we construct minimax estimators accordingly whether σ is known or considered as a nuisance, in which case we can estimate it and achieve asymptotic minimaxity by plug-in. The statistical reconstruction exhibits a competition between a direct and a deconvolution problem. Our study reveals in particular the existence of a least favourable intermediate inference scale, a phenomenon that seems to be new.

Bibliography

[1] Hoffmann, M., Trabs, M., *Dispersal density estimation across scales*, arXiv: 2108.05279, 2021.

Changepoint detection in large factor models

Monday, 3.07
16:45–17:15
IS 01
Room 0.03

Matteo Barigozzi[1], **Haeran Cho**[2], **Lorenzo Trapani**[3]

[1] University of Bologna (Italy)
[2] University of Bristol (UK)
[3] University of Leicester (UK)

Abstract

We study changepoint detection in aÂ large factor model, proposing both offline and online detection methods. In all cases, we build on the idea that, in aÂ factor model, under the maintained assumption of the homoscedasticity of the common factors, in the presence of aÂ changepoint the second moment matrix of the common factors changes. Hence, aÂ high-dimensional problem can be conveniently cast into aÂ low-dimensional problem, based on checking for breaks in the second moment matrix of the estimated common factors. Our results build on aÂ maximal inequality for the partial sums of the second moment matrix of the estimated common factors. By virtue of this, we are able to derive strong invariance principles for the partial sums of the estimated second moment matrix of the common factors. Hence, we propose aÂ family of weighted CUSUM statistics for the offline detection of change points, including the standardised CUSUM process (for which we derive aÂ Darling-Erdos theorem) and even more heavily weighted statistics. Our tests have power versus breaks occurring also close to the sample endpoints. We further study the asymptotics of the MOSUM process, and of the maximally selected LR statistic. In addition, we investigate the problem of sequential, online detection of changepoints, proposing aÂ family of detectors which ensure procedure-wise size control and short detection delays; as aÂ by-product, we derive the limiting distribution of the detection delay.

Concentration analysis of multivariate elliptic diffusions

Monday, 3.07
16:05–16:25
CS 03
Room 2180

Lukas Trottner[1], **Cathrine Aeckerle–Willems**[2], **Claudia Strauch**[1]

[1] Aarhus University (Aarhus)
[2] University of Mannheim (Germany)

Abstract

We prove concentration inequalities and associated PAC bounds for continuous- and discrete-time additive functionals for possibly unbounded functions of multivariate, nonreversible diffusion processes. Our analysis relies on an approach via the Poisson equation allowing us to consider a very broad class of subexponentially ergodic processes. These results add to existing concentration inequalities for additive functionals of diffusion processes which have so far been only available for either bounded functions or for unbounded functions of processes from a significantly smaller class. We demonstrate the usefulness of the results by applying them in the context of high-dimensional drift estimation and Langevin MCMC for moderately heavy-tailed target densities.

i-learner: an orthogonal learner of the conditional counterfactual outcome mean with output space constraints

Thursday, 6.07
14:00–14:20
CS 24
Room 1.40

Stijn Vansteelandt, Pawel Morzywolek

Ghent University (Belgium)

Abstract

Orthogonal meta-learners, such as DR-learner, R-learner and IF-learner, are increasingly used to estimate conditional average treatment effects. However, these meta-learners have a key limitation in that they apply standard learners to transformed outcome data, which can lead them to disregard the possibly constrained output space. This can be particularly problematic for dichotomous outcomes, which typically get transformed to unconstrained values outside the unit interval, making it difficult for standard learners to guarantee predictions within the unit interval. To address this issue, we propose a new approach to constructing orthogonal meta-learners for estimating conditional counterfactual means that respect the constrained output space. Our approach provides broader insight into the construction of orthogonal learners for other estimands. By respecting the outcome space, it is more generally expected to outperform existing learners, as we confirm empirically in simulation studies and an analysis of critical care data.

Single coordinate semi-parametric Bernstein-von-Mises theorems in high-dimensional linear regression

Wednesday, 5.07
12:00–12:30
IS 03
Room 0.06

Ismaël Castillo, Stéphanie van der Pas, Kolyan Ray, Aad van der Vaart, Lasse Vuursteen

Abstract

Consider the setting of high-dimensional linear regression where there is one particular covariate of interest, for which we want uncertainty quantification. We study conditions under which the coefficient of this covariate satisfies a Bernstein-von-Mises theorem when the rest of the covariates are drawn from a model selection prior. We show that under conditions where a Bernstein-von-Mises type of result fails to hold for the model selection prior, peturbing the prior in with an estimate of the least favourable direction allows for obtaining a Bernstein-von-Mises type of result for the coordinate of interest still.

Inference for multivariate random fields

Monday, 3.07
16:45–17:15
IS 04
Room 0.06

Jonas Wallin

Abstract

Inference for multivariate spatial data requires modeling both the marginal and cross-covariance functions. While the marginal covariance functions are always symmetric, the cross-covariance functions may not be. This can occur due to a lag between the fields, meaning that the maximum correlation between observations of two fields is not obtained by setting the same location for both fields. Instead, the location of the second field is shifted by some vector, known as the shift parameter. In this study, we investigate the asymptotic properties of the shift parameters under infill asymptotics. That is how an increasing number of observations within a fixed area affects parameter estimation.

On polynomial-time mixing of MCMC for high-dimensional posterior distributions

Friday, 7.07
15:00–15:30
IS 05
Room 0.03

Sven Wang

Massachusetts Institute of Technology (USA)

Abstract

This work considers the problem of generating random samples from high-dimensional posterior distributions. We will discuss two recent papers which respectively consider (i) conditions under which diffusion-based MCMC algorithms converge in polynomial-time and (ii) situations in which MCMC suffers from an exponentially long mixing time. The first paper [1] is joint work with R. Nickl, while the second is joint work with A. Bandeira, A. Maillard and R. Nickl [2].

We focus on the setting of non-linear inverse regression models. Our positive results on polynomial-time mixing in part (i) are derived under local 'gradient stability' assumptions on the forward map, which can be verified for a range of well-known non-linear inverse problems involving elliptic PDE (see [3]). Our results are valid under the assumption that a sufficiently good initializer is available, and we prove that such initializers can be computed in polynomial-time, for a prototypical non-linear inverse problem involving the steady-state Schrödinger equation. Our negative results on exponentially long mixing times in part (ii) hold for 'cold-start' MCMC. We show that there exist non-linear regression models in which the posterior distribution is unimodal, but there exists a so-called 'free entropy barrier', which local Markov chains take an exponentially long time to traverse. Our counter-examples hold for general MCMC schemes based on gradient or random walk steps, and the theory is illustrated for Metropolis-Hastings adjusted methods such as pCN and MALA.

Bibliography

[1] Nickl, R. and Wang, S. *On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms*, Journal of the European Mathematical Society, to appear.

[2] A.S. Bandeira, A. Maillard, R. Nickl and S. Wang, *On free energy barriers in Gaussian priors and failure of MCMC for high-dimensional unimodal distributions*, Philosophical Transactions of the Royal Society A 381, 2023.

[3] Nickl, R., *Bayesian non-linear statistical inverse problems*, European Mathematical Society Press, in press.

Contributions of Hugo Steinhaus to Probability Theory

Friday, 7.07
11:00–11:30
IS 16
Room 0.03

Aleksander Weron

Wrocław University of Science and Technology (Poland)

Abstract

The purpose of this talk is to explain what happened exactly one hundred years ago and how probability theory, an important new sub-discipline of mathematics, came into being [1]-[4]. Interesting enough, all four papers were published in one volume of the Polish journal Fundamenta Math. 4 (1923)! We will focus on the role of Hugo Steinhaus' article [2]. In addition to "discovering" Stefan Banach in 1916 and collaborating with him in the field of functional analysis, Steinhaus pioneered the foundations of probability theory, anticipating Kolmogorov [4, 5], and game theory, anticipating von Neumann. See [6] for details.

Bibliography

[1] Łomnicki M., Nuoveaux fondements de la theorie des probabilites. Defnition de la probabilite fondee sur la theorie des ensembles), Fundamenta Math. 4 (1923), 34–71.
[2] Steinhaus H., Les probabilites denombrables et leur rapport a la theorie de la mesure, Fundamenta Math. 4 (1923), 286–310,
[3] Frechet A., Familles et fonctions additives d'ensembles abstraits, Nuoveaux fondements de la theorie des probabilites. Defininition de la probabilite fondee sur la theorie des ensembles, Fundamenta Math. 4 (1923), 329–365.
[4] Kolmogorov A.N., Une serie de Fourier-Lebesgue divergent presque partout. Fundamenta Math. 4 (1923), 324–328.
[5] Kolmogorov A.N., Grundbergrife der Wahrscheinlichkeitsrechnung. Springer, Berlin 1933.
[6] Steinhaus H., Mathematician for All Seasons. Recollections and Notes, Vol. 1 (1887-1945) Birkhauser 2015 and Vol.2 (1945–1968), Birkhauser 2016.

Graphical Dirichlet type priors for negative-multinomial/multinomial graphical models

Thursday, 6.07
11:00–11:20
CS 19
Room 1.01

Jacek Wesołowski

Warsaw University of Technology (Poland)

Abstract

We introduce a multivariate distribution, called G-Dirichlet, where G is a decomposable graph. It interpolates between product of univariate beta laws and the classical Dirichlet law on the unit simplex. The distribution has intriguing independence properties, which extend neutralitities of the classcial Dirichlet. These and other features are due to properties of the clique polynomial

$$\Delta_G(x) = \sum_{C \in \mathcal{C}_{G^*}} (-1)^{|C|} \prod_{i \in C} x_i,$$

which is the core factor in the density. In the formula above G^* is a graph complementary to G, i.e. $v \sim w$ in G^* iff $v \not\sim w$ in G, $\mathcal{C}_{G^*}$ is the set of cliques in G^* and $|C|$ is the number of vertices in clique C.

We show that G-Dirichlet law works nicely as a conjugate prior in a parametric discrete Markov graphical models of the negative-multinomial type, which we also introduce, relating it to the polynomial Δ_G. Negative-multinomial graphical model interpolates between product of univariate negative binomial distribution and the classical negative multinomial law. They also exhibit a surprising connection to the Cartier-Foata version of the McMahon Master theorem.

This is a joint project with Bartosz Kołodziejek (Politechnika Warszawska) and Xiaolin Zeng (Universite de Strasbourg).

How to estimate time-dependent parameters in SDEs-based models via artificial neural networks

Friday, 7.07
15:00–15:20
CS 35
Room 0.06

Andrzej Kałuża[1], **Paweł M. Morkisz**[1], **Bartłomiej Mulewicz**[2], **Paweł Przybyłowicz**[1], **Martyna Wiącek**[1]

[1] AGH University of Science and Technology (Poland)
[2] Aigorithmics sp. z o. o. (Poland)

Abstract

Estimation of parameters in SDE-based models is a complex problem that has important practical applications in many fields, for example in finances or energy prices or consumption forecasting. Especially, estimating time-dependent parameters is a challenge due to the multitude of estimated values, and the most common approach is simplification using a piecewise-constant functions. We propose a novel method for estimation time-dependent parameters that is based on neural networks and which extends the approach known from [1] for the regression in the heteroscedasticity case.

In our algorithm we define a suitable loss function based on the maximum likelihood approach, which enables us to translate our approximation task into an optimization problem. This enable us to use deep learning techniques. Moreover, we prove that under certain conditions

$$\sup_{0\le t\le T}\Big(\mathbb{E}\|X_1(t)-X_2(t)\|^2\Big)^{1/2} \le C\Big(\|\Theta_1-\Theta_2\|_\infty^{\alpha_1}+\|\Theta_1-\Theta_2\|_\infty^{\alpha_2}\Big),$$

where X_1 is the solution process of the underlying SDE with the real parameter function $\Theta_1=\Theta_1(t)$ and X_2 is the solution process of the same SDE but with the parameter function $\Theta_2=\Theta_2(t)$ that is obtained from our neural network and which approximates Θ_1.

We demonstrate the effectiveness of our approach through a series of numerical experiments using the Deep Learning framework Tensorflow. We compare trajectories of solutions generated with two sets of parameter functions – real and approximations obtained by the neural network. Moreover, we also compare the distributions of $X_1(T)$ and $X_2(T)$.

Bibliography

[1] O. Dürr, B. Sick, E. Murina, *Probabilistic Deep Learning with Python, Keras and TensorFlow Probability*, Manning, 2020.

[2] S.N. Cohen, R.J. Elliot, *Stochastic calculus and applications, 2nd ed.*, Probability and its applications, New York, Birkhäuser, 2015.

[3] R. Situ, *Theory of Stochastic Differential Equations with Jumps and Applications*, Springer Science+Business Media, 2005.

Modelling Predictive Uncertainty in Probabilistic and Black-box Models

Tuesday, 4.07
14:40–15:00
CS 09
Room 1.01

Priyantha Wijayatunga

Umeå University (Sweden)

Abstract

Probabilistic modelling is popular framework for uncertainty handling and reasoning. Especially, well-known probabilistic Bayesian networks (BNs) are a better tool for the task [1]. However, ideally their predictions should be presented with respective uncertainties, especially when they are used for subsequent decision making. Here we show a simple way to express the uncertainties in their predictions, which is otherwise done in complicated ways or ignored. We show that, in discrete predictions (classifications) uncertainty of them can be represented by defining corresponding virtual parameters for the BN. Note that the BN is defined through a set of parameters which represents the uncertainties of the learnt probabilities since one can define probability distributions for them. When a virtual parameter (vector) is defined for a prediction, then it should represent the predictive uncertainties. The virtual parameter for any given prediction (which itself is a probability) can be defined in the similar way to defining any learnable parameter. However, its "precision" (e.g. in Dirichlet case) should be selected appropriately. We show how to do it in a simple way. So, our approach is an easy but logical and computational method.

And then we propose a similar method for representing the uncertainties of predictions in black-box models such as neural networks, support vector machines, random forest models, etc. [2]. Even though these models have immense predictive power they are poor in terms of explainability and giving uncertainties of predictions [3]. Both are regarded as serious drawbacks as far as applications in medicine, biology, etc. are concern. Our method is based on defining a multinomial-Dirichlet model for the learnt black-box models. Then we propose how to use this model in individual predictions in a simple and logical way. Since many black-box models do not give prediction probabilities it is challenging to define them for the individual cases. However, as in the case of the BNs, the precisions of these predictions could be defined in terms of the training data classification results. We discuss these challenges in details and present our solutions.

Bibliography

[1] Cowell R. G., Dawid P., Lauritzen S. L., Spiegelhalter D. J., *Probabilistic Networks and Expert Systems*, Springer-Verlag, 1999.

[2] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[3] Moloud A., Farhad P., et al., *A review of uncertainty quantification in deep learning: Techniques, applications and challenges*, Information Fusion, 76 (2021), 243–297.

Causal regularization

Wednesday, 5.07
11:30–12:00
IS 02
Room 0.03

Lucas Kania[1], **Ernst C. Wit**[2]

[1] Carnegie Mellon University (USA)
[2] Università della Svizzera italiana (Switzerland)

Abstract

In recent decades, several data analytic ways of dealing with causality have been introduced, such as propensity score matching [1], the PC algorithm [2] and Causal Dantzig [4]. Although originally hailed for their interpretational appeal, here we study the identification of causal-like models from in-sample data that provide out-of-sample risk guarantees when predicting a target variable from a set of covariates.

Whereas ordinary least squares provides the best in-sample risk with limited out-of-sample guarantees, causal models have the best out-of-sample guarantees by sacrificing in-sample risk performance. We introduce *causal regularization*, by defining a trade-off between these properties. As the regularization increases, *causal regularization* provides estimators whose risk is more stable at the cost of increasing their overall in-sample risk. The increased risk stability is shown to result in out-of-sample risk guarantees. We provide finite sample risk bounds for all models and prove the adequacy of cross-validation for attaining these bounds. Details of the method can be found in [3].

Bibliography

[1] Austin P.C., *A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003*, Statistics in medicine, 27(12), 2037–2049, 2008.
[2] Glymour C., Zhang K., Spirtes P., *Review of causal discovery methods based on graphical models*, Frontiers in genetics, 10, 524, 2019.
[3] Kania L., Wit E.C., *Causal Regularization: A trade-off between in-sample risk and out-of-sample risk guarantees*, arXiv preprint arXiv:2205.01593, 2022.
[4] Rothenhäusler D., Bühlmann P., Meinshausen N., *Causal Dantzig*, The Annals of Statistics, 47(3), 1688–1722, 2019.

Verifying the validity of exponentiality

Thursday, 6.07
15:20–15:40
CS 25
Room 2180

Grzegorz Wyłupek

Institute of Mathematics, University of Wrocław (Poland)

Abstract

The paper proposes a new goodness-of-fit test for the composite null hypothesis of exponentiality. The new solution is based on an appropriately weighted empirical process with the estimated parameter. The test statistic is an Anderson-Darling type solution while the weight is selected in such a manner that the random variables being the values of the introduced empirical process are asymptotically $N(0,1)$ distributed under the null model. As a result, they serve as a tool allowing one to distinguish different types of deviations from exponentiality. Such an approach sheds a new light on a possible classification of the popular alternatives considered in the literature. The related test is universally consistent. An extensive simulation study and real data examples demonstrate the behaviour of the new test in practice.

Bibliography

[1] Baringhaus, L., Henze, N., *A class of consistent tests for exponentiality based on the empirical Laplace transform*, Annals of the Institute of Statistical Mathematics, 43 (1991), 551–564.

[2] Cox D. R., Oakes D., *Analysis of Survival Data*, Chapman and Hall, (1984).

[3] Csörgő, S., *Weighted correlation tests for scale families*, TEST, 11 (2002), 219–248.

[4] de Wet, T., *Goodness-of-fit tests for location and scale families based on a weighted L2-Wasserstein distance measure*, TEST, 11 (2002), 89–107.

[5] de Wet, T., Venter, J. H., *A goodness of fit test for a scale parameter family distributions*, South African Statistical Journal, 7 (1973), 35–46.

[6] Epps, T. W., Pulley, L. B., *A test for exponentiality vs. monotone hazard alternatives derived from the empirical characteristic function*, Journal of the Royal Statistical Society, B, 48 (1986), 206–213.

[7] Gail, M. H., Gastwirth, J. L., *A scale-free goodness-of-fit test for the exponential distribution based on the Gini statistic*, Journal of the Royal Statistical Society, B, 40 (1978), 350–357.

[8] Grzegorzewski, P., Wieczorkowski, R., *Entropy-based goodness-of-fit test for exponentiality*, Communications in Statistics - Theory and Methods, 28 (1999), 1183–1202.

[9] Henze, N., Meintanis, S. G., *Goodness-of-fit tests based on a new characterization of the Exponential distribution*, Communications in Statistics - Theory and Methods, 31 (2002), 1479–1497.

[10] Kallenberg, W. C. M., Ledwina, T., *Data-driven smooth tests when the hypothesis is composite*, Journal of the American Statistical Association, 92 (1997), 1094–1104.

[11] Shapiro, S. S., Wilk, M. B., *An analysis of variance test for the exponential distribution (complete samples)*, Technometrics, 14 (1972), 355–370.

[12] Wyłupek, G., *Verifying the validity of exponentiality*, Under review, (2023).

Detecting changes in a distributed system in real time with unknown parameters: from Gaussian to mixed-type data

Wednesday, 5.07
11:20–11:40
CS 14
Room 1.01

Ziyang Yang, Idris Eckley, Paul Fearnhead

Lancaster University (United Kingdom)

Abstract

With the advent of the Internet of Things, it is increasingly common to have large networks of sensors. Each sensor has a limited computing ability and the ability to transmit data to a central cloud. However, the communication between sensors and the cloud can be very expensive. Often we are interested in detecting events that lead to changes in properties of the data at one or more sensors. This leads to a new challenge: to detect changes within such a network in real-time with as little communication and computation as possible. Two distributed detection have been developed for this setting [2], [5]. However, these methods assume known distributions for the data both and after the change, and can lose power if these are set incorrectly.

A recently proposed algorithm - FOCuS [3], can detect changes in Gaussian data but relaxes the assumptions of known pre and post-change means, yet is still computationally efficient and suitable for detecting changes in real-time. We have extended FOCuS into a distributed setting. Simulation results suggest that our method, disFOCuS, can achieve similar performance to the idealised setting, where we have no constraints on communication between sensors, but substantially reduces the transmission costs.

Furthermore, driven by applications in the field of the internet of things that sensors processing different data types, we extend disFOCuS to mixFOCuS, which can detect changes in mixed-type data, where data from different sensors follow different, possibly non-Gaussian, distributions. This algorithm is based on the work - expFOCuS [4] that can detect a change for exponential family models in a constant time. Simulation results suggest that mixFOCuS can provide more reliable results than disFOCuS. We also demonstrate its utility on the AnoML-IoT dataset [1] where sensors monitor temperature, humidity, light, loudness and air quality in a room.

Bibliography

[1] Kayan, H., Majib, Y., Alsafery, W., Barhamgi, M., Perera, C., *AnoML-IoT: An end to end re-configurable multi-protocol anomaly detection pipeline for Internet of Things*, Internet of Things, 16 (2021), p. 100437.

[2] Liu, K., Zhang, R., Mei, Y., *Scalable sum-shrinkage schemes for distributed monitoring large-scale data streams*, Statistica Sinica, 29.1 (2019), 1–22.

[3] Romano, G., Eckley, I., Fearnhead, P., Rigaill, G., *Fast online changepoint detection via functional pruning CUSUM statistics*, arXiv preprint arXiv:2110.08205, 2021.

[4] Ward, K., Romano, G., Eckley, I., Fearnhead, P,*A Constant-per-Iteration Likelihood Ratio Test for Online Changepoint Detection for Exponential Family Models*, arXiv preprint arXiv:2302.04743, 2023.

[5] Zhang R., Mei Y., *Asymptotic statistical properties of communication-efficient quickest detection schemes in sensor networks*, Sequential Analysis, 37.3 (2018), 375–96.

Riemannian Functional Regression and Reproducing Kernel Tensor Hilbert Spaces

Tuesday, 4.07
15:00–15:20
CS 10
Room 1.40

Ke Yu, James W. Taylor

University of Oxford (UK)

Abstract

In many scientific fields, data arise in the form of smooth functions on Riemannian manifolds. Analyzing the relationship between a Riemannian functional response and a Riemannian functional predictor has become increasingly important. We develop a Riemannian function-on-function regression model under the reproducing kernel tensor Hilbert space (RKTHS) framework. As an extension of vector-valued reproducing kernel Hilbert spaces, the RKTHS we construct consists of functions taking values in tangent spaces along a curve on a manifold, and is able to capture the intrinsic geometry of the manifold. We prove that the estimator of the regression coefficient achieves the optimal rate of convergence in mean prediction. Moreover, we propose a method to compare objects from different tensor Hilbert spaces based on Hilbert manifolds. Potential problems caused by nonnegative sectional curvatures of manifolds are also studied. Our work is an extension of the models developed by [1, 2].

Simulation studies demonstrate the numerical advantages of our RKTHS-based approach over the function-on-function regression based on Riemannian functional PCA ([3, 4]). The proposed method is applied to tropical cyclone data to predict trajectories, and brain imaging data of preterm and full-term infants in the Developing Human Connectome Project (dHCP) to study the linear relationship between homologous white matter fiber tracts in two hemispheres of the brain.

Bibliography

[1] Cai, T. T. and Yuan, M., *Minimax and Adaptive Prediction for Functional Linear Regression*, Journal of the American Statistical Association, 107 (2012), 1201–1216.

[2] Sun, X., Du, P., Wang, X., and Ma, P., *Optimal Penalized Function-On-Function Regression Under a Reproducing Kernel Hilbert Space Framework*, Journal of the American Statistical Association, 113 (2018), 1601–1611.

[3] Dai, X. and Müller, H.-G., *Principal Component Analysis for Functional Data on Riemannian Manifolds and Spheres*, The Annals of Statistics, 46 (2018), 3334–3361.

[4] Lin, Z. and Yao, F., *Intrinsic Riemannian Functional Data Analysis*, The Annals of Statistics, 47 (2019), 3533–3577.

Mirror Descent Strikes Again: Optimal Stochastic Convex Optimization under Infinite Noise Variance

Tuesday, 4.07
16:00–18:00
Poster

Nuri Mert Vural[1], **Lu Yu**[1], **Krishnakumar Balasubramanian**[2], **Stanislav Volgushev**[1], **Murat A. Erdogdu**[1]

[1] University of Toronto
[2] University of California, Davis

Abstract

We study stochastic convex optimization under infinite noise variance. Specifically, when the stochastic gradient is unbiased and has uniformly bounded $(1 + \kappa)$-th moment, for some $\kappa \in (0, 1]$, we quantify the convergence rate of the Stochastic Mirror Descent algorithm with a particular class of uniformly convex mirror maps, in terms of the number of iterations, dimensionality and related geometric parameters of the optimization problem. Interestingly this algorithm does not require any explicit gradient clipping or normalization, which have been extensively used in several recent empirical and theoretical works. We complement our convergence results with information-theoretic lower bounds showing that no other algorithm using only stochastic first-order oracles can achieve improved rates. Our results have several interesting consequences for devising online/streaming stochastic approximation algorithms for problems arising in robust statistics and machine learning

Incorporating covariates in single-cell clustering with dependent hierarchical Dirichlet processes

Tuesday, 4.07
16:00–18:00
Poster

Huizi Zhang, Sara Wade, Natalia Bochkina

The University of Edinburgh (UK)

Abstract

Single-cell RNA-sequencing (scRNA-seq) technologies have enabled measurements of transcriptome profiles for individual cells across thousands of genes. Unsupervised clustering is one of the most important analysis of such data, allowing for discovery of rare cell types. Over the years, many algorithms have been proposed to cluster single cells from a single dataset. In a recent work (see [2]), it has been shown that the hierarchical Dirichlet processes ([3]) can be applied to cluster cells nonparametrically from different datasets, yielding shared and unique clusters across multiple groups. However, most of the existing methods still rely on the use of the gene expression data that records the total number of mRNA, whilst there may be more information hidden inside the scRNA-seq data that can be helpful for clustering.

Recently, there has been a growing interest in RNA velocity and latent time (see [1]) that describe cellular dynamics and position in the biological process being studied, both of which can be estimated by further categorizing the total mRNA into nascent and mature mRNA. In our work, we extend the model in [2], by proposing a covariate-dependent hierarchical Dirichlet process (HDP) that can include other meaningful covariates, such as latent time, to cluster subjects from different datasets. Specifically, covariates can be flexibly incorporated by combining kernel functions with HDP. For efficient inference, we construct a novel MCMC algorithm that utilizes latent variables to cope with the intractable normalized weights. We demonstrate that our model can capture the relationship between the clusters and covariates, and identify reasonable clusters across groups in both simulated and real data. We will also discuss a joint Bayesian model for clustering and RNA velocity.

Bibliography

[1] Bergen, V., Lange, M., Peidli, S., Wolf, F. A., Theis, F. J., *Generalizing RNA velocity to transient cell states through dynamical modeling*, Nature Biotechnology, 38 (2020), 1408–1414.

[2] Liu, J., Wade, S., Bochkina, N., *Shared Differential Clustering across Single-cell RNA Sequencing Datasets with the Hierarchical Dirichlet Process*, Unpublished manuscript, Department of Mathematics, University of Edinburgh, 2022.

[3] Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M., *Hierarchical Dirichlet Processes*, Journal of the American Statistical Association, 101 (2006), 1566–1581.

Testing heteroscedasticity for high dimensional asymmetric least-squares regression

Tuesday, 4.07
14:00–14:20
CS 11
Room 2180

Jing Zhou[1], **Hui Zou**[2]

[1] University of East Anglia (UK)
[2] University of Minnesota Twin Cities

Abstract

Heteroscedasticity is commonly observed in real-world data and has been understudied. We explore this issue for a typical high-dimensional regression setting where the number of predictive variables p exceeds the sample size n. Specifically, we focus on testing regression error heteroscedasticity in high dimensions and propose a test using asymmetric least-squares regression. The asymmetry of the least-squares loss function is determined by an addition parameter $\tau \in (0, 1)$, which is relevant to the proportion of the data. We assume the variance of the regression errors depends on the linear combination of the predictive variables combined by a p-vector γ_0.

The proposed test exploits the conditional distribution of the response variable and is constructed by contrasting asymmetric least-squares parameter estimations at τ_1, τ_2 under heteroscedasticity. The main focus is finding a small collection of predictive variables that are associated with heteroscedasticity by testing $H_0 : \gamma_{0,j} = 0$ versus $H_a : \gamma_{0,j} \neq 0$. We propose to use the approximate message passing (AMP) algorithm to perform an asymptotic analysis for the proposed test, assuming $n/p \to \delta \in (0, 1)$. Further, the power of the test is investigated analytically. The numerical performance of the proposed test is validated by various simulation settings. The application of the proposed test is demonstrated by the Bardet-Biedl gene expression dataset.

Speeding up Monte Carlo integration: nearest neighbor estimates as control variates

Tuesday, 4.07
11:40–12:00
CS 07
Room 2180

Rémi Leluc[1], **François Portier**[2], **Johan Segers**[3], **Aigerim Zhuman**[3]

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris (France)
[2] CREST, ENSAI (France)
[3] ISBA, LIDAM, UCLouvain (Belgium)

Abstract

The method of control variates is a powerful technique that allows to reduce the variance of the Monte Carlo estimate [1] of a multivariate integral by introducing auxiliary functions with known expectations, called control variates [2]. We propose to use nearest neighbor estimates [3] as control variates in order to speed up the convergence rate of the Monte Carlo integration procedure. Our novel estimate, called the Control Neighbor estimate, achieves the optimal convergence rate of the order $n^{-1/2}n^{-1/d}$ for Lipschitz functions, where d is the dimension of the domain of the integrand. In addition, a non-asymptotic bound on the probabilistic error of the procedure is obtained via an extension of McDiarmid's inequality for functions with bounded differences on a high probability set [4]. Moreover, several numerical experiments confirm the good performance of the proposed estimate.

Bibliography

[1] Glasserman, P., *Monte Carlo Methods in Financial Engineering*, Springer, 2003.
[2] Portier, F. and Segers, J., *Monte Carlo integration with a growing number of control variates*, Journal of Applied Probability, 56 (2019), pp. 1168–1186.
[3] Biau, G. and Devroye, L., *Lectures on the nearest neighbor method*, 246, Springer, 2015.
[4] Combes, R., *An extension of McDiarmid's inequality*, arXiv preprint arXiv:1511.05240, 2015.

Nonparametric methods for randomized controlled trials with multiple endpoints: Beyond O'Brien-Wei-Lachin

Thursday, 6.07
11:40–12:00
CS 20
Room 1.40

Guangyong Zou[1], **Lily Zou**[2]

[1] Western University (Canada)
[2] University of Waterloo (Canada)

Abstract

Randomized controlled trials are usually conducted to answer the question: Will a patient have a better outcome when treated with a new intervention than with a placebo? We can thus quantify the treatment effect using the win probability (WinP) that a randomly selected patient from treatment group has better scores than a randomly selected patient from the control group. Focusing on the WinP, Zou et al [1, 2] proposed nonparametric methods for the design and analysis of trials with a single primary endpoint.

Multiple endpoints on different scales are needed to evaluate treatment effects for complex disorders. O'Brien [3] proposed a rank-sum test to analyze studies with multiple endpoints on different scales. To perform the test, the data for each endpoint are pooled over the treatment groups and each observation is replaced by its corresponding ranking. Then, for each subject, the ranks are summed across different endpoints. The rank-sums are compared using a 2-sample t-test. We point out that the underlying parameter of this test is the mean of WinPs. Wei and Lachin [4] proposed an equivalent test, which first obtains endpoint-specific WinPs and their variance-covariance, and then conducts a test for the mean of WinPs. We refer to these two procedures as the O'Brien-Wei-Lachin (OWL) procedure, which is known to have inability to adjust for baseline measurements.

The goal of this work is to extend our regression approach [1, 2] to trials having a pretest-posttest design with multiple endpoints and missing observations. We develop and evaluate methods for confidence interval estimation and sample size planning for the mean of WinPs. We demonstrate that our procedure reduces to the OWL procedure when baseline measurements are ignored. Simulation results indicate that our methods perform very well in terms of confidence interval coverage, tail errors, and assurance to achieving the pre-specified precision. We illustrate our new methods with data arising from a randomized controlled trials which used three endpoints of different types to evaluate treatment effects. We implement our methods in SAS, Stata, and R.

Bibliography

[1] Zou G., Zou L., Choi, Y-H., *Distribution-free approach to the design and analysis of randomized stroke trials with the modified Rankin scale*, Stroke, 53 (2022), 3025–3031.

[2] Zou G., Zou L., Qiu S-F., *Parametric and nonparametric methods for confidence intervals and sample size planning for win probability in parallel-group randomized trials with Likert item and Likert scale data*, Pharmaceutical Statistics, 22 (2023) (in press).

[3] O'Brien P.C., *Procedures for comparing samples with multiple endpoints*, Biometrics, 40 (1984), 1079–1087.

[4] Wei L.J., Lachin J.M., *Two-sample asymptotically distribution-free tests for incomplete multivariate observations*, Journal of the American Statistical Association, 79 (1984), 653–661.

Index