NORDSTAT 2023 GOTHENBURG

19-22 June 2023

https://nordstat2023.org/

@nordstat2023

DOCUMENT VERSION 17 June 2023

CHALMERS | UNIVERSITY OF GOTHENBURG

Foreword

There are many people and organizations that we, the Organizing Committee, wish to thank for the setting-up of NORDSTAT 2023.

Firstly, we would like to thank all the 300 participants, and the 173 oral speakers and 43 poster presenters that enriched the conference with their work. Thank you for sharing your research with us!

We cannot but be very grateful to the Scientific Committee (see next page) chaired by Professor Mattias Villani, for the excellent work in overseeing the quality of the scientific program, for selecting themes for the invited sessions and their sessions organizers, and for the difficult but necessary work of scrutinizing and selecting the many contributed talks and posters. Communication with all of you has been a pleasure and we thank you very much for your service.

We thank the organizers of the 18 invited sessions for allowing the conference to benefit from the presence of so many excellent scholars.

We thank our conference secretariat MEETX for continuous advice and professional support before and during the conference: in particular, we thank the MEETX project responsible Maria Lindbäck, Marie Eriksson, Pärnilla Thompson and Nick Witthall.

We thank the NORDSTAT support team formed by PhD students from the Department of Mathematical Sciences at Chalmers and the University of Gothenburg (see next pages), for their indispensable help in setting up many logistics, IT supervision at the many parallel sessions, and for their enthusiasm.

We are grateful for generous funding from: AstraZeneca, Chalmers AI Research Centre (CHAIR), the Royal Society of Arts and Sciences in Gothenburg, the Swedish Statistical Society, the Wilhelm and Martina Lundgren's Science Foundation.

Thanks to the board of the Scandinavian Journal of Statistics (SJS) for sponsoring the SJS lecture and to the Swedish Statistical Society (SSS) for sponsoring the SSS Lecture.

We thank the City of Gothenburg for generously hosting the Welcome Reception.

Thanks to Springer for donating books used e.g. as prizes.

We thank the many colleagues at Chalmers and University of Gothenburg that have generously provided time and expertise: (in alphabetic order) communications officer Anneli Andersson, IT officer Ulf Andersson, administrator Marie Kühn, administrator Lisellote Fernström, administrator Jeanette Montell-Westerlin, financial officer Ai-Linh Nguyen.

We thank our colleagues Professor Peter Jagers and Professor Sergey Zuev for precious advice.

We thank AV-Service at Chalmers and especially Sebastian Pousette and Nikola Velijanoski for IT advice and support.

Finally, we thank the chair of the NORDSTAT 2021 organizing committee Prof. Sigrunn Holbek Sørbye, for generous advice whenever requested.

Sincerely,
The NORDSTAT 2023 Organizing Committee

Organizing Committee (all at Dept. Mathematical Sciences, Chalmers and the University of Gothenburg)

Umberto Picchini (committee chair)
Setta Aspström
Marina Axelson-Fisk
Ottmar Cronie
Johan Jonasson
Rebecka Jörnsten
Annika Lang
Holger Rootzen
Serik Sagitov
Moritz Schauer
Aila Särkkä

Scientific Committee

Mattias Villani (committee chair, Stockholm University)

Krista Fischer (University of Tartu)

Annika Lang (Chalmers/University of Gothenburg)

Lasse Leskelä (Aalto University)

Umberto Picchini (Chalmers/University of Gothenburg)

Valeria Vitelli (University of Oslo)

Rasmus Waagepetersen (Aalborg University)

NORDSTAT 2023 Support Team (PhD students at Dept. Mathematical Sciences, Chalmers and the University of Gothenburg)

Julia Jansson (chair of the team)	Erik Karlsson Nordling
Isac Boström	Konstantinos Konstantinou
Kasper Bågmark	Vincent Molin
Oskar Eklund	Ioanna Motschan-Armen
Erik Jansson	Selma Tabakovic
Petar Jovanovski	Johan Ulander

The NORDSTAT 2023 logo was designed by Moritz Schauer.

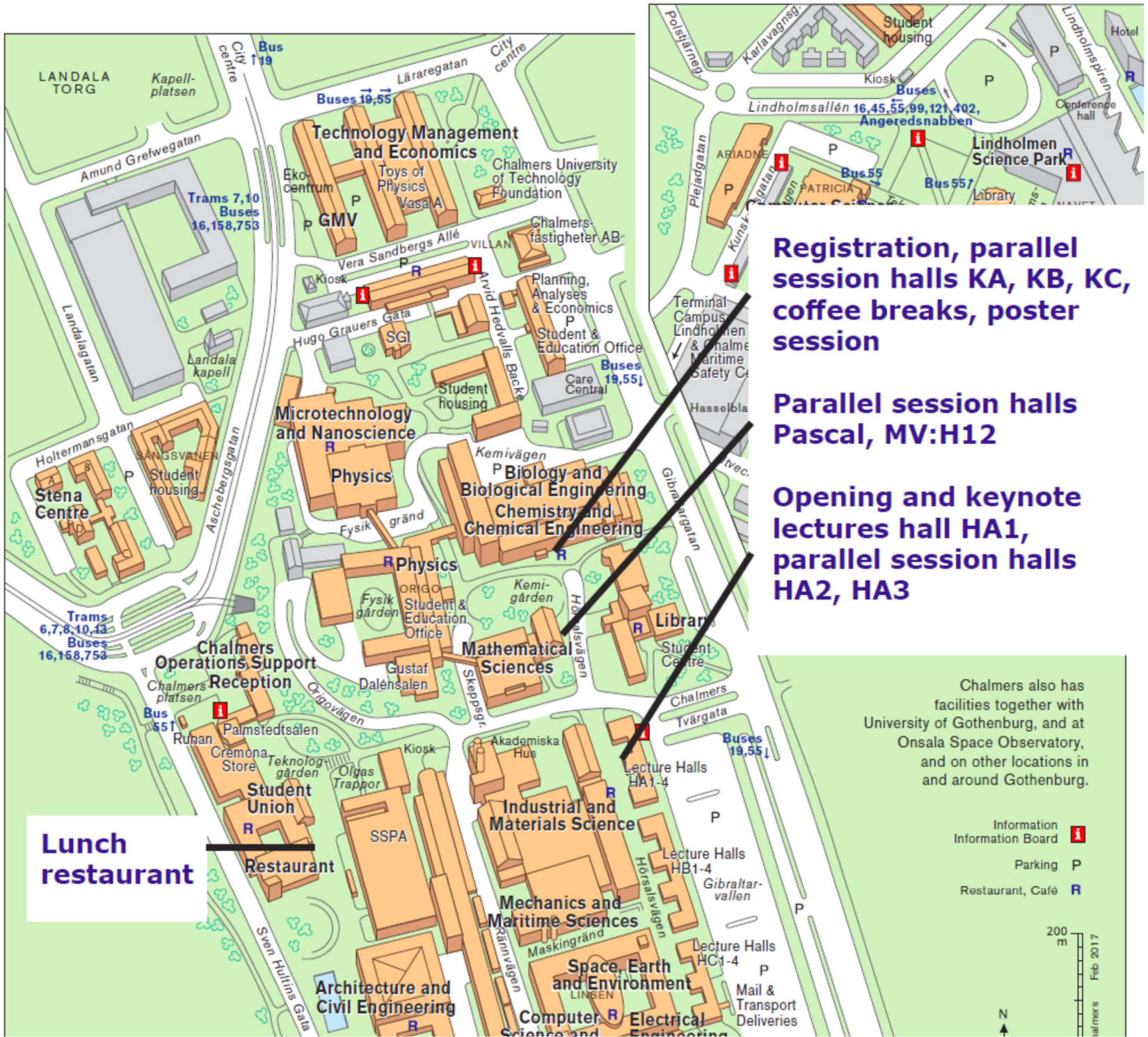
NORDSTAT 2023 acknowledges the financial support by:

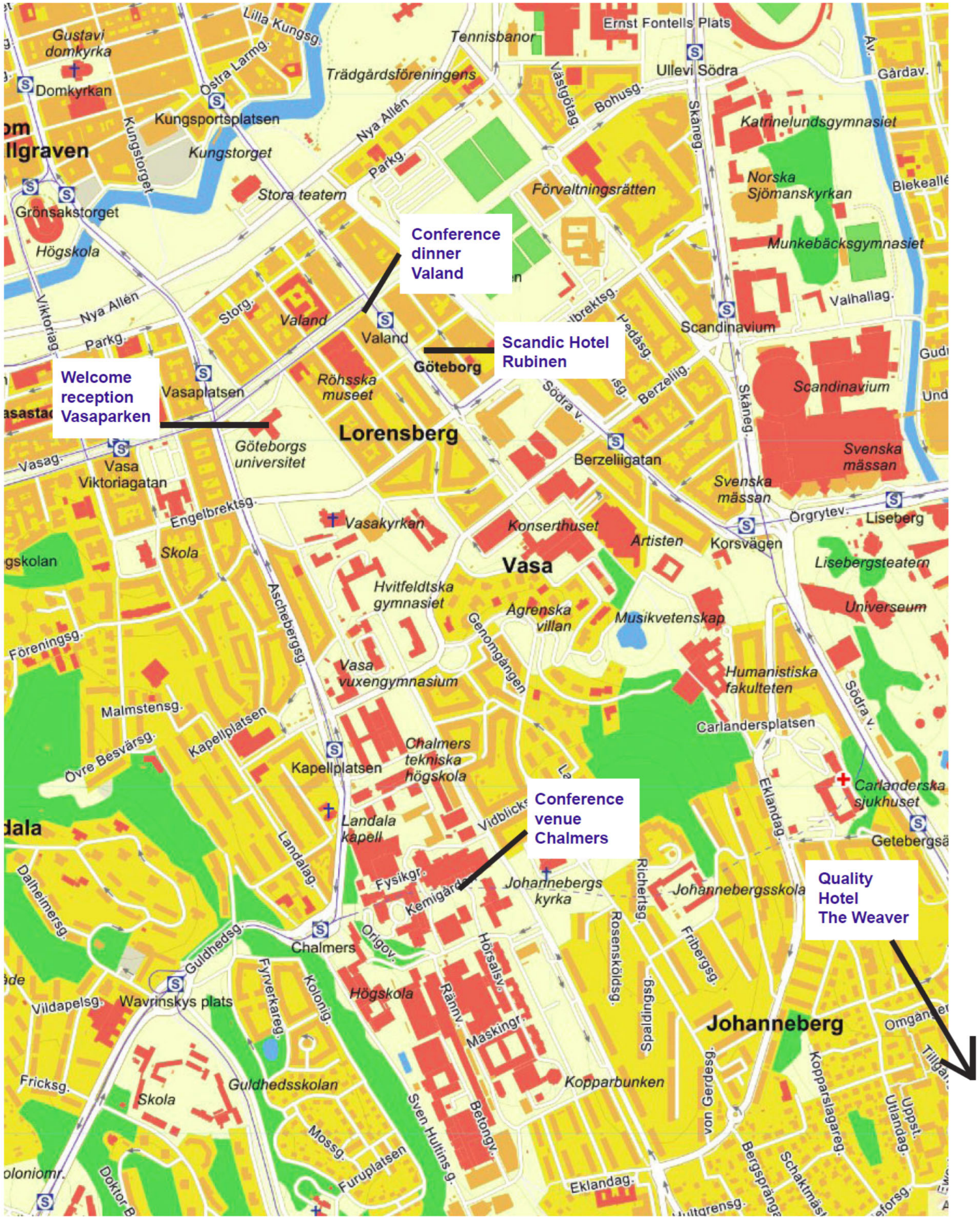
AstraZeneca
Chalmers AI Research Centre (CHAIR)
the Royal Society of Arts and Sciences in Gothenburg
the Swedish Statistical Society
Wilhelm and Martina Lundgren's Science Foundation

The Welcome Reception is hosted in collaboration with the City of Gothenburg.

Maps and addresses

Conference check-in: Kemigården 4, Chalmers Campus Johanneberg
Welcome Reception (June 19 at 19.30): Universitetsplatsen 1, University of Gothenburg Vasaparken
Conference dinner (June 21 at 19.00): Vasagatan 41, Valand Restaurang
Scandic Rubin hotel: Kungsportsavenyn 24
Quality Hotel The Weaver: Göteborgsvägen 91


Registration, parallel session halls KA, KB, KC, coffee breaks, poster session
Parallel session halls Pascal, MV:H12
Opening and keynote lectures hall HA1, parallel session halls HA2, HA3
Lunch restaurant
Chalmers also has facilities together with University of Gothenburg, and at Onsala Space Observatory, and on other locations in and around Gothenburg.
Information Information Board
Parking P
Restaurant, Café R
LANDALA TORG
Kapell-platsen
Amund Grefwegatan
Landalagatan
Holtermansgatan
Aschebergsgatan
Trams 7,10 Buses 16,158,753
Bus 19
Buses 19,55
Läraregatan
City centre
Technology Management and Economics
Eko-centrum
Toys of Physics
Vasa A
GMV
Chalmers University of Technology Foundation
Chalmers-fastigheter AB
VILLAN
Vera Sandbergs Allé
Kiosk
Arvid Hedvalls Backe
Planning, Analyses & Economics
Student & Education Office
Hugo Grauers Gata
SGI
Care Central
Buses 19,55
Landala kapell
Student housing
SÄNGSVANEN
Stena Centre
Microtechnology and Nanoscience
Physics
Kemivägen
Biology and Biological Engineering
Chemistry and Chemical Engineering
Fysik gränd
Physics
ORIGO
Student & Education Office
Fysik gården
Kemi-gården
Mathematical Sciences
Gustaf Dalénsalen
Skeppsgr.
Origovägen
Hörsalsvägen
Library
Student Centre
Chalmers Tvärgata
Trams 6,7,8,10,13 Buses 16,158,753
Chalmers Operations Support Reception
Chalmers platsen
Bus 55
Ruhan
Palmstedtsalen
Cremona Store
Teknologgården
Olgas Trappor
Kiosk
Student Union
Restaurant
SSPA
Akademiska Hus
Lecture Halls HA1-4
Industrial and Materials Science
Lecture Halls HB1-4
Gibraltar-vallen
Mechanics and Maritime Sciences
Maskingränd
Rännvägen
Lecture Halls HC1-4
Space, Earth and Environment
LINSEN
Mail & Transport Deliveries
Architecture and Civil Engineering
Sven Hultins Gata
Computer Science and
Electrical
Polstjärneg.
Karlavagnsg.
Student housing
Hotel
Lindholmspiren
Conference hall
Lindholmsallén
Buses 16,45,55,99,121,402, Angeredsnabben
Lindholmen Science Park
ARIADNE
PATRICIA
Bus 55
Library
Plejadgatan
Terminal Campus Lindholmen
Hasselbla
200 m
Feb 2017
N



Conference dinner Valand
Scandic Hotel Rubinen
Welcome reception Vasaparken
Conference venue Chalmers
Quality Hotel The Weaver
Gustavi domkyrka
Domkyrkan
Kungsportsplatsen
Kungstorget
Stora teatern
Grönsakstorget
Högskola
Trädgårdsföreningens
Tennisbanor
Ernst Fontells Plats
Ullevi Södra
Katrinelundsgymnasiet
Förvaltningsrätten
Norska Sjömanskyrkan
Munkebäcksgymnasiet
Valand
Göteborg
Scandinavium
Vasaplatsen
Röhsska museet
Lorensberg
Göteborgs universitet
Berzeliigatan
Svenska mässan
Vasa Viktoriagatan
Vasakyrkan
Konserthuset
Artisten
Korsvägen
Liseberg
Lisebergsteatern
Vasa
Hvitfeldtska gymnasiet
Ågrenska villan
Musikvetenskap
Universeum
Humanistiska fakulteten
Vasa vuxengymnasium
Carlandersplatsen
Chalmers tekniska högskola
Kapellplatsen
Carlanderska sjukhuset
Landala kapell
Johannebergs kyrka
Johannebergsskola
Chalmers
Högskola
Wavrinskys plats
Guldhedsskolan
Kopparbunken
Johanneberg
Skola

PROGRAM

Invited sessions are denoted with an “I” and contributed sessions are denoted with a “C”. Notice the full version of the program with titles and abstracts is at the conference page.

Room names are given between parentheses.

MONDAY 19 JUNE

Opening at 13.00 in HA1-HA2

Keynote by Sofia Olhede 13.30-14.30 in HA1-HA2

Parallel sessions 15.00-16.30

I1	I2	C1	C2	C3	C4
Functional data analysis (KA)	**Percolation and related topics (KC)**	**Gaussian processes (KB)**	**Finance (Pascal)**	**Bayesian modelling (HA2)**	**AI 1 (HA3)**
Alessia Pini Vaidotas Characiejus Anton Rask Lundborg	Nina Gantert Daniel Ahlberg Erik Broman	Louise Kimpton Henrik Bengtsson Sahoko Ishida	Vilhelm Niklasson Erik Lindström Alexander Herbertsson	Javier Aguilar Aliaksandr Hubin Sofia Stroustrup Ida Scheel	Florian Frommlet Olof Zetterqvist Oskar Allerbo Chiara Amorino

MONDAY 19 JUNE

Parallel sessions 17.00-18.30

I3	I4	I5	C5	C6	C7
Reinforcement learning (KA)	**Stochastic differential equations (KC)**	**Computationally intensive methods in biostatistics (KB)**	**Causality (Pascal)**	**Spatial point processes 1 (HA3)**	**Special topics in Statistics 1 (HA2)**
Alexandre Proutiere Yassir Jedra Sadegh Talebi	Mark Podolskij Thuan Nguyen Jean-François Chassagneux	Eszter Lakatos Magnus Neuman Mattias Rantalainen	Maximilian Scholz Søren Wengel Mogens Shimeng Huang Lucas Kook	Ottmar Cronie Julia Jansson Christophe Biscio Anne Marie Svane	Krzysztof Bartoszek Martin Schlather Karl Sigfrid Fredrik Lundvall Wollbraaten

Welcome Reception 19.30-21.30 in Universitetsplatsen 1 (entrance hall of the University of Gothenburg main building).
The welcome reception is kindly hosted in collaboration with the City of Gothenburg.

TUESDAY 20 JUNE

Keynote by Fred Espen Benth, 9.30-10.30 in HA1-HA2

Parallel sessions 11.00-12.30

I6	C8	C9	C10	C11	C12
Heavy tails / Extreme values (KA)	**Optimal design (KB)**	**Change point (HA2)**	**Bayesian computation (HA3)**	**Survival analysis (Pascal)**	**Probability (KC)**
Kirstin Strokorb Chen Zhou Holger Rootzen	Evangelos Evangelou Maryna Prus Henrik Imberg	Susanne Ditlevsen Per August Moen Anica Kostic Elias Erdtman	Dennis Christensen Umberto Picchini Jimmy Olsson Henri Pesonen	Thomas Matcham Céline Cunen Clara Bertinelli Salucci	Stefan Geiss Gabija Liaudanskaite Orimar Sauri

TUESDAY 20 JUNE

Parallel sessions 14.00-15.30

I7	I8	C13	C14	C15	C16
Stochastic models for evolution (KA)	**Multivariate Statistics (KC)**	**Functional data (Pascal)**	**Spatio-temporal models (HA2)**	**Parameter inference for SDEs (HA3)**	**Special topics in Statistics 2 (KB)**
Daniela Bertacchi	Krzysztof Podgorski	Mehdi Moradi	Christopher Dörr	Celia García-Pareja	Matyas Barczy
Anton Wakolbinger	Kristi Kuljus	Hiba Nassar	Malte Jahn	Predrag Pilipovic	Torkel Erhardsson
Asger Hobolth	Stepan Mazur	Johan Strandberg	Mike Pereira	Marc Vaisband	Marco Loog
	Katarzyna Filipiak		Henrik Ekström	Petar Jovanovski	Denise Uwamariya

TUESDAY 20 JUNE

Parallel sessions 16.00-17.30

I9	I10	C17	C18	C19	C20
Likelihood-free inference (KA)	**Statistical modelling of infectious disease outbreaks (KC)**	**Spatial stats & Gaussian random fields (KB)**	**AI 2 (HA2)**	**Model selection (Pascal)**	**Filtering (HA3)**
Solveig Engbretsen	Felix Günther	Erik Jansson	Lars Henry Berge Olsen	Sebastian Arnold	Alessandro Mastrototaro
Luigi Acerbi	Geir Storvik	John Paige	Olle Häggström	Ingrid Dæhlen	Matti Vihola
Dennis Prangle	Ganna Rozhnova	Johan Lindström	Claudio Meggio	Ivan Hejný	Cédric Travelletti
		Geir-Arne Fuglstad	Väinö Yrjänäinen	Nils Lid Hjort	Håkon Gryvill

Poster session 17.30-19.30 in the foyer of Kemihuset (the same building where KA-KB-KC are)

WEDNESDAY 21 JUNE

Keynote by Janine Illian (*"Scandinavian Journal of Statistics Lecture"*) 9.30-10.30 in HA1-HA2

Parallel sessions 11.00-12.30

I11	I12	C21	C22	C23	C24
Causal inference in complex data structures (KA)	**Gaussian processes on networks and graphs (KC)**	**Epidemics & Neuronal networks (HA3)**	**Time series (Pascal)**	**Stochastic partial differential equations (KB)**	**Random Graphs (HA2)**
Shuangning Li Nicola Gnecco Dominik Rothenhaeusler	Jesper Møller Joel Oskarsson Alexander Simas	Tom Britton Vasilii Goriachkin Mikhail Shubin Mohamed El Khalifi	Vytaute Pilipauskaite Chandler Ross Niklas Ahlgren Mads Stehr	Thorben Pieper Andreas Petersson Annika Lang	Joel Danielsson Tiffany Yin Yuan Lo Mindaugas Bloznelis Cecilia Holmgren

WEDNESDAY 21 JUNE

Parallel sessions 14.00-15.30

I13	I14	I15	C25	C26	C27
Spatial point processes – modelling, estimation, and prediction (KB)	**Random graphs and related topics (KA)**	**Stochastic Finance (KC)**	**Local methods for functional data (HA3)**	**Computational statistics (HA2)**	**Regression and applications (Pascal)**
Jakob Gulddahl Rasmussen Thordis Thorarinsdottir Tuomas Rajala	Svante Janson Matas Šileikis Tatyana Turova	Sigrid Källblad Giulia di Nunno Teemu Pennanen	Helle Sørensen Simone Vantini Valeria Vitelli Niels Olsen	Petter Mostad Alexander Freudenberg Max Raner Miika Kailas	Léna Gozé Szymon Urbas Ingrid Hobæk Haff Vladimir Pastukhov

15.30-19.00 Free time (see the conference page for suggestions)

19.00-00.00 Conference dinner in Valand restaurant (Vasagatan 41), for those that are registered for the dinner.

THURSDAY 22 JUNE

Keynote by Jonas Peters 9.30-10.30 in HA1-HA2

Parallel sessions 11.00-12.30

I16	I17	I18	C28	C29	C30
Learning for stochastic differential equation models (KB) Frank van der Meulen Stefan Sommer Frank Schäfer	**Probabilistic machine learning and AI (KA)** Fredrik Lindsten Judith Bütepage Pierre-Alexandre Mattei	**Random matrices (KC)** Taras Bodnar Alexey Onatskiy Antoine Maillard	**Spatial point processes 2 (HA2)** Rasmus Waagepetersen Magnus Ekström Konstantinos Konstantinou Diala Hawat	**Latent variable modelling for the analysis of preference data (Pascal)** Matteo Ventura Ambra Macis Francesco Amato Antonio D'Ambrosio	**Special topics in Statistics 3 (HA3)** Gabriel Wallin Per Gösta Andersson Janis Valeinis Christos Dimitrakakis

Keynote by Peter Diggle (*"Swedish Statistical Society Lecture"*) 14.00-15.15 in HA1-HA2

Poster Prizes and Closing 15.15 in HA1-HA2

POSTER SESSION ON 20 JUNE at 17:30: LIST OF AUTHORS

(the full list with titles and abstracts is at the conference page)

Reinis Alksnis
Kasper Bågmark
Elizabeth Baker
Sam Beaton
Reza Belaghi
Bayu Brahmantio
Timo Braun
Marc Corstanje
Johan de Aguas
Mara Delesa-Velina
Svitlana Drin
Oskar Eklund
Jonas Evaeus
Henrik Häggström
Henrik Imberg
Alexandra Jauhiainen
Natoya Jourdain
Hao Chi Kiang
Konstantinos Konstantinou
Vera Kvisgaard
Rolf Larsson
Klara Leffler
Marie Levakova
Karina Lilleborge
Gustav Lindwall
Alexander Mey
Vincent Molin
Philip Mostert

Ioanna Motschan
Helga Kristin Olafsdottir
Hannu Reittu
Axel Ringh
Nataliya Shchestyuk
Andrey Shternshis
Dana Sylvan
Selma Tabakovic
Johan Ulander
Emelyne Umunoza Gasana
Per Westerlund
Väinö Yrjänäinen
Shokoufa Zeinali

Efficient Bayesian inference with noisy, black-box likelihoods via Variational Bayesian Monte Carlo

Luigi Acerbi[1]

[1] *University of Helsinki, Finland, luigi.acerbi@helsinki.fi*

This talk presents an integrated overview of the recent advances in Bayesian inference techniques for dealing with expensive, black-box likelihoods and noisy log-likelihood evaluations, focusing on the Variational Bayesian Monte Carlo (VBMC) framework. VBMC is a sample-efficient inference method that combines variational inference with Gaussian-process based, active-sampling Bayesian quadrature [1]. VBMC handles well noisy log-likelihood evaluations arising from simulation-based models, thanks to robust acquisition functions [2]. Noisy estimators of the log-likelihood include, among others, the *synthetic likelihood* and *inverse binomial sampling* – an efficient, unbiased estimator of the log-likelihood recently proposed for use in computational neuroscience [3]. Finally, we present Sparse Variational Bayesian Monte Carlo (SVBMC) [4], a *post-process* Bayesian inference method that leverages sparse Gaussian processes and active learning to quickly build a posterior approximation out of preliminary runs of other algorithms (e.g., MAP optimization, exploratory MCMC runs). The talk discusses the effectiveness and applicability of these techniques, highlighting their potential as novel tools for Bayesian inference with black-box, noisy likelihoods in various scientific domains.

Software: VBMC and related algorithms are available as open-source, well-documented software packages in both MATLAB and Python [5]; please see: `https://github.com/acerbilab`

References

[1] Acerbi, L., Variational Bayesian Monte Carlo, *NeurIPS* 31, 2018.

[2] Acerbi, L., Variational Bayesian Monte Carlo with Noisy Likelihoods, *NeurIPS* 33, 2020.

[3] van Opheusden, B., Acerbi, L., & Ma, W. J., Unbiased and Efficient Log-Likelihood Estimation with Inverse Binomial Sampling, *PLoS Computational Biology* 16(12):e1008483, 2020.

[4] Li, C., Clarté, G., & Acerbi, L., Fast post-process Bayesian inference with Sparse Variational Bayesian Monte Carlo, *arXiv preprint*: 2303.05263, 2023.

[5] Huggins, B., Li, C., Tobaben, M., Aarnos, M. J., & Acerbi, L., PyVBMC: Efficient Bayesian inference in Python, *arXiv preprint*: 2303.09519, 2023; `https://acerbilab.github.io/pyvbmc/`

Intuitive Joint Priors for Bayesian Multilevel Models: The R2D2M2 prior

Javier Aguilar[1] and Paul Bürkner [2]

[1] *University of Stuttgart, Cluster of Excellence SimTech*
javier.aguilar-romero@simtech.uni-stuttgart.de
[2] *University of Stuttgart, Cluster of Excellence SimTech*
paul-christian.buerkner@simtech.uni-stuttgart.de

The training of high-dimensional regression models on comparably sparse data is an important yet complicated topic, especially when there are many more model parameters than observations in the data. From a Bayesian perspective, inference in such cases can be achieved with the help of shrinkage prior distributions, at least for generalized linear models. However, real-world data usually possess multilevel structures, such as repeated measurements or natural groupings of individuals, which existing shrinkage priors are not built to deal with.

We generalize and extend one of these priors, the R2D2 prior by Zhang [1], to linear multilevel models leading to what we call the R2D2M2 prior. The proposed prior enables both local and global shrinkage of the model parameters. It comes with interpretable hyperparameters, which we show to be intrinsically related to vital properties of the prior, such as rates of concentration around the origin, tail behavior, and amount of shrinkage the prior exerts.

We offer guidelines on how to select the prior's hyperparameters by deriving shrinkage factors and measuring the effective number of non-zero model coefficients. Hence, the user can readily evaluate and interpret the amount of shrinkage implied by a specific choice of hyperparameters.

Finally, we perform extensive experiments on simulated and real data, showing that our inference procedure for the prior is well calibrated, has desirable global and local regularization properties and enables the reliable and interpretable estimation of much more complex Bayesian multilevel models than was previously possible.

References

[1] Zhang, Y. D., Naughton, B.P, Bondell, H. D, Bayesian Regression Using a Prior on the Model Fit: The R2-D2 Shrinkage Prior, Journal of the American Statistical Association, Taylor & Francis.

Chaos and concentration in spatial growth models

Daniel Ahlberg[1], Mia Deijfen[1] and Matteo Sfragara[1]

[1] *Department of Mathematics, Stockholm University*

The occurrence of chaotic phenomena in models of disordered systems such as spin glasses was predicted by physicists in the 1980s. A rigorous connection between anomalous fluctuations and a chaotic behaviour of the ground state in certain Gaussian disordered systems was established by Chatterjee about a decade and a half ago. The purpose of this paper is to show that Chatterjee's work gives evidence of a more general principle, by establishing an analogous connection between fluctuations and chaos in the context of first-passage percolation. More precisely, we resample a small proportion of the edge weights, and find that a vanishing fraction of the edges on the distance-minimising path still belongs to the distance-minimising path obtained after resampling.

References

[1] Ahlberg, D., Deijfen, M. and Sfragara, M., Chaos, concentration and multiple valleys in first-passage percolation. Preprint.

[2] Ahlberg, D., Deijfen, M. and Sfragara, M., From stability to chaos in last-passage percolation. Preprint.

A new GARCH model with a deterministic time-varying intercept

Alexander Back[1], Niklas Ahlgren[1] and Timo Teräsvirta[2]

[1] *Hanken School of Economics, alexander.back@hanken.fi, niklas.ahlgren@hanken.fi*
[2] *Aarhus University, tterasvirta@econ.au.dk*

We show that a popular multiplicative decomposition of volatility has an interpretation as a GARCH model augmented by a time-varying intercept. We parameterize the intercept by a logistic transition function with rescaled time as the transition variable, which provides a flexible and simple way of capturing deterministic non-linear changes in the conditional and unconditional variances. It is common for financial time series to exhibit these types of shifts. The time-varying intercept makes the model globally nonstationary but locally stationary. We use the theory of locally stationary processes to derive the asymptotic properties of the quasi maximum likelihood estimator (QMLE) of the parameters of the model. We show that the QMLE is consistent and asymptotically normally distributed. To corroborate the results of the analysis, we provide a small simulation study. An empirical application on Oracle Corporation returns demonstrates the usefulness of the model. We find that the persistence implied by the workhorse GARCH(1,1) parameter estimates is reduced by incorporating a time-varying intercept.

Bartlett corrections for two-sample Empirical Likelihood quantile inference

Reinis Alksnis and Jānis Valeinis

University of Latvia, Latvia, reinis.alksnis@lu.lv, janis.valeinis@lu.lv

One of the main advantages of the empirical likelihood method over other nonparametric methods is its ability to be Bartlett-corrected. In a seminal paper by DiCiccio et al. [1] it was shown that Bartlett corrections can be applied to empirical likelihood in a single sample setting with independent observations, when the parameter of interest can be expressed as a smooth function of means. Chen and Hall [2] developed Bartlett corrections for smoothed empirical likelihood dealing with quantile inference. Since then this theory has been developed further in many directions, however, little attention has been devoted to two-sample problems. Only difference of two means has been addressed so far ([3], [4]). Furthermore, although the possibility of Bartlett's corrections is interesting in itself from a theoretical point of view, as it relates empirical likelihood to its parametric counterpart, it is not always clear whether it can also provide a significant practical improvement. Our objective is to develop Bartlett's corrections for two-sample problems concerning quantile differences. Various measures in the field of finance as well as typical biomarkers in medicine are defined through the respective quantiles, hence it would also be of practical significance to develop a better performing small sample test.

References

[1] DiCiccio, Thomas, Peter Hall, and Joseph Romano. "Empirical likelihood is Bartlett-correctable." the Annals of Statistics (1991): 1053-1061.

[2] Chen, Song Xi, and Peter Hall. "Smoothed empirical likelihood confidence intervals for quantiles." The Annals of Statistics (1993): 1166-1181.

[3] Liu, Yukun, Changliang Zou, and Runchu Zhang. "Empirical likelihood for the two-sample mean problem." Statistics Probability Letters 78, no. 5 (2008): 548-556.

[4] Liu, Yukun, and Chi Wai Yu. "Bartlett correctable two-sample adjusted empirical likelihood." Journal of Multivariate Analysis 101, no. 7 (2010): 1701-1711.

Solving Kernel Ridge Regression with Gradient Descent

Oskar Allerbo[1], Rebecka Jörnsten[2]

[1] *Chalmers/University of Gothenburg, allerbo@chalmers.se*
[2] *Chalmers/University of Gothenburg, jornsten@chalmers.se*

We present a novel, equivalent formulation for the objective function of kernel ridge regression (KRR), that opens up for studying KRR from the perspective of gradient descent. Utilizing gradient descent with infinitesimal step size, allows us to formulate a new regularization for kernel regression through early stopping.
The new formulation of KRR also enables us to explore other penalties than the ridge penalty. Specifically, we explore the ℓ_1 and ℓ_∞ penalties and show that these correspond to two flavours of gradient descent, thus alleviating the need of computationally heavy proximal gradient descent algorithms. We show theoretically and empirically how these formulations correspond to signal-driven and robust regression, respectively.
The gradient descent formulation of KRR allows us expand to a time dependent stationary kernel, where we decrease the bandwidth to zero during training. This circumvents the need of hyper parameter selection. Furthermore, we are able to achieve both zero training error and a double descent behaviour, phenomena that do not occur for KRR with constant bandwidth, but are known to appear for neural networks, thus showcasing how the reformulated KRR problem provides new insights.

Clustering Longitudinal Ordinal Data

Francesco Amato[1], Julien Jacques[1]

[1] Univ Lyon, Univ Lyon 2, ERIC, Lyon.
{francesco.amato, julien.jacques}@univ-lyon2.fr

In social sciences, studies are often based on questionnaires asking participants to express ordered responses several times over a study period. We present a model-based clustering algorithm for such longitudinal data. Assuming that an ordinal variable is the discretization of a underlying latent continuous variable, the model relies on a mixture of matrix-variate normal distributions [1], accounting simultaneously for within- and between-time dependence structures. An EM algorithm is considered for parameter estimation. An evaluation of the model through synthetic data show its estimation abilities and its advantages when compared to competitors. A real-world application concerning preferences for grocery shopping during the Covid-19 pandemic period in France will be presented.

References

[1] Cinzia Viroli. "Finite mixtures of matrix normal distributions for classifying three-way data". In: Statistics and Computing 21.4 (2011), pp. 511–522.

Minimax optimal convergence rates under componentwise local differential privacy

Chiara Amorino[1], and Arnaud Gloter [2]

[1] *Université du Luxembourg, L-4364 Esch-Sur-Alzette, Luxembourg, chiara.amorino@uni.lu*
[2] *Laboratoire de Mathématiques et Modélisation d'Evry, CNRS, Univ Evry, Université Paris-Saclay, 91037, Evry, France, arnaud.gloter@univ-evry.fr*

A major challenge in statistical inference consists in balancing statistical utility and the privacy of individuals from whom data is obtained.
In our analysis, we examine data from n individuals that provides information about d different aspects of their lives, where each component is made public separately. This approach may be advantageous because different features of the same individual may require varying levels of privacy.
We propose a formal framework for analyzing the tradeoff between statistical utility and local differential privacy, based on the classical minimax theory. Our main objective is to understand how the optimal rate of estimation changes as the privacy levels for the different components are varied, for different types of estimation problems.
We develop several general techniques for deriving minimax bounds under componentwise local differential privacy constraints, which enable us to determine the minimum amount of information that can be extracted from the data while preserving a certain level of privacy.
By presenting concrete examples and statistical applications of such bounds, we aim to showcase their versatility and practicality. In particular, we will detail the estimation of the covariance and the estimation of the density, in a locally private and multivariate context.

Confidence interval construction for a binomial p: a trivial task? Hardly!

Per Gösta Andersson[1]

[1] *Stockholm University, Sweden, per.gosta.andersson@stat.su.se*

Over the years there have been many suggestions how to construct a confidence interval for p in the binomial distribution, based on a sequence of independent Bernoulli trials, each with probability p of success. Here we will have a look at three intervals, which are all constructed from the assumption of approximate normality of the point estimator: the relative frequency of successes.

The first is the Wald interval, which for too long hs been accepted as being useful and has become a ”standard”, much due to its simplicity. The Wilson (score) interval is actually not much more complicated, but has substantially improved coverage properties. Lastly, the AN (Andersson-Nerman) interval, based on a general principle for adjusting a pivot, is quite closely related to the Wilson interval, but differs systematically from the latter in terms of both increased coverage and expected length.

There is a vast amount of literature on this subject and the most relevant articles for his presentation are [4] and [5], which have provided much of the inspiration to [1], [2] and [3].

References

[1] Andersson, P.G., Approximate confidence intervals for a binomial p — once again. *Statistical Science* 37(4), (2022) 598—606.

[2] Andersson, P.G., The Wald confidence interval for a binomial p as an illuminating ”bad” example. Published online 21 March, 2023 under ”Latest articles” in *The American Statistician*.

[3] Andersson, P.G., A comparison of the Wilson and AN confidence intervals for the binomial p. (2023) *Submitted manuscript*.

[4] Brown, L.D., Cai, T.T. and DasGupta, A., Interval estimation for a binomial proportion. *Statistical Science* 16, (2001) 101—133.

[5] Brown, L.D., Cai, T.T. and DasGupta, A., Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics* 30, (2002) 160—210.

Sequential model confidence sets

Sebastian Arnold[1] and Johanna Ziegel[1]

[1] *Institute of Mathematical Statistics and Actuarial Science, University of Bern, sebastian.arnold@unibe.ch, johanna.ziegel@unibe.ch*

In most forecasting situations, we are given a whole set of different and possibly competing models. Naturally we want to select the best models amongst all available ones, where *best* is understood in terms of appropriate loss functions. The *model confidence set* (MCS) algorithm by Hansen et. al [1] provides a powerful solution to this problem. However the MCS algorithm only allows for inference over an evaluation period that is fixed in advance. We adapt and extend the MCS algorithm since forecasting and forecast evaluation are inherently sequential tasks: We collect and accumulate data sequentially over time and want to draw inference on a regular basis, as e.g. a weather prediction institution that wants to decide which models have performed best by the end of each year.
We provide a sequential version of the MCS algorithm that allows to compare and select the models sequentially over time incorporating the possibility of time-varying performances of the models. Our approach is based on *e-processes* which allow for safe anytime-valid inference and have recently found great attention in the statistical literature, see [2] for a recent summary.

References

[1] Hansen, P. R, Lunde, A. and Nason, J. M.,The model confidence set. *Econometrica*, 79:453–497, 2011.

[2] Ramdas, A., Grünwald, P., Vovk, V. and Shafer, G. Game-theoretic statistics and safe anytime-valid inference. *Preprint, arXiv: 2210.01948*, 2022.

An energy-based deep splitting method for the nonlinear filtering problem

Kasper Bågmark[1], Adam Andersson[1,2] and Stig Larsson[1]

[1] *Chalmers University of Technology and University of Gothenburg, Gothenburg, bagmark@chalmers.se, stig@chalmers.se*

[2] *Saab AB Radar Solutions, Gothenburg, adam.andersson@chalmers.se*

The problem of estimating the probability density of a continuous state given noisy measurements is called the filtering problem. In the case when the system of states and observations is nonlinear the problem cannot be solved analytically (except in a few special cases). Classical methods, namely particle filters, suffer under the curse of dimensionality in the underlying dimension of the state space. Deep learning is a powerful tool in creating scalable approximations for similar problems. The proposed method combines a deep splitting method, previously used for PDEs and SPDEs [1, 2], with an energy-based approach [4], in order to approximate the solution to the Zakai equation. This is a linear SPDE, whose solution is in fact an unnormalized filtering density. This results in a computationally fast filter that takes observations as input and that does not require re-training when new observations are received [3]. The method is tested on four examples, two linear in one and twenty dimensions and two nonlinear in one dimension. The method shows promising performance when benchmarked against the Kalman filter and the bootstrap particle filter.

References

[1] Beck, C., Becker, S., Cheridito, P., Jentzen, A., & Neufeld, A. (2021). Deep splitting method for parabolic PDEs. *SIAM Journal on Scientific Computing, 43*(5), A3135-A3154.

[2] Beck, C., Becker, S., Cheridito, P., Jentzen, A., & Neufeld, A. (2020). Deep learning based numerical approximation algorithms for stochastic partial differential equations and high-dimensional nonlinear filtering problems. arXiv:2012.01194.

[3] Bågmark, K., Andersson, A., & Larsson, S. (2022). An energy-based deep splitting method for the nonlinear filtering problem. arXiv:2203.17153.

[4] Gustafsson, F. K., Danelljan, M., Bhat, G., & Schön, T. B. (2020). Energy-based models for deep probabilistic regression. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16* (pp. 325-343). Springer International Publishing.

Gaussian Processes in Shape Spaces

Elizabeth Baker[1], Thomas Besnier[2] and Stefan Sommer[1]

[1] *Department of Computer Science (DIKU), University of Copenhagen, Denmark, elba@di.ku.dk, sommer@di.ku.dk*
[2] *Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France, thomas.besnier@centrale.centralelille.fr*

Modelling morphological changes in evolutionary biology requires stochastic models for shape evolutions. By conditioning on observations, the stochastic models allows inference of phylogenetic trees based on morphological data, as opposed to traditional methods in biology that rely on hand-picked traits.

Previous work has focused on defining bridges between shapes using landmarks or point approximations of the shapes [1]. The process of marking important landmarks on shapes is time-consuming, has potential for bias, and can lead to loss of important information present in the full shape. To overcome these limitations, we propose a new method that defines bridges on whole shapes using spatiotemporal Gaussian processes.

We treat shapes as functions in a Sobolev space between an underlying manifold and Euclidean space. Using a Sobolev space provides regularity of the shapes and allows for easier implementations. In this space, we define a spatiotemporal Gaussian process with a separable covariance kernel. We then condition the resulting Gaussian process, facilitating bridges between whole shapes.

Separable Gaussian processes allow for efficient implementations and are easy to sample from. We believe this method will lead to new inference methods on morphological data.

References

[1] Alexis Arnaudon, Frank van der Meulen, Moritz Schauer, and Stefan Sommer. Diffusion bridges for stochastic Hamiltonian systems and shape evolutions. *SIAM Journal on Imaging Sciences*, 15(1):293–323, March 2022. arXiv:2002.00885 [physics].

Existence and uniqueness of weighted generalized ψ-estimators

Mátyás Barczy[1] and Zsolt Páles[2]

[1] *University of Szeged, Szeged, barczy@math.u-szeged.hu*
[2] *University of Debrecen, Debrecen, pales@science.unideb.hu*

We introduce the notions of generalized and weighted generalized ψ-estimators as unique points of sign change of some appropriate functions, and we give necessary as well as sufficient conditions for their existence. We also derive a set of sufficient conditions under which the so-called ψ-expectation function has a unique point of sign change. We present several examples from statistical estimation theory, where our results are well-applicable. For example, we consider the cases of empirical quantiles, empirical expectiles, some ψ-estimators that are important in robust statistics, and some examples from maximum likelihood theory as well. Further, we introduce Bajraktarević-type (in particular, quasi-arithmetic-type) ψ-estimators. Our results specialized to ψ-estimators with a function ψ being continuous in its second variable provide new results for (usual) ψ-estimators (also called Z-estimators). The talk is based on our paper [1].

Mátyás Barczy has been supported by the project TKP2021-NVA-09. Project no. TKP2021-NVA-09 has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme.

References

[1] Barczy, M., Páles Zs., Existence and uniqueness of weighted generalized ψ-estimators, *ArXiv* 2211.06026.

Centred kernel quadratic stochastic operators: asymptotics and simulations

Krzysztof Bartoszek[1] and Małgorzata Pułka[2]

[1] *Linköping University, Linköping, krzysztof.bartoszek@liu.se; krzbar@protonmail.ch*
[2] *Gdańsk University of Technology, Gdańsk, malpulka@pg.edu.pl*

We consider a subclass of *centred kernel quadratic stochastic operators* (CKQSO) acting on probability density functions, i.e., operators of the form

$$\mathbf{Q}(f_1, f_2)(z) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f_1(x) f_2(y) g(z - \frac{x+y}{2}) \mathrm{d}x \mathrm{d}y.$$

We assume that the random variables, associated with f_1, f_2, and g have finite variance, and furthermore the one associated with g has mean 0. Inside this class the random variables associated with the CKQSO converge almost surely and in L^2, while previously in the more general case, we only had weak convergence results [1]. We have exact formulæ for the Wasserstein distance between iterates of the CKQSO, and their strong limit. We will discuss the case of a Gaussian kernel in detail and present simulation algorithms.

References

[1] Bartoszek, K., Domsta J. and Pułka M., Weak Stability of Centred Quadratic Stochastic Operators, *Bulletin of the Malaysian Mathematical Sciences Society*, 42(4), (2019) 1813-1830.

Using functional principal component analysis to characterize infant neurodevelopment in rural Gambia.

Sam Beaton[1], Sam McCann[1], Anna Blasi[2] and Sophie Moore[1]

[1] *King's College London, London, UK; Medical Research Council Unit, The Gambia at the London School of Hygiene and Tropical Medicine, The Gambia; samuel.beaton@kcl.ac.uk, samantha.mccann@kcl.ac.uk, sophie.moore@kcl.ac.uk*

[2] *University College London, London, UK; a.blasi@ucl.ac.uk*

The period from conception to two years of age is particularly crucial for neurodevelopment [1], emphasizing the importance of robust, sensitive assessment methods. Habituation and novelty detection provide foundations for fundamental aspects of learning; methods that accurately measure these processes may give valuable insight into brain development [2]. A recent study in rural Gambia (the BRIGHT Study) [3] employed functional near-infrared spectroscopy [4] in a longitudinal cohort of infants aged 1- to 24 months, to measure habituation and novelty detection objectively [5].

We propose the use of functional principal component analysis (fPCA) [6] to capture latent sources of variation in the set of response curves. We subsequently propose to use the change in functional principal component scores to characterize the change in cortical activation, corresponding to infant habituation and novelty detection. Using examples derived from the BRIGHT study, we will present data highlighting the context, challenges, preliminary findings and future research avenues: extending the work to high-density, tomographic imaging and the investigation of contributory factors in brain development. Such analyses may permit discrimination between different patterns of habituation and novelty detection, serving as early markers of neurodevelopment.

References

[1] Mattei D, Pietrobelli A. Micronutrients and Brain Development. Curr Nutr Rep. 2019 Jun 1;8(2):99-107.

[2] Sicard-Cras I, Rioualen S, Pellae E, Misery L, Sizun J, Roué J. A review of the characteristics, mechanisms and clinical significance of habituation in foetuses and newborn infants. Acta Paediatrica. 2022 Feb;111(2):245–58.

[3] Katus L, Hayes N, McCann S, Mason L, Blasi A, Darboe MK, et al. Implementing neuroimaging and eye tracking methods to assess neurocognitive development of young infants in low-and middle-income countries. Gates Open Research. 2019;3.

[4] Lloyd-Fox S, Blasi A, Elwell CE. Illuminating the developing brain: The past, present and future of functional near infrared spectroscopy. Neuroscience and Biobehavioral Reviews. 2010 Feb;34(3):269-84.

[5] Lloyd-Fox S, Blasi A, McCann S, Rozhko M, Katus L, Mason L, et al. Habituation and novelty detection fNIRS brain responses in 5- and 8-month-old infants: The Gambia and UK. Developmental Science. 2019 Mar 13;e12817.

[6] Hyndman RJ, Ullah S. Robust forecasting of mortality and fertility rates: A functional data approach. Computational Statistics and Data Analysis. 2007 Jun;51(10):4942-56.

discFA package R package: An introduction to discrete factor analysis

Reza Belaghi[1], Rolf Larsson[1] Yasin Asar[2]

[1] *Uppsala University, Uppsala, reza.belaghi@math.uu.se*
[2] *Necmettin Erbakan University, yasinasar@hotmail.com*

Literature suggested that using the traditional factor analysis for the count data might generate incorrect results [1]. With that in mind, discrete factor analysis builds on fitting systems of dependent discrete random variables to data. The data should be in the form of non-negative counts. They data may be truncated at some positive integer value. The discFA package in R allows for two distributions: Poisson and Negative Binomial, in combination with possible zero inflation and truncation, hence all in all, eight different alternatives. A forward search algorithm is employed to find the model with the lowest AIC [2]. In this paper, we will introduce the discFA package. Then, we present several different real examples from psychology, agriculture, car industry, and simulated data to compare the results provided by discFA and the traditional data analysis

References

[1] Basto M, Pereira JM (2012). "An SPSS R-menu for ordinal factor analysis." Journal of statistical software, 46, 1–29.

[2] Larsson R (2020). "Discrete factor analysis using a dependent Poisson model." Computational Statistics, 35(3), 1133–1152.

Excursion probability approximation using Clipped Slepian process

Henrik Bengtsson[1] and Krzysztof Podgórski[2]

[1] *Department of Statistics, Lund University, henrik.bengtsson@stat.lu.se*
[2] *Department of Statistics, Lund University, krzysztof.podgorski@stat.lu.se*

Excursion times for stochastic processes is a long-standing research topic in probability theory and its importance in fields such as risk analysis and statistical physics or extreme value theory is well-known. Due to the difficulty of obtaining analytical solutions, approximation methods are commonly used. The independent interval approximation method (IIA) has gained popularity in physics. However, this approximation is typically only used to approximate the tail behavior of the excursion time distribution by estimating the so-called persistency coefficient.

A modified version of the IIA method has been developed that is based on the Slepian process [1]. Recent results [2] provide the theoretical foundations for this new approximation method for a large class of covariance functions. For this class, there exists an explicit stochastic representation of the excursion time distribution which is useful for efficient simulations and numerical methods of evaluating persistency. The Slepian based method also provides better approximations for some processes, compared to the ordinary IIA. This is illustrated with the random acceleration process, which is one of the very few processes with the analytically known persistency exponent.

References

[1] Bengtsson H and Podgórski K, The Slepian model and independent interval approximation of persistency and exceedance distributions. 2023. Manuscript in preparation.

[2] Bengtsson, H. Characteristics of the switch process and geometric divisibility, 2022. URL: https://arxiv.org/abs/2205.06103.

Stochastic modelling of forward prices in commodity and energy markets with applications

Fred Espen Benth[1], Nils Detering[2], Luca Galimberti[3], Dennis Schroers[4], Carlo Sgarra[5] and Almut Veraart[6]

[1] *University of Oslo, fredb@math.uio.no*
[2] *University of California at Santa Barbara, detering@uscb.edu*
[3] *Norwegian University of Science and Technology, luca.galimberti@ntnu.no*
[4] *University of Bonn, dschroer@uni-bonn.de*
[5] *Politecnico di Milano, carlo.sgarra@polimi.it*
[6] *Imperial College London, a.veraart@imperial.ac.uk*

In this talk we survey some recent advances on stochastic modelling of forward prices in commodity and energy markets. Forward prices in these markets can be represented as space-time random fields, where the space variable is "time-to-maturity" of the forward contracts. Their dynamics satisfies a class of stochastic partial differential equations, i.e., the so-called term structure dynamics (or forward curves) is a stochastic process taking values in a function space (typically being a Hilbert space). We present infinite-dimensional stochastic volatility models as operator-valued processes including the possibility of leverage ([5]), and discuss the question of option pricing in this context. Neural networks in Hilbert spaces provide an attractive numerical method to price options on forward curves, where stylised shapes of the curves are used as additional information in the training ([1, 2]). Finally, we present limit theorems on the realised variation of forward curves, which can be used for estimation ([3, 4]).

References

[1] Benth, F. E., Detering, N., and Galimberti, L., Neural networks in Fréchet spaces, *Annals of Mathematics and Artificial Intelligence*, 91, (2023) 75-103.

[2] Benth, F. E., Detering, N., and Galimberti, Pricing of options on flow forwards by neural networks in Hilbert space, *arxiv:2202.11606*.

[3] Benth, F. E., Schroers, D. and Veraart, A., A weak law of large numbers for realised covariation in a Hilbert space setting, *Stochastic Processes and their Applications*, 145, 241-268, 2022.

[4] Benth, F. E., Schroers, D. and Veraart, A., A feasible central limit theorem for realised covariation of SPDEs in the context of functional data, *arxiv:2205.03927*.

[5] Benth, F. E., and Sgarra, C., A Barndorff-Nielsen and Shephard model with leverage in Hilbert space for commodity forwards, *ssrn: 3835053*.

An evolution model with uncountably many alleles

Daniela Bertacchi[1], Jüri Lember[2] and Fabio Zucca[3]

[1] *University of Milano-Bicocca, Italy, daniela.bertacchi@unimib.it*
[2] *University of Tartu, Estonia, juri.lember@ut.ee*
[3] *Politecnico di Milano, Italy, fabio.zucca@polimi.it*

We study a class of evolution models, where the breeding process involves an arbitrary exchangeable process, allowing for mutations to appear. The population size n is fixed, hence after breeding, selection is applied. Individuals are characterized by their genome, picked inside a set $\mathcal{X}$ (which may be uncountable), and there is a fitness associated to each genome. Being less fit implies a higher chance of being discarded in the selection process. The stationary distribution of the process can be described and studied. We are interested in the asymptotic behavior of this stationary distribution as n goes to infinity. Choosing a parameter $\lambda > 0$ to tune the scaling of the fitness when n grows, we prove limiting theorems both for the case when the breeding process does not depend on n, and for the case when it is given by a Dirichlet process prior. In both cases, the limit exhibits phase transitions depending on the parameter λ.

Time-to-event data analysis with boosted first-hitting-time models based on a Gamma process

Clara Bertinelli Salucci[1], Azzeddine Bakdi[2], Ingrid Kristine Glad[3], Erik Vanem[4] and Riccardo De Bin[5]

[1] *University of Oslo, Department of Mathematics, Norway, clarabe@math.uio.no*
[2] *Corvus Energy, Porsgrunn, Norway, azzeddib@math.uio.no*
[3] *University of Oslo, Department of Mathematics, Norway, glad@math.uio.no*
[4] *University of Oslo, Department of Mathematics, DNV Group Technology and Research, Norway, erik.vanem@dnv.com*
[5] *University of Oslo, Department of Mathematics, Norway, debin@math.uio.no*

In time-to-event data analysis, the focus is on the ending point of some evolving process which is often unknown and, hence, goes unconsidered [1]. First-Hitting-Time models, in which the ending point is represented by the first passage time of an underlying stochastic process to a threshold, aim at considering the evolving process, too, and are appealing in many situations. Choosing a suitable stochastic process is indeed crucial for the analysis. In [2], a boosting algorithm for First-Hitting-Time models based on a Wiener process with drift is proposed. Since the Wiener process is characterised by Gaussian increments, the process is not guaranteed to be monotonic. In many cases, however, one deals with irreversible phenomena (e.g. nontreatable cancer, fatigue accumulation), therefore a monotonic process represents the most suitable alternative. In this work, we propose an extension of the boosting algorithm for First-Hitting-Time process based on an underlying Gamma process, to ensure the process monotonicity. We illustrate the performance of our algorithm, and its versatility, in real data examples from both engineering and biomedical applications, as well as a simulation study.

References

[1] Aalen, O. O., and Gjessing, H. K., Understanding the shape of the hazard rate: a process point of view, *Statistical Science*, 16, (2001) 1–14.

[2] De Bin, R. and Stikbakke, V., A boosting first-hitting-time model for survival analysis in high-dimensional settings, *Lifetime Data Analysis*, 29, (2023) 420-440.

Non-parametric intensity estimation of spatial point processes by random forests

Christophe A.N. Biscio[1], Frédéric Lavancier[2]

[1] Aalborg University, Denmark, christophe@math.aau.dk
[2] University of Nantes, France, frederic.lavancier@univ-nantes.fr

The intensity of a spatial point process is one of the first quantity of interest to estimate in presence of real-data. When no covariate is observed, non-parametric kernel estimation is routinely used, but comes with some drawbacks: it adapts poorly to non-convex domain, and the estimation is not consistent in an increasing domain asymptotic regime. When the intensity depends on observed covariates, most estimation methods are parametric. Non-parametric kernel estimation has been extended to this situation, but it appears to be efficient only for a few numbers of covariates, and provides no indication on the importance of each covariate, which is crucial for interpretation.

In this talk, we show how to adapt random forest regression to circumvent these drawbacks and estimate non-parametrically the intensity of a spatial point process with or without covariates, while measuring the importance of each variable in the latter case. Our approach allows to handle non-convex domain together with a large number of covariates. From a theoretical side, we prove that in the case of purely random forests, our method is consistent in both infill and increasing domain asymptotic regime, and may achieve a minimax rate of convergence.

Approximate subgraph count: asymptotic normality and jackknife estimate of variance

Mindaugas Bloznelis

Vilnius university, Vilnius, Lithuania, mindaugas.bloznelis@mif.vu.lt

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.

References

[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.

Resurrecting pseudoinverse: Asymptotic properties of large Moore-Penrose inverse with applications

Taras Bodnar[1] and Nestor Parolya[2]

[1] *Stockholm University, Stockholm, taras.bodnar@math.su.se*
[2] *Delft University of Technology, Delft, N.Parolya@tudelft.nl*

In this paper, we derive high-dimensional asymptotic properties of the Moore-Penrose inverse of the sample covariance matrix (see, e.g., [1], [2]), i.e, the case when number of variables p is larger than the sample size n. We prove the convergence results related to the traces of weighted moments of the Moore-Penrose inverse matrix, which involve both its eigenvalues and eigenvectors. We extend previous findings in several directions: (i) first, the population covariance matrix is not assumed to be a multiple of the identity; (ii) second, the assumptions of normality is not used in the derivation, only existence of the 4th moments is required; (iii) third, the asymptotic properties of the weighted moments are derived under the high-dimensional asymptotic regime $p \to \infty$ and $n \to \infty$ such that $p/n \to c > 1$. Our findings allow the construction of the optimal linear shrinkage estimators for the large precision matrix, for the vector of coefficients in the high-dimensional linear model and for the weights of optimal portfolio under l_2 regularization. Finally, the finite sample properties of the derived theoretical results are investigated via an extensive simulation study.

References

[1] Bodnar, T., Dette, H., Parolya, N. (2016). Spectral analysis of the Moore–Penrose inverse of a large dimensional sample covariance matrix. Journal of Multivariate Analysis, 148, 160-172.

[2] Imori, S., von Rosen, D. (2020). On the mean and dispersion of the Moore-Penrose generalized inverse of a Wishart matrix. The Electronic Journal of Linear Algebra, 36, 124-133.

An R package for Bayesian inference of mixed Gaussian phylogenetic models

Bayu Brahmantio[1]

[1] *Linköping University, Sweden, bayu.brahmantio@liu.se*

The evolution of continuous traits is often modelled using stochastic differential equations that combine deterministic change of a trait through time with noise that represents different unobservable evolutionary pressures. Two of the most popular choices are Brownian Motion and Ornstein-Uhlenbeck processes, which belong to the $\mathcal{G}_{LInv}$ family of models [1], i.e., models with a Gaussian transition probability whose expectation is linear with respect to ancestral value and variance is invariant with respect to it. Using this framework, it is possible to set different $\mathcal{G}_{LInv}$ models into different parts of a phylogenetic tree to do parameter inferences and model comparisons.

In this work, a Bayesian scheme is implemented as an extension of the maximum likelihood framework to include uncertainties in the parameters estimate and prior knowledge that are more biologically relevant. The method is written as an R package that applies MCMC with C++ backend to perform inferences. The package also features custom user-defined priors and Bayesian model selection.

References

[1] Mitov, Venelin, et al. "Fast likelihood calculation for multivariate Gaussian phylogenetic models with shifts." *Theoretical Population Biology* 131 (2020): 66-78.

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

Timo C. Braun[1], Anders Kvellestad[2] and Riccardo De Bin[3]

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

[2] *Department of Physics, University of Oslo, Norway, anders.kvellestad@fys.uio.no*

[3] *Department of Mathematics, University of Oslo, Norway, debin@math.uio.no*

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. To be able to speed up the computations and provide an estimate of the likelihood value in a reasonable time, 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 novel algorithm is illustrated on data from the GAMBIT project [2].

References

[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*

Semi-directed networks and their relation to epidemic models

Tom Britton[1]

[1] *Stockholm University University, Sweden, tom.britton@math.su.se*

First we describe the configuration model for a semi-directed network consisting of both directed and undirected edges, and some large population properties of such networks. Then we show that the final outcome of a surprisingly large class of epidemic models can be characterized by such semi-directed networks.

Higher dimensional stick percolation

Erik Broman[1]

[1] *Gothenburg University and Chalmers University of Technology, Gothenburg, Sweden, broman@chalmers.se*

The so-called stick percolation model was introduced by Roy in 1991 (see [2]). In its most basic form, the sticks are line-segments of lengths 1, the centers of the sticks are placed in R^2 according to a Poisson point process with intensity $\lambda > 0$, while the orientations of the sticks are chosen uniformly and independently.

Since the introduction of this model, only a short mathematical note (by Popov and Vachkovskaia 2002, see [3]) has been written concerning the model. However, within the physics and chemistry community there has been much more interest (see e.g. Mietta, Negri and Tamborenea in [1] and Tarasevich and Eserkepov in [4], and the references therein). In contrast to Roy's model, the sticks there have a non-zero width and "live" in both two-dimensional and three-dimensional space.

In this talk, I consider two variants of the stick percolation model in dimensions $d \geq 2$. In both, the sticks have width 1 and length L. In the first variant, the orientation distribution is uniform, while in the second variant, all sticks are oriented along the same direction.

Letting $\lambda_c(L)$ denote the critical intensity for the percolation phase transition, the focus of this project has been to establish the asymptotic behavior of $\lambda_c(L)$ as L diverges. For both variants, the exact scaling exponent was obtained.

References

[1] Mietta J. L., Negri R. M. and Tamborenea P. I., Numerical Simulations of Stick Percolation: Application to the Study of Structured Magnetorheological Elastomers *J. Phys. Chem. C*, Vol **118** (2014), No 35, 20594–20604.

[2] Roy R., Percolation of poisson sticks on the plane. *Probab. Theory Related Fields*, Vol **89** (1991), No 4, 503—517.

[3] Popov, S. and Vachkovskaia, M., A note on percolation of Poisson sticks. *Braz. J. Probab. Stat.*, Vol **16** (2002), no. 1, 59-–67.

[4] Tarasevich Y. Y. and Eserkepov A.V., Percolation thresholds for discorectangles: numerical estimation for a range of aspect ratios. *Phys. Rev. E*, Vol **101** (2020), No 2, 022108.

Beyond text and images: Training and evaluating generative models when both data and good metrics are hard to find

Judith Bütepage, Mónica Villanueva Aylagas, Mattias Teye, Héctor Leon, Sergi Andreu, Konrad Tollmar[1]

[1] *SEED - Electronic Arts, Stockholm, jbutepage@ea.com, mvillanuevaaylagas@ea.com, mteye@ea.com, hleon@ea.com, sergi.andreu.boix@gmail.com, ktollmar@ea.com*

In 2022, deep generative models have impressed not only the machine learning community but also the general public with their ability to generate images and coherent text. Images and natural language are well-suited playgrounds for these models because we have a large amount of training data and both of these modalities are easily interpretable by laymen and specialists alike. While quantitative evaluation remains challenging [1], qualitative evaluation of a model's samples is therefore relatively straightforward even for laymen. This is not always the case in other domains. In this talk, I will highlight the difficultly of training and evaluating generative models on small datasets in domains that are not easily interpretable by laymen, namely sound generation and facial animation. I will highlight how model and training scheme design can partially overcome the problems of small datasets and the lack of an appropriate training metric [2]. Additionally I will discuss the importance of expert domain knowledge [3] and how heuristics can help to select the best model even if no expert knowledge is available.

References

[1] Theis, L and van den Oord, A and Bethge, M, A note on the evaluation of generative models, *International Conference on Learning Representations (ICLR 2016)*, 2016, 1–10.

[2] Sergi, A and Villanueva Aylagas, M, Neural Synthesis of Sound Effects Using Flow-Based Deep Generative Models, Proceedings of the Eighteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 2022, 1.

[3] Villanueva Aylagas, M and Anadon Leon, H and Teye, M and Tollmar, K, Voice2Face: Audio-driven Facial and Tongue Rig Animations with cVAEs, Computer Graphics Forum, 2022, 255–265.

The maximum of the periodogram of a sequence of functional data

Vaidotas Characiejus[1], Clément Cerovecki[2] and Siegfried Hörmann[3]

[1] *University of Southern Denmark, Denmark, characiejus@imada.sdu.dk*
[2] *Katholieke Universiteit Leuven, Belgium, clement.cerovecki@kuleuven.be*
[3] *Graz University of Technology, Austria, shoermann@tugraz.at*

The detection of periodic signals in functional time series is investigated when the length of the period is not assumed to be known. A natural test statistic for the detection of periodicities is the maximum over all fundamental frequencies of the Hilbert-Schmidt norm of the periodogram operator. Using recent advances in Gaussian approximation theory, we show that under certain assumptions, the appropriately standardised test statistic belongs to the domain of attraction of the Gumbel distribution. The asymptotic results allow us to construct tests for hidden periodicities. We demonstrate the performance of our methodology in a simulation study, and we also illustrate the usefulness of our approach by examining periodicities in the air quality data from Graz, Austria and showing that our approach is not only able to detect the presence of periodic signals, but it is also able to reveal the structure of periodicities in the data. The talk is based on the paper by Cerovecki, Characiejus, and Hörmann [1].

References

[1] Clément Cerovecki, Vaidotas Characiejus, and Siegfried Hörmann. The maximum of the periodogram of a sequence of functional data. *Journal of the American Statistical Association*, to appear. URL `https://doi.org/10.1080/01621459.2022.2071720`.

Reflected backward SDEs in non-convex domain

Jean-François Chassagneux[1]

[1] *Universite Paris Cite, chassagneux@lpsm.paris*

In this talk, I will first present results on the well-posedness of reflected BSDEs in non-convex that we have recently obtained with S. Nadtochiy and A. Richou, under some structural assumptions on the domain and the terminal condition. These results were obtained using tools and estimates based on the Euclidean structure of $\mathbb{R}^d$. In a work in progress with M. Arnaudon, S. Nadtochiy and A. Richou, we improve these results, at least in dimension 2, by seeing our domain as a flat manifold with a boundary and relying on stochastic geometry tools already developed in other context. In a the second part of my talk, I will explain this new approach and the kind of results we are able to obtain.
Based on joint work with M. Arnaudon, S. Nadtochiy and A. Richou.

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

Dennis Christensen[1,2]

[1] *University of Oslo, Oslo*
[2] *Norwegian Defence Research Establishment (FFI), Kjeller, dennis.christensen@ffi.no*

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, 3]. There has been a large interest in applying Bayesian nonparametric models to such data since the early 1970s [4, 5, 6]. For models based on Dirichlet processes, posterior inference is possible via Markov chain Monte Carlo (MCMC) methods [6, 7, 8]. However, for many modern model choices, such MCMC based methods fail [9, 10, 11, 12]. Here we present a new importance sampling algorithm for nonparametric models given exchangeable binary response data [13]. 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.

References

[1] Morgan, B. J. T., Analysis of Quantal Response Data, Chapman & Hall, New Delhi, 1st edition, 1992.

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

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

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

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

[6] Doss, H., Bayesian nonparametric estimation for incomplete data via successive substitution sampling, *The Annals of Statistics*, 22:4, (1994) 1763–1786.

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

[8] Walker, S. G., Sampling the Dirichlet mixture model with slices, *Communications in Statistics - Simulation and Computation*, 36:1, (2007) 45–54.

[9] Murray, I.; MacKay, D. and Adams, R. P., The Gaussian process density sampler. In *Advances in Neural Information Processing Systems (NIPS)*, (2008) volume 21.

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

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

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

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

Likelihood-based inference for partially observed chemical reaction processes

Marc Corstanje[1] and Frank van der Meulen[2]

[1] *Vrije Universiteit, Amsterdam, m.a.corstanje@vu.nl*
[2] *Vrije Universiteit, Amsterdam, f.h.van.der.meulen@vu.nl*

Chemical reaction networks can be used to study a wide class of biological, physical and chemical processes that evolve over time. Such processes are typically modeled as d-dimensional stochastic jump processes with inhomogeneous jump intensities. We suppose that X denotes such a process and aim to study the statistical problem of performing inference on parameters that govern the dynamics of X based on partial observations $v_i = L_i X(t_i)$, where L_i is an matrix of dimension $m_i \times d$ with $m_i \leq d$ for $i = 1, \ldots, n$.

The approach we present is to simulate a so-called guided process presented in [1] that is equivalent to $(X \mid L_i X(t_i) = v_i,\, i = 1, \ldots, n)$ with closed form Radon-Nikodym derivative from which we then deduce a likelihood. In this talk, I will briefly introduce chemical reaction processes and the approach of conditioning by guiding. Next I will discuss the statistical problem and demonstrate various results for parameter inference.

References

[1] Corstanje, M. A., van der Meulen, F.H and Schauer, M., Conditioning continuous time Markov processes by guiding, *Stochastics*, (2022).

Point process learning: A cross-validation-based approach to statistics for point processes

Ottmar Cronie[1], Mehdi Moradi[2] and Christophe A.N. Biscio[3]

[1] *Department of Mathematical Sciences, Chalmers University of Technology & University of Gothenburg, Sweden, ottmar@chalmers.se*

[2] *Department of Mathematics and Mathematical Statistics, Umeå University, Sweden, mehdi.moradi@umu.se*

[3] *Department of Mathematical Sciences, Aalborg University, Denmark, christophe@math.aau.dk*

Point processes generalise classical iid samples by allowing i) the sample size to be random and/or ii) the sample points to be dependent. They have therefore become ubiquitous in the modelling of spatial/temporal event data, e.g. earthquakes and disease cases. This talk presents a new cross-validation-based statistical theory for general point processes [1], which is motivated by cross validation's general ability to reduce overfitting and mean square error. It is based on the combination of two novel concepts for general point processes: cross validation and prediction errors. Our cross-validation approach uses thinning to split a point process/pattern into pairs of training and validation sets, while our prediction errors measure discrepancy between two point processes. The new statistical approach exploits the prediction errors to measure how well a given model predicts validation sets using associated training sets. Due to its connection to the general idea of empirical risk minimisation, it is referred to as Point process learning. Having discussed its components and properties, we discuss why it generally outperforms the state of the art in most spatial statistical settings. We finally illustrate this numerically in one such setting.

References

[1] Cronie, O., Moradi, M., Biscio, C.A.N., A cross-validation-based statistical theory for point processes, *Biometrika* (under review), (2023).

Blending forecasts for the time-to-frost

Céline Cunen[1]

[1] *Norwegian Computing Center, Oslo, celine.cunen@gmail.com*

In this talk I will present some general methods for combining probabilistic forecasts from two or more sources. Ideally, the final blended forecast should be well-calibrated and more precise than any of the single-source forecasts. We have developed a new blending method, called the Gaussian Forecast Filter (GFF), which seems to have good properties. The methods will be illustrated through a real application: attempting to precisely forecast the time-to-frost some months ahead by blending seasonal and subseasonal forecasts. This talk is based on joint work with Thea Roksvåg, Claudio Heinrich-Mertsching and Alex Lenkoski.

An alternative approach to rank aggregation

Antonio D'Ambrosio[1], Maurizio Romano[2] and Claudio Conversano[2]

[1] *University of Naples Federico II, antdambr@unina.it*
[2] *University of Cagliari, maurizio.romano@unica,it, claudio.conversano@unica.it*

Rank aggregation is maybe the main task in analyzing preference rankings, as it concerns producing the so-called central (or median, or consensus) ranking, namely that ranking (or those rankings) that best represents the consensus opinion of a set of judges. In this work we propose an alternative optimization algorithm dealing with a large number of items. We follow the Kemeny's axiomatic approach [1], hence we deal with the universe of rankings containing all the possible ties, not constraining the solution to be necessarily a permutation.

References

[1] Kemeny, J. G., and Snell, L. *Mathematical models in the social sciences*. Boston: Ginn and Company, 1962.

A simple concentration inequality for minimum weight combinatorial optimisation problems

Joel Larsson Danielsson[1]

[1] *Lund University, joel.danielsson@stat.lu.se*

Given the complete graph K_n with i.i.d. edge costs/weights X_e and a family $\mathcal{F}$ of subgraphs, we study the problem of finding an $F \in \mathcal{F}$ of minimal total cost $M(\mathcal{F}) := \min_{F\in\mathcal{F}} \sum_{e\in E(F)} X_e$. In a large number of cases (e.g. [1, 2, 3, 4]), this minimum is known to be sharply concentrated. For instance, when $X_e \sim U(0,1)$ and $\mathcal{F}$ is the family of spanning trees, it is known that $M(\mathcal{F}) \to \zeta(3) \approx 1.202$ (in probability, as $n \to \infty$)[3].

We present a concentration inequality for the minimum cost $M(\mathcal{F})$, with easily checkable conditions on $\mathcal{F}$ and X_e. Roughly speaking, suppose all $F \in \mathcal{F}$ has at most m edges. If $O(\sqrt{m})$ edges are deleted from the $F \in \mathcal{F}$ attaining the minimum $M(\mathcal{F})$, how cheaply can we repair this damage? If this cost is $\ll \mathbf{E}M(\mathcal{F})$, then $M(\mathcal{F})$ is sharply concentrated.

The inequality is obtained by applying Talagrand's concentration inequality [6] to a dual problem. Although the proof is fairly elementary, the resulting inequality is often easier to apply than alternative methods. The inequality was used in [4] (joint with Federico) and will appear in [5].

References

[1] Aldous, D. The $\zeta(2)$ limit in the random assignment problem, *Random Structures & Algorithms*, 2001, 18(4):381-418

[2] Benjamini, I., Lubetzky, E. and Peled, Y., Minimum weight disk triangulations and fillings, *Transactions of the American Mathematical Society*, 2021, 374(5):3265–3287

[3] Frieze, A. M., On the value of a random minimum spanning tree problem, *Discrete Applied Mathematics*, 1985, 10(1):47–56

[4] Federico L. and Larsson Danielsson, J., Minimal H-factors and covers, *Preprint*, 2023, arXiv:2302.12184

[5] Larsson Danielsson, J., A concentration inequality for combinatorial minimum-weight problems, *In preparation*, 2023

[6] Talagrand, M., New concentration inequalities in product spaces, *Inventiones mathematicae*, 1996, 126(3):505-563

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

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

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.

References

[1] E. Bareinboim and J. Pearl. 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] J. Correa, J. Tian, and E. Bareinboim. Generalized adjustment under confounding and selection biases. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.12125. URL https://ojs.aaai.org/index.php/AAAI/article/view/12125.

Covariate adjustment based on empirical likelihood and robust estimators

Māra Delesa-Vēliņa[1], Alvīne Helēna Janševska[1]

[1] University of Latvia, Latvia, mara.delesa-velina@lu.lv

Covariate adjustment is often used in statistical analysis of randomized experiments to increase efficiency of estimators of treatment effects. Usually the classical ANCOVA method is employed, however, when the assumptions of normality and variance homogeneity are violated simultaneously, it is not robust and can yield test size different from the nominal.

A nonparametric method for covariate adjustment based on generalized empirical likelihood (EL) for the mean treatment effects was proposed in [1]. However, EL-based inference for the mean can have low power when sampling from heavy-tailed distributions or in the presence of outliers. We propose an extension to [1], using robust estimators for the treatment effects and for covariate adjustment. In particular, we consider trimmed means and M-estimators, since they are shown to have desirable robustness properties in two-sample comparison problems [2]. We evaluate the proposed methods in terms of RMSE, empirical coverage of the CIs and the test power.

References

[1] Chaussé, P., Liu, J. & Luta, G. A simulation-based comparison of covariate adjustment methods for the analysis of randomized controlled trials. International Journal of Environmental Statistics and Public Health, 13, (2016) 414.

[2] Delesa-Velina, M., Valeinis, J. & Luta, G. Comparing two independent populations using a test based on empirical likelihood and trimmed means. Lithuanian Mathematical Journal, 61, (2021) 199–216.

Using Model-based Geostatistical Methods to Support Neglected Tropical Disease Control Programmes

Peter J Diggle[1], **Claudio Fronterre**[1], **Emanuele Giorgi**[1]

[1] *Lancaster University, p.diggle@lancaster.ac.uk*

Neglected Tropical Disease Control Programmes aim to reduce the prevalence of NTDs to a level at which they can be considered to have been eliminated as a public health problem or, more ambitiously, eradicated altogether. The core statistical challenge in this endeavour is to design and analyse prevalence surveys that achieve pre-agreed performance levels (typically, acceptably high positive and negative predictive values for declaration of elimination at a specified level of spatial resolution) at minimum cost.

The NTD community is beginning to recognise that geospatial statistical methods are better suited to this task than are the currently prevailing methods based on classical survey sampling methodology (Diggle et al, 2023). In this talk I will review the basic components of model-based geostatistics (Diggle, Moyeed and Tawn, 1998), show how these have recently been applied to a national survey of lymphatic filariasis prevalence in Guyana, and discuss other applications (some incomplete) where extensions of the basic methodology are required.

References

[1] Diggle, P.J., Fronterre, C., Gass, K., Hundley, L., Niles-Robin, R., Sampson, A., Morice, A. and Carvalho Scholte, R., Modernising the design and analysis of prevalence surveys for Neglected Tropical Diseases. *Philosophical Transactions of the Royal Society*, to appear, (2023)

[2] Diggle, P.J., Moyeed, R.A. and Tawn, J.A., Model-based geostatistics (with Discussion). *Applied Statistics*, 47, (1998) 299-350.

Minimax-Bayes reinforcement learning

Thomas Kleine Buening,[1] **Christos Dimitrakakis**[1 2 3], Hannes Eriksson[2], Divya Grover and Emilio Jorge[2]

[1] *University of Oslo*
[2] *Chalmers University of Technology*
[3] *University of Neuchâtel*

While the Bayesian decision-theoretic framework offers an elegant solution to the problem of decision making under uncertainty, one question is how to appropriately select the prior distribution. One idea is to employ a worst-case prior, known as the *minimax-Bayes* formulation of the decision problem. While this has been studied intensively in simple statistical estimation problems, its study is difficult when it is combined with adaptive data collection. In this talk, I will explain the minimax-Bayes idea in terms of a zero-sum game between the statistician and Nature, and outline recent solution methods for such games. I will then discuss sequential statistical decision problems, and in particular, Bayesian reinforcement learning. Finally, I will go over some algorithms we developed for minimax-Bayes methods for reinforcement learning. I demonstrate (sometimes approximate) minimax-Bayes solutions for various reinforcement learning problems to gain insights into the properties of the corresponding priors and policies. Our findings show that while the worst-case prior depends on the setting, the corresponding minimax policies are more robust than those that assume a standard (i.e. uniform) prior.

Sandwiched Volterra volatility models and hedging

Giulia di Nunno[1] and Anton Yurchenko-Tytarenko[2]

[1] *University of Oslo, giulian@math.uio.no*
[2] *University of Oslo, antony@math.uio.no*

We present the Sandwiched Volterra Volatility (SVV) model and we discuss some of its characteristic properties. We then move to consider hedging and consider the explicit computation of quadratic hedging strategies. The theoretical solution formula is well-known in terms of the non-anticipating derivative for all square integrable claims, the fact that these models are typically non-Markovian provides a concrete difficulty in the direct computation of conditional expectations at the core of the explicit hedging strategy. To overcome this difficulty, we propose a Markovian approximation of the model which stems from an adequate approximation of the kernel in the Volterra noise. We show some numerical simulations performed with different methods.

References

[1] Di Nunno, G. and A. Yurchenko-Tytarenko *ArXiv*: 2209.13054

Prediction of tipping times

Susanne Ditlevsen[1] and Peter Ditlevsen[2]

[1] *Department of Mathematical Sciences, University of Copenhagen, Denmark, susanne@math.ku.dk*
[2] *Niels Bohr Institute, University of Copenhagen, Denmark*

In recent years there has been an increasing awareness of the risks of collapse or tipping points in a wide variety of complex systems, ranging from human medical conditions, pandemics, ecosystems to climate, finance and society. They are characterized by variations on multiple spatial and temporal scales, leading to incomplete understanding or uncertainty in modelling of the dynamics. Even in systems where governing equations are known, such as the atmospheric flow, predictability is limited by the chaotic nature of the system and by the limited resolution in observations and computer simulations. In order to progress in analyzing these complex systems, assuming unresolved scales and chaotic dynamics beyond the horizon of prediction as being stochastic has proven itself efficient and successful.

When complex systems undergo critical transitions by changing a control parameter through a critical value, a structural change in the dynamics happens, the previously statistically stable state ceases to exist and the system moves to a different statistically stable state. To establish under which conditions an early warning for tipping can be given, we consider a simple stochastic model, which can be considered a generic representative of many complex two state systems. We show how this provides a robust statistical method for predicting the time of tipping. The method is used to give a warning of a forthcoming collapse of the Atlantic meridional overturning circulation.

References

[1] Ditlevsen, P.D., Ditlevsen, S. (2022). Warning of a forthcoming collapse of the Atlantic meridional overturning circulation, *submitted*, 2022.

Conditional Negative Definiteness in Multivariate Geostatistics revisited: the Pseudo Cross-Variogram

Christopher Dörr[1] and Martin Schlather

[1] *University of Mannheim, Mannheim, Germany, christopher.doerr@students.uni-mannheim.de*
[2] *University of Mannheim, Mannheim, Germany, schlather@uni-mannheim.de*

Pseudo cross-variograms are a useful tool for the analysis of multivariate random fields, and appear naturally in spatial extreme value theory. We give a necessary and sufficient criterion for a matrix-valued function to be a pseudo cross-variogram, and further present a Schoenberg-type result connecting pseudo cross-variograms and multivariate correlation functions. This connection adds another dimension to the range of applications of pseudo cross-variograms, that is, the construction of covariance models for multivariate random fields. We illustrate this point via several examples.
The underlying key ingredient for the above results is a stronger notion of conditional negative definiteness for matrix-valued functions than the one for cross-variograms, which have been predominantly used in geostatistical literature. We discuss this stronger notion, and present a general construction principle for conditionally negative definite matrix-valued kernels, which we use to reinterpret previous covariance modelling proposals.

A recommendation system with reinforcement for determining the price of a product

Svitlana Drin[1], Anastasia Kriuchkova[2]

[1] *School of Business, Örebro University, Örebro, svitlana.drin@oru.se*
[2] *Department of Mathematics, National University of Kyiv-Mohyla Academy, Ukraine, anastasia.kriuchkova@ukma.edu.ua*

Recommender systems belong to a fairly new concept with a wide scope of application and the property of adaptation to non-classical optimization problems. Thus, recommender systems can be used in e-commerce to predict demand and price for a new product in a chain of stores. LightGBM is a reinforcement recommendation system based on regression trees. The Boosting technology for the new tree $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda f^k(x)$ is selected with the specified parameter value λ. A typical value of λ is 0.001 to 0.1. But despite its simplicity and many obvious advantages, for many real big data with long-term dependence, classic Boosting techniques should not be used in real models. For example, if there are 900 stores and 7000 SKUs in the chain, then we have more than 6 million combinations per day in the event horizon. In the LightGBM technique, trees of different structure q are combined, where $q : R^m \rightarrow T$. And in the known tree structure $q(x)$, the optimal weight of each leaf is obtained by minimizing the loss function $L'^{(t)} = \sum_{i=1}^{T} \left[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2\right] + \gamma T$.

It is worth noting that the LightGBM algorithm, which is characterized by its efficiency, accuracy and speed, creates histograms and uses the generated classes instead of the entire range of values of each variable, achieving a significant reduction in training time. The generated classes are the result of the Gradient One Side Sampling (GOSS) method. The main idea of the GOSS methodology focuses on the fact that not all observations contribute equally to the learning of the algorithm, since those with a small first derivative of the loss function learn better than those with a large first derivative of the loss function.

In the case of forecasting the price of goods that do not have a sales history, we will use embedding in NLP in our model. The main purpose of using embedding is the distance between vectors of products to a new product that have a similar nature of sales, and this distance should be minimal. Results will be evaluated using the MAE performance metric.

References

[1] Demertzis, K., Kostinakis, K., Morfidis, K., Iliadis, L., A Comparative Evaluation of Machine Learning Algorithms for the Prediction of R/C Buildings' Seismic Damage, *arXiv preprint arXiv:2203.13449*, (2022).

[2] Yaqi Mao, Yonghui Duan, Yibin Guo, Xiang Wang, A Study on the Prediction of House Price Index in First-Tier Cities in China Based on Heterogeneous Integrated Learning Model, *Journal of Mathematics*, (2022)

Accurate bias estimation with applications to focused model selection

Ingrid Dæhlen[1], Nils Lid Hjort[1] and Ingrid Hobæk Haff[1]

[1] *University of Oslo, Norway, ingrdae@math.uio.no, nils@math.uio.no, ingrihaf@math.uio.no*

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].

References

[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.

Guided smoothing and control for diffusion processes

Oskar Eklund[1], Annika Lang[1], Moritz Schauer[1]

[1] *Chalmers University of Technology and University of Gothenburg, Gothenburg, ekoskar@chalmers.se, annika.lang@chalmers.se, smoritz@chalmers.se*

To solve the smoothing problem for general diffusion processes given by a stochastic differential equation, we extend the theory of guided processes. The smoothing distribution is the conditional distribution of the process in the space of trajectories given noisy observations made continuously in time. It is generally difficult to sample from this distribution. Here guided processes are versatile tools because they mimic how the drift of a stochastic process changes given observations. Motivated by an application of the Bellman principle from optimal control, we use the theory of enlargement of filtrations to show that the optimally guided process resembles the unobserved process, with an additional drift term derived from the backward filtering distribution that is moving or guiding the process towards the observations. We derive Markov Chain Monte Carlo and sequential Monte Carlo algorithms to sample from the smoothing distribution using similarly guided processes. The choice of the guided process is discussed and evaluated. Earlier, guided processes have been used for discrete-time observation processes. See e.g. [1] and [2] for previous work on guided processes and smoothing and see e.g. [3] for a similar model to ours and some derivations we need for our framework to work.

References

[1] F. van der Meulen and M. Schauer. Automatic Backward Filtering Forward Guiding for Markov processes and graphical models. 2020. doi: 10.48550/ARXIV.2010.03509. url: https://arxiv.org/abs/2010.03509

[2] M. Mider, M. Schauer, and F. van der Meulen. “Continuous-discrete smoothing of diffusions”. In: Electronic Journal of Statistics 15.2 (Jan. 2021). doi: 10.1214/21-ejs1894. url: https://doi.org/10.1214/21-ejs1894.

[3] B. Delyon and Y. Hu. Simulation of conditioned diffusions. 2006. doi: 10.48550/ARXIV.MATH/0602455. url: https://arxiv.org/abs/math/0602455

Cellular Automata modelling excitatory and inhibitory neurons

Henrik Ekström[1], Tatyana Turova[1]

[1] *Lund University, Henrik.Ekstrom@matstat.lu.se, Tatyana.Turova@matstat.lu.se*

We consider a cellular automaton model for the propagation of electrical impulses or activity in a neuronal network. The vertices of a square lattice represent neurons, and the edges of the lattice represent the synaptic connections. Each vertex can be active or inactive and is additionally assigned a fixed type, either inhibitory or excitatory, which dictates the effect it will have on its neighbours. The dynamics of propagation of the initial activity captures features of the "integrate-and-fire" model. We study the spread of activation in a large network and describe possible spatio-temporal limiting patterns depending on the initial activation. The rich palette of the limits with qualitatively different properties, including expanding patterns, fixed patterns, and patterns of constant size moving across the network, allows us to argue that this is a versatile model for the study of associative memory.

Model-based estimation and mapping of plant density based on remote sensing and presence/absence data

Magnus Ekström[1,2], Léna Gozé[1], Jörgen Wallerman[1], Jonas Dahlgren[1], Bengt Gunnar Jonsson[3], Saskia Sandring[1], and Göran Ståhl[1]

[1] *Swedish University of Agricultural Sciences, Umeå, Magnus.Ekstrom@slu.se, Lena.Goze@slu.se, Jorgen.Wallerman@slu.se, Jonas.Dahlgren@slu.se, Saskia.Sandring@slu.se, Goran.Stahl@slu.se*

[2] *Umeå University, Umeå, Magnus.Ekstrom@umu.se*

[3] *Mid Sweden University, Sundsvall, Bengt-Gunnar.Jonsson@miun.se*

Inventories of populations are of great importance in ecological research. For monitoring plants, presence/absence (P/A) sampling is simple to conduct, since only presence or absence of a species on a plot needs to be registered. Estimates of plant density may be obtained from P/A data but need to be based on model assumptions about the spatial distribution of plants. We have showed, in a model-based setting, how to use P/A and remote sensing data for estimating and mapping plant density for regions and subregions, where the model assumes plant locations to follow an inhomogeneous Poisson point process. To guard against model misspecifications, we derived a test for assessing the plausibility of the model assumptions of the inhomogeneous Poisson point process. Using empirical data from the Swedish National Forest Inventory as well as artificial plant population data, we evaluated the performance of the suggested estimators of plant density and the proposed test of model assumptions, respectively.

SIRS model with gradual waning of immunity

Mohamed El Khalifi[1] and Tom Britton[1]

[1] *Dept of Mathematics, Stockholm University, Stockholm,*
mohamed.elkhalifi@math.su.se, tom.britton@math.su.se

The SIRS epidemic model for the spreading of infectious diseases makes the extreme assumption that individual immunity drops all at once. Here, a novel mathematical model is presented allowing for gradual waning of immunity following linear or exponential waning modes. The two new models and the SIRS model are compared assuming all three models have the same cumulative immunity. When no intervention is put in place, we find that the long-term prevalence is higher for the models with gradual waning. If aiming for herd immunity by continuous vaccination, it is shown that larger vaccine quantities are required when immunity wanes gradually compared with results obtained from the SIRS model, and this difference is biggest for the most realistic assumption of exponentially waning of immunity.

References

[1] El Khalifi, M., Britton, T. Extending sirs epidemics to allow for gradual waning of immunity. arXiv preprint arXiv:2211.09062, 2022.

[2] Inaba, H. Kermack and McKendrick revisited: the variable susceptibility model for infectious diseases. Japan Journal of Industrial and Applied Mathematics, 18(2):273–292, 2001.

[3] Martcheva, M. An introduction to mathematical epidemiology, volume 61. Springer, 2015.

Real-time inference in stochastic metapopulation models: the case of COVID-19 in Norway

Solveig Engebretsen[1]

[1] *Norwegian Computing Center, Norway, solveig.engebretsen@nr.no*

In this talk, I will present the regional model which has been used by the Norwegian Institute of Public Health for real-time surveillance of COVID-19 in Norway. Regional heterogeneity has been a key trait of the COVID-19 pandemic, due to e.g. regionally varying individual behaviour and mobility, locally implemented interventions and vaccination. Timely information about the local effects of most recent interventions is crucial for operational policy decisions. The model is informed by real-time mobility estimates from mobile phone data, confirmed imported cases from abroad, laboratory-confirmed cases and hospitalisation incidence. Inference for many parameters is necessary in a regional model, and to obtain estimates in useful time has been one of the key challenges during the pandemic surveillance in Norway. We developed a novel sequential Approximate Bayesian Computation motivated by real-time operational usefulness. The idea is to split the calibrations into different periods and move chronologically, restricting past parameters to samples from the posterior distributions, allowing to focus the computational time during operation on the most recent transmissibility parameters. In addition, for computational efficiency, we let the priors for the regional transmissibility parameters be informed by posteriors from separate, simpler, lower-dimensional calibration settings for each region.

Variance change point detection with a Binomial method

Elias Erdtman[1]

[1] *Linköping University, Department of Mathematics, elias.erdtman@liu.se*

We consider a simple method for detecting a change with respect to the variance in a sequence of independent normally distributed observations with a constant mean. The method filters out extreme values and divides the sequence into equally large subsequences. For each subsequence, the count of extreme values is translated as a binomial random variable which is tested toward the expected number of extremes. The expected number of extremes comes from prior knowledge of the sequence and a specified probability of how common an extreme value should be. Then specifying the significance level of the goodness of fit test gives how many extreme observations are needed to detect a change.

The approach is extended to a sequence of independent multivariate normally distributed observations by using the Mahalanobis distance to transform the sequence into a univariate sequence and apply the same approach. Also, some other statistics, e.g., changes in eigenvalues or information measures such as the Kullback-Leibler divergence and the Bhattacharyya distance, are considered.

Normal approximation for mixtures of normal distributions: error bounds via Stein's method

Krzysztof Bartoszek[1], **Torkel Erhardsson**[2]

[1] Krzysztof Bartoszek, Linköping University, krysztof.bartoszek@liu.se

[2] Torkel Erhardsson, Linköping University, torkel.erhardsson@liu.se

The distribution of a random variable X is called a mixture of normal distributions if the conditional distributions of X given some σ-algebra $\mathscr{G}$ are normal. We here give explicit bounds for the Kolmogorov and Wasserstein distances between a mixture of normal distributions and a normal distribution with properly chosen parameter values. The bounds depend only on the first two moments of $\mathbb{E}(X|\mathscr{G})$ and $\mathbb{E}(X^2|\mathscr{G})$. The proof makes use of Stein's method and couplings. As an application, we consider a model from evolutionary biology, in which an Ornstein-Uhlenbeck process evolves on top of a random tree given by a Yule process.

References

[1] Bartoszek, K. and Erhardsson, T., Normal approximation for mixtures of normal distributions and the evolution of phenotypic traits, *Adv. in Appl. Probab.*, 53, (2021) 162–188.

Inference of time-varying parameters in epidemiological models

Jonas Evaeus[1], Stefan Engblom[1]

[1] *Uppsala University, Uppsala, jonas.evaeus@it.uu.se, stefane@it.uu.se*

Reliable and efficient parameter inference in mechanistic epidemiological models is an important goal, as it would allow for well-informed decision making during ongoing epidemics. Meanwhile, the low information content of typical data sources make parameter inference challenging, particularly when considering small population sizes for which stochastic effects play an important role. We present and compare three methods for inferring time-varying parameters in stochastic epidemiological models, paying special attention to the related uncertainty estimates. These methods include a guided SMC-ABC (Sequential Monte Carlo-Approximate Bayesian Computation) scheme [1], an approach inspired by the IMM (Interacting Multiple Model) filter [2], and a dynamic optimization procedure inspired by MPC (Model Predictive Control) which was first introduced in [3]. The study concerns computational experiments for which the ground truth is available, allowing for a straightforward comparison between the three algorithms with regard to their accuracy, uncertainty estimation, and overall computational efficiency. Finally, we discuss common features of the three algorithms, offer insight into key points that should be considered when inferring dynamic parameters in epidemiological models, and point to remaining challenges.

References

[1] Picchini, U. and Tamborrino, M., Guided sequential schemes for intractable Bayesian models, *arXiv preprint arXiv:2206.12235*, 2022.

[2] Bar-Shalom, Y. and Li, X-R., Multitarget-Multisensor Tracking: Principles and Techniques, *Storrs, CT: YBS Publishing*, 1995.

[3] Marin, R., Runvik H., Medvedev, A., and Engblom, S., Bayesian Monitoring of COVID-19 in Sweden, *arXiv preprint arXiv:2205.00859*, 2022.

A design utility approach for preferentially sampled spatial data

Evangelos Evangelou[1] and Elizabeth J. Gray[2]

[1] *University of Bath, ee224@bath.ac.uk*
[2] *Imperial College London, e.gray@imperial.ac.uk*

[1] introduced the concept of spatial preferential sampling to refer to the situation in which the choice of sampling locations is stochastically dependent on the values of the spatial process of interest. Traditional geostatistical methods ignore this dependence, leading to potentially inaccurate inferences. I will present a general framework, based on design utilities, for modelling the preferences of the experimenter jointly with the spatial process of interest in order to adjust for this bias. This framework allows us to dispense with the unrealistic assumption (required by existing methods) of conditional independence of the sampling locations, and which may encapsulate an arbitrarily wide range of preferences. I will present techniques for generating designs under the proposed model and their incorporation in a wider Monte Carlo fitting scheme.

References

[1] Diggle, P. J., Menezes, R., & Su, T. L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 59(2), 191-232.

On covariance matrix tests under T-distributed models

Katarzyna Filipiak[1] and Tõnu Kollo[2]

[1] *Poznań University of Technology, katarzyna.filipiak@put.poznan.pl*
[2] *University of Tartu, Estonia, tonu.kollo@ut.ee*

In this presentation the likelihood ratio, Rao score and Wald test statistics to verify the hypothesis related to the specific form of covariance matrix under the multivariate T-distribution are studied. Convergence to the asymptotic chi-square distribution under the null hypothesis is examined in extensive simulation experiments. It is shown that the Rao score test behaves most adequately compared to the remaining tests in the situation when the dimensionality is growing.

Accelerated estimation of genomic values in animal breeding

Alexander Freudenberg[1], Martin Schlather[1], Jeremie Vandenplas[2]

[1] *University of Mannheim, alexander.freudenberg@uni-mannheim.de, martin.schlather@uni-mannheim.de*
[2] *Wageningen University and Research, jeremie.vandenplas@wur.nl*

In recent years, the growing size of datasets used for genomic selection in animal breeding has led to the need for more efficient statistical estimation of genomic values. This talk presents novel approaches to accelerating the estimation of genomic values by employing tailored algorithms designed to solve the assoiated mixed model equations more efficiently. We illustrate how breeding programs can benefit from these advances and discuss avenues for further enhancements in the future.

A neutral comparison of algorithms to minimize L_0 penalties for high-dimensional variable selection

Florian Frommlet[1]

[1] *Medical University Vienna, Austria, Florian.Frommlet@meduniwien.ac.at*

Variable selection methods based on L_0 penalties have excellent theoretical properties to select sparse models in a high-dimensional setting. There exist modifications of BIC which either control the family wise error rate (mBIC) or the false discovery rate (mBIC2) in terms of which regressors are selected to enter a model [1]. However, the minimization of L_0 penalties comprises a mixed inte-ger problem which is known to be NP hard and therefore becomes computationally challenging with increasing numbers of regressor variables. The last few years have seen some real progress in developing new algorithms to minimize L_0 penalties. In a neutral comparison study simulations covering a wide range of scenarios inspired by genetic association studies were performed to compare the performance of different algorithms. Based on these simulations and a real data example concerned with eQTL mapping we provide a clear recommendation which algorithms to use in practice.

References

[1] Bogdan, M., Frommlet, F. Identifying Important Predictors in Large Data Bases - Multiple Testing and Model Selection., *Handbook of multiple comparisons*, Chapman and Hall/CRC, 2021. 139-182.

Spatially Varying Anisotropy in Three Dimensions with Applications to Adaptive Oceanographic Sampling

Geir-Arne Fuglstad[1], and Martin Outzen Berild[1]

[1] *Norwegian University of Science and Technology, Norway, geir-arne.fuglstad@ntnu.no, martin.o.berild@ntnu.no*

Gaussian random fields (GRFs) with isotropic covariance structures are prolific in spatial statistics, but are only pragmatic approximations. E.g., in the ocean, the variability of a response may vary with depth, and ocean currents may lead to spatially varying anisotropy. In this talk, we discuss how to construct and how to parametrize a class of GRFs in three-dimensional space, which can incorporate such non-stationarity. Specifically, we parametrize spatially varying anisotropy through vector fields which control the coefficients of a stochastic partial differential equation, and make efficient computations using a Gaussian Markov random field approximation.

The new model is used to estimate a stationary GRF prior and a non-stationary GRF prior for salinity in an ocean mass outside Trondheim, Norway, based on the complex ocean model SINMOD. Then these GRF priors are evaluated using in-situ measurements collected with an autonomous underwater vehicle. We find that the non-stationary model outperforms the stationary anisotropic GRF prior for real-time prediction of unobserved locations both in terms of root mean square error and continuous rank probability score.

Biased random walk on dynamical percolation

Nina Gantert[1], Sebastian Andres[2] Perla Sousi[3] and Dominik Schmid[4]

[1] *Technical University of Munich, gantert@ma.tum.de*
[2] *University of Manchester, sebastian.andres@manchester.ac.uk*
[3] *Cambridge University, p.sousi@statslab.cam.ac.uk*
[4] *University of Bonn, d.schmid@uni-bonn.de*

We study the biased random walk for dynamical percolation on Z^d. We establish a law of large numbers and an invariance principle for the random walk using regeneration times. Moreover, we verify that the Einstein relation holds, and we investigate the speed of the walk as a function of the bias. While for $d = 1$ the speed is increasing, we show that in general this fails in dimension $d \geq 2$. As our main result, we establish two regimes of parameters, separated by a critical curve, such that the speed is either eventually strictly increasing or eventually strictly decreasing. This is in sharp contrast to the biased random walk on a static supercritical percolation cluster, where the speed is known to be eventually zero.

Unbiased maximum likelihood estimation for Wright-Fisher diffusion processes

Celia García-Pareja[1] and Fabio Nobile[1]

[1] *Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland, celia.garciapareja@epfl.ch, fabio.nobile@epfl.ch*

In this talk I will present an unbiased Monte Carlo maximum likelihood estimator for discretely observed Wright-Fisher diffusions. Our approach is based on exact simulation techniques that are of special interest for diffusion processes defined on a bounded domain, where numerical methods typically fail to remain within the required boundaries. We start by building unbiased maximum likelihood estimators for scalar diffusions and later present an extension to the multidimensional case. Consistency results of our proposed estimator are also presented and the performance of our method will be illustrated through numerical examples.

Directional gradient estimates on the Lévy-Itô space

Stefan Geiss[1] and Thuan Nguyen [2]

[1] *University of Jyväskyä, Finland, stefan.geiss@jyu.fi*
[2] *Saarland University, Saarbrücken, Germany, nguyen@math.uni-sb.de*

Assume a pure jump Lévy process $X = (X_t)_{t\in[0,T]}$ with Lévy measure ν and a Borel function $f : \mathbf{R} \to \mathbf{R}$ with $f(x + X_s) \in L_1$ for $(s, x) \in [0, T] \times \mathbf{R}$. Define $F : [0, T] \times \mathbf{R} \to \mathbf{R}$ by $F(t, x) := Ef(x + X_{T-t})$ and the vector-valued gradient

$$D_J F : [0,T) \times \mathbf{R} \to \mathrm{L}_0(\mathbf{R} \setminus \{0\}) \quad \text{by} \quad \mathrm{D_J F(t,x)} := \left\{ \mathrm{z} \mapsto \frac{\mathrm{F(t,x+z) - F(t,x)}}{\mathrm{z}} \right\}$$

known from Malliavin calculus and non-local PDEs. If ρ is a finite Borel measure on $\mathbf{R}$ sharing the small ball estimate $\rho([-r, r]) \leq cr^{\varepsilon}$ for some $\varepsilon \geq 0$ and if the coupling property $\|P_{z+X_s} - P_{X_s}\|_{TV} \leq d|z|s^{-\frac{1}{\beta}}$ holds for some $\beta \in (0, 2]$, then in [1] we prove

$$\left\| (T-t)^{\alpha} \sup_{x\neq 0} \left| \int_{\mathbf{R}\setminus\{0\}} (\partial_J F(t,x))(z) d\rho(z) \right| \right\|_{L_q((0,T],\frac{dt}{T-t})} \leq C\|f\|_{\mathrm{Hoel}_{\eta,q}},$$

where f belongs to the Besov space $\mathrm{Hoel}_{\eta,q}(\mathbf{R})$ with $(\eta, q) \in (0, 1-\varepsilon) \times [1, \infty]$, $X \subseteq L_{\eta+\gamma}$ for some $\gamma > 0$, and for $\alpha := \frac{1-(\varepsilon+\eta)}{\beta} > 0$. The exponent α is best possible. The estimate applies to stable like processes. Applications to the predictable representation property on the Lévy-Itô space and the path-regularity of the gradient $D_J F$ are given.

References

[1] S. Geiss and T. Nguyen: On Riemann-Liouville type operators, BMO, gradient estimates in the Lévy-Itô space, and approximation, arXiv:2009.00899.

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

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

[1] *University of Copenhagen, Copenhagen, Denmark [ng@math.ku.dk]*
[2] *University of Geneva, Geneva, Switzerland*
[3] *ETH Zurich, Zurich, Geneva, Switzerland*

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.

References

[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.

Decay of connection probabilities with distance in 2D and 3D neuronal networks

Vasilii Goriachkin[1] and Tatyana Turova[2]

[1] *Lund University, Lund, vasilii.goriachkin@matstat.lu.se*
[2] *Lund University, Lund, tatyana.turova@matstat.lu.se*

We study connectivity in a model of a growing neuronal network in dimensions 2 and 3. Although the axon-to-dendrite proximity is an insuffcient condition for establishing a functional synapse, it is still a necessary one. Therefore we study connection probabilities at short distances between the randomly grown axon trees and somas as probabilities of potential connections between the corresponding neurons. Our results show that, contrary to a common belief, these probabilities do not necessarily decay polynomially or exponentially in distance, but there are regimes of parameter values when the probability of proximity is not sensitive to the distance. In particular, in 3 dimensions the Euclidean distance between the neuronal cell body centers of neurons seems to play a very subtle role, as the probabilities of connections are practically constant within a certain finite range of distance.

The model has a sufficient number of parameters to assess networks of neurons of different types. Our results give a firm basis for further modelling of the neuronal connectivity taking into account some realistic bouton distributions for establishing synaptic connections.

Estimation of plant density based on presence/absence data using hybrid inference

Léna Gozé[1], Magnus Ekström[1 2], Saskia Sandring[1], Bengt-Gunnar Jonsson[3], Jörgen Wallerman[1] and Göran Ståhl[1]

[1] *Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden, lena.goze@slu.se, magnus.ekstrom@slu.se, saskia.sandring@slu.se, jorgen.wallerman@slu.se, goran.stahl@slu.se*

[2] *Department of Statistics, USBE, Umeå University, Umeå, Sweden, magnus.ekstrom@slu.se*

[3] *Department of Natural Sciences, Design and Sustainable Development, Mid Sweden University, Sundsvall, Sweden, Bengt-Gunnar.Jonsson@miun.se*

Monitoring of plant populations has become more and more important, especially in the current context of environmental change. In this paper, we propose methods to estimate plant density from presence/absence surveys. Presence/absence sampling is a useful and relatively simple method for monitoring state and change of plant communities. Moreover, it has advantages compared to traditional plant cover assessment, the latter being more prone to surveyor judgement error. We present a hybrid estimation framework, that combines model- and design-based inference features, in which a generalised linear model (for binary presence/absence data) and an inhomogeneous Poisson model (for plant locations) are used to estimate plant density in an area of interest. We look at two different cases, the first one with a known area of interest and the second one where the area is unknown and must be estimated. Our methods are applied to real data on *Vaccinium vitis-idaea* from the Swedish National Forest Inventory as well as simulated data to assess the performance of our estimators of plant density and corresponding variance estimators. The results obtained are promising and indicate that this method has a potential to add considerable analytic strength to monitoring programmes that collect presence/absence data.

A sparse matrix formulation of model-based ensemble Kalman filter

Håkon Gryvill[1], Håkon Tjelmeland[1]

[1] *Norwegian University of Science and Technology, Trondheim, hakon.gryvill@ntnu.no, haakon.tjelmeland@ntnu.no*

The ensemble Kalman filter (EnKF) introduced in [1] has been used successfully in many applications to do data assimilation in state-space models. EnKF represents the knowledge about the latent process in the state-space model through a set of realisations. The filter alternates between a prediction step and an update step. The prediction step can be performed analytically, while the update step must be approximated. In [2] a Bayesian model is to used to perform the update step.

Say we want to update the m'th realisation according to the Bayesian model. We assume that the m'th realisation and the state variable are iid and Gaussian, that our data comes from a linear Gaussian observation model, and that the model parameters are assigned conjugate priors. Moreover, [2] derive a set of updating procedures that are valid according to this model, and finds which of these procedures that are most robust against model assumptions made. However, the procedure introduced in [2] is computationally demanding. We therefore propose a computationally efficient approximation of this filter. Simulation studies suggest that the reduction in computational demands is considerable and that the approximation error in the computationally efficient approach is small compared to the overall variability.

References

[1] Evensen, G., Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Journal of Geophysical research*, 99, (1994) 10143-10162

[2] Loe, M. K., Tjelmeland, H., A generalised and fully Bayesian ensemble updating framework, *technical report*, https://hdl.handle.net/11250/2777303, (2021)

Disentangling the importance of population behavior and seasonality for SARS-CoV-2 spread: a case study of Norway and Sweden

Felix Günther[1], Hilde Kjelgaard Brustad[2], Arnoldo Frigessi[2] and Tom Britton[1]

[1] *Department of Mathematics, Stockholm University, Sweden, felix.gunther@math.su.se, tom.britton@math.su.se*

[2] *Department of Biostatistics, University of Oslo, Norway, h.k.brustad@medisin.uio.no, arnoldo.frigessi@medisin.uio.no*

Voluntary contact reduction, government recommendations and non-pharmaceutical interventions led to substantial behavioral changes during the SARS-CoV-2 pandemic.

In theory, it is clear that reductions in contacts lead to reduced disease spread. However, it is of interest to understand and demonstrate empirically how behavioral changes in different settings are associated with disease spread. Key challenges are the complexity of measuring population activity and the existence of different interrelated driving factors of infection dynamics.

A new source of information on population activity during the pandemic was the Google COVID-19 community mobility reports. We propose a principled model to assess the relationship between these data and the transmissibility of SARS-CoV-2 using an extended SIR model for representing disease spread. We assess the effect of population behavior in different settings, as well as the role of seasonality and virus variants. We account for age-prioritised vaccination and fit our model to weekly hospital admissions from February 2020 to July 2021 on the county level.

Our results show that mobility reports explain 25-75% of the changes in contacts depending on the county - tending to be more in Swedish counties than in Norwegian ones. The transit station and workplace setting appear to play the largest role. Furthermore, we estimate that transmissibility is reduced by around 40% in summer compared to winter.

A copula-based extension of logistic regression

Ingrid Hobæk Haff[1] and Simon Boge Brant[1]

[1] *University of Oslo, Norway, ingrihaf@math.uio.no, simonbb@math.uio.no*

Discriminative modelling, that aims to model $Y|\mathbf{X} = \mathbf{x}$, Y being binary and $\mathbf{X}$ a covariate vector, may be done with a discriminative or a generative approach, such as logistic regression or Naive Bayes, respectively, the main difference being how the parameters are estimated. The generative approach performs best when the assumed model is (close to) correct, whereas the discriminative one is more robust to misspecification [1, 2]. If the covariate distribution is complex, it is unrealistic to specify a tractable model that is close to correct, whereas the model-robust logistic regression with only linear main effects may be too simplistic. We propose to extend the simple logistic regression model to account for non-linearities and interaction effects by specifying the model on generative form, with a combination of certain marginal distributions and a vine copula [3, 4], accounting for the dependence, for each of the two classes. However, parameters are estimated as in the discriminative framework, maximising the likelihood of $Y|\mathbf{X} = \mathbf{x}$. Further, dependence between covariates is only included when it differs sufficiently in the two classes. Hence, the purpose is not to find the best models for the covariates in the two classes, but the model that discriminates between the two classes the best.

References

[1] Efron, B., The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis, *JASA*, 70, (1975) 892–898.

[2] O'Neill, T.J., The General Distribution of the Error Rate of a Classification Procedure with Application to Logistic Regression Discrimination, *JASA*, 75, (1980), 154–160.

[3] Bedford, T. and Cooke, R.M., Vines: A new graphical model for dependent random variables, *Annals of Statistics*, 30, (2002), 1031–1068.

[4] Aas, K. and Czado, C. and Frigessi, A. and Bakken, H., Pair-copula constructions of multiple dependence, *Insurance: Mathematics and economics*, 44, (2009), 182–198.

AI risk and AI alignment in the light of 2023 developments

Olle Häggström[1]

[1] *Chalmers University of Technology, olleh@chalmers.se*

We are at a crucial moment in human history, when we are automating and offloading to machines the one key asset that has brought our tremendous success so far: our intelligence. What could possibly go wrong? A lot. But there is hope: the field that has become known as AI alignment aspires to make sure that advanced AI has goals and values aligned with ours and compatible with human flourishing. I will discuss some challenges in AI alignment in the context of the large language models and other AI that are currently being developed and deployed. I abstain here from further detail which would likely be misleading, as the breakneck speed of these developments would likely overthrow most such detail between March 30 (abstract deadline) and June 19 (conference start).

Monte Carlo variance reduction with the Pushed Poisson point process

Diala Hawat[1 2], Raphaël Lachièze-Rey[2] and Rémi Bardenet[1]

[1] *Université de Lille, CNRS, Centrale Lille UMR 9189 – CRIStAL, F-59000 Lille, France, diala.hawat@univ-lille.fr*

[2] *Université de Paris Cité, MAP5, Paris, France*

The Monte Carlo method is a widely used computational technique for estimating the integral of a function by summing its values at points from a Poisson point process. Recently, an alternative approach using Determinantal Point Processes (DPP) has garnered attention due to its reduced variance compared to the classical Monte Carlo method. However, the high computational cost of sampling from a DPP makes it impractical for some applications.

To overcome this limitation, we introduce a one-step dynamic that moves the points of a homogeneous Poisson point process while maintaining a sampling complexity that is lower than that of a DPP. We proved that our proposed approach results in a Monte Carlo variant with reduced variance, compared to the classical Monte Carlo method. We conducted several numerical experiments to evaluate the effectiveness of our method and compared our results with other Monte Carlo variants. Our proposed Monte Carlo method offers a promising variance reduction method with a tractable computational time.

Weak pattern convergence for SLOPE and its robust versions

Ivan Hejný[1], Jonas Wallin[1] and Małgorzata Bogdan[1,2]

[1] *Department of Statistics, Lund University, ivan.hejny@stat.lu.se, jonas.wallin@stat.lu.se*
[2] *Institute of Mathematics, University of Wroclaw, malgorzata.bogdan@stat.lu.se*

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.

Joint default models via generalized Cox frameworks

Alexander Herbertsson[1]

[1] *University of Gothenburg, alexander.herbertsson@cff.gu.se*

We develop a continuous time credit portfolio model which allows for joint (simultaneous) defaults among obligors in the portfolio called the secondary group. The joint default model is created by letting the default times τ_i for obligors in the secondary group be the first time an increasing stochastic process Λ_t^i passes an exponential random variable, and where the increasing process Λ_t^i consist of an increasing absolute continuous term and a compound point process. The jumps in the compound point process are positive random variables that occur at the default times in an exogenously given group of defaultable entities, called the primary group. The default times in the primary group can come from any type of credit portfolio model as long as they are independent of the processes that drives the absolute continuous part in Λ_t^i for each obligor in the the secondary group. This construction will allow simultaneous defaults in the secondary group, that is $P[\tau_i = \tau_j] > 0$. We consider both the case with arbitrary positive jump terms and exchangeable jump terms in the compound point process. We then derive both the conditional and unconditional survival distributions for the default times τ_i in the secondary group and find explicit expressions for the joint default probabilities $P[\tau_i = \tau_j]$. Furthermore, we study the multivariate distributions for the default times in the secondary group. Several numerical examples are given, and we in particular consider the valuation of different credit related quantities in such models.

How many were killed in Guatemala, 1978-1996?

Nils Lid Hjort[1]

[1] *University of Oslo, Norway, nils@math.uio.no*

There are no complete lists given us the names or numbers of the too many tens of thousands of persons killed during the Guatemala civil wars, 1978-1996. I develop methods for estimating this total number, along with a confidence curve, based on the incomplete information from three official lists (REMHI, CEH, CIIDH, with acronyms reflecting their Spanish language names). The statistical challenge is to count the uncounted, the size of the zero-box. This is related to estimating the probability p_0 in a multinomial setup for counts $(N_0, N_1, \ldots, N_k)$, with probabilities $(p_0, p_1, \ldots, p_k)$, but where only $N_1, \ldots, N_k$ are observed. Can we estimate how many words Shakespeare actually knew, based on counting and categorising those found in his collective writing?

Matrix-analytical sampling theory for the infinite sites model

Asger Hobolth[1], Andreas Futschik[2], Simon Boitard[3] and Raphael Leblois[4]

[1] *Aarhus University, asger@math.au.dk*
[2] *Johannes Kepler University, andreas.futschik@jku.at*
[3] *Centre de Biologie et Gestion des Populations, Montpellier, simon.boitard@inrae.fr*
[4] *Centre de Biologie et Gestion des Populations, Montpellier, raphael.leblois@inrae.fr*

The infinite sites model can be characterized as a two-step procedure: A coalescent process that describes the ancestral relatedness of the samples, and a sprinkling of mutations in separate sites on the ancestral tree according to a Poisson process. The coalescent process can be defined using multivariate phase-type theory. The requirements are a rate matrix that determines the coalescent rates between the ancestral states, an initial state probability vector, and a reward matrix that informs about the characteristics of the ancestral states. For example, this information could be the number of singleton, doubleton or higher-order lineages in the state or a topological tree property such as which leaves have coalesced. In this paper, we analyse the probability generating function for the infinite sites model as a function of the initial state probability vector, the coalescent rate matrix, the reward matrix, and the mutation rate. We describe a general matrix-analytical method for calculating the probability of a population genetic data set. We demonstrate that the method is computationally feasible for a small number of mutations and provide a simple and easy-to-implement algorithm for determining the probability of a sample from the infinite sites model. The method only involves a single inverse matrix operation, matrix multiplication, and matrix addition. We convey comprehensive understanding of the procedure by detailed calculations and discussions of several elementary examples. These examples include the site frequency spectrum, a fully labelled sample, and a two-island model. Finally, we apply the procedure to calculate probabilities of site frequency spectra with a given number of segregating sites and mutation rate, and compare the probabilities of the spectra to a ranking based on Tajimas' D-statistics.

The asymptotic distribution of cluster sizes for supercritical percolation on random split trees

Cecilia Holmgren[1] and Gabriel Berzunza [2]

[1] *Uppsala University, cecilia.holmgren@math.uu.se*
[2] *University of Liverpool, Gabriel.Berzunza-Ojeda@liverpool.ac.uk*

This talk is based on our paper [1]. We consider the model of random trees introduced by Devroye [2], the so-called random split trees. The model encompasses many important randomized algorithms and data structures. We then perform supercritical Bernoulli bond-percolation on those trees and obtain the asymptotic distribution for the sizes of the largest clusters.

References

[1] Berzunza, G. and Holmgren, C., The asymptotic distribution of cluster sizes for supercritical percolation on random split trees. *Random Structures and Algorithms*, 60, (2022) 631–652.

[2] Devroye, L., Universal limit laws for depth in random trees. *SIAM J. Comput.*, 28, (1998) 409–432.

Causal change point detection and localization

Shimeng Huang, Rikke Søndergaard, Jonas Peters and Niklas Pfister

University of Copenhagen, Denmark,
{shimeng, rsn, jonas.peters, np}@math.ku.dk

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.

Fractional Polynomials Models as Special Cases of Bayesian Generalized Nonlinear Models

Aliaksandr Hubin[1,2], Georg Heinze[3] and Riccardo De Bin[2]

[1] *NMBU, Ås, aliaksandr.hubin@nmbu.no*
[2] *University of Oslo, Norway, aliaksah@math.uio.no, debin@math.uio.no* [3] *Medical Unversity of Vienna, georg.heinze@meduniwien.ac.at*

We propose a framework for fitting fractional polynomials models as special cases of Bayesian Generalized Nonlinear Models [1], using an adapted version of the Genetically Modified Mode Jumping Markov Chain Monte Carlo algorithm [2] to fit them. The broad generality of the Bayesian Generalized Nonlinear Models allows us to extend the use of a Bayesian version of the fractional polynomials [4] models to any supervised learning task, including regression, classification and time-to-event data analysis. We show through a simulation study that our novel approach outperforms the original version of Bayesian fractional polynomials [4] and performs similarly to the classical frequentist fractional polynomials [3] approach in terms of variable selection, identification of the correct functional forms and prediction ability, while providing, in contrast to its frequentist version, a coherent inference framework. Real data examples provide further evidence in favour of our approach and show its flexibility.

References

[1] Hubin, A., Storvik, G., and Frommlet, F. (2021). Flexible Bayesian nonlinear model configuration. *Journal of Artificial Intelligence Research*, 72:901–942.

[2] Hubin, A., Storvik, G., and Frommlet, F. (2020a). A novel algorithmic approach to Bayesian logic regression (with discussion). *Bayesian Analysis*, 15:263–333.

[3] Sauerbrei, W. and Royston, P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162:71–94.

[4] Sabanés Bové, D. and Held, L. (2011). Bayesian fractional polynomials. *Statistics and Computing*, 21:309–324.

Sequential Bayesian inference with intractable likelihoods

Henrik Häggström[1], Umberto Picchini[2]

[1] *Chalmers University of Technology, Gothenburg, henhagg@chalmers.se*

[2] *Department of Mathematical Sciences, Chalmers University of Technology & University of Gothenburg, Gothenburg, picchini@chalmers.se*

Likelihood-free methods perform parameter inference in models where evaluating the likelihood is intractable but sampling data from a generative model is possible. Recent approaches such as SNPE-C [1] learn the posterior by sequential updates of neural-network based density estimators. While these methods perform well they require a network architecture to be specified and training the neural-network can be computationally demanding and time consuming. In this work we present a Bayesian inference method which, in place of neural networks, uses Gaussian mixtures sequentially learned through an expectation-maximization procedure. Posterior samples are then obtained via MCMC through an informative and self-tuned proposal sampler. Only the number of components in the Gaussian mixture needs to be specified to run the algorithm. We show the feasibility of this method and benchmark it against state-of-the-art Bayesian methods in several simulation studies.

References

[1] Greenberg, D., Nonnemacher, M. and Macke, J., Automatic Posterior Transformation for Likelihood-Free Inference, *Proceedings of the 36th International Conference on Machine Learning* in *Proceedings of Machine Learning Research* 97:2404-2414, 2019.

Optimal subsampling designs for analysis of massive datasets and measurement-constrained experiments

Henrik Imberg, Marina Axelson-Fisk and Johan Jonasson

Chalmers University of Technology and University of Gothenburg, Gothenburg
imbergh@chalmers.se, marina.axelson-fisk@chalmers.se, jonasson@chalmers.se

Subsampling is commonly used to overcome computational and economical bottlenecks in the analysis of finite populations and massive datasets. Existing methods are often limited in scope and use optimality criteria (e.g., A-optimality) with well-known deficiencies, such as lack of invariance to the measurement-scale of the data and parameterisation of the model. We present a general theory of optimal design for data subsampling problems, including finite population inference, parametric density estimation, and regression modelling. Our theory encompasses and generalises most existing methods in the field of optimal subdata selection based on unequal probability sampling and inverse probability weighting. We derive optimality conditions for a general class of optimality criteria, and present corresponding algorithms for finding optimal sampling schemes under Poisson and multinomial sampling designs. We present a novel class of transformation- and parameterisation-invariant linear optimality criteria which enjoy the best of two worlds: the computational tractability of A-optimality and invariance properties similar to D-optimality. The methodology is illustrated on an application in the traffic safety domain. In our experiments, the proposed invariant linear optimality criteria achieve 90–99% D-efficiency with 90–95% lower computational demand. In contrast, the A-optimality criterion has only 46% and 60% D-efficiency on two of the examples.

An active sampling framework for finite population inference with optimal subsamples

Henrik Imberg[1], Xiaomi Yang[2], Carol Flannagan[2,3] and Jonas Bärgman[2]

[1] *Chalmers University of Technology and University of Gothenburg, imbergh@chalmers.se*
[2] *Chalmers University of Technology, Gothenburg, xiaomi@chalmers.se, bargman@chalmers.se*
[3] *University of Michigan, Ann Arbor, cacf@umich.edu*

Data subsampling has become widely recognised as a tool to overcome computational and economic bottlenecks in analysing massive datasets and measurement-constrained experiments. However, traditional subsampling methods often suffer from the lack of information available at the design stage. We propose an active sampling strategy that iterates between estimation and data collection with optimal subsamples, guided by machine learning predictions on yet unseen data. Using a martingale central limit theorem, we establish consistency and asymptotic normality of the active sampling estimator of a finite population characteristic, under mild assumptions on the sampling design. Consistent variance estimators are derived, and our theoretical results confirmed empirically in experiments. The method is illustrated for an application on scenario generation for virtual safety assessment of an advanced driver assistance system. Substantial improvements over traditional importance sampling methods are demonstrated, with sample size reductions of 10–50% for the same level of performance in terms of the root mean squared error of the estimator.

Efficient implementation of additive Gaussian process models using Kronecker method

Sahoko Ishida[1], and Wicher Bergsma[1]

[1] *London School of Economics and Political Science, London, s.ishida@lse.ac.uk, w.p.bergsma@lse.ac.uk*

This project presents an efficient implementation of additive Gaussian process (GP) models using Kronecker method, applicable when the dataset has a multidimensional grid structure. While this method has gained popularity in the GP literature, the existing approach is limited to when the covariance function of the process has a tensor product structure, which implies a model with only the highest interaction, or a saturated model. We extend the method to a more general class of additive GP models. More specifically, we propose a hierarchical ANOVA kernel that combines tensor product and additive kernels to allow for more flexible modeling and straightforward interpretation. By using a centered kernel in constructing the hierarchical ANOVA kernel, we demonstrate that the Kronecker product structure can still be exploited for significant computational gain. This approach has broad applications in time series, spatial analysis, image analysis, and repeated measurements in medical science and psychology. Our contribution to the field is providing a more flexible and interpretable approach to Gaussian process modeling for large-scale data, allowing researchers to gain deeper insights into the underlying data structure. We illustrate the proposed approach by analyzing high-frequency NO2 concentrations data collected in London during the COVID-19 lockdown period.

A flexible likelihood-based neural network extension of the classic spatio-temporal model

Malte Jahn[1]

[1] *Helmut-Schmidt University, Hamburg, jahnma@hsu-hh.de*

A new framework for combining spatio-temporal regression techniques and artificial neural network (ANN) regression models is presented. The key idea is to use the universal approximation property of the ANN function to account for an arbitrary spatial pattern in the dependent variable by including geographic coordinate variables as regressors. Moreover, the implicit location-specific effects are allowed to exhibit arbitrary interaction effects with other regressors such as a time variable. In contrast to most machine learning approaches for spatio-temporal data, the likelihood framework of the classic (linear) spatio-temporal model from [1] is preserved. This allows, inter alia, for inference regarding marginal effects and associated confidence. The framework also allows for non-normal conditional distributions, conditional spatial correlation, arbitrary trend and seasonality. These features are demonstrated in a simulation study and two data examples, using linear spatio-temporal models as a reference.

References

[1] LeSage, J., Pace, R.K. Introduction to spatial econometrics, Chapman & Hall/CRC, Taylor & Francis, 2009.

The number of descendants in a random directed acyclic graph

Svante Janson

Uppsala University

Consider a random directed acyclic graph, obtained by recursively adding vertices, where each new vertex has a fixed outdegree d and the endpoints of the d edges from it are chosen uniformly at random among previously existing vertices. For simplicity we assume $d = 2$.

We study the set of vertices that are descendants of a given vertex n (asymptotically, for large n). The main result is that the number of such vertices, divided by $\sqrt{n}$, converges in distribution together with all moments. The limit distribution is, up to a constant factor, a chi distribution $\chi(4)$.

The talk is based on a recent preprint [1].

References

[1] Svante Janson. The number of descendants in a random directed acyclic graph. `arXiv:2302.12467`.

Generating Gaussian random fields using surface finite element methods

Erik Jansson[1], Annika Lang [1] and Mike Pereira[2]

[1] *Chalmers University of Technology & University of Gothenburg, Gothenburg, erikjans@chalmers.se, annika.lang@chalmers.se*
[2] *Mines Paris - PSL University, Fontainebleau, mike.pereira@mines-paristech.fr*

In the middle of the previous century, Peter Whittle demonstrated that Whittle–Matérn Gaussian random fields on Euclidean domains can be obtained as solutions to a fractional elliptic stochastic partial differential equation (SPDE). The SPDE–Random field connection can be leveraged to generate random fields on other domains, such as manifolds — by solving an SPDE on the domain. We consider how the computational technique of surface finite elements can be utilized to generate random fields on surfaces. This talk is based on joint work with Annika Lang and Mike Pereira.

Point process learning for Gibbs process models

Julia Jansson[1], Ottmar Cronie[2]

[1] *Department of Mathematical Sciences, Chalmers University of Technology & University of Gothenburg, Sweden, juljans@chalmers.se*

[2] *Department of Mathematical Sciences, Chalmers University of Technology & University of Gothenburg, Sweden, ottmar@chalmers.se*

In classical statistics, there is usually an underlying i.i.d. assumption. However, in the field of spatial and temporal statistics we often have dependencies for example in forestry, seismology and epidemiology. A point process may be viewed as a generalized random sample where we allow dependence and a random sample size. Recently in [1] the authors proposed a prediction-based statistical theory called point process learning (PPL). It is based on the combination of two novel concepts for general point processes: cross-validation and prediction errors.

We here explore how PPL can be exploited to model a range of common Gibbs point process models, like the hard-core process, the Strauss process and the Geyer saturation process. Through a simulation study, we compare the performance of PPL in terms of MSE and MAPE to pseudo-likelihood methods, which is the state of the art in the context of the studied models. We found that PPL outperforms pseudo-likelihood for all considered models. In addition, we propose a procedure to obtain resample-based confidence regions for the model parameters. The empirical coverage was studied for these so-called PPL-confidence intervals, and we found that there is merit to the confidence region idea.

References

[1] Ottmar Cronie, Mehdi Moradi, and Christophe AN Biscio. Statistical learning and cross-validation for point processes. *arXiv preprint arXiv:2103.01356*, 2021.

Latent process location-scale mixed-effects model for intensively sampled ordinal patient reported outcomes

Alexandra Jauhiainen[1], Agnieszka Król[2], Jacob Leander[3], Robert Palmér[3] and Cécile Proust-Lima[4]

[1] *BioPharma Early Biometrics and Statistical Innovation, Data Science & AI, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden, alexandra.jauhiainen@astrazeneca.com*

[2] *BioPharma Early Biometrics and Statistical Innovation, Data Science & AI, BioPharmaceuticals R&D, AstraZeneca, Warsaw, Poland, agnieszka.krol@astrazeneca.com*

[3] *Clinical Pharmacology and Quantitative Pharmacology, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg, Sweden, jacob.leander@astrazeneca.com, robert@binarybrains.se*

[4] *Biostatistics Team, INSERM CR1219, University of Bordeaux, Bordeaux, France, cecile.proust-lima@u-bordeaux.fr*

Patient reported outcomes (PROs) are collected on a daily basis in clinical trials to measure patients' quality of life, e.g. symptoms. Often these data are reported in a small range ordinal scale and analyzed without considering their longitudinal aspect.The emergence of electronic data collection methods for home-based measurements has enabled routine capture of various symptom scores such as breathlessness in respiratory diseases, highlighting the need to develop statistical methods for the analysis of these intensive ordinal longitudinal data. Indeed, both the level, long-term fluctuations and short-term variability of these outcomes are expected to be linked to the disease progression and to be affected by treatment. To model the dynamics of ordinal PROs, we propose a location-scale probit mixed-effects model which includes two types of variability: individual trajectories alowing for flexibiility, e.g. using splines and the short-term variability with an error where the variance is expressed as a linear structure of covariates such as treatment arm and a patient-specific random intercept. The model is estimated in the Maximum Likelihood framework with an interface in R. The burdensome high-dimensional intractable integrals in the optimization are approximated using the Quasi-Monte Carlo method. We apply the methodology to a randomized Phase III COPD clinical trial to evaluate the effect of treatment on dynamics of the respiratory symptoms and their variability.

Nearly Optimal Latent State Decoding in Block MDPs

Yassir Jedra[1], based on joint work with Juhnghyun Lee[2], Alexandre Proutiere[1], and Se-Young Yun[2]

[1] *KTH Royal Institute Of Technology, Stockholm, jedra@kth.se, alepro@kth.se*
[2] *Kim Jaechul Graduate School of AI, KAIST, Seoul, jh_lee00@kaist.ac.kr, yunseyoung@kaist.ac.kr*

We investigate the problem of model estimation in episodic Block MDPs. In these MDPs, the decision maker has access to rich observations or contexts generated from a small number of latent states. Our objective is to recover the latent state decoding function (the mapping from the observations to latent states) based on data gathered under a fixed behavior policy. We provide an instance-dependent information-theoretical lower bound on the error rate for estimating this function and devise an algorithm with a performance approaching this fundamental limit. As a byproduct of our algorithm, we also provide estimates of the MDP dynamics. We further apply our results to the problem of learning near-optimal policies in the reward-free setting. Based on our efficient model estimation algorithm, we show that we can infer a policy converging (as the number of collected samples grows large) to the optimal policy at the best possible asymptotic rate. Our analysis provides necessary and sufficient conditions under which exploiting the block structure yields improvements in the sample complexity for identifying near-optimal policies. When these conditions are met, the sample complexity in the minimax reward-free setting is improved by a multiplicative factor n, where n is the number of contexts.

A Spatial-statistical model to analyse historical rutting data

Natoya O.A.S. Jourdain[1], Ingelin Steinlsland [1], Mamoona Birkhez Shami [1], Alex Klein-Paste [1], Dagfin Gryteselv [2] and Doreen Siebert [2]

[1] *Norwegian University of Science and Technology (NTNU), Trondheim, Norway, natoya.jourdain@ntnu.no, ingelin.steinsland@ntnu.no, mamoona.b.shami@ntnu.no, alex.klein-paste@ntnu.no*

[2] *Statens Vegvesen, Trondheim, Norway, dagfin.gryteselv@vegvesen.no, doreen.siebert@vegvesen.no*

Pavement rutting is a major problem in flexible pavements. It may occur at the pavement infrastructure's various layers (surface or subgrade) and stages, and it is one of the major causes of why pavements need asphalt resurfacing. We propose, fit and evaluate statistical models for pavement rutting in a Bayesian framework of linear mixed models with spatial components. The annual average daily traffic, referred to as traffic intensity, pavement type, rut depth, road width and slope height are included in spatial latent Gaussian models to account for and estimate the contribution of annual observations to rutting. The proposed models quantify the uncertainty and identify locations potentially in the greatest need of maintenance. The models are fitted to eleven years (2010-2020) of data collected from the European route E14 — Stjørdal, Norway to Storlien on the Swedish border. We find that pavement type together with traffic intensity is the main driver of rutting. Further, there are spatial dependencies and road stretches with more than 1 millimeter rutting than expected annually. This means that the expected lifetime is halved at these stretches (at least more than two years). We provide maps with expected rutting, and some locations have been identified for accelerated rutting, with reduction in pavement life expectancy of at least 10 years.

Approximate Bayesian Computation with Backward Simulation for Discretely Observed Diffusions

Petar Jovanovski[1], Andrew Golightly[2] and Umberto Picchini[1]

[1] *Department of Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, Sweden, petarj@chalmers.se, picchini@chalmers.se*
[2] *Department of Mathematical Sciences, Durham University, UK, andrew.golightly@durham.ac.uk*

Stochastic differential equations (SDE) are employed in many areas of science as a powerful tool for modelling processes that are subject to random fluctuations. Bayesian inference for a large class of SDEs is challenging due to the analytic intractability of the likelihood function. Nevertheless, forward simulation via numerical methods is straightforward, motivating the use of approximate Bayesian computation (ABC). We propose a simulation scheme for SDE models that is based on processing the observation in both the forward and backward direction, effectively utilizing the information provided by the observed data. This leads to the simulation of sample paths that are consistent with the observations, by building on [2], thereby increasing the ABC acceptance rate. We additionally leverage partial exchangeability of Markov processes and employ invariant neural networks to learn the summary statistics that are needed in ABC, following [1]. These are sequentially learned by exploiting a sequential Monte Carlo ABC sampler, which provides new training data at each iteration. Therefore, our novel contribution is a learning tool for SDE model parameters while simultaneously learning the summary statistics. Using synthetic data generated from the Chan-Karaolyi-Longstaff-Sanders SDE family, we show that our approach accelerates inference considerably, compared to standard (forward-only) methods, while preserving inference accuracy.

References

[1] Wiqvist, S., Mattei, P. A., Picchini, U., & Frellsen, J. (2019, May). Partially exchangeable networks and architectures for learning summary statistics in approximate Bayesian computation. In *International Conference on Machine Learning* (pp. 6798-6807). PMLR.

[2] Del Moral, P., Murray, L. M. (2015). Sequential Monte Carlo with highly informative observations. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1), 969-997.

Online mass matrix adaptation for Hamiltonian Monte Carlo

Miika Kailas[1] and Matti Vihola[1]

[1] *University of Jyväskylä, Jyväskylä, miika.p.kailas@jyu.fi, matti.s.vihola@jyu.fi*

We consider adaptive Markov Chain Monte Carlo methods within the Hamiltonian Monte Carlo (HMC) sampler and its dynamic variant, the No U-Turn Sampler (NUTS). In particular we study strategies for full-rank mass matrix adaptation and make two primary contributions. First, we study regularization strategies for online estimates relating to full-rank mass matrix adaptation in HMC and variants. Second and more importantly, we propose a novel adaptation target for the mass matrix. Contrasting with the usual choice of choosing the mass matrix as the inverse of (an estimate of) the covariance matrix of the target distribution, a global quantity, our alternative proposal is instead an average over local geometric quantities relating to the stability of discretized Hamiltonian dynamics. The proposed target and its estimators are computationally cheap and simple to implement, and our empirical studies show that the proposed adaptation strategies are applicable to challenging problems in hundreds of dimensions.

Title: Adapted Wasserstein distance between the laws of SDEs

Sigrid Källblad[1], Julio Backhoff-Veraguas[2] and Ben Robinson[2]

[1] *KTH Royal Institute of Technology, Stockholm, sigrid.kallblad@math.kth.se*
[2] *University of Vienna, Vienna, julio.backhoff@univie.ac.at, robinsonb20@univie.ac.at*

We consider an adapted optimal transport problem between the laws of Markovian stochastic differential equations (SDEs) and establish optimality of the so-called synchronous coupling between the given laws. The proof of this result is based on time-discretisation methods and reveals an interesting connection between the synchronous coupling and the celebrated discrete-time Knothe–Rosenblatt rearrangement. We also provide a related result on equality of various topologies when restricting to certain laws of continuous-time stochastic processes. The result is of relevance for the study of stability with respect to model specification in mathematical finance.

Unbiased Estimation of Phylogenetic Multivariate Brownian Motion and Ornstein-Uhlenbeck Models

Hao Chi Kiang[1]

[1] *Linköping university, Sweden, hao.chi.kiang@liu.se*

Brownian Motion and Ornstein-Uhlenbeck models along phylogeneies are fundamental building blocks of phylogenetic comparative methods, in which these stochastic process models are typically used to model trait evolution [1][2]. The standard practice is to compute their maximum likelihood estimates; other estimation methods are much less explored. My on-going work investigates unbiased estimation and linear-algebraic properties of these models.

The analysis reveals the exact distribution of the maximum-likelihood estimator of the phylogenetic Brownian Motion models' covariance matrix and a novel unbiased estimator of the Ornstein-Uhlenbecks' covariance matrix when the drift rate is known, along with some other linear algebraic properties of the models.

References

[1] J. Felsenstein, 'Phylogenies and the Comparative Method', The American Naturalist, vol. 125, no. 1, pp. 1–15, 1985.

[2] T. F. Hansen, 'Stabilizing Selection and the Comparative Analysis of Adaptation', Evolution, vol. 51, no. 5, pp. 1341–1351, 1997.

Cross-Validation Based Adaptive Sampling for Multi-Level Gaussian Process Models

Louise Kimpton, Peter Challenor, Tim Dodwell, James Salter and Hossein Mohammadi

University of Exeter, UK, l.m.kimpton@exeter.ac.uk

Uncertainty quantification is a vital part of decision making, which often makes use of complex computer codes (or digital twins). Such computer codes, or models, can often be run in a hierarchy of different levels of fidelity ranging from the basic to the sophisticated. The top levels in this hierarchy are expensive to run, limiting the number of possible runs. To make use of runs over all levels, and crucially improve emulation at the top level, we use multi-level Gaussian process emulators (GPs). The accuracy of the GP depends on the design of the training points. We present a multi-level adaptive sampling algorithm to sequentially increase the set of design points to optimally improve the fit of the GP. The normalised expected leave-one-out cross-validation error is calculated at all unobserved locations, and a new design point is chosen using expected improvement and a repulsion function. This criterion is calculated for each model level weighted by the cost for running the code at that level. Hence, at each iteration, our algorithm optimises for both the new point location and the level the model is run at. We provide examples, and show how the algorithm can be adapted for batch designs.

Global tests for quantile regression with applications in modeling distributions

Konstantinos Konstantinou[1], Tomáš Mrkvička[2], Mari Myllymäki[3] and Mikko Kuronen[3]

[1] *Chalmers University of Technology and University of Gothenburg, Sweden, konkons@chalmers.se*

[2] *University of South Bohemia, České Budějovice, mrkvicka.toma@gmail.com*

[3] *Natural Resources Institute Finland (Luke), Finland, mari.myllymaki@luke.fi, mikko.kuronen@luke.fi*

Statistical tests allowing for simultaneous inference of the quantile regression process [1], with graphical interpretation are proposed. The tests are based on the global envelope tests [2], initially developed in spatial statistic for solving the multiple testing problem. The proposed global quantile regression (permutation) tests can determine not only if there is a difference, but it can also determine for which quantiles the difference is significant on the global significance level. The case where the effect of a factor (e.g., a categorical factor giving the group) on the distribution functions is of interest but confounded with other factors affecting the distributions is studied. An extensive simulation study is conducted to compare the global quantile regression tests with classical graphical tests based on the Kolmogorov-Smirnov test statistic.

References

[1] Koenker, Roger, and Gilbert Bassett Jr. ”Regression quantiles.” Econometrica: Journal of the Econometric Society (1978): 33-50.

[2] Myllymäki, Mari, et al. ”Global envelope tests for spatial processes.” Journal of the Royal Statistical Society. Series B (Statistical Methodology) (2017): 381-404.

Mathematical modeling of nerve mortality caused by diabetic neuropathy

Konstantinos Konstantinou[1], Farnaz Ghorbanpour[2], Umberto Picchini[1], Adam Loavenbruck[3], and Aila Särkkä[1]

[1] *Chalmers University of Technology and University of Gothenburg, Sweden, konkons@chalmers.se, picchini@chalmers.se, aila@chalmers.se*
[2] *Allameh Tabataba'i University, Tehran, Iran, farnazghorbanpour@yahoo.com*
[3] *Kennedy Laboratory, University of Minnesota, Minneapolis, Minnesota, loave001@umn.edu*

Diabetic neuropathy is a disorder characterized by impaired nerve function and reduction of the number of epidermal nerve fibers per epidermal surface. Additionally, as neuropathy related nerve fiber loss and regrowth progresses over time, the two-dimensional spatial arrangement of the nerves becomes more clustered. These observations suggest that with development of neuropathy, the spatial pattern of diminished skin innervation is defined by a thinning process which remains incompletely characterized. We regard samples obtained from healthy controls and subjects suffering from diabetic neuropathy as realisations of planar point processes consisting of nerve entry points and nerve endings, and propose point process models based on spatial thinning to describe the change as neuropathy advances. Initially, the hypothesis that the nerve removal occurs completely at random is tested using independent random thinning of healthy patterns. Then, a dependent parametric thinning model that favors the removal of isolated nerve trees is proposed. Approximate Bayesian computation is used to infer the distribution of the model parameters, and the goodness-of-fit of the models is evaluated using both non-spatial and spatial summary statistics. Our findings suggest that the nerve mortality process changes behaviour as neuropathy advances.

Invariant causal prediction for non-additive noise models

Lucas Kook, Sorawit Saengkyongam and Jonas Peters

University of Copenhagen, Denmark
lucasheinrich.kook@gmail.com, {jp, ss}@math.ku.dk

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`.

Multiple hypothesis testing from the change-point detection viewpoint

Anica Kostic[1], Piotr Fryzlewicz[1]

[1] *London School of Economics and Political Science, UK, a.kostic@lse.ac.uk, p.fryzlewicz@lse.ac.uk*

Estimating the proportion of false null hypotheses among independently tested hypotheses is a crucial problem in the multiple testing literature. By integrating a proportion estimator, most multiple testing methods that control the false discovery rate (FDR) can be made adaptive, leading to increased power while still maintaining control over the FDR. Although many proportion estimators have been proposed in the literature, the approach by Schweder and Spjøtvoll [1], more commonly known as Storey's estimator, is the most widely used due to its simplicity. We present a novel approach that tunes Storey's estimator using ideas from the change-point literature. Our approach aims to identify the change-point in the slope of the p-value plot to distinguish between true and false null hypotheses. The proposed method is fast, simple, and improves on existing tuning parameters. The theoretical results for our estimator are based on strong limit theorems for quantile processes. Our simulations demonstrate that our method outperforms existing methods, particularly in sparse cases with a small false null proportion. Additionally, we discuss the potential usefulness of our method in addressing the common issue of superuniform p-values, which typically causes Storey's estimator to fail.

References

[1] T. Schweder, E. Spjøtvoll, Plots of P-values to evaluate many tests simultaneously, *Biometrika*, 69 (3), 1982, 493–502,

Model validation for multivariate distributions with maximum spacing method

Kristi Kuljus[1] and Bo Ranneby[2]

[1] *University of Tartu, Estonia, kristi.kuljus@ut.ee*
[2] *Swedish University of Agricultural Sciences, Sweden, bo.ranneby@slu.se*

Maximum spacing (MSP) method is a parameter estimation method for continuous distributions based on spacings of the distribution functions. In the univariate case the observations can be ordered and spacings are defined through the nearest neighbour to the right of each observation. In the multivariate case there is no natural way of ordering observations and the probabilistic approach with nearest neighbour balls can be used to generalize spacings. Besides being a useful estimation method, the MSP method provides an important complementary tool to model selection and evaluation methods. Together with parameter estimation it gives a possibility to check whether the assigned model class is suitable. The MSP method can be applied with different information measures (e.g. the Kullback-Leibler information measure, the Hellinger distance, Jeffrey's divergence measure), which is essential in model validation context because parameter estimates obtained with different divergence measures have different properties. The talk is mainly based on articles [1] and [2].

References

[1] Kuljus, K. and Ranneby, B., Generalized maximum spacing estimation for multivariate observations, *Scandinavian Journal of Statistics*, 42, (2015) 1092-1108.

[2] Kuljus, K. and Ranneby, B., Asymptotic normality of generalized maximum spacing estimators for multivariate observations, *Scandinavian Journal of Statistics*, 47, (2020) 968-989.

Bayesian estimation of causal effects from observational discrete data

Vera Kvisgaard[1], Johan Pensar[2]

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

We address the problem of estimating causal effects and identifying the strongest pairwise causal relationships among a set of discrete variables based on observational data. 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][4][2][3]. 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].

References

[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.

LiquidCNA: tracking cancer evolution through therapy using copy number alterations

Eszter Lakatos[1,2], Helen Hockings[3], Maximilian Mossner[2], Michelle Lockley[3] and Trevor Graham[2]

[1] *Chalmers University of Technology, Gothenburg, Sweden; eszter.lakatos@icr.ac.uk*
[2] *Institute of Cancer Research, London, UK*
[3] *Barts Cancer Centre, London, UK*

Next-generation sequencing of cell-free DNA (cfDNA) obtained from blood samples ('liquid biopsies') is revolutionising cancer care. It is enabling longitudinal monitoring through therapy, so that we can now investigate the compositional changes that lead to treatment resistance. However, the state-of-the-art for assessing tumour composition is deep sequencing: looking for known variants specific to cancer cells that resist therapy. As this is expensive and requires high quality samples and knowledge of these variants, frequent cfDNA-based sampling/monitoring is still infeasible with current methods.

Here we present liquidCNA, a bionformatic algorithm to track cancer composition from cfDNA samples based on copy number alterations (CNAs). CNAs play a central role in cancer evolution and can be detected using cost-effective shallow genome sequencing even in low quality material with high contamination from non-tumour DNA. LiquidCNA tracks cancer burden and simultaneously genotype and quantify an emerging/resistant sub-population without requiring any prior knowledge of the cancer genetics.

We demonstrate the accuracy of liquidCNA in synthetically generated sample sets and *in vitro* and *in silico* mixtures of cancer cell lines. We apply liquidCNA to patients with ovarian cancer, and reveal the progressive emergence of a novel tumour sub-population indicative of disease progression.

The stochastic wave equation on the sphere: properties and simulation

Annika Lang[1] and David Cohen[2]

[1] *Chalmers & University of Gothenburg, Gothenburg, annika.lang@chalmers.se*
[2] *Chalmers & University of Gothenburg, Gothenburg, david.cohen@chalmers.se*

The stochastic wave equation driven by isotropic Gaussian noise is considered on the unit sphere. We solve this stochastic partial differential equation and discuss properties of the derived solutions. These are used in the developed approximation scheme based on spectral methods and its convergence analysis. We derive strong, weak, and almost sure convergence rates for the proposed algorithm and show that these rates depend only on the smoothness of the driving noise, the initial conditions, and the test functions. Numerical experiments confirm the theoretical rates. Finally we discuss extensions to more general domains and equations that can be treated in a similar way.

Confidence distributions for the autoregressive parameter

Rolf Larsson[1]

[1] *Uppsala University, Sweden, rolf.larsson@math.uu.se*

The notion of confidence distributions is applied to inference about the parameter in a simple autoregressive model, allowing the parameter to take the value one. This makes it possible to compare to asymptotic approximations in both the stationary and the non stationary cases at the same time. The main point, however, is to compare to a Bayesian analysis of the same problem. In the sense of Jeffreys, a non informative prior for a parameter is given as the ratio of the confidence density and the likelihood. In this way, the similarity between the confidence and non-informative Bayesian frameworks is exploited. It is shown that, in the stationary case, asymptotically the so induced prior is flat. However, if a unit parameter is allowed, the induced prior has to have a spike at one of some size. Simulation studies and two empirical examples illustrate the ideas.

Compressed sensing for low-count PET denoising in measurement space

Klara Leffler[1]

[1] *Umeå University, Umeå, klara.leffler@umu.se*

Low-count positron emission tomography (PET) data suffer from high noise levels, leading to poor image quality and reduced diagnostic accuracy. Compressed sensing (CS) based denoising methods have shown potential in medical imaging. This study investigates the performance of CS-based denoising methods on PET sinograms.

Three simulated datasets were used in this study, including circular phantom, patient pelvis phantom, and patient brain phantom. Ten sampling levels were employed to investigate the effect of data reduction on diagnostic accuracy. CS-based denoising methods were applied pre-reconstruction, and a conventional Gaussian post-filter was used for comparison. Performance measures included rRMSE, SSIM, SNR, line profiles, and FWHM.

Overall, the proposed CS-based denoising methods performed similarly to the benchmark in terms of lesion contrast, spatial resolution, and noise texture. The proposed methods outperformed the benchmark in low-count situations by suppressing background noise and preserving contrast better.

The results of this study demonstrate that CS-based denoising methods in the sinogram domain can improve the quality of low-count PET images, particularly in suppressing background noise and preserving contrast. These findings suggest that CS-based denoising could be a promising solution for improving the diagnostic accuracy of low-count PET data.

Network analysis in EEG data: A cointegration approach

Marie Levakova[1] and Susanne Ditlevsen[2]

[1] *University of Copenhagen, male@math.ku.dk*
[2] *University of Copenhagen, susanne@math.ku.dk*

EEG recordings are an invaluable source of information about brain activity. Our work aims to enrich the statistical toolbox with a methodology based on cointegration, which has been originally developed with econometrics applications in mind [1], however, the idea to use it in realm of oscillating systems in physics [2] and in neuroscience [3] has emerged recently. We assume that observations of EEG are an integrated vector autoregressive (VAR) process. The idea is to discern, which part of the trending behavior can be attributed to stochastic trends, and which part stems from long-term linear equilibrium relationships, termed cointegration relationships.

The estimation procedure provides estimates of several parameters useful for interpretation: the cointegration rank gives the number of independent cointegration relationships and the number of independent stochastic trends; the cointegration matrix contains coefficients of cointegration relationships; and the loading matrix describes how the system reacts to deviations from the cointegration relationships. Most importantly, the product of the loading matrix and the cointegration matrix describes the functional network structure of the channels. Several technical aspects specific to EEG data had to be resolved, such as dimension far larger than dimensions commonly encountered in previous applications of cointegration analysis.

References

[1] Granger, C. W., Some properties of time series data and their use in econometric model specification. *J. Econometrics*, 16(1), (1981) 121-130.

[2] Dahlhaus, R., Kiss, I. Z., Neddermeyer, J. C., On the Relationship between the Theory of Cointegration and the Theory of Phase Synchronization. *Statist. Sci.*, 33(3), (2018) 334-357.

[3] Østergaard, J., Rahbek, A., Ditlevsen, S., Oscillating systems with cointegrated phase processes. *J. Math. Biol.*, 75(4), (2017) 845-883.

[4] Levakova, M., Christensen, J. H., Ditlevsen, S., Classification of brain states that predicts future performance in visual tasks based on co-integration analysis of EEG data. *Roy. Soc. Open Sci.*, 9, (2022) 220621.

Experimenting under Stochastic Congestion

Shuangning Li[1], Ramesh Johari[2], Stefan Wager[2] and Kuang Xu[2]

[1] *Harvard University, shuangningli@fas.harvard.edu*
[2] *Stanford University, rjohari@stanford.edu, swager@stanford.edu, kuangxu@stanford.edu*

Stochastic congestion, a phenomenon in which a system becomes temporarily overwhelmed by random surges in demand, occurs frequently in service applications. While randomized experiments have been effective in gaining causal insights and prescribing policy improvements in many domains, using them to study stochastic congestion has proven challenging. This is because congestion can induce interference between customers in the service system and thus hinder subsequent statistical analysis. In this paper, we aim at getting tailor-made experimental designs and estimators for the interference induced by stochastic congestion. In particular, taking a standard queueing system as a benchmark model of congestion, we study how to conduct randomized experiments in a service system that has a single queue with an outside option. We study switchback experiments and a local perturbation experiment and propose estimators based on the experiments to estimate the effect of a system parameter on the average arrival rate. We establish that the estimator from the local perturbation experiment is asymptotically more accurate than the estimators from the switchback experiments because it takes advantage of the structure of the queueing system.

Estimates in Wasserstein metric for some weakly dependent random variables

Gabija Liaudanskaitė[1]

[1] *Vilnius University, Vilnius, gabija.liaudanskaite@mif.vu.lt*

The sum of symmetric three-point 1-dependent nonidentically distributed random variables is approximated by a compound Poisson distribution in the Wasserstein norm. The sum $S_n = X_1 + \ldots + X_n$ is analyzed. Here $X_j = \xi_j\xi_{j+1}$, $\mathrm{P}(\xi_j = -1) = p_j$, $\mathrm{P}(\xi_j = 0) = 1 - 2p_j$, and $\mathrm{P}(\xi_j = 1) = p_j$. It is proved that, if $\max p_j \le \frac{1}{24}$,

$$\|F_n - G\|_{\mathrm{W}} \le C\varepsilon \min\left(\sigma^{-3}, 1\right).$$

Here F_n is a distribution of S_n, $\sigma^2 = \mathrm{Var}S_n$, $\varepsilon = \sum_{k=2}^{n-1}(p_{k-1}p_kp_{k+1} + (p_kp_{k+1})^2)$, $G = \prod_{k=1}^{n} G_k$, $G_k = \exp\{2p_kp_{k+1}(I_1 + I_{-1} - 2I)\}$.

In addition, lower bound estimates are derived in Wasserstein norm. The characteristic function method is used in the proofs.

References

[1] Čekanavičius V., Vellaisamy, P., Compound Poisson approximations in ℓ_p-norm for sums of weakly dependent vectors, *Journal of Theoretical Probability*, 34, (2020) 2241-2264.

[2] Liaudanskaitė G., Čekanavičius V., Approximations for sums of three-valued 1-dependent summetric random variables, *Nonlinear Analysis: Modelling and Control*, 25, (2020) 675-691.

[3] Liaudanskaitė G., Čekanavičius V., Compound Poisson approximations to sums of extrema of Bernoulli variables, *Lithuanian Mathematical Journal*, 62, (2022) 481-499.

Understanding the spatio-temporal variability of speed distribution in public transport data

Karina Lilleborge[1], Sara Martino[2], Ingelin Steinsland[3] and Geir-Arne Fuglstad[4]

[1] *Norwegian University of Science and Technology, Trondheim, karina.lilleborge@ntnu.no*
[2] *Norwegian University of Science and Technology, Trondheim, sara.martino@ntnu.no*
[3] *Norwegian University of Science and Technology, Trondheim, ingelin.steinsland@ntnu.no*
[4] *Norwegian University of Science and Technology, Trondheim, geir-arne.fuglstad@ntnu.no*

Public transport (PT) is at an all-time high in Norway, and is expected to grow in popularity in the coming years [1]. Users of PT agree that reliability is the key factor when assessing the quality of the PT service [2]. To ensure high reliability for the users, causes of delays in the PT network must be identified. A delay in the PT network can be recognized by a lower section speed between two consecutive stops. By defining valid spatio-temporal models for the speed of a single bus line, it is possible to identify certain space-time areas more prone to delays, and make adjustments to the bus schedule. The spatial dependency is modelled through a stochastic partial differential equation, which allows for efficient and inexpensive computations, by taking advantage of the sparsity of the precision matrix [3]. The time dependency is assumed to follow an autoregressive process. The spatio-temporal effects are combined by a Kronecker delta product, allowing for flexibility in the space-time dependency, while still maintaining sparsity in the precision matrix. The model fitting and evaluation is performed using INLA (www.r-inla.org)/inlabru [4, 5]. Grouped cross validation with mean square error and continuous ranked probability score is used for model evaluation [6].

References

[1] Transportøkonomisk institutt. "Cost developments in Norwegian public transport – documentation report". Rapport 1582b/2017. `https://www.toi.no/publikasjoner/kostnadsdrivere-i-kollektivtransporten-dokumentasjonsrapport-article34546-8.html`

[2] Rodriguez-Deniz, Hector, and Mattias Villani. "Robust Real-Time Delay Predictions in a Network of High-Frequency Urban Buses." IEEE Transactions on Intelligent Transportation Systems 23, no. 9 (2022): 16304-16317.

[3] Finn Lindgren, Haavard Rue, and Johan Lindstrom. "An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach (with discussion)". Journal of the Royal Statistical Society, Series B, 73 (4):423–498, 2011.

[4] Haavard Rue, Sara Martino, and Nicolas Chopin. "Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion)". Journal of the Royal Statistical Society, Series B, 71(2):319-392, (2009).

[5] David Borchers, Janine Illian, Steve Buckland, Joyce Yuan, Finn Lindgren, and Fabian E. Bachl. "inlabru" (R package). `https://sites.google.com/inlabru.org/inlabru/home`

[6] Zhedong Liu, and Haavard Rue. "Leave-Group-Out Cross-Validation For Latent Gaussian Models." arXiv preprint arXiv:2210.04482 (2022).

Self-supervised representation learning by clustering in latent space

Fredrik Lindsten[1], Hariprasath Govindarajan[1,2], Per Sidén[2], and Jacob Roll[2]

[1] *Linköping University, Linköping, Sweden, {fredrik.lindsten,hariprasath.govindarajan}@liu.se*
[2] *Qualcomm Technologies, Inc., Linköping, Sweden, {hargov,psiden,jroll}@qti.qualcomm.com*

Self-supervised pre-training is a key methodology for the success of contemporary machine learning, enabling semantically meaningful and transferable representations to be learned from large quantities of unlabeled data. In this talk I will present self-supervised learning (SSL) from the perspective of classical unsupervised learning methods. Specifically, I will discuss clustering in latent space as an efficient approach to SSL and show that popular methods such as DINO and iBOT can be interpreted as clustering models based on mixtures of von Mises–Fisher distributions. Using this insight we identify a missing normalization term when computing the cluster assignment probabilities in the standard DINO formulation. Correcting for this normalization constant we propose DINO-vMF [1], resulting in improved training stability and performance on a range of downstream tasks when compared to DINO. We obtain similar improvements for iBOT-vMF vs iBOT and thereby show the relevance of our proposed modification also for other methods derived from DINO.

References

[1] Govindarajan, H., Sidén, P., Roll, J., and Lindsten, F., DINO as a von Mises-Fisher mixture model, *International Conference on Learning Representations*, Kigali, Rwanda, 2023. `https://openreview.net/forum?id=cMJo1FTwBTQ`

What Drives Cryptocurrency Returns?
A Sparse Statistical Jump Model Approach

Federico P. Cortese[1], Petter N. Kolm[2] and Erik Lindström[3]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy, federico.cortese@unimib.it
[2] Courant Institute of Mathematical Sciences, New York, NY 10012, USA, petter.kolm@nyu.edu
[3] Centre for Mathematical Sciences, Lund University, Sweden, erik.lindstrom@matstat.lu.se

We consider the statistical sparse jump model, a recently developed, robust and interpretable regime switching model, to identify features that drive the return dynamics of the largest cryptocurrencies.

The approach simultaneously performs feature selection, parameter estimation, and state classification. Our large number of candidate features comprises cryptocurrency, sentiment, and financial market-based time series that previously have been identified in the emerging literature as influencing cryptocurrency returns, as well as new ones.

Our empirical study indicates that a three-state model offers the most accurate description of the cryptocurrency returns dynamics. These states have straightforward market-based interpretations as they correspond to bull, neutral, and bear market regimes, respectively. Using the data-driven feature selection methodology, we are able to determine which features are important and which ones are not. Our findings reveal that, among the set of candidate features, the first moments of returns, features that represent trends and reversal signals, market activity, and public attention are key drivers of crypto market dynamics.

Spline smoothing in 3D with positivity constraints: An application to flame tomography

Johan Lindström[1], David Sanned[2], Adrian Roth[2]

[1] *Centre for Mathematical Sciences, Lund University, Sweden*
johan.lindstrom@matstat.lu.se,
[2] *Combustion Physics, Lund University, Sweden*

Tomography is a linear reconstruction problem where each observation is represented by the sum of intensities for all voxels along a line. The aim is to use these observations to reconstruct the observed 3D-shape, represented by intensities in a number of voxels. The problem is often underdetermined and a smoothness prior is imposed to make the solution identifiable [1]. Expressing the smoothness prior as a squared penalty on derivatives the problem can be rephrased as a latent Gaussian Markov random field [2, 3]. To avoid memory and computational limitations iterative methods and FFT are used.

The flame intensity has to be positive, and the observations give information regarding voxels that are known to equal zero. Reconstruction and simulation of Gaussian fields under these constraints can be formulated as a constrained convex optimisation problem. Reconstructions and parameter estimates are obtained using a MC-EM algorithm that utilises iterative solvers for fast simulation in the E-step. Finally we note that correct discretization of the smoothness penalty allows an expression of the Gaussian field that accounts for voxel size, allowing parameters estimated at a coarse spatial scale to be used for reconstructions at higher resolutions, speeding up the parameter estimation.

References

[1] Sanned, D., Lindström, J., Roth, A., Aldén, M. & Richter, M., (2022). Arbitrary position 3D tomography for practical application in combustion diagnostics, *Measurement Science and Technology*, 33(12).

[2] Eilers, P.H.C, Currie, I.D., Durbán, M., (2006) Fast and compact smoothing on large multidimensional grids, *Computational Statistics and Data Analysis*, 50.

[3] Nychka, D.W., (2000) Spatial-process estimates as smoothers, *Smoothing and Regression: Approaches, Computation, and Application*, Wiley, New York, 393–424

Fast and precise inference on diffusivity in interacting particle systems

Gustav Lindwall[1] and Philip Gerlee[2]

[1] *Chalmers University of Technology, Gothenburg, guslindw@chalmers.se*
[2] *Chalmers University of Technology, Gothenburg, gerlee@chalmers.se*

Systems made up of interacting agents is a popular model used in a vast array of applications, not the least in biology where the agents can for example represent single cells. Usually, the particles are assumed to undergo some type of random movements, and a popular way to model this is by using Brownian motion. The magnitude of random motion is often quantified using mean squared displacement, which provides a simple estimate of the diffusion coefficient. However, this method often fails when data is sparse or interactions between agents frequent. In order to address this, we derive a conjugate relationship in the diffusion term for large interacting particle systems undergoing isotropic diffusion, giving us an efficient inference method. The method accurately accounts for emerging effects such as anomalous diffusion stemming from mechanical interactions. We apply our method to an agent-based model with a large number of interacting particles, and the results are contrasted with a naive MSD-based approach. We find a significant improvement in performance when using the higher-order method over the naive approach. This method can be applied to any system where agents undergo Brownian motion and will lead to improved estimates of diffusion coefficients compared to existing methods.

The expected degree distribution in transient duplication divergence models

A. D. Barbour[1] and **Tiffany Y. Y. Lo**[2]

[1] *Institut für Mathematik, Universität Zürich, a.d.barbour@math.uzh.ch*
[2] *Matematiska institutionen, Uppsala universitet, yin_yuan.lo@math.uu.se*

We study the degree distribution of a randomly chosen vertex in a duplication–divergence graph, paying particular attention to what happens when a non-trivial proportion of the vertices have large degrees, establishing a central limit theorem for the logarithm of the degree distribution. Our approach, as in [3] and [2], relies heavily on the analysis of related birth–catastrophe processes. This talk is based on [1].

References

[1] Barbour, A. D. and Lo, T. Y. Y., The expected degree distribution in transient duplication divergence models, *Latin American Journal of Probability and Mathematical Statistics*, 19, (2022) 69-107.

[2] Hermann. F. and Pfaffelhuber, P., Large-scale behavior of the partial duplication random graph, *Latin American Journal of Probability and Mathematical Statistics*, 13, (2016) 687-710.

[3] Jordan, J., The connected component of the partial duplication graph, *Latin American Journal of Probability and Mathematical Statistics*, 15, (2018) 1431-1445.

Learning Curve Behavior: Some Results, Some Open Problems

Marco Loog[1], Alexander Mey[2], and Tom J. Viering[3]

[1] *Radboud University, The Netherlands, marco.loog@ru.nl*
[2] *Eindhoven University of Technology, The Netherlands, a.mey@tue.nl*
[3] *Delft University of Technology, The Netherlands, t.j.viering@tudelft.nl*

Plotting a learner's expected performance against the training set size results in a so-called learning curve. Studying such curves on one or more data sets is a way to get to a better understanding of the generalization properties of this learner. Beyond the standard generalization bounds [1, 2], however, the behavior of learning curves is not well understood and can display (for many a researcher) quite unexpected behavior [3, 4]. We consider the notion of learning curve monotonicity (or smart learners, as Devroye et al. [2] would call them), which asks the performance to not deteriorate with increasing training set sizes in expectation over the training sample. Subsequently, we show surprising learning curve behaviors for various learners. In particular, we consider learners that rely on maximum likelihood and Vapnik's empirical risk minimization [1]. It turns out that these can act nonmonotonically irrespective of the training sample size [5]. Theoretical underpinning can be provided for specific instantiations from classification, regression, density estimation, and clustering. We point out the significance of these findings and conclude with a brief overview and discussion of open problems that, we believe, warrant further theoretical and empirical investigation.

References

[1] Vapnik, V.V. Estimation of Dependences Based on Empirical Data, Springer, 1982.

[2] Devroye, L., Györfi, L., and Lugosi, G. A Probabilistic Theory of Pattern Recognition, Springer, 1996.

[3] Viering, T.J. and Loog, M., The shape of learning curves: a review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2022).

[4] Loog, M., Viering, T.J., A survey of learning curves with bad behavior: or how more data need not lead to better performance, *arXiv*, (2022).

[5] Loog, M., Viering, T.J., Mey, A., Minimizers of the empirical risk and risk monotonicity, *Advances in Neural Information Processing Systems*, (2019).

Conditional Independence Testing in Hilbert Spaces with Applications to Functional Data Analysis

Anton Rask Lundborg[1], Rajen D. Shah[2], Jonas Peters[3]

[1] *University of Copenhagen, Denmark, arl@math.ku.dk*
[2] *University of Cambridge, England, r.shah@statslab.cam.ac.uk*
[3] *ETH Zurich, Switzerland, jonas.peters@stat.math.ethz.ch*

We study the problem of testing the null hypothesis that X and Y are conditionally independent given Z, where each of X, Y and Z may be functional random variables. This generalises testing the significance of X in a regression model of scalar response Y on functional regressors X and Z. We show, however, that even in the idealised setting where additionally (X, Y, Z) has a Gaussian distribution, the power of any test cannot exceed its size. Further modelling assumptions are needed and we argue that a convenient way of specifying these assumptions is based on choosing methods for regressing each of X and Y on Z. We propose a test statistic involving inner products of the resulting residuals that is simple to compute and calibrate: type I error is controlled uniformly when the in-sample prediction errors are sufficiently small. We show this requirement is met by ridge regression in functional linear model settings without requiring any eigen-spacing conditions or lower bounds on the eigenvalues of the covariance of the functional regressor. We apply our test in constructing confidence intervals for truncation points in truncated functional linear models and testing for edges in a functional graphical model for EEG data.

Modelling rating data: exploring the relationship between CUB and CUM models

Ambra Macis[1], Matteo Ventura[1], Marica Manisera [1] and Paola Zuccolotto [1]

[1] *University of Brescia, Italy, ambra.macis@unibs.it, matteo.ventura@unibs.it, marica.manisera@unibs.it, paola.zuccolotto@unibs.it*

Questionnaires are a common tool for investigating respondents' perceptions by means of ratings, which are assumed to be the result of a latent decision process (DP). The DP differs when respondents are asked to express their opinions on a Likert or a Semantic Differential scale. A possible paradigm to formalize the DP is presented in [1], based on the presence of a *feeling* and an *uncertainty* latent component, originally proposed as the foundations of the CUB (Combination of Uniform and shifted Binomial) class [2].
According to this paradigm, respondents expressing ratings on a Likert scale start their reasoning by focusing on the bottom of the scale and moving upward according to their sensations. Differently, when using a Semantic Differential scale, respondents are assumed to start from the middle and move upward or downward.
The CUM (Combination of Uniform and Multinomial), a new model in the CUB class, derived from the above mentioned DP, has been proposed to analyse rating data expressed on a Semantic Differential scale [3].
This work aims at exploring analytically the relationships between the basic CUB and the CUM models, focusing on the possible overlapping of their parametric spaces.

References

[1] Manisera, M., and Zuccolotto, P. (2014). Modeling rating data with nonlinear CUB models. *Computational Statistics and Data Analysis*, 78, 100-118.

[2] D'Elia, A., and Piccolo, D. (2005). A mixture model for preferences data analysis. *Computational Statistics and Data Analysis*, 49(3), 917-934.

[3] Manisera, M., and Zuccolotto, P. (2022). A mixture model for ordinal variables measured on semantic differential scales. *Econometrics and Statistics*, 22, 98-123.

Some advances on extensive-rank matrix denoising and factorization.

Antoine Maillard[1], V. Erba[2], F. Krzakala[2], M. Mézard[3], E. Troiani[2] and L. Zdeborová[2]

[1] *ETH Zürich, Switzerland, antoine.maillard@math.ethz.ch.*
[2] *École Polytechnique Fédérale de Lausanne, Switzerland.*
[3] *Bocconi University, Milan, Italy.*

Factorization of matrices where the rank of the two factors diverges linearly with their sizes has many applications in diverse areas such as unsupervised representation learning, dictionary learning or sparse coding. We consider a setting where the two factors are generated from known component-wise independent prior distributions, and the statistician observes a (possibly noisy) component-wise function of their matrix product. In the limit where the dimensions of the matrices tend to infinity, but their ratios remain fixed, we expect to be able to derive closed form expressions for the optimal mean squared error on the estimation of the two factors. However, this remains a very involved mathematical and algorithmic problem. A related, but simpler, problem is extensive-rank matrix denoising, where one aims to reconstruct a matrix with extensive but usually small rank from noisy measurements. In this talk we will start by approaching the latter problem, and relate previously-known rotationally-invariant estimators to Bayes-optimal estimation, and show how exact asymptotic calculations of the minimal error can be performed using extensive-rank Harish-Chandra-Itzykson-Zuber integrals. We will then highlight the additional difficulties posed by the matrix factorization setting, and will layout a systematic way to derive the corrections to existing approximations, using high-temperature expansions at fixed order parameters which originated in statistical physics. This talk is based on [1] and [2].

References

[1] Maillard Antoine, Krzakala Florent, Mézard Marc and Zdeborová Lenka, Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising, *Journal of Statistical Mechanics: Theory and Experiment* (2022) 083301.

[2] Troiani Emanuele, Erba Vittorio, Krzakala Florent, Maillard Antoine and Zdeborová Lenka, Optimal denoising of rotationally invariant rectangular matrices, *Mathematical and Scientific Machine Learning*, 97–112, (2022), PMLR.

Adaptive online variance estimation in particle filters: the ALVar estimator

Alessandro Mastrototaro[1], Jimmy Olsson[1]

[1] *KTH Royal Institute of Technology, Stockholm, alemas@kth.se, jimmyol@kth.se*

We present a new approach—the ALVar estimator—to estimation of asymptotic variance in sequential Monte Carlo methods, or, particle filters. The method, which adjusts adaptively the lag of the estimator proposed in [2] applies to very general distribution flows and particle filters, including auxiliary particle filters with adaptive resampling. The algorithm operates entirely online, in the sense that it is able to monitor the variance of the particle filter in real time and with, on the average, constant computational complexity and memory requirements per iteration. Crucially, it does not require the calibration of any algorithmic parameter. Estimating the variance only on the basis of the genealogy of the propagated particle cloud, without additional simulations, the routine requires only minor code additions to the underlying particle algorithm. Finally, we prove that the ALVar estimator is consistent for the true asymptotic variance as the number of particles tends to infinity and illustrate numerically its superiority to existing approaches.

References

[1] A. Mastrototaro and J. Olsson. Adaptive online variance estimation in particle filters: the ALVar estimator. *arXiv preprint arXiv:2207.09590*, 2022.

[2] J. Olsson and R. Douc. Numerically stable online estimation of variance in particle filters. *Bernoulli*, 25(4):1504–1535, 2019.

On Concordance Indices for Models with Time-Varying Risk

Thomas Matcham[1] and Axel Gandy[2]

[1] *Imperial College London, thomas.matcham14@imperial.ac.uk*
[2] *Imperial College London, a.gandy@imperial.ac.uk*

Harrel's concordance index is a commonly used discrimination metric for survival models, particularly for models where the relative ordering of the risk of individuals is time-independent, such as the proportional hazards model. There are several suggestions, but no consensus, on how it could be extended to models where risk varies over time, e.g. in case of crossing hazard rates. We show that asymptotically, concordance is maximized if and only if the risk score is concordant with the hazard rate, in the sense that for a comparable pair where the first event time is observed, the risk score is concordant with the hazard rate at this first event time. Thus, we suggest using the hazard rate as the risk score when calculating concordance. Through simulations, we demonstrate situations in which other concordance indices can lead to incorrect models being selected over a true model, justifying the use of our suggested risk prediction in both model selection and in loss functions in, e.g., deep learning models.

Are ensembles getting better all the time?

Pierre-Alexandre Mattei[1], Damien Garreau[1]

[1] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, France*

Ensemble methods combine the predictions of several base models. We study whether or not including more models in an ensemble always improve its average performance. Such a question depends on the kind of ensemble considered, as well as the predictive metric chosen. We focus on situations where all members of the ensemble are a priori expected to perform as well, which is the case of several popular methods like random forests or deep ensembles. In this setting, we essentially show that ensembles are getting better all the time if, and only if, the considered loss function is convex. More precisely, when the loss function is convex, the average loss of the ensemble is a decreasing function of the number of models. This result generalises and unifies previous theorems, and has a very short proof based on Jensen's inequality. When the loss function is nonconvex, we show a series of results that can be summarised by the insight that ensembles of good models keep getting better, and ensembles of bad models keep getting worse. To this end, we use large deviation techniques to prove a new result on the monotonicity of tail probabilities that may be of independent interest.

Matrix variate generalized asymmetric Laplace distributions

Stepan Mazur[1,2], Tomasz J. Kozubowski[3] and Krzysztof Podgórski[4]

[1] *Unit of Statistics, Örebro University, Örebro, Sweden, stepan.mazur@oru.se*
[2] *Department of Economics and Statistics, Linnaeus University, Växjö, Sweden, stepan.mazur@lnu.se*
[3] *Department of Mathematics and Statistics, University of Nevada, RENO, USA, tkozubow@unr.edu*
[4] *Department of Statistics, Lund University, Lund, Sweden, Krzysztof.Podgorski@stat.lu.se*

The generalized asymmetric Laplace (GAL) distribution, also known as the *variance/mean-gamma* model, is a popular flexible class of distributions that can account for peakedness, skewness, and heavier than normal tails. We consider extensions of the GAL distribution to the matrix variate case, which arise as covariance mixtures of matrix variate normal distributions. Two different mixing mechanisms connected with the nature of the random scaling matrix are considered, leading to what we term matrix variate GAL distributions of Type I and II. While Type I matrix variate GAL distribution has been studied before, there is no comprehensive account of Type II in the literature, except for their rather brief treatment as a special case of matrix variate generalized hyperbolic distributions. With this work we fill this gap, and present an account for basic distributional properties of Type II matrix variate GAL distributions. In particular, we derive their probability density function and the characteristic function, as well as provide stochastic representations related to matrix variate gamma distribution. We also show that this distribution is closed under linear transformations, and study the relevant marginal distributions. In addition, we also briefly account for Type I and discuss the intriguing connections with Type II.

Pattern Boosting, a statistical learning method for chemistry applications

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*

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 a property related to the structure of the graph. Taking advantage of the iterative nature of the boosting, we propose an algorithm that at each step expands only the most significant path selected in the previous step to keep the search space limited. The decision on which path to select is made in a gradient boosting fashion. Emphasis is put on explainability and our algorithm provides a measure of how much each individual path contributes to the final prediction of the outcome. We apply our algorithm to a chemical dataset known as ”tmQMg”[1], which contains 60k molecular graphs, with the goal of predicting quantum properties like the HOMO/LUMO gap.

References

[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

Causal Discovery Using Causally Invariant Locally Linear Models

Alexander Mey[1], Rui Castro[1]

[1] *Technical University Eindhoven, Eindhoven, The Netherlands, a.mey@tue.nl, r.m.pires.da.silva.castro@tue.nl*

Identifying causal relationship is an often desired, but difficult, task, and generally only possible under specific assumptions. In this paper we are considering the task of identifying the causal parents of a target variable from a set of potential candidates, given only observational data. The main assumption of our work is that the potential candidate variables were observed in different environments. Different environments may arise as different settings of a machine, different countries in which we collected data or different time intervals in a dynamical process. We introduce a model and a set of assumptions that allow us to view and use the different environments as interventions on the observed system, which in turn allows us to draw causal conclusions. In particular we introduce a locally linear model class that can identify causal structures based on the following two assumptions: the observed data is sufficiently heterogeneous across environments and the underlying causal graph is the same for all environments. In this way we extend the work of [1], who assumed a global linear model, in several directions. We also propose a, in this context novel, hypothesis test to detect the causal parents based on extreme values that can still make use of the different environments, even if the data heterogeneity is sparse across environments. We finally analyze the behavior of the model with respect to the strength of the heterogeneity, the number of environments, and the number of samples per environment.

References

[1] Peters, J., Bühlman, P., Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B* (Statistical Methodology), 78(5):947-1012,2016.

COX PROCESSES DRIVEN BY TRANSFORMED GAUSSIAN PROCESSES ON LINEAR NETWORKS

Jesper Møller[1] and Jakob G. Rasmussen[2]

[1] *Aalborg University, Denmark, jm@math.aau.dk*
[2] *Aalborg University, Denmark, jgr@math.aau.dk*

There is a lack of point process models on linear networks. For an arbitrary linear network, we use isotropic covariance functions with respect to the geodesic metric or the resistance metric to construct new models for isotropic Gaussian processes and hence new models for various Cox processes with isotropic pair correlation functions. In particular we introduce three model classes given by log Gaussian, interrupted, and permanental Cox processes on linear networks, and consider for the first time statistical procedures and applications for parametric families of such models. Moreover, we construct new simulation algorithms for Gaussian processes on linear networks and discuss whether the geodesic metric or the resistance metric should be used for the kind of Cox processes studied in this talk.

Efficient sparsity adaptive change point estimation

Per August Moen[1], Ingrid K. Glad[2] and Martin Tveten[2]

[1] *University of Oslo, Oslo, pamoen@math.uio.no, glad@math.uio.no*
[2] *Norwegian Computing Center, Oslo, tveten@nr.no*

We study detection and estimation of an unknown number of changes in the mean-vector sequence of high-dimensional isotropic Gaussian vectors. We present a novel change point estimation algorithm, called ESAC, short for **E**fficient **S**parsity **A**daptive **C**hange point estimator. To detect and estimate an unknown number of 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. ESAC is also multiscale, in the sense that the spacing between consecutive change points, as well as the affected coordinates at each change point, may differ. For both single and multiple change point estimation, we prove finite sample theoretical results 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 show that ESAC is highly competitive in terms of estimation accuracy and computational complexity.

References

[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] Kovács, S., Bühlmann, P., *Seeded binary segmentation: a general methodology for fast and optimal changepoint detection* (2022), Biometrika, 110(1), 249-256.

Learning Granger-causal graphs from partially observed multivariate time series

Søren Wengel Mogensen

Lund University, soren.wengel_mogensen@control.lth.se

Granger causality is an asymmetric notion of independence which describes how a system of stochastic processes evolves over time. Let A, B, and C be three subsets of coordinate processes. Intuitively speaking, A is Granger-noncausal for B given C if knowing the past of both A and C is not more informative about the present of B than knowing the past of C only.

We use directed mixed graphs to represent the Granger causality structure of the observed coordinate processes. Several directed mixed graphs may induce the same Granger noncausalities and therefore it is of interest to characterize such equivalence classes. Directed mixed graphs satisfy a certain maximality property and this allows a simple graphical representation of an entire equivalence class. The equivalence class can be learned from data and its graphical representation concisely describes which underlying graphs could have generated the observations.

We also extend the maximality result to so-called weak equivalence classes. This enables tractable learning of equivalence classes of Granger-causal graphs, even in large networks. Moreover, this can be done in a modular fashion such that data can be seen as providing separate evidence for or against the inclusion of each edge in the output graph.

Learning distributions on tangent bundles with generative adversarial networks

Vincent Molin[1], Moritz Schauer[12] and Axel Ringh[12]

[1] *Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, molinv@chalmers.se, smoritz@chalmers.se, axelri@chalmers.se*

[2] *University of Gothenburg, Gothenburg.*

Given a data set sampled from some underlying distribution of interest, generative adversarial networks[1] (GANs) are a class of deep generative models that during training learn an approximate low-dimensional parametrization of the data distribution. Assuming that the data distribution is concentrated on a smooth submanifold of the data space, we consider the case when data points are coupled with samples from the tangent space at that point. Such distributions on the tangent bundle can, for instance, arise naturally when each point is paired with a Gaussian distributed according to a Riemannian metric tensor. We present an extension of GAN training with correct gradient penalties[2][3] where the discriminative network sees both points and tangents and the generative network outputs both the primal evaluation as well as a directional derivative. The training is coupled with suitable regularizations and architectural choices. Preliminary experiments show that our approach leads to more stable training and better results.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems 27*, 2014.

[2] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron C. Courville. Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems 30*, 2017.

[3] Lars Mescheder, Andreas Geiger, Sebastian Nowozin. Which Training Methods for GANs do actually Converge?. *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80:3481-3490, 2018.

Hierarchical group-wise multivariate change-point detection

Mehdi Moradi[1], Ottmar Cronie[2], Pérez-Goya[3] and Jorge Mateu[4]

[1] *Umeå University, Umeå, Sweden, mehdi.moradi@umu.se*
[2] *Chalmers University of Technology & University of Gothenburg, Gothenburg, Sweden*
[3] *Public University of Navarre, Pamplona, Spain*
[4] *University Jaume I, Castellón, Spain*

Detecting change-points in multivariate settings is usually carried out by analysing all marginals either independently, via univariate methods, or jointly, through multivariate approaches. The former discards any inherent dependencies between different marginals and the latter may suffer from domination/masking among different change-points of distinct marginals. As a remedy, we propose an approach which groups marginals with similar temporal behaviours, and then performs group-wise multivariate change-point detection. Our approach groups marginals based on hierarchical clustering using distances which adjust for inherent dependencies. Through a simulation study, considering various scenarios where domination/masking happens, we show that our approach, by preventing domination/masking, significantly enhances the general performance of the employed multivariate change-point detection method. Finally, we apply our approach to two datasets: i) Land Surface Temperature in Spain, during the years 2000–2021, and ii) The WikiLeaks Afghan War Diary data.

References

[1] Moradi, M., O. Cronie, U. Pérez-Goya, and J. Mateu (2023). Hierarchical spatio-temporal change-point detection. *The American Statistician*. doi: 10.1080/00031305.2023.2191670.

Efficient HMM computations using symmetries, with an application to linked genetic markers and pedigree inference

Petter Mostad[1], Andreas Tillmar[2] and Daniel Kling[2]

[1] *Chalmers University, Göteborg, mostad@chalmers.se*
[2] *Department of Forensic Genetics and Toxicology, Linköping, andreas.tillmar@rmv.se, daniel.kling@rmv.se*

Sequential models, or hidden Markov models, are extremely useful tools, but efficient computation can be crucial for their applicability. We study models where a group of symmetries act on the hidden chain in a way that commutes with the transition functions. We show how, when the emission probabilities are invariant under some subgroup, this can be used to simplify and speed up computations. In our single example, the Markov chain follows a chromosome, and its variables represent DNA markers inherited in a pedigree. Computations with the Lander-Green algorithm for linked markers can then be sped up significantly, in particular for large pedigrees. In applications, we show how pedigree inference is possible for such pedigrees and for DNA data that contain large amounts of uncertainty.

PointedSDMs: an R Package to Facilitate the Modeling of Integrated Species Distribution Models

Philip S. Mostert[1] and Robert O'Hara[2]

[1] *Norwegian University of Science and Technology, Trondheim, philip.s.mostert@ntnu.no*
[2] *Norwegian University of Science and Technology, Trondheim, bob.ohara@ntnu.no*

Ecological data are being collected at a large scale from a multitude of different source, each with their own sampling protocols and assumptions. As a result, the integration of disparate datasets is a rapidly growing area in quantitative ecology, and is subsequently becoming a major asset in understanding the shifts and trends in species' distributions. However, the tools and software available to construct statistical models to integrate these disparate datasets into a unified framework is lacking. This has made these methods inaccessible to general practitioners, and has stagnated the growth of data integration in more applied settings. We therefore present *PointedSDMs*: an easy to use *R* package used to construct integrated species distribution models. It provides functions to easily format the data, fit the models in a computationally efficient way, and presents the output in a format that is convenient for additional work. This presentation illustrates a selection of diverse case studies which use the package for their analyses.

Simulation of the stochastic heat equation on the sphere

Ioanna Motschan-Armen[1], Annika Lang

[1] *Chalmers and University of Gothenburg, Gothenburg, ioannamo@chalmers.se*
[2] *Chalmers and University of Gothenburg, Gothenburg, annika.lang@chalmers.se*

We consider the stochastic heat equation on the sphere driven by isotropic Wiener noise. Using the spherical harmonic functions, we calculate the solution up to some truncation level leading to a spectral approximation in space.

For the time discretization the exact solution in time is compared to the standard forward and backward Euler scheme. We simulate strong and weak errors to show convergence rates of the different methods with and without Monte Carlo approximations. Special emphasis is given to the second moment as an important quantity to describe the properties of the solution.

To illustrate the generated solutions, time evolving sample paths will be integrated via links. As an application we use these samples to generate evolving stochastic surfaces.

This is joint work with Annika Lang (Chalmers & University of Gothenburg).

ZB splinets for representation of distributional data

Hiba Nassar[1], Ivana Pavlu[2], Jana Burkotova[2], Jitka Machalova[2] and Karel Hron[2]

[1] *Technical University of Denmark, Denmark, hibna@dtu.dk*
[2] *Palacký University Olomouc, Czech Republic, ivana.pavlu@upol.cz, jana.burkotova@upol.cz, jitka.machalova@upol.cz, karel.hron@upol.cz*

Distributional data such as probability density functions can be treated as functional data objects in Bayes spaces satisfying the unit integral constraint. A centered log-ratio (clr) transformation allows to process such data in the standard L^2 space. However, the unit integral property of the density functions in the Bayes spaces induces a zero integral constraint in L^2 space. Therefore a spline basis in L^2_0, namely ZB-splines basis, is constructed for proper representation of density functions. As a result of this paper, we describe an efficient method of obtaining an orthogonal ZB-splines basis, which is called a ZB-splinets. Advantages of ZB-splinets approach such as orthogonality and locality of splinets supports are desirable for interpretability e.g. in the context of principal component analysis. The proposed approach is presented on the empirical dataset describing the demographic composition of 114 municipalities in Upper Austria.

References

[1] Liu, Xijia, Hiba Nassar, and Krzysztof Podgórski. "Dyadic diagonalization of positive definite band matrices and efficient B-spline orthogonalization." Journal of Computational and Applied Mathematics 414 (2022): 114444.

[2] Machalová, Jitka, et al. "Compositional splines for representation of density functions." Computational Statistics 36 (2021): 1031-1064.

Module-based regularization improves Gaussian graphical models when observing noisy data

Magnus Neuman[1], Joaquín Calatayud[2], Viktor Tasselius[3] and Martin Rosvall[1]

[1] *Integrated Science Lab, Department of Physics, Umeå University, Umeå, Sweden, magnus.neuman@umu.se, martin.rosvall@umu.se*

[2] *Departamento de Biología, Geología, Física y Química Inorgánica, Universidad Rey Juan Carlos, Madrid, Spain, joaquin.calatayud@urjc.es*

[3] *School of Public Health and Community Medicine, University of Gothenburg, Gothenburg, Sweden, viktor.tasselius@gu.se*

Researchers often represent relations in multi-variate correlational data using Gaussian graphical models, which require regularization to sparsify the models. Acknowledging that they often study the modular structure of the inferred network, we suggest integrating it in the cross-validation of the regularization strength to balance under- and overfitting. Using synthetic and real data, we show that this approach allows us to better recover and infer modular structure in noisy data compared with the graphical lasso, a standard approach using the Gaussian log-likelihood when cross-validating the regularization strength.

Föllmer–Schweizer decomposition and discretization with jump correction in exponential Lévy models

Thuan Nguyen[1]

[1] *Saarland University, Germany, nguyen@math.uni-sb.de*

We discuss in this talk two hedging problems in exponential Lévy models.

First, we provide an explicit representation for the Föllmer–Schweizer decomposition of European type options under mild conditions, which implies a closed-form expression of the corresponding local risk-minimizing strategies. This representation can be regarded as a counterpart of that in [2].

Secondly, we discretize stochastic integrals driven by an exponential Lévy process using a new jump-corrected approximation method to achieve L_p-estimates, $p \in (2, \infty)$, for the discretization errors. To do that, we employ the weighted BMO-approach which was introduced in [1] and the effect of a change of measure satisfying a reverse Hölder inequality is also addressed.

As an application, the error caused by discretizing the stochastic integral component in the Föllmer–Schweizer decomposition is investigated. In particular, when the small jump activity of the underlying Lévy process behaves like an α-stable process with $\alpha \in (1, 2)$, our results achieve under a regular regime the same convergence rate for the error as in [4]. Moreover, our approach extends to the case $\alpha \in (0, 1]$ and to the L_p-setting, $p \in (2, \infty)$, which are not treated there.

This talk is based on the manuscript [3].

References

[1] S. Geiss, Weighted BMO and discrete time hedging within the Black-Scholes model, *Probab. Theory Related Fields* 132, 2005, 13–38.

[2] F. Hubalek, J. Kallsen, L. Krawczyk, Variance-optimal hedging for processes with stationary independent increments, *Ann. Appl. Probab.* 16(2), 2006, 853–885.

[3] T. Nguyen, Explicit Föllmer–Schweizer decomposition and discretization with jump correction in exponential Lévy models, 2022. arXiv:2009.04328.

[4] M. Rosenbaum, P. Tankov, Asymptotically optimal discretization of hedging strategies with jumps, *Ann. Appl. Probab.* 24(3), 2014, 1002–1048.

Incorporating different sources of information for Bayesian optimal portfolio selection

Vilhelm Niklasson[1], Olha Bodnar [2] and Taras Bodnar [1]

[1] *Örebro University, Örebro, olha.bodnar@oru.se*
[2] *Stockholm University, Stockholm, niklasson@math.su.se, taras.bodnar@math.su.se*

In this paper, Bayesian inference procedures are developed for the tangency portfolio by utilizing the subjective conjugate prior and the objective Jeffreys prior, which are derived for its weights. The application of the conjugate prior allows to incorporate external sources of information directly into the decision process, together with historical data of asset returns. This constitutes a great advantage of the suggested approach by implementing both the expert knowledge and the available information into the construction of optimal portfolios. Via en extensive bootstrapping investigation and a large empirical study, the established new approaches are compared to the existing ones in terms of several performance measures. It is concluded that the Bayesian tangency portfolio based on the conjugate prior outperforms the other trading strategies in most of the considered cases.

Local tail-scale invariance in scoring rules for extreme value forecasts

Helga Kristin Olafsdottir[1], Holger Rootzén[1] and David Bolin[2]

[1] *Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden, khelga@chalmers.se*

[2] *Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*

Within the field of statistical analysis of extreme values, one main objective is to make a prediction of the behavior of future events. Proper scoring rules measure the quality of these forecast distributions and can be used to compare models. The choice of scoring rules can affect the result, since they emphasize different factors. Locally scale invariant scoring rules are indifferent to the prediction uncertainty when measuring the quality of forecasts and can therefore be appropriate for spatial forecasts with varying scale. However, this can be an unnecessarily strict requirement for extreme value forecasts. Instead, we propose the concept of local tail-scale invariance, fulfilling local scale invariance for large events. We develop and study a scaled version of the weighted Continuous Ranked Probability score (swCRPS) that fulfills this criteria. We show that this is a suitable alternative to the wCRPS for scoring extreme value models over areas with varying scale of extreme events, through simulations as well as on real world data using models on extreme water levels in the Great Lakes and annual maximum rainfall in the Northeastern United States.

On Graph Limits as Models for Interaction Data

Sofia Olhede[1]

[1] *École Polytechnique Fédérale de Lausanne, Switzerland, sofia.olhede@epfl.ch*

Network data has become a staple in many different applications, ranging from ecology, to neuroscience and systems biology. Its inference will of course depend on the application where we collect the network data, but I will discuss some general principles based on probabilistic symmetries such as permutation invariance. Just like other probabilistic invariances, the distributional invariance to permuting indices of a matrix of interactions implies a representation theorem (the Aldous-Hoover theorem). This representation is in terms of a graph limit function, or graphon. I will discuss the representation, how to make inferences based on this representation, what to do if distributional permutation invariance does not hold, and what to do if we have additional information such as time stamp of interactions, multiple interactions or additional covariate data.

Techniques for Estimating Conditional Shapley Values and When to Use Them

Lars Henry Berge Olsen[1], Ingrid Kristine Glad[1], Martin Jullum[2], and Kjersti Aas[2]

[1] *University of Oslo, Oslo, Norway, lholsen@math.uio.no, glad@math.uio.no*
[2] *Norwegian Computing Center, Oslo, Norway, jullum@nr.no, kjersti@nr.no*

Complex machine learning (ML) models often obtain accurate predictions for supervised learning problems in numerous fields but at the cost of interpretability. Not understanding the input's influence on the ML-model's output is a significant drawback; hence, the field of explainable artificial intelligence has proposed several types of explanation frameworks [1, 2]. The most commonly used framework is *Shapely values*, a promising model-agnostic explanation methodology with a solid mathematical foundation and unique theoretical properties from cooperative game theory [3, 4].

There are several 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 must model the input variables' dependencies [5, 6]. [7] states that an important future research direction is developing new approaches and systematically evaluating existing approaches. In our work, we extend and develop new methods and systematize them into classes for comparison and evaluation. We conduct extensive simulation studies to evaluate how precisely the techniques estimate the conditional expectations/Shapley values. Finally, we apply the techniques to real-world data examples and provide recommendations for when to use the different techniques/approaches.

References

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

[2] Covert I., Lundberg S., Lee S., *Explaining by removing: A unified framework for model explanation*, Journal of Machine Learning Research, 22:209 (2021), 1-90.

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

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

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

[6] 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.

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

Local inference for functional data on manifolds using permutation tests

Niels Lundtorp Olsen[1], Alessia Pini[2] and Simone Vantini[3]

[1] *Technical University of Denmark, Lyngby, nalo@dtu.dk*
[2] *Università Cattolica del Sacro Cuore, Milano, alessia.pini@unicatt.it*
[3] *Politecnico di Milano, Milano, simone.vantini@polimi.it*

Pini and Vantini [1] introduced the interval-wise testing procedure which performs local inference for functional data defined on an interval domain, where the output is an adjusted p-value function that controls for type I errors.

We extend this idea to a general setting where the domain is a Riemannian manifold. This requires new methodology such as how to define adjustment sets on product manifolds and how to approximate the test statistic when the domain has non-zero curvature. We introduce a new parameter that defines the amount of type I error control, with pointwise tests as one extreme, and control on the entire domain as the other extreme. We propose to use permutation tests for inference. We do a simulation study on a "chameleon-shaped" manifold and note the tradeoff between type I and type II errors: increasing the adjustment set reduces the type I error but also results in smaller areas of significance. Finally we apply the procedure in two applications related to climate change where the manifolds are a complex subset of S^2 and $S^2 \times S^1$, respectively.

References

[1] Pini, A., and Vantini, S., Interval-wise testing for functional data, *Journal of nonparametric statistics*, 29, (2017) 407–424.

Effortless bias reduction in self-normalized importance sampling

Jimmy Olsson[1], Gabriel Cardoso[2], Eric Moulines[2], Sergey Samsonov[3], and Achille Thin[4]

[1] *KTH Royal Institute of Technology, Stockholm, Sweden, jimmyol@kth.se*
[2] *Ecole polytechnique, Paris–Saclay, France* [3] *HSE University, Moscow, Russia* [4] *AgroParisTech, Paris, France*

Self-normalized importance sampling (SNIS) is a classical method for approximating expectations under a given target distribution, known up to a normalization constant, on the basis of independent samples from some proposal distribution and associated importance weights. In the era of artificial intelligence and statistical machine learning, characterized by large computational resources and Bayesian inference, SNIS is enjoying a revival. While the use of self-normalization can have a positive effect on the dispersion of the estimator, it introduces bias. In this talk, I propose a new estimator whose complexity and variance are essentially the same as for SNIS, but whose bias is significantly reduced. The estimator is a wrapper in the sense that it includes the same proposal samples and importance weights as SNIS, but makes clever use of iterated sampling–importance resampling (i-SIR) and Gibbs sampling to form a bias-reduced version of the SNIS estimator. The proposed algorithm is furnished with rigorous theoretical results, which are illustrated numerically. Finally, I will discuss the extension of the proposed sampling technology to Feynman–Kac particle models and the so-called PaRIS algorithm. The talk is based on the recent papers [1, 2].

References

[1] Cardoso, G., Samsonov, S., Thin, A., Moulines, E., and Olsson, J., "BR-SNIS: Bias-reduced self-normalized importance sampling", *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[2] Cardoso, G., Moulines, E., and Olsson, J., "Particle-based, rapid incremental smoother meets particle Gibbs". *Statistica Sinica* (2023, to appear).

Trace test for high-dimensional cointegration

Alexei Onatski[1]

[1] *University of Cambridge, UK, ao319@cam.ac.uk*

This paper studies Johansen's [1] trace test for cointegration in high-dimensional data. We show that when both cross-sectional and temporal dimension of the data go to infinity proportionally, the shifted and scaled modified trace statistic converges to a Gaussian random variable. We give explicit formulae for the shift and scale parameters, as well as for the mean and variance of the Gaussian limit. Monte Carlo analysis shows excellent size properties of the asymptotic test, which is an improvement over the Bartlett-corrected versions of the original trace test, especially for relatively large ratios of the dimensionality to the sample size. The Monte Carlo also reveals a non-motonicity of the power of the test. We comment on the source of such a non-monotonicty.

References

[1] Johansen, S. (1988) "Statistical Analysis of Cointegrating Vectors," *Journal of Economic Dynamics and Control* 12, 231-254

Bayesian Learning on Graphs using Deep Gaussian Markov Random Fields

Joel Oskarsson[1], Per Sidén[2], and Fredrik Lindsten[1]

[1] *Linköping University, Sweden, joel.oskarsson@liu.se, fredrik.lindsten@liu.se*
[2] *Qualcomm Technologies, Inc., psiden@qti.qualcomm.com*

Machine learning methods on graphs have proven useful in many applications due to their ability to handle data with general structure. The framework of Gaussian Markov Random Fields (GMRFs) provides a principled way to define Gaussian models on graphs by utilizing their sparsity. Deep GMRFs [2, 1] are flexible GMRF models constructed using a multi-layer structure. The layers are designed to allow for efficient training using variational inference and existing software frameworks for graph neural networks. This allows the models to scale to large graphs, without making any restrictive assumptions on the graph structure. This talk will give an introduction to Deep GMRFs, their application to graph-structured data and the Bayesian inference involved.

References

[1] J. Oskarsson, P. Sidén, and F. Lindsten. Scalable deep Gaussian Markov random fields for general graphs. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[2] P. Sidén and F. Lindsten. Deep Gaussian Markov random fields. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

A spatial fusion model for jittered and geomasked data

John Paige[1], Geir-Arne Fuglstad[1] and Andrea Riebler[1]

[1] *NTNU, Trondheim, Norway, john.paige@ntnu.no, gier-arne.fuglstad@ntnu.no, andrea.riebler@ntnu.no*

Positional error, error in the locations of spatial data, can bias a spatial model's parameter estimates and spatial predictions when improperly accounted for, and is relevant in applications from public health to paleoseismology. Existing methods that account for positional error frequently either rely on non-generalizable parametric assumptions, employ ad hoc techniques, or use computationally intensive MCMC. We show that a newly introduced method addressing these issues can be extended to account for arbitrary positional error distributions including jittering and geomasking, the censoring of each observation point location up to the area containing it. We further provide a flexible numerical integration scheme accounting for spatial covariate information.

We apply the method to women's secondary education completion data in the 2018 Nigeria demographic and health survey (NDHS) containing point locations jittered via random radial displacements, and the 2016 Nigeria multiple indicator cluster survey (NMICS) containing geomasked locations. Both surveys add positional error intentionally for confidentiality purposes. In this setting where high quality survey data is sparse, we show via validation that the spatial fusion of these two datasets in a statistically rigorous way improves parameter estimates and spatial prediction precision.

Fused Lasso Nearly-Isotonic Signal Approximation in General Dimensions

Vladimir Pastukhov

Chalmers University of Technology, Gothenburg, Sweden, vlapas@chalmers.se

We introduce and study fused lasso nearly-isotonic signal approximation, which is a combination of fused lasso and generalized nearly-isotonic regression. We show how these three estimators relate to each other and derive solution to the general problem. Our estimator is computationally feasible and provides a trade-off between monotonicity, block sparsity and goodness-of-fit. Next, we prove that fusion and near-isotonisation in one dimensional case can be applied interchangably, and this step-wise procedure gives the solution to the original optimization problem. This property of the estimator is very important, because it provides a direct way to construct path solution when one of the penalization parameters is fixed. Also, we derive an unbiased estimator of the degrees of freedom of the estimator.

Cashflow-driven investment beyond expectations

Sergio Alvares Maffra[1], **Teemu Pennanen**[2]

[1] *King's college London, sergio.maffra@kcl.ac.uk*
[2] *King's College London, teemu.pennanen@kcl.ac.uk*

This paper presents a computationally tractable optimization model for cashflow-driven investment where the aim is to find asset portfolios whose future payouts cover given liability payments as well as possible. While current industry solutions are largely based on expected future cash flows, we use a stochastic optimization model that seeks portfolios that give the best possible match across time as well as scenarios. In our model, cashflow matching across scenarios is controlled by the risk aversion and the timing is controlled by illiquidity of money markets. When the illiquidity increases, the hedging strategy quickly shifts towards portfolios suggested by deterministic models, but significant uncertainty remains. The model can incorporate hundreds of quoted instruments in the construction of optimal hedging strategies. The hedging strategies are able to employ any statistical connections between the liabilities and publicly quoted assets. The model is solved with simple Monte-Carlo approximations and off-the-shelf convex optimization software. Besides optimal hedging strategies, we find the least cost of hedging which provides a market-consistent valuation based on the current quotes and the liquidity factors as well as the views and risk preferences of the investor/regulator. The approach is illustrated by the pricing and hedging of defined benefit pension liabilities which depend on uncertain longevity developments and the consumer price index. The hedging strategies are constructed from 128 publicly quoted instruments including index-linked bonds and equities. We find that the optimized hedging strategies achieve lower risk at a lower cost than the strategies obtained by matching expected cashflows. The hedging-based liability valuations are robust with respect to model parameters and the additional risk reduction achieved by optimization does not add much to the overall hedging cost.

Gaussian processes for spatio-temporal data on surfaces

Mike Pereira[1]

[1] *Mines Paris - PSL University, Fontainebleau (France), mike.pereira@minesparis.psl.peu*

In the context of geostatistical modeling of spatio-temporal data, it is usual to assume that these data can be represented by a Gaussian random field defined over a spatial domain. This approach allows not only to model the uncertainty one might have on the data, but also to perform such tasks as simulations and predictions while accounting for the spatio-temporal correlations observed in the data. In this talk, we present Gaussian process models to describe spatio-temporal data in the particular case where these lie on a surface. We propose to define the Gaussian random fields used to model the data as solutions of a stochastic partial differential equation (SPDE) defined on the surface. This SPDE approach to Gaussian processes has been extensively used to model spatio-temporal data on Euclidean domains [2], and extended to model spatial data on surfaces (see eg. [3]) and more generally on Riemannian manifolds [1]. Our aim is now to further generalize this last approach to the spatio-temporal setting by considering a diffusion SPDE, and propose matrix-free algorithms for sampling and predicting the resulting random fields.

This work is conducted in collaboration with A. Lang (Chalmers University and University of Gothenburg)

References

[1] A. Lang and M. Pereira. Galerkin–Chebyshev approximation of Gaussian random fields on compact Riemannian manifolds. *arXiv preprint arXiv:2107.02667*, 2021.

[2] F. Lindgren, D. Bolin, and H. Rue. The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spatial Statistics*, 50:100599, 2022.

[3] A. F. Mejia, Y. Yue, D. Bolin, F. Lindgren, and M. A. Lindquist. A Bayesian general linear modeling approach to cortical surface fMRI data analysis. *Journal of the American Statistical Association*, 115(530):501–520, 2020.

Stratified sampling improves the efficiency of sequential approximate bayesian computation

Henri Pesonen[1] and Jukka Corander[2,3,4]

[1] *Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Norway*
henri.pesonen@medisin.uio.no
[2] *Oslo Centre for Biostatistics and Epidemiology, Oslo University, Norway,*
jukka.corander@medisin.uio.no
[3] *Helsinki Institute of Information Technology, Department of Mathematics and Statistics, University of Helsinki, Finland*
[4] *Parasites and Microbes, Wellcome Sanger Institute, UK*

Approximate bayesian computation (ABC) methods are standard tools for inferring parameters of complex models when related likelihood functions are analytically intractable. A common approach to accelerate the acceptance rate of the rejection sampling based ABC methods is to use sequential monte carlo (ABC SMC) to produce an adapting sequence of proposal distributions. Proposal distributions for the next iteration round are most often calculated empirically from the weighted set of samples of the current iteration round. We investigate methods that improve the acceptance rate by taking the simulated observation values into account in construction of the proposal distribution and introduce a method called stratified distance ABC SMC. Algorithm stratifies samples based on their distance between corresponding simulated and the true observation and estimates a distinct proposal distribution for each set of samples. Further efficiency is gained by introducing a novel stopping rule for the sequential process based on the stratified posterior samples.

Invariant models in real life: falsifiability, dynamical systems, and policy learning

Jonas Peters[1]

[1] *ETH Zurich, jonas.peters@stat.math.ethz.ch*

Assume that we observe data from a response Y and a vector X of covariates under different experimental conditions (or environments). Rather than focusing on the model that is most predictive, it has been suggested to take into account the invariance of a model. This can help us to infer causal structure (Which covariates are causes of Y?) or find models that generalize better (How well does the model perform on an unseen environment?). We consider these ideas in the light of three applications and discuss how to adapt the corresponding methods.

SPDE bridges with observation noise and their spatial approximation

Andreas Petersson[1], Giulia Di Nunno[1] and Salvador Ortiz-Latorre[1]

[1] *University of Oslo, Norway, andreep@math.uio.no*

In this talk, we introduce SPDE bridges with observation noise and analyze their spatially semidiscrete approximations. The SPDEs are linear and driven by an additive cylindrical Wiener process, and the observational noise is cylindrical and Gaussian. We establish SPDE bridges via conditional distributions of Gaussian random variables in Hilbert spaces and provide a general framework for their spatial discretization. Noisily observed stochastic reaction-diffusion equations on bounded domains is the canonical example. We derive explicit convergence rates for spectral and finite element-based methods, demonstrating that, given sufficiently rough observation noise, these rates align with the discretization of the original SPDE. This talk is based on [1].

References

[1] Di Nunno, G., Ortiz-Latorre, S., and Petersson, A., SPDE bridges with observation noise and their spatial approximation, *Stoch. Process. Their Appl.*, 158, (2023) 170-207.

Guided sequential ABC schemes for intractable Bayesian models

Umberto Picchini[1], and Massimiliano Tamborrino[2]

[1] *Chalmers University of Technology and University of Gothenburg, Sweden, picchini@chalmers.se*

[2] *Department of Statistics, University of Warwick, UK, massimiliano.tamborrino@warwick.ac.uk*

Sequential algorithms such as sequential importance sampling (SIS) and sequential Monte Carlo (SMC) have proven fundamental in Bayesian inference. However, probabilistic models often do not admit a readily available likelihood function or one that is computationally cheap to approximate. In the last 20 years, simulation-based approaches have flourished to bypass the likelihood intractability by implicitly making use of it via model simulations. The most studied class of simulation-based inference methods is arguably approximate Bayesian computation (ABC). For ABC, sequential Monte Carlo (SMC-ABC) is the state-of-art sampler. However, since the ABC paradigm is intrinsically wasteful, sequential ABC schemes can benefit from well-targeted proposal samplers that efficiently avoid improbable parameter regions. In [1] we construct novel proposal samplers that are conditional to summary statistics of the data. In a sense, the proposed parameters are "guided" to rapidly reach regions of the posterior surface that are compatible with the observed data. This speeds up the convergence of these sequential samplers, thus reducing the computational effort, while preserving the accuracy in the inference. We provide a variety of guided samplers easing inference for challenging case-studies, including multimodal posteriors, highly correlated posteriors, hierarchical models with high-dimensional summary statistics.

References

[1] Picchini, Umberto, and Tamborrino, Massimiliano. "Guided sequential schemes for intractable Bayesian models." arXiv:2206.12235 (2022).

Bayesian Computation for Discretely Observed Stochastic Partial Differential Equations

Thorben Pieper[1]

[1] *Technical University Delft, The Netherlands, T.Pieper@tudelft.nl*

This talk concerns the estimation of model parameters θ for a discretely observed, semi-linear stochastic partial differential equation (SPDE). When observations of the associated process X are sparse, both space and/or time, such the statistical problem becomes notoriously difficult due to intractable likelihoods.

From a computational Bayesian viewpoint, this can be considered a *missing data problem*, in which case a natural solution is the *data augmentation algorithm.*

In its simplest form, it is a Gibbs sampler that iterates between updating the missing path segments of X and parameters θ. This solution faces two key challenges: Firstly, a careless implementation of the data augmentation algorithm leads to reducible Markov chains.

Secondly, the Gibbs sampler requires proposals of the missing path, i.e. an infinite dimensional diffusion bridge. As these are in general intractable, computational efficiency of the algorithm relies on 'good' proposals that closely resemble the true path conditioned on the observations.

These challenges are known from the finite dimensional case and have been investigated in the past twenty years. In this talk I will present current progress of lifting previous solutions from the finite to the infinite dimensional case. Particular attention will be paid to the construction of 'guided processes' for the simulation of infinite dimensional diffusion bridges, based on previous work by [1]. This requires a solid understanding of such bridge processes - an open problem that has so far received surprisingly little attention, with few exceptions such as the linear case ([2]).

The performance of the proposed solutions will be presented in numerical experiments. This talk is based on a project with Frank van der Meulen and Aad van der Vaart.

References

[1] Frank van der Meulen and Moritz Schauer, Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals, *Electronic Journal of Statistics*, 11(1), (2017) 2358-2306.

[2] B. Goldys and B. Maslowski,The ornstein-uhlenbeck bridge and applications to markov semigroups, *Stochastic Processes and their Applications*, 118(10), (2008), 1738-1767.

Parameter estimation of discretely observed interacting particle systems

Vytautė Pilipauskaitė[1], Chiara Amorino[2], Akram Heidari[2] and Mark Podolskij[2]

[1] *Aalborg University, Denmark, vypi@math.aau.dk*
[2] *University of Luxembourg, Luxembourg*

In this talk, we discuss the problem of joint estimation of parameters in the drift and diffusion coefficients of a system of N interacting particles associated with the McKean–Vlasov SDE. We propose a contrast function for discrete observations of the system over a fixed time interval $[0, T]$, using the likelihood of the corresponding Euler approximation. We prove that our minimum contrast estimator is consistent when the discretization step Δ_n and the number of particles N satisfy $\Delta_n \to 0$ and $N \to \infty$, and is asymptotically normal when the additional condition $N\Delta_n \to 0$ holds. The talk is based on [1].

References

[1] Amorino, C., Heidari, A., Pilipauskaitė, V., Podolskij, M. Parameter estimation of discretely observed interacting particle systems, `arXiv:2208.11965 [math.ST]`.

Parameter Estimation in Second-Order SDEs using Strang Estimator

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

[1] *University of Copenhagen, Copenhagen, predrag@math.ku.dk, susanne@math.ku.dk*
[2] *University of Grenoble Alpes, Grenoble, adeline.leclercq-samson@univ-grenoble-alpes.fr*

Second-order stochastic differential equations (SDEs) arise naturally in various fields, such as physics and biology, where the underlying process is described by position and velocity. We consider second-order SDEs and propose a parameter estimation method by transforming the second-order SDE into a system of first-order SDEs, introducing an auxiliary variable representing velocity. This transformation results in a hypoelliptic first-order system, where the noise terms are only on the velocity variables. Additionally, the resulting model is partially observed, as the velocity variables are unobserved. Standard methods such as the Euler-Maruyama scheme fail to estimate the parameters accurately in this setup. Instead, we propose two Strang splitting scheme estimators, for both complete and partially observed cases, and demonstrate that both estimators are consistent and asymptotically normal. The estimator in the complete case is efficient with the standard asymptotic distribution. In contrast, the estimator in the partial case has a slightly larger variance for the diffusion parameters due to lost information. We illustrate the theoretical results of the stochastic Jansen and Rit neural mass model (JR-NMM).

Local inference for the mean and covariance of functional data

Alessia Pini[1]

[1] *Università Cattolica del Sacro Cuore, Milan, alessia.pini@unicatt.it*

When testing hypotheses on functional data, it is often important to select the portions of the domain of the data responsible for the rejection of the null hypothesis. This problem gave birth to a lively area of research in FDA that is typically referred to as local inference. Statistical methods in this area are of two types: inference can be tackled either by developing confidence bands for functional data (e.g., [2]), or by testing the null hypothesis locally along the data domain. This talk focuses on the latter type of methods, which are based on the definition of an adjusted p-value function on the data domain. In particular, I will start by presenting methods that can be used to test hypotheses on the mean of functional data, which are based on a control of either the family-wise error rate ([4, 1]) or of the false discovery rate [3]. Finally, I will present a method to perform local inference on the covariance function of data.

References

[1] Abramowicz, K., Häger, C.K., Pini, A., Schelin, L., Sjöstedt de Luna, S., Vantini, S.: Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scand. J. Stat.* 45 (4) (2018)

[2] Liebl, D., Reimherr, M.: Fast and Fair Simultaneous Confidence Bands for Functional Parameters. *arXiv:1910.00131* (2019)

[3] Lundtorp Olsen, N., Pini, A., Vantini, S.: False discovery rate for functional data. *TEST* 30, (2021) 789–809

[4] Pini, A., Vantini, S.: Interval-wise testing for functional data. *J. Nonparametr. Stat.* 29(2), (2017) 407–424

Inversion of large dimensional sparse positive definite matrices for functional data analysis

Krzysztof Podgórski[1]

[1] *Lund University, Sweden, Krzysztof.Podgorski@stat.lu.se*

The problem of inverting large sparse positive definite matrices is at the center of many high-dimensional statistical methods. Recently, a highly efficient dyadic algorithm has been proposed for diagonalization and inversion of the band matrices, i.e. matrices that are zero outside of a narrow band of the entries around the main diagonal. Random permutation of a band matrix creates a sparse matrix for which it is easy to find the permutation that reverses it back to the band form. Using this simple observation, we propose an algorithm that allows efficient diagonalization and thus also inversion of a class of sparse matrices that can be decomposed into a small dimensional block and a number of not connected blocks of permuted band matrices. We formulate and prove mathematical results as well as compare efficiency of our algorithms with other existing methods of inverting high-dimensional sparse matrices. The relation to functional data analysis by the means of orthogonalized splines is discussed.

References

[1] Liu, X., Nassar, H., Podgórski, K. (2022) Dyadic diagonalization of positive definite band matrices and efficient B-spline orthogonalization. J. Comp. Appl. Math., 414, DOI: 10.1016/j.cam.2022.114444

[2] Basna, R., Nassar, H., Podgórski, K. (2022) Data driven orthogonal basis selection for functional data analysis. J. Mult. Analysis , 189, pp. 104868, DOI: 10.1016/j.jmva.2021.104868

On Lasso estimation for the drift function in diffusion models

Gabirela Ciolek[1], Dmytro Marushkevych[2] and **Mark Podolskij**[3],

[1] *University of Luxembourg, Luxembourg, gabriela.ciolek@uni.lu*
[2] *University of Copenhagen, Copenhagen, dmytro.m@math.ku.dk*
[3] *University of Luxembourg, Luxembourg, mark.podolskij@uni.lu*

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.

Transport ABC

Dennis Prangle[1] and Cecilia Viscardi[2]

[1] *University of Bristol, UK, dennis.prangle@bristol.ac.uk*
[2] *University of Florence, Italy, cecilia.viscardi@unifi.it*

We describe an efficient method to propose parameter values within a ABC-SMC algorithm (Approximate Bayesian Computation within the Sequential Monte Carlo framework) based on learning a transport map. As a new transport map must be trained automatically in each ABC-SMC iteration, a naive approach can easily fail due to over-training or under-training. We describe methodology to avoid these problems. We also present examples where the method provides efficiency gains over standard ABC-SMC.

Optimal Exploration in Reinforcement Learning via Frank-Wolfe

Po-An Wang[1], Ruo-Chun Tzeng[1] and **Alexandre Proutiere**[1]

[1] *KTH, Royal Institute of Technology, Sweden, {wang9, rctzeng, alepro}@kth.se*

We study the problem of active pure exploration with fixed confidence in Reinforcement Learning, where the learner's objective is to identify an efficient policy with a given level of certainty while minimizing her sampling budget. For this problem, instance-specific lower bounds on the expected sample complexity reveal the optimal exploration process an Oracle algorithm would apply. This process solves an optimization problem whose tractability strongly depends on the structural properties of the environment, but may be instrumental in the design of efficient learning algorithms. We give several examples of this optimization problems. We devise Frank-Wolfe-based Sampling (FWS), a simple algorithm whose sample complexity matches the lower bounds for a wide class of structured bandit problems. The algorithm is computationally efficient as, to learn and track the optimal proportion of arm draws, it relies on a single iteration of Frank-Wolfe algorithm applied to the lower-bound optimization problem. We apply FWS to various pure exploration tasks, including best arm identification in unstructured, thresholded, linear, and Lipschitz bandits. Despite its simplicity, FWS is competitive compared to state-of-art algorithms.

Optimizing the allocation of trials to sub-regions in multi-environment crop variety testing for multi-annual experiments

Maryna Prus[1], Hans-Peter Piepho[2]

[1] *Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Stuttgart, maryna.prus@uni-hohenheim.de, maryna.prus@ovgu.de*
[2] *Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Stuttgart, hans-peter.piepho@uni-hohenheim.de*

New crop varieties are usually evaluated for their performance in a target population of environments (TPE). This evaluation requires conducting randomized field trials at several environments sampled from the TPE. Such trials are called multi-environment trials (MET). If the TPE is large and can be suitably stratified along geographical borders or agro-ecological zonations, it may be advantageous to subdivide the TPE into sub-regions. If the same set of genotypes is tested at a number of locations in each of the sub-regions, a linear mixed model may be fitted with random genotype-within-sub-region effects. The first analytical results to optimizing allocation of trials to sub-regions have been obtained in [1]. That paper considers only a single year of trials. However, in practice the responses are usually being observed during several years. In this work we consider the extended linear mixed model that incorporates the influence of the years. We propose an analytical solution for optimal allocations of trials and illustrate the obtained results by a real data example.

References

[1] Prus, M. and Piepho, H.-P. (2021). Optimizing the allocation of trials to sub-regions in multi-environment crop variety testing. Journal of Agricultural, Biological and Environmental Statistics, 26, 267–288.

Periodogram for spatial point patterns: 59 years later

Tuomas Rajala[1], Jake Grainger[2], Sofia Olhede[2] and David J Murrell[3]

[1] *Natural Resources Institute Finland, Finland, tuomas.rajala@luke.fi*
[2] *École polytechnique fédérale de Lausanne, Switzerland, jake.grainger@epfl.ch, sofia.olhede@epfl.ch*
[3] *University College London, United Kingdom, d.murrell@ucl.ac.uk*

Since its introduction in 1964 by Bartlett, the spatial point pattern periodogram, a second order summary statistic, has seen only minor usage in statistical applications. The main problem in using the periodogram, it seems, is that it is quite hard to understand what its target estimand, the spectral density aka power spectrum, actually means for point processes. Classical interpretations from time-series analysis such as ”energy of the signal” or ”variance of the spectral dual process” make no immediate sense when the ”signal” is a finite set of random locations. Spatial product-densities, most notably the pair correlation function, have instead become popular in applications thanks to their direct interpretability. But even though the spectral density is linked to the pair correlation function by essentially being its Fourier pair, a carefully estimated power spectrum might be able to provide some benefits over its spatial twin.

We revisit Bartlett's periodogram, and introduce modifications that improve its statistical properties [1]. We discuss the challenges of defining spectral analysis in a spatial point process setting, and demonstrate spectral properties that have potential use in spatial analysis.

References

[1] Rajala T, Olhede S, Grainger J and Murrell DJ (2023): What is the Fourier Transform of a Spatial Point Process? https://arxiv.org/abs/2009.01474

Goodness of fit in beta regression

Max Raner[1]

[1] *Uppsala University, Uppsala, max.raner@math.uu.se*

Interval-bounded outcomes, such as rates and proportions, often exhibit asymmetry and heteroscedasticity in a regression context. One way in which to model this, is by so called beta regression. [1] proposed a model similar to a GLM by using a reparameterised beta distribution in terms of its mean and a precision parameter, for the response, which then naturally models both skew and varying dispersion. It has since grown increasingly popular, with several extensions developed for it, and made easily available by e.g. the R package `betareg` [2]. No distributional tests for overall fit are available for beta regression, but a goodness of fit statistic can be repurposed for beta regression from the leCessie test for logistic regression [3]. By simulation the test statistic is shown to behave well under the null-hypothesis, and preliminary results would seem to suggest similar performance as in the logistic regression case in terms of power—at least for some types of model deviation.

References

[1] Silvia Ferrari and Francisco Cribari-Neto. "Beta Regression for Modelling Rates and Proportions." Journal of Applied Statistics 31, no. 7 (2004): 799-815.

[2] Cribari-Neto, Francisco and Achim Zeileis (2010). "Beta Regression in R". In: Journal of Statistical Software 34.2, pp. 1–24.

[3] Cessie, S. le, and J. C. van Houwelingen. "A Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods." Biometrics 47, no. 4 (1991): 1267–82.

AI-based precision pathology: applications of deep learning to extract novel prognostic information from gigapixel breast cancer histopathology images in large retrospective cohort studies

Mattias Rantalainen[1]

[1] *Karolinska Institutet, Department of Medical Epidemiology and Biostatistics, mattias.rantalainen@ki.se*

Precision medicine has the potential to substantially improve cancer patient outcomes, but it requires precision diagnostic solutions to be effective. However, molecular diagnostics remains expensive, limiting patient access and imposing a high economic burden on healthcare systems.

To address these challenges we develop, validate, translate and implement AI-based histopathology image analysis solutions for image-based phenotyping and patient stratification, with applications both in the clinical setting and for cancer research.

I will present results from a few recent studies based on large retrospective population representative breast cancer cohorts that include gigapixel histopathology images, registry-based clinical information, and molecular profiling data. We apply multiple deep learning strategies to learn histopathology image representations that provide independent prognostic value, aiming to improve risk stratification of breast cancer patients with intermediate risk for progression. Validation results (time-to-event) are presented from independent external validation studies.

Testing the presence of directions in point processes on directed linear networks

Jakob G. Rasmussen[1]

[1] *Aalborg University, Aalborg, Denmark, jgr@math.aau.dk*

A linear network is a set of line segments approximating real life networks, such as road or river networks. The locations of events or objects on such networks can be modeled by point processes on linear networks (see e.g. the overview paper [1] for various models).

Many networks have natural directions, such as the direction of the flow in a river network, and it may often be natural to take these directions into account when constructing a point process model. However, point processes on linear networks typically do not incorporate information on directions. On the other hand, [2] assumes a directed linear network, and develops point process models incorporating directions.

This poses a natural question: If a point pattern dataset has been observed on a directed linear network, has the point pattern evolved following the directions or not? Answering this question provides a good starting point for choosing an appropriate class of models for the data. In this talk, Markov properties imposed by directionality are considered, and used for developing tests for whether or not a point pattern dataset has evolved according to the directions. The tests are applied to simulated data coming from models with or without directionality.

References

[1] A. Baddeley, G. Nair, S. Rakshit, G. McSwiggan, T. M. Davies, Analysing point patterns on networks - A review, *Spatial Statistics*, 42, (2021).

[2] J. G. Rasmussen and H. S. Christensen, Point Processes on Directed Linear Networks, *Methodol Comput Appl Probab*, 23, (2021) 647-667.

A new quantum graph community detection method for reducing the complexity of the molecular structure computation

Hannu Reittu[1], Werner Dobrautz[2] and Ville Kotovirta[1]

[1] *VTT Technical Research Centre of Finland, first.last@vtt.fi*
[2] *Chalmers University of Technology, Sweden, dobrautz@chalmers.se*

The Authors [1], suggest a modularity maximization based graph community detection method for finding approximate diagonal blocks in molecular Hamiltonian matrices, H, allowing computation of the corresponding energy spectrum of the molecule. The graph has a weight matrix W with zero diagonal and with off-diagonal elements $w_{ij} = |H_{ij}|$. In this work we suggest to uses a different approach for community detection based on Szemerédi's regularity lemma [2, 3]. A basic concept of this theorem is so called ϵ-regularity of a bipartite graph. The corresponding decision problem can be formulated as a quadratic unconstrained binary optimization problem (QUBO), which can be solved on quantum computers. For a graph with k communities the method needs $k-1$ rounds of partitioning of the graph in two. Each partition corresponds to finding minimum of the following QUBO $L(S) = \sum_{ij}(s_i s_j(d-w_{ij})+s_i(1-s_j)(w_{ij}-d))$, d is arithmetic mean of off-diagonal weights and $s_i \in \{0,1\}$ are binary variables for each node. Variables s corresponding to the minimum are indicators of one part. Next, both parts are treated as single communities or otherwise they are further partitioned until no new communities can be found.

Our method is well suited for quantum computing. First experiments yielded promising result for the water molecule.

References

[1] Susan. M Mniszewski et al.: Reduction of the molecular hamiltonian matrix using quantum community detection, *Sci Rep* , 11, 4099 (2021).

[2] Reittu, H., Kotovirta, V., Leskelä, L., Rummukainen, H., Räty, T. Towards analyzing large graphs with quantum annealing, In C. Baru, et al. (Eds), Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019 (pp. 2457-2464).

[3] T. Tao: Szemerédi's regularity lemma revisited, *Contributions to Discrete Mathematics* 1(1), 2006.

Inverse optimal control for linear-quadratic systems

Axel Ringh[1], and Han Zhang[2]

[1] *Department of Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, Gothenburg, Sweden, axelri@chalmers.se*
[2] *Department of Automation, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, zhanghan_tc@sjtu.edu.cn*

Dynamical systems are essential for modeling and analysis in many different areas, including engineering, biology, and economics. In fact, behavior observed in nature can often be modeled by dynamical systems that operate by minimizing cost or maximizing reward; this type of behavior is known as optimal control. However, a problem in applications of optimal control is to design the cost function. The reason is that the cost function must be carefully selected and tuned to the contextual environment in order to induce an appropriate control response. Inverse optimal control is a method to try to overcome this issue. In the inverse problem, the task is to identify the underlying cost function with respect to which observed behavior is optimal.

Here, I will present part of our work on inverse optimal control for linear-quadratic systems [1, 2]. In particular, we formulate an estimator for the underlying (unknown) quadratic cost function from observed optimal state trajectories. This estimator is the minimizing argument to a convex optimization problem, and we prove that the estimator is statistically consistent, that is, that it converges in probability to the true underlying cost as the number of observed trajectories goes to infinity.

References

[1] Han Zhang and Axel Ringh. Inverse linear-quadratic discrete-time finite-horizon optimal control for indistinguishable homogeneous agents: A convex optimization approach. *Automatica*, 148:110758, 2023.

[2] Han Zhang and Axel Ringh. Statistically consistent inverse optimal control for discrete-time indefinite linear-quadratic systems. *arXiv preprint arXiv:2212.08426*, 2022.

Is there a cap on how long a human can live? Truncation, censoring and extreme value modelling

Holger Rootzen[1]

[1] *Chalmers and Gothenburg University hrootzen@chalmers.se*

There is sustained and widespread interest in understanding the limit, if any, to the human lifespan. Apart from its intrinsic interest, changes in survival at extreme ages, say 105 and over, have implications for the biology of ageing and for the sustainability of social security systems. Recent analyses of data on the oldest human lifespans have led to competing claims about survival and to controversy, often due to misunderstandings about the selection of data and to inappropriate use of statistical methods. One central question is whether the endpoint of the underlying lifetime distribution is finite. This talk discusses the particularities associated with data on extreme lifespans, presents models from Extreme Value Statistics and Demography for their analysis, and outlines ways of handling the truncation and censoring often present in the data. It provides a critical assessment of earlier work and illustrates the ideas through novel analysis of new datasets on 105+ year lifetimes. The talk is based on a 2022 review paper in Annual Review of Statistics and its Application, written together with Léo Belzile, Anthony Davison, Jutta Gampe and Dimitrii Zholud, and on three earlier papers written together with subsets of this group.

Non-parametric modelling of microbial community dynamics from short time series

Chandler Ross[1], Ville Laitinen[1], Guilhem Sommeria-Klein[1] and Leo Lahti[1]

[1] *University of Turku, Finland, cmross@utu.fi*

Mechanistic models of microbial community dynamics are often limited in their ability to characterise the variation observed in naturally occurring microbiomes. An alternative approach is to characterize the dynamical landscape using non-parametric models. Modelling the observed fluctuations in longitudinal time series can inform us about stability, resilience and other key aspects of community dynamics. However, existing approaches rely on access to long, dense and continuous time series, whereas available time series usually consist of a few time points. We show that this limitation can be partly overcome by combining information across multiple short time series. We propose an approach to decompose the dynamics into drift (deterministic) and diffusion (stochastic) components from short time series data using Gaussian processes. We discuss how this information allows us to predict the locations of the stable and tipping points of the system and quantify resilience. We benchmark the model on simulated data and then demonstrate the approach by characterising the dynamical landscape of the human gut microbiome based on short time series of the relative abundances of 130 bacterial genera in 76 individuals. We show that the model correctly captures the expected dynamics, including the bistability of certain bacterial genera.

Estimating the effect of observed heterogeneity in replication studies

Ying Jin[1], Kevin Guo[1] and **Dominik Rothenhäusler**[1]

[1] *Stanford University, Stanford, ying531@stanford.edu, kxguo@stanford.edu, rdominik@stanford.edu*

Several scientific disciplines have recently conducted large-scale reproducibility projects wherein independent laboratories attempt to replicate key findings in the field. A common finding in these projects is that many studies cannot be systematically replicated. Many authors have identified treatment effect heterogeneity as a contributor to these low replication rates. The idea is that large effects in one population might be reduced or even absent in replication populations if treatment efficacy varies across individuals or contexts. In this talk, we discuss how to use statistical methods from the "generalizability" literature to estimate the impact of *measured* heterogeneity in a given experiment-replication pair. We use these methods to decompose discrepancies for several directly-replicated experiments in behavioral science.

Age-specific transmission dynamics of SARS-CoV-2 during the first two years of the pandemic

Ganna Rozhnova[1,2,3], Otilia Boldea[4], Amir Alipoor[4], Sen Pei[5] and Jeffrey Shaman[5,6]

[1] *Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands, g.rozhnova@umcutrecht.nl*

[2] *Center for Complex Systems Studies (CCSS), Utrecht University, Utrecht, The Netherlands*

[3] *BioISI—Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal*

[4] *Department of Econometrics and OR and CentER, Tilburg School of Economics and Management, Tilburg University, Tilburg, The Netherlands, o.boldea@tilburguniversity.edu, a.h.noroozalipoorbaboli@tilburguniversity.edu*

[5] *Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, USA, sp3449@cumc.columbia.edu, jls106@cumc.columbia.edu*

[6] *Columbia Climate School, Columbia University, New York, USA*

The COVID-19 pandemic manifested as multiple waves shaped by complex interactions between virus variants, interventions, and immunity in the population. Understanding age-specific transmission dynamics of SARS-CoV-2 is crucial for informing policy decisions. We developed an inference-based model to reconstruct the burden of true infections and hospitalizations in children, adolescents and adults over the seven waves of four variants (wild-type, Alpha, Delta, Omicron BA.1) during the first two years of the pandemic, using the Netherlands as the motivating example [1]. We find that reported cases are a considerable underestimate and a generally poor predictor of true infection burden, especially because case reporting differs by age. The contribution of children and adolescents to total infection and hospitalization burden increased with successive variants and was largest during the Omicron BA.1 period. Before the Delta period, almost all infections were primary infections occurring in naive individuals. During the Delta and Omicron BA.1 periods, primary infections were common in children but relatively rare in adults who experienced either re-infections or breakthrough infections. Our approach can be used to understand age-specific epidemiology through successive waves in other countries where random community surveys uncovering true SARS-CoV-2 dynamics are absent but basic surveillance and statistics data are available.

References

[1] Boldea, O., Alipoor, A., Pei, S., Shaman, J., Rozhnova, G. (2022). Age-specific transmission dynamics of SARS-CoV-2 during the first two years of the pandemic. https://doi.org/10.48550/arXiv.2212.13470.

Distribution Error of the Euler Scheme for Stochastic Delay Differential Equations with Fractional Additive Noise

Orimar Sauri[1]

[1] *Aalborg University, Department of Mathematical Sciences, DK, osauri@math.aau.dk*

In this talk we consider the Euler scheme for a class of stochastic delay differential equations (SSDE) driven by a linear fractional β-stable Lévy motion with index $H \in (0,1)$. We stablish the consistency of the scheme and study the limit distribution of the normalized error process. We show that in the rough case, i.e. when $H < 1/\beta$, the rate of convergence of the simulation error is of order $\Delta_n^{H+1-1/\beta}$ (Δ_n is the step-size of the scheme) and that the limit process is once again the solution of an SDDE. We further illustrate how our results can be used in the context of option pricing under rough volatility.

Computing Derivatives of Programs with Continuous and Discrete Randomness

Frank Schäfer[1]

[1] *CSAIL, Massachusetts Institute of Technology, Cambridge, MA 02139, USA, franksch@mit.edu*

Stochastic differential equations (SDEs) arise naturally in quantum information processing and metrology tasks when the quantum system is subjected to an observation process. Often, these tasks require the design of feedback controllers that are robust to inherent fluctuations or the estimation of system parameters. The time evolution of the filtering distribution is described by quantum filtering theory. In the first part of this talk, we model a controller by a neural network and discuss how its parameters can be optimized by gradient descent. To compute the relevant gradients of the objective function with respect to the neural-network parameters, we use automatic differentiation (AD) on the SDE solver operations and memory-efficient custom rules [1]. If the programs that we differentiate exhibit discrete stochastic behaviors, traditional AD fails. In the second part, we discuss our new reparameterization-based methodology for programs containing discrete randomness [2]. We showcase how this method gives an unbiased and low-variance estimator which is as automated as traditional AD mechanisms.

References

[1] Schäfer, F., *et al.*, Control of Stochastic Quantum Dynamics by Differentiable Programming, *Mach. Learn.: Sci. Technol.*, **2**, 035004 (2021).

[2] Arya, G., *et al.*, Automatic Differentiation of Programs with Discrete Randomness, *Adv. Neural Inf. Process. Syst.*, 10435–10447 (2022).

Graphical and summary diagnostics for node level adequacy in Bayesian hierarchical models

Ida Scheel[1], Jørund Gåsemyr[1]

[1] *Department of Mathematics, University of Oslo idasch@uio.no*

In order to check the adequacy of Bayesian hierarchical models, one can look for conflict between all information informing a parameter of the model by representing the parameters as nodes in a directed acyclic graph. This is useful since all neighbours in the graph potentially provide, potentially conflicting, information on a node. Building on this idea originating from [1], [2, 3] constructed node based conflict measures that have been shown to be well calibrated. Sharing the idea of considering potential conflict between separate information sources for each node, [4] constructed a graphical diagnostic lcp. This can be used both to identify conflict at each node, and provide insight into the nature of the conflict, hence being more informative than summary diagnostics. In the present paper we link these two ideas together by constructing a new diagnostic plot iiic-lcp that is supplementary to lcp, but builds on the framework of [3]. It has the advantage over the lcp with the possibility to display curves corresponding to different parameters in the same plot, saving space and easing comparisons, particularly useful for sets of parameters representing exchangeable groups. We show how to visually read some of the conflict measures defined in [3] from iic-lcp, combining graphical and summary diagnostics.

References

[1] O'Hagan, A. (2003). HSSS model criticism. In Green, P. J., Richardson, S., and Hjort, N. L. (eds.), *Highly Structured Stochastic Systems*, 423–444. Oxford: Oxford University Press.

[2] Dahl, F. A., Gåsemyr, J., and Natvig, B. (2007). A robust conflict measure of inconsistencies in Bayesian hierarchical models. *Scand. J. Stat.*, 34: 816–828.

[3] Gåsemyr, J. and Natvig, B. (2009). Extensions of a conflict measure of inconsistencies in Bayesian hierarchical models. *Scand. J. Stat.*, 36: 822–838.

[4] Scheel, I., Green, P. J., and Rougier, J. C. (2011). A graphical Ddiagnostic for identifying influential model Choices in Bayesian hierarchical models. *Scand. J. Stat.*, 38: 529–550.

A semi-group approach to PCA for extremes

Martin Schlather[1] and Felix Reinbott[2]

[1] *Universität Mannheim, Mannheim, martin.schlather@uni-mannheim.de*
[2] *Otto-von-Guericke-Universität Magdeburg, Magdeburg, felix.reinbott@ovgu.de*

Principal Component Analysis for extremes has been an open problem for a long time and has been dealt with by several authors in meanwhile. Since the maximum-operator defines algebraically a semi-group, we offer here a general approach to PCA based on semi-groups that allows for further generalization, so that the classical PCA reappears as the Gaussian case and variable selection in regression analysis is also included as special case.

Prediction Can Be Safely Used as a Proxy for Explanation in Causally Consistent Bayesian Generalized Linear Models

Maximilian Scholz[1] and Paul-Christian Bürkner[1]

[1] *Cluster of Excellence SimTech, University of Stuttgart, Germany,*
maximilian.scholz@simtech.uni-stuttgart.de

Bayesian modeling provides a principled approach to quantifying uncertainty in model parameters and structure and has seen a surge of applications in recent years. Despite a lot of existing work on an overarching Bayesian workflow, many individual steps still require more research to optimize the related decision processes. One such practice is the use of prediction as a proxy for explanation in the context of latent inferential goals. In this talk we discuss the relationship of predictive performance and parameter recoverability and present results from a large simulation study of Bayesian generalized linear models that indicate that for causally consistent models predictive performance can be safely used as a proxy for parameter recoverability.

Parameter estimation for time changed fractional Brownian motion

Nataliya Shchestyuk[1,2], Stepan Mazur[1] Mark Podolskij[3] and Farrukh Javed[4]

[1] *School of Business, Örebro University, Örebro, nataliya.shchestyuk@oru.se*
[2] *Department of Mathematics, "Kyiv-Mohyla Academy" University, Kyiv*
[3] *Department of Mathematics, Luxembourg University, Luxembourg, mark.podolskij@uni.lu*
[4] *Department of Statistics, Lund University, Lund, farrukh.javed@stat.lu.se*

Fractional Brownian motion (fBM) belongs to the class of long-range dependent systems with self-similarity property and has been widely used in different applications. Mathematically speaking, the scaled fBm is fully characterised by its scaling parameter $\sigma > 0$ and Hurst parameter $H \in (0, 1)$. But in spite of many obvious advantages, for many real-life data with long-range dependence, the classical fBM with Gaussian property cannot be considered an appropriate model. For example the classical fBM cannot model the real time series with apparent constant time periods (called also trapping events), which are often observed in data sets recorded within various fields. One of the possible solutions is the time-changed fBM $B^H_{L_t}$ with α-stable Lévy subordinator $(L_t)_{t\geq 0}$.

We construct consistent estimators of the parameters (σ, α, H) for the time-changed fBM $X_t = B^H_{L_t}$. Our approach is based on the limit theory for stationary increments of a linear fractional stable motion [1]. We use these techniques, combine negative power variation statistics and their empirical expectations and covariances to obtain consistent estimates of (σ, α, H). We show that X_t is a symmetric H/α-stable Lévy process and for $p < \alpha$ we deduce the law of large numbers. The above law of large numbers immediately gives a consistent estimator of the self-similarity parameter H/α of X_t. In order to estimate the other parameters of the model we use the some identities, which has been shown in [2]. Finally we present the statistical inference and prove some weak limit theorems for the all parameters (σ, α, H) using classical delta-method.

References

[1] Mazur, S., Otryakhin D. and Podolskij M., Estimation of the linear fractional stable motion, Bernoulli, 26(1), (2020) 226-252.

[2] Dang, T.T.N., Istas, J., Estimation of the Hurst and the stability indices of a H-self-similar stable process, Electronic Journal of Statistics, 11(2), (2018) 4103-4150.

Variance of entropy estimation for testing time-varying regimes

Andrey Shternshis[1] and Piero Mazzarisi[2]

[1] *Scuola Normale Superiore, Pisa, Italy, andrey.shternshis@sns.it*
[2] *University of Siena, Siena, Italy, piero.mazzarisi@unisi.it*

Shannon entropy is the most common metric to measure the degree of randomness of time series. Real-world systems may be in general non stationary, with an entropy value that is not constant in time. We propose a hypothesis testing procedure to test the null hypothesis of constant Shannon entropy for time series. To this aim, we find an unbiased approximation of the variance of the entropy's estimator up to the order $O(n^{-4})$ with n the sample size. In order to characterize the variance of the estimator, we obtain the explicit formulas of the central moments for the multinomial distribution, which describe the distribution of the Shannon entropy. We propose a novel criterion to find the optimal length of the rolling window used for estimating the time-varying Shannon entropy. We corroborate our findings by using the novel methodology to test for time-varying regimes of entropy for stock price dynamics, in particular considering the case of meme stocks in 2020 and 2021.

The effect of cross-border mobility on the spread of COVID-19 in Nordic countries

Mikhail Shubin[1], Hilde Kjelgaard Brustad[2], Felix Günther[3], Mikko Kivelä[1], Gianpaolo Scalia Tomba[2,4], Lasse Leskelä[1]

[1] *Aalto University, Espoo, Finland, mikhail.shubin@aalto.fi*
[2] *University of Oslo, Norway*
[3] *Stockholm University, Sweden*
[4] *University of Rome Tor Vergata, Italy*

The countries of Denmark, Finland, Norway, and Sweden have a significant cross-border traffic (the number of yearly border crossing compatible to the counties' population sizes) including the period of the COVID-19 pandemic. Such a high mobility could have impacted the spread of the virus.

To asses this impact, we build a Nordic model combining the disease dynamic and data from Denmark, Finland, Norway, and Sweden. We estimate the number of infectious people crossing the borders, the fraction of infections caused by cross-border transmission and the potential burden of these infections. We then extrapolate the possible effects of various kinds of possible interventions, such as border closures or reopening. The model and data integration could be used in the future for supporting decisions on interventions of future epidemics.

Item response theory for recommender systems: A multi-dimensional test that is adaptive and interpretable

Karl Sigfrid[1]

[1] *Stockholm University, Stockholm, karl.sigfrid@stat.su.se*

Questionnaire-based recommender systems rely on user input to match user preferences to options in a choice situation. Providing the input can be time consuming, which may negatively affect participation. The participation threshold may be lowered if there is an option to conclude the test without answering all question items. The test result should be close to the result that the participant would have gotten after answering all question items. In this talk we propose a method that extends the graded response model and the Maximum Information Criterion, used in Item Response Theory (IRT). The aim is to allow the user to control the length of the test. Furthermore, we want a simpler interpretation of multidimensional parameter estimates than we get from traditional multivariate IRT. To achieve this, we propose an adaptive algorithm selecting question items from multiple unidimensional scales. With simulated data and response data from a voting advice application project, we evaluate the accuracy of shorter tests with our adaptive method. When only a few test items are answered, our proposed method outperforms a static-order IRT test of equal length in identifying the best match. We expect that implementation of the method can increase participation in questionnaire-based recommender systems.

Graph flip processes related to dynamical systems

Matas Šileikis[1],

[1] *Institute of Computer Science of the Czech Academy of Sciences, Prague, Czechia, matas@cs.cas.cz*

We define a (discrete time) *graph flip process* starting with an arbitrary n-vertex graph G_0 where G_i is obtained form G_{i-1} by altering a subgraph induced by a uniformly chosen set of k vertices (with k fixed) according to a predefined rule that uses only the information about the induced subgraph. For example, the rule may be "if you see a clique, remove its edges" or "whatever graph you see, replace it by a random graph". Fixing constant $T > 0$, we discuss the concentration of trajectories $(G_0, \cdots, G_{Tn^2})$, as $n \to \infty$, from the perspective of the theory of dense graph limits. We show that the mean trajectory can be interpreted as a solution of a dynamical system on the space of graphons, which are elements of $L^\infty([0,1]^2)$. We describe properties of a few specific processes (including Erdős–Rényi process and triangle removal process) and state a couple of open questions.

Presented work appears in papers [1, 2], joint with P. Araújo (CSI CAS), F. Garbe (MU Brno), J. Hladký (CSI CAS), E. K. Hng (CSI CAS), F. Skerman (Uppsala U).

References

[1] Garbe, F., Hladký, J., Šileikis, M., Skerman, F. From flip processes to dynamical systems on graphons, arXiv:2201.12272.

[2] Araújo, P., Hladký, J., Hng, E.K., Šileikis, M., Prominent examples of flip processes, arXiv:2206.03884.

Regularity, Markov properties and numerical approximation of generalized Whittle-Matérn fields on metric graphs

Alexandre B. Simas[1]

[1] *Statistics Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), alexandre.simas@kaust.edu.sa*

In this talk we will consider the generalized Whittle-Matérn fields on compact metric graphs, which are defined through the fractional-order stochastic differential equation $L^\alpha(\tau u) = \mathcal{W}$ on Γ, where $L = \kappa^2 - \nabla(a\nabla)$ for (sufficiently nice) functions κ, a, and $\mathcal{W}$ is Gaussian white noise. The differential operator L is augmented with some suitable vertex conditions. We demonstrate the existence of a unique solution for a general class of vertex conditions and derive the regularity of the solution in the specific case of Kirchhoff vertex conditions. Furthermore, we propose a numerical solution based on a finite element approximation combined with a rational approximation of the fractional power $L^{-\beta}$. For the resulting approximation, strong mean squared error as well as the $L_2(\Gamma \times \Gamma)$-error of the covariance function of the solution are analyzed in the stochastic setting. Explicit rates of convergences are derived for all cases. We then study Markov properties of these fields. More precisely, we show that the generalized Whittle-Matérn field with exponent α is Markov if and only if $\alpha \in \mathbb{N}$. Further, if $\alpha \in \mathbb{N}$, a generalized Whittle–Matérn field u is Markov of order α, which means that the field u in one region $S \subset \Gamma$ is conditionally independent of u in $\Gamma \setminus S$ given the values of u and its $\alpha - 1$ derivatives on ∂S. The Markov property thus allows us to obtain an explicit characterization of u on a fixed edge e, revealing that the conditional distribution of u on e given the values at the two vertices connected to e is independent of the geometry of Γ. Finally, some examples and numerical illustrations will be given using our forthcoming `MetricGraph` package for the statistical software `R` [4].

This work is based upon [1, 2] and [3]. This is a joint work with David Bolin, Mihály Kovvcs, Vivek Kumar and Jonas Wallin. The `MetricGraph` package is a joint work with David Bolin and Jonas Wallin.

References

[1] Bolin, D., M. Kovács, V. Kumar, and A. B. Simas (2023). Regularity and numerical approximation of fractional elliptic differential equations on compact metric graphs. Preprint, arXiv:2302.03995.

[2] Bolin, D., A. B. Simas, and J. Wallin (2022). Gaussian Whittle–Matérn fields on metric graphs. Preprint, arXiv:2205.06163.

[3] Bolin, D., A. B. Simas, and J. Wallin (2023). Markov properties of Gaussian random fields on compact metric graphs. Preprint, arxiv:2304.03190.

[4] R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Manifold Bridges and Diffusion Means

Stefan Sommer[1]

[1] *University of Copenhagen, sommer@di.ku.dk*

We present schemes for simulating conditioned semimartingales taking values in Riemannian manifolds, Lie groups, and homogeneous spaces. Extending a guided bridge proposal approach used for simulating Euclidean bridges, the scheme replaces the drift of the conditioned process with an approximation in terms of a scaled radial vector field. This handles the fact that transition densities are generally intractable in geometric spaces. We prove the validity of the scheme by a change of measure argument, specifically handling the discontinuity of the radial vector field at the cut locus of the target.

The schemes can be directly used for numerically simulating bridges on manifolds, in importance sampling, and for approximating otherwise intractable transition densities. Furthermore, we discuss application of the bridge simulation techniques for estimating the underlying metric from data, and for finding the diffusion mean of observations. The latter can be done directly by optimizing the data likelihood approximated with the simulation scheme, or, with n observations, by conditioning the process to hit the diagonal of an n-times product of the manifold. The diagonal conditioning scheme avoids the nested iterated optimization algorithms that are often used for computing means in general geometric spaces such as for finding the Fréchet mean.

References

[1] Eltzner, B., Hansen P., Huckemann, S., Sommer, S., Diffusion Means in Geometric Spaces, *accepted for Bernoulli*, (2022)

[2] Jensen, M. H. and Sommer, S., Mean Estimation on the Diagonal of Product Manifolds, *Algorithms*, 15 (2022)

[3] Jensen, M. H. and Sommer, S., Simulation of Conditioned Semimartingales on Riemannian Manifolds, *arXiv:2105.13190*, (2021)

[4] Jensen, M. H., Joshi, S., and Sommer, S., Discrete-Time Observations of Brownian Motion on Lie Groups and Homogeneous Spaces: Sampling and Metric Estimation, *Algorithms*, (2022)

Local inference for functional linear mixed models

Helle Sørensen[1], Alessia Pini[2], Anders Tolver [3] and Simone Vantini[4]

[1] *University of Copenhagen, Denmark, helle@math.ku.dk*
[2] *Università Cattolica del Sacro Cuore, Milan, Italy, alessia.pini@unicatt.it*
[3] *University of Copenhagen, Denmark, tolver@math.ku.dk*
[4] *Politecnico di Milano, Italy, simone.vantini@polimi.it*

Local inference for functional data aims at identifying intervals of the time domain with statistical significance for a certain hypothesis. In principle, this involves infinitely many hypothesis tests. Interval-wise testing (IWT) is a permutation-based method that incorporates intra-curve correlation to compute an adjusted p-value function, defined on the same time domain as the data. In the talk, the basic IWT method for independent, one-dimensional data is extended to correlated and multivariate functional data, using a functional linear mixed model framework. More specifically, think of situations where several treatments are tested on the same subjects such that dependence between (possibly multivariate) data curves from the same subjects must be taken into account in order to have valid inference. I illustrate the method on gait data from horses.

Extremes of regularly varying stochastic volatility fields

Mads Stehr[1] and Anders Rønn-Nielsen[1]

[1] *Copenhagen Business School, Department of Finance, mast.fi@cbs.dk, aro.fi@cbs.dk*

We consider a spatial stationary stochastic volatility field $(Y_v Z_v)$, where Z is regularly varying and Y has lighter tails and is independent of Z. We make—relative to existing literature—rather general assumptions on the dependence structure of both fields. In particular this allows Y to be non-ergodic, in contrast to the typical assumption that it is i.i.d., and Z to be given by an infinite moving average.

Considering the stochastic volatility field on a general sequence of increasing index sets, we show the existence and form of a Y-dependent extremal functional generalizing the classical extremal index. More precisely, conditioned on the field Y, the extremal functional shows exactly how the extremal clustering of the (conditional) stochastic volatility field is given in terms of the extremal clustering of the regularly varying field Z and the realization of Y. Having obtained this, we show that two different types of cluster counting processes converge to a Poisson point process with intensity given in terms of the extremal functional.

Sequential Monte Carlo for infectious disease models - the Covid-19 case

Geir Storvik[1] and Alfonso Diz-Lois Palomares[2]

[1] *University of Oslo, geirs@math.uio.no*
[2] *Norwegian Institute of Public Health, Alfonso.Diz-LoisPalomares@fhi.no*

Sequential Monte Carlo (SMC) algorithms is a powerfull class of computational methods for performing inference in state space models. In [1], SMC was applied on a stochastic SEIR model for inference on national time-varying reproduction numbers for the Covid-19 epidemic in Norway. The method was in regular use in Norway during the pandemics and appears to be a powerful instrument for epidemic monitoring and management. Although some parameters were estimated as part of the SMC algorithm, most parameters were pre-estimated from other sources and thereafter fixed.

In this talk we will present the algorithms applied for performing estimates of *regional* reproduction numbers. We will further show how particle MCMC methods can be applied to estimate some of the parameters that previously have been fixed, making uncertainty bounds more reliable. Results will be based on daily hospitalisation and positive test incidences within the 11 counties in Norway.

References

[1] **Storvik, G.**, **Engebretsen S.**, **de Blasio, B.**, et al. (2023). A sequential Monte Carlo approach to estimate a time varying reproduction number in infectious disease models: the Covid-19 case. Accepted as discussion paper in JRSSA. *arXiv:2201.07590*.

Inferential methods for spatially dependent functional data with application to EEG data

Johan Strandberg[1], Lina Schelin [1] and Sara Sjöstedt de Luna[2]

[1] *Umeå School of Business, Economics and Statistics, Umeå University, johan.strandberg@umu.se, lina.schelin@umu.se*
[2] *Department of Mathematics and Mathematical Statistics, Umeå University, sara.sjostedt.de.luna@umu.se*

This work focuses on testing the equality of population means of two groups of curves in a situation where data are realizations from several functional random fields. The study is motivated by an investigation of event-related potential (ERP) curves collected using electroencephalography (EEG). The hypothesis is that an audio stimulus will provoke anxiety-related arousal that could be captured with EEG. The observed data consists of ERP signals at different channels (locations at the head) for a set of individuals under two different types of stimuli, and we want to test for functional differences between the two types. We propose to use a non-parametric approach based on permutations that takes into account the spatial dependence between functional data in different ways. An extensive simulation study is conducted to evaluate the performance of four different methods to perform the tests under different types of data-generating mechanisms. The results show that there is a large difference between the four methods on the resulting power of the test, which indicates the importance of a methodological choice. In general, a method that allows individuals to have different dependency structures works best in most scenarios, producing high power even in scenarios without spatial dependence.

Graphical models for infinite measures with applications to extremes and and Lévy processes

Kirstin Strokorb[1], Sebastian Engelke[2] and Jevgenijs Ivanovs[2]

[1] *Cardiff University, Wales, UK, StrokorbK@cardiff.ac.uk*
[2] *University of Geneva, Switzerland, sebastian.engelke@unige.ch*
[3] *Aarhus University, Denmark, jevgenijs.ivanovs@math.au.dk*

Conditional independence and graphical models are well studied for probability distributions on product spaces. We propose a new notion of conditional independence for any measure Λ on the punctured Euclidean space $\mathbf{R}^d \setminus \{0\}$ that explodes at the origin. The importance of such measures stems from their connection to infinitely divisible and max-infinitely divisible distributions, where they appear as Lévy measures and exponent measures, respectively. We characterize independence and conditional independence for Λ in various ways through kernels and factorization of a modified density, including a Hammersley-Clifford type theorem for undirected graphical models. As opposed to the classical conditional independence, our notion is intimately connected to the support of the measure Λ. Our general theory unifies and extends recent approaches to graphical modeling in the fields of extreme value analysis and Lévy processes. Our results for the corresponding undirected and directed graphical models lay the foundation for new statistical methodology in these areas.

Bayesian inference for stochastic shape models on phylogenetic trees

Sofia Stroustrup[1], Frank van der Meulen [2], Stefan Horst Sommer [3] and Rasmus Nielsen [1 4]

[1] *Center for GeoGenetics, University of Copenhagen*
[2] *Department of Mathematics, Vrije Universiteit Amsterdam*
[3] *Department of Computer Science, University of Copenhagen*
[4] *Departments of Integrative Biology and Statistics, University of California, Berkeley*

We consider the problem of modelling how form and structure of biological organisms i.e., morphology evolve over time. As morphological data is inherently correlated and non-linear, most naïve methods for analyzing morphology in an evolutionary context fail to capture the complexities of evolving shapes. We propose a novel approach for analyzing the evolution of morphology based on recent advances in mathematical shape analysis. We directly model the evolution of shapes as evolving according to a diffusion process in non-linear shape space and let the process evolve on a phylogenetic tree describing the evolutionary relationship between our data points. We apply Bayesian inference and a Markov Chain Monte Carlo sampling scheme for learning the parameters and the diffusion process conditioned on the data. Specifically, we construct a Metropolis-Hastings algorithm for jointly inferring the bridge process and the parameters by sequentially sampling either parameters or paths. We propose paths simulated from a guided process evolving on a fork and infer the posterior over the full tree by iteratively updating parts of the tree according to the Metropolis-Hastings scheme. For our sampling scheme we apply the Automatic Backward Filtering Forward Guiding algorithm [1] [2] to construct the guided processes from which we sample paths.

References

[1] Frank van der Meulen & Moritz Schauer. ”Automatic backward filtering forward guiding for Markov processes and graphical models.” arXiv preprint (2020).

[2] Marcin Mider, Moritz Schauer & Frank van der Meulen. ”Continuous-discrete smoothing of diffusions.” Electronic Journal of Statistics.

Normal approximation for Gibbs processes via disagreement coupling

Christian Hirsch [1], Moritz Otto [1] and **Anne Marie Svane** [2]

[1] *Aarhus University, hirsch@math.au.dk, otto@math.au.dk*
[2] *Aalborg University, annemarie@math.aau.dk*

Obtaining central limit theorems for functionals of point processes arising in topological data analysis is challenging, since these are not naturally expressed in terms of exponentially stabilizing score functions, which is the assumption of most existing approaches. For this reason, we derive a central limit theorem for weakly stabilizing functionals of Gibbs point processes, which applies to persistent Betti numbers. The result is obtained by combining a martingale approach originally developed for Poisson processes [1] with a recent construction of Gibbs processes called disagreement coupling [2]. The latter provides a coupling between Gibbs processes in finite and infinite domains by linking them to a common non-percolating Boolean model.

References

[1] M. D. Penrose, J. E. Yukich. Central limit theorems for some graphs in computational geometry. *Ann. Appl. Prob.*, 11:1005–1041, 2001.

[2] G. Last, M. Otto (2023): Disagreement coupling of Gibbs processes with an application to Poisson approximation. *Ann. Appl. Probab.*, to appear.

Applications of geostatistics in baseball

Dana Sylvan

Hunter College of the City University of New York, dsylvan@hunter.cuny.edu

Baseball is a game of stats, more so in recent years with the advent of "big data" and the emerging field of data science. Extensive statistical literature has been devoted to baseball data, see for example the recent text [2] and the references therein. However, comparatively less is known about the utility of computational tools specific to geostatistics to analyze performance of baseball players. The freely available Sportvision PITCHf/x data provides continuous location coordinates of each pitch using high-speed cameras. This detailed spatial information can be employed to visualize a batter's ability across regions in and around the strike zone.

This presentation summarizes classic geostatistics methodology, shows how it can be applied to baseball data, and presents comprehensive heat maps based on pre-pandemic Major League Baseball pitches. The stochastic process underlying batting ability is assumed to be a spatial Gaussian random field with isotropic covariance that is estimated from the aforementioned data. Spatial interpolation (kriging) is then used to obtain best estimates of heat maps of batting ability for individual players. Uncertainty of these estimates is assessed by using Monte Carlo simulations and resampling. An analog of the five-number summary in this spatial context is used to illustrate findings, showing estimated percentile heat maps for visualization, see [1] and [3] for details.

References

[1] Cross, J. and Sylvan, D., Modeling spatial batting ability using a known covariance matrix. *Journal for Quantitative Analysis of Sports*, 11 (3), (2015), 155-167.

[2] Marchi, M., Albert, J. and Baumer, B. Analyzing baseball data with R. Chapman & Hall, 2019.

[3] Sylvan, D., Using geostatistics to model and visualize batting ability in baseball. Proceedings, Mathsport International IX, Reading, UK, (2022), 107-111.

Improving sepsis treatment with personalized reinforcement learning using a multi branch dueling double deep Q Network

Selma Tabakovic[1] and Marina Axelson-Fisk [2]

[1] *Chalmers University of Technology, Gothenburg, selmat@chalmers.se*
[2] *Chalmers University of Technology, Gothenburg, marina.axelson-fisk@chalmers.se*

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection, and remains a leading cause of death in intensive care units worldwide [1]. An optimal treatment strategy is still unknown, leading to a significant variability in sepsis treatment [2].

Recently, deep reinforcement learning have shown promise as a decision-aiding tool for the administration of intravenous fluids and vasopressors to septic patients [3, 4]. However, these models are limited in their ability to accommodate high-risk and low-risk patients, and thus fail to provide personalized treatment recommendations [5].

To address this limitation, we propose a Multi-Branch Dueling Double Deep Q-Network (MB-DDDQN) model that incorporates patient characteristics to enable more personalized treatment recommendations. The MB-DDDQN model has multiple output layers, each of which is optimized for a specific patient profile. The model is trained using the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) database, which is a freely available database of intensive care unit patients admitted between 2001 and 2012 at the Beth Israel Deaconess Medical Center in Boston.

In conclusion, the proposed MB-DDDQN model has the potential to improve the efficacy of sepsis treatment by providing personalized recommendations based on patient profiles, ultimately leading to better treatment policies.

References

[1] Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., ... & Angus, D. C. (2016). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama*, 315(8), 801-810.

[2] Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11), 1716-1720.

[3] Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., & Ghassemi, M. (2017). Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*.

[4] Roggeveen, L., El Hassouni, A., Ahrendt, J., Guo, T., Fleuren, L., Thoral, P., ... & Elbers, P. W. (2021). Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. *Artificial Intelligence in Medicine*, 112, 102003.

[5] Lu, M., Shahn, Z., Sow, D., Doshi-Velez, F., & Li-wei, H. L. (2020). Is deep reinforcement learning ready for practical applications in healthcare? A sensitivity analysis of duel-DDQN for hemodynamic management in sepsis patients. In *AMIA Annual Symposium Proceedings* (Vol. 2020, p. 773). American Medical Informatics Association.

Exploration in reward machines with near-optimal regret

Hippolyte Bourel[1], Anders Jonsson[2], Odalric-Ambrym Maillard[3], **Sadegh Talebi**[4]

[1] *University of Copenhagen, hippolyte.bourel@di.ku.dk, m.shahi@di.ku.dk*
[2] *Universitat Pompeu Fabra, anders.jonsson@upf.edu*
[3] *Inria, odalric.maillard@inria.fr*

We study reinforcement learning for decision processes with Markovian dynamics but non-Markovian rewards, in which high-level knowledge in the form of a finite-state automaton is available to the learner. Such an automaton, often called *Reward Machine (RM)* [1], determines reward functions based on its internal state as well as events that are detected at various states in the environment. The associated decision processes is called an MDPRM, and we focus on average-reward MDPRMs in the regret setting. For a given MDPRM, there is an *equivalent cross-product* MDP, to which one can apply provably efficient off-the-shelf algorithms *obliviously* to the structure induced by the MDPRM. However, this would lead to a large regret in view of the large state-space of the cross-product MDP.

We formalize regret minimization in average-reward MDPRMs, and establish a first regret lower bound for MDPRMs. We present a model-based algorithm that efficiently exploits the structure in MDPRMs, and analyze its regret non-asymptotically. Like the lower bound, our bound is independent of Q, the number of RM states. Further, it improves over regret bound of the existing baselines (e.g., UCRL2 [2] applied to the cross-product MDP) by up to a factor of $Q^{3/2}$. Finally, we report numerical experiments that demonstrate the superiority of the proposed algorithm over existing baselines in practice.

References

[1] Toro Icarte, R., Klassen, T., Valenzano, R., and McIlraith, S., *Using reward machines for high-level task specification and decomposition in reinforcement learning*. In Proc: International Conference on Machine Learning, 2018, pp. 2107–2116.

[2] Jaksch, T., Ortner, R., and Auer, P., *Near-optimal regret bounds for reinforcement learning*, Journal of Machine Learning Research, 11:1563–1600, 2010.

Validation of point process predictions with proper scoring rules

Thordis L. Thorarinsdottir[1], Claudio Heinrich-Mertsching[2], Peter Guttorp[2] and Max Schneider[3]

[1] *University of Oslo, Oslo, Norway, thordist@math.uio.no*
[2] *Norwegian Computing Center, Oslo, Norway, claudio@nr.no, peterg@nr.no*
[3] *United States Geological Survey, Mountain View, California, United States, mschneider@usgs.gov*

In prediction settings, model validation methods are needed to rank competing models according to their predictive performance. Such rankings are typically obtained by proper scoring rules. A challenge for applying known scoring rules to point process predictions is that mathematical properties, such as densities or moment measures, are intractable for many point process models. We introduce a class of proper scoring rules for evaluating point process predictions based on summary statistics. These scoring rules rely on Monte-Carlo approximations of expectations and can therefore easily be evaluated for any point process model that can be simulated. The scoring rules allow for evaluating the calibration of a model to specific aspects of a point process, such as its spatial distribution or tendency towards clustering.

Covariance Estimation Techniques for large-scale Ensemble Kalman Filtering, with Applications to Paleoclimate Reconstruction

Cédric Travelletti[1], Lorenz Hilfiker[1], David Ginsbourger[1], Alberto Terranova[1], Veronika Valler[2], Jörg Franke[2], and Stefan Brönnimann[2]

[1] *Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland, cedric.travelletti@unibe.ch*

[2] *Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland*

Ensemble Kalman filtering (EnKF) is a data assimilation technique that is successful at handling large state vectors, but suffers from drawbacks when the background covariance matrix has to be regularized. Such regularization is paramount in applications where only few ensemble members are available, but owing to the size of the matrices involved, only basic regularization techniques can be applied; moreover, in this context, regularization can introduce data-ordering dependencies in the assimilation. In this talk, we present a distributed computing framework for handling large covariance matrices in EnKFs and show how it allows for more sophisticated regularization procedures to be brought to bear in a large-scale setting and how it alleviates ordering dependencies by assimilating the data in a non-sequential fashion. We demonstrate our techniques on a paleoclimate reconstruction problem [1], comparing them to the current state of the art.

References

[1] Valler et al., Impact of different estimations of the background-error covariance matrix on climate reconstructions based on data assimilation, *Climate of the Past*, 15(4), (2019) 1427-1441.

Geometric Random Graphs

Tatyana Turova

Lund University, Sweden, tatyana@maths.lth.se

We consider random graphs on the set of vertices placed on the discrete d-dimensional torus. The edges between pairs of vertices are independent, and their probabilities depend on the distance between the vertices, more precisely, they decay polynomially. Hence, the probabilities of connections are naturally scaled with the total number of vertices via distance. The graphs under consideration have finite degrees. It was proved in [1] that despite similarities with the classical random graphs this class of geometric random graphs exhibit also phase transitions with respect to the number of small subgraphs, triangles in particular. We investigate the effect of such phase transitions on the properties of bootstrap percolation. It is shown that the process of the bootstrap percolation is determined not only by the amount of initial activation but also by the location of the active initial states.

References

[1] G. M. Napolitano and T. Turova Geometric Random Graphs vs Inhomogeneous Random Graphs, Markov Processes Relat. Fields 25, 615–637, 2019.

Boundary-preserving schemes for a stochastic Allen-Cahn equation

Johan Ulander[1]

[1] *Chalmers University of Technology and University of Gothenburg, Sweden, johanul@chalmers.se*

We propose boundary-preserving Lie–Trotter–Lamperti splitting schemes with finite difference discretisation in space for approximating solutions to the stochastic Allen-Cahn equation with multiplicative diffusion term of the form (1-u)*(1+u). For such drift and diffusion term, the solution remains in the interval $[-1, 1]$ almost surely. This property is shared by the numerical approximations. Preliminary numerical experiments indicate a mean-square convergence of order 1 in time for a fixed spatial discretisation under a CFL condition.

Edgeworth-type expansion of the density of the classifier when growth curves are classified via likelihood

Emelyne Umunoza Gasana[1], Dietrich von Rosen[2] and Martin Singull[1]

[1] *Linköping University, Linköping, emelyne.umunoza.gasana@liu.se, martin.singull@liu.se*
[2] *Swedish University of Agricultural Sciences, Uppsala, dietrich.von.rosen@slu.se*

When classifying repeated measurements using the Growth Curve model, also known as bilinear regression model, it can happen that the observations to classify might not belong to any of the two predetermined populations. [1] derived a two-step classification rule taking into account this perspective. Probabilities of misclassification of the two-step likelihood-based discriminant rule are established for the classification of growth curves where the distribution for the classifier is approximated using an Edgeworth-type expansion.

References

[1] von Rosen, D. and Singull, M., Classification of repeated measurements using growth curves, *Linköping University Electronic Press*, 2022.

Bayesian partial least squares regression for prediction from spectral data

Szymon Urbas[1], Donagh Berry[2] and Isobel Claire Gormley[1]

[1] *University College Dublin, Ireland, szymon.urbas@ucd.ie, claire.gormley@ucd.ie*
[2] *Teagasc, Fermoy, Co. Cork, Ireland, dberry@moorepark.teagasc.ie*

Partial least squares (PLS) [1] regression is a popular method for making predictions in the field of chemometrics. The method is based on maximising the covariance between the predictor and response vectors through a latent-space decomposition. However, the method does not arise from the statistical inference of any particular probabilistic model. Thus it cannot account for parameter uncertainty, and it does not easily lend itself to modifications. Furthermore, the choice of the dimension, Q, of the latent space can be sensitive to the training dataset and the choice process can be computationally expensive. We introduce a Bayesian latent-variable model which emulates the desirable properties of PLS. To eschew the need to choose the latent dimension, we employ a multiplicative gamma process prior [2] which enforces shrinkage on the latent variables. The proposed Bayesian partial least squares regression framework is very flexible and allows for modifications; for example, it allows for different sparsity structures in the regression. We show how the point predictions produced by the proposed approach are at least as accurate as the industry-standard PLS methods. In addition to accurate point predictions, the statistical model provides prediction intervals with the correct coverage.

References

[1] Wold, H. (1973). Nonlinear iterative partial least squares (NIPALS) modelling: some current developments. In *Multivariate analysis–III* (pp. 383-407). Academic Press.

[2] Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306.

Large deviations of extremal eigenvalues of $\beta-$Laguerre ensembles

Denise Uwamariya[1] and Xiangfeng Yang[1]

[1] *Linköping University, Linköping, denise.uwamariya@liu.se, xiangfeng.yang@liu.se*

For any $\beta > 0$, in the paper we consider general β-Laguerre ensembles which extend the classical (real with $\beta = 1$, complex with $\beta = 2$ and quaternion with $\beta = 4$) Wishart matrices. More specifically, large deviations for the extremal eigenvalues of general β-Laguerre ensembles are established under the assumption that the dimension p and the sample size n are comparable $p/n \to \kappa \in (0, 1)$. The study generalizes several results in the literature.

Assumption-free SDE inference using Universal Differential Equations

Marc Vaisband[1,2], Valentin von Bornhaupt[1], Izdar Abulizi[1], Maren Philipps[1], and Jan Hasenauer[1]

[1] *Life & Medical Sciences Institute, University of Bonn, vaisband@uni-bonn.de, bornhaupt@uni-bonn.de, s6izabul@uni-bonn.de, maren.philipps@uni-bonn.de, jan.hasenauer@uni-bonn.de*

[2] *Department of Internal Medicine III with Haematology, Medical Oncology, Haemostaseology, Infectiology and Rheumatology, Oncologic Center, Salzburg Cancer Research Institute - Laboratory for Immunological and Molecular Cancer Research (SCRI-LIMCR); Paracelsus Medical University, Cancer Cluster Salzburg, Salzburg, Austria*

Inferring the dynamics of stochastic differential equations (SDEs) from observed sample paths is an extremely difficult task, owing to their innate randomness and challenging mathematical properties. There exists a number of methods for parametric inference when the analytical structure of the true drift and diffusion coefficients is known; however the assumption-free case is significantly more problems. Nonetheless, it is highly relevant from an application perspective, as SDEs can represent intra-population variability better than ordinary differential equations, while remaining computationally more tractable than microscopic jump processes. We present an inference approach inspired by the Universal Differential Equation framework proposed by Rackauckas et al. [1], parametrising the coefficient functions of an SDE by a neural network.

In similar setups, optimisation is typically done by back-propagating through a differentiable SDE solver and comparing summary statistics of observed and generated trajectories. We propose an alternative method that utilises pseudo-likelihoods, a popular approximation in parametric SDE inference, based on the transition densities of the simple Euler-Maryama scheme. Our results demonstrate that the two can be hybridised for a greatly increased capacity to learn unknown model dynamics, both in synthetic problems and with real-life experimental data.

References

[1] Rackauckas, C, et al., Universal differential equations for scientific machine learning., *arXiv preprint* arXiv:2001.04385 (2020).

Inference for structural relationship models

Janis Valeinis[1], **Leonora Pahirko**[1] **and Jelizaveta Sokolovska**[2]

[1] *University of Latvia, Faculty of Physics, Mathematics and Optometry, Riga, janis.valeinis@lu.lv, leonora.pahirko@lu.lv*

[2] *University of Latvia, Faculty of Medicine, Riga, jelizaveta.sokolovska@lu.lv*

Structural relationship models were introduced by Freitag and Munk [1] to describe a semi-parametric relationship between two distribution functions. It contains the classical two-sample location-scale model and the Lehmann alternatives model. Our goal is to use empirical likelihood-based methods to make inference for structural relationship models. For this purpose, we propose and compare two novel approaches: 1) simultaneous inference for the structural relationship parameter using several empirical likelihood-based test statistics; 2) graphical empirical likelihood-based goodness-of-fit testing of the structural relationship model by plug-in estimation of the unknown structural parameter.

We validate our methods by some simulation study and demonstrate the application on Latvian diabetes patient data. The location-scale transformation was used to reduce the variability in observations measured on several plates. Moreover, we tested the location-scale model between control and diabetes patients comparing such biomarkers as high sensitivity C-reactive protein, lipopolysaccharide and endogenous anti-endotoxin core antibodies.

References

[1] Freitag, G. and Munk, A. On Hadamard differentiability in k-sample semiparametric models—with applications to the assessment of structural relationships, *Journal of multivariate analysis*, 94(1), 123-158.

Inference for partially observed stochastic processes by message passing

Frank van der Meulen[1] and Moritz Schauer[2]

[1] *Vrije Universiteit Amsterdam, The Netherlands, f.h.van.der.meulen@vu.nl*
[2] *Chalmers University of Technology — University of Gothenburg, Sweden, smoritz@chalmers.se*

We consider the problem of parameter inference for a continuous-time stochastic processes that is only observed discretely in time. A partially observed diffusion process is a key example. In this setting, the likelihood is typically intractable. From a computational perspective, likelihood based inference is usually based on a data-augmentation approach, where the latent paths are imputed. I will discuss a structured way for this imputation which leads to closed form expressions of transition densities. The expressions derived contain a term of the form $\mathbb{E}\Psi(X^\circ)$, where X° is a tractable process. Such results can be used subsequently in MCMC, MCEM or variation inference for example.

From a computational point of view, the results give rise to a bidirectional data flow, where information from a backward filtering step is passed as a message to a forward sampling step.

References

[1] Mider, Schauer & Van der Meulen, *Continuous-discrete smoothing of diffusions*, Electronic Journal of Statistics 15, (2021) 4295–4342.

[2] F.H. van der Meulen and M. Schauer, *Automatic Backward Filtering Forward Guiding for Markov processes and graphical models*, arXiv:2010.03509 (preprint), 2022.

[3] C. Corstanje, F.H. van der Meulen and M. Schauer (2022), *Conditioning continuous-time Markov processes by guiding*, Stochastics (2022).

Finite-sample exact prediction bands for functional data

Jacopo Diquigiovanni[1], Matteo Fontana[2] and **Simone Vantini**[3]

[1] *Department of Statistical Sciences, University of Padova, Italy, jacopo.diquigiovanni@phd.unipd.it*
[2] *MOX - Dept of Mathematics, Politecnico di Milano, Italy, now at Joint Research Centre, European Commission, Ispra (VA), Italy, matteo.fontana@ec.europa.eu*
[3] *MOX - Dept of Mathematics, Politecnico di Milano, Italy, simone.vantini@polimi.it*

The talk will deal with the key challenge of creating prediction bands for a new observation in the functional data framework given a training set of observed functional data and possibly in presence of covariates, either scalar, categorical, or functional. Starting from the investigation of the literature concerning this topic, we propose an innovative approach building on top of Conformal Prediction and Functional Data Analysis able to overcome the main drawbacks associated to the existing approaches. Under minimal distributional assumptions (i.e., exchangeability of the random functions), we will show how the new proposed nonparametric method (i) is able to provide prediction regions which could visualized in the form of bands, (ii) is guaranteed with exact coverage probability also for finite sample sizes, and finally (iii) is computational efficient. Different specifications of the method will be compared in terms of efficiency in some simulated and real case scenarios also in the case of multi-dimensional domain and/or codomain.

References

[1] Diquigiovanni, J., Fontana, M., Vantini, S. (2021): The importance of being a band: Finite-sample exact distribution-free prediction sets for functional data. arXiv:2102.06746

[2] Diquigiovanni, J., Fontana, M., Vantini, S., (2022): Conformal Prediction Bands for Multivariate Functional Data, *Journal of Multivariate Analysis*, 189, 104879.

Synesthetic experience in museums: a statistical analysis based on multi-point semantic differential scales

Matteo Ventura[1], Ambra Macis[1], Marica Manisera [1] and Paola Zuccolotto [1]

[1] *University of Brescia, Italy, matteo.ventura@unibs.it, ambra.macis@unibs.it, marica.manisera@unibs.it, paola.zuccolotto@unibs.it*

Using statistical methods for mapping visitors' behaviours, experiences and perceptions in cultural places is essential to increase knowledge of organizations, decision makers and citizens. In the project *Data Science for Brescia - Arts and Cultural Places* (https://bodai.unibs.it/ds4bs/), a survey investigated the sensory sensations and emotions experienced by visitors at the Tosio-Martinengo Gallery in Brescia, Italy.
To explore visitors' sensory experiences, we used the concepts of synesthesia (stimulation of one sense leads to involuntary experiences in other senses) and ideasthesia (activation of general concepts evokes perception-like experiences). By using a questionnaire with responses given on a multi-point semantic differential scale, visitors were asked to rate their experiences by positioning themselves on a rating between two bipolar adjectives (rough/soft; spicy/fruity, ...).
We analyzed the resulting rating data with the CUM model (Combination of discrete Uniform and a - linearly transformed - Multinomial random variables) [2], belonging to the CUB (Combination of discrete Uniform and shifted Binomial) class [1].
Results show interesting insights about the visitor multisensory experience that can be combined with results about tangible aspects of the museum experience, such as expectations and satisfaction, and then exploited in order to improve the quality of citizens' and tourists' experiences.

References

[1] D'Elia, A., and Piccolo, D. (2005). A mixture model for preferences data analysis. *Computational Statistics and Data Analysis*, 49(3), 917-934.

[2] Manisera, M., and Zuccolotto, P. (2022). A mixture model for ordinal variables measured on semantic differential scales. *Econometrics and Statistics*, 22, 98-123.

Conditional particle filters with bridge backward sampling

Matti Vihola[1], Santeri Karppinen[1], Sumeetpal S. Singh[2]

[1] *University of Jyväskylä, Finland,* {*matti.s.vihola,santeri.j.karppinen*} *@ jyu.fi*
[2] *University of Wollongong, Australia, sumeetpals @ uow.edu.au*

The conditional particle filter (CPF) [1] and its backward/ancestor sampling versions [5, 4] are Markov transitions targeting the smoothing distribution of a hidden Markov model. The performance of the backward/ancestor sampling CPF is often impressive even with long data records. Two known exceptions are when the observations are weakly informative and when the dynamic model is slowly mixing. These are both present when sampling finely time-discretised continuous-time path integral models, but can occur with hidden Markov models too. Multinomial resampling, which is commonly employed in the (backward sampling) CPF, resamples excessively for weakly informative observations and thereby introduces extra variance. A slowly mixing dynamic model renders the backward sampling step ineffective. We resolve [3] the former issue by replacing multinomial resampling by a conditional version of recently suggested variant of systematic resampling [2]. To avoid the degeneracy issue of backward sampling, we introduce a generalisation that involves backward sampling with an auxiliary 'bridging' CPF step, which is parameterised by a blocking sequence. We present practical tuning strategies for choosing an appropriate blocking. Our experiments demonstrate that the CPF with a suitable resampling and the developed 'bridge backward sampling' can lead to substantial efficiency gains in the weakly informative regime.

References

[1] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(3):269–342, 2010.

[2] N. Chopin, S. S. Singh, T. Soto, and M. Vihola. On resampling schemes for particle filters with weakly informative observations. *Ann. Statist.*, 50(6):3197–3222, 2022.

[3] S. Karppinen, S. S. Singh, and M. Vihola. Conditional particle filters with bridge backward sampling. *Preprint* arXiv:2205.13898, 2022.

[4] F. Lindsten, M. I. Jordan, and T. B. Schön. Particle Gibbs with ancestor sampling. *J. Mach. Learn. Res.*, 15(1):2145–2184, 2014.

[5] N. Whiteley. Discussion on "Particle Markov chain Monte Carlo methods". *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(3):306–307, 2010.

Joint Sparse Clustering and Alignment of Functional Data

Valeria Vitelli[1]

[1] *University of Oslo, Norway; valeria.vitelli@medisin.uio.no*

Finding sparse solutions to clustering problems has emerged as a hot topic in statistics in recent years, and sparse clustering approaches have been proposed also for the case of functional data, when it is often of interest to select the portion of the curves' domain mostly exhibiting a clustering structure. Functional sparse clustering can be analytically defined as a well-posed variational problem, with a hard-thresholding constraint ensuring the sparsity of the optimal solution, which is also proved to be unique.

When dealing with curve clustering we cannot forget the presence of misalingment: this additional source of variation along the abscissa of the functions (named phase variability) can heavily affect the sparse clustering results, potentially leading to meaningless conclusions. Many methods to jointly cluster and align curves, which efficiently decouple amplitude and phase variability, have already been proposed in the literature on functional data. I propose a possible approach to deal with sparse functional clustering while also aligning the curves, therefore jointly performing all these tasks: functional clustering, curves' alignment, and domain selection [1, 2]. The method is studied in its well-posedness, and its performance and parameter tuning are explored in a variety of simulated scenarios and real case studies.

References

[1] Vitelli, V., A novel framework for joint sparse clustering and alignment of functional data, *Journal of Nonparametric Statistics*, 2023, 1–30.

[2] Vitelli, V., SparseFunClust: Sparse Functional Clustering. *R package version 1.0.0*, `https://cran.r-project.org/web/packages/SparseFunClust/`.

Second-order semi-parametric inference for multivariate point processes

Rasmus Waagepetersen[1]

[1] *Department of Mathematical Sciences, Aalborg University, Denmark, rw@math.aau.dk*

We consider a new approach to inferring the second-order properties of a multivariate log Gaussian Cox process (LGCP) with a complex intensity function. We assume a semi-parametric model for the multivariate intensity function containing an unspecified complex factor common to all types of points. Given this model, we construct a second-order conditional composite likelihood to infer the pair correlation and cross pair correlation functions of the LGCP. Crucially this likelihood does not depend on the unspecified part of the intensity function. We also introduce a cross-validation method for model selection and an algorithm for regularized inference that can be used to obtain sparse models for cross pair correlation functions. The methodology is applied to simulated data as well as data examples from microscopy and criminology. This shows how the new approach outperforms existing alternatives where the intensity functions are estimated non-parametrically.

References

[1] Hessellund, K. B., Xu, G., Guan, Y. and Waagepetersen, R., Second-order semi-parametric inference for multivariate log Gaussian Cox processes, *Journal of the Royal Statistical Society Series C*, 71 (2022), 244-268.

Muller's ratchet with tournament selection: click rate and type frequency profile

Anton Wakolbinger[1]

[1] *Goethe-University Frankfurt am Main, wakolbinger@math.uni-frankfurt.de*

Muller's ratchet is a prototype model in mathematical population genetics. In an asexual population of constant size N, individual lineages are assumed to slowly acquire slightly deleterious mutations over the generations. Due to randomness, every once and a while the individuals with the currently smallest number of mutations disappear from the population; this is a click of the ratchet. The classical variant of the model, which assumes so-called proportional selection, so far has resisted against a fully rigorous asymptotic analysis of the clicking rate. We overcome this hurdle by considerng tournament (instead of proportional) selection, where selective competition within pairs is won by the fitter individual. By means of a graphical representation we obtain a dual process which for a wide range of selection and mutation parameters allows to analyse (as $N \to \infty$) the click rates and the quasi-stationary type frequency profile. For the latter, an interesting connection with first passage percolation in Poisson decorated Yule trees turns out to be relevant.
The talk is based on joint work with Adrián González-Casanova, Charline Smadi, and Jan Lukas Igelbrink.

Rotation to Sparse Loadings using L^p Losses and Related Inference Problems

Gabriel Wallin[1,2,a], Xinyi Liu[1], Yunxiao Chen[1] and Irini Moustaki[1]

[1] *London School of Economics and Political Science*
[2] *Umeå University*
[a] *G.A.Wallin@lse.ac.uk*

Researchers have widely used exploratory factor analysis (EFA) to learn the latent structure underlying multivariate data. Rotation and regularised estimation are two classes of methods in EFA that are often used to find interpretable loading matrices. In this presentation, a new family of oblique rotations based on component-wise L^p loss functions ($0 < p \leq 1$) that is closely related to an L^p regularised estimator is presented. We develop model selection and post-selection inference procedures based on the proposed rotation. When the true loading matrix is sparse, the proposed method tends to outperform traditional rotation and regularised estimation methods in terms of statistical accuracy and computational cost. Since the proposed loss functions are nonsmooth, we develop an iteratively reweighted gradient projection algorithm for solving the optimisation problem. We also develop theoretical results that establish the statistical consistency of the estimation, model selection, and post-selection inference. We evaluate the proposed method and compare it with regularised estimation and traditional rotation methods via simulation studies. We further illustrate it using an application to the Big Five personality assessment.

Different distance functions for clustering of a matrix containing failure modes and diagnostic methods

Per Westerlund[1]

[1] *Luleå University of Technology, Skellefteå, per.westerlund@ltu.se*

Planning of maintenance should take into account the condition of the different pieces of equipment instead of having a prescribed interval of certain number of years.

The working group A3.43 of CIGRE, an organisation for high-voltage electrical technology, studies maintenance strategies for circuit breakers, a key component in electric grids. It has compiled a 46x39 matrix, which maps the connection between failure modes and diagnostic methods as estimated by experienced engineers. Each element is a number from 0 to 3, where a higher number indicates a clearer connection between them.

In order to understand better how the failure modes and the diagnostic methods are related to each other, a clustering method is applied to the matrix. The chosen method is a hierarchical clustering, where both rows and columns are reordered using the Euclidean and the Manhattan distance function.[1, p 301] The result is presented as a heatmap with dendrograms showing the possible clusters in the columns and the rows, respectively.

The results are graphs that give a comprehensive picture of how the failure modes and the diagnostic methods are related. This is valuable for newcomers in the field. The Manhattan distance function gives less value to elements in the matrix.

References

[1] Cha, S-H, Comprehensive Survey on Distance/Similary Measures between Probability Functions, *International Journal och Mathematical Models and Methods in Applied Sciences*, 4, (2007), 300-307.

Temporal Event Boosting: gradient boosting for marked temporal point processes

Fredrik Lundvall Wollbraaten[1] and Geir Storvik[2]

[1] *University of Oslo, fredrlwo@math.uio.no*
[2] *University of Oslo, geirs@math.uio.no*

Temporal point process data arise in many real-world settings, e.g system log messages, neural spiking activity, online customer behavior data or events in football matches. We consider Marked Temporal Point Processes (MTPP), where each arrival time t_i has a discrete mark $m_i \in \{1, ..., M\}$. Considering the MTPP as a multivariate point process, we propose Temporal Event Boosting (TEB) for estimating mark-specific conditional intensities depending on the history. Despite the success of gradient boosting, its extension to MTPP has not been considered. TEB is a gradient boosting approach for MTPP based on discretizing time, encoding the history of the process using interevent time lags, whereafter gradient boosting is applied to the count observations which have a Poisson likelihood. Using simulated multivariate Hawkes Process data, we show that TEB performs as well or better than a recurrent neural network-based competitor [1], the main machine learning method in the MTPP literature, in terms of test-set log-likelihood and recovering the true excitation functions. Notably, TEB proves to be particularly advantageous in scenarios with sparse infectivity matrices or cross-excitation functions exhibiting distinct parametric shapes.

References

[1] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song, *Recurrent marked temporal point processes: Embedding event history to vector*, In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1555–1564, 2016.

Word Embedding Uncertainty Estimation

Väinö Yrjänäinen[1], Isac Boström[2], Johan Jonasson[2] and Måns Magnusson[1]

[1] *Uppsala University, Sweden, vaino.yrjanainen@statistics.uu.se*
[2] *Chalmers University of Technology, Sweden*

Word embeddings are an increasingly popular method for inference from large textual data. For reliable inference, estimating the uncertainty of the embeddings is integral. So far only bootstrap and approximate Bayes methods have proven to be practically feasible for estimating the uncertainty of word embeddings. As alternatives for these, we investigate (1) Laplace approximation around the posterior mode and (2) Hamiltonian Monte Carlo sampling of the posterior. For both methods, we present ways to combat their prohibitive computational cost on large datasets, which are commonly used with word embedding methods. We go on to compare the methods with each other as well as with variational inference and bootstrap on both static embeddings and dynamic word embeddings. This is done with both simulated and real data.

Contextualized Word Embeddings with Laplacian Graph Priors

Väinö Yrjänäinen[1], Andreas Östling[1], Jonas Wallin[2] and Måns Magnusson[1]

[1] *Uppsala University, Sweden, vaino.yrjanainen@statistics.uu.se*
[2] *Lund University, Sweden*

Recently, BERT has emerged as the default language model for contextualized word embeddings. However, its standard version does not take into account additional available information about time, authors and semantic relations between the words in the text. We introduce Contextualized Word Embeddings with Laplacian Graph Priors (CWELP) as a way to incorporate this information, which combines priors with the likelihood of the BERT and RoBERTa models. We empirically study if we can improve masked language model performance by including dictionary information as well as temporal and group metadata about the text. We evaluate the performance on multiple standard datasets, including the US Congress and the Arxiv corpora. We furthermore show how some recently proposed contextual embeddings can be formulated using a graph Laplacian prior and that previously proposed models can be seen as a special case.

Optimizing Hydraulic Pressure in a Wave Power Plant

Shokoufa Zeinali[1], Johan Lindström[2] , Magnus Wiktorsson[3] ,Georg Lindgren[4] and Jan Forsberg[5]

[1] *Lund University, Lund, shokoufa.zeinali@matstat.lu.se*
[2] *Lund University, Lund, johan.lindstrom@matstat.lu.se*
[3] *Lund University, Lund, magnus.wiktorsson@matstat.lu.se*
[4] *Lund University, Lund, Georglindgren@telia.com*
[5] *Karlstad University, Karlstad , jan.forsberg@fluiddynamik.se*

This study aims to determine the optimal pressure for the accumulator tank in a wave energy converter (WEC) to maximize electricity generation. A simulation-based approach was employed to analyze the buoy's performance with wave data generated from the Wafo toolbox using the generalized Pierson-Moskowitwitz spectrum. These data were then fed into a Simulink model of the WEC system. After filtering the data, concave regression was applied to model the relationship between tank pressure and power output. Subsequently, a regression analysis was performed to model the optimal pressure as a function of wave steepness. The findings indicate a strong dependency of optimal pressure on wave steepness, suggesting that adjusting the accumulator tank pressure according to wave conditions can significantly enhance the WEC's energy output. This research contributes to the ongoing efforts to improve the efficiency of wave energy converters and provides valuable insights for developing adaptive control strategies in WEC systems.

Noise Sensitivity and Stability of Deep Neural Networks for Binary Classification

Olof Zetterqvist[1], Johan Jonasson[1] and Jeffrey E. Steif[1]

[1] *Chalmers University of Technology and University of Gothenburg, Gothenburg, olofze@chalmers.se, jonasson@chalmers.se, steif@chalmers.se*

A first step is taken towards understanding often observed non-robustness phenomena of deep neural net (DNN) classifiers. This is done from the perspective of Boolean functions by asking if certain sequences of Boolean functions represented by common DNN models are noise sensitive or noise stable, concepts defined in the Boolean function literature. Due to the natural randomness in DNN models, these concepts are extended to annealed and quenched versions. Here we sort out the relation between these definitions and investigate the properties of two standard DNN architectures, the fully connected and convolutional models, when initiated with Gaussian weights.

Graphical LASSO for extremes

Chen Zhou[1] and Phyllis Wan[2]

[1] *Erasmus University Rotterdam, zhou@ese.eur.nl*
[2] *Erasmus University Rotterdam, wan@ese.eur.nl*

Gaussian graphical LASSO is a statistical tool for detecting sparse dependence structure in data following a sparse multivariate Gaussian distribution. To model multivariate extremes with a spare dependence structure, the Hüsler-Reiss graphical model has been recently proposed as an alternative. However, the adaptation of graphical LASSO to the estimation of the sparse Hüsler-Reiss graphical model is not straightforward. In this paper, we propose a graphical LASSO method for extremes through re-parametrizating the Hüssler-Reiss graphcial model. The estimator enjoys theoretical properties similar to the Gaussian graphical LASSO: concentration equalities, fast computation and the ability to scale up to large dimensions.