2023 IMS International Conference on Statistics and Data Science (ICSDS)

December 18-21, 2023
Lisbon, Portugal

Program Book

SPE
Sociedade Portuguesa de Estatística

Centro de Estatística e Aplicações
Universidade de Lisboa

Fundação
para a Ciência
e a Tecnologia

ISBN: 978-0-940600-86-7
December, 2023

Editors: Ivette Gomes, Teresa Oliveira, Amilcar Oliveira, Pedro Pestana, Min Xu

Preface

In response to the call from the 2021 IMS (Institute of Mathematical Statistics) Survey report to expand membership from emerging areas of data science, underrepresented groups, and from regions outside of North America, the IMS has launched the annual IMS *International Conference on Statistics and Data Science* (ICSDS). Following the success of 2022 ICSDS in Florence Italy, December 13-16. 2022, the second and this 2023 ICSDS is held on December 18-21, in Lisbon, Portugal. In addition to plenary sessions, invited, contributed and poster sessions, ICSDS offers a student paper competition for 12 Student Travel Awards. There are also Junior Researcher Support Funds for travel support for junior researchers. We gratefully acknowledge here the generous support for both awards from the funds of *Industry Friends of* IMS (IFoIMS). Students and young researchers are strongly encouraged to participate and utilize this support.

The ICSDS conference has been thoughtfully organized to provide platforms to facilitate discovery dissemination and foster collaborations among researchers from a wide range of research and practice areas in statistics and data science, and from academia, industry and government. We are gratified to be able to welcome to the ICSDS more than 550 participants coming from more than 40 countries. The goal of the ICSDS is to afford such a diverse crowd a stimulating setting for exchanging ideas on the developments of modern statistics and data science, broadly defined, in all aspects of theory, methods and applications.

The Local Organizing Committee in Lisbon, the beautiful capital city of Portugal, has worked hard to find the modern *Cultural Center of Belém* (CCB) in Belém area for us to host the ICSDS this year. CCB is located near where Tagus River meets the Atlantic Ocean, and it is surrounded by many magnificent landmarks, including Tower of

Belém and Mosteiro dos Jerónimos. This location conveniently affords the conference participants rich social programs, including conference tours to Mosteiro dos Jerónimos and Tower of Belém, both are UNESCO World Heritage Sites, and are historical and architectural treasures for the world. We are also able to organize a conference banquet in the charming Casa do Alentejo to enjoy the local cuisine and the renowned Portuguese traditional Fado music performance.

Participants are also highly encouraged to take advantage of the close proximity to "Pastéis de Belém" from CCB to enjoy the world famous Portuguese tarts. Many other wonderful things to do, eat and see in Lisbon can also be found in `https://www.timeout.com/lisbon`.

We wish that the 2023 ICSDS proves to be another productive conference that will successfully cultivate more fruitful exchanges and collaborations for years to come.

Acknowledgement: Needless to say, an international conference of this scale, with its coverage of wide-ranging subjects and size of broad participants from various disciplines across the world, would not have been possible without the collective efforts of many.

We would like to thank the program committee, of 50 members from 29 countries, for helping establish the rich program. The local Organizing Committee: Eunice Carrasquinha (Co-chair, CEAUL-FCUL), Ivette Gomes (Co-chair, CEAUL-FCUL), Tiago Marques (University of St Andrews, UK), Teresa A. Oliveira (Co-chair, CEAUL and Universidade Aberta), Soraia Pereira (CEAUL, Universidade de Lisboa), Giovani Silva (IST, CEAUL, Universidade de Lisboa), Lisete Sousa (Universidade de Lisboa), have contributed their tremendous effort and time in every aspect of this endeavor, from helping organize sessions, finding conference venue to all the logistics up to the very last minute, including banquet and tours. In particular, Ivette and Teresa brilliantly turned the messy and long conference program into a beautifully organized program book.

Our heartfelt thanks also go to Elyse Gustafson (IMS Executive Director) for her working overtime on the financial issues and other related formalities on behalf of the IMS, and to Arlene Gray (Administrator, ICSDS) for her patiently taking care of the nonstop inquiries and tracking numerous responses and requests from the participants and the conference organizing team.

Finally, we would be remiss not to acknowledge the invaluable con-

tributions behind the scenes from Min Xu (IMS, Rutgers University), from setting up and managing the conference website, negotiating IT support with the conference venue, to setting up the program and readying all the slides. He bravely and efficiently met head on all kinds of unexpected challenges, technical as well as personal. It suffices to say that Min did all the heavy-lifting to make the conference program a reality for us all to enjoy.

2023 ICSDS Program Co-chairs:
Regina Liu (IMS Past-President) and Annie Qu (IMS Program Secretary)

Program Co-Chairs

- Regina Liu — *USA, rliu@stat.rutgers.edu*
- Annie Qu — *USA, aqu2@uci.edu*

Local Arrangement Committee

- Ivete Gomes — *Co-chair, Portugal*
- Eunice Carrasquinha — *Co-chair, Portugal*
- Teresa Oliveira — *Co-chair, Portugal*
- Soraia Pereira — *Portugal*
- Giovani Silva — *Portugal*
- Lisete Sousa — *Portugal*
- Tiago Marques — *Portugal*

Scientific Committee

- Genevera Allen — *USA*
- Serena Arima — *Italy*
- Arne Bathke — *Austria*
- Howard Bondell — *Australia*
- Eunice Carrasquinha — *Portugal*
- Jinyuan Chang – *China*
- Probal Chaudhuri – *India*
- Radu Craiu – *Canada*
- Sally Cripps – *Australia*
- Juan Cuesta Albertos – *Spain*
- Elena Di Bernardino – *France*
- Susana Eyheramendy – *Chile*

- Yingying Fan – *USA*
- Arnoldo Frigessi – *Norway*
- Ricardo Friman – *Uruguay*
- Haoda Fu – *USA*
- Irène Gijbels – *Belgium*
- M. Ivette Gomes – *Portugal*
- Pauliina Ilmonen – *Finland*
- Rebecka Jrnsten – *Sweden*
- Julie Josse – *France*
- Roger Koenker – *UK*
- Eric Laber – *USA*
- Jialiang Li – *Singapore*
- Chae Young Lim – *S. Korea*
- Ivy Liu – *New Zealand*
- Yan Liu – *Japan*
- Pamela Llop – *Argentina*
- Po-Ling Loh – *UK*
- Tiago Marques – *UK*
- Fabrizia Mealli *Italy*
- Axel Munk – *Germany*
- Sofia Olhede – *Switzerland*
- Teresa Oliveira – *Portugal*
- Hernando Ombao – *Saudi Arabia*
- Phillip Otto – *Germany*

- Nicole Pashley – *USA*
- Byeong Park – *S. Korea*
- Daniel PE – *Spain*
- Soraia Pereira – *Portugal*
- Saharon Rosset – *Israel*
- Dominik Rothenhler – *USA*
- Richard Samworth – *UK*
- Bodhisattva Sen – *USA*
- Giovani Silva – *Portugal*
- Helle Srensen – *Denmark*
- Pragya Sur – *USA*
- Sheng-Tsaing Tseng – *Taiwan*
- Junhui Wang – *Hong Kong*
- Yin Xia – *China*
- Min Xu – *USA*

Sponsors

- *Fundação para a Ciência e a Tecnologia* (FCT)
- *Centro de Estatística e Aplicações* (CEAUL)
- *Dept. de Estatística e Investigação Operacional* (DEIO)
- *Faculdade de Ciências de Lisboa*

Contents

Chapter 1

Plenary Talks

Causality meets Representation Learning

Caroline Uhler

Massachusetts Institute of Technology, USA

Abstract. Massive data collection holds the promise of a better understanding of complex phenomena and ultimately, of better decisions. Representation learning has become a key driver of deep learning applications, since it allows learning latent spaces that capture important properties of the data without requiring any supervised annotations. While representation learning has been hugely successful in predictive tasks, it can fail miserably in causal tasks including predicting the effect of an intervention. This calls for a marriage between representation learning and causal inference. An exciting opportunity in this regard stems from the growing availability of interventional data (in medicine, advertisement, education, etc.). However, these datasets are still miniscule compared to the action spaces of interest in these applications (e.g. interventions can take on continuous values like the dose of a drug or can be combinatorial as in combinatorial drug therapies). In this talk, we will present initial ideas towards building a statistical and computational framework for causal representation learning and its application towards optimal intervention design in the context of the biomedical sciences.

Data Science at the Singularity

David Donoho

Stanford University, USA

Abstract. A purported "AI Singularity" has been much in the public eye recently, especially since the release of ChatGPT last November, spawning social media "AI Breakthrough" threads promoting Large Language Model (LLM) achievements. Alongside this, mass media and national political attention focused on "AI Doom" hawked by social media influencers, with twitter personalities invited to tell congresspersons about the coming "End Times". In my opinion, "AI Singularity" is the wrong narrative; it drains time and energy with pointless speculation. We do not yet have general intelligence, we have not yet crossed the AI singularity, and the remarkable public reactions signal something else entirely. Something fundamental to science really has changed in the last ten years. In certain fields which practice Data Science according to three principles I will describe, progress is simply dramatically more rapid than in those fields that don't yet make use of it. Researchers in the adhering fields are living through a period of very profound transformation, as they make a transition to frictionless reproducibility. This transition markedly changes the rate of spread of ideas and practices, and marks a kind of singularity, because it affects mindsets and paradigms and erases memories of much that came before. Many phenomena driven by this transition are misidentified as signs of an AI singularity. Data Scientists should understand what's really happening and their driving role in these developments.

Network archaeology: models and some recent results

Gábor Lugosi

Pompeu Fabra University, Spain

Abstract

Large networks that change dynamically over time are ubiquitous in various areas such as social networks, and epidemiology. These networks are often modeled by random dynamics which, despite being relatively simple, give a quite accurate macroscopic description of real networks. "Network archaeology" is an area of combinatorial statistics in which one studies statistical problems of inferring the past properties of such growing networks. In this talk we discuss some simple network models and review recent results on revealing the past of the networks.

Statistical Inference, Asymmetry of Information, and Statistical Contract Theory

Michael I. Jordan

University of California, Berkeley, USA

Abstract

Contract theory is the study of economic incentives when parties transact in the presenceof private information. We augment classical contract theory to incorporate a role forlearning from data, where the overall goal of the adaptive mechanism is to obtain desiredstatistical behavior. We consider applications of this framework to problems in federatedlearning, the delegation of data collection, and principal-agent regulatory mechanisms.

Chapter 2

Invited Talks

Weak Limits for Empirical Entropic Optimal Transport: Beyond Smooth Costs

Alberto González Sanz[1], **Shayan Hundrieser**[2]

[1] *Department of Statistics, Columbia University, US, ag4855@columbia.edu*
[2]*Institute for Mathematical Stochastics, University of Göttingen, Germany, s.hundrieser@math.uni-goettingen.de*

Abstract

We establish weak limits for the empirical entropy regularized optimal transport cost, the expectation of the empirical plan and the conditional expectation. Our results require only uniform boundedness of the cost function and no smoothness properties, thus emphasizing the far-reaching regularizing nature of entropy penalization. To derive these results, we employ a novel technique that sidesteps the intricacies linked to empirical process theory and the control of suprema of function classes determined by the cost. Instead, we perform a careful linearization analysis for entropic optimal transport with respect to an empirical L^2-norm, which enables a streamlined analysis. As a consequence, our work gives rise to new implications for a multitude of transport-based applications under general costs, including pointwise distributional limits for the empirical entropic optimal transport map estimator, kernel methods as well as regularized colocalization curves. Overall, our research lays the foundation for an expanded framework of statistical inference with empirical entropic optimal transport.

Keywords

Central Limit Theorem; Entropy Regularization; Infinite order V-statistics; Optimal transport.

Valid statistical inference with privacy constraints

Aleksandra (Seša) Slavković

Penn State University, Statistics, USA, sesa@psu.edu

Abstract

A vast amount of sensitive data (e.g., health, financial, genomic, survey data) is collected and archived by corporations, government agencies, health networks, and social networking websites. The social benefits of analyzing these data are significant, and include support for open data access and reproducibility. However, the release of these data and/or analyses can be devastating to the privacy of individuals and organizations. Limiting the disclosure risk of sensitive data and statistical analyses is a long-standing problem in statistics. Differential privacy (DP) and its variants now provide a useful framework for mathematically provable privacy guarantees in a transparent manner in support of releasing summary statistics and synthetic data. DP methods/mechanisms require the introduction of randomness which potentially reduces the utility of statistical results especially in finite samples. In this talk, I will give an overview of challenges associated with protecting confidential data, of statistical data privacy and its links to DP. I will describe a general framework, built on sound statistical principles from measurement error, robustness and the likelihood-based inference, and give few examples of how to achieve optimal statistical inference under formal privacy.

Keywords

Disclosure limitation, differential privacy, inference, synthetic data.

Evaluating causal effects on time-to-event outcomes in an RCT in Oncology with treatment discontinuation due to adverse events

Veronica Ballerini[1], **Björn Bornkamp**[2], **Alessandra Mattei**[3], **Fabrizia Mealli**[4], **Craig Wang**[5], **Yufen Zhang**[6]

[1] *University of Florence, Italy, veronica.ballerini@unifi.it*
[2] *Global Drug Development, Novartis Pharmaceuticals Corporation, Switzerland, bjoern.bornkamp@novartis.com*
[3] *University of Florence, Italy, alessandra.mattei@unifi.it*
[4] *European University Institute, Italy, Fabrizia.Mealli@eui.eu*
[5] *Global Drug Development, Novartis Pharmaceuticals Corporation, Switzerland, craig.wang@novartis.com*
[6] *Global Drug Development, Novartis Pharmaceuticals Corporation, Switzerland, yufen.zhang@novartis.com*

Abstract

In clinical trials, patients sometimes discontinue study treatments prematurely due to reasons such as adverse events. Since treatment discontinuation occurs after the randomization as an intercurrent event, it makes causal inference more challenging. The Intention-To-Treat (ITT) analysis provides valid causal estimates of the effect of treatment assignment; still, it does not take into account whether or not patients had to discontinue the treatment prematurely. We propose to deal with the problem of treatment discontinuation using principal stratification, which is recognised in the ICH E9(R1) addendum as a strategy for handling intercurrent events. Under this approach, we can decompose the overall ITT effect into *principal* causal effects for groups of patients defined by their potential discontinuation behaviour in continuous time. In this framework, we must consider that discontinuation happening in continuous time generates an infinite number

of principal strata; furthermore, discontinuation time is not defined for patients who would never discontinue. An additional complication is that discontinuation time and time-to-event outcomes, which are often the main endpoints in clinical trials, are subject to administrative censoring. We employ a flexible model-based Bayesian approach to deal with such complications. We apply the Bayesian principal stratification framework to analyze synthetic data based on a recent clinical trial in Oncology, aiming to assess the causal effects of a new investigational drug combined with standard of care versus standard of care alone on progression-free survival. We simulate data under different assumptions that reflect real situations where patients' behaviour depends on critical baseline covariates. Finally, we highlight how such an approach makes it straightforward to characterise patients' discontinuation behaviour with respect to the available covariates with the help of a simulation study.

Keywords

Causal inference, Censoring, MCMC, Potential outcomes, Principal stratification.

Mechanistic knowledge, machine learning and causal inference

Alexander Volfovsky[1]

[1] *Duke University, Department of Statistical Science,*
alexander.volfovsky@duke.edu

Abstract

At their core, the assumptions needed for causal inference are concerned with removing the effects of potentially unobserved quantities. We may know that a drug is given but maybe not when, we may observe where a disease is transmitted but maybe not exactly from whom, yet in both these settings we might be interested in causal questions: Does the drug have an effect? Does a mitigation strategy work to prevent future transmission? Because these processes are governed by established biological and social mechanisms, mechanistic models can provide invaluable insights into the interactions between biological and social objects (drug diffusion in the body, transmission probabilities between individuals, information diffusion over networks). Conditioning on these models can provide more credibility to the necessary assumptions for causal inference. We present two case studies of leveraging these types of models within advanced machine learning pipelines for causal inference: (1) we analyze observational data of critically ill patients and identify the effect of seizures if they were not treated, and (2) we employ a mechanistic model of disease transmission to help design a trial for evaluating a non-pharmaceutical intervention.

Keywords

Causal inference, mechanistic model, stochastic process, interventions.

Robust estimation in finite state space hidden Markov models

Alexandre Lecestre[1]

University of Luxembourg, Department of Mathematics, Luxembourg, alexandre.lecestre@uni.lu

Abstract

Hidden Markov models (HMMs) are latent variable models where the latent variables form a Markov chain. Their flexibility is quite useful to model observations with dependence. HMMs have been used for a wide variety of applications, in particular for speech recognition or character recognition. Most applications consider homogeneous finite state space HMMs, i.e. when the latent Markov chain is homogeneous and only take a finite number of values. We consider the problem of estimating the different parameters of a finite state space stationary HMM. In addition, we want our estimators to be robust to misspecification, contamination and outliers. Existing results show that it is possible to deduce the different parameters from the stationary law of L consecutive observations, for L large enough. Therefore the problem can be reduced to the (robust) estimation of a stationary distribution. We base our estimation procedure on ρ-estimators developed by Baraud et al. (2017) and Baraud & Birgé (2018). Those estimators are quite general and proven to be robust in the context of independent independent observations. We show that they are also robust to dependence, in the sense that the performance of the estimator is not worse when the observations are "almost independent". We can use the mixing properties of finite state space HMMs to select a subset of the observations that are "almost independent". Then the problem is to realize a compromise between the sample size and the distance from independence. We obtain the desired results of robustness and optimal convergence rates. Our method for the estimation of a stationary distribution is not specific to hidden Markov models and can

be applied to Markovian processes with similar mixing properties. We illustrate it with the estimation of the invariant distribution from discrete observations for a class of diffusion processes.

Keywords

Robust estimation, hidden Markov models, diffusion process.

References

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

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

Lecestre, A. (2023). Robust estimation for ergodic Markovian processes. arXiv:2307.03666 [math:ST].

Fuzzy Statistical Depth

Alicia Nieto-Reyes[1,2], **Luis González-de la Fuente**[1,3], **Pedro Terán**[4]

[1] *University of Cantabria, Department of Mathematics, Statistics and Computer Science, Spain,*
[2] *alicia.nieto@unican.es,*
[3] *luis.gonzalezd@unican.es*
[4] *University of Oviedo, Department of Statistics, Operations Research and Didactics of Mathematics, Spain, teranpedro@uniovi.es*

Abstract

A statistical data depth function orders the elements of a space with respect to a distribution. The notion of statistical data depth consists of a series of properties that vary according to the underlying space. Here we will comment on the suitable properties for fuzzy spaces and will provide extensions of multivariate depth functions to the fuzzy framework. We will analyze which of these extensions do satisfy the considered properties.

Keywords

Fuzzy Data, Nonparametric Statistics, Robust Statistics, Statistical Data Depth.

Examining the validity and fairness of societally high-stakes decision-making algorithms

Amanda Coston

Abstract

Automated decision systems are used for decision-making in societally high-stakes settings from child welfare and criminal justice to healthcare and consumer lending. Concerns around the suitability and equity of these systems require urgent attention. Much of the current discourse on responsible use focuses on fairness and ethics, often overlooking first-order questions of validity. In this talk, we explore the important role validity plays in responsible use and consider its implications for fairness. Drawing on validity theory from the social sciences, we develop a taxonomy of challenges that threaten validity in the algorithmic decision-making context. We delve into a couple common challenges — selection bias and missing data — in two societally consequential domains, consumer credit lending and child welfare screening. We illustrate how failure to properly address these issues can invalidate standard fairness assessments and undermine fairness corrective interventions. To resolve these issues, we present an alternative method for conducting fairness assessments and corrective interventions that addresses common forms of selection bias and missing data using techniques from causal inference. We conclude by considering the broader question of governance of high-stakes decision-making algorithms.

Clustering for dynamically generated stochastic processes

Ana Cristina Moreira Freitas

[1] *Universidade do Porto, Faculdade de Economia & Centro de Matemática, Portugal, amoreira@fep.up.pt*

Abstract

We consider stochastic processes arising from dynamical systems by evaluating an observable function along the orbits of the system and analyse the possible extreme value laws for these stochastic processes. In this context, the existence of an extremal index less than 1 is associated to the occurrence of periodic phenomena. For generic points, the exceedances, in the limit, are isolated and occur at Poisson times. But around periodic points the picture is different: the respective point processes of exceedances converge to a compound Poisson process, so instead of isolated exceedances, we have entire clusters of exceedances occurring at Poisson times with a geometric distribution ruling its multiplicity. The extremal index usually coincides with the reciprocal of the mean of the limiting cluster size distribution. Here, we build dynamically generated stochastic processes with an extremal index for which that equality does not hold. The mechanism we use is based on considering observable functions maximized at at least two points of the phase space, where one of them is an indifferent periodic point. For the second point at which the observable function is maximized we consider either a periodic or a nonperiodic point. We explore two such examples and, for each of them, we compute the extremal index and present the corresponding cluster size distribution.

Keywords

Dynamical Systems, Extremal Index, Point Processes of Rare Events.

Ackowledgements

This work has been financed by Portuguese public funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., in the framework of the project 2022.07167.PTDC and CMUP's project with reference UIDB/00144/2020.

Understanding posting behavior on social media using functional data analysis

Xiaoxia Champonr[1], Ana-Maria Staicu[2], Anthony Weishampel[3], Chathura Jayalah[4], William Rand[5]

[1] *North Carolina State University, Department of Statistics, USA, xzhao17@ncsu.edu*
[2] *North Carolina State University, Department of Statistics, USA, astaicu@ncsu.edu*
[3] *North Carolina State University, Department of Statistics, USA, aweishampel@gmail.com*
[4] *University of Central Florida, Department of Industrial and Engineering Management Systems, USA, acj@knights.ucf.edu*
[5] *North Carolina State University, Poole College of Management, usa, wmrand@ncsu.edu*

Abstract

Social media provides more insight into consumer behavior than companies have ever had, and firms can interact with consumers on social media to increase their brand loyalty and solve concerns they might have. However, it is critical for companies to evaluate whether the consumers they interact with have the potential to positively promote the firm. This can be challenging, especially when limited information is available about social media users. Our research proposes a flexible methodology to cluster many Twitter users based on their behavioral differences to solve this problem. We provide a novel framework that views users' high-frequency postings during a specified timeframe as densely-observed categorical functional data, and proposes to cluster them using latent user-specific characteristics. This leads to an interpretable and computationally-efficient algorithm and enables us to gain insights into the posting behavior of social media users. While our methods are inspired by a Twitter application they can be applied to understand posting behavior across various social media platforms.

Finite-sample properties of the methods are investigated through simulations.

Keywords

Categorical functional data analysis, clustering, multivariate latent process, Twitter.

Assessing Privacy and Security Risk via Composite Metrics

Anand N. Vidyashankar[1], **Lei Li**[2], **Lucy Doyle**[3], **Crissa Marshburn**[4]

[1] *George Mason University, Department of Statistics, USA, avidyash@gmu.edu*
[2] *George Mason University, Department of Statistics, USA, sddxlilei@gmail.com*
[3] *McKesson Corporation, Strategic information management, lucyd0491@gmail.com*
[4] *McKesson Corporation, Global privacy office, de-identification, Crissa.Clark@McKesson.com*

Abstract

Privacy and security laws are continuously evolving, with states adopting and refining existing regulatory guidelines. Recently, within the U.S., Delaware passed a comprehensive data privacy law, joining twelve other states to provide consumers with privacy rights. These privacy laws tend to include additional security requirements and recommendations. Data warehouses that process personally identifiable information (PII) adopt disclosure control mechanisms to adhere to federal and state privacy and security regulatory guidelines. Thus, it is beneficial to identify metrics that detect loss of privacy due to security vulnerabilities and vice-versa. In this presentation, we describe a new class of composite metrics integrating key performance indicators (KPIs) of security policies and privacy risk measures to determine privacy loss. We study the statistical properties of the proposed metrics and provide an uncertainty assessment that facilitates the development of policies and procedures for data sharing.

Keywords

Personally identifiable information, Key performance indicators, Composite metrics, Uncertainty assessment.

Measurement Error Models for Spatial Network Lattice Data: Analysis of Car Crashes in Leeds

Andrea Gilardi[1], **Riccardo Borgoni**[2], **Luca Presicce**[2], **Jorge Mateu**[3]

[1] *Politecnico di Milano, Department of Mathematics, Milan (IT), andrea.gilardi@polimi.it*
[2] *Università degli Studi di Milano - Bicocca, Department of Economics, Management and Statistics (DEMS), Milan (IT), riccardo.borgoni@unimib.it - l.presicce@campus.unimib.it*
[3] *Universitat Jaume I, Department of Mathematics, Castellon (ES), mateu@uji.es*

Abstract

Road casualties represent an alarming concern for modern societies, especially in poor and developing countries. In the last years, several authors developed a series of statistical approaches to help authorities implement new policies and mitigate the problem. These models are typically developed taking into account socio-demographic variables such as population density or traffic volumes, but they usually ignore that these external factors may be suffering from measurement error which can severely bias the statistical inference. Therefore, we propose a Bayesian hierarchical model to analyse car crashes occurrences at the network lattice level taking into account measurement error in the spatial covariates. The suggested methodology is exemplified by considering the collisions that occurred in the road network of Leeds (UK) from 2011 to 2019. Traffic volumes are approximated at the street segment level using an extensive set of road counts obtained from mobile devices and the estimates are adjusted using a spatial measurement error correction. Our results show that ignoring the measurement error considerably worsens the model's fit and attenuates the effects of the imprecise covariates.

Keywords

Bayesian Hierarchical Models, Car Crashes, Network Lattice, Measurement Error, Spatial Networks.

Estimating a geodesic normal distribution on the sphere with elliptical contours

Andrea Meilán-Vila[1] **and José E. Chacón**[2]

[1] *Universidad Carlos III de Madrid, Department of Statistics, Spain, ameilan@est-econ.uc3m.es*
[2] *Universidad de Extremadura, Department of Mathematics, Spain, jechacon@unex.es*

Abstract

The classical von Mises-Fisher distribution on the sphere belongs to the class of so-called extrinsic normal distributions since it is based on the Euclidean distance inherited by the sphere when embedded in three-dimensional Euclidean space. More recently, intrinsic spherical normal distributions have been introduced, which rely on the more natural geodesic distance on the sphere, taking into account its curvature. The isotropic versions of these geodesic normal distributions are intrinsically defined on the sphere, but it is necessary to project them onto the tangent space to define their anisotropic counterparts. In this work, we introduce a new geodesic normal distribution on the sphere, which is defined in a fully intrinsic manner. The density level sets of this distribution are true spherical ellipses, allowing it to be considered anisotropic. A procedure for estimating the parameters of the distribution is introduced and validated in practice.

Keywords

Anisotropic distribution, directional data, geodesic distance.

Network Comparison via Optimal Transport of Markov Chains

Andrew Nobel[1], **Bongsoo Yi**[2], **Kevin O'Connor**[3], **Kevin McGoff**[4]

[1] *University of North Carolina at Chapel Hill, Department of Statistics and Operations Research, nobel@email.unc.edu*
[2] *University of North Carolina at Chapel Hill, Department of Statistics and Operations Research*
[3] *University of North Carolina at Chapel Hill, Department of Statistics and Operations Research*
[4] *Department of Mathematics and Statistics, University of North Carolina at Charlotte*

Abstract

Networks are frequently used to represent and study pairwise interactions between a collection of objects or individuals under study, and have become objects of study in their own right. In this talk I will describe a procedure called NetOTC (network optimal transition coupling) that can be used to compare and align two networks. The networks of interest may be directed or undirected, weighted or unweighted, and may have distinct vertex sets of different sizes. Given two networks and a cost function relating their vertices, NetOTC finds a transition coupling of their associated random walks having minimum expected cost. The minimizing cost quantifies the difference between the networks, while the optimal transport plan itself provides alignments of the vertices and edges of the two networks. Coupling of the full random walks ensures that NetOTC captures local and global information about the networks and preserves edges. I will review some useful theoretical properties of NetOTC, and present numerical experiments supporting its performance.

Keywords

Graph alignment, graph comparison, optimal transport, isomorphism.

Simple Binary Hypothesis Testing: Optimal Non-asymptotic Rates

Ankit Pensia[1], **Po-Ling Loh**[2], **Varun Jog**[3]
[1] *IBM Research, ankitp@ibm.com*
[2] *University of Cambridge, pll28@dppms.cam.ac.uk*
[3] *University of Cambridge, vj270@dppms.cam.ac.uk*

Abstract

Simple binary hypothesis testing is a fundamental problem in statistics, where the goal is to distinguish between the two candidate distributions (say, the distributions p and q) given i.i.d. samples. We study the standard Bayesian setup, where the true distribution is generated by a prior distribution (in this case, Bernoulli distribution over $\{p, q\}$). Sample complexity is a non-asymptotic quantity that refers to the number of samples that are necessary and sufficient so that the probability of incorrect detection is small. A fundamental result in statistics is that the best error exponent, asymptotically, is the Chernoff information (for all priors), hinting that the sample complexity should be characterized by the Chernoff information (for all priors). This observation is further backed by the well-known fact when the prior is uniform, the sample complexity is indeed characterized by Chernoff information (equivalent to Hellinger divergence, up to constants). However, when the priors are unequal, simple examples show that the sample complexity does not depend on Chernoff information for many regimes; moreover, it crucially depends on the prior. Our main contribution is a tight characterization of sample complexities of simple binary hypothesis testing in nearly all regimes, where we show that the sample complexity is given by a different (prior-dependent) f-divergence. Our results extend to various other formulations: prior-free hypothesis testing, robust binary hypothesis testing, and hypothesis testing under local communication and privacy constraints.

Keywords

Hypothesis testing, instance-optimality, non-asymptotic rates, Chernoff information.

Mode-wise Principal Subspace Pursuit and Matrix Spiked Covariance Model

Anru Zhang[1]

[1] *Duke University, Durham, NC, USA; email: anru.zhang@duke.edu*

Abstract

In this talk, we introduce a novel framework called Mode-wise Principal Subspace Pursuit (MOP-UP) to extract hidden variations in both the row and column dimensions for matrix data. To enhance the understanding of the framework, we introduce a class of matrix-variate spiked covariance models that serve as inspiration for the development of the MOP-UP algorithm. The MOP-UP algorithm consists of two steps: Average Subspace Capture (ASC) and Alternating Projection (AP). These steps are specifically designed to capture the row-wise and column-wise dimension-reduced subspaces which contain the most informative features of the data. ASC utilizes a novel average projection operator as initialization and achieves exact recovery in the noiseless setting. We analyze the convergence and non-asymptotic error bounds of MOP-UP, introducing a blockwise matrix eigenvalue perturbation bound that proves the desired bound, where classic perturbation bounds fail. The effectiveness and practical merits of the proposed framework are demonstrated through experiments on both simulated and real datasets. This is a joint work with Runshi Tang and Ming Yuan.

Keywords

Alternating projection, average projection operator, dimensionality reduction, mode-wise principal subspace pursuit, principal component analysis.

An Empirical Exploration of the Law or Large Numbers

Armin Schwartzman[1]

[1] *University of California, San Diego, Halıcıoğlu Data Science Institute, USA, armins@ucsd.edu*

Abstract

Frequentist estimation theory, including estimation consistency and calculation of standard errors, is based fundamentally on the law of large numbers (LLN). The LLN relies on the assumption that the observations or noise elements are iid, or some variation thereof. However, the iid assumption is difficult to enforce and check in practice. In fact, it is an idealized model and often does not hold in real data, either because the data is collected observationally or because the data hides unmeasured variables. In this work, I explore the use of subsampling to directly assess whether the LLN holds empirically, without having to check the iid assumption. Interestingly, for the sample mean, the empirical convergence rate of the standard error ranges from n^{-1} to $n^{-1/2}$ depending on the degree of (apparent) randomness of the data sequence. Practical examples are given from fMRI and genomics data.

Keywords

Frequentist statistics, bootstrap, subsampling, random processes, convergence rate.

Effectively Combining Nonparametric Functionals

Arne Bathke[1], **Jonas Beck**[1], **Patrick Langthaler**[1]

[1] *Intelligent Data Analytics Lab, Department of Artificial Intelligence and Human Interfaces, Faculty of Digital and Analytical Sciences, University of Salzburg, Austria, Arne.Bathke@plus.ac.at*

Abstract

Nonparametric statistical methods are usually characterized by rather generous invariance properties, as well as robustness against departures from narrow model classes. This has made them very popular in the last decades, and the attractivity of nonparametric methods transfers to many data science applications where specific parametric models are not justifiable. However, a shortcoming of all those nonparametric procedures that rely on the relative effect as their base functional is their inability to capture differences between distributions that cannot be described by a stochastic tendency. To this end, we have introduced a functional describing distributional overlap and derived a consistent estimator along with its asymptotic distribution, even jointly with that of the relative effect estimator. Combining these two functionals allows for much more versatile inference, which we will demonstrate in this presentation. Also, we will try to address the issue of interpretability of the resulting effect measures, as straightforward interpretability is key to their usability in practice.

Keywords

Niche Overlap, Probabilistic Index, Relative Effect, Resampling, Wilcoxon-Mann-Whitney Functional.

From limited patient data, to high frequency synthetic data, to the differential equation of a breast tumour growth

Arnoldo Frigessi[1]

[1] *Integreat, University of Oslo, Norway, frigessi@uio.no*

Abstract

Let $X(t)$ be the volume of a solid tumour in a patient at time $t \in [0, T]$, over a period of T =three weeks, after a treatment was given at time 0. What is the ordinary differential equation that this volume solves for this tumour of this patient? This equation, or parts of it, might be useful as biomarkers for success of the therapy. If we would be able to measure $X(t)$ quite often during the three weeks, then one could estimate the differential equation using for example symbolic regression and genetic programming, see for example [Qian, Z., Kacprzyk, K., & van der Schaar, M. (2022, January). D-code: Discovering closed-form odes from observed trajectories, in International Conference on Learning Representations]. However, cancer volumes are measured by imaging only in the start and end of treatment, possibly a few more times, which is insufficient. We developed a mechanistic model of the cancer of the patient, a digital twin of that tumour, which allows to simulate synthetic volume trajectories, see [Lai, X., Geier, O. M., Fleischer, T., Garred, Ø., Borgen, E., Funke, S. W., ... & Frigessi, A. (2019). Toward personalized computer simulation of breast cancer treatment: A multiscale pharmacokinetic and pharmacodynamic model informed by multitype patient data. Cancer research, 79(16), 4293-4304.]. This is a stochastic model. By repeated running it we produce several synthetic trajectories, which can then be sampled at any wished frequency and so after used to estimate the differential equation. If we will be able, we will show the first differential equation of a breast cancer volume after chemotherapy treatment.

This is joint work with Håkon Taskén, Alvaro Köhn-Luque of Integreat, the University of Oslo and Jasmine Foo, Kevin Leder, Universtity of Minnesota, among others.

Keywords

Synthetic data generation by mechanistic modelling, Symbolic regression, Digital Twin, Closed-form differential equation.

Bayesian Plant-Capture Methods for Estimating Population Size from Uncertain Plant Captures

Yiran Wang[1], **Martin Lysy**[2], **Audrey Beliveau**[3]

[1] *University of Waterloo, Department of Statistics and Actuarial Science, Canada, yiran.wang2@uwaterloo.ca*
[2] *University of Waterloo, Department of Statistics and Actuarial Science, Canada, mlysy@uwaterloo.ca*
[3] *University of Waterloo, Department of Statistics and Actuarial Science, Canada, audrey.beliveau@uwaterloo.ca*

Abstract

Plant-capture is a capture-recapture method used to estimate the size of a population. In this method, decoys referred to as "plants" are introduced into the population. These plants are assumed to seamlessly blend with the population and are indistinguishable from regular individuals. The proportion of plants that get captured provides an estimate the capture probability. The inverse of this capture probability is then employed to extrapolate the population count into an estimate of the population size. However, existing plant-recapture methods typically do not consider the uncertainty in the capture status of the plants, particularly in the context of point-in-time surveys of the homeless. In this research, we introduce a range of Bayesian models and computational approaches designed to formally address this uncertainty. We validate the statistical performance of these models through simulation studies. Furthermore, we apply this novel methodology to estimate the homeless population size in various U.S. cities in the context of the Shelter and Street-Night (S-night) survey conducted by the U.S. Census Bureau.

Keywords

Bayesian modeling, capture-recapture, homeless, hierarchical modeling, missing at random.

Optimal Transport Based Colocalization Analysis

Axel Munk[1]

[1] *Georg August Universität Göttingen, Department of Mathematics and Computer Science, Germany, munk@math.uni-goettingen.de*

Abstract

Super-resolution fluorescence microscopy is a widely used technique in cell biology which enables the recording of multiple-color images with subdiffraction resolution. The enhanced resolution leads to new challenges regarding colocalization analysis of macromolecule distributions, i.e. the investigation of spatial arrangements of proteins and their interactions. We demonstrate that well-established methods for the analysis of colocalization in diffraction-limited datasets are not equally well suited for the analysis of high-resolution images. We propose optimal transport colocalization, which measures the minimal transporting cost below a given spatial scale to match two protein intensity distributions. Its validity on simulated data as well as on dual-color STED microscopy recordings of yeast and mammalian cells is demonstrated. This is extended to multiple color images (MultiMatch) and our methodology is illustrated on chain-like particle arrangements. We showcase that MultiMatch is able to consistently recover all present chain structures in three-color STED recordings of DNA origami nanorulers. MultiMatch statistically incorporates incomplete labeling enabling inference on existent, but not fully observable particle chains.

Keywords

Colocalization, multimarginal optimal transport, cell biology, combinatorial inference, resampling.

References

Tameling, C., Stoldt., S., Stephan, T., Naas, J., Jakobs, S. and Munk, A. (2021). Colocalization for super-resolution microscopy via optimal transport. *Nature Computational Science 1*, 199-2011.

Naas, J., Nies, G., Li, H., Stoldt, S., Schmitzer, B., Jakobs, S. and Munk, A. (2023) MultiMatch: Optimal matching colocalization in multi-color super-resolution microscopy. *submitted*.

Statistical analysis of non-convexity measures

Alejandro Cholaquidis[1], Ricardo Fraiman[2], Leonardo Moreno[3], Beatriz Pateiro-López[4]

[1] *Centro de Matemática, Facultad de Ciencias, Universidad de la República, Uruguay, acholaquidis@cmat.edu.uy*

[2] *Centro de Matemática, Facultad de Ciencias, Universidad de la República, Uruguay, rfraiman@cmat.edu.uy*

[3] *Instituto de Estadística, Departamento de Métodos Cuantitativos, FCEA, Universidad de la República, Uruguay, leonardo.moreno@fcea.edu.uy*

[4] *CITMAga, Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, Spain, beatriz.pateiro@usc.es*

Abstract

Several measures of non-convexity (departures from convexity) have been introduced in the literature, both for sets and functions. Some of them are of geometric nature, while others are more of topological nature. We address the statistical analysis of some of these measures of non-convexity of a set S, by dealing with their estimation based on a sample of points in S. We introduce also a new measure of non-convexity. We discuss briefly about these different notions of non-convexity, prove consistency and find the asymptotic distribution for the proposed estimators. We also consider the practical implementation of these estimators and illustrate their applicability to a real data example.

Keywords

Convex hull, Convexity measure, Shape analysis, Skin lesions.

References

Cholaquidis, A., R. Fraiman, L. Moreno, and B. Pateiro-López (2023) Statistical analysis of measures of non-convexity. *TEST*, doi:10.1007/s11749-023-00889-4

A Mean Field Approach to Empirical Bayes Estimation in High-dimensional Linear Regression

Sumit Mukherjee[1], **Bodhisattva Sen**[2], **Subhabrata Sen**[3]

[1] *Columbia University, Department of Statistics, USA,*
sm3949@columbia.edu
[2] *Columbia University, Department of Statistics, USA,*
bodhi@stat.columbia.edu
[3] *Harvard University, Department of Statistics, USA,*
subhabratasen@fas.harvard.edu

Abstract

We study empirical Bayes estimation in high-dimensional linear regression. To facilitate computationally efficient estimation of the underlying prior, we adopt a variational empirical Bayes approach, introduced originally in Carbonetto and Stephens (2012) and Kim et al. (2022). We establish asymptotic consistency of the nonparametric maximum likelihood estimator (NPMLE) and its (computable) naive mean field variational surrogate under mild assumptions on the design and the prior. Assuming, in addition, that the naive mean field approximation has a dominant optimizer, we develop a computationally efficient approximation to the oracle posterior distribution, and establish its accuracy under the 1-Wasserstein metric. This enables computationally feasible Bayesian inference; e.g., construction of posterior credible intervals with an average coverage guarantee, Bayes optimal estima- tion for the regression coefficients, estimation of the proportion of non-nulls, etc. Our analysis covers both deterministic and random designs, and accommodates correlations among the features. To the

best of our knowledge, this provides the first rigorous nonparametric empirical Bayes method in a high-dimensional regression setting without sparsity.

Keywords

Consistency of nonparametric maximum likelihood, evidence lower bound, posterior inference, variational approximation.

The Cost of Data Bias: A Model of the Diminishing Value of Noisy Information

Boris Babic[1], **Robin Gong**[2]

[1] *University of Toronto, Statistics and Philosophy, Canada*
[2] *Rutgers University, Statistics, USA*

Abstract

The purpose of this project is to provide a more realistic assessment of the true worth of "big data", and to temper the sometimes unduly optimistic expectations about its information value. We are going to do this by leveraging and combining some classic literature on (a) imperfect survey sampling and (b) imprecise probabilities in a novel and, we believe, insightful way. Our basic point is simple: all data comes with errors and there is a price to be paid in not accounting for those errors. We will demonstrate in several ways how the information value of data substantially diminishes when we take errors into account.

Keywords

Data, Bias, Noise, Misclassification.

Sampling depth trade-off in function estimation under a two-level design

Akira Horiguchi[1], **Li Ma**[1], **Botond Szabo**[2]

[1] *Department of Statistical Science, Duke University, USA*
[2]*Department of Decision Sciences and Institute for Data Science and Analytics, Bocconi University, Italy*

Abstract

Many modern statistical applications involve a two-level sampling scheme that first samples subjects from a population and then samples observations on each subject. These schemes often are designed to learn both the population-level functional structures shared by the subjects and the functional characteristics specific to individual subjects. Common wisdom suggests that learning population-level structures benefits from sampling more subjects whereas learning subject-specific structures benefits from deeper sampling within each subject. Oftentimes these two objectives compete for limited sampling resources, which raises the question of how to optimally sample at the two levels. We quantify such sampling-depth trade-offs by establishing the L_2 minimax risk rates for learning the population-level and subject-specific structures under a hierarchical Gaussian process model framework where we consider a Bayesian and a frequentist perspective on the unknown population-level structure. These rates provide general lessons for designing two-level sampling schemes. Interestingly, subject-specific learning occasionally benefits more by sampling more subjects than by deeper within-subject sampling. We also construct estimators that adapt to unknown smoothness and achieve the corresponding minimax rates. We conduct two simulation experiments validating our theory and illustrating the sampling trade-off in practice, and apply these estimators to two real datasets.

Keywords

Sampling design, nonparametrics, Gaussian processes, hierarchical models, Bayesian models.

Bayesian computational methods for spatial models with intractable likelihoods

Brian J Reich[1]

[1] *North Carolina State University, Department of Statistics, USA, bjreich@ncsu.edu*

Abstract

Extreme value analysis is critical for understanding the effects of climate change. Exploiting the spatiotemporal structure of climate data can improve estimates by borrowing strength across nearby locations and provide estimates of the probability of simultaneous extreme events. A fundamental probability model for spatially-dependent extremes is the max-stable processes. While this model is theoretically justified, it leads to an intractable likelihood function. We propose to use deep learning to overcome this computational challenge. The approximation is based on simulating millions of draws from the prior and then the data-generating process, and then using deep learning density regression to approximate the posterior distribution. We verify through extensive simulation experiments that this approach leads to reliable Bayesian inference, and discuss extensions to other spatial processes with intractable likelihoods including the autologistic model for binary data and SIR model for the spread of an infectious disease.

Keywords

Deep learning, Machine learning, Max-stable process, Simulation-based inference.

Inference for model-agnostic longitudinal variable importance

Brian D. Williamson[1]

[1] *Biostatistics Division, Kaiser Permanente Washington Health Research Institute, brian.d.williamson@kp.org*

Abstract

In prediction settings where data are collected over time, it is often of interest to understand both the importance of variables for predicting the response at each time point and the importance summarized over the time series. Building on recent advances in estimation and inference for variable importance measures, we define summaries of variable importance trajectories. These measures can be estimated and the same approaches for inference can be applied regardless of the choice of the algorithm(s) used to estimate the prediction function. We propose a nonparametric efficient estimation and inference procedure as well as a null hypothesis testing procedure that are valid even when complex machine learning tools are used for prediction. Through simulations, we demonstrate that our proposed procedures have good operating characteristics, and we illustrate their use by investigating the longitudinal importance of risk factors for suicide attempt.

Keywords

Intrinsic variable importance, prediction, machine learning, longitudinal data.

An Extension of the Unified Skew-Normal Family of Distributions and Application to Bayesian Binary Regression

Brunero Liseo[1], **Paolo Onorati**[2]

[1] *Sapienza Università di Roma, MEMOTEF, Italy,*
brunero.liseo@uniroma1.it
[2] *Sapienza Università di Roma, MEMOTEF, Italy,*
p.onorati@uniroma1.it

Abstract

We consider the general problem of Bayesian binary regression and we introduce a new class of distributions, the Perturbed Unified Skew Normal (pSUN), which generalizes the Unified Skew-Normal (SUN) class. We show that the new class is conjugate to any binary regression model, provided that the link function may be expressed as a scale mixture of Gaussian densities. We discuss in detail the popular logit case, and we show that, when a logistic regression model is combined with a Gaussian prior, posterior summaries such as cumulants and normalizing constants can be easily obtained through the use of an importance sampling approach, opening the way to straightforward variable selection procedures. For more general priors, the proposed methodology is based on a simple Gibbs sampler algorithm. We also claim that, in the $p > n$ case, the proposed methodology shows better performances - both in terms of mixing and accuracy - compared to the existing methods. We illustrate the performance through several simulation studies and two data analyses.

Keywords

Importance Sampling, Kolmogorov Distribution, Logistic Regression, Scale Mixture of Gaussian Densities.

Follow the Arrow ... Plot

Carina Silva[1,2], **Maria Antónia Amaral Turkman**[2], **Lisete Sousa**[2,3]

[1] *H&TRC—Health & Technology Research Center, ESTeSL—Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Portugal, carina.silva@estesl.ipl.pt*
[2] *Centro de Estatística e Aplicações (CEAUL), Universidade de Lisboa, Portugal, antonia.turkman@ciencias.ulisboa.pt*
[3] *Faculdade de Ciências da Universidade de Lisboa, DEIO, Portugal, lmsousa@ciencias.ulisboa.pt*

Abstract

A common task in analyzing genomic data is to determine which genes are differentially expressed under two (or more) kinds of tissue samples or samples submitted under different experimental conditions. It is well known that biological samples are heterogeneous due to factors such as molecular subtypes or genetic background, which are often unknown to the researcher. For instance, in experiments which involve molecular classification of tumors it is important to identify significant subtypes of cancer. Bimodal or multimodal distributions often reflect the presence of subsamples mixtures. Consequently, truly differentially expressed genes on sample subgroups may be lost if usual statistical approaches are used. A graphical tool which identifies genes with up and down regulation, as well as genes with differential expression revealing hidden subclasses, that are usually missed if current statistical methods are used. This tool, Arrow Plot, is based on two measures, namely the overlapping coefficient (OVL) between two densities and the area under (AUC) the receiver operating characteristic (ROC) curve. Our results indicate that the Arrow Plot represents a flexible and useful tool for the analysis of gene expression profiles.

Acknowledgements
This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020.

Keywords

Arrow Plot, AUC, OVL, subclasses, gene expression.

Shape Preserving Differential Privacy

Carlos J. Soto

University of Massachusetts Amherst, Department of Mathematics and Statistics, U.S.A., carlossoto@umass.edu

Abstract

Motivated by the problem of statistical shape analysis, this work considers the problem of producing sanitized differentially private estimates through the K-norm Gradient Mechanism (KNG) when the data or parameters live on a Riemannian manifold. In particular, Kendall's 2D shape space is a Riemannian manifold which is positively curved. KNG requires an objective function and produces a sanitized estimate by favoring values which produce gradients close to zero of the objective. This work extends KNG to consider objective functions which take on manifold-valued data or parameters. Respecting the nature of the data leads to utility gains when compared to sanitization in an ambient space, as well as removing the need for post-processing. Specifically, this work proposes sanitizing the Fréchet mean for the sphere, symmetric positive definite matrices, and Kendall's 2D shape space under pure differentially private framework with an application to corpus callosum data.

Keywords

Differential Privacy, Shape Analysis, Manifolds, Shape Space.

Wasserstein Autoregressive Models for Density Time Series

Chao Zhang[1], **Piotr Kokoszka**[2], **Alexander Petersen**[3]

[1] *University of California, Santa Barbara, Department of Statistics and Applied Probability, Santa Barbara, CA, USA, czhang@pstat.ucsb.edu*

[2] *Colorado State University, Department of Statistics, Fort Collins, CO, USA, piotr.kokoszka@colostate.edu*

[3] *Brigham Young University, Department of Statistics, Provo, UT, USA, petersen@stat.byu.edu*

Abstract

Data consisting of time-indexed distributions of cross-sectional or intraday returns have been extensively studied in finance, and provide one example in which the data atoms consist of serially dependent probability distributions. Motivated by such data, we propose an autoregressive model for density time series by exploiting the tangent space structure on the space of distributions that is induced by the Wasserstein metric. The densities themselves are not assumed to have any specific parametric form, leading to flexible forecasting of future unobserved densities. The main estimation targets in the order-p Wasserstein autoregressive model are Wasserstein autocorrelations and the vector-valued autoregressive parameter. We propose suitable estimators and establish their asymptotic normality, which is verified in a simulation study. The new order-p Wasserstein autoregressive model leads to a prediction algorithm, which includes a data driven order selection procedure. Its performance is compared to existing prediction procedures via application to four financial return data sets, where a variety of metrics are used to quantify forecasting accuracy. For most metrics, the proposed model outperforms existing methods in two of the data sets, while the best empirical performance in the other two data sets is attained by existing methods based on functional transformations of the densities.

Keywords

Random Densities, Wasserstein Metric, Time Series, Distributional Forecasting.

Low-Rank and Sparse Decomposition for Brain Functional Connectivity in Naturalistic fMRI Data

Chee-Ming Ting[1], **Jeremy I. Skipper**[2], **Fuad Noman**[1], **Steven L. Small**[3], **Hernando Ombao**[4]

[1] *Monash University, Malaysia Campus, Malaysia, ting.cheeming@monash.edu, fuad.noman@monash.edu*
[2] *University College London, UK, jeremy.skipper@ucl.ac.uk*
[3] *University of Texas at Dallas, USA, small@utdallas.edu*
[4] *King Abdullah University of Science and Technology, Saudi Arabia, hernando.ombao@kaust.edu.sa*

Abstract

We consider the challenges in extracting stimulus-related neural dynamics from other intrinsic processes and noise in naturalistic functional magnetic resonance imaging (fMRI). Most studies rely on inter-subject correlations (ISC) of low-level regional activity and neglect varying responses in individuals. We propose a novel, data-driven approach based on low-rank plus sparse (L+S) decomposition to isolate stimulus-driven dynamic changes in brain functional connectivity (FC) from the background noise, by exploiting shared network structure among subjects receiving the same naturalistic stimuli. The time-resolved multi-subject FC matrices are modeled as a sum of a low-rank component of correlated FC patterns across subjects, and a sparse component of subject-specific, idiosyncratic background activities. To recover the shared low-rank subspace, we introduce a fused version of principal component pursuit (PCP) by adding a fusion-type penalty on the differences between the rows of the low-rank matrix. The method improves the detection of stimulus-induced group-level homogeneity in the FC profile while capturing inter-subject variability. We develop an efficient algorithm via a linearized alternating direction

method of multipliers to solve the fused-PCP. Simulations show accurate recovery by the fused-PCP even when a large fraction of FC edges are severely corrupted. When applied to natural fMRI data, our method reveals FC changes that were time-locked to auditory processing during movie watching, with dynamic engagement of sensorimotor systems for speech-in-noise. It also provides a better mapping to auditory content in the movie than ISC.

Keywords

Low-rank plus sparse decomposition, brain network analysis, fMRI data.

A new p-value based multiple testing procedure with arbitrary dependence for generalized linear models

Joseph Rilling[1], **Cheng Yong Tang**[1]

[1] *Department of Statistics, Operations, and Data Science, Temple University*

Abstract

We address the challenge of addressing multiple testing issues in the context of generalized linear models where there is arbitrary dependence in the design matrix. Based on the concept of model-X knockoffs, we introduce an innovative approach to generate two independent sets of paired p-values for the purpose of testing model coefficients. These paired p-values serve as the foundation for a new multiple testing procedure, which has demonstrated its ability to effectively control the false discovery rate. Our empirical findings underscore the promising performance of this novel method.

Keywords:

False discovery rate, Generalized linear models, Multiple testing.

Semiparametric estimation of the transformation model by leveraging external aggregate data in the presence of population heterogeneity

Yu-Jen Cheng[1], **Yen-Chun Liu**[2], **Chang-Yu Tsai**[1], **Chiung-Yu Huang**[3]

[1] *Institute of Statistics, National Tsing Hua University, Hsin-Chu 300, Taiwan*

[2] *Department of Statistical Science, Duke University, Durham, North Carolina 27710, U.S.A.*

[3] *Department of Epidemiology & Biostatistics, University of California at San Francisco, San Francisco, California 94158, U.S.A., ChiungYu.Huang@ucsf.edu*

Abstract

Leveraging information in aggregate data from external sources to improve estimation efficiency and prediction accuracy with smaller-scale studies has drawn a great deal of attention in recent years. Yet, conventional methods often either ignore uncertainty in the external information or fail to account for the heterogeneity between internal and external studies. This article proposes an empirical likelihood-based framework to improve the estimation of the semiparametric transformation models by incorporating information about the t-year subgroup survival probability from external sources. The proposed estimation procedure incorporates an additional likelihood component to account for uncertainty in the external information and employs a density ratio model to characterize population heterogeneity. We establish the consistency and asymptotic normality of the proposed estimator and show that it is more efficient than the conventional pseudo-partial likelihood estimator without combining information. Simulation studies show that the proposed estimator yields little bias and outperforms

the conventional approach even in the presence of information uncertainty and heterogeneity. The proposed methodologies are illustrated with an analysis of a pancreatic cancer study.

Keywords

Aggregate data; Empirical likelihood, Information uncertainty, Leveraging information, Meta-analysis, Population heterogeneity, Semiparametric transformation model.

A Multiverse of Decisions: Fairness Implications of Algorithmic Profiling Schemes

Christoph Kern[1], **Jan Simson**[1], **Florian Pfisterer**[1]

[1] *LMU Munich, Department of Statistics, Germany,*
christoph.kern@stat.uni-muenchen.de

Abstract

A vast number of systems across the world use algorithmic decision-making (ADM) to (partially) automate decisions that have previously been made by humans. While these systems promise more objective and efficient decision-making, they are also susceptible to feeding forward biases that may be present in training data. In this talk, we highlight how such biases can be mitigated or reinforced along the modeling pipeline dependent on the decisions made during the ADM design. We first present an empirical use case of algorithmic profiling on the labor market and systematically compare and evaluate differently designed profiling models and predictions with respect to fairness metrics to illustrate vulnerabilities to modeling decisions. Motivated by this example, we introduce the method of multiverse analysis for algorithmic fairness that draws on insights from the field of psychology. In our proposed method, we turn implicit design decisions into explicit ones and demonstrate their fairness implications. By combining decisions, we create a grid of all possible "universes" of decision combinations. For each of these universes, we compute metrics of fairness and performance. The resulting dataset allows for nuanced evaluations of how (which) design decisions impact fairness. We illustrate how decisions during the design of a machine learning system can have surprising effects on its fairness and how to detect these effects using multiverse analysis.

Keywords

Statistical profiling, multiverse analysis, algorithmic fairness.

Bayesian Causal Inference with Uncertain Physical Process Interference

Nathan Wikle[1], **Corwin Zigler**[2]

[1] *University of Iowa, Department of Statistics, USA, nathan-wikle@uiowa.edu*
[2] *University of Texas at Austin, Department of Statistics and Data Sciences, USA, cory.zigler@austin.utexas.edu*

Abstract

Causal inference with spatial environmental data is often challenging due to the presence of interference: outcomes for observational units depend on some combination of local and non-local treatment. This is especially relevant when estimating the effect of power plant emissions controls on population health, as pollution exposure is dictated by (i) the location of point-source emissions, as well as (ii) the transport of pollutants across space via dynamic physical-chemical processes. In this work, we estimate the effectiveness of air quality interventions at coal-fired power plants in reducing two adverse health outcomes in Texas in 2016: pediatric asthma ED visits and Medicare all-cause mortality. We develop methods for causal inference with interference when the underlying network structure is not known with certainty and instead must be estimated from ancillary data. We offer a Bayesian, spatial mechanistic model for the interference mapping which we combine with a flexible non-parametric outcome model to marginalize estimates of causal effects over uncertainty in the structure of interference. Our analysis finds some evidence that emissions controls at upwind power plants reduce asthma ED visits and all-cause mortality, however accounting for uncertainty in the interference renders the results largely inconclusive.

Keywords

Causal inference, Interference, Air pollution, Networks.

Large Contingency Tables

Chong Wu[1], **Yisha Yao**[2], **Cun-Hui Zhang**[3]

[1] *Rutgers University, Department of Statistics, USA,*
chong.wu@rutgers.edu
[2] *Columbia University, Department of Statistics, USA,*
yy3381@columbia.edu
[3] *Rutgers University, Department of Statistics, USA,*
czhang@rutgers.edu

Abstract

We provide necessary and sufficient conditions for the chi-squared and normal approximations of Pearson's chi-squared statistics for the test of independence and the goodness-of-fit test when the cell probabilities of the multinomial data are in general pattern and the dimension is allowed to diverge with the sample size. Necessary and sufficient conditions are provided as well for the normal approximation of the likelihood ratio and Hellinger statistics for testing goodness of fit. This theory applies continuously throughout the low- and high-dimensional regimes. A cross-sample version of Pearson's test of independence is shown to be more robust than its classic version in two-way contingency tables with diverging dimensions. A degrees-of-freedom adjusted chi-squared approximation improves accuracy by matching Pearson's chi-squared statistic in both the mean and variance. Compared with traditional guidelines, this theory allows the sample size to be much smaller than the total number of cells in the contingency table, implying the dominance of empty cells, provided the presence of a diverging number of cells with more than a single count in any deterministic fractional subsets of cells. Specific examples are provided to demonstrate the asymptotic normality of test statistics when the classical regularity conditions for the chi-squared and normal approximations are violated. Simulation results support the theoretical findings and demonstrate that the chi-squared and normal approximations are more

robust for the likelihood ratio and Hellinger statistics, compared with Pearson's chi-squared statistics.

Keywords

Test of independence, Goodness of fit, Chi-squared approximation, Contingency tables.

The out-of-sample prediction error of the square-root lasso and related estimators

José Luis Montiel Olea[1], **Amilcar Velez**[2], **Cynthia Rush**[3], **Johannes Wiesel**[4]

[1] *Cornell University, Department of Economics, USA,*
montiel.olea@gmail.com
[2] *Northwestern University, Department of Economics, USA,*
amilcarvelezsalamanca2025@u.northwestern.edu
[3] *Columbia University, Department of Statistics, USA,*
cynthia.rush@columbia.edu
[4] *Carnegie Mellon University, Department of Mathematics, USA,*
jwiesel@andrew.cmu.edu

Abstract

We study the classical problem of predicting an outcome variable, Y, using a linear combination of a d-dimensional covariate vector, X. We are interested in linear predictors whose coefficients solve: $\inf_\beta(\mathbb{E}[(Y - \langle \beta, X\rangle)^r])^{1/r} + \delta\|\beta\|$, where $r > 1$ and $\delta > 0$ is a regularization parameter. We provide conditions under which linear predictors based on these estimators minimize the worst-case prediction error over a ball of distributions determined by a type of max-sliced Wasserstein metric. A detailed analysis of the statistical properties of this metric yields a simple recommendation for the choice of regularization parameter. The suggested order of δ, after a suitable normalization of the covariates, is typically d/n, up to logarithmic factors. Our recommendation is computationally straightforward to implement, pivotal, has provable out-of-sample performance guarantees, and does not rely on sparsity assumptions about the true data generating process.

Keywords

Wasserstein, square-root LASSO, distributionally robust optimization, generalization bounds.

Network Regression and Supervised Centrality Estimation

Junhui Cai[1], **Dan Yang**[2], **Wu Zhu**[3], **Haipeng Shen**[4], **Linda Zhao**[5]

[1] *University of Notre Dame, USA, jcai2@nd.edu*
[2] *The University of Hong Kong, Hong Kong, dyanghku@hku.hk*
[3] *Tsinghua University, China, zhuwu@sem.tsinghua.edu.cn*
[4] *The University of Hong Kong, Hong Kong, haipeng@hku.hk*
[5] *University of Pennsylvania, USA, lzhao@wharton.upenn.edu*

Abstract

The centrality in a network is often used to measure nodes' importance and model network effects on a certain outcome. Empirical studies widely adopt a two-stage procedure, which first estimates the centrality from the observed noisy network and then infers the network effect from the estimated centrality, even though it lacks theoretical understanding. We propose a unified modeling framework, under which we first prove the shortcomings of the two-stage procedure, including the inconsistency of the centrality estimation and the invalidity of the network effect inference. Furthermore, we propose a supervised centrality estimation methodology, which aims to simultaneously estimate both centrality and network effect. The advantages in both regards are proved theoretically and demonstrated numerically via extensive simulations and a case study in predicting currency risk premiums from the global trade network.

Keywords

Hub centrality, Authority centrality, Measurement error, Global trade network, Currency risk premium.

Likelihood-based finite mixture models for ordinal data

Daniel Fernández[1], **Richard Arnold**[2] **& Shirley Pledger**[2]

[1] *Universitat Politècnica de Catalunya – BarcelonaTech (UPC), Department of Statistics and Operations Research, Spain, daniel.fernandez.martinez@upc.edu*

[2] *Victoria University of Wellington, School of Mathematics and Statistics, New Zealand, richard.arnold@vuw.ac.nz; shirley.pledger@vuw.ac.nz*

Abstract

Many dimensionality reduction methods for data matrices rely on mathematical techniques, such as distance-based algorithms, matrix decomposition, and eigenvalues. However, these techniques lack an underlying probability model, making it challenging to employ statistical inferences or determine model appropriateness using information criteria. Ordinal data, such as Likert or pain scale measurements, is prevalent in various domains. To address the specific challenges posed by ordinal data, recent research (Fernández, D. *et al.,* 2016) has introduced likelihood-based finite mixture models tailored for such datasets. This innovative approach applies clustering through finite mixtures to the ordered stereotype model (Anderson, J.A., 1984). By utilizing expectation-maximization (EM) algorithms and Bayesian methods (Reversible-Jump MCMC sampler), it enables fuzzy allocation of rows, columns (one-dimensional clustering), and rows and columns simultaneously (two-dimensional clustering, also known as biclustering or block clustering) into corresponding clusters. To illustrate the practical application of this approach, we will provide examples using real-world ordinal data sets. Additionally, we will demonstrate various visualisation tools designed to depict the inherent fuzziness of the clustering results for ordinal data.

Keywords

Clustering, EM-algorithm, Finite mixture model, Ordinal data, Stereotype model.

References

Fernández, D., Arnold, R., and Pledger, S. (2016). Mixture-based clustering for the ordered stereotype model. *Computational Statistics & Data Analysis*, *93*, 46–75.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *46(1)*, 1–22.

Matrix-Variate Canonical Correlation Analysis

Daniel Kessler[1], **Elizaveta Levina**[2]

[1] *University of Washington, Department of Statistics, United States, dakess@uw.edu*
[2] *University of Michigan, Department of Statistics, United States, elevina@umich.edu*

Abstract

We consider the extension of Canonical Correlation Analysis (CCA) to the matrix-variate setting, where one or both of the random vectors of classical CCA is replaced by random matrices. The goal remains the identification of pairs of linear functions that transform the data into maximally correlated canonical variates. We exploit matrix-specific structure by seeking low-rank representations through the use of a nuclear norm penalty. Although generally applicable to matrix-variate data, this approach is motivated by applications in network neuroscience, where the matrix-variate data is a participant-specific connectivity matrix of spatial correlations. When applied to network data, these low-rank canonical directions can be understood as seeking latent network structure. We show in synthetic data that our approach is effective at recovering low rank signals even in noisy cases with relatively few observations, and we apply the method to human neuroimaging data.

Keywords

Covariance matrix, matrix computations, network data, penalization, neuroimaging.

Nonparametric shrinkage estimation in high dimensional generalized linear models via Polya trees

Asaf Weinstein[1], **Jonas Wallin**[2], **Daniel Yekutieli**[3], **Małgorzata Bogdan**[4]

[1] *Department of Statistics and Data Science, Hebrew University of Jerusalem, asaf.weinstein@mail.huji.ac.il*
[2] *Department of Statistics, Lund University, jonas.wallin@stat.lu.se*
[3] *Department of Statistics and Operations Research, Tel Aviv University, yekutiel@tauex.tau.ac.il*
[4] *Department of Mathematics, University of Wroclaw and Department of Statistics, Lund University, Malgorzata.Bogdan@math.uni.wroc.pl*

Abstract

In a given generalized linear model with fixed effects, and under a specified loss function, what is the optimal estimator of the coefficients? We propose as a contender an ideal (oracle) shrinkage estimator, specifically, the Bayes estimator under the particular prior that assigns equal mass to every permutation of the true coefficient vector. We first study this ideal shrinker, showing some optimality properties in both frequentist and Bayesian frameworks by extending notions from Robbins's compound decision theory. To compete with the ideal estimator, taking advantage of the fact that it depends on the true coefficients only through their *empirical distribution*, we postulate a hierarchical Bayes model, that can be viewed as a nonparametric counterpart of the usual Gaussian hierarchical model. More concretely, the individual coefficients are modeled as i.i.d. draws from a common distribution π, which is itself modeled as random and assigned a Polya tree prior to reflect indefiniteness. We show in simulations that the posterior mean of π approximates well the empirical distribution of

the true, *fixed* coefficients, effectively solving a nonparametric deconvolution problem. This allows the posterior estimates of the coefficient vector to learn the correct shrinkage pattern without parametric restrictions. We compare our method with popular parametric alternatives on the challenging task of gene mapping in the presence of polygenic effects. In this scenario, the regressors exhibit strong spatial correlation, and the signal consists of a dense polygenic component along with several prominent spikes. Our analysis demonstrates that, unlike standard high-dimensional methods such as ridge regression or Lasso, the proposed approach recovers the intricate signal structure, and results in better estimation and prediction accuracy in supporting simulations.

Keywords

Empirical Bayes, Shrinkage estimation, nonparametric Bayes.

Optimal minimax random designs for weighted least squares estimators

David Azriel[1]

[1] *The Technion, Faculty of Data and Decision Sciences, Israel, davidazr@technion.ac.il*

Abstract

Consider an experimental design problem where the values of a predictor variable, denoted by x, are to be determined with the goal of estimating a function $m(x)$, which is observed with noise. A linear model is fitted to $m(x)$ but it is not assumed that the model is correctly specified. It follows that the quantity of interest is the best linear approximation of $m(x)$, which is denoted by $l(x)$. It is shown that in this framework the ordinary least squares estimator typically leads to an inconsistent estimation of $l(x)$, and rather weighted least squares should be considered. An asymptotic minimax criterion is formulated for this estimator, and a design that minimizes the criterion is presented. The results are illustrated for polynomial regression models.

Keywords

Experimental design, I-optimality, Robust regression.

Synthetic Data in Chemistry: Deterministic, Evolutionary, and Generative

David Balcells[1]

[1] *University of Oslo, Department of Chemistry, Hylleraas Center of Excellence for Quantum Molecular Sciences, Norway, david.balcells@kjemi.uio.no*

Abstract

Chemistry plays a central role in the application of the physical sciences to many key technologies. Examples include drug discovery in the pharmaceutical industry and the development of catalysts enabling green processes based on renewable energies. The role played by data in these research fields is becoming increasingly important. In this talk, I will present the work of my group in the generation of synthetic data from three different perspectives: deterministic, evolutionary, and generative. In the deterministic approach, we formulate chemical spaces *ad hoc* with the aim of creating the datasets needed to optimize predictive machine learning models. When we instead need to generate molecules or materials that satisfy multiple objectives, we turn to the genetic algorithms that we recently developed to gain full control over the direction and scope of the optimization, while maximizing diversity. Finally, we recently started a new line of research in which we are developing a variational autoencoder tackling the difficulties associated with chemical compounds containing metal elements.

Keywords

Chemical spaces, genetic algorithms, multiobjective optimization, variational autoencoders, inverse design.

References

Friederich, P.; Gomes, G. d. P.; de Bin, R.; Aspuru-Guzik, A.; Balcells, D. (2020). Machine Learning Dihydrogen Activation in the Chemical Space Surrounding Vaska's Complex. *Chem. Sci. 11*, 4584–4601.

Kneiding, H.; Nova, A.; Balcells, D. (2023). Directional Multiobjective Optimization of Metal Complexes at the Billion-Scale with the tmQMg-L Dataset and PL-MOGA Algorithm. *ChemRxiv*, DOI: 10.26434/chemrxiv-2023-k3tf2-v2

Jin, W.; Barzilay, R.; Jaakkola, T. (2019). Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv*, DOI: 10.48550/arXiv.1802.04364

Detection and Estimation of Jumps, Bumps, and Kinks

D. Siegmund[1]

[1] *Stanford University, Department of Statistics, USA,*
siegmund@stanford.edu

Abstract

I will describe problems of segmentation of (usually normal) observations according to changes in their mean values, which can occur continuously: (i) a change in slope, (ii) a bump, or discontinuously: (iii) a jump in the level of the observations. Theoretical results will be illustrated by applications to copy number changes, historical weather records, and COVID-19: (a) daily incidence, (b) wastewater analysis, and (c) excess deaths. Particular attention will be paid to vector observations with (approximately) concurrent changes in several coordinates. I will also discuss online detection, confidence regions for the change-points, and difficulties associated with dependent observations. Aspects of this research involve collaboration with Fang Xiao, Li Jian, Liu Yi, Nancy Zhang, Benjamin Yakir, Keith Worsley, and Li (Charlie) Xia.

Keywords

Change-point, online detection, dependence.

Domain Adaptation meets Individual Fairness. And they get along

Debarghya Mukherjee
Jointly with Felix Petersen, Yuekai Sun and Mikhail Yurochkin

Abstract

Many instances of algorithmic bias are caused by distributional shifts. For example, machine learning (ML) models often perform worse on demographic groups that are underrepresented in the training data. In this paper, we leverage this connection between algorithmic fairness and distribution shifts to show that algorithmic fairness interventions can help ML models overcome distribution shifts, and that domain adaptation methods (for overcoming distribution shifts) can mitigate algorithmic biases. In particular, we show that (i) enforcing suitable notions of individual fairness can improve the out-of-distribution accuracy of ML models, and that (ii) it is possible to adapt representation alignment methods for domain adaptation to enforce (individual) fairness. The former is unexpected because individual fairness interventions were not developed with distribution shifts in mind. The latter is also unexpected because representation alignment is not a common approach in the individual fairness literature.

Matrix–free Conditional Simulation of Gaussian Random Fields

Debashis Mondal[1], **Somak Dutta**[2]

[1] *Washington University, Department of Statistics and Data Science, USA, mondal@wustl.edu*
[2] *Iowa State University, Department of Statistics, USA, somakd@iastate.edu*

Abstract

In recent years, interest in spatial statistics has increased significantly. However, for large data sets, statistical computations for spatial models have remained a challenge, as it is extremely difficult to store a large covariance or an inverse covariance matrix, and compute its inverse, determinant or Cholesky decomposition. In this talk, we shall focus on spatial mixed models and discuss a new algorithm for fast matrix-free conditional samplings for their inference. This new algorithm relies on 'rectangular' square roots of the inverse covariance matrices and covers a large class of spatial models including spatial models based on Gaussian conditional and intrinsic autoregressions, and fractional Gaussian fields. We shall show that the algorithm outperforms sparse Cholesky, and other existing conditional simulation methods. We demonstrate usefulness of this algorithm by analyzing groundwater arsenic contamination in Bangladesh, and by mapping simultaneous exceedance regions.

Keywords

BLUP, Conditional autoregression, De Wijs process, Fractional Gaussian fields, Lanczos algorithm, Preconditioning, Rectangular square root, REML.

Fighting Noise with Noise: Causal Inference with Many Candidate Instruments

Xinyi Zhang[1], **Linbo Wang**[1], **Stanislav Volgushev**[1], **Dehan Kong**[1]

[1] *University of Toronto, Department of Statistical Sciences, Canada*

Abstract

Instrumental variable methods provide useful tools for inferring causal effects in the presence of unmeasured confounding. To apply these methods with large-scale data sets, a major challenge is to find valid instruments from a possibly large candidate set. In practice, most of the candidate instruments are often not relevant for studying a particular exposure of interest. Moreover, not all relevant candidate instruments are valid as they may directly influence the outcome of interest. In this article, we propose a data-driven method for causal inference with many candidate instruments that addresses these two challenges simultaneously. A key component of our proposal is a novel resampling method, which constructs pseudo variables to remove irrelevant candidate instruments having spurious correlations with the exposure. Synthetic data analyses show that the proposed method performs favourably compared to existing methods. We apply our method to a Mendelian randomization study estimating the effect of obesity on health-related quality of life.

Keywords

A/B tests; Mendelian randomization; Selection bias; Spurious correlation.

Transport ABC: improving the efficiency of ABC SMC using normalizing flows

D Prangle[1], **S Ragy**[1], **C Viscardi**[2]

[1] *University of Bristol, School of Mathematics, United Kingdom,*
dennis.prangle@bristol.ac.uk
[2] *University of Florence, DiSIA, Italy*

Abstract

We describe an efficient method to update particles in an ABC-SMC algorithm (Approximate Bayesian Computation within the Sequential Monte Carlo framework). The main contribution is to learn a parameter proposal distribution via a transport map, implemented using normalizing flows. 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. A secondary contribution is an efficient variation of the "r-hit" ABC-MCMC kernel which can be used when parameter proposals are independent of the current state. We also present examples where the method provides efficiency gains over standard ABC-SMC.

Keywords

ABC, flows, SMC, MCMC.

Transfer learning under random distribution shifts

Yujin Jeong[1], **Dominik Rothenhäusler**[2]

[1] *Department of Statistics, Stanford, USA, yujinj@stanford.edu*
[2] *Department of Statistics, Stanford, USA, rdominik@stanford.edu*

Abstract

We consider estimation in the setting of transfer learning where, in addition to partial observations from the target distribution, auxiliary observations from different but related distributions are available. If the auxiliary distributions are similar to the target distribution, it is sensible to estimate the parameter on the pooled data. If the auxiliary distributions are very different from each other, one would want to only use observations from the distribution that is closest to the target distribution. We introduce a model for random distribution shifts that interpolates between these two settings and propose a procedure that aggregates the heterogeneous data sources in an optimal fashion. Perhaps surprisingly, the proposed procedure is closely related to synthetic controls in causal inference. More broadly speaking, the framework provides a new language for distributional shifts and brings clarity regarding how we should infer parameters, predict outcomes, perform model selection, and quantify uncertainty under randomly shifted distributions.

Keywords

Distribution shift, robustness, transfer learning, heterogeneous data.

Surrogate method for partial association between mixed data with application to well-being survey analysis

Shaobo Li[1], **Zhaohu Fan**[2], **Ivy Liu**[3], **Philip S. Morrison**[4], **Dungang Liu**[5]

[1] *University of Kansas, School of Business, USA, shaobo.li@ku.edu*
[2] *Georgia Institute of Technology, Scheller College of Business, USA, jonathan.fan@scheller.gatech.edu*
[3] *Victoria University of Wellington, School of Mathematics and Statistics, New Zealand, ivy.liu@vuw.ac.nz*
[4] *Victoria University of Wellington,, School of Geography, Environment and Earth Sciences, New Zealand, philip.morrison@vuw.ac.nz*
[5] *University of Cincinnati, Lindner College of Business, USA, dungang.liu@uc.edu*

Abstract

This paper is motivated by the analysis of a survey study focusing on college student well-being before and after the COVID-19 pandemic outbreak. A statistical challenge in well-being studies lies in the multidimensionality of outcome variables, recorded in various scales such as continuous, binary, or ordinal. The presence of mixed data complicates the examination of their relationships when adjusting for important covariates. To address this challenge, we propose a unifying framework for studying partial association between mixed data. We achieve this by defining a unified residual using the surrogate method. The idea is to map the residual randomness to a consistent continuous scale, regardless of the original scales of outcome variables. This framework applies to parametric or semiparametric models for covariate adjustments. We validate the use of such residuals for assessing partial association, introducing a measure that generalizes classical

Kendall's tau to capture both partial and marginal associations. Moreover, our development advances the theory of the surrogate method by demonstrating its applicability without requiring outcome variables to have a latent variable structure. In the analysis of the college student well-being survey, our proposed method unveils the contingency of relationships between multidimensional well-being measures and micro personal risk factors (e.g., physical health, loneliness, and accommodation), as well as the macro disruption caused by COVID-19.

Key words:

Covariate adjustment, COVID-19 pandemic, Kendall's tau, mental health, moderation effect, partial correlation, surrogate residual.

Joint Dynamic Models and Statistical Inference for Recurrent Competing Risks, Longitudinal Marker, and Health Status

Lili Tong[1], **Piaomu Liu**[2], **Edsel A. Peña**[3]

[1] *Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE 68198*
[2] *Department of Mathematical Sciences, Bentley University, Waltham, MA 02452, PLIU@bentley.edu*
[3] *Department of Statistics, University of South Carolina, Columbia, SC 21208, pena@stat.sc.edu*

Abstract

Consider a subject or unit in a longitudinal biomedical, public health, engineering, economic or social science study which is being monitored over a possibly random duration. Over time this unit experiences recurrent events of several types and a longitudinal marker transitions over a discrete state-space. In addition, its "health" status also transitions over a discrete state-space with at least one absorbing state. A vector of covariates will also be associated with this unit. Of major interest for this unit is the time-to-absorption of its health status process, which could be viewed as the unit's lifetime. Aside from being affected by its covariate vector, there could be associations among the recurrent competing risks processes, the longitudinal marker process, and the health status process in the sense that the time-evolution of each process is associated with the other processes. A joint dynamic stochastic model for these components is proposed and statistical inference methods are developed. This joint model, formulated via counting processes and continuous-time Markov chains, has the potential of facilitating 'personalized' interventions. This could enhance, for example, the implementation and adoption of precision medicine in medical settings. Semi-parametric and likelihood-based inferential methods for the model parameters are developed when a sample of these units is

available. Finite-sample and asymptotic properties of estimators of the finite- and infinite-dimensional model parameters will be presented in this talk.

Keywords

Continuous-time Markov chain; Counting processes; Dynamic models; Intensity-based model; Personalized medicine; Semi-parametric estimation.

Hippocampus shape analysis via skeletal models and kernel smoothing

Eduardo García-Portugués[1], **Andrea Meilán-Vila**[1]

[1] *Universidad Carlos III de Madrid, Department of Statistics, Spain, edgarcia@est-econ.uc3m.es*
[2] *Universidad Carlos III de Madrid, Department of Statistics, Spain, ameilan@est-econ.uc3m.es*

Abstract

Skeletal representations (s-reps) have been successfully adopted to parsimoniously parametrize the shape of three-dimensional objects, and have been particularly employed in analyzing hippocampus shape variation. Within this context, we provide a fully-nonparametric dimension-reduction tool based on a new kernel density estimator for determining the main source of variability of hippocampus shapes parametrized by s-reps. The methodology introduces density ridges for data on the polysphere $(S^d)^r$ and involves addressing high-dimensional computational challenges. Asymptotic properties of the novel kernel density estimator will also be discussed, as well as a k-sample test stemming from it.

Keywords

Directional data, kernel density estimation, skeletal representations.

Fusing Sufficient Dimension Reduction with Neural Networks

Efstathia Bura[1], **Daniel Kapla**[2], **Lukas Fertl**[3]

[1] *Institute of Statistics and Mathematical Methods in Economics, Faculty of Mathematics and Geoinformation, TU Wien, Austria, efstathia.bura@tuwien.ac.at*

[2] *Institute of Statistics and Mathematical Methods in Economics, Faculty of Mathematics and Geoinformation, TU Wien, Austria, daniel.kapla@tuwien.ac.at*

[3] ***d-fine*** *Austria GmbH, Lukas.Fertl@d-fine.at*

Abstract

Neural networks are combined with sufficient dimension reduction methodology in order to remove the limitation of small p and n of the latter. NN-SDR applies when the dependence of the response Y on a set of predictors $\boldsymbol{X}$ is fully captured by the regression function $g(\boldsymbol{B}'\boldsymbol{X})$, for an unknown function g and low rank parameter $\boldsymbol{B}$ matrix. It is shown that the proposed estimator is on par with competing sufficient dimension reduction methods, such as *minimum average variance estimation* and *conditional variance estimation*, in small p and n settings in simulations. Its main advantage is its scalability in regressions with large data, for which the other methods are infeasible.

Keywords

Large p and n, Regression, Mean subspace, Nonparametric, Prediction.

Efficiency loss with binary pre-processing of continuous monitoring data

Elizabeth Juarez-Colunga[1], **Paula Langner**[2], **John Rice**[3], **Gary Grunwald**[4]

[1] *University of Colorado Anschutz Medical Campus, Department of Biostatistics and Informatics, USA, elizabeth.juarez-colunga@cuanschutz.edu*

[2] *US Department of Veterans Affairs, Rocky Mountain Regional VA Medical Center, USA, Paula.Langner@va.gov*

[3] *University of Michigan, Department of Biostatistics, USA, jdrice@umich.edu*

[4] *University of Colorado Anschutz Medical Campus, Department of Biostatistics and Informatics, USA, gary.grunwald@cuanschutz.edu*

Abstract

In studies with a recurrent event outcome, events may be captured as counts during subsequent intervals or follow-up times either by design or for ease of analysis. In many cases, recurrent events may be further coarsened such that only an indicator of one or more events in an interval is observed at the follow-up time, resulting in a loss of information relative to a record of all events. In this paper, we examine efficiency loss when coarsening longitudinally observed counts to binary indicators and aspects of the design which impact the ability to estimate a treatment effect of interest. The investigation is motivated by a study of patients with Cardiac implantable electronic devices in which investigators aimed to examine the effect of a treatment on events detected by the devices over time. In order to study components of such a recurrent event process impacted by data coarsening, we derive the asymptotic relative efficiency of a treatment effect estimator utilizing a coarsened binary outcome relative to the count outcome, which represents a longitudinal recurrent event process. We compare the efficiencies and consider conditions where the binary process maintains good efficiency in estimating a treatment effect.

Keywords

Longitudinal, Poisson process, Counting process, Panel count.

Missing data with causal and statistical dependence

Elizabeth L. Ogburn[1]

[1] *Johns Hopkins University, Biostatistics, USA, eogburn@jhsph.edu*

Abstract

In this talk I will describe two recent projects on causal inference in the presence of missing data. In one project (with Brian Gilbert and Abhirup Datta) we harness spatial dependence to help create proxies for missing confounders; in this setting the presence of statistical dependence is assumed to lend structure to the unmeasured confounder and this structure facilitates the identification of causal effects. In the other project (with Ranjani Srinivasan, Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser) we identify a new kind of missing data process in which missingness indicators can exhibit causal dependence across units; this kind of dependence undermines existing identification results for missing data and requires new graphical model based results.

Keywords

Missing data, causal inference, dependence, spatial statistics, interference.

Learning extreme Expected Shortfall with neural networks. Application to cryptocurrency data

Michaël Allouche[1], **Stéphane Girard**[2], **Emmanuel Gobet**[3]

[1] *Kaiko - Quantitative Data. 2 rue de Choiseul 75002 Paris, France, michael.allouche@kaiko.com*

[2] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, stephane.girard@inria.fr*

[3] *Centre de Mathématiques Appliquées (CMAP), CNRS, Ecole Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau Cedex, France, emmanuel.gobet@polytechnique.edu*

Abstract

We propose new parametrizations for neural networks in order to estimate extreme Expected Shortfall in heavy-tailed settings as a function of confidence levels. The proposed neural network estimator is able to extrapolate in the distribution tails thanks to an extension of the usual extreme-value second-order condition to an arbitrary order. The convergence rate of the uniform error between the extreme log-Expected Shortfall and its neural network approximation is established. The finite sample performance of the neural network estimator is compared to bias-reduced extreme-value competitors on simulated data. It is shown that our method outperforms them in difficult heavy-tailed situations where other estimators almost all fail. Finally, the neural network estimator is tested on real data to investigate the behavior of cryptocurrency extreme loss returns.

Keywords

Extreme-value theory, heavy-tailed distribution, risk measure estimation, neural networks.

A bridge between PLS and GME estimators in the SEM framework

Enrico Ciavolino[1,2,*], **Mario Angelelli**[1,2],

[1] *University of Salento, Department of Human and Social Sciences, Italy*
[2] *CAMPI - Centre for Applied Mathematics and Physics for Industry, Lecce, Italy*
[*] *enrico.ciavolino@unisalento.it*

Abstract

Structural Equation Modeling (SEM) serves as a versatile analytical framework for evaluating intricate relationships between unobservable latent variables and their measurable indicators. Within the realm of SEM, two prevalent estimation methodologies are regularly utilized. These include a parametric approach, which adopts the Maximum Likelihood Estimation (MLE) method, and a non-parametric alternative, known as Partial Least Squares (PLS). In recent years, there has been a surge of interest in innovative estimation techniques in SEM. These advancements include Generalized Structured Component Analysis (GSCA), offering a structured approach to model evaluation, and Generalized Maximum Entropy (GME), which presents a semi-parametric approach. This study explores the potential synergy between these non-parametric and semi-parametric approaches, seeking to establish a bridge that connects PLS and GME estimators within the SEM framework. The contribution provides an in-depth overview of PLS and GME, highlighting their respective advantages and limitations. PLS, known for its emphasis on predictive power and its ability to handle reflective and formative measurement models, is contrasted with GME, which excels in addressing multi-collinearity, endogeneity, and latent variable reliability while accommodating complex structural relationships. We present a novel integration approach that merges the strengths of PLS and GME. The methodology for implementing this bridge is detailed, providing guidance to researchers

seeking to combine these estimators effectively. To illustrate the practical utility of our integrated approach, we provide a simulation study and empirical example, showcasing the benefits of this bridge in real-world applications.

Keywords

Structural Equation Modeling, Generalized Maximum Entropy, Partial Least Squares.

Optimal treatment regimes under partially ordered surrogates

Eric Laber[1], **Yinyihong Liu**[1], **Marc Brooks**[2]

[1] *Duke University, Statistical Science, USA, eric.laber@duke.edu*
[2] *University of Michigan, Statistics, USA,*

Abstract

Behavioral interventions delivered by mobile-health (mHealth) often seek to affect long-term outcomes such as body mass index (BMI), prolonged substance-use cessation, resting heart rate, and VO2 max. Often, the effect of any single behavioral intervention on the long-term outcome of interest is imperceptible relative to momentary fluctuations. Thus, adaptive algorithms for mHealth typically work through multiple surrogate outcomes; e.g., an mHealth intervention for BMI reduction in overweight individuals might target surrogates such as step count, calorie intake, or mindfulness. While the relationship between surrogates and the outcome of interest is typically unknown, one often has access to partial directional information; e.g., in the context of BMI reduction, more steps are better than fewer steps, and more app engagement is better than less, but it is not clear if taking five-hundred additional steps is better than spending five additional minutes engaged with the app. We consider optimal sequential decision masking with surrogates that admit a partial ordering. We show that a partial ordering on the surrogates corresponds to a stochastic partial order over the space of treatment regimes from which we derive an estimator of the set of maximal treatment regimes. Empirical experiments show that bandit algorithms that make use of partial ordering by restricting to maximal treatment regimes improve patient outcomes, especially when momentary treatment effects are small. We provide an illustrative example based on the ADAPT mHealth study for BMI reduction in overweight subjects with type 1 diabetes (T1D).

Keywords

Precision medicine, isotonic regression, partial orderings

Differentially Private Linear Regression with Linked Data

Shurong Lin[1], **Elliot Paquette**[2], **Eric D. Kolaczyk**[2]

[1] *Boston University, Department of Mathematics & Statistics, USA, shrlin@bu.edu*

[2] *McGill University, Department of Mathematics & Statistics, Canada, {elliot.paquette,eric.kolaczyk}@mcgill.ca*

Abstract

There has been increasing demand for establishing privacy-preserving methodologies for modern statistics and machine learning. Differential privacy, a mathematical notion from computer science, is a rising tool offering robust privacy guarantees. Recent work focuses primarily on developing differentially private versions of individual statistical and machine learning tasks, with nontrivial upstream pre-processing typically not incorporated. An important example is when record linkage is done prior to downstream modeling. Record linkage refers to the statistical task of linking two or more datasets of the same group of entities without a unique identifier. This probabilistic procedure brings additional uncertainty to the subsequent task. In this paper, we present two differentially private algorithms for linear regression with linked data. In particular, we propose a noisy gradient method and a sufficient statistics perturbation approach for the estimation of regression coefficients. We investigate the privacy-accuracy tradeoff by providing finite-sample error bounds for the estimators, which allows us to understand the relative contributions of linkage error, estimation error, and the cost of privacy. The variances of the estimators are also discussed. We demonstrate the performance of the proposed algorithms through simulations and an application to synthetic data.

Keywords

Differential privacy, record linkage, data integration, privacy-preserving record linkage, gradient descent.

Naive imputation implicitly regularizes high-dimensional linear models

Alexis Ayme[1], **Claire Boyer**[1], **Aymeric Dieuleveut**[2], **Erwan Scornet**[1]

[1] *Sorbonne University, LPSM, France,*
erwan.scornet@sorbonne-universite.fr
[2] *École polytechnique, CMAP, France*

Abstract

Two different approaches exist to handle missing values for prediction: either imputation, prior to fitting any predictive algorithms, or dedicated methods able to natively incorporate missing values. While imputation is widely (and easily) use, it is unfortunately biased when low-capacity predictors (such as linear models) are applied afterwards. However, in practice, naive imputation exhibits good predictive performance. In this paper, we study the impact of imputation in a high-dimensional linear model with MCAR missing data. We prove that zero imputation performs an implicit regularization closely related to the ridge method, often used in high-dimensional problems. Leveraging on this connection, we establish that the imputation bias is controlled by a ridge bias, which vanishes in high dimension. As a predictor, we argue in favor of the averaged SGD strategy, applied to zero-imputed data. We establish an upper bound on its generalization error, highlighting that imputation is benign in the $d \gg \sqrt{n}$ regime. Experiments illustrate our findings.

Keywords

Missing values, linear models, high dimension, imputation, implicit regularization.

Reconciling model-X and doubly robust approaches to conditional independence testing

Ziang Niu[1], **Abhinav Chakraborty**[2], **Oliver Dukes**[3], **Eugene Katsevich**[4]

[1] *University of Pennsylvania, Department of Statistics and Data Science, USA, ziangniu@wharton.upenn.edu*
[2] *University of Pennsylvania, Department of Statistics and Data Science, USA, abch@wharton.upenn.edu*
[3] *Ghent University, Department of Applied Mathematics, Computer Science and Statistics, Belgium, oliver.dukes@ugent.be*
[4] *University of Pennsylvania, Department of Statistics and Data Science, USA, ekatsevi@wharton.upenn.edu*

Abstract

Model-X approaches to testing conditional independence between a predictor and an outcome variable given a vector of covariates usually assume exact knowledge of the conditional distribution of the predictor given the covariates. Nevertheless, model-X methodologies are often deployed with this conditional distribution learned in sample. We investigate the consequences of this choice through the lens of the distilled conditional randomization test (dCRT). We find that Type-I error control is still possible, but only if the mean of the outcome variable given the covariates is estimated well enough. This demonstrates that the dCRT is doubly robust, and motivates a comparison to the generalized covariance measure (GCM) test, another doubly robust conditional independence test. We prove that these two tests are asymptotically equivalent, and show that the GCM test is optimal against (generalized) partially linear alternatives by leveraging semiparametric efficiency theory. In an extensive simulation study, we compare the dCRT to the GCM test. These two tests have broadly

similar Type-I error and power, though dCRT can have somewhat better Type-I error control but somewhat worse power in small samples or when the response is discrete. We also find that post-lasso based test statistics (as compared to lasso based statistics) can dramatically improve Type-I error control for both methods.

Keywords

Model-X, conditional independence, conditional randomization test, generalized covariance measure, doubly robust.

Is Motivation to Change Alcohol Use a State or a Trait? An Investigation of Mobile Health Investigation

Eun-Young Mun[1], **Feng Geng**[2], **Michael S. Businelle**[3], **Scott T. Walters**[4]

[1] *The University of North Texas Health Science Center, School of Public Health, USA, eun-young.mun@unthsc.edu*
[2] *Fort Worth, Texas, USA, fgeng.tx@gmail.com*
[3] *University of Oklahoma Health Sciences Center, Stephenson Cancer Center, USA, Michael-Businelle@ouhsc.edu*
[4] *The University of North Texas Health Science Center, School of Public Health, USA, Scott.Walters@unthsc.edu*

Abstract

Prevailing evidence-based intervention and treatment strategies for alcohol misuse and alcohol use disorders attempt to motivate clients to reduce resistance and engage in change talks by utilizing Motivational Interviewing (MI) consistent language during in-person counseling sessions or in tech-adapted interventions across platforms. For example, Lyssn, an artificial intelligence (AI) platform start-up to assess and improve fidelity to evidence-based practices, is built on the premise that spoken language AI can be trained to understand empathy and can be used to provide real-time feedback on therapeutic conversations between therapists and clients. However, the available evidence of whether alcohol interventions change motivation is limited (see Tan et al., 2023) for important theoretical and study design reasons (see also the commentaries by Magill, 2023 and Richards, 2023). The current presentation reviews available evidence from clinical trials and presents data from an mHealth intervention trial for adults experiencing homelessness with alcohol misuse (Businelle et al., 2020; Mun et al., 2021; Walters et al., 2022). Participants wore a wearable alcohol sensor to detect alcohol use in real time. They also responded

to an ecological momentary assessment (EMA) of drinking goals (i.e., no goal, stay sober, drink less) and self-reported alcohol consumption several times a day for one month. Following each EMA response, participants received MI consistent intervention messages. We investigated whether and how drinking goals changed within and across days depending on participants' alcohol consumption and whether MI consistent intervention messages changed participants' drinking goals over time. Supported by NIH grants R01 AA019511, K02 AA028630, and R34 AA024584.

Keywords

Alcohol Interventions, Motivational Interviewing, Wearable Alcohol Sensor, Tech-adapted Interventions, mHealth.

Detecting when the available data does not allow reliable inference

Fanny Yang[1], **Alexandru Tifrea**[1], **Eric Staravache**[1], **Piersilvio de Bartolomeis**[1], **Javier Abad Martinez**[1]

[1] *ETH Zurich, Department of Computer Science, Switzerland*

Abstract

In this talk, we consider two settings where the goal is to decide whether inference can be made reliably. First, we discuss novel class detection in machine learning, where the aim is to flag samples from classes that were not in the training set. A learned model should refrain from prediction on these flagged examples and instead, forward them to human experts for inspection. For our second use case in causal inference, we propose a test that can detect when the amount of unobserved confounding in observational datasets is too strong. If that is the case, the observational study should be adjusted by experts, e.g. by including additional variables, for more reliable treatment effect estimation. Our test leverages randomized control trials that are for example available in post-marketing surveillance, and uses techniques from sensitivity analysis.

Keywords

Out-of-distribution detection, causal inference, sensitivity analysis.

A Low-Rank Perspective on Structured Output Prediction

Luc Brogat-Motte[1&2], Tamim El Ahmad[1], Pierre Laforgue[3], Junjie Yang[1] & Florence d'Alché-Buc[1]

[1] *Télécom Paris, IP Paris, LTCI, France,*
[2] *INRIA, Sierra, France,*
[3] *Università degli Studi di Milano, Italy*
florence.dalche@telecom-paris.fr

Abstract

Surrogate regression methods offer a powerful and flexible solution to structured output prediction by embedding the output variable in a Hilbert space. In this talk we focus on one the simplest and oldest surrogate approach that leverages kernels in the input space as well as in the output space. While enjoying strong statistical guarantees, these surrogate kernel methods require important computations, for training as well as for inference. We propose to re-visit them by applying low-rank projections in the input and output feature spaces to reduce their complexity. Low-rank projection operators based on spectral decomposition as well as sketching are presented and the statistical properties of the resulting novel estimators are studied in terms of excess risk bounds. From a computational perspective, we show that the two approximations have distinct but complementary impacts: low-rank approximation in the input space reduces training time, while in the output space it decreases the inference time. In conclusion, we identify other surrogate models including neural networks where this approach might be relevant as well.

Keywords

Structured Output Prediction, Surrogate approaches, Kernels, Low-Rank approximation, Sketching.

Uncovering Regions of Maximum Dissimilarity on Random Process Data

Miguel de Carvalho[1], **Gabriel Martos**[2]

[1] *University of Edinburgh, School of Mathematics, UK, miguel.decarvalho@ed.ac.uk*
[2] *Universidad Torcuato Di Tella, Departamento de Matemática y Estadística, Argentina, gmartos@utdt.edu*

Abstract

The comparison of local characteristics of two random processes can shed light on periods of time or space at which the processes differ the most. This paper proposes a method that learns about regions with a certain volume, where the marginal attributes of two processes are less similar. The proposed methods are devised in full generality for the setting where the data of interest are themselves stochastic processes, and thus the proposed method can be used for pointing out the regions of maximum dissimilarity with a certain volume, in the contexts of functional data, time series, and point processes. The parameter functions underlying both stochastic processes of interest are modeled via a basis representation, and Bayesian inference is conducted via an integrated nested Laplace approximation. The numerical studies validate the proposed methods, and we showcase their application with case studies on criminology, finance, and medicine.

Keywords

Functional Parameters, Multi-objective Optimization, Pairs of Random Processes, Kolmogorov metric, Set Function Optimization.

Holdout Predictive Checks for Bayesian Model Criticism

Gemma E. Moran[1], **David M. Blei**[2], **Rajesh Ranganath**[3]

[1] *Rutgers University, Department of Statistics, USA, gm845@stat.rutgers.edu*
[2] *Columbia University, Department of Statistics, Department of Computer Science, USA*
[3] *New York University, Center for Data Science, USA*

Abstract

Bayesian modeling helps applied researchers articulate assumptions about their data and develop models tailored for specific applications. Thanks to good methods for approximate posterior inference, researchers can now easily build, use, and revise complicated Bayesian models for large and rich data. These capabilities, however, bring into focus the problem of model criticism. Researchers need tools to diagnose the fitness of their models, to understand where they fall short, and to guide their revision. In this paper we develop a new method for Bayesian model criticism, the Holdout Predictive Check (HPC). HPCs are built on Posterior Predictive Checks (PPCs), a seminal method that checks a model by assessing the posterior predictive distribution on the observed data. However, PPCs use the data twice—both to calculate the posterior predictive and to evaluate it—which can lead to uncalibrated p-values. HPCs, in contrast, compare the posterior predictive distribution to a draw from the population distribution, a heldout dataset. This method blends Bayesian modeling with frequentist assessment. Unlike the PPC, we prove that the HPC is properly calibrated. Empirically, we study HPC on classical regression, a hierarchical model of text data, and factor analysis.

Keywords

Bayesian model checking, predictive checks.

Leave-One-Out Confidence Intervals for Feature Importance: A Fast and Powerful Approach Using Minipatch Ensembles

Luqin Gan[*1], **Lili Zheng**[*2], **Genevera I. Allen**[3]

[1] *Department of Statistics, Rice University, USA, (glq.gan@gmail.com)*
[2] *Department of Electrical and Computer Engineering, Rice University, USA, (lz67@rice.edu)*
[3] *Departments of Electrical and Comptuer Engineering, Statistics, and Computer Science, Rice University, USA, (gallen@rice.edu)*
[*] *Denotes equal contribution.*

Abstract

Feature importance inference has been a long-standing statistical problem that helps promote scientific discoveries. Instead of testing for parameters that are only interpretable for specific models, there has been increasing interest in model-agnostic methods, often in the form of feature occlusion or leave-one-covariate-out (LOCO) inference. However, existing approaches often make limiting distributional assumptions or require model refitting and data splitting. Instead, we leverage minipatch ensemble learning, which creates an ensemble based on random subsamples of observations and features, to develop a novel LOCO inference procedure that is computationally efficient and statistically powerful. Despite the dependencies induced by using minipatch ensembles, we show that our approach provides valid asymptotic coverage for the feature importance score of any regression or classification model under mild assumptions. Finally, our same procedure can also be leveraged to provide valid predictive intervals, hence providing fast, simultaneous uncertainty quantification of both predictions and feature importance. We validate our approach on a series of synthetic and real data examples, demonstrating its computational and statistical advantages over existing methods.

Keywords

Feature Importance, Conformal Inference, Model-Agnostic Inference, Leave-One-Covariate-Out Inference, Minipatch Learning.

Covariate adjustment in rare diseases

Georg Zimmermann[1], **Konstantin Emil Thiel**[1], **Arne C. Bathke**[2]

[1] *Team Biostatistics and Big Medical Data, IDA Lab Salzburg, Paracelsus Medical University, Salzburg, Austria,* georg.zimmermann@pmu.ac.at

[2] *IDA Lab Salzburg, Department of Artificial Intelligence and Human Interfaces, Salzburg, Austria,* arne.bathke@plus.ac.at

Abstract

In clinical trials that are targeted at evaluating the comparative efficacy of treatments for patients with a rare disease, sample sizes are usually quite small, due to the fact that the disease is rare, and since there are several further challenges with respect to recruitment (e.g., high trial burden for patients). Therefore, statistical methods which allow for making effective use of the limited data are much needed. Generally speaking, as frequently recommended also in methodological guidance documents, adjusting the treatment effects for the impact of key covariates (e.g., baseline measurements of the outcome) might lead to an increase in statistical power. More specifically, in particular semi- and nonparametric approaches might be promising alternatives compared to the more restrictive, classical parametric models (e.g., parametric ANCOVA). Combining the former with resampling techniques may yield statistical inference methods which maintain the prespecified levels and have a good performance in terms of power even in small samples. In my talk, I will present some of these methods, including a discussion of the underlying theory as well as their appropriateness for practical applications, especially in research on rare diseases.

This is joint work with Konstantin Emil Thiel and Arne C. Bathke.

Keywords

Nonparametric statistics, resampling, ANCOVA, ordinal data.

Survival Kernets: Scalable and Interpretable Deep Kernel Survival Analysis with an Accuracy Guarantee

George H. Chen[1]

[1] *Carnegie Mellon University, Heinz College of Information Systems and Public Policy, USA,* georgechen@cmu.edu

Abstract

Kernel survival analysis models estimate individual survival distributions with the help of a kernel function, which measures the similarity between any two data points. Such a kernel function can be learned using deep kernel survival models. In this paper, we present a new deep kernel survival model called a *survival kernet*, which scales to large datasets in a manner that is amenable to model interpretation and also theoretical analysis. Specifically, the training data are partitioned into clusters based on a recently developed training set compression scheme for classification and regression called *kernel netting* that we extend to the survival analysis setting. At test time, each data point is represented as a weighted combination of these clusters, and each such cluster can be visualized. For a special case of survival kernets, we establish a finite-sample error bound on predicted survival distributions that is, up to a log factor, optimal. Whereas scalability at test time is achieved using the aforementioned kernel netting compression strategy, scalability during training is achieved by a warm-start procedure based on tree ensembles such as XGBoost and a heuristic approach to accelerating neural architecture search. On three standard survival analysis datasets of varying sizes (up to roughly 3 million data points), we show that survival kernets are highly competitive with the best of baselines tested in terms of concordance index. Our code is available at: https://github.com/georgehc/survival-kernets

Keywords

Survival analysis, kernel methods, neural networks, scalability, interpretability.

Multiple change Point Detection in High Dimensional Low Rank Models

George Michailidis[1]

[1] *UCLA, Department of Statistics and Data Science, USA,*
gmichail@ucla.edu

Abstract

We study the problem of detecting and locating change points in high-dimensional models with low rank structure, such as networks exhibiting community structure, or tensor data. We develop a simple two step algorithm for the problem at hand and establish performance guarantees in the form of finite sample bounds for the accuracy of the estimated locations of the change points and the underlying model parameters. We illustrate the detection strategy on data from three different domains that employ different statistical models: macroeconomics, neuroimaging and political science.

Keywords

Low Rank Models, Change Point Analysis, Fast Algorithms.

A Generic Approach for Reproducible Model Distillation

Giles Hooker[1], **Yunzhe Zhou**[2], **Peiru Xu**[3]

[1] *University of Pennsylvania, Department of Statistics and Data Science, USA, ghooker@wharton.upenn.edu*
[2] *University of California, Berkeley, Department of Biostatistics, USA, ztzyz615@berkeley.edu*
[3] *University of California, Berkeley, Department of Statistics, USA,xpr2019@berkeley.edu*

Abstract

Model distillation has been a popular method for producing interpretable machine learning. It uses an interpretable "student" model to mimic the predictions made by the black box "teacher" model. However, when the student model is sensitive to the variability of the data sets used for training even when keeping the teacher fixed, the corresponded interpretation is not reliable. Existing strategies stabilize model distillation by checking whether a large enough corpus of pseudo-data is generated to reliably reproduce student models, but methods to do so have so far been developed for a specific student model. In this paper, we develop a generic approach for stable model distillation based on central limit theorem for the average loss. We start with a collection of candidate student models and search for candidates that reasonably agree with the teacher. Then we construct a multiple testing framework to select a corpus size such that the consistent student model would be selected under different pseudo samples. We demonstrate the application of our proposed approach on three commonly used intelligible models: decision trees, falling rule lists and symbolic regression. Finally, we conduct simulation experiments on Mammographic Mass and Breast Cancer datasets and illustrate the testing procedure throughout a theoretical analysis with Markov process.

Keywords

Interpretability, Model Distillation, Multiple Testing, Reproducibility, Explainable Artificial Intelligence.

References

Shafer, G. (2023). Statistical testing with optional continuation. Working Paper 63 at probabilityandfinance.com.

Shafer, G., Vovk, V. (2019). *Game-Theoretic Foundations for Probability and Finance, Wiley.*

Extrapolation Trees for domain generalization

Gloria Buriticá[1], **Sebastian Engelke**[2]

[1] *Research Center for Statistics, University of Geneva, Switzerland, gloria.buriticaborda@unige.ch*
[2] *Research Center for Statistics, University of Geneva, Switzerland, sebastian.engelke@unige.ch*

Abstract

In numerous climate science applications, the goal is to predict the future impact of covariates on a climate variable. Even though the relationship among drivers typically follows invariant physical laws, the long-term climate is sensitive to distribution changes like mean or variance shifts. In this case, counting on regression strategies adapted to the entire covariates' domain is essential. A precise domain generalisation in regression is also crucial for short-term predictions; for example, if the environmental covariates reach unprecedented levels, then accurate predictions under extreme climate conditions are necessary to mitigate the hazards. Standard machine learning methods are popular among regression strategies because they impose minor assumptions on the training model; however, they are only reliable if the test points lie inside the range of the training data. Leveraging extreme value theory, we propose a tree-based algorithm for the conditional median function that enhances prediction at test points from unfrequented regions of the training distribution. We apply our extrapolation tree-based algorithm to simulated and real data, and show that compared to classical machine learning methods, it significantly improves the prediction error on extremal regions of the predictor space.

Keywords

Extreme value theory, conditional extreme value model, quantile regression, random forest, covariate shift, environmental applications.

On model-based clustering with entropic optimal transport

Gonzalo Mena[1]

[1] *Carnegie Mellon University, Department of Statistics and Data Science, Pittsburgh, PA, USA, gmena@andrew.cmu.edu*

Abstract I develop new methodology for model-based clustering. The standard approach, based on the optimization of the likelihood, provides a principled statistical framework for clustering where solutions are found via the the EM algorithm. However, as the log-likelihood is nonconvex, convergence to only local optima can be guaranteed, and practitioners rely on the use of several starting points with the hope that one of them will converge to the global solution. I consider a new loss based on entropic optimal transport that shares the same global optimum as the log-likelihood but has a much better behaved landscape so that spurious local optima configurations that are known to be pervasive for the log-likelihood are avoided. Similar to the EM algorithm for the log-likelihood, this new loss can be optimized by the so-called Sinkhorn EM algorithm, that I show to enjoy similar convergence guarantees as EM. By analyzing extensive numerical experiments as well as two real world applications on image segmentation for C.elegans microscopy and clustering in spatial transcriptomics experiments I show that this new loss improves upon optimization of the log-likelihood so it is a valuable clustering alternative for practitioners.

Keywords

Optimal Transport, Entropic Optimal Transport, Clustering, EM algorithm, Sinkhorn algorithm.

Generalized network structured models with mixed responses subject to measurement error and misclassification

Qihuang Zhang[1], **Grace Yi**[2]

[1] *McGill University, Department of Biostatistics, Canada,*
qihuang.zhang@mcgill.ca
[2] *University of Western Ontario, Department of Statistical and Actuarial Sciences and Department of Computer Science, Canada,*
gyi5@uwo.ca

Abstract

Research of associations between covariates with a complex association structure and multiple responses has attracted increasing attention. A great challenge in analyzing such data is posited by the presence of the network structure in covariates that is typically unknown. Moreover, mismeasurement of responses introduces additional complexity to distort usual inferential procedures. In this talk, I will discuss the problem with mixed binary and continuous responses that are subject to mismeasurement and associated with complex structured covariates. I will start with the case where data are precisely measured and describe a generalized network structured model. The development will be further extended to accommodate mismeasured responses, where the information on mismeasurement is either known or estimated from a validation sample. Theoretical results are established and numerical studies are conducted to evaluate the finite sample performance of the proposed methods.

Keywords

Gaussian graphical model, gene regulatory network, generalized estimating equation, measurement error, misclassification.

A new species distribution modeling approach for biased citizen science data

Greta Panunzi[1], **Jafet Belmont**[2], **Janine Illian**[2] **and Sara Martino**[3]

[1] *Department of Statistical Sciences, Sapienza University of Rome, Italy, greta.panunzi@uniroma1.it*
[2] *School of Mathematics and Statistics, University of Glasgow, U.K.*
[3] *Department of Mathematical Sciences, Norway.*

Abstract

In the field of ecology, understanding and predicting species distribution is crucial for effective conservation strategies. Citizen Science initiatives have revolutionized the collection of species occurrence data, providing a cost-effective method for monitoring wildlife across various spatiotemporal scales. However, the lack of standardized sampling protocols within CS programs presents analytical challenges, resulting in biased sampling efforts that favor regions that are easily accessible or extensively studied. This study conducts a case study on the national butterfly monitoring program in the UK, utilizing a marked point process framework that integrates spatial and temporal covariates to analyze the spatial and temporal patterns of butterfly species occurrences. Our approach combines two essential components: (i) the locations visited by volunteers participating in these schemes and (ii) the presence records of a species of interest, using the INLA/inlabru framework. Our investigation highlights the importance of addressing sampling bias to enhance the accuracy and reliability of species distribution modeling. By combining the strengths inherent in CS data with rigorous modeling techniques, our approach paves the way for a better understanding of species distribution dynamics, strengthening the foundation for more effective conservation efforts.

Keywords

Citizen Science; INLA; Integrated Models; Occupancy; Preferential Sampling.

References

Martino, S., Pace, D. S., Moro, S., Casoli, E., Ventura, D., Frachea, A., ... others (2021). Integration of presence-only data from several sources: a case study on dolphins' spatial distribution. *Ecography, 44 (10)*, 1533–1543.

Pollard, E. (1977). A method for assessing changes in the abundance of butterflies. *Biological conservation, 12 (2)*, 115–134.

Yuan, Y., Bachl, F. E., Lindgren, F., Borchers, D. L., Illian, J. B., Buckland, S. T.,... Gerrodette, T. (2017). Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales.

Optimal (?) Monte Carlo Methods for Nested Structural Problems

Guanyang Wang[1]

[1] *Rutgers University, Department of Statistics, U.S.,*
guanyang.wang@rutgers.edu

Abstract

The estimation of repeatedly nested expectations is a challenging task that arises in many applications. However, existing methods generally suffer from high computational costs when the number of nestings becomes large. This talk will introduce a novel Monte Carlo algorithm which achieves an optimal cost of $\mathcal{O}(1/\varepsilon^2)$ to obtain an estimator within ε-accuracy. We will also discuss its applications, caveats, and potential generalizations. This talk is based on joint works with Jose Blanchet, Peter Glynn, Yasa Syed, and Zhengqing Zhou.

Keywords

Monte Carlo, nested expectation, optimal cost.

High-dimensional clustering of compound precipitation and wind extremes over Europe

Gwladys Toulemonde[1], **Alexis Boulin**[2], **Elena Di Bernardino**[3], **Thomas Laloë**[3]

[1] *Univ Montpellier, CNRS, IMAG and Inria Lemon, France, gwladys.toulemonde@umontpellier.fr*
[2] *Université Côte d'Azur, CNRS, LJAD and Inria Lemon, France*
[3] *Université Côte d'Azur, CNRS, LJAD, France*

Abstract

Disastrous climate events such as floods, wildfires, and heatwaves often occur due to the simultaneous extreme behaviour of several interacting processes. Since in these compound events several spatio-temporal factors are jointly extreme and by their very nature are of high dimension, it is for a proper understanding of them to develop dependence summary measures that are appropriate for extreme value random vectors. These latter is a key ingredient to propose spatial clustering of these temporal processes. Based on the recent development of an algorithm specifically tailored for AI-block models (see Boulin et al., 2023) we propose in this talk a clustering method adapted to compound extreme events. We exemplify this method proposing a regionalization task. More precisely we identify regions based on gridded data from observations and climate model ensembles over Europe. This approach uses daily precipitation sums and daily maximum wind speed data from the ERA5 reanalysis dataset from 1979 to 2022.

Keywords

Spatial clustering, compound extremes, extremal dependence.

References

Boulin, A., Di Bernardino, E., Laloë, T., Toulemonde G. (2023). High-dimensional variable clustering based on sub-asymptotic maxima of a weakly dependent random process. *arXiv:2302.00934*.

Matrix denoising and completion based on Kronecker product approximation

Chencheng Cai[1], **Rong Chen**[2], **Han Xiao**[2]

[1] *Washington State University, Department of Mathematics and Statistics, USA, chencheng.cai@wsu.edu*
[2] *Rutgers University, Department of Statistics, USA, rongchen@stat.rutgers.edu, hxiao@stat.rutgers.edu*

Abstract

We consider the problem of matrix denoising and completion induced by the Kronecker product decomposition. Specifically, we propose to approximate a given matrix by the sum of a few Kronecker products of matrices, which we refer to as the Kronecker product approximation (KoPA). Because the Kronecker product is an extension of the outer product from vectors to matrices, KoPA extends the low rank matrix approximation, and includes it as a special case. Comparing with the latter, KoPA also offers a greater flexibility, since it allows the user to choose the configuration, which are the dimensions of the two smaller matrices forming the Kronecker product. On the other hand, the configuration to be used is usually unknown, and needs to be determined from the data in order to achieve the optimal balance between accuracy and parsimony. We propose to use extended information criteria to select the configuration. Under the paradigm of high dimensional analysis, we show that the proposed procedure is able to select the true configuration with probability tending to one, under suitable conditions on the signal-to-noise ratio. We demonstrate the superiority of KoPA over the low rank approximations through numerical studies, and several benchmark image examples.

Keywords

Matrix denoising, matrix completion, Kronecker product, low rank approximation.

Test-Fairness Deep Learning with Influence Score

Henry Horng-Shing Lu[1]

[1] *National Yang Ming Chiao Tung University, Taiwan,*
henryhslu@nycu.edu.tw

Abstract

We develop a new method on the deep learning model to make the model fair while keeping high prediction performance. We adopt Influence Score in the proposed model, which is the feature selection algorithm that takes the interaction between multiple variables into consideration, and the influential features will be included in follow-up predictions. Through this method, the fair model only contains the important features not influenced by the discriminatory factor and keeps its high prediction performance. In the experiment, we apply the method to the ISIC 2019 and Asan skin lesion datasets. The ISIC 2019 dataset is mainly collected from Western, and the Asan dataset is mainly collected from South Korea. Because in the skin lesion literature they show that the probability of diagnosed skin lesion in Western is significantly higher than Asian, we consider the area information as the bias of the prediction model. The result shows that the fair model can fairly and precisely classify the types of skin lesions by eliminating the discriminatory information. This is a joint work with collaborators, including Professor Shaw-Hwa Lo in Columbia University and others.

Keywords

Test-Fairness, Deep Learning, Influence Score.

Tensor quantile regression with low-rank tensor train estimation

Zihuan Liu[1], **Cheuk Yin Lee**[2], **Heping Zhang**[3]

[1] *Yale University, Biostatistics, USA, liuzihua@msu.edu*
[2] *Chinese University of Hong Kong, School of Science and Engineering, leecheukyin@cuhk.edu.cn*
[3] *Yale University, Biostatistics, USA, heping.zhang@yale.edu*

Abstract

Neuroimaging studies often involve predicting a scalar outcome from an array of images collectively called tensor. The use of magnetic resonance imaging (MRI) provides a unique opportunity to investigate the structures of the brain. To learn the association between MRI images and human intelligence, we formulate a scalar-on-image quantile regression framework. However, the high dimensionality of the tensor makes estimating the coefficients for all elements computationally challenging. To address this, we propose a low-rank coefficient array estimation algorithm based on tensor train (TT) decomposition which we demonstrate can effectively reduce the dimensionality of the coefficient tensor to a feasible level while ensuring adequacy to the data. Our method is more stable and efficient compared to the commonly used, Canonic Polyadic rank approximation-based method. We also propose a generalized Lasso penalty on the coefficient tensor to take advantage of the spatial structure of the tensor, further reduce the dimensionality of the coefficient tensor, and improve the interpretability of the model. The consistency and asymptotic normality of the TT estimator are established under some mild conditions on the covariates and random errors in quantile regression models. The rate of convergence is obtained with regularization under the total variation penalty. Extensive numerical studies, including both synthetic and real MRI imaging data, are conducted to examine the empirical performance of the proposed method and its competitors.

Keywords

Tensor regression, conditional quantile, tensor train decomposition, total variation.

Overview of Functional Dependence in Brain Networks

Hernando Ombao[1]

[1] *King Abdullah University of Science and Technology, Statistics Program, Saudi Arabia, hernando.ombao@kaust.edu.sa*

Abstract

Brain activity is complex. A full understanding of brain activity requires careful study of its multi-scale spatial-temporal organization (from neurons to regions of interest; and from transient events to long-term temporal dynamics). Motivated by these challenges, we will explore some characterizations of dependence between components of a brain network. This is potentially interesting because alterations in functional brain connectivity are associated with mental and neurological diseases. In this talk, we provide an overview of functional dependence measures. We present a general framework for exploring dependence through the oscillatory activities derived from each component of the tine series. The talk will draw connections of this framework to some of the classical notions of spectral dependence such as coherence, partial coherence, and dual-frequency coherence. Moreover, this framework provides a starting point for exploring potential non-linear cross-frequency interactions. These interactions include the impact of phase of one oscillatory activity in one component on the amplitude of another oscillation. The proposed approach captures lead-lag relationships and hence can be used as a general framework for spectral causality. Under this framework, we will also present some recent work on inference using spectral mutual information and entropy measures. This is joint work with Marco Pinto (UC Irvine), Paolo Redondo (KAUST) and Raphael Huser (KAUST).

Keywords

Brain Networks, Coherence, Spectral Analysis, Transfer Entropy.

Is there a cap on how long a human can live? Truncation, censoring and extreme value modelling

Holger Rootzén[1]

[1] *Chalmers and Gothenburg University, Mathematical Sciences, Sweden, hrootzen@chalmers.se*

Abstract

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 review paper in Annual Review of Statistics and its Application, written together with Léo Belzile, Anthony Davison, Jutta Gampe and Dimitrii Zholud.

Keywords

Censoring, data validation, extreme old age, generalized Pareto distribution, Gompertz distribution, Lexis diagram, survival analysis, supercentenarian, truncation.

On the Need for a Language Describing Distribution Shifts

Jiashuo Liu, Tianyu Wang, Peng Cui, Hongseok Namkoong[1]

[1] *Decision, Risk, and Operations Division, Columbia Business School,* namkoong@gsb.columbia.edu

Abstract

Different distribution shifts require different algorithmic and operational interventions. Methodological research must be grounded by the specific shifts they address. Although nascent benchmarks provide a promising empirical foundation, they implicitly focus on covariate shifts, and the validity of empirical findings depends on the type of shift, e.g., previous observations on algorithmic performance can fail to be valid when the $Y|X$ distribution changes. We conduct a thorough investigation of natural shifts in 5 tabular datasets over 86,000 model configurations, and find that $Y|X$-shifts are most prevalent. To encourage researchers to develop a refined language for distribution shifts, we build WhyShift, an empirical testbed of curated real-world shifts where we characterize the type of shift we benchmark performance over. Since $Y|X$-shifts are prevalent in tabular settings, we identify covariate regions that suffer the biggest $Y|X$-shifts and discuss implications for algorithmic and data-based interventions. Our testbed highlights the importance of future research that builds an understanding of how distributions differ.

Transfer Learning with Random Coefficient Ridge Regression

Hongzhe Zhang[1], **Hongzhe Li**[1]

[1] *University of Pennsylvania, USA, hongzhe@upenn.edu*

Abstract

Ridge regression with random coefficients provides an important alternative to fixed coefficients regression in high dimensional setting when the effects are expected to be small but not zeros. This paper considers estimation and prediction of random coefficient ridge regression in the setting of transfer learning, where in addition to observations from the target model, source samples from different but possibly related regression models are available. The informativeness of the source model to the target model can be quantified by the correlation between the regression coefficients. Two estimators of regression coefficients of the target model using the weighted sum of the ridge estimates of both target and source models are developed, where the weights can be determined by minimizing the limiting estimation risk or prediction risk. Using random matrix theory, the limiting values of the optimal weights are derived under the setting when $p/n \to \gamma$, where p is the number of the predictors and n is the sample size, which leads to an explicit expression of the estimation or prediction risks. Simulations show that these limiting risks agree very well with the empirical risks. An application to predicting the polygenic risk scores for lipid traits shows such transfer learning methods lead to smaller prediction errors than the single sample ridge regression or Lasso-based transfer learning.

Keywords

Genetic correlation, random matrix theory, genome wide association studies, polygenic risk score.

Bayesian Empirical Likelihood Inference for Estimating Equations

Weichang Yu and Howard Bondell

University of Melbourne, School of Mathematics and Statistics, Australia weichang.yu@unimelb.edu.au howard.bondell@unimelb.edu.au

Abstract

Bayesian inference typically relies on specification of a likelihood as a key ingredient. Recently, likelihood-free approaches have become popular to avoid specification of potentially intractable likelihoods. Alternatively, in the Frequentist context, estimating equations are a popular choice for inference corresponding to an assumption on a set of moments (or expectations) of the underlying distribution, rather than its exact form. Common examples are in the use of generalised estimating equations with correlated responses, or in the use of M-estimators for robust regression avoiding the distributional assumptions on the errors. In this talk, I will discuss some of the motivation behind empirical likelihood, and how it can be used to incorporate a fully Bayesian analysis into these settings where only specification of moments is desired. This allows one to then take advantage of prior distributions that have been developed to accomplish various shrinkage tasks, both theoretically and in practice. I will further discuss computational issues that arise due to non-convexity of the support of this likelihood and the corresponding posterior, and show how this can be rectified to allow for MCMC and variational approaches to perform posterior inference.

Keywords

Empirical likelihood, Posterior computation, Variational Bayes.

Design and Analysis of Quantitative Mass-Spectrometry Proteomics Experiments

Huaying Fang[1,2], **Lihua Jiang**[1], **Michael P. Snyder**[1], **Hua Tang**[1]

[1] *Stanford University School of Medicine, Department of Genetics, USA hyfang@stanford.edu & huatang@stanford.edu*
[2] *Academy of Multidisciplinary Studies, Capital Normal University, China.*

Abstract

Recent advancements in multiplex mass-spectrometry technologies have facilitated high-throughput quantitative profiling of the proteome. However, several experimental variables may increase sample variability and introduce systematic biases. In a recent proteogenomic investigation, we introduced two innovative features in experimental design: First, two reference samples were incorporated within each mass-spectrometry run to serve as internal standards, and second, each specimen was assayed as technical replicates in two distinct mass-spectrometry runs. In this context, we propose model and computational methods to harness the potential of these supplementary experimental components. We use both simulated and real data to evaluate the benefits of these enhanced experimental design features.

Keywords

Proteomics, mass-spectrometry, linear mixed model, normalization.

Probabilistic prediction for spatial processes through deep learning

Pratik Nag[1], **Ying Sun**[2], **Huixia Judy Wang**[3]

[1] *KUAST, Saudi Arabia, pratik.nag@kaust.edu.sa*
[2] *KAUST, Saudi Arabia, ying.sun@kaust.edu.sa*
[3] *The George Washington University, USA, judywang@gwu.edu*

Abstract

In spatial statistics, the kriging predictor is the best linear predictor at unsampled locations, but not the optimal predictor for non-Gaussian or nonstationary processes. In this talk, I will introduce an indicator deep kriging method for univariate and bivariate spatial processes. The method is based on thresholding the spatial observations at a given set of quantile values and a deep neural network framework. The developed method does not require any parametric assumptions on the marginal distribution and, thus, is more flexible than existing methods. The method can provide the entire predictive distribution function at a new location, allowing for both point and interval predictions. I will present some numerical results to demonstrate the method's efficacy compared to existing approaches.

Keywords

Bivariate, deep learning, kriging, spatial.

Optimizing Two-Variable Gamma Accelerated Degradation Tests with a Semi-Analytical Approach

Hung-Ping Tung[1]

[1] *Department of Industrial Engineering and Management, National Yang Ming Chiao Tung University, Taiwan, hptung@nycu.edu.tw*

Abstract

Gamma accelerated degradation tests are widely used to assess timely lifetime information of highly reliable products when the degradation path of quality characteristic of products follows a monotonic process. In this talk, a semi-analytical approach is proposed to determine the optimal designs for two-variable gamma accelerated degradation tests under three criteria: D-optimality, A-optimality and V-optimality. We first use general equivalence theorem to prove that the optimal approximate designs only allocate test units at the four vertices of a rectangular design region, and the corresponding optimal proportion of total number of measurements at each stress level is derived. Next, we apply the concept of prescribed accuracy level and total experimental cost to further determine optimal integer designs. More specifically, a numerical approach is used to resolve the number of test units and number of measurements at each stress level.

Keywords

Accelerated degradation tests, gamma process, general equivalence theorem, optimal design.

Optimal Designs of Accelerated Degradation Tests with Unequal Measurement Intervals

I-Chen Lee

National Cheng Kung University, Taiwan, iclee@ncku.edu.tw

Abstract

The accelerated degradation test (ADT) is widely used to assess the lifetime information for highly reliable products. To obtain the failure information more efficiently, how to design an efficient ADT plan is a critical task for real applications. In this study, we mainly focus on the determination of sample size allocation, and we assume that the degradation path follows gamma degradation models, including the fixed effect degradation models and random effect degradation models. For the constraints of conducting an ADT, we allow the intervals between two consecutive measurements under different settings of stress levels can be different. The theoretical or numerical solutions are provided, and the results demonstrate that the different lengths of measurement intervals in the ADTs indeed affects the optimal sample allocation. Through simulation studies, we validate that the proposed strategy is exactly an optimal design and is better than the existing strategies.

Keywords

Accelerated degradation test, Gamma degradation models, Optimal design.

Statistical Inference via Sample Splitting

Ilmun Kim[1], **Shubhanshu Shekhar**[2], **Aaditya Ramdas**[3]

[1] *Yonsei University, Department of Statistics and Data Science, South Korea, ilmun@yonsei.ac.kr*
[2] *Carnegie Mellon University, Department of Statistics and Data Science, USA, shubhan2@andrew.cmu.edu*
[3] *Carnegie Mellon University, Department of Statistics and Data Science, USA, aramdas@stat.cmu.edu*

Abstract

Classical asymptotic theory for statistical inference usually involves calibrating a statistic by fixing the dimension d while letting the sample size n increase to infinity. Recently, much effort has been dedicated towards understanding how these methods behave in high-dimensional settings, where d and n both increase to infinity simultaneously. This often leads to different inference procedures, depending on the assumptions about the dimensionality, leaving the practitioner in a bind. Motivated by this critical issue, we introduce a simple yet powerful approach whose validity does not depend on the assumption on d versus n. At the heart of our proposal are sample splitting and studentization: after splitting the dataset into two parts, we use the first split to learn a direction with strong signal, and then project the points in the second split along that direction. The projected random variables are aggregated via studentization that leads to a dimension-agnostic limiting distribution. In this talk, we exemplify our technique for a handful of classical and modern problems and discuss directions for future research.

Keywords

Sample Splitting, Studentization, Dimension-agnostic inference, Kernel-based tests, Minimax power.

Monitoring Machine Learning Forecasts for Platform Data Streams

Jeroen Rombouts[1], **Ines Wilms**[2]

[1] *ESSEC Business School, France, rombouts@essec.edu*
[2] *Maastricht University, Department of Quantitative Economics, the Netherlands, i.wilms@maastrichtuniversity.nl*

Abstract

Data stream forecasts are essential inputs for decision making at digital platforms. Machine learning (ML) algorithms are appealing candidates to produce such forecasts. Yet, digital platforms require a large-scale forecast framework that can flexibly respond to sudden performance drops. Re-training ML algorithms at the same speed as new data batches enter is usually computationally too costly. On the other hand, infrequent re-training requires specifying the re-training frequency and typically comes with a severe cost of forecast deterioration. To ensure accurate and stable forecasts, we propose a simple data-driven monitoring procedure to answer the question when the ML algorithm should be re-trained. Instead of investigating instability of the data streams, we test if the incoming streaming forecast loss batch differs from a well-defined reference batch. Using 15-min data from an on-demand logistics platform operating in London, we apply the monitoring procedure to popular ML algorithms including random forest, XGBoost and lasso. We show that monitor-based re-training produces accurate forecasts compared to viable benchmarks while preserving computational feasibility. Moreover, the choice of monitoring procedure is more important than the choice of ML algorithm, thereby permitting practitioners to combine the proposed monitoring procedure with one's favorite forecasting algorithm.

Keywords

Forecasting, Machine Learning, Monitoring, Platform econometrics, Streaming data, Time series.

Synthetic data with vine-copulas – balancing utility and privacy

Ingrid Hobæk Haff[1]

[1] *University of Oslo, Department of Mathematics, Norway*

Abstract

When data are highly sensitive and cannot be published, which is often the case for instance in medical applications, synthetic data can be the solution, provided that privacy is taken into consideration when generating them. Privacy concerns are typically handled by introducing "noise" into the data used to train the synthetic data generator ([1],[2],[3]). The more noise is introduced, the better the privacy, but on the other hand, it makes the resulting synthetic data much less useful, thereby impairing its utility. Our solution is to generate synthetic data from original data using vine copulas, in a way that balances privacy against utility. To do this is in a good manner, we take the knowledge about the intended use of the synthetic data, which we assume is classification or regression, into account. Vine copulas are very flexible statistical models, that are able to capture a wide range of complex dependencies, and are composed of a joint dependence structure modelled by bivariate building blocks, namely pair-copulas, and marginal distributions ([4],[5],[6]), built from a nested sequence of trees consisting of bivariate building blocks, namely pair-copulas. Our idea is to design the vine copula in such that the important dependencies between the response Y and the features $\mathbf{X}$ are captured in the first trees, which is obtained by using a so-called C-vine. Further, we truncate the vine copula, meaning that we cut out all trees after a certain level. By doing so, we emphasise the fit to the part of the model that is important for the utility of the synthetic data, namely classification or regression performance, while introducing noise into parts of the model that are not important for the utility, but may contain sensitive features. Hence, noise is introduced into the model in a targeted way,

as opposed to more uniformly, as in current best practices. The result is synthetic data with comparable level of privacy, but much better utility.

This is joint work with Elisabeth Griesbauer, Arnoldo Frigessi and Claudia Czado.

Keywords

C-vine, truncation, classification, regression.

References

[**1**] Dwork, C., McSherry, F., Nissim, K. and Adam Smith (2006): Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, 265–284. Springer.

[**2**] Dwork, C. and Roth, A. (2014): The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*, 211–407.

[**3**] Gambs, S., Ladouceur, F., Laurent, A. and Roy-Gaumond, A. (2021): Growing synthetic data through differentially-private vine copulas. *Proc. Priv. Enhancing Technol.*, *3*, 122–141.

[**4**] Bedford, T. and Cooke, R.M. (2001): Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial intelligence*, *32*, 245–268.

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

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

Copula based Cox proportional hazards models for dependent censoring

N.W. Deresa[1], **I. Van Keilegom**[2]

[1] *Universiteit Hasselt, Data Science Institute, Belgium, negerawakgari.deresa@uhasselt.be*
[2] *KU Leuven, ORSTAT, Belgium, ingrid.vankeilegom@kuleuven.be*
UCLouvain, LIDAM, Belgium

Abstract

Most existing copula models for dependent censoring in the literature assume that the parameter defining the copula is known. However, prior knowledge on this dependence parameter is often unavailable. In this paper we propose a novel model under which the copula parameter does not need to be known. The model is based on a parametric copula model for the relation between the survival time (T) and the censoring time (C), whereas the marginal distributions of T and C follow a semiparametric Cox proportional hazards model and a parametric model, respectively. We show that this model is identified, and propose estimators of the nonparametric cumulative hazard and the finite-dimensional parameters. It is shown that the estimators of the model parameters and the cumulative hazard function are consistent and asymptotically normal. We also investigate the performance of the proposed method using finite-sample simulations. Finally, we apply our model and estimation procedure to a follicular cell lymphoma data set.

Keywords

Association, asymptotic theory, identifiability, semiparametric regression, survival analysis.

Semiparametric Variable Selection in Kernel Machine Survival Model

Inyoung Kim[1]

[1] *Virginia Tech, Department of Statistics, USA, inyougk@vt.edu*

Abstract

Motivated by a breast cancer gene-pathway data set, which exhibits the "small n, large p" characteristics, we propose a semiparametric variable selection method for Bayesian kernel survival model to simultaneously study the effects of both clinical covariates and gene expression levels within a pathway on survival time and also identify important variables associated to survival time. We model the unknown high-dimension functions of pathways via Gaussian kernel machine to consider the possibility that genes within the same pathway interact with each other. To address the multiple comparisons problem under a full Bayesian setting, we propose a similarity-dependent procedure based on Bayes factor to control the family-wise error rate. We demonstrate the outperformance of our approach under various simulation settings and pathways data.

Keywords

Gaussian Process, Kernel Hilbert Space, Kernel Machine Regression.

Navigating Spatial Confounding in a Bayesian Framework: Assessment, Approaches, and Practical Recommendations for Researchers

Isa Marques[1], **Emiko Dupont**[2], **Thomas Kneib**[3], **Paul Wiemann**[4]

[1] *School of Mathematics and Statistics, University of Glasgow, UK, Isa.Marques@glasgow.ac.uk*
[2] *Department of Mathematical Sciences, University of Bath, UK, eahd20@bath.ac.uk*
[3] *Chair of Statistics and Campus Institute Data Science, University of Göttingen, Germany, tkneib@uni-goettingen.de*
[4] *Department of Statistics, University of Wisconsin-Madison, US, paul.wiemann@wisc.edu*

Abstract

Spatial confounding arises from the interplay of collinearity among covariates and spatial random effects within regression models, as well as a direct result of smoothing in the model fitted. In this presentation, we delve into the drivers of spatial confounding from a theoretical point of view. Subsequently, we explore the potential of Bayesian methodology in alleviating spatial confounding and leveraging the understanding of how such confounding originates in the construction of prior distributions. Lastly, but perhaps often overlooked, we shed light on the practical application of these conceptual insights to analyzing real-world datasets in forestry and ecology. We aim to identify situations prone to potential confounding and determine the most suitable models for specific circumstances.

Confounding bias, spatial regression, spatial random effects, smoothing

The Myth of the Kraken:
When Mythology Meets EVT

Jessica Silva Lomba[1,2], **Isabel Fraga Alves**[1],

[1] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, mialves@ciencias.ulisboa.pt*
[2] *MC Sonae, Matosinhos, Portugal, jslomba@ciencias.ulisboa.pt*

Abstract

The *Architeuthis*, also known as Giant Squid and often associated with the mythological Kraken, is one of the most elusive great sea creatures, and much about their biology and behaviour is still shrouded in mystery. Much debate exists about exactly how long can these creatures be. Here, we develop an Extreme Value analysis with the aim of obtaining a statistical answer to this question, a novelty approach to the small data set of Giant Squid measurements available in the literature. We focus on records of the *Total Length* (TL) of the *Architeuthis*, as well as on the *Mantle Length* (MtL) – data set in Paxton (2016). The goal is on questions such as: *What is the expected size of a specimen observed on average once every 1000 records? What is the maximum possible size of the Architeuthis?* To answer these questions we employ several techniques from EV Statistics, both *parametric* and *semi-parametric*. The basic assumption that the underlying distributions of TL and MtL of Giant Squids, respectively F_{TL} and F_{MtL}, belong to some max-domain of attraction (DoA), followed by testing the plausibility of the assumption that F_{TL} and F_{MtL} have a finite right endpoint x^F, i.e., if there is statistical evidence of a finite maximum possible length for Giant Squids (cf. Neves and Fraga Alves, 2008). In a *parametric* approach, given the selected level u^* – see Silva Lomba and Fraga Alves (2020) for the *Automatic L-moment Ratio Selection Method* (ALRSM) and Northrop and Coleman(2014) for the *Score Test Selection Method* (STSM) – the GPd fit to the n^* excesses over u^* is obtained via maximum likelihood (ML). We also

followed an exploratory *semi-parametric* estimation of the assumed-to-exist right endpoints. As the interest in this analysis is pivoted towards studying the existence of a finite right endpoint, we employed the EVI-independent procedures devised by Neves and Pereira (2010), for testing the hypothesis of finiteness of the right endpoint. The preliminary assessment seems to be coherent with the existence of a statistical upper bound for the Giant Squid size, regarding both TL and MtL. Afterwards, we looked at the sample paths for the *general right endpoint estimator* $\hat{x}^F_{FAN}$. These *semi-parametric* tools were presented by Fraga Alves *et al.*(2017), in the context of supercentenarian women lifespan. Returning to both questions above, and based on the data set of Paxton (2016), we concluded that it is expected that a Giant Squid of about $17m$ (*reliable*) to $20m$ (*contested*) of TL, or around $3m$ of MtL, may be observed, on average, once every 1000 measurements; moreover, statistically there is evidence that the maximum size of the *Architeuthis* may be about $18m$(*reliable*) to $22m$ (*contested*) of TL, or between $3m$ (*semi-parametric*) and $5m$(*parametric*) of MtL.

Keywords

Architeuthis, EVT, POT, L-moment theory, Automatic Threshold Selection.

Acknowledgment

This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020.

References

Fraga Alves, I., Neves, C., and Rosário, P. (2017). A general estimator for the right endpoint with an application to supercentenarian women's records. *Extremes*, 20(1):199-237. doi:10.1007/s10687-016-0260-6

Neves, C. and Fraga Alves, M. I. (2008). Testing extreme value conditions – an overview and recent approaches. *REVSTAT - Statistical Journal*, 6(1):83-100. doi:10.57805/revstat.v6i1.59

Neves, C. and Pereira, A. (2010). Detecting finiteness in the right endpoint of light-tailed distributions. *Statistics & Probability Letters*, 80(5-6):437-444. doi:10.1016/j.spl.2009.11.021

Northrop, P. J. and Coleman, C. L. (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes*, 17(2):289-303. doi:10.1007/s10687-014-0183-z

Paxton, C. G. M. (2016). Unleashing the Kraken: on the maximum length in giant squid (*Architeuthis* sp.). *Journal of Zoology*, 300(2):82-88. doi:10.1111/ jzo.12347

Silva Lomba, Jessica. *Contributions to Inference in Extremes based on Moment type Statistics*. PhD Thesis University Lisbon. 2023.

Silva Lomba, J. and Fraga Alves, M. I. (2020). L-moments for automatic threshold selection in extreme value analysis. *Stochastic Environmental Research and Risk Assessment*, 34(3):465-491. doi:10.1007/s00477-020-01789-x

Further Tales on the Role of Tails in Risk Assessment

M. Ivette Gomes[1], **Frederico Caeiro**[2], **Lígia Henriques-Rodrigues**[3]

[1] *CEAUL, University of Lisbon, Portugal, migomes@ciencias.ulisboa.pt*
[2] *CMA, NOVA University Lisbon, Portugal, fac@fct.unl.pt*
[3] *CIMA, University of Évora, Portugal, ligiahr@uevora.pt*

Abstract

The *Weibull tail-coefficient* (WTC), being the reciprocal of the index of regular variation of a regularly varying cumulative hazard function, $H(x) = -\log(1 - F(x))$, can have a high relevance in the assessment of risk. Due to the specificity of the WTC, and its deep and clear link to a positive EVI, any estimator of a positive EVI, like all *generalised means* (GMs) EVI-estimators (see Gomes and Martins, 2001, and Caeiro *et al.*, 2016, among others), generalizing the classical Hill (1975) estimator, can be used for the estimation of the WTC. We mention the WTC-estimators in Caeiro, Gomes and Henriques-Rodrigues (2022), Caeiro, Henriques-Rodrigues and Gomes (2022) and Henriques-Rodrigues, Caeiro and Gomes (2023). Contrarily to the WTC, these estimators are scale invariant but not location invariant. With PORT standing for *peaks over random threshold*, new classes of PORT WTC-estimators are now introduced and studied. These classes are dependent on an extra tuning parameter s, $0 \leq s < 1$, and they are both location and scale invariant. The asymptotic normal behaviour of those PORT classes is derived. These WTC-estimators are further studied for finite samples, through a Monte-Carlo simulation study. An adequate choice of the tuning parameters under play is put forward, as part of an interesting '*tale of the tails*'. Some concluding remarks are further provided.

Keywords

Risk assessment, semi-parametric estimation, statistics of extremes, Weibull tail-coefficient.

Acknowledgement

Research partially supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under the projects UIDB/00006/2020 (CEAUL), UIDB/00297/2020 (CMA/UNL), UIDB/04674/2020 (CIMA), and HiTEc Cost Action CA21163.

References

Caeiro, F., Gomes, M.I., Beirlant, J. and T. de Wet (2016). Mean-of-order-p reduced-bias extreme value index estimation under a third-order framework. *Extremes*, 19:4, 561–589. https://doi.org/10.1007/s10687-016-0261-5

Gomes, M.I. and M.J. Martins (2001). Generalizations of the Hill estimator — asymptotic versus finite sample behaviour. *J. Statist. Planning and Inference* **93**, 161-180. https://doi.org/10.1016/S0378-3758(00)00201-9

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3, 1163–1174. https://doi.org/10.1214/aos/1176343247

Caeiro, F., Gomes, M.I. and L. Henriques-Rodrigues (2022). Estimation of the Weibull Tail Coefficient through the Power Mean-of-Order-p. In Bispo R *et al.* (Eds.), *Recent Developments in Statistics and Data Science*, Springer Proceedings in Mathematics and Statistics, vol 398, Springer, Cham, section 4, pp. 41–53. https://doi.org/10.1007/978-3-031-12766-3_4

Caeiro, F., Henriques-Rodrigues, L. and M.I. Gomes (2022).The use of Generalized Means in the Estimation of the Weibull Tail Coefficient. *Computational and Mathematical Methods*, Article ID 7290822, 12 pages. https://doi.org/10.1155/2022/7290822

Henriques-Rodrigues L., Caeiro F. and M.I. Gomes (2023). *Improvements in the Estimation of the Weibull Tail Coefficient—a Comparative Study.* http://arxiv.org/abs/2310.01072

Recurrent event analysis: basic concepts and some recent contributions

Ivo Sousa-Ferreira[1,3], **Cristina Rocha**[2,3] **and Ana Maria Abreu**[1,4]

[1] *Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal*
[2] *Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[3] *CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[4] *CIMA – Centro de Investigação em Matemática e Aplicações, Portugal*
ivo.ferreira@staff.uma.pt

Abstract

Over the past decade, substantial efforts have been devoted to the development of survival models for gap times between recurrent events. An emerging approach consists in proposing parametric rate models derived from a non-homogeneous Poisson process, allowing to deduce the conditional distribution of each gap time given the previous recurrence time. In order to capture how the rate function changes over time, exploring different baseline distributions is usually a good option. Furthermore, assessing the impact of unobserved heterogeneity is important, as it induces within-subject correlation and can lead to biased estimators. This aspect encourages the inclusion of a shared frailty (i.e., a random effect) into the model. The aim of this talk is to discuss basic concepts and some recent contributions in recurrent event analysis, while addressing different distributional assumptions in the formulation of gap time models. Specifically, models based on a new lifetime distribution (Sousa-Ferreira et al., 2022, 2023) or on restricted cubic splines are presented.

Keywords

Gap times, non-homogeneous Poisson process, parametric models, recurrent events, survival analysis.

Acknowledgement

This work is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the projects UIDB/00006/2020 (CEAUL) and UIDB/04674/2020 (CIMA).

References

Sousa-Ferreira, I., Abreu, A.M. and Rocha , C. (2023). The extended Chen-Poisson lifetime distribution. *REVSTAT-Statistical Journal*, *21*(2), 173–196.

Sousa-Ferreira, I., Rocha, C. and Abreu, A.M. (2022). The extended Chen-Poisson marginal rate model for recurrent gap time data. In: Bispo, R., Henriques-Rodrigues, L., Alpizar-Jara, R., de Carvalho, M. (eds.) Recent Developments in Statistics and Data Science, pp. 337–351. Springer Proceedings in Mathematics & Statistics, *398*. Springer, Cham.

Semi-supervised clustering for ordered categorical data

Ying Cui[1,4], **Louise McMillan**[2,4], **Ivy Liu**[3,4]

[1] *School of Mathematics and Statistics, Victoria University of Wellington, New Zealand, cuiying@myvuw.ac.nz*
[2] *School of Mathematics and Statistics, Victoria University of Wellington, New Zealand, Louise.McMillan@vuw.ac.nz*
[3] *School of Mathematics and Statistics, Victoria University of Wellington, New Zealand, Ivy.Liu@vuw.ac.nz*
[4] *Centre for Data Science and Artificial Intelligence, Victoria University of Wellington, New Zealand*

Abstract

There are many methods for supervised clustering (cluster memberships are known already) and unsupervised clustering (memberships are unknown). This talk will discuss a model-based method for semi-supervised clustering focusing on ordinal response data, where a subset of memberships are known in advance. This situation may arise when cluster memberships can only be determined by an expensive investigation. We will demonstrate the methodology using an aquaculture example to determine the health status of fish from less costly variables. Often the true health status can only be determined by destructive features that lead to a small proportion of data labelled. Our method allows us to use all data (both labelled and unlabelled) to identify clusters of fish.

Keywords

Clustering, Ordered categorical data, Semi-supervised.

Generalized Data Thinning Using Sufficiency

Ameer Dharamshi[1], Anna Neufeld[2], Keshav Motwani[1], Lucy L. Gao[4], Daniela Witten[1,2], Jacob Bien[6]

[1] *University of Washington, Department of Biostatistics, USA*
[2] *University of Washington, Department of Statistics, USA*
[3] *University of British Columbia, Department of Statistics, USA*
[4] *University of Southern California, Department of Data Sciences and Operations, USA, jbien@usc.edu*

Abstract

Sample splitting is one of the most tried-and-true tools in the data scientist toolbox. It breaks a data set into two independent parts, allowing one to perform valid inference after an exploratory analysis or after training a model. Recent work has provided a remarkable alternative to sample splitting, which is attractive in situations where sample splitting is not possible. This alternative, called convolution-closed data thinning, proceeds very differently from sample splitting, and yet it also produces two statistically independent data sets from the original. In this talk, we will show that sufficiency is the key underlying principle that makes this approach possible. This insight leads naturally to a new framework, which we call generalized data thinning. This generalization unifies both sample splitting and convolution-closed data thinning as different applications of the same procedure. Furthermore, we show that this generalization greatly widens the scope of distributions where thinning is possible.

Keywords

Cross-validation, sample splitting, exponential families, selective inference, model validation.

Wasserstein-Quantile PCA

Jaesung Park[1], **Sungkyu Jung**[2]

[1] *Seoul National University, Department of Statistics, South Korea, wotjdddd@snu.ac.kr*
[2] *Seoul National University, Department of Statistics, South Korea, Sungkyu@snu.ac.kr*

Abstract

This paper introduces Quantile Principal Component Analysis (Quantile PCA), which is based on the PCA of quantile functions. Quantile PCA reconstructs the quantile function using the result of PCA, with the restriction that it must be a quantile function again. The key property of Quantile PCA is that all collections of square-integrable quantile functions called the quantile space, is a closed convex subset of a Hilbert space. To establish consistency of Quantile PCA, a novel framework called the "generalized Fréchet mea" is proposed and shown to be consistent as the sample size increases. Geometric properties of the quantile space are established to establish this framework. Discretized optimization problems and numerical algorithms are presented to solve for Quantile PCA. The superiority of the proposed algorithm over existing ones is demonstrated and several optimization problems are compared using simulated data and real data applications. Overall, the results propose that Quantile PCA is a promising approach for feature selection and dimension reduction in one-dimensional distribution datasets.

Keywords

PCA of density functions, Wasserstein Space, Consistency of Fréchet means.

A Spatio-Temporal Dirichlet Process Mixture Model for Coronavirus Disease-19

Jaewoo Park[1,2], **Seorim Yi**[1], **Won Chang**[3], **Jorge Mateu**[4]

[1] *Department of Statistics and Data Science, Yonsei University*
[2] *Department of Applied Statistics, Yonsei University*[3] *Division of Statistics and Data Science, University of Cincinnati*[4] *Department of Mathematics, University Jaume I*

Abstract

Understanding the spatio-temporal patterns of the coronavirus disease 2019 (COVID-19) is essential to construct public health interventions. Spatially referenced data can provide richer opportunities to understand the mechanism of the disease spread compared to the more often encountered aggregated count data. We propose a spatio-temporal Dirichlet process mixture model to analyze confirmed cases of COVID-19 in an urban environment. Our method can detect unobserved cluster centers of the epidemics, and estimate the space-time range of the clusters that are useful to construct a warning system. Furthermore, our model can measure the impact of different types of landmarks in the city, which provides an intuitive explanation of disease spreading sources from different time points. To efficiently capture the temporal dynamics of the disease patterns, we employ a sequential approach that uses the posterior distribution of the parameters for the previous time step as the prior information for the current time step. This approach enables us to incorporate time dependence into our model in a computationally efficient manner without complicating the model structure. We also develop a model assessment by comparing the data with theoretical densities, and outline the goodness-of-fit of our fitted model.

Keywords

Bayesian hierarchical model, Dirichlet process Gaussian mixture, Infectious diseases, Markov chain Monte Carlo, Spatio-temporal point patterns.

A Geometric Perspective on Bayesian and Generalized Fiducial Inference

Jan Hannig[1], **Yang Liu**[2], **Alexander C. Murph**[3]

[1] *University of North Carolina at Chapel Hill, USA, jan.hannig@unc.edu*
[2] *University of Maryland, USA, yliu87@umd.edu*
[3] *Los Alamos National Laboratory, USA, murph290@gmail.com*

Abstract

Post-data statistical inference concerns making probability statements about model parameters conditional on observed data. When a priori knowledge about parameters is available, post-data inference can be conveniently made from Bayesian posteriors. In the absence of prior information, we may still rely on objective Bayes or generalized fiducial inference (GFI). Inspired by approximate Bayesian computation, we propose a novel characterization of post-data inference with the aid of differential geometry. Under suitable smoothness conditions, we establish that Bayesian posteriors and generalized fiducial distributions (GFDs) can be respectively characterized by absolutely continuous distributions supported on the same differentiable manifold: The manifold is uniquely determined by the observed data and the data generating equation of the fitted model. Our geometric analysis not only sheds light on the connection and distinction between Bayesian inference and GFI, but also allows us to sample from posteriors and GFDs using manifold Markov chain Monte Carlo algorithms.

Keywords

Riemannian manifold, Bayesian inference, Generalized fiducial inference, foundations of statistics.

Error Reduction from Stacked Regressions

Xin Chen[1], **Jason M. Klusowski**[2], **Yan Shuo Tan**[3]

[1] *Princeton University, Department of Operations Research and Financial Engineering, USA, xc5557@princeton.edu*
[2] *Princeton University, Department of Operations Research and Financial Engineering, USA, jason.klusowski@princeton.edu*
[3] *National University of Singapore, Department of Statistics and Data Science, Singapore, yanshuo@nus.edu.sg*

Abstract

Stacking regressions is an ensemble technique that forms linear combinations of different regression estimators to enhance predictive accuracy. The conventional approach uses cross-validation data to generate predictions from the constituent estimators, and least-squares with nonnegativity constraints to learn the combination weights. We learn these weights analogously by minimizing an estimate of the population risk subject to a nonnegativity constraint. When the constituent estimators are linear least-squares projections onto nested subspaces separated by at least three dimensions, we show that thanks to a shrinkage effect, the resulting stacked estimator has strictly smaller population risk than best single estimator among them. Here "best" refers to an estimator that minimizes a model selection criterion such as AIC or BIC. In other words, in this setting, the best single estimator is inadmissible. Because the optimization problem can be reformulated as isotonic regression, the stacked estimator requires the same order of computation as the best single estimator, making it an attractive alternative in terms of both performance and implementation.

Keywords

Stacking; ensemble learning; model selection; shrinkage; isotonic regression.

Designing Experiments for Marketplaces and Other Bipartite Graphs

Jean Pouget-Abadie

Google Research, NYC, USA, jeanpa@google.com

Abstract

When the treatment assignment of one unit affects the outcome of another, we say there is interference. Interference is especially prevalent in marketplaces, where buyer and seller interactions lead to complex dependence structures. As a violation of the stable unit treatment value assumption (SUTVA), the presence of interference can lead to bias of standard estimators under naive randomized designs. In this talk, we will cover a set of design and estimation paradigms to conduct causal inference research in a bipartite graph setting, inspired from—but not limited to—marketplace experiments, with specific attention to clustered randomized designs under different randomization constraints and bias corrections to standard estimators.

Keywords

Causal inference, bipartite experiments, marketplaces, interference, leakage, violations of SUTVA, clustering.

References

Boyarsky, A., Namkoong, H., and Pouget-Abadie, J. (2023). Modeling interference using experiment roll-out. *Proceedings of the 24th ACM Conference on Economics and Computation, 298*

Brennan, J., Mirrokni, V., and Pouget-Abadie, J. (2022) Cluster randomized designs for one-sided bipartite experiments. *Advances in Neural Information Processing Systems 35*, 37962–37974.

Harshaw, C., Savje, F., Eisenstat, D., Mirrokni, V., Pouget-Abadie, J. (2023). Design and Analysis of Bipartite Experiments under a Linear Exposure-Response Model. *Electronic Journal of Statistics 17*, 464–518.

Pouget-Abadie, J., Aydin, K., Schudy, W., Brodersen, K., and Mirrokni, V. (2019). Variance reduction in bipartite experiments through correlation clustering. *Advances in Neural Information Processing Systems 32*.

Goodness of fit for Bayesian generative models

Guillaume Le Mailloux[1,2], **Jean-Michel Marin**[1], **Paul Bastide**[1], **Arnaud Estoup**[2]

[1] *IMAG, Univ Montpellier, Cnrs, Montpellier, France*
[2] *CBGP, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France*

Abstract

Goodness-of-fit methods (GOF) aim at evaluating the level of adequacy between the observed dataset and a given model of interest, typically using an hypothesis-testing approach. In an Approximate Bayesian Computation context, this question can be re-framed as a novelty detection problem, in which one seeks to evaluate to which extent the observed dataset is an outlier compared to the simulated datasets. Many scores have been used as metrics to construct GOF test statistics and have been extensively tested in the literature. Here we propose a score based on the Local Outlier Factor.

Keywords

Generative models, goodness-of-fit, novelty detection, likelihood-free.

Network Regression and Supervised Centrality Estimation

Cai, J.[1], **Yang, D.**[2], **Zhu, W.**[3], **Shen, H.**[2], **Zhao, L.**[4]

[1] *University of Notre Dame, Department of Information Technology, Analytics, and Operations, USA*
[2] *The University of Hong Kong, Innovation and Information Management, Faculty of Business and Economics, Hong Kong, China*
[3] *Tsinghua University, Department of Finance, School of Economics and Management, China*
[4] *University of Pennsylvania, Department of Statistics and Data Science, USA*

Abstract

The centrality in a network is often used to measure nodes' importance and model network effects on a certain outcome. Empirical studies widely adopt a two-stage procedure, which first estimates the centrality from the observed noisy network and then infers the network effect from the estimated centrality, even though it lacks theoretical understanding. We propose a unified modeling framework, under which we first prove the shortcomings of the two-stage procedure, including the inconsistency of the centrality estimation and the invalidity of the network effect inference. Furthermore, we propose a supervised centrality estimation methodology, which aims to simultaneously estimate both centrality and network effect. The advantages in both regards are proved theoretically and demonstrated numerically via extensive simulations and a case study in predicting currency risk premiums from the global trade network.

Keywords

Hub centrality, Authority centrality, Measurement error, Global trade network, Currency risk premium.

Dynamic Split Random Forest

Jelena Bradic[1], **Weijie Ji**[2], **Yuqian Zhang**[3]

[1] *University of California San Diego, Department of Mathematics and Halicioglu Data Science Institute, USA, jbradic@ucsd.edu*
[2] *University of California San Diego, Department of Mathematics, USA, w6ji@ucsd.edu*
[3] *Renmin University of China, Institute of Statistics and Big Data, China, yuqianzhang@ruc.edu.cn*

Abstract

In this study, we introduce a dynamically adaptable split direction mechanism for honest random forests, which leads to significantly faster convergence rates compared to median random forests. Remarkably, our approach achieves minimax optimality for Lipschitz functions, a milestone that centered forests cannot attain when splitting directions are chosen randomly, especially in cases where the covariate vector's dimension exceeds one. Our data-dependent splitting criterion allows for the effective utilization of information from the dataset during the tree-growing process, enhancing empirical performance, albeit at the cost of greater theoretical complexity in the analysis of such data-dependent forests. We showcase applications to Average Treatment Effects to illustrate the benefits of the newly proposed method.

Keywords

Optimal minimax estimation, Average Treatment Effects, Nonparametrics.

Clustering High-Dimensional Noisy Categorical and Numerical Data with Applications in Reliability

J. Tang

Purdue University, Daniels School of Business, U.S.A.,
jtang@purdue.edu

Abstract

Clustering is a widely used unsupervised learning technique that groups subjects into homogeneous clusters based on their similarity. In this talk, we first propose a general categorical data encoding method and a computationally efficient spectral-based algorithm to cluster high-dimensional noisy categorical data (nominal or ordinal). Under a general statistical model for data, we establish a sufficient condition that ensures our algorithm's capability to exactly recover the true clusters with high probability in high dimensions. We discuss an extension of our method to handle mixed data, where some attributes are numerical, while others are categorical. To illustrate our proposed method, we apply it to the fatigue-crack-growth data from Hudak, Saxena, Bucci, and Malcolm (1978) and Bogdanoff and Kozin (1985), which was also analyzed in Lu and Meeker (1993). In our analysis, we identify two distinct clusters within the data. One cluster contains the censored samples, while the other cluster comprises uncensored samples. This is a joint work with Z. Tian (IQVIA) and J. Xu (Duke University).

Keywords

Clustering, mixed data, spectral algorithm, clustering accuracy, degradation data.

Density estimation on Lie groups in the presence of measurement error without auxiliary data

Jeong Min Jeon[1] **and Ingrid Van Keilegom**[2]

[1] *Seoul National University, Department of Statistics, South Korea, jeongmin.jeon.stat@gmail.com*
[2] *KU Leuven, ORSTAT, Belgium, ingrid.vankeilegom@kuleuven.be*

Abstract

In this talk, we introduce density estimation on a general Lie group when data contain measurement errors and the distribution of measurement error is unknown. We estimate the target density without additional observations such as an observable random sample from the measurement error distribution or repeated measurements. To achieve this, we take a semiparametric approach, which assumes that the measurement error distribution belongs to a parametric family. We also investigate a fully parametric approach for the case where the target density is also parametric. We establish an identifiability result for a measurement error model on the general Lie group and derive various asymptotic properties of our density estimators. Simulation studies are performed to demonstrate the superior performance of our estimators.

Keywords

Density estimation, Lie group, Unknown measurement error distribution.

Stochastic optimal transport in Banach spaces for regularized estimation of multivariate quantiles

Bernard Bercu[1], **Jérémie Bigot**[2], **Gauthier Thurin**[3]

[1] *Université de Bordeaux, Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), France, bernard.bercu@u-bordeaux.fr*
[2] *Université de Bordeaux, Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), France, jeremie.bigot@u-bordeaux.fr*
[3] *Université de Bordeaux, Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), France, gauthier-louis.thurin@u-bordeaux.fr*

Abstract

It is proposed to focus on recent contributions in statistics on the definition of a notion of multivariate quantiles (for random vectors) which extends the usual notion of quantile for probability measures supported on the real line. This approach is based on tools from optimal transport theory. In this framework, we introduce a new stochastic algorithm to solve the entropic optimal transport problem between two absolutely continuous probability measures. Our work is motivated by the specific framework of Monge-Kantorovich quantiles where the source measure is either the uniform distribution on the hypercube or the spherical uniform distribution. Using knowledge of the source measure, we propose to parameterize a dual Kantorovich potential by its Fourier coefficients. In this way, each iteration of our stochastic algorithm reduces to two Fourier transforms which allow us to use the fast Fourier transform in order to implement an efficient numerical method to solve the entropic optimal transport. We study the almost sure convergence of our stochastic algorithm which takes its values in an infinite dimensional Banach space. Then, using numerical experiments, we illustrate the performance of our approach on the calculation of regularized Monge-Kantorovich quantiles. In particular, we investigate the potential benefits of entropy regularization for

smooth estimation of multivariate quantiles using data sampled from the target measure.

Keywords

Entropic Optimal Transport, Monge-Kantorovich quantiles, Multivariate quantiles, Stochastic optimization in a Banach space, Multiple Fourier Series.

Private Treatment Assignment for Causal Experiments

Jeremy Seeman[1]

[1] *University of Michigan, Michigan Institute for Data Science (MIDAS) and Institute for Social Research (ISR), United States, jhseeman@umich.edu*

Abstract

Open science efforts like experimental preregistration help to improve the reproducibility and external validity of causal inferences. However, research involving human data subjects can raise numerous privacy concerns, especially in healthcare contexts where covariates contain personal health information. Existing methods at the intersection of differential privacy (DP) and causal inference traditionally focus on how to calculate causal estimates while satisfying DP. Such work fails to address how transparent experimental designs may themselves leak information about participants, as the probability of treatment assignment often depends on confidential covariate values and their dependencies amongst participants. This work provides tools to publish experimental designs for open science efforts while satisfying DP. First, we show how many experimental designs can leak information about participants. We analyze the design trade-off between observed covariate balance and robustness to worst case (or adversarial) potential outcomes, noting that the latter enables 0-DP algorithms while most designs for the former fail to satisfy DP for any finite privacy loss. We then establish DP alternatives based on discrepancy theory where covariate dependency is algorithmically captured by experimental design decisions and privacy loss bounds. Doing so ensures that privacy loss is spent efficiently relative to practitioner needs for covariate balance. Finally, we discuss how to use these experimental designs in downstream DP inferences with valid, finite sample interval estimators. Such work will allow scientists to integrate privacy protections into end-to-end open science experiments.

Keywords

Differential Privacy, Experimental Design, Causal Inference.

Data Science Ethics for Statistics Education and Practice

Jessica Utts[1]

[1] *University of California, Irvine, Statistics, USA, jutts@uci.edu*

Abstract

Statisticians have always been concerned with the ethics of our practice, but as our methods have become more complex so have our ethical issues. This talk will focus on a few of those issues, and what statisticians can do to address them. An important aspect of addressing ethical issues is incorporating them into statistics education, early and often. Part of this talk will focus on ideas for discussing ethics in class, even at the introductory statistics level.

Keywords

Ethics, Statistics education, Communicating statistics.

A Latent Space Model for Hypergraphs with Diversity and Heterogeneous Popularity

Xianshi Yu[1], **Ji Zhu**[2]

[1] *University of Michigan, Department of Statistics, U.S.A.,* *xsyu@umich.edu*
[2] *University of Michigan, Department of Statistics, U.S.A.,* *jizhu@umich.edu*

Abstract

While relations among individuals make an important part of data with scientific and business interests, existing statistical modeling of relational data has mainly been focusing on dyadic relations, i.e., those between two individuals. This work addresses the less studied, though commonly encountered, polyadic relations that can involve more than two individuals. In particular, we propose a new latent space model for hypergraphs using determinantal point processes, which is driven by the diversity within hyperedges and each node's popularity. This model mechanism is in contrast to existing hypergraph models, which are predominantly driven by similarity rather than diversity. Additionally, the proposed model accommodates broad types of hypergraphs, with no restriction on the cardinality and multiplicity of hyperedges. Consistency and asymptotic normality of the maximum likelihood estimates of the model parameters have been established. The proof is challenging, owing to the special configuration of the parameter space. Simulation studies and an application to the What's Cooking data show the effectiveness of the proposed model.

Keywords

Hypergraph embedding, determinantal point process, network analysis.

Empirical Likelihood MLE for Joint Modeling Right Censored Survival Data with Longitudinal Covariates

Jian-Jian Ren[1], **Yuyin Shi**[2]

[1] *University of Maryland, Department of Mathematics, USA,*
jjren@umd.edu
[2] *U.S. Food and Drug Administration, USA*

Abstract

Up to now, almost all existing methods for joint modeling survival data and longitudinal data rely on parametric/semiparametric assumptions on longitudinal covariate process, and the resulting inferences critically depend on the validity of these assumptions that are difficult to verify in practice. The kernel method based procedures rely on choices of kernel function and bandwidth, and none of the existing methods provides estimate for the baseline distribution in proportional hazards model. This article proposes a proportional hazards model for joint modeling right censored survival data and intensive longitudinal data taking into account of within-subject historic change patterns. Without any parametric/semiparametric assumptions or use of kernel function, we derive empirical likelihood-based maximum likelihood estimators for the regression parameter and the baseline distribution function. Also, we develop stable computing algorithms and present some simulation results. Analyses of real data sets are conducted for smoking cessation data and liver disease data.

Keywords

Empirical likelihood; intensive longitudinal data; maximum likelihood estimator; proportional hazards model; right censored data.

UTOPIA: Universally Trainable Optimal Prediction Intervals Aggregation

Jianqing Fan[1], **Jiawei Ge**[2], **Debarghya Mukherjee**[3]

[1] *Princeton University, Department of Operations Research and Financial Engineering, USA, jqfan@princeton.edu*
[2] *Princeton University, Department of Operations Research and Financial Engineering, USA, jg5300@princeton.edu*
[3] *Boston University, Department of Statistics, mdeb@umich.edu*

Abstract

Uncertainty quantification for prediction is an intriguing problem with significant applications in various fields, such as biomedical science, economic studies, and weather forecasts. Numerous methods are available for constructing prediction intervals, such as quantile regression and conformal predictions, among others. Nevertheless, model misspecification (especially in high-dimension) or sub-optimal constructions can frequently result in biased or unnecessarily-wide prediction intervals. In this work, we propose a novel and widely applicable technique for aggregating multiple prediction intervals to minimize the average width of the prediction band along with coverage guarantee, called Universally Trainable Optimal Predictive Intervals Aggregation (UTOPIA). The method also allows us to directly construct predictive bands based on elementary basis functions. Our approach is based on linear or convex programming which is easy to implement. All of our proposed methodologies are supported by theoretical guarantees on the coverage probability and optimal average length, which are detailed in this paper. The effectiveness of our approach is convincingly demonstrated by applying it to synthetic data and two real datasets on finance and macroeconomics.

Keywords

Average predictive widths, predictive interval constructions, coverage probability, basis expansions, kernel tricks, neural networks.

The statistical triangle

Jiashun Jin

Carnegie Mellon University, Department of Statistics, USA,
jiashun@stat.cmu.edu

Abstract

In his Fisher's Lecture in 1996, Efron suggested that there is a philosophical triangle in statistics with "Bayesian", "Fisherian", and "Frequentist" being the three vertices, and many representative statistical methods can be viewed as a convex linear combination of the three philosophies. We collected and cleaned a data set consisting of the citation and bibtex (e.g., title, abstract, author information) data of 83,331 papers published in 36 journals in statistics and related fields, spanning 41 years. Using the data set, we constructed 21 co-citation networks, each for a time window between 1990 and 2015. We propose a dynamic Degree-Corrected Mixed- Membership (dynamic-DCMM) model, where we model the research interests of an author by a low-dimensional weight vector (called the network memberships) that evolves slowly over time. We propose dynamic-SCORE as a new approach to estimating the memberships. We discover a triangle in the spectral domain which we call the Statistical Triangle, and use it to visualize the research trajectories of individual authors. We interpret the three vertices of the triangle as the three primary research areas in statistics: "Bayes", "Biostatistics" and "Nonparametrics". The Statistical Triangle further splits into 15 sub- regions, which we interpret as the 15 representative sub-areas in statistics. These results provide useful insights over the research trend and behavior of statisticians.

Keywords

SCORE-normalization, mixed-membership, simplex, vertex hunting, research map, research trajectory.

Estimating heritability of time-to-event traits using censored multiple variance component model

Jin Zhou[1]

UCLA[1]

Abstract

Genome-wide association studies (GWASs) have spotlighted genetic variants linked to the onset age for diseases such as Type 2 diabetes, Alzheimer's, and heart disease. Central to GWASs is heritability, which represents the proportion of phenotypic variation attributable to genetic variation. Historical approaches rooted in animal breeding are inadequate for modern human genomic data. Alternatively, heritability relies on binary disease-status traits that fail to capture the temporal aspect of disease development. In this context, our research presents the censored multiple variance component model (CMVC) based on an accelerated failure time model with syntenic variables. Designed for individual-level genotype data, it's scalable for biobank data and handles time-to-event data with random right-censoring. Simulations affirm its unbiased nature for right-censored outcomes. We offer heritability assessments for various diseases, explore per-allele effect sizes across genomic segments, and apply our methods to UK Biobank. Conclusively, our study introduces an advanced method tailored for human genomic data, shedding profound insights into the genetic determinants of disease onset and progression.

SOFARI: High-Dimensional Manifold-Based Inference

Zemin Zheng[1], **Xin Zhou**[2], **Yingying Fan**[3], **Jinchi Lv**[4]

[1] *University of Science and Technology of China, Department of Statistics and Finance, China, zhengzm@ustc.edu.cn*
[2] *University of Science and Technology of China, Department of Statistics and Finance, China, zx1120@mail.ustc.edu.cn*
[3] *University of Southern California, Data Sciences and Operations Department, USA, fanyingy@marshall.usc.edu*
[4] *University of Southern California, Data Sciences and Operations Department, USA, jinchilv@marshall.usc.edu*

Abstract

Multi-task learning is a widely used technique for harnessing information from various tasks. Recently, the sparse orthogonal factor regression (SOFAR) framework, based on the sparse singular value decomposition (SVD) within the coefficient matrix, was introduced for interpretable multi-task learning, enabling the discovery of meaningful latent feature-response association networks across different layers. However, conducting precise inference on the latent factor matrices has remained challenging due to orthogonality constraints inherited from the sparse SVD constraint. In this paper, we suggest a novel approach called high-dimensional manifold-based SOFAR inference (SOFARI), drawing on the Neyman near-orthogonality inference while incorporating the Stiefel manifold structure imposed by the SVD constraints. By leveraging the underlying Stiefel manifold structure, SOFARI provides bias-corrected estimators for both latent left factor vectors and singular values, for which we show to enjoy the asymptotic mean-zero normal distributions with estimable variances. We introduce two SOFARI variants to handle strongly and weakly orthogonal latent factors, where the latter covers a broader range of applications. We illustrate the effectiveness of SOFARI and justify our theoretical results through

simulation examples and a real data application in economic forecasting. This is a joint work with Yingying Fan, Zemin Zheng and Xin Zhou.

Keywords

Multi-task learning, Sparse SVD, Orthogonality constraints, Manifold-based inference, Neyman near-orthogonality.

Testing independence and conditional independence in high dimensions via coordinatewise Gaussianization

Jinyuan Chang[1,2], **Yue Du**[1], **Jing He**[1]

[1] *Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, China*
[2] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China*

Abstract

Testing independence and conditional independence (given a random vector $Z \in R^m$) between two random vectors $X \in R^p$ and $Y \in R^q$ are two fundamental problems in statistics. In this paper, we propose two new tests for solving these problems in high dimensions. More specifically, we first propose a coordinatewise Gaussianization procedure to transform each component of the observed data into marginally normal distributed variables, then construct the L_∞-type test statistics for these two hypothesis testing problems based on the transformed data, and finally determine the critical values by the multiplier bootstrap. The proposed tests have two advantages: (i) it does not require any moment conditions on each component of X, Y and Z, and (ii) it allows p, q and m diverging exponentially fast with the sample size. Extensive numerical simulations and a real data analysis show that the proposed tests work promisingly.

Keywords

Conditional independence test, coordinatewise Gaussianization, Gaussian approximation, high-dimensional statistical inference, independence test, multiplier bootstrap.

ELSA: Efficient Label Shift Adaptation through the Lens of Semiparametric Models

Jiwei Zhao[1]

[1] *University of Wisconsin-Madison, USA, jiwei.zhao@wisc.edu*

Abstract

We study the domain adaptation problem with label shift in this work. Under the label shift context, the marginal distribution of the label varies across the training and testing datasets, while the conditional distribution of features given the label is the same. Traditional label shift adaptation methods either suffer from large estimation errors or require cumbersome post-prediction calibrations. To address these issues, we first propose a moment-matching framework for adapting the label shift based on the geometry of the influence function. Under such a framework, we propose a novel method named Efficient Label Shift Adaptation (ELSA), in which the adaptation weights can be estimated by solving linear systems. Theoretically, the ELSA estimator is $\sqrt{n}$-consistent (n is the sample size of the source data) and asymptotically normal. Empirically, we show that ELSA can achieve state-of-the-art estimation performances without post-prediction calibrations, thus, gaining computational efficiency.

Keywords

Domain adaptation, label shift, semiparametric statistics.

Bivariate Tail Probability Approximations

J.E. Kolassa[1] **and D. Lee**[2]

[1] *Rutgers, the State University of New Jersey, Department of Statistics, USA, kolassa@stat.rutgers.edu*

[2] *Rutgers, the State University of New Jersey, Department of Statistics, USA, dl990@stat.rutgers.edu*

Abstract

This presentation extends the univariate Lugannani and Rice saddlepoint tail approximation to multiple dimensions. The resulting approximation uses a multivariate Gaussian approximation to the distribution of the signed roots of the log likelihood statistics. As in the univariate case, the next correction term involves the difference between reciprocals of the signed root of the likelihood statistics and the analogous Wald statistics. This approximation also uses the curvature of the boundary of the tail region in terms of signed root of likelihood ratio statistics. Extensions to lattice variables and conditional distributions are provided.

Keywords

Saddlepoint approximation, Bivariate tail probability, multiple comparisons.

An introduction and application of random projections

Juan A. Cuesta-Albertos

Universidad de Cantabria, Department of Matemáticas, Estadística y Computación, Spain, cuestaj@unican.es

Abstract

In the first part of the talk, I will describe the reasons that make it possible to test p-dimensional hypotheses by replacing the hypothesis to be tested by its "projected" counterpart on just one randomly chosen 1-dimensional subspace. Obviously, this procedure implies a loss of power. I will comment on the two proposed ways to alleviate this problem: 1) Use several 1-dimensional projections. 2) If the dimension is low, it is also possible to integrate over all possible projections. In the second part of the talk, I will comment on the use of the latter idea to obtain a family of uniformity tests on the p-dimensional sphere. This family includes some well-known uniformity tests, but it also allows to extend some circular tests to higher dimensions as well as to introduce some new ones. Results in the second part have been obtained in cooperation with E. García-Portugués (U. Carlos III, Spain) and P. Navarro-Esteban (U. de Cantabria, Spain).

Keywords

Random projections, Goodness of fit tests, Uniformity, p-dimensional sphere.

A projection based approach for interactive fixed effects panel data models

Georg Keilbar[1], **Juan M. Rodriguez-Poo**[2], **Alexandra Soberón**[3], **Weining Wang**[4],

[1] *Humboldt-Universität zu Berlin, Chair of Statistics, Germany, georg.keilbar@hu-berlin.de*
[2] *University of Cantabria, Department of Economics, Spain, juan.rodriguez@unican.es*
[3] *University of Cantabria, Department of Economics, Spain, alexandra.soberon@unican.es*
[4] *University of York, Department of Economics and Related Studies, UK, weining.wang@york.ac.uk*

Abstract

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

Keywords

Cross-sectional dependence, semiparametric factor models, principal components, sieve approximation, large panels.

Fourier analysis of spatial point processes

Junho Yang[1], **Yongtao Guan**[2]

[1] *Institute of Statistical Science, Academia Sinica, Taiwan, junhoyang@stat.sinica.edu.tw*
[2] *School of Data Science, Chinese University of Hong Kong–Shenzhen, China, guanyongtao@cuhk.edu.cn*

Abstract

In this presentation, we discuss a comprehensive frequency domain methods for the estimation and inference of the second-order structure of spatial point processes. The main element here is the discrete Fourier transform (DFT) of the observed point pattern and its tapered version. Under stationarity, we show that both the DFTs and the tapered DFTs, evaluated at different frequencies (which can have the same limit), are asymptotically independent centered Gaussian. Based on this result, we prove the central limit theorem for the statistics that can be written as quadratic forms of the tapered DFT. As an application, we introduce a frequency domain inferential method for stationary point processes. The resulting parameter estimator is computationally tractable and provides meaningful interpretations, even in cases where the model is mis-specified. Lastly, we extend our frequency domain methods to a class of non-stationary spatial point processes.

Keywords

Cox cluster processes, discrete Fourier transform, spatial point processes, parameter estimation.

Adaptive Merging and Efficient Estimation in Longitudinal Networks

Haoran Zhang[1] **and Junhui Wang**[2]

[1] *Department of Statistics and Data Science*
Southern University of Science and Technology
[2] *Department of Statistics*
The Chinese University of Hong Kong

Abstract

Longitudinal network consists of a sequence of temporal edges among multiple nodes, where the temporal edges are observed in real time. It has become ubiquitous with the rise of online social platform and e-commerce, but largely under-investigated in literature. In this talk, we present an efficient estimation framework for longitudinal network, leveraging strengths of adaptive network merging, tensor decomposition and point process. It merges neighboring sparse networks so as to enlarge the number of observed edges and reduce estimation variance, whereas the estimation bias introduced by network merging is controlled by exploiting local temporal structures for adaptive network neighborhood. A projected gradient descent algorithm is proposed to facilitate estimation, where the upper bound of the estimation error in each iteration is established. Theoretical analysis of the proposed method shows that it can significantly reduce the estimation error and also provides guideline for network merging under various scenarios. We further demonstrate the advantage of the proposed method through extensive numerical experiments on synthetic datasets and a militarized interstate dispute dataset.

Keywords

Dynamic network, embedding, multi-layer network, point process, tensor decomposition.

Statistical Computing, Robust Methods, and Data Displays: Critical tools for Big Data

Karen Kafadar[1], **Jordan Rodu**[2]

[1] *University of Virginia, Department of Statistics, U.S.A.,*
kk3ab@virginia.edu
[2] *University of Virginia, Department of Statistics, U.S.A.,*
kk3ab@virginia.edu

Abstract

Today's massive datasets make statistical computing and displays even more needed than when the terms "Statistical Computing" and "Statistical Graphics" evolved as disciplines 50 years ago. Because the central goals of data analysis are insight and inference, and because rarely should all data be displayed, we need algorithms and data displays that meet both these objectives. Further, 'big data' are even more likely to require robust techniques, due to exotic values, outliers, or mixtures of distributions. Finally, more data does not imply more confidence, especially when they are non-representative of their target populations. Robust statistical methods are essential to these displays, in sampling the dataset, estimating key quantities, and communicating insights and inferences. This talk will discuss some recently analyzed datasets and new displays that demonstrate that statistical methods in the 'data science' era remain critical for analyzing 'big data.'

Keywords

Classification, Sampling, Quantiles, 'Black-box' algorithms.

Doubly-robust causal inference using matching, with application to the effect of e-cigarette use on smoking cessation

Ruifeng Chen[1], **Karen Messer**[2]

[1] *Regeneron Pharmaceuticals, New York, USA, ruifench@gmail.com*
[2] *University of California San Diego, Division of Biostatistics, USA, kmesser@ucsd.edu*

Abstract

Propensity score (PS) matching can be an attractive approach to causal inference because of its potential transparency, interpretability, and robustness. PS matching is often followed by regression modeling, a practice which yields a doubly-robust estimator. The prognostic score (PGS), which is the mean outcome conditional on covariates for control subjects, can also be used for matching, and double matching on both the PS and the PGS can yield a doubly-robust estimator. We provide a concise exposition of assumptions needed for such a doubly-matched estimator to be doubly robust. We compare the performance of the doubly-matched estimator with the estimator obtained by regression modeling after PS matching, and we study methods of confidence interval construction for these estimators. We present a case study using a double-matching approach to estimate the causal effect of e-cigarettes that are used for smoking cessation among US smokers. Double matching can be a useful approach to doubly-robust estimation.

Keywords

Causal inference, matching, doubly robust.

Estimating network-mediated causal effects via spectral embeddings

Alex Hayes[1], **Mark M. Fredrickson**[2], **Keith Levin**[1]

[1] *University of Wisconsin–Madison, Department of Statistics, USA,*
{alex.hayes,kdlevin}@wisc.edu
[2] *University of Michigan, Department of Statistics, USA,*
mfredric@umich.edu

Abstract

Causal inference for network data is an area of active interest in the social sciences. Unfortunately, the complicated dependence structure of network data presents an obstacle to many causal inference procedures. We consider the task of mediation analysis for network data, and present a model in which mediation occurs in a latent embedding space. Under this model, node-level interventions have causal effects on nodal outcomes, and these effects can be partitioned into a direct effect independent of the network, and an indirect effect induced by homophily. To estimate network-mediated effects, we embed nodes into a low-dimensional space and fit two regression models: (1) an outcome model describing how nodal outcomes vary with treatment, controls, and position in latent space; and (2) a mediator model describing how latent positions vary with treatment and controls. We prove that the estimated coefficients are asymptotically normal about the true coefficients under a sub-gamma generalization of the random dot product graph, a widely-used latent space model. We show that these coefficients can be used in product-of-coefficients estimators for causal inference. Our method is easy to implement, scales to networks with millions of edges, and can be extended to accommodate a variety of structured data.

Keywords

Networks, mediation analysis, spectral methods.

Regression Models with Interval-Censored Covariates

Klaus Langohr[1], **Andrea Toloba López-Egea**[1], **Guadalupe Gómez Melis**[1]

[1] *Universitat Politècnica de Catalunya, Department of Statistics and Operations Research, Spain,* `klaus.langohr@upc.edu`

Abstract

Interval censoring is typically encountered in the analysis of times to a silent event. In these cases, the time that the event occurs, for instance, the moment of an infection with a certain virus, cannot be observed exactly. Methods to analyse such data have been extensively studied; however, scientific literature on regression models with an interval-censored covariate is scarce, because times to an event of interest are, most often, rather a study's response than one of the explanatory variables. Gómez et al. (2003) presented a linear regression model with an interval-censored covariate in the context of a clinical trial for HIV-infected persons and proposed an Expectation-Maximization (EM)-type algorithm, the so-called GEL (Gómez–Espinal–Lagakos) algorithm, to jointly estimate the model parameters and the marginal distribution of the interval-censored covariate. This algorithm was implemented in R by Langohr and Gómez (2014) and Gómez et al. (2022) applied the algorithm to generalized linear models. In this work, we present the GEL algorithm for generalized linear models and illustrate the method with a gamma regression model applied to data from a metabolomic study. Residuals for such models will be sketched and are, presently, under study.

Keywords

Generalized linear models, interval-censored covariates, metabolomic studies.

References

Gómez, G., A. Espinal, A., and S.W. Lagakos (2003). Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine*, *22*, 409–425.

Gómez, G., M. Marhuenda-Muñoz, and K. Langohr, K. (2022). Regression analysis with interval-censored covariates. Application to liquid chromatography. In: Sun J., Chen DG. (eds). *Emerging topics in modeling interval-censored survival data.* ICSA Book Series in Statistics. Springer, Cham.

Langohr, K. and Gómez, G. (2014). Estimation and residual analysis with R for a linear regression model with an interval-censored covariate. *Biometrical Journal*, *56*, 867–885.

Adaptive Linear Estimating Equations

Mufang Ying[1], **Koulik Khamaru**[1], **Cun-Hui Zhang**[1]

[1] *Rutgers University, USA*

Abstract

Sequential data collection has emerged as a widely adopted technique for enhancing the efficiency of data gathering processes. Despite its advantages, such data collection mechanism often introduces complexities to the statistical inference procedure. For instance, the ordinary least squares (OLS) estimator in an adaptive linear regression model can exhibit non-normal asymptotic behavior, posing challenges for accurate inference and interpretation. In this paper, we propose a general method for constructing debiased estimator which remedies this issue. It makes use of the idea of adaptive linear estimating equations, and we establish theoretical guarantees of asymptotic normality, supplemented by discussions on achieving near-optimal asymptotic variance. A salient feature of our estimator is that in the context of multi-armed bandits, our estimator retains the non-asymptotic performance of the least square estimator while obtaining asymptotic normality property. Consequently, this work helps connect two fruitful paradigms of adaptive inference: a) non-asymptotic inference using concentration inequalities and b) asymptotic inference via asymptotic normality.

Keywords

Sequential inference, Asymptotic normality.

Doubly Robust Sequential Quantile Off-Policy Inference

Yang Xu[1], **Chengchun Shi**[2], **Shikai Luo**[3], **Lan Wang**[4], **Rui Song**[5]

[1] *North Carolina State University, Department of Statistics, USA, yxu63@ncsu.edu*
[2] *London School of Economics and Political Science, Department of Statistics, USA, c.shi7@lse.ac.uk*
[3] *ByteDance, China, sluo198912@163.com*
[4] *University of Miami, Miami Herbert Business School, USA, lxw611@miami.edu*
[5] *Amazon, USA, songray@gmail.com*

Abstract

Sequential decision making plays crucial roles in various domains, ranging from healthcare to technology industries. We introduce a novel doubly-robust inference procedure for the quantiles of a target policy's reward in multi-stage sequential decision making problems. While previous research has focused on doubly-robust estimation of the mean reward, our work addresses two key advancements. Firstly, existing literature primarily evaluates the mean reward of a given policy, disregarding the variability of outcomes. However, in many applications, alternative criteria, such as quantile-based metrics, are more suitable, especially when the reward distribution is skewed and asymmetric. Quantile-based metrics offer robustness and interoperability. Secondly, doubly-robust inference is a more challenging problem than doubly-robust estimation and has received less attention. Even in the context of estimating mean effects, a doubly-robust estimation procedure does not automatically guarantee doubly-robust inference. To address the challenges associated with nonsmoothness and parameter-dependent nuisance functions, we leverage deep conditional generative learning methods and develop a doubly-robust estimator for the asymptotic

variance of the policy value estimator. We demonstrate the advantages of this approach through both simulations and a real-world data example from a short-video platform. In particular, we observe that the proposed estimator outperforms classical off-policy evaluation (OPE) estimators for the mean in the presence of heavy-tailed reward distributions.

Keywords

Doubly-robust inference Sequemtial decision making, Quantile.

Using Auxiliary Information in Probability Survey Data to Improve Pseudo-Weighting in Non-Probability Samples: A Copula Model Approach

Tingyu Zhu[1], **Lan Xue**[2], **Virginia Lesser**[3]

[1] *Oregon State University, Department of Statistics, USA, zhuti@oregonstate.edu*
[2] *Oregon State University, Department of Statistics, USA, xuel@oregonstate.edu*
[3] *Oregon State University, Department of Statistics, USA, Virginia.Lesser@oregonstate.edu*

Abstract

Although probability sampling has long been regarded as the gold standard for survey methods, nonprobability sampling, such as online opt-in surveys, have gained popularity due to their convenience and cost-effectiveness. However, nonprobability samples can introduce estimation bias due to the unknown nature of the underlying selection mechanism. In this talk, we present a parametric approach to integrate probability and nonprobability samples using shared ancillary variables. It assumes that the joint distribution of ancillary variables follows a latent Gaussian copula model, and logistic regression is used to model the mechanism by which population units enter the nonprobability sample. The unknown parameters in the copula model are estimated through the pseudo maximum likelihood approach, and the logistic regression model is estimated by maximizing the sample likelihood constructed from the nonprobability sample. Our simulation results demonstrate that the proposed method effectively corrects selection bias in nonprobability sample by consistently estimating the underlying inclusion mechanism. By leveraging additional information from the nonprobability sample, the combined method provides a

more efficient estimation of the population mean than using the probability sample alone. A real data application is provided to illustrate the practical use of the proposed method.

Keywords

Ancillary variable, Inclusion probability, Panel sample, Pseudo likelihood, Sample likelihood.

Estimating heterogenous spillover effects on network neighbors to identify influential and susceptible individuals

Laura Forastiere[1]

[1] *Department of Biostatistics, Yale University, USA,*
laura.forstiere@yale.edu

Abstract

With behavioral interventions, due to peer influence the intervention (e.g., training) received by a unit is likely to affect behavioral outcomes of other connected units. Under interference, spillover effects have been defined by contrasting potential outcomes under a different number of treated units or under different treatment allocations in the interference set. In contrast, we define casual estimands representing the influence effect of having one network neighbor treated vs not treated, while the treatment of other units in the interference set is fixed or randomly assigned under an allocation strategy. This can also be defined from the influencer's perspective as the effect of being treated vs not on the outcome of their neighbors. We compare the two estimands, from the influencer's or the influencee's point of view, in a finite sample and super-population perspective. Given the definition of influence effects, we are interested on its heterogeneity with respect to the characteristics of the influencer, to identify influential individuals, and those of the influencee, to identify susceptible individuals. We develop IPW estimators for average and heterogeneous influence effects, with marginal structural models for continuous covariates, and a statistical test for the heterogeneity. We then use our estimators to investigate influencers and susceptible individuals with respect to the purchase of a weather insurance for farmers in China.

Keywords

Causal inference, interference, social networks, spillover effects, heterogeneity.

Non-parametric estimation of diffusion coefficient function in certain SPDE-systems

Lauri Viitasaari[1]

[1] *Uppsala University, Department of Mathematics, Sweden,*
lauri.viitasaari@math.uu.se

Abstract

We consider a stochastic partial differential equation

$$\partial_t u = L_x u + \sigma(u)\dot{W},$$

where L_x is a suitable elliptic differential operator acting on the spatial variable x, σ is the unknown function depending on the solution u, and $\dot{W}$ is a Gaussian multiplicative noise that is white in time and possibly correlated in space. We consider the estimation of the unknown diffusion function σ based on multiple (independent) observations $u^k, k = 1, \dots, n$ of the solution u. In particular, we provide consistent estimator together with rate of convergence results.

Keywords

Non-parametric estimation, diffusion function, stochastic partial differential equations.

A quantitative Heppes Theorem and multivariate Bernoulli distributions

Ricardo Fraiman[1], **Leonardo Moreno**[2], **Thomas Ransford**[3]

[1] *Centro de Matemática, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay, rfraiman@cmat.edu.uy*
[2] *Instituto de Estadística, Departamento de Métodos Cuantitativos, FCEA, Universidad de la República, Montevideo, Uruguay, leonardomoreno@fcea.edu.uy*
[3] *Département de mathématiques et de statistique, Université Laval, Québec City (Québec), G1V 0A6, Canada, ransford@mat.ulaval.ca*

Abstract

Using some extensions of a theorem of Heppes on finitely supported discrete probability measures, we address the problems of classification and testing based on projections. In particular, when the support of the distributions is known in advance (as for instance for multivariate Bernoulli distributions), a single suitably chosen projection determines the distribution. Several applications of these results are considered.

Keywords

Classification, Discrete tomography, Heppes theorem, Multivariate Bernoulli, Random projections.

Functional Data Analysis in the Bures-Wasserstein Space

Leonardo V. Santoro[1]

[1] *Institut de Mathématiques, École Polytechnique Fédérale de Lausanne, Switzerland, leonardo.santor@epfl.ch*

Abstract

We develop a statistical framework for conducting inference on collections of time-varying covariance operators (covariance flows) over a general Hilbert space, possibly infinite dimensional. We model the intrinsically non-linear structure of covariances by means of the Bures-Wasserstein geometry, and make use of the manifold-like structure induced by this metric to gradually build up a FDA framework. We define a notion of mean of a random flow, and address questions pertaining to its existence, uniqueness, regularity and estimation. We use the tangent-bundle structure of this space to define a notion of covariance of a random flow, and develop an associated Karhunen-Loève expansion. We then treat the problem of estimation and construction of functional principal components from a finite collection of covariance flows. The use of these tools for inference is further demonstrated by a flow-on-scalar regression model. Our methods combine tools from Optimal Transport theory, Functional Analysis and Riemmanian Geometry. Our theoretical results are motivated by modern problems in functional data analysis, where one observes operator-valued random processes — for instance when analysing dynamic functional connectivity and fMRI data, or when analysing multiple functional time series in the frequency domain.

Keywords

Functional Data Analysis, Bures-Wasserstein space, Karhunen-Loève expansion.

Statistical Inference using Deep Generative Learning

Lexin Li, jointly with Chengchun Shi and others

[1] *University of California, Berkeley, Department of Biostatistics and Epidemiology, USA, e-mail: lexinli@berkeley.edu*
[2] *London School of Economics and Political Science, Department of Statistics, UK, e-mail: C.Shi7@lse.ac.uk*

Abstract

Deep generative learning has undergone remarkable growth in recent years, presenting exciting opportunities. This talk explores the applications of some cutting-edge generative tools in addressing classic statistical problems. We delve into several case studies, employing generative adversarial networks to tackle some fundamental statistical inference questions. Examples include testing for conditional independence, evaluating the Markov property in time series, and examining the structure of directed acyclic graphs. This exploration not only highlights the capabilities of these advanced tools, but also opens up new avenues of harnessing the power of deep learning for statistical problem solving.

Keywords

Deep Generative Learning; Double Robustness; Statistical Inference.

Distributed Heterogeneity Learning from Big Spatial Data

Shan Yu[1], **Guannan Wang**[2], **Lily Wang**[3]

[1] *The University of Virginia, Department of Statistics, Charlottesville, VA 22904, USA, sy5jx@virginia.edu*
[2] *William & Mary, Department of Mathematics, Williamsburg, VA 23185, USA, gwang01@wm.edu*
[3] *George Mason University, Department of Statistics, Fairfax, VA 22030, USA, lwang41@gmu.edu*

Abstract

Spatial heterogeneity is a ubiquitous challenge in social, economic, and environmental science studies. Spatially varying coefficient models are a popular and effective technique for addressing spatial regression heterogeneity. However, accounting for heterogeneity comes at the cost of reducing model parsimony. This presentation introduces a class of generalized partially linear spatially varying coefficient models that enable the inclusion of both constant and spatially varying covariate effects while balancing flexibility and parsimony. In addition, to address the challenge of large spatial datasets collected from modern technologies, we propose a novel distributed heterogeneity learning (DHL) method based on multivariate spline smoothing over a triangulation of the domain. The DHL algorithm has a simple, scalable, and communication-efficient implementation scheme that can almost achieve linear speedup. The DHL framework is theoretically supported by demonstrating the asymptotic normality of DHL linear estimators and DHL spline estimators' convergence rate equivalent to that of global spline estimators obtained from the entire dataset. The performance of the proposed DHL method is evaluated through simulation studies and analyses of the U.S. loan application data.

Keywords

Domain decomposition, Distributed learning inference, Semiparametric spatial regression, Triangulation.

Nonparametric Testing for Survival Data With Time-dependent Covariates

Ying Cui[1], **Limin Peng**[2]

[1] *Emory University, Department of Biostatistics and Bioinformatics, USA, ying.cui@emory.edu*
[2] *Emory University, Department of Biostatistics and Bioinformatics, USA, lpeng@emory.edu*

Abstract

A time-dependent covariate (e.g., time-varying treatment or exposure) is often encountered in survival studies in biomedical research. The evolving nature of the time-dependent covariate along with the survival outcome poses extra complications in the assessment of the corresponding covariate-outcome association. In this work, we propose a new nonparametric testing framework that is designed to robustly evaluate the effect of a time-dependent covariate on a survival outcome. By adopting the landmark perspective and utilizing a new interval quantile correlation index, our testing procedure does not require parametric or semiparametric modeling of the relationship between the time-dependent covariate and the time-to-event outcome, while flexibly accommodating dynamic covariate effects on the survival outcome. We provide theoretical justifications for our proposals. The new method is applied to probe the effect of time-varying feeding patterns on the pulmonary outcomes of infants with cystic fibrosis.

Keywords

Nonparametric hypothesis testing, time-dependent covariate, time-to-event outcome.

The Promises of Parallel Outcomes

Ying Zhou[1], **Dingke Tang**[2], **Dehan Kong**[3], **Linbo Wang**[4]

[1] *University of Connecticut, Department of Statistics, U.S.A., yzhou@uconn.edu*
[2] *University of Toronto, Department of Statistical Sciences, Canada, dingke.tang@mail.utoronto.ca*
[3] *University of Toronto, Department of Statistical Sciences, Canada, dehan.kong@utoronto.ca*
[4] *University of Toronto, Department of Statistical Sciences, Canada, linbo.wang@utoronto.ca*

Abstract

A key challenge in causal inference from observational studies is the identification and estimation of causal effects in the presence of unmeasured confounding. In this paper, we introduce a novel approach for causal inference that leverages information in multiple outcomes to deal with unmeasured confounding. An important assumption in our approach is conditional independence among multiple outcomes. In contrast to existing proposals in the literature, the roles of multiple outcomes in the conditional independence assumption are symmetric, hence the name parallel outcomes. We show nonparametric identifiability with at least three parallel outcomes and provide parametric estimation tools under a set of linear structural equation models. Our proposal is evaluated through a set of synthetic and real data analyses.

Keywords

Causal inference; Latent confounding; Multivariate outcome; Nonparametric identification.

Personalized Reinforcement Learning with Applications to Recommender System

Linda Zhao[1]

[1] *University of Pennsylvania, Data Science and Statistics, USA, lzhao@wharton.upenn.edu*

Abstract Reinforcement learning (RL) has achieved remarkable success across various domains; however, its applicability is often hampered by challenges in practicality and interpretability. Many real- world applications, such as in healthcare and business settings, have large and/or continuous state and action spaces and demand personalized solutions. In addition, the interpretability of the model is crucial to decision-makers so as to guide their decision-making process while incorporating their domain knowledge. To bridge this gap, we propose a personalized reinforcement learning framework that integrates personalized information into the state- transition and reward-generating mechanisms. We develop an online RL algorithm for our framework. Specifically, our algorithm learns the embeddings of the personalized state- transition distribution in a Reproducing Kernel Hilbert Space (RKHS) by balancing the exploitation-exploration tradeoff. We further provide the regret bound of the algorithm and demonstrate its effectiveness in recommender systems.

Joint work with Cai, J., Chen, R., Wainwright, M.

Keywords

Personalized Reinforcement Learning, Online Learning, Reproducing Kernel Hilbert. Space (RKHS)

Convergence rates for density trees and forests

Linxi Liu[1], **Li Ma**[2]

[1] *University of Pittsburgh, Department of Statistics, USA,*
linxi_liu@pitt.edu
[2] *Duke University, Department of Statistical Science, USA,*
li.ma@duke.edu

Abstract

Density estimation is a useful statistical tool for sketching variations of data, profiling information content, and making risk optimal decisions. It consequently plays a fundamental role in a wide spectrum of statistical analyses and applications, such as two-sample comparison, data compression, and nonparametric evaluation of disease risk. In this work, we focus on tree based methods for density estimation under the Bayesian framework, and consider the optional Pólya tree (Wong and Ma, 2010) prior or its variations on individual trees. First we show that the density tree can achieve minimax (up to a logarithmic term) convergence over the anisotropic Besov class, which implies that the method can adapt to spatially inhomogeneous features of the underlying density function, and can achieve fast convergence as the dimension increases. We also introduce a novel Bayesian model for density forests, and show that for a class of Hölder continuous functions, such type of density forests can achieve faster convergence than trees. The convergence rate is adaptive in the sense that to achieve such a rate we do not need any prior knowledge of the smoothness level of the density. The Bayesian framework naturally endows a stochastic search scheme over either the tree space or the forest one. For both density trees and forests, we provide several numerical results to illustrate their performance in the moderatel high-dimensional case.

Keywords

Multivariate density estimation, convergence rate, spatial adaptation.

LassoNet: A Neural Network with Feature Sparsity

Ismael Lemhadri[1], Feng Ruan[2], Louis Abraham[3], Robert Tibshirani[4]

[1]*Department of Statistics, Stanford University, Stanford, U.S.A., lemhadri@stanford.edu*
[2] *Department of Statistics, University of California, Berkeley, USA, fengruan@berkeley.edu*
[3] *Université Paris 1 Panthéon-Sorbonne, Paris, France, louis.abraham@yahoo.fr*
[4] *Departments of Biomedical Data Sciences, and Statistics, Stanford University, Stanford, U.S.A., tibs@stanford.edu*

Abstract

Much work has been done recently to make neural networks more interpretable, and one approach is to arrange for the network to use only a subset of the available features. In linear models, Lasso (or ℓ_1-regularized) regression assigns zero weights to the most irrelevant or redundant features, and is widely used in data science. However the Lasso only applies to linear models. Here we introduce LassoNet, a neural network framework with global feature selection. Our approach achieves feature sparsity by adding a skip (residual) layer and allowing a feature to participate in any hidden layer only if its skip-layer representative is active. Unlike other approaches to feature selection for neural nets, our method uses a modified objective function with constraints, and so integrates feature selection with the parameter learning directly. As a result, it delivers an entire regularization path of solutions with a range of feature sparsity. We apply LassoNet to a number of real-data problems and find that it significantly outperforms state-of-the-art methods for feature selection and regression. LassoNet uses projected proximal gradient descent, and generalizes directly to deep networks. It can be implemented by adding just a few lines of code to a standard neural network.

Keywords

Neural Networks, Feature Selection, Strong Hierarchy, Proximal Gradient Descent.

`clustglm` and `clustord`: R packages for clustering with covariates for binary, count, and ordinal data

Louise McMillan[1], **Daniel Fernández**[2], **Shirley Pledger**[3], **Richard Arnold**[3], **Ivy Liu**[3], **Murray Efford**[4]

[1] *Victoria University of Wellington, School of Mathematics and Statistics, New Zealand, louise.mcmillan@vuw.ac.nz*
[2] *Universitat Politècnica de Catalunya–BarcelonaTech, Department of Statistics and Operations Research, Spain*
[3] *Victoria University of Wellington, School of Mathematics and Statistics, New Zealand*
[4] *ORCID: 0000-0001-5231-5184*

Abstract

We present two R packages for model-based clustering with covariates. Both packages can perform clustering and biclustering (clustering sites and species simultaneously, for example). Both use likelihood-based methods for clustering, which enables users to compare models using AIC and BIC as measures of relative goodness of fit. The models implemented in both packages use linear predictor terms, and so look more like regression models than clustering models. This allows for the incorporation of regression-style covariates alongside clustering effects, and enables the use of models suitable for non-continuous data. Both `clustglm` and `clustord` can include the effects of numerical or categorical covariates alongside cluster effect, or can fit pattern-detection models that include individual-level effects alongside cluster effects. For example, when applied to presence/absence data, you can cluster sites and species while also taking into account any single-species effects, and any additional covariates. `clustglm` implements techniques from Pledger and Arnold (2014) for handling binary and count data. It leverages `glm` and can accommodate balanced and non-balanced designs. It provides the clustering equivalent of biplots, and also profile

plots. `clustord` handles ordinal categorical data, using techniques outlined in Matechou et al. (2016), Fernández et al. (2016) and Fernández et al. (2019). It can fit the proportional odds model or the ordered stereotype model, a more flexible model whose fitted parameters can reveal when two ordinal categories are effectively equivalent to each other. We will illustrate the use of `clustglm` and `clustord` with a selection of ecological or medical datasets, though they are even more widely applicable than that. For example, `clustord` is useful for analysing Likert-scale survey data, giving participants' answers on a scale from 1 to 5.

Keywords

Clustering, model-based clustering, categorical data, binary data, count data.

Floodgate: A Swiss Army Knife for Regression Inference

Lu Zhang[1], **Lucas Janson**[1]

[1] *Harvard University, Department of Statistics, USA,*
ljanson@fas.harvard.edu

Abstract

A recently introduced method called floodgate can provide asymptotic inference for the importance of a covariate in a (possibly high-dimensional) regression. The measure of importance that serves as the inferential target is interpretable yet completely model-free, capturing arbitrary nonlinearities and interactions in the conditional relationship between a covariate and the response given the other covariates. The floodgate method is based on the novel idea of a floodgate function, which gives a flexible deterministic yet unobservable lower-bound for the inferential target, but is much easier to provide inference for than the original target. This talk will generalize the floodgate approach to infer many other low-dimensional quantities of interest for high-dimensional regression applications, such as the quality of a dimensionality reduction or representation of the covariates, the heterogeneity of a treatment effect or interaction between two covariates, the non-monotonicity of the regression function, or the maximum or minimum of the regression function.

Keywords

Interactions, heterogeneity, non-linearity, shape constraints, feature engineering.

Inference for Latent Variable Interpretations, with Application to Single-Cell RNA Sequencing Data

Anna Neufeld[1], Lucy Gao[2], Joshua Popp[3], Alexis Battle[4], Daniela Witten[5]

[1] *Fred Hutchinson Cancer Center, Public Health Sciences Division, USA, aneufeld@fredhutch.org*
[2] *University of British Columbia, Statistics, Canada, lucygao@ubc.ca*
[3] *Johns Hopkins University, Biomedical Engineering, USA, jpopp4@jhu.edu*
[4] *Johns Hopkins University, Biomedical Engineering, USA, USA, ajbattle@jhu.edu*
[5] *University of Washington, Statistics and Biostatistics, USA, dwitten@uw.edu*

Abstract

Latent variable estimation methods (e.g. clustering and principal component analysis) are ubiquitously applied across fields to preprocess, compress, visualize, and explore large data sets. It is often of importance and interest to interpret the estimated latent variable. While this problem is of broad importance in many application areas, we will focus on a specific example in this talk: the analysis of single-cell RNA sequencing data, where researchers often first characterize the variation between cells by estimating a latent variable the cells, then test each gene for association with the estimated latent variable, hoping to use the information from the test of association to lend support to interpreting the estimated latent variable as specific cell types or position along a specific developmental trajectory. If the same data are used for both latent variable estimation and inference, then standard methods for computing p-values and confidence intervals will fail to achieve statistical guarantees such as Type 1 error control or nominal coverage. Furthermore, sample splitting cannot be fruitfully applied

to address the problem. Instead, we apply data thinning (Neufeld et al. 2023, Dharamshi et al. 2023), an operation that decomposes a realization of a random variable into two or more independent pieces, to the data from each cell. This enables valid inference in this setting, under a Poisson or negative binomial assumption. The general sketch of our approach to the problem is applicable to data coming from many distributional families, including Gaussian, gamma, binomial distributions.

Keywords

Latent variable, unsupervised learning, selective inference.

RM-SMOTE: A robust balancing technique

M. Rosário Oliveira[1], **Rasool Taban**[1], **Cláudia Nunes**[1]
[1] *CEMAT and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Portugal,*
rosario.oliveira@tecnico.ulisboa.pt, rasooltaban@tecnico.ulisboa.pt, cnunes@math.tecnico.ulisboa.pt

Abstract

Imbalanced data is a common characteristic in classification problems. The imbalance of the data may lead to biased results, due to the difficulty of learning classes with few observations, called Minority classes. Balancing techniques, and more specifically oversampling techniques, are a common strategy to overcome imbalanced data situations, where synthetic Minority class observations are generated to balance the dataset. However, if existing, atypical observations can be amplified by the oversampling technique and have a stronger and even more disturbing impact on the balanced dataset resulting in biased and distorted classification results. Being so, a robust approach must be considered. In this talk, we propose a robust approach to imbalanced learning, called RM-SMOTE, that automatically accommodates potential atypical Minority class observations, making use of robust Mahalanobis distance (Taban et al., 2023). The performance of the RM-SMOTE is evaluated using several simulation scenarios, with different levels of contamination, and also using a set of imbalanced benchmark datasets. The results indicate the superiority of RM-SMOTE when handling different proportions of outliers while balancing the dataset. In cases where the observations are not linearly separable, this superiority is more significant. In the case of the benchmark datasets, the results support the statistical superiority of RM-SMOTE compared to the state-of-the-art balancing techniques.

Keywords

Imbalanced classification, robust Mahalanobis distance, oversampling, SMOTE, minimum covariance determinant estimator.

Acknowledgements: This work was supported by Fundação para a Ciência e Tecnologia, Portugal, through the project UIDB/04621/2020 and BIGMATH, from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 812912.

References

Taban, R. Nunes, C. and Oliveira, M.R. (2023). RM-SMOTE: A new robust balancing technique. *PREPRINT (Version 1) available at Research Square* `https://doi.org/10.21203/rs.3.rs-3256245/v1`.

Estimation of risk measures at extreme levels: an overview

M. Manuela Neves[1], **Clara Cordeiro**[2], **Dora Prata Gomes**[3]

[1] *Instituto Superior de Agronomia and CEAUL, Universidade de Lisboa, Portugal, manela@isa.ulisboa.pt*
[2] *Faculdade de Ciências e Tecnologia, Universidade do Algarve, and CEAUL, Universidade de Lisboa, Portugal, ccordei@ualg.pt*
[3] *NOVA School of Science and Technology (FCT NOVA), and Center for Mathematics and Applications (CMA), FCT NOVA, Portugal, dsrp@fct.unl.pt*

Abstract

Climate change brings with it new risks to many areas such as hidrology, the environment and finance and can have catastrophic impacts on everyday life. One of the standard approaches to studying risks is Extreme Value Theory (EVT). A set of procedures for estimating risk measures related to extreme events such as quantiles or conditional tail expectations are reviewed in this work. These procedures involve EVT parameter estimation jointly with adequate time series modeling in the presence of heavy tails. EVT statistical inference is performed through parametric and non-parametric approaches. The estimation of risk measures through EVT, time series modeling and resampling procedures reavealed a good compromise to obtain more efficient results. Our approach will be applied to a real dataset using R software.

Keywords

Extreme value theory, risk measures, semiparametric estimation, time series.

Acknowledgements

Manuela Neves and Clara Cordeiro are partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020. Dora Prata Gomes is financed by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UIDB/00297/2020 and UIDP/00297/2020 (Center for Mathematics and Applications).

Test and Visualization of Covariance Properties for Multivariate Spatio-Temporal Random Fields

Huang Huang[1], **Ying Sun**[1], **Marc G. Genton**[1]

[1] *King Abdullah University of Science and Technology (KAUST), Statistics Program, Saudi Arabia, huang.huang@kaust.edu.sa, ying.sun@kaust.edu.sa, marc.genton@kaust.edu.sa*

Abstract

The prevalence of multivariate space-time data collected from monitoring networks and satellites, or generated from numerical models, has brought much attention to multivariate spatio-temporal statistical models, where the covariance function plays a key role in modeling, inference, and prediction. For multivariate space-time data, understanding the spatio-temporal variability, within and across variables, is essential in employing a realistic covariance model. Meanwhile, the complexity of generic covariances often makes model fitting very challenging, and simplified covariance structures, including symmetry and separability, can reduce the model complexity and facilitate the inference procedure. However, a careful examination of these properties is needed in real applications. In the work presented here, we formally define these properties for multivariate spatio-temporal random fields and use functional data analysis techniques to visualize them, hence providing intuitive interpretations. We then propose a rigorous rank-based testing procedure to conclude whether the simplified properties of covariance are suitable for the underlying multivariate space-time data. The good performance of our method is illustrated through synthetic data, for which we know the true structure. We also investigate the covariance of bivariate wind speed, a key variable in renewable energy, over a coastal and an inland area in Saudi Arabia.

Keywords

Bivariate Wind Vector, Functional Boxplot, Multivariate Spatio-Temporal Data, Rank-Based Test, Separability, Symmetry.

Nonparametric Measure-Transportation-Based Multiple-Output Center-Outward Quantile Regression

Marc Hallin

ECARES and Department of Mathematics
Université libre de Bruxelles, Belgium

Abstract

Exploiting novel measure-transportation-based concepts of multivariate quantiles (Hallin et al., *Annals of Statistics* 49, 1139–1165 (2021)), we are considering the problem of nonparametric multiple-output quantile regression. Our approach defines nested *conditional center-outward quantile regression contours* and *regions* with given conditional probability content irrespective of the underlying distribution; their graphs constitute nested *center-outward quantile regression tubes*. Empirical counterparts of these concepts are constructed, yielding interpretable empirical regions and contours that are shown to consistently reconstruct their population versions in the Pompeiu-Hausdorff topology. Our method is entirely non-parametric and performs well in simulations including heteroskedasticity and nonlinear trends; its potential as a data-analytic tool is illustrated on some real datasets.

Based on joint work with Tasio del Barrio and Alberto González-Sanz.

Keywords

Multivariate quantiles, Measure transportation, Nonparametric regression, Quantile regression.

References

del Barrio, E., González-Sanz, A., and Hallin, M. (2022). Nonparametric multiple-output center-outward quantile regression, arXiv.2204.11756.

Segmenting toroidal time-series by inhomogeneous hidden semi-Markov models

Francesco Lagona[1] **and Marco Mingione**[1]

[1] *Roma Tre University, Department of Political Sciences, Italy,*
francesco.lagona@uniroma3.it
[2] *Roma Tre University, Department of Political Sciences, Italy,*
marco.mingione@uniroma3.it

Abstract

A nonhomogeneous hidden semi-Markov model is proposed to segment toroidal time series according to a finite number of latent regimes and, simultaneously, estimate the influence of time-varying covariates on the process' survival under each regime. The model is a mixture of toroidal densities, whose parameters depend on the evolution of a semi-Markov chain, which is in turn modulated by time-varying covariates through a proportional hazards assumption. Parameter estimates are obtained using an EM algorithm that relies on an efficient augmentation of the latent process. The proposal is illustrated on a time series of wind and wave directions recorded during winter.

Keywords

Circular data, dwell times, hidden semi-Markov model, proportional hazards.

On the risk of cancer recurrence based on tumor's clonal structure

Marek Kimmel[1]

[1] *Rice University, Departments of Statistics and Bioengineering, USA, kimmel@rice.edu*

Abstract

One of the tenets of cancer prevention is that early detection of cancer is reducing mortality. We examine this question in view of what is known about the clonal structure of tumors, using lung and bladder cancers as examples. Mathematical and biological details are largely based on [Dinh et al. Statistical Science 2020] and [Bondaruk et al. iScience 2022]. Briefly, analysis of tumor genomes, using models based on coelascent theory, may allow estimation of the growth rates of distinct clones. This in turn may help classify the early-detected tumors into these that will recur and those that will not.

Keywords

coalescent, Griffiths-Tavare model, DNA sequencing, risk of cancer recurrence.

Using a parametric model to improve nonparametric density estimation on the sphere

María Alonso-Pena[1,2], **Gerda Claeskens**[1,4], **Irène Gijbels**[3,4]

[1] *KU Leuven, ORSTAT, Belgium,*
maria.alonsopena@kuleuven.be, gerda.claeskens@kuleuven.be
[2] *University of Granada, IMAG, Spain*
[3] *KU Leuven, Department of Mathematics, Belgium*
irene.gijbels@kuleuven.be
[4] *KU Leuven, LStat, Belgium*

Abstract

We consider the problem of estimating, nonparametrically, the density function of a hyperspherical variable, by including a parametric model as guide. This guide allows us to reduce the bias of the nonparametric estimator, while the variance stays asymptotically the same, as long as the guide is not too far from the true model. In addition, we show how, with some specific kernel functions, the parametrically guided estimator performs at least as well as the classical kernel density estimator even if the guide is very far from the true density. On top of deriving the asymptotic properties of the new estimator, we show some simulations that support these results and illustrate the use of the estimator with real datasets.

Keywords

Hyperspherical data, Kernel density estimation, Parametric guide.

Poisson Kernel-Based Clustering on the d-dimensional Sphere: Convergence Properties, Identifiability and Methods of Sampling

Marianthi Markatou[1]

[1] *University at Buffalo, Department of Biostatistics, USA, markatou@buffalo.edu*

Abstract

Many scientific fields produce data that are directional in nature and can be analyzed as unit vectors on the d-dimensional sphere. Specific examples include text mining, in particular clustering of documents, gene expression analysis and the study of comets. Model-based clustering methods for directional data are proposed in the literature. These methods use mixtures of von Mises-Fisher distributions (vMF), inverse stereographic projections of multivariate normal and Watson distributions. A limitation of these algorithms is that there are no closed form expressions for the estimates of some of the parameters of the mixture models on which the algorithms are based. We present a clustering method based on mixtures of Poisson Kernel-Based Densities (PKBD) on the d-dimensional sphere. We first discuss the relationship of PKBDs (or exit on the sphere densities) with the Brownian motion and other, defined on the d-dimensional sphere, densities, and investigate the identifiability of various forms of the PKBD mixture model. We then prove convergence of the generalized EM algorithm, and study its operational characteristics. Finally, we briefly discuss sampling methods from the PKBD densities. Our experimental results show that Poisson Kernel-Based clustering exhibits superior performance over other clustering algorithms in the presence of noise in the data, when performance is evaluated by Adjusted Rand Inder (ARI), macro-precision and macro-recall. An additional advantage is that no approximations are needed to estimate the parameters. The algorithm

is implemented in the package QuadratiK (Saraceno et.al., 2023), in the function "pkbc".

Keywords

Clustering, Directional Data, Poisson Kernel-Based Density, Spherical Data.

Construction of an intelligent based CT-scan model to predict response of asthmatic patient

Marie-Félicia Béclin[1], **Pierre Lafaye de Micheaux**[2], **Nicolas Molinari**[3]

[1] *IDESP, Université de Montpellier, PreMeDICaL, Inria-Inserm, marie-felicia.beclin@umontpellier.fr*
[2] *IDESP, Université Paul Valéry Montpellier, PreMeDICaL, Inria-Inserm, pierre.lafaye-de-micheaux@univ-montp3.fr*
[3] *IDESP, Université de Montpellier, PreMeDICaL, Inria-Inserm, nicolas.molinari@inserm.fr*

Abstract

The objective is to ascertain the efficacy of Benralizumab, a medication to treat asthma. Practicians rely on specific biomarkers and clinical data. The challenge is to devise an informative feature derived from medical imaging to evaluate treatment response and predict patient's response. The images are thoracic scans in expiration and inspiration before and after one year of treatment. The hypothesis is that patients with improved conditions, will exhibit enhanced expiration scans after treatment. It is manifested by higher Hounsfield Unit values. This improvement is indicated by a shift to the right in the histograms between the pre-and post-treatment images. We construct a model mimicking the classical linear regression model, based on the histograms. From histograms, quantiles are computed for the regression approach. The method introduced by Irpino and Verde[1] do not propose confidence interval and law of the estimators. So, we propose a way to define estimators by maximum of likelihood, law of the estimators and confidence interval. This approach has limitations, including the loss of spatial information and the assumption of linear relationships between voxel distributions. Investigation is needed to develop a more general distribution-on-distribution regression method, such as

the work by Chen and Ghodrati and the work of Panaretos. Another limitation is the inability to incorporate clinical covariates. Ongoing research aims to predict 2D-histograms post-treatment from scans in inspiration and expiration after registration, along with corresponding pre-treatment histograms, while including scalar covariates.

Keywords

Distribution on Distribution Regression, Imaging-derived biomarker, Treatment prediction, Histograms.

Combining classification algorithms with pre- and post-processing techniques to handle imbalanced data for an accurate screening of familial hypercholesterolemia

M. Antunes[1], J. Albuquerque[1], A.M. Medeiros[2], A.C. Alves[2], M. Bourbon[2]

[1] *CEAUL, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal, marilia.antunes@ciencias.ulisboa.pt*
[2] *Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa, Portugal*

Abstract

Familial Hypercholesterolemia (FH) is an inherited disorder of cholesterol metabolism that can be diagnosed through a genetic test. Current criteria for FH screening, such as the Simon Broome (SB) criteria, lead to high false positive rates, implying significant expenditure. The aim of this work was to explore alternative procedures for FH screening based on different biological and biochemical indicators that could be cost-effective and easily understood and used in clinical practice. For this purpose, logistic regression (LR), the naive Bayes classifier (NB), random forest (RF), and extreme gradient boosting (XGB) algorithms were combined with Synthetic Minority Oversampling Technique (SMOTE) or threshold adjustment by maximizing the Youden index (YI) and compared. The performance of the models was assessed through a 10×10 repeated k-fold cross-validation procedure. The LR model demonstrated overall better performance, as evidenced by the areas under the receiver operating characteristics (AUROC) and precision-recall (AUPRC) curves, as well as several operating characteristics (OC), regardless of the strategy used to address class imbalance. When either data processing technique was adopted, significantly higher accuracy (Acc), G-mean, and F1 score values were found for all classification algorithms compared to the SB

criteria ($p < 0.01$), indicating a more balanced predictive ability for both classes and higher effectiveness in classifying FH patients. The adjustment of the cutoff values through pre- or post-processing methods revealed a considerable gain in sensitivity (Sens) values ($p < 0.01$). Although the performance of pre- and post-processing strategies was similar, SMOTE does not cause the model's parameters to lose interpretability. These results suggest that an LR model combined with SMOTE can be an optimal approach for use as a widespread screening tool. As a result of this work, a web-based app was developed and can be used to assist clinicians in their decision to recommend the genetic test.

Keywords

Extreme gradient boosting, logistic regression, naive Bayes, random forest, synthetic minority oversampling technique.

Acknowledgement

The current work was supported by the programme Norte2020 [Grant Number NORTE-08-5369-FSE-000018] and by Fundação para a Ciência e Tecnologia (FCT) [Grant Number UID/MAT/00006/2020].

Awareness and maturity in Big Data initiatives: atypical behaviour in latent trait models through Bayesian IRT

Mario Angelelli[1,2,*], **Massimiliano Gervasi**[2], **Enrico Ciavolino**[1,2]

[1] *University of Salento, Department of Human and Social Sciences, Italy*
[2] *CAMPI - Centre for Applied Mathematics and Physics for Industry, Lecce, Italy*
[*] *mario.angelelli@unisalento.it*

Abstract

The evaluation of latent traits is a fundamental objective in psychometry and soft metrology, with the aim of defining mathematical and statistical methods to measure abstract constructs and latent variables. When the subject of measurement is a concept rather than a tangible entity, uncertainty becomes a central concern. This paper discusses two approaches to exploring uncertainty in psychometric measurements: inconsistencies in preferences and atypical behaviours. First, we introduce a method based on information decomposition that combines multiple probability distributions to describe deviations from rational behaviour formalised by Luce's axioms. We test this approach with Ellsberg's paradox, a classic case of decision-making under ambiguity. Second, we investigate the symmetries (translations and permutations) underlying the previous method in relation to identifiability issues in a Bayesian framework. Specifically, Item Response Theory (IRT) and Graded Response Models (GRM) are used to specify the latent traits that explain individuals' responses through a Bayesian hierarchical model. To enhance identifiability and detection of atypical behaviours in responses, we explore the use of global-local priors for latent traits. Finally, we present preliminary evidence of the proposed Bayesian model's application in the management of big data

initiatives, where dedicated measurement tools (maturity models) are designed to recognise the aware adoption of emerging technologies and leverage atypical behaviours as potential sources of innovation.

Keywords

Item Response Theory, Global-Local, Information decomposition, Uncertainty modelling, Ambiguity.

Telling Cause From Effect for Categorical Variables

Mário A. T. Figueiredo

Instituto de Telecomunicações and Lisbon ELLIS Unit
Instituto Superior Técnico, Universidade de Lisboa, Portugal
mario.figueiredo@tecnico.ulisboa.pt

Abstract

Causal discovery addresses the problem of identifying cause-effect relationships between variables. Although, in principle, finding these causal relationships requires interventions, this is often impossible, impractical, or unethical, which has stimulated much research on causal discovery from purely observational data. Arguably, the simplest instance of this class of problems is to distinguish cause from effect from observations of a pair of random variables. Most proposals to handle this task, most notably *additive noise models* (ANM), are only adequate for quantitative data, since, as the name implies, they rely on a notion of addition. Consequently, these methods cannot be used when the variables are categorical. In this talk, I will begin by briefly reviewing the cause-effect problem and several approaches that have been proposed to tackle it. I will then describe our recent proposal of a method to address this problem with categorical variables (living in sets with no meaningful order). The method is based on an information-theoretic view of conditional probability mass functions as discrete memoryless channels. Specifically, we propose to select as the most likely causal direction that for which the corresponding channel is closer to being a so-called *uniform channel* (UC). The rationale is that, in a UC, as in an ANM, the conditional entropy (of the effect given the cause) does not depend on the cause distribution, in agreement with one of the foundational principles of causal discovery: *independence of cause and mechanism*. The proposed approach, which

is supported by an identifiability proof, can thus be seen as extending the ANM rationale to categorical variables. Finally, the proposed method compares favorably with recent state-of-the-art alternatives in experiments on synthetic, benchmark, and real data.

Keywords

Causal discovery, cause-effect problem, categorical variables, additive noise models, communication channels, entropy, independence of cause and mechanism.

On Lamperti transformation and characterizations of discrete random fields

Marko Voutilainen[1], **Lauri Viitasaari**[2], **Pauliina Ilmonen**[3]

[1] *University of Turku, Department of Accounting and Finance, Finland, mtvout@utu.fi*
[2] *Uppsala University, Department of Mathematics, Sweden, lauri.viitasaari@math.uu.se*
[3] *Aalto University, Department of Mathematics and Systems Analysis, Finland, pauliina.ilmonen@aalto.fi*

Abstract

We discuss vector-valued fields indexed by the set of N-dimensional integers. We introduce Lamperti transformation providing a one-to-one correspondence between strictly stationary fields and Θ-self-similar fields, where Θ is a N-tuple of commuting positive definite matrices. Moreover, we give a one-to-one correspondence between Θ-self-similar fields and a class of stationary rectangular increment fields denoted by $\mathcal{G}_\Theta$. As a corollary, the composition of these two transformations is a bijection between stationary fields and fields in $\mathcal{G}_\Theta$. We also characterize strictly stationary fields via AR(1) type of equations, where the noise process belongs to $\mathcal{G}_\Theta$. The connection between the noise and the stationary solution is given by the above-mentioned bijection. In addition, we discuss how our approach can be applied in order to define stationary fractional Ornstein-Uhlenbeck fields in $\mathbb{Z}^N$.

Keywords

Stationary fields, self-similar fields, fractional Ornstein-Uhlenbeck fields, Lamperti transformation.

Identifying Brain Tumor Gene Signatures through Multi-Omics Network Inference and Classification

Marta B. Lopes[1,2,3], **Roberta Coletti**[2], **João Carrilho**[1]

[1]*NOVA School of Science and Technology (NOVA FCT), Caparica, Portugal*
[2]*Center for Mathematics and Applications (NOVA Math), NOVA FCT, Caparica, Portugal*
[3]*UNIDEMI, Department of Mechanical and Industrial Engineering, NOVA FCT, Caparica, Portugal*
marta.lopes@fct.unl.pt; roberta.coletti@fct.unl.pt; jf.carrilho@campus.fct.unl.pt

Abstract

Gliomas are primary brain tumors known for their inter and intra-tumoral heterogeneity nature, and generally poor prognosis. Ensuring precise diagnosis and effective treatment of glioma patients is of paramount importance in managing these tumors effectively. The analysis of multi-omics datasets is expected to uncover molecular changes, offering promising molecular candidates for glioma diagnosis and therapeutic targets. Identifying novel cancer diagnostic biomarkers and understanding the intricate role they play within the broader molecular network is an active research front in cancer. However, these goals face significant challenges stemming from the high-dimensionality inherent to multi-omics datasets, which can be overcome by employing sparsity-inducing methods. Graphical methods for network inference, such as graphical lasso (Friedman et al., 2008) and joint graphical lasso (Danaher et al., 2014), estimate undirected relations encouraging edge-sparsity through a lasso regularization term. Also, model regularization approaches for feature selection based on the Elastic net penalty (Tay et al., 2023; Friedman et al, 2010) have been proposed to select

relevant subsets of features in multi-class problems. Based on RNA-sequencing and DNA-methylation glioma datasets available from The Cancer Genome Atlas (TCGA) data portal, we designed a pipeline for multi-omics biomarker discovery via sparse network inference and classification. After grouping the patients into the three main glioma types, the methodology encompasses two steps: i) the two omics are integrated in a network-based variable selection procedure performed by joint and classical graphical lasso; ii) the diagnostic role of the identified features is investigated through regularized multinomial logistic regression with the Elastic net penalty. Our study identified potential multi-omics biomarkers specific to each glioma type, with the majority of these markers having been previously reported in the literature as associated with glioma or cancer. Subsequent investigations are required to validate these findings in biological models.

Keywords

Glioma, networks, multi-omics, graphical lasso, regularized classification.

Acknowledgements

These results obtained are based on data from the TCGA Research Network (https://www.cancer.gov/tcga). This work was funded by the Portuguese Foundation for Science and Technology (FCT, I.P.), as part of the MONET project (PTDC/CCI-BIO/4180/2020), and through the references UIDB/00297/2020 and UIDP/00297/2020 (NOVA Math), UIDB/00667/2020 and UIDP/00667/2020 (UNIDEMI), and CEECINST/00042/2021.

References

Friedman, J.; Hastie, T.; and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, v.9, n.3, 432–441.

Danaher, P.; Wang, P.; and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, v.76, n.2, 373–397.

Tay, J.K.; Narasimhan, B., and T. Hastie (2023). Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software*, v.106, n.1, 1–31.

Friedman, J.H.; Hastie, T., and R. Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, v.33, n.1, 1–22.

On the Pointwise Behavior of Recursive Partitioning and Its Implications for Heterogeneous Causal Effect Estimation

Matias D. Cattaneo[1], **Jason M. Klusowski**[2], **Peter M. Tian**[3]

[1] *Princeton University, ORFE, USA, cattaneo@princeton.edu*
[2] *Princeton University, ORFE, USA, jason.klusowski@princeton.edu*
[3] *Two Sigma, USA, ptian@princeton.edu*

Abstract

Decision tree learning is increasingly being used for pointwise inference. Important applications include causal heterogenous treatment effects and dynamic policy decisions, as well as conditional quantile regression and design of experiments, where tree estimation and inference is conducted at specific values of the covariates. In this paper, we call into question the use of decision trees (trained by adaptive recursive partitioning) for such purposes by demonstrating that they can fail to achieve polynomial rates of convergence in uniform norm, even with pruning. Instead, the convergence may be poly-logarithmic or, in some important special cases, such as *honest* regression trees, fail completely. We show that random forests can remedy the situation, turning poor performing trees into nearly optimal procedures, at the cost of losing interpretability and introducing two additional tuning parameters. The two hallmarks of random forests, subsampling and the random feature selection mechanism, are seen to each distinctively contribute to achieving nearly optimal performance for the model class considered.

Keywords

Recursive partitioning, decision trees, random forests, pointwise estimation, causal inference, heterogeneous treatment effects.

On the possibility of doubly robust root-n inference

Matteo Bonvini[1], **Edward H. Kennedy**[2]

[1] *Rutgers University, Department of Statistics, U.S.A., mb1662@stat.rutgers.edu*
[2] *Carnegie Mellon University, Department of Statistics & Data Science, U.S.A, edward@stat.cmu.edu*

Abstract

We study the problem of constructing an estimator of the average treatment effect (ATE) that exhibits doubly-robust asymptotic linearity (DR-AL). This is a stronger requirement than doubly-robust consistency. In fact, a DR-AL estimator can yield asymptotically valid Wald-type confidence intervals even in the case when the propensity score or the outcome model is inconsistently estimated. On the contrary, the celebrated doubly-robust, augmented-IPW estimator requires consistent estimation of both nuisance functions for root-n inference. Previous authors have considered this problem and provided sufficient conditions under which their proposed estimators are DR-AL. Such conditions are typically stated in terms of "high-level nuisance error rates" needed for root-n inference. In this paper, we build upon their work and establish sufficient and more explicit smoothness conditions under which a DR-AL estimator can be constructed. We also consider the case of slower-than-root-n convergence rates and clarify the connection between DR-AL estimators and those based on higher-order influence functions. We complement our theoretical findings with simulations.

Keywords

Bias correction, semiparametric efficiency theory, observational study.

Approximate filtering via discrete dual processes

Kon Kam King, G.[1], **Pandolfi, A.**[2], **Piretto, M.**[3], **Ruggiero, M.**[4]

[1] *Université Paris-Saclay - INRAE, France*
[2] *Bocconi University, Italy*
[3] *BrandDelta, UK*
[4] *University of Torino and Collegio Carlo Alberto, Italy*

Abstract

We consider the task of filtering a dynamic parameter evolving as a diffusion process, given data collected at discrete times from a likelihood which is conjugate to the marginal law of the diffusion, when a generic dual process on a discrete state space is available. Recently, it was shown that duality with respect to a death-like process implies that the filtering distributions are finite mixtures, making exact filtering and smoothing feasible through recursive algorithms with polynomial complexity in the number of observations. Here we provide general results for the case of duality between the diffusion and a regular jump continuous-time Markov chain on a discrete state space, which typically leads to filtering distribution given by countable mixtures indexed by the dual process state space. We investigate the performance of several approximation strategies on two hidden Markov models driven by Cox?Ingersoll? Ross and Wright?Fisher diffusions, which admit duals of birth-and-death type, and compare them with the available exact strategies based on death-type duals and with bootstrap particle filtering on the diffusion state space as a general benchmark.

Keywords

Duality; Filtering; Bayesian inference; Diffusion; Particle filtering; Smoothing.

Adaptive conformal classification with noisy labels

Matteo Sesia[1], **Y. X. Rachel Wang**[2], **Xin Tong**[1]

[1] *University of Southern California, Department of Data Sciences and Operations, USA.*
[2] *University of Sydney, School of Mathematics and Statistics, Australia.*

Abstract

This work presents novel conformal prediction methods for classification tasks that can automatically adapt to random label contamination in the calibration sample, enabling more informative prediction sets with stronger coverage guarantees compared to state-of-the-art approaches. This is made possible by a precise theoretical characterization of the effective coverage inflation (or deflation) suffered by standard conformal inferences in the pres- ence of label contamination, which is then made actionable through new calibration algorithms. Our solution is flexible and can leverage different modeling assumptions about the label contamination process, while requir- ing no knowledge about the data distribution or the inner workings of the machine-learning classifier. The advantages of the proposed methods are demonstrated through extensive simulations and an application to object classification with the CIFAR-10H image data set.

Keywords

Classification, Conformal prediction, Contaminated data, Label noise.

Future prospects for spatial statistics

Michael L. Stein[1]

[1] *Rutgers University, Department of Statistics, USA,*
ms2870@stat.rutgers.edu

Abstract

Data with a spatial aspect arises in many areas of the natural and social sciences. Arguably the central problem in spatial statistics is the development of models that appropriately describe the spatial dependence between observations taken at different locations. Stationary Gaussian processes are a natural starting point for continuously varying spatial quantities, but increasing dataset sizes generally make it clear that neither stationarity nor Gaussianity are tenable assumptions in many applications. Additionally, large datasets create computational problems when using such models that will not be solved by bigger and faster computers. This talk will review the present state of spatial and space-time statistics models and important directions for future research in theory, computation and methodology for spatial statistics, including to what extent modern machine learning methods might provide a way forward on some of the more difficult challenges in working with large spatially indexed datasets.

Keywords

Space-time statistics, Gaussian process, non-stationary process.

Bayesian multi-species N-mixture models for large scale spatial data in community ecology

Michele Peruzzi[1]

[1] *University of Michigan, Department of Biostatistics, United States, peruzzi@umich.edu*

Abstract

Community ecologists seek to model the local abundance of multiple animal species while taking into account that observed counts only represent a portion of the underlying population size. Analogously, modeling spatial correlations in species' latent abundances is important when attempting to explain how species compete for scarce resources. We develop a Bayesian multi-species N-mixture model with spatial latent effects to address both issues. On one hand, our model accounts for imperfect detection by modeling local abundance via a Poisson log-linear model. Conditional on the local abundance, the observed counts have a binomial distribution. On the other hand, we let a directed acyclic graph restrict spatial dependence in order to speed up computations, and use recently developed gradient-based Markov-chain Monte Carlo methods to sample a posteriori in the multivariate non-Gaussian data scenarios in which we are interested.

Keywords

Bayesian, ecology, latent Gaussian, Gaussian Process, spatial.

Semiparametric Bayesian Modeling of Nonstationary Joint Extremes

Miguel de Carvalho[1] **and Vianey Palacios Ramirez**[2]

[1] *University of Edinburgh, School of Mathematics, Portugal, Miguel.deCarvalho@ed.ac.uk*
[2] *Newcastle University, School of Mathematics, Statistics and Physics, UK, Vianey.Palacios-Ramirez@newcastle.ac.uk*

Abstract

In this talk, I will propose a novel Bayesian model for inferring about the intensity of observations in the joint tail over time, and for assessing if two stochastic processes are asymptotically dependent. To model the intensity of observations exceeding a high threshold, I will develop a Bayesian nonparametric approach that defines a prior on the space of what we define as EDI (Extremal Dependence Intensity) functions. In addition, a parametric prior is set on the coefficient of tail dependence. An extensive battery of experiments on simulated data showcases that the proposed methods are able to recover the true targets in a variety of scenarios. An application of the proposed methodology to a set of big tech stocks—known as FAANG—sheds light on some interesting features on the dynamics of their combined losses over time.

Keywords

Mixture of Polya trees, Statistics of extremes, Multivariate extreme values, Nonparametric prior, Nonstationary extremal dependence.

Spectral methods for clustering signed and directed networks and heterogeneous group synchronization

Mihai Cucuringu[1][2][3]

[1] *University of Oxford, Department of Statistics & Mathematical Institute, Oxford, UK, mihai.cucuringu@stats.ox.ac.uk*
[2] *Oxford-Man Institute of Quantitative Finance, Oxford, UK*
[3] *The Alan Turing Institute, London, UK*

Abstract

Graph clustering problems typically arise in settings where there exists a discrepancy in the edge density within different parts of the graph. In this work, we consider several problem instances where the underlying cluster structure arises as a consequence of a signal present on the edges or on the nodes of the graph, and is not driven by edge density. We first consider the problem of clustering in two important families of networks: signed and directed, both relatively less well explored compared to their unsigned and undirected counterparts. Both problems share an important common feature: they can be solved by exploiting the spectrum of certain graph Laplacian matrices or derivations thereof. In signed networks, the edge weights between the nodes may take either positive or negative values, encoding a measure of similarity or dissimilarity. We consider a generalized eigenvalue problem involving graph Laplacians, and provide performance guarantees under the setting of a Signed Stochastic Block Model, along with regularized versions to handle very sparse graphs (below the connectivity threshold), a regime where standard spectral methods are known to underperform. We also propose a spectral clustering algorithm for directed graphs based on a complex-valued representation of the adjacency matrix, which is able to capture the underlying cluster structures, for which the information encoded in the direction of the edges is crucial. We evaluate the proposed algorithm in terms

of a cut flow imbalance-based objective function, which, for a pair of given clusters, it captures the propensity of the edges to flow in a given direction. We analyze its theoretical performance on a Directed Stochastic Block Model for digraphs in which the cluster-structure is given not only by variations in edge densities, but also by the direction of the edges. Finally, we discuss an extension of the classical angular synchronization problem that aims to recover unknown angles $\theta_1, \ldots, \theta_n$ from a collection of noisy pairwise measurements of the form $(\theta_i - \theta_j) \bmod 2\pi$, for each $\{i, j\} \in E$. We consider a generalization to the heterogeneous setting where there exist k unknown groups of angles, and the measurement graph has an unknown edge-disjoint decomposition $G = G_1 \cup G_2 \ldots \cup G_k$, where the G_i's denote the subgraphs of noisy edge measurements corresponding to each group. We propose a probabilistic generative model for this problem, along with a spectral algorithm for which we provide a detailed theoretical analysis in terms of robustness against both sampling sparsity and noise.

Keywords

Spectral methods, signed/directed networks, group synchronization.

Neural Likelihood Surface Estimation for Intractable Spatial Models

Julia Walchessen[1], **Amanda Lenzi**[2], **Mikael Kuusela**[3]

[1] *Carnegie Mellon University, Department of Statistics and Data Science, USA,* jwalches@andrew.cmu.edu
[2] *University of Edinburgh, School of Mathematics, UK,* amanda.lenzi@ed.ac.uk
[3] *Carnegie Mellon University, Department of Statistics and Data Science, USA,* mkuusela@andrew.cmu.edu

Abstract

Likelihood-based inference tends to be computationally intensive or wholly intractable for many common models in spatial statistics. Examples include Gaussian processes for large data sets and models for spatial extremes. Recent work has used neural networks to predict parameters in these models, circumventing the intractability of likelihood computations. Prediction, however, depends on the choice of a prior on the parameters and does not provide a straightforward means of frequentist uncertainty quantification. In this talk, I will demonstrate how to use tools from likelihood-free inference to learn the likelihood function of intractable spatial processes using convolutional neural networks. In cases where the exact likelihood is available, the method provides similar point estimation and uncertainty quantification performance as exact likelihood computations at a fraction of the computational cost. When the likelihood is unavailable, this method can learn the otherwise intractable likelihood function, providing inferences that are superior to existing approximations. The method is applicable to any spatial process on a grid from which fast forward simulations are available and can also be adapted to complex models in other areas of statistics.

Keywords

Neural inference, deep neural networks, likelihood-free inference, simulation-based inference, spatial extremes.

Root and Community Inference on the Latent Growth Process of a Network

Harry Crane[1], **Min Xu**[1]

[1] *Rutgers University, Department of Statistics, USA,*
gm845@stat.rutgers.edu

Abstract

Many existing statistical models for networks overlook the fact that many real world networks are formed through a growth process. To address this, we introduce the PAPER (Preferential Attachment Plus Erdős–Rényi) model for random networks, where we let a random network G be the union of a preferential attachment (PA) tree T and additional Erdős–Rényi (ER) random edges. The PA tree component captures the underlying growth/recruitment process of a network where vertices and edges are added sequentially, while the ER component can be regarded as random noise. Given only a single snapshot of the final network G, we study the problem of constructing confidence sets for the early history, in particular the root node, of the unobserved growth process; the root node can be patient zero in a disease infection network or the source of fake news in a social media network. We propose an inference algorithm based on Gibbs sampling that scales to networks with millions of nodes and provide theoretical analysis showing that the expected size of the confidence set is small so long as the noise level of the ER edges is not too large. We also propose variations of the model in which multiple growth processes occur simultaneously, reflecting the growth of multiple communities, and we use these models to provide a new approach to community detection.

Keywords

Network model, community detection, node centrality, inference.

Exact Inference for Common Odds Ratio in Meta-Analysis with Zero-Total-Event Studies

Xiaolin Chen[1], Jerry Cheng[2], Lu Tian[3], Min-ge Xie[3]

[1]*Qufu Normal University, School of Statistics and Data Science, P. R. China*
[2]*New York Institute of Technology, Department of Computer Science, USA*
[3]*Stanford University, School of Medicine, USA*
[4]*Rutgers University, Department of Statistics, USA*
mxie@stat.rutgers.edu

Abstract

Stemming from the high profile publication of Nissen and Wolski (2007) and subsequent discussions with divergent views on how to handle observed zero-total-event studies, defined to be studies which observe zero events in both treatment and control arms, the research topic concerning the common odds ratio model with zero-total-event studies remains to be an unresolved problem in meta-analysis. In this article, we address this problem by proposing a novel repro samples method to handle zero-total-event studies and make inference for the parameter of common odds ratio. The development explicitly accounts for sampling scheme and does not rely on large sample approximation. It is theoretically justified with a guaranteed finite sample performance. The empirical performance of the proposed method is demonstrated through simulation studies. It shows that the proposed confidence set achieves the desired empirical coverage rate and also that the zero-total-event studies contains information and impacts the inference for the common odds ratio. The proposed method is also applied to combine information in the Nissen and Wolski study.

Keywords

Exact confidence interval; Meta-analysis; Odds ratio; Repro samples method; Zero-total-event studies.

Confidence Sets for Causal Orderings

Y. Samuel Wang[1], **Mladen Kolar**[2], **Mathias Drton**[3]

[1] *Cornell University, ysw7@cornell.edu*
[2] *University of Chicago, mkolar@chicagobooth.edu*
[3] *TU Munich, mathias.drton@tum.de*

Abstract

Causal discovery procedures aim to deduce causal relationships among variables in a multivariate dataset. While various methods have been proposed for estimating a single causal model or a single equivalence class of models, less attention has been given to quantifying uncertainty in causal discovery in terms of confidence statements. The primary challenge in causal discovery is determining a causal ordering among the variables. Our research offers a framework for constructing confidence sets of causal orderings that the data do not rule out. Our methodology applies to structural equation models and is based on a residual bootstrap procedure to test the goodness-of-fit of causal orderings. We demonstrate the asymptotic validity of the confidence set constructed using this goodness-of-fit test and explain how the confidence set may be used to form sub/supersets of ancestral relationships as well as confidence intervals for causal effects that incorporate model uncertainty.

Keywords

Causal discovery, linear structural equation models, confidence sets, quantification of uncertainty, goodness-of-fit.

Kernelized CODEC: A Family of Correlation Coefficients

Mona Azadkia[1], **Fang Han**[2]

[1] *London School of Economics and Political Science, Department of Statistics, United Kingdom, m.azadkia@lse.ac.uk*
[2] *University of Washington, Department of Statistics, United States, fanghan@uw.edu*

Abstract

In 2021, Chatterjee proposed an ingenious estimator using nearest neighbours for the non-parametric measure of dependence between random variables $\mathbf{X}$ and $\mathbf{Y}$ introduced by Dette *et al.* This dependence measure takes values between 0 and 1, where 0 or 1 corresponds to the case that a pair of random variables are independent or one is a measurable function of the other, almost surely. In this paper, we propose a new family of estimators for this measure and its expanded version, which allows $\mathbf{X}$ and $\mathbf{Y}$ to be multi-dimensional. We use the kernel smoothing technique to incorporate more information regarding the number of data points and the smoothness of the underlying relationship. We show the consistency of these estimators under certain assumptions on the kernel and provide the rate of convergence. For a special choice of kernel, we show the asymptotic normality of the estimator under independence.

Keywords

Measure of association, independence, correlation, Kernel Smoother.

References

Chatterjee, S. (2021). A new coefficient of correlation. *J. Amer. Statist. Assoc. 116*, 2009–2022.

Dette, H., Siburg, K. F., and Stoimenov, P. A. (2013). A copula-based nonparametric measure of regression dependence. *Scand. J. Stat., 40(1)*, 21–41.

Engression: Extrapolation for Nonlinear Regression?

Xinwei Shen[1], **Nicolai Meinshausen**[1]

[1] *Seminar für Statistik, D-MATH, ETH Zurich, Switzerland,*
{*xinwei.shen,meinshausen*}*@stat.math.ethz.ch*

Abstract

Extrapolation is crucial in many statistical and machine learning applications, as it is common to encounter test data outside the training support. However, extrapolation is a considerable challenge for nonlinear models. Conventional models typically struggle in this regard: while tree ensembles provide a constant prediction beyond the support, neural network predictions tend to become uncontrollable. This work aims at providing a nonlinear regression methodology whose reliability does not break down immediately at the boundary of the training support. Our primary contribution is a new method called 'engression' which, at its core, is a distributional regression technique for pre-additive noise models, where the noise is added to the covariates before applying a nonlinear transformation. Our experimental results indicate that this model is typically suitable for many real data sets. We show that engression can successfully perform extrapolation under some assumptions such as a strictly monotone function class, whereas traditional regression approaches such as least-squares regression and quantile regression fall short under the same assumptions. We establish the advantages of engression over existing approaches in terms of extrapolation, showing that engression consistently provides a meaningful improvement. Our empirical results validate these findings. The software implementations of engression are available in both R and Python.

Keywords

High-dimensional regression, neural networks, quantile regression.

Optimal Transport Based Denoising

Nicolás García Trillos[1], **Bodhisattva Sen**[2]

[1] *University of Wisconsin-Madison, Department of Statistics, USA,*
garciatrillo@wisc.edu
[2] *Columbia University, Department of Statistics, USA,*
bodhi@stat.columbia.edu

Abstract

In the standard formulation of the classical denoising problem, one is given a probabilistic model of latent variables and observations, and the goal is to construct a map to recover the latent variables from the observations. While there are many classical approaches for building denoising estimators, including the posterior mean, these estimators are often unable to adapt to the geometric structure of the prior distribution of latent variables. In this talk, I will present a new perspective on the denoising problem inspired by optimal transport (OT) theory. We propose new estimands that are motivated by theoretical considerations, first assuming that the prior distribution of latent variables is known. We rigorously prove that, under general assumptions on the model, these estimands are mathematically well-defined and are closely connected to solutions to Monge OT problems. After this, we discuss approaches for recovering our theoretically defined estimands in realistic settings and in particular discuss that, when the likelihood model is an exponential family of distributions, and assuming additional identifiability of the model, our estimands can be recovered solely from information of the marginal distribution of observations after solving a linear relaxation of the original problem that is reminiscent to standard multi-marginal OT. Our family of OT-like relaxations is of interest in its own right and for the denoising problem suggests alternative numerical methods inspired by the rich literature on computational OT.

Keywords

Denoising, Optimal transport.

Modeling considerations when optimizing adaptive experiments under the reinforcement learning framework

Nina Deliu[1,2], **Bibhas Chakraborty**[3,4]

[1] *MEMOTEF, Sapienza University of Rome, Italy, nina.deliu@uniroma1.it*
[2] *MRC – Biostatistics Unit, University of Cambridge*
[3] *Department of Statistics and Data Science, National University of Singapore, Singapore, bibhas.chakraborty@duke-nus.edu.sg*
[4] *Department of Biostatistics and Bioinformatics, Duke University*

Abstract

Artificial intelligence tools powered by machine learning have shown considerable improvements in a variety of experimental domains, from education to healthcare. In particular, the reinforcement learning (RL) and the multi-armed bandit (MAB) framework hold great promise for defining sequential designs with the aim of delivering optimized adaptive interventions with outcomes and resource (e.g., cost, time, or sample size) benefits. In this work, we discuss the opportunity offered by RL and MABs to current trends in healthcare experimentation, as well as the specific modeling challenges posed by this framework. Motivated by three case studies–in mobile health, digital mental health, and clinical trials–differing in their type of outcome, we illustrate our methodological contribution to this framework by integrating elements of traditional statistics. Specifically, we combine common offline data models for count and rating scale outcomes, increasingly common in digital and mobile health, with the Thompson sampling technique, which is possibly the most popular MAB algorithm. We discuss the theoretical properties of some of the proposed solutions and evaluate their empirical advantages in terms of balancing the exploitation (outcome performance) and exploration (learning performance) trade-off typical of reinforcement learning problems. Further considerations are provided under the unique challenging case of

small samples, where parametric assumptions are often unrealistic. In such settings, we demonstrate how RL-based solutions combined with bootstrap approaches represent a flexible yet improved strategy for achieving a near-optimal balance between patient benefit within the study and enhancing statistical operating characteristics in a small population.

Keywords

Reinforcement learning, Multi-armed bandits, Adaptive experiments.

Convex loss selection via score matching

Yu-Chun Kao[1], **Oliver Y. Feng**[2], **Min Xu**[3] and **Richard J. Samworth**[4]

[1] *Department of Statistics, Rutgers University, Piscataway, NJ 08854-8019, yk495@scarletmail.rutgers.edu*
[2] *Department of Mathematical Sciences, University of Bath, UK, BA2 7AY, of402@bath.ac.uk*
[3] *Department of Statistics, Rutgers University, Piscataway, NJ 08854-8019, mx76@stat.rutgers.edu*
[4] *Statistical Laboratory, University of Cambridge, UK, CB3 0WB, r.samworth@statslab.cam.ac.uk*

Abstract

We consider a linear regression model in which the regression coefficients are estimated by minimising the empirical risk based on a convex loss function. The accuracy of the estimator depends on the choice of loss function; for instance, when the errors are non-Gaussian, ordinary least squares can be outperformed by estimators based on alternative loss functions. A natural question then is how to select a data-driven convex loss function that leads to optimal downstream estimation of the regression coefficients. We propose a nonparametric approach that approximates the derivative of the log-density of the noise distribution by a monotone function, and explicitly identifies the convex loss function for which the asymptotic variance of the resulting M-estimator is minimal. We show that this optimisation problem is equivalent to a version of score matching, which corresponds to a log-concave projection of the noise distribution not in the usual Kullback–Leibler sense, but instead with respect to the so-called Fisher divergence.

Keywords

Score matching, robust regression, antitonic projection, antitonic efficiency.

The Dynamics of Firm Size Inequality: The Role of Acquisition and Innovation

Ou Liu[1]

[1] *Rutgers University-Newark, Department of Economics, USA,*
ouliu@rutgers.edu

Abstract

I study how the interaction between large acquirers and small targets shapes upper tail firm size inequality via acquisition and innovation. Empirically, I compile a new dataset, tracking dynamic ownership of public and private firms, along with their patents from in-house development or acquisitions. I identify three innovation channels through which acquisitions drive firm growth: (i) acquirers develop more innovations using target firms' patents; (ii) acquirers use acquisitions to expand into new areas; (iii) acquisitions shield acquirers' innovations from becoming technologically obsolete. Using firm random growth theories, I derive the dynamics of firm size distribution based on the dynamics of individual firms, modeling the growth of acquiring firms as a jump-diffusion process consistent with the empirically uncovered innovation channels. I find that acquisitions increase stationary firm size inequality and lead to a faster rise in inequality at the upper tail of firm size distributions.

Keywords

M&A, firm dynamics, Pareto inequality, transition dynamics.

Empowering Astronomy through Transformers: Time Series Classification and Text-to-SQL Challenges

Pablo A. Estevez[1,2]

[1] *University of Chile, Department of Electrical Engineering, Chile*
[2] *Millennium Institute of Astrophysics, Chile, pestevez@cec.uchile.cl*

Abstract

ALeRCE is an astronomical broker (Carrasco-Davis *et al.*, 2021) that is processing the alert stream from the Zwicky Transient Facility, and starting in 2025, it will begin processing the data streaming from the Vera C. Rubin Observatory. We use machine learning to process time series to classify transient, stochastic, and periodic stellar objects (Sánchez-Sáez *et al.*, 2021). In this work, we introduce deep attention-based models (transformers) using natural language processing concepts. We apply transformers to process astronomical time series to classify supernovae events and compare results with state-of-the-art machine learning models such as recurrent neural networks and random forests (Pimentel *et al.*, 2023). Next, we introduce ATAT, a multimodal transformer that combines time series and tabular data (Astorga *et al.*, 2023). ATAT is applied to the synthetic dataset from the ELAsTICC challenge, reaching a macro F1-score of 82.9 ± 0.4 using 20 classes. We apply large language models to the text-to-SQL parsing problem in the ALeRCE astronomical dataset and discuss future challenges in the field.

Keywords

Transformers, Light Curves, Tabular data, Text-to-SQL parsing.

Acknowledgments

The author acknowledges funding from ANID-Chile ICN12009 and Fondecyt 1220829

References

Pimentel, O., Estevez, P.A., Förster, F. (2023). Deep Attention-Based Supernovae Classification of Multi-Band Light-Curves. *The Astronomical Journal 165:18*.

Astorga, N., Reyes, I., Cabrera, G., et al. (2023). ATAT: Astronomical Transformer for time series And Tabular data *https://assets.researchsquare.com/files/rs-2395110/v1/0bb4b30c3ad9bf06573458e3.pdf?c=1673457093*

Carrasco-Davis et al. (2021). Alert Classification for the ALeRCE Broker System: The Real-time Stamp Classifier, *The Astronomical Journal 162:231*.

Sánchez-Sáez, P. et al. (2021) Searching for Changing-state AGNs in Massive Data Sets. I. Applying Deep Learning and Anomaly-detection Techniques to Find AGNs with Anomalous Variability Behaviors. *AJ 162:206*.

A New Dependence Measure for Extremal Brain Connectivity

Paolo Victor Redondo[1], **Jordan Richards**[2], **Raphaël Huser**[3], **Hernando Ombao**[4]

[1] *King Abdullah University of Science and Technology, CEMSE Division, Saudi Arabia, paolovictor.redondo@kaust.edu.sa*
[2] *King Abdullah University of Science and Technology, CEMSE Division, Saudi Arabia, jordan.richards@kaust.edu.sa*
[3] *King Abdullah University of Science and Technology, CEMSE Division, Saudi Arabia, raphael.huser@kaust.edu.sa*
[4] *King Abdullah University of Science and Technology, CEMSE Division, Saudi Arabia, hernando.ombao@kaust.edu.sa*

Abstract

Execution of voluntary muscular movements requires coordination between brain regions that carry out different cognitive functions. These movements may be reflected in recorded neuronal activities, such as electroencephalograms (EEGs), in the form of large amplitude signals. The interest now is to assess the complex connectivity patterns associated with these extreme signals, that may be hidden when dependence is explored for the entire data distribution. Thus, we develop new extremal dependence measures that quantify the extent to which large amplitude signals from one channel is associated with signals from another channel. By defining our measures based on ranks, one advantage is that our estimators involve fast computations. Another contribution is a fully nonparametric test for assessing extremal independence, which is an alternative to tests based on asymptotic null distribution with difficult sample size requirements, especially in the context of extremes. Lastly, we highlight the advantages of the proposed measures through numerical experiments and report interesting findings on the analysis of EEG recordings linked to a motor movement and imagery experiment.

Keywords

Conditional extremal dependence, Electroencephalogram, Motor imagery, Nonparametric test, Tail correlation.

Density estimation via JKO-flow generative models with guarantees

Chen Xu[1], **Xiuyuan Cheng**[2], **Yao Xie**[1]

[1] *Georgia Institute of Technology, H. Milton Stewart School of Industrial and Systems Engineering, USA, cx9711@gatech.edu, yao.xie@isye.gatech.edu*[2] *Duke University, Department of Mathematics, USA, xiuyuan.cheng@duke.edu*

Abstract

Recent strides in computational methodologies, encompassing areas such as optimization, neural networks, and optimal transport, have ushered in innovative approaches to tackle statistical inference challenges. These advancements present alternatives to the conventional Maximum Likelihood Estimation (MLE), particularly in contexts involving high-dimensional and modern complex data. In this talk, I will present a new continuous ODE flow-based generative model, called JKO-flow, which represents data density via a sequence of optimal transport maps parametrized by residual neural networks applied to a multivariate Gaussian and achieves stable learning with Wasserstein-2 regularized gradient flow in probability density space. The connection of our JKO-flow method with proximal gradient descent in the Wasserstein-2 space enables us to prove a performance guarantee with an exponential convergence rate.

Keywords

Density estimation, generative models, JKO, flow-based models.

Large scale outcome-guided Bayesian mixture models for cluster analysis of EHR datasets

Paul D.W. Kirk[1], **Sylvia Richardson**[1]

[1] *MRC Biostatistics Unit, University of Cambridge, United Kingdom, paul.kirk@mrc-bsu.cam.ac.uk*

Abstract

Motivated by the analysis of electronic health records (EHRs) for characterising patterns of multi-morbidity in UK populations, we present a divide-and-conquer approach for using Dirichlet process mixtures to model large-scale datasets. We review existing approaches, and consider the general challenge of matching cluster labels across data shards. We present a novel technique for cluster matching and assessing cluster stability, and introduce a recursive refinement approach to ensure that smaller clusters are not lost. We assess our method in simulations, and illustrate its application in the context of outcome-guided clustering to define patient populations that have similar patterns of co-occurring long-term conditions.

Keywords

Divide-and-conquer, Bayesian mixture models, cluster analyses.

A Bayesian Nonparametric Generative Model for Large Multivariate Non-Gaussian Spatial Fields

Paul F.V. Wiemann[1], **Matthias Katzfuss**[2]

[1] *University of Wisconsin–Madison, Department of Statistics, USA,* *paul.wiemann@wisc.edu*

[2] *University of Wisconsin–Madison, Department of Statistics, USA,* *katzfuss@gmail.com*

Abstract

Multivariate spatial fields are of interest in many domains, notably in climate model emulation. These fields can exhibit nonstationarity within individual marginal fields, and moreover, the dependence structure among the marginal fields might vary substantially. Our method extends a recently proposed Bayesian approach developed for nonstationary univariate spatial fields. By applying a triangular transport map, we can estimate the multivariate spatial field by solving a series of independent Gaussian process (GP) regression problems with Gaussian errors. Due to the potential nonlinearity in the conditional means, the joint distribution modeled can be non-Gaussian. The nonparametric Bayesian methodology scales well to high-dimensional spatial fields and works well when the number of replicates is relatively small. Inference in the model is carried out within an empirical Bayesian framework, utilizing a stochastic gradient method that is highly scalable. The approach can use mini-batching and profits from easy parallelization. To demonstrate the generative model, we apply it to analyze hydrological variables derived from non-Gaussian climate-model output.

Keywords

Climate-model emulation, Gaussian process, Generative modeling, Multivariate spatial field, spatial statistics.

Detecting outliers in large sets of time series

Pedro Galeano[1], **Daniel Peña**[2], **Ruey S. Tsay**[3]

[1] *Universidad Carlos III de Madrid, Departamento de Estadística, Spain, pedro.galeano@uc3m.es*
[2] *Universidad Carlos III de Madrid, Departamento de Estadística, Spain, daniel.pena@uc3m.es*
[3] *University of Chicago, Booth School of Business, USA, ruey.tsay@chicagobooth.edu*

Abstract

A fast and powerful procedure to individually identify outliers in time series from a large and heterogeneous database is presented. The approach is highly flexible as it can handle databases with different definitions, frequencies, and sample sizes across the various series. The process examines the residuals of the observed series after robust model fitting for outliers and, then, uses saturated regression models that account for all observations as potential outliers. The Orthogonal Greedy Algorithm (OGA) for saturated linear regression models is used to detect these effects within the observed series. The method is automatic and parallelizable, allowing fast and efficient identification of outliers over many long time series. The performance of the procedure is demonstrated in several simulations and the analysis of a real data example.

Keywords

Additive outliers, Level shifts, Orthogonal greedy algorithm, Outliers, Time series.

Deep exponential families for single-cell data analysis

Pedro F. Ferreira[1,2], **Jack Kuipers**[1,2,*], **Niko Beerenwinkel**[1,2,*]

[1] *Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland*
[2] *SIB Swiss Institute of Bioinformatics, Basel, Switzerland*
[*] *Corresponding authors: jack.kuipers@bsse.ethz.ch, niko.beerenwinkel@bsse.ethz.ch*

Abstract

Single-cell gene expression data characterizes the complex heterogeneity of living systems. This heterogeneity is composed of various cells with diverse cell states driven by different sets of genes. Cell states can often belong to different categories according to biological hierarchies. Hierarchical modelling therefore not only improves functional interpretation, but can also be leveraged to ensure that gene signatures are less influenced by noise and batch effects. We present single-cell Deep Exponential Families (scDEF), a multi-level Bayesian matrix factorization model which extracts hierarchical gene signatures from single-cell RNA-sequencing data. scDEF can additionally make use of prior information to guide the hierarchy. It can be used for dimensionality reduction, gene signature identification, and batch integration. We validate scDEF with simulations and multiple annotated real data sets, and show that scDEF recovers meaningful hierarchies in single- and multiple-batch scenarios.

Keywords

Probabilistic matrix factorization, Bayesian modelling, variational inference, scRNA-seq, hierarchical modelling.

In-Context Learning Linear Models with Transformers

Peter Bartlett*

* *UC Berkeley and Google DeepMind, USA, peter@berkeley.edu*

Abstract

Transformer networks have demonstrated a remarkable ability at in-context learning (ICL): given a short prompt sequence of labeled data, they can behave like supervised learning algorithms. We investigate the dynamics of ICL in transformers with a single linear self-attention layer trained by gradient flow on linear regression tasks. We show that despite non-convexity, gradient flow with a suitable initialization finds a global optimum and achieves prediction error competitive with the best linear predictor over the test prompt distribution, but it is not robust to shifts in the covariate distribution. For a simplified parameterization, we establish a statistical complexity bound for attention model pretraining using stochastic gradient descent, showing how in-context learning performance improves with the number of independent tasks.

Based on joint work with Ruiqi Zhang and Spencer Frei, and with Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, and Quanquan Gu.

Keywords

In-context learning, transformer models, deep learning, non-convex optimization.

Some new algorithms and old theory for Independent Component Analysis (ICA)

S. Kumar[1], **P. Sarkar**[2], **P. Bickel**[3]

[1] *UT Austin, Computer Science, USA, syamantak@utmail.utexas.edu*
[2] *UT Austin, Statistics and Data Sciences, USA, purna.sarkar@austin.utexas.edu*
[3] *UC Berkeley, Statistics, USA, bickel@stat.berkeley.edu*

Abstract

ICA was proposed (without Gaussian noise), by Comon (1994) as a formal model for blind source separation in engineering. The generalized model for (noisy) ICA or (non-Gaussian) factor analysis is to observe $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ i.i.d. vectors in d dimensional space distributed as:

$$\boldsymbol{x} = B\boldsymbol{z} + \boldsymbol{g}$$

where $\boldsymbol{z}$ is a latent k dimensional vectors with unknown non-Gaussian components, $B \in \mathbb{R}^{d \times k}$ $(d \geq k)$ is an unknown mixing matrix, and $\boldsymbol{g}$ is a multivariate Gaussian vector. As shown by (Comon, 1994) if $\boldsymbol{g} = 0$ (noiseless case), $d = k$, for suitably normalized B, the above model, unlike in PCA, is identifiable. Furthermore, separating the sources often produces interpretable factors. The key idea in ICA is the development of a contrast function which is minimized for Gaussian random vectors and leads to an objective function that is maximized when one learns the directions of the independent components. The major algorithms FastICA (Hyvarinen, 1999) and JADE (Cardoso, 1993) do this using functions based on univariate or multivariate kurtosis. These algorithms can give severely biased results if $\boldsymbol{z}$ is heavy-tailed or has no higher moments (Anderson et al., 2017; Chen and Bickel, 2005; Hyvarinen, 1997) or the kurtosis of the components of $\boldsymbol{z}$ is small in magnitude. Moreover in the noisy case biases arise from low signal-to-noise ratios. Following Eriksson and Koivunen (2003), Chen and Bickel (2005) proposed minimization of a criterion that did not have the above issues in the noiseless case. In this work, we propose:

1. New contrast functions that work in situations where the usual algorithms fail but can fail in situations where the others work.

2. An extension of the Chen-Bickel criterion to the noisy case with a new goal: optimize among the various fast contrast function-based algorithms.

We validate our approach using simulations.

Keywords

Independent Component Analysis, Non-Gaussianity, Contrast functions.

References

[1] J. Anderson, N. Goyal, A. Nandi, and L. Rademacher. Heavy-tailed analogues of the covariance matrix for ica. In S. P. Singh and S. Markovitch, editors, *AAAI*, pages 1712–1718. AAAI Press, 2017.

[2] J. Cardoso. Blind beamforming for non-gaussian signals. *IEE Proceedings F (Radar and Signal Processing)*, 140:362–370(8), December 1993. ISSN 0956-375X. URL `https://digital-library.theiet.org/content/journals/10.1049/ip-f-2.1993.0054`.

[3] A. Chen and P. Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53(10):3625–3632, 2005. doi: 10.1109/TSP.2005.855098.

[4] P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[5] J. Eriksson and V. Koivunen. Characteristic-function-based independent component analysis. *Signal Processing*, 83(10):2195–2208, 2003. ISSN 0165-1684. doi; https://doi.org/10.1016/S0165-1684(03)00162-2. URL `https://www.sciencedirect.com/science/article/pii/S0165168403001622`.

[6] A. Hyvarinen. One-unit contrast functions for independent component analysis: a statistical analysis. In *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*, pages 388–397, 1997. doi:10.1109/NNSP.1997.622420.

[7] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3): 626–634, 1999.

Fast Linear Model Trees by PILOT

Peter J. Rousseeuw[1], **Jakob Raymaekers**[2], **Tim Verdonck**[3], **Ruicong Yao**[1]

[1] *Section of Statistics and Data Science, University of Leuven, Belgium, {peter.rousseeuw,ruicong.yao}@kuleuven.be*
[2] *Department of Quantitative Economics, Maastricht University, The Netherlands, j.raymaekers@maastrichtuniversity.nl*
[3] *Department of Mathematics, University of Antwerp, Belgium, tim.verdonck@uantwerpen.be*

Abstract

Linear model trees are regression trees that incorporate linear models in the leaf nodes. This preserves the intuitive interpretation of decision trees and at the same time enables them to better capture linear relationships, which is hard for standard decision trees. But most existing methods for fitting linear model trees are time consuming and therefore not scalable to large data sets. In addition, they are more prone to overfitting and extrapolation issues than standard regression trees. In this paper we introduce PILOT, a new algorithm for linear model trees that is fast, regularized, stable and interpretable. PILOT trains in a greedy fashion like classic regression trees, but incorporates an L^2 boosting approach and a model selection rule for fitting linear models in the nodes. The abbreviation PILOT stands for **PI**ecewise **L**inear **O**rganic **T**ree, where 'organic' refers to the fact that no pruning is carried out. PILOT has the same low time and space complexity as CART without its pruning. An empirical study indicates that PILOT tends to outperform standard decision trees and other linear model trees on a variety of data sets. Moreover, we prove its consistency in an additive model setting under weak assumptions. When the data is generated by a linear model, the convergence rate is faster.

Keywords

Consistency, Piecewise linear model, Regression trees, Scalable algorithms.

Inferring Asymmetric Relations via Cross-fitting Data Analytics

Soumik Purkayastha[1], **Peter X. K. Song**[1]

[1] *University of Michigan, Department of Biostatistics, USA,*
soumikp@umich.edu, pxsong@umich.edu

Abstract

Causal investigations in observational studies pose a great challenge in scientific research where the fundamental assumptions of causality are not testable. Leveraging Shannon's information theory, we develop a new statistical inference theory in the analysis of predictive asymmetry, a central concept in information geometric causal inference. Asymmetric relation may be regarded as a low-dimensional projection of causality, which is characterized by an estimand between association and causality. In essence, predictive asymmetry enables assessment of whether a variable X is a stronger predictor of another variable Y or *vice-versa*. Such asymmetric relation or direction of association is often graphically depicted by an arrowed-edge in causal diagrams and deemed practically appealing. Our proposed estimand pertains to a new metric called *Asymmetric Mutual Information* (AMI) whose key statistical properties are established, including estimation consistency and the inference theory via cross-fitting techniques. The AMI is not only able to detect complex non-linear association patterns, but also is able to detect and quantify predictive asymmetry or direction of association. To estimate AMI, we propose a scalable non-parametric density estimation through fast Fourier transformation, which is manyfold faster than the classical bandwidth-based density estimation, with no loss of estimation accuracy. We illustrate the performance of the AMI methodology through simulation studies as well as multiple real data examples.

Keywords

Copula, data splitting, entropy, nonparametric estimation.

Multinomial Logistic Regression: Asymptotic Normality on Null Covariates in High-Dimensions

Kai Tan[1], **Pierre C. Bellec**[1]

[1] *Rutgers University, Department of Statistics, NJ, USA*

Abstract

This paper investigates the asymptotic distribution of the maximum-likelihood estimate (MLE) in multinomial logistic models in the high-dimensional regime where dimension and sample size are of the same order. While classical large-sample theory provides asymptotic normality of the MLE under certain conditions, such classical results are expected to fail in high-dimensions as documented for the binary logistic case in the seminal work of Sur and Candès [2019]. We address this issue in classification problems with 3 or more classes, by developing asymptotic normality and asymptotic chi-square results for the multinomial logistic MLE (also known as cross-entropy minimizer) on null covariates. Our theory leads to a new methodology to test the significance of a given feature. Extensive simulation studies on synthetic data corroborate these asymptotic results and confirm the validity of proposed p-values for testing the significance of a given feature.

Keywords

Multinomial logistic regression, confidence intervals, confidence ellipsoids, proportional regime.

Differentially private penalized M-estimation via noisy optimization

Marco Avella Medina[1], **Zheng Liu**, **Po-Ling Loh**[2]

[1] *Columbia University, Department of Statistics, USA, marco.avella@columbia.edu*
[2] *University of Cambridge, Department of Pure Mathematics and Mathematical Statistics, UK, pll28@cam.ac.uk*

Abstract

We propose a noisy composite gradient descent algorithm for differentially private statistical estimation in high dimensions. We begin by providing general rates of convergence for the parameter error of successive iterates under assumptions of local restricted strong convexity and local restricted smoothness. Our analysis is local, in that it ensures a linear rate of convergence when the initial iterate lies within a constant-radius region of the true parameter. At each iterate, multivariate Gaussian noise is added to the gradient in order to guarantee that the output satisfies Gaussian differential privacy. We then derive consequences of our theory for linear regression and mean estimation. Motivated by M-estimators used in robust statistics, we study loss functions which downweight the contribution of individual data points in such a way that the sensitivity of function gradients is guaranteed to be bounded, even without the usual assumption that our data lie in a bounded domain. We prove that the objective functions thus obtained indeed satisfy the restricted convexity and restricted smoothness conditions required for our general theory. We then show how the private estimators obtained by noisy composite gradient descent may be used to obtain differentially private confidence intervals for regression coefficients, by leveraging work in Lasso debiasing proposed in high-dimensional statistics. We complement our theoretical results with simulations that illustrate the favorable finite-sample performance of our methods.

Keywords

Differential privacy, robust statistics, high-dimensional statistics, Lasso debiasing.

Sensitivity Analysis of Observational Studies via Stochastic Programming

Qingyuan Zhao[1]

[1] *Statistical Laboratory, University of Cambridge, United Kingdom, qyzhao@statslab.cam.ac.uk*

Abstract

Any observational study of causal relationships relies on untestable assumptions. Thus, the credibility of a study crucially depends on the extent that its assumptions can be defended. In this talk, I will formulate sensitivity analysis as an stochastic programming problem, which provides a unified conceptual framework for a large number of statistical models and methods in the literature. I will also review some recent progress in this area and highlight the variety of theoretical, methodological, and practical difficulties involved in this framework. Part of this talk is based on joint work with Yao Zhang and Tobias Freidling.

Keywords

Causal inference, Optimization, Sensitivity analysis.

Autoregressive networks and some stylized features of network data

Qiwei Yao

Department of Statistics, London School of Economics and Political Science, q.yao@lse.ac.uk

Abstract

We propose a first-order autoregressive model for dynamic network processes in which edges change over time while nodes remain unchanged. The model depicts the dynamic changes explicitly. It also facilitates simple and efficient statistical inference such as the maximum likelihood estimators which are proved to be (uniformly) consistent and asymptotically normal. The model diagnostic checking can be carried out easily using a permutation test. We also elucidate how the AR model can accommodate node heterogeneity, edge sparsity, transitivity, homophily and other stylized features in network data.

Keywords

Dynamic networks, node heterogeneity, edge sparsity, transitivity, edge dependence.

Scalable Minimum Distance Estimator with Universal Performance Guarantees

Raazesh Sainudiin[1] **and Axel Sandstedt**[1]

[1] *Uppsala University, Department of Mathematics, Sweden,*
raazesh.sainudiin@math.uu.se

Abstract

We construct multivariate histogram estimators known as minimum distance estimatators with universal performance guaranteesfrom arbitrarily large sample sizes. We provide new distributed algorithms which minimize network communication taking place between computers in a clutser by exploiting properties of sparse binary trees. Finally using arithmetic over such trees we are able to conduct conditional density regression and estimate tail probabilities to detect typicality or anomaly. Open source library along with simulation results over terabytes of sample data are provided.

Keywords

L1 Density Estimation, Yatracos Class, Scheffe Sets, Big Data, Sparse Binary Trees.

Rank-transformed subsampling: Inference for multiple data splitting and exchangeable p-values

F.R. Guo[1] **and R.D. Shah**[2]

[1] *University of Cambridge, Statistical Laboratory, UK,*
r.shah@statslab.cam.ac.uk
[2] *University of Cambridge, Statistical Laboratory, UK,*
ricguo@statslab.cam.ac.uk

Abstract

Many testing problems are readily amenable to randomised tests such as those employing data splitting, which divide the data into disjoint parts for separate purposes. However despite their usefulness in principle, randomised tests have obvious drawbacks. Firstly, two analyses of the same dataset may lead to different results. Secondly, the test typically loses power because it does not fully utilise the entire sample. As a remedy to these drawbacks, we study how to combine the test statistics or p-values resulting from multiple random realisations such as through random data splits. We introduce rank-transformed subsampling as a general method for delivering large sample inference about the combined statistic or p-value under mild assumptions. We apply our methodology to a range of problems, including testing unimodality in high-dimensional data, testing goodness-of-fit of parametric quantile regression models, testing no direct effect in a sequentially randomised trial and calibrating cross-fit double machine learning confidence intervals. For the latter, our method improves coverage in finite samples and for the testing problems, our method is able to derandomise and improve power. Moreover, in contrast to existing p-value aggregation schemes that can be highly conservative, our method enjoys type-I error control that asymptotically approaches the nominal level.

Keywords

Data-splitting, Goodness-of-fit, Rejection sampling, Subsampling.

Spatio-temporal modelling of fish species distribution

Raquel Menezes[1], **Daniela Silva**[2], **Susana Garrido**[3]

[1] *Centre of Mathematics (CMAT), Minho University, Guimarães, Portugal, rmenezes@math.uminho.pt*
[2] *Centre of Mathematics (CMAT), Minho University, Braga, Portugal, danyelasylva2@gmail.com*
[3] *Division of Modelling and Management of Fishery Resources, Portuguese Institute for the Sea and Atmosphere (IPMA), Lisboa, Portugal, susana.garrido@ipma.pt*

Abstract

Scientific tools capable of identifying species distribution patterns are crucial, as they contribute to advancing our understanding of the factors driving species fluctuations. Species distribution data often exhibit residual spatial autocorrelation and temporal variability, making both components essential for studying the evolution of species distribution from an ecological perspective. Fishery data typically originate from two primary sources: fishery-independent data, often obtained from commercial fleets, and fishery-independent data, typically collected through research surveys. Research surveys are conducted once or twice a year over a broader spatial region and involve standardized sampling designs that cover fewer spatial locations. In contrast, data collected from commercial fleets often exhibit a higher recurrence, with more sampled locations within a smaller region due to a preferential selection of these locations. While these two data sources may offer distinct yet valuable information, they can be used complementarily. Jointly modeling these two sources requires an approach capable of accommodating the differing sampling designs. Classical tools excel at handling standardized sampling designs but are ill-equipped to address the preferential nature of commercial data. The current presentation shares preliminary results on a proposed joint model capable

of handling both preferential and non-preferential sampling designs. Furthermore, when modeling fish species distribution, it's essential to consider zero-inflated data, a common occurrence in research surveys data. The discussed models, addressing these challenges, are designed to provide a comprehensive understanding of species distribution patterns and fluctuations, offering a promising tool for ecological research in the context of fisheries management.

Keywords

Geostatistics, joint modelling, preferential sampling, species distribution model.

Composite likelihood for space-time point processes

Rasmus Waagepetersen[1]

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

Abstract

Large space-time point pattern data sets are collected by rain forest ecologists to study, e.g., factors underlying the high biodiversity of rain forests. The dynamics of a rain forest is extremely complex involving births, deaths and growth of trees with complex interactions between hundreds of species of trees, animals, climate, and environment. Building a realistic stochastic model for the space-time evolution of a rain forest is therefore very challenging if possible at all. We pursue a more modest goal where the objective is to infer selected aspects of rain forest growth while minimizing model assumptions. We hence aim at inferring parameters in regression models for recruitment and deaths of trees between consecutive rain forest censuses. We infer regression parameters using composite likelihood functions that only involve the specified first order properties of the data. To evaluate parameter estimation uncertainty we construct estimators of covariance matrices that also only rely on the first order moments. Time series of point patterns from rain forest censuses are quite short while each point pattern covers a fairly big spatial region. To obtain asymptotic results we therefore exploit a central limit theorem (Jalilian et al., 2023) for the fixed timespan - increasing domain asymptotic setting. Conveniently, if suffices to impose weak dependence assumptions on the innovations of the space-time process. We investigate the proposed methodology by simulation studies and applications to rain forest data.

Keywords

Composite likelihood, increasing domain asymptotics, intensity function, logistic regression, point process, space-time, spatio-temporal.

References

Jalilian, A., Xu, G., Poinas, A. and Waagepetersen, R. (2022). A central limit theorem for a sequence of conditionally centered random fields. Submitted for publication.

Single-index mixture cure models. An application to a study of cardiotoxicity in breast cancer patients

Ricardo Cao[1], **Beatriz Piñeiro-Lamas**[2], **Ana López-Cheda**[3]

[1] *Universidade da Coruña, CITIC, A Coruña, Spain, ricardo.cao@udc.es*
[2] *Universidade da Coruña, CITIC, A Coruña, Spain, b.pineiro.lamas@udc.es*
[3] *Universidade da Coruña, CITIC, A Coruña, Spain, ana.lopez.cheda@udc.es*

Abstract

Standard survival models assume that the event of interest would always happen if there was a sufficient follow-up time. However, this is not always realistic. For instance, HER2-positive breast cancer patients usually receive trastuzumab. Although this therapy has antitumor efficacy, it can cause a problem in the heart, known as cardiotoxicity, in some patients. In this context, there will be a fraction of individuals that will never suffer the side effect, just because they are not susceptible to it. They are said to be cured, in the sense that no matter how long you observe them, they will never experience the final event. To study the time until the cardiotoxicity appears, mixture cure models are appropriate. They allow to estimate both the probability of being cured and the survival function of the uncured population, depending on some covariates. In the literature, nonparametric estimation of both functions is limited to continuous unidimensional covariates (A. López-Cheda *et al.*, 2017a,b). We fill this important gap by considering multidimensional covariates, and proposing a single-index model for dimension reduction. A dataset related to cardiotoxicity from the University Hospital of A Coruña is considered.

Keywords

Cure model, single-index model, survival analysis, cardiotoxicity.

Acknowledgement
This research has been supported by MINECO Grant PID2020-113578RB-I00, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020- 14 and Centro de Investigación del Sistema Universitario de Galicia ED431G 2019/01), all of them through the ERDF. The second author acknowledges financial support from Axudas Predoutorais da Xunta de Galicia, with reference ED481A-2020/290, and from Centro de Investigación en Tecnoloxías da Información e das Comunicacións (CITIC) of the University of A Coruña, funded by Xunta de Galicia and the European Union (ERDFGalicia 2014-2020 Program), by Grant ED431G 2019/01.

References

A. López-Cheda, R.Cao, M.A. Jácome, I. Van Keilegom (2017a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics & Data Analysis*, v.105, 144–165.

A. López-Cheda, M.A. Jácome, R. Cao (2017b). Nonparametric latency estimation for mixture cure models. *TEST* v.26, 353–376.

Clustering Multivariate Time Series using Energy Distance

Richard A. Davis[1], **Leon Fernandes**[2], **Konstantinos Fokianos**[3]

[1] *Columbia University, Department of Statistics, USA, davis.richarda@gmail.com*
[2] *Columbia University, Department of Statistics, USA, lf2607@columbia.edu*
[3] *University of Cyprus, Department of Mathematics and Statistics, Cyprus, fokianos.konstantinos@ucy.ac.cy*

Abstract

A novel methodology is proposed for clustering multivariate time series data that is based on energy distance defined in Szekely and Rizzo (2013). Specifically, a dissimilarity matrix is formed using the energy distance statistic to measure separation between the finite dimensional distributions for the component time series. Once the pairwise dissimilarity matrix is calculated, a hierarchical clustering method is then applied to obtain the final clustering. This procedure is completely non-parametric as the dissimilarities between stationary distributions are directly calculated without making any model assumptions. In order to justify this procedure, asymptotic properties of the energy distance estimates are derived for general stationary and ergodic time series. The method is illustrated in a simulation study for various component time series that are either linear or nonlinear. Finally, the methodology is applied to two examples; one involves GDP of selected countries and the other is population size of various states in the U.S.A from 1900-1999.

Keywords
Characteristic function; dissimilarity measure; energy distance; hierarchical clustering.

Isotonic subgroup selection

Manuel M. Müller[1], Henry W. J. Reeve[2], Timothy I. Cannings[3] and Richard J. Samworth[4]

[1] *University of Cambridge, Statistical Laboratory, United Kingdom, mm2559@cam.ac.uk*

[2] *University of Bristol, School of Mathematics, United Kingdom, henry.reeve@bristol.ac.uk*

[3] *University of Edinburgh, School of Mathematics, United Kingdom, timothy.cannings@ed.ac.uk*

[4] *University of Cambridge, Statistical Laboratory, United Kingdom, r.samworth@statslab.cam.ac.uk*

Abstract

Given a sample of covariate-response pairs, we consider the subgroup selection problem of identifying a subset of the covariate domain where the regression function exceeds a pre-determined threshold. We introduce a computationally-feasible approach for subgroup selection in the context of multivariate isotonic regression based on martingale tests and multiple testing procedures for logically-structured hypotheses. Our proposed procedure satisfies a non-asymptotic, uniform Type I error rate guarantee with power that attains the minimax optimal rate up to poly-logarithmic factors. Extensions cover classification, isotonic quantile regression and heterogeneous treatment effect settings. Numerical studies on both simulated and real data confirm the practical effectiveness of our proposal, which is implemented in the `R` package `ISS`.

Keywords

Isotonic regression, multiple testing, post-selection inference, subgroup selection, superlevel set estimation.

Uncertainty Quantification in Synthetic Controls with Staggered Treatment Adoption

Matias D. Cattaneo[1], **Yingjie Feng**[2], **Filippo Palomba**[3], **Rocio Titiunik**[4]

[1] *Department of Operations Research and Financial Engineering, Princeton University*
[2] *School of Economics and Management, Tsinghua University*
[3] *Department of Economics, Princeton University*
[4] *Department of Politics, Princeton University*

Abstract

We propose principled prediction intervals to quantify the uncertainty of a large class of synthetic control predictions (or estimators) in settings with staggered treatment adoption, offering precise nonasymptotic coverage probability guarantees. From a methodological perspective, we provide a detailed discussion of different causal quantities to be predicted, which we call *causal predictands*, allowing for multiple treated units with treatment adoption at possibly different points in time. From a theoretical perspective, our uncertainty quantification methods improve on prior literature by (i) covering a large class of causal predictands in staggered adoption settings, (ii) allowing for synthetic control methods with possibly nonlinear constraints, (iii) proposing scalable robust conic optimization methods and principled data-driven tuning parameter selection, and (iv) offering valid uniform inference across post-treatment periods. We illustrate our methodology with an empirical application studying the effects of economic liberalization in the 1990s on GDP for emerging European countries. Companion general-purpose software packages are provided in `Python`, `R` and `Stata`.

Acknowledgements
We thank Alberto Abadie, Simon Freyaldenhoven, and Bartolomeo

Stellato for many insightful discussions. Cattaneo and Titiunik gratefully acknowledge financial support from the National Science Foundation (SES-2019432 and SES-2241575), Cattaneo gratefully acknowledges financial support from the National Institute of Health (R01 GM072611-16

Keywords

Causal inference, synthetic controls, staggered treatment adoption, prediction intervals, non-asymptotic inference.

Dynamic Matrix/Tensor Factor Models for High Dimensional Time Series

Rong Chen[1]

[1] *Rutgers, Department of Statistics, Rutgers University, USA, rongchen@stat.rutgers.edu*

Abstract

In many applications fields such as economics, finance and engineering, matrix or tensor observations are now commonly observed over time. Often the dimensions of these time series data high. Factor models have been used as a dimension reduction tool, resulting in matrix/tensor factor models. Such factor models do not assume any specific dynamic structures of the latent factor process, hence is not generative and cannot be used to make predictions. In this paper, we propose a dynamic matrix/tensor factor model that extends the factor model by modeling the latent low dimensional factor process with an autoregressive structure. A two stage procedure is used for estimation. Theoretical properties and finite sample empirical properties of the estimator are presented, along with application examples.

Keywords

Time Series, Matrix, Tensor, Factor.

Spatial data fusion adjusting for preferential sampling using INLA and SPDE

Ruiman Zhong[1], **André Victor Ribeiro Amaral**[2],**Paula Moraga**[2]

[1] *King Abdullah University of Science and Technology, Saudi Arabia, ruiman.zhong@kaust.edu.sa*

[2] *King Abdullah University of Science and Technology, Saudi Arabia*

Abstract

Spatially misaligned data can be fused by using a Bayesian melding model that assumes that underlying all observations there is a spatially continuous Gaussian random field process. This model can be used, for example, to predict air pollution levels by combining point data from monitoring stations and areal data from satellite imagery. However, if the data presents preferential sampling, that is, if the observed point locations are not independent of the underlying spatial process, the inference obtained from models that ignore such a dependence structure might not be valid. In this paper, we present a Bayesian spatial model for the fusion of point and areal data that takes into account preferential sampling. The model combines the Bayesian melding specification and a model for the stochastically dependent sampling and underlying spatial processes. Fast Bayesian inference is performed using the integrated nested Laplace approximation (INLA) and the stochastic partial differential equation (SPDE) approaches. The performance of the model is assessed using simulated data in a range of scenarios and sampling strategies that can appear in real settings. The model is also applied to predict air pollution in the USA.

Keywords

Spatial misalignment,Preferential Sampling , log Gaussian Cox process, Point patterns, INLA, SPDE .

Policy Learning with Continuous Actions Under Unmeasured Confounding

Yuhan Li[1], **Eugene Han**[2], **Wenzhuo Zhou**[3], **Zhengling Qi**[4], **Yifan Cui**[5], **Ruoqing Zhu**[6]

[1] *University of Illinois Urbana-Champaign, Department of Statistics, USA,* yuhanli8@illinois.edu

[2] *University of Illinois Urbana-Champaign, Department of Statistics, USA,* eugeneh2@illinois.edu

[3] *University of California Irvine, Department of Statistics, USA,* wenzhuz3@uci.edu

[4] *The George Washington University, Department of Decision Sciences, USA,* qizhengling@gwu.edu

[5] *Zhejiang University, Center for Data Science, China,* cuiyf@zju.edu.cn

[6] *University of Illinois Urbana-Champaign, Department of Statistics, USA,* rqzhu@illinois.edu

Abstract

In the field of reinforcement learning applied to personalized medicine, unmeasured confounding variables often hinder the optimization of treatment policies, particularly in offline settings. While most existing methods focus on off-policy evaluation (OPE), they are generally not directly suited for learning optimal policies. For example, common assumptions that the behavior policy depends solely on unobserved state variables can be practically violated in real-world medical scenarios. In this study, we introduce a novel identification framework to accurately estimate policy values. This is achieved by identifying a set of variables that are not involved in policy determination but can potentially affect the reward. By appropriately constructing bridge functions, we can learn an optimal policy based on observed states, thereby enabling practical implementation. Our framework additionally tackles the dose-finding problem in personalized medicine by considering a continuous action space. We also explore the asymptotic properties of our proposed estimators under suitable conditions.

Keywords

Reinforcement Learning, Personalized Medicine, Policy Optimization, Unmeasured Confounding, Dose-finding.

A new reading of the parameters in Markov switching stereotype models

Roberto Colombi[1], **Sabrina Giordano**[2]

[1] *Department of Management, Information and Production Engineering, University of Bergamo, Italy, roberto.colombi@unibg.it*
[2] *Department of Economics, Statistics and Finance "Giovanni Anania", University of Calabria, Italy, sabrina.giordano@unical.it*

Abstract

When assessing attitudes and perceptions using Likert scales, individuals frequently display response tendencies such as orienting towards the midpoint or extremes, irrespective of the actual content. These behavioral patterns, referred to as middle, extremes, acquiescence, and disacquiescence response styles, can introduce biases in the results and are thoroughly investigated in psychometric literature. Our innovative approach, particularly in the context of longitudinal ordered categorical data, involves the simultaneous consideration of temporal response dynamics (observable ordinal variables) and latent response behaviors influenced by response styles. We achieve this through a Markov switching logit model with two latent components. One component addresses serial dependence and respondent-specific unobserved heterogeneity, while the other component accounts for the responding attitude (whether influenced by response styles or not). To model the influence of covariates on responses, we employ a stereotype logit model, which is chosen for its flexibility as an extension of the proportional odds logit model while retaining the advantage of utilizing a single parameter to describe the effect of a regressor. We introduce a novel interpretation of the stereotype model's parameters by defining allocation sets as intervals within the linear predictor values that identify the most likely response.

Keywords

Longitudinal ordinal data, response styles, latent variables.

Integrated shape-sensitive functional metrics

Sami Helander[1], **Petra Laketa**[2], **Pauliina Ilmonen**[3], **Stanislav Nagy**[2], **Germain Van Bever**[4], **Lauri Viitasaari**[5]

[1]*Aalto University, School of Business, Finland*
[2]*Charles University, Czech Republic*
[3]*Aalto University, School of Science, Finland*
[4]*University of Namur, Belgium*
[5]*Uppsala University, Sweden*

Abstract

In the paper, we develop a new integrated ball (pseudo)metric which provides an intermediary between a chosen starting (pseudo)metric d and the L_p distance in general function spaces. Selecting d as the Hausdorff or Fréchet distances, we introduce integrated shape-sensitive versions of these supremum-based metrics. The new metrics allow for finer analyses in functional settings, not attainable applying the non-integrated versions directly. Moreover, convergent discrete approximations make computations feasible in practice.

Keywords

Fréchet distance, Functional Data Analysis, Hausdorff distance, Pseudometric.

Modeling Shapes and Surfaces

Sayan Mukherjee[1], **Several co-authors**

[1] *University of Leipzig, Max Planck Institute for Mathematics in the Sciences, Duke University, sayan.mukherjee@mis.mpg.de*

Abstract

We will consider modeling shapes and fields via topological and lifted-topological transforms. Specifically, we show how the Euler Characteristic Transform and the Lifted Euler Characteristic Transform can be used in practice for statistical analysis of shape and field data. We also state a moduli space of shapes for which we can provide a complexity metric for the shapes. We also provide a sheaf theoretic construction of shape space that does not require diffeomorphisms or correspondence. A direct result of this sheaf theoretic construction is that in three dimensions for meshes, 0-dimensional homology is enough to characterize the shape. We will also discuss Gaussian processes on fiber bundles and applications to evolutionary questions about shapes. Applications in biomedical imaging and evolutionary anthropology will be stated throughout the talk.

Keywords

Topological data analysis, stochastic processes on geometric objects, biomedical and geometric applications.

Interpretable Classification of Categorical Time Series Using the Spectral Envelope and Optimal Scalings

Zeda Li[1], **Scott A. Bruce**[2], **and Tian Cai**[3]

[1] *Baruch College, The City University of New York, Paul H. Chook Department of Information System and Statistics, USA,* *zeda.li@baruch.cuny.edu*

[2] *Texas A&M University, Department of Statistics, USA,* *sabruce@tamu.edu*

[3] *The City University of New York, The Graduate Center, USA,* *tcai@gradcenter.cuny.edu*

Abstract

This article introduces a novel approach to the classification of categorical time series under the supervised learning paradigm. To construct meaningful features for categorical time series classification, we consider two relevant quantities: the spectral envelope and its corresponding set of optimal scalings. These quantities characterize oscillatory patterns in a categorical time series as the largest possible power at each frequency, or spectral envelope, obtained by assigning numerical values, or scalings, to categories that optimally emphasize oscillations at each frequency. Our procedure combines these two quantities to produce an interpretable and parsimonious feature-based classifier that can be used to accurately determine group membership for categorical time series. Classification consistency of the proposed method is investigated, and simulation studies are used to demonstrate accuracy in classifying categorical time series with various underlying group structures. Finally, we use the proposed method to explore key differences in oscillatory patterns of sleep stage time series for patients with different sleep disorders and accurately classify patients accordingly. The code for implementing the proposed method is available at https://github.com/zedali16/envsca.

Keywords

Categorical time series, classification, optimal scaling, multiple time series, spectral envelope.

Equality and equity in performative prediction

Seamus Somerstep[1], **Ya'acov Ritov**[1], **Yuekai Sun**[1]

[1] *Department of Statistics, University of Michigan, Ann Arbor, MI*

Abstract

In social prediction problems, the very act of making predictions can affect the prediction target. This is known as performativity, and it often manifests as distribution shifts in machine learning problems. We develop a notion of algorithmic fairness for performative prediction that takes advantage of performativity to resolve some incompatibilities between group fairness definitions. In particular, we show that it is possible for policymakers to treat workers in the Coate-Loury labor market model equally *and* steer them to an equitable state.

Keywords

Group fairness, performative prediction, statistical discrimination.

Step-Stress degradation Model for Lifetime Prediction of Rechargeable Batteries

Sheng-Tsaing Tseng[1]

[1] *Institute of Statistics, National Tsing Hua University, Taiwan, sttseng@stat.nthu.edu.tw*

Abstract

Conducting a cost-efficient lifetime-testing plan to timely assess lifetime information of recurrent event data is a challenging task in the industry. Motivated by a case study of rechargeable batteries, this work introduces a multi-run k-level step-stress experiment (in which different stresses are repeated in a cycle fashion) to collect the degradation data of lithium-ion batteries. We formulate the battery capacity over recharge cycles as a counting process and then adopt a trend renewal process (TRP) to characterize the degradation patterns of capacity varying with the stress level of the accelerated factor. By assuming a Markov property hold in our counting process, the degradation data observed in a multi-run klevel step-stress TRP model can be equivalently converted to their counterparts in k constant-stress TRP models. Based on this connection, we estimate the parameter using maximum likelihood and infer lithium-ion batteries' end-of-performance (EOP) at normal-use conditions with uncertainty quantification.

Keywords

Trend renewal process, trend renewal data, a multi-run step-stress experiment, lithium-ion batteries, end of performance (EOP).

A Copula Model for Trivariate Circular Data

Shogo Kato[1], **Christophe Ley**[2], **Sophia Loizidou**[2].

[1] *Institute of Statistical Mathematics, Japan*
skato@ism.ac.jp
[2] *Department of Mathematics, University of Luxembourg, Luxembourg*
christophe.ley@uni.lu *sophia.loizidou@uni.lu*

Abstract

We propose a new family of distributions for trivariate circular data. The probability density function of the proposed distribution can be expressed in simple form without involving infinite sums or integrals. The univariate marginals of this distribution are the uniform distributions on the circle, making the presented family a copula model for trivariate circular data. The bivariate marginals of the presented distribution belong to the family of Wehrly and Johnson (1980). The univariate and bivariate conditional distributions are the wrapped Cauchy distributions and the distributions of Kato and Pewsey (2015), respectively. An efficient algorithm is presented to generate random variates from our model. A closed-form expression is available for trigonometric moments. Maximum likelihood estimation for the presented distribution is numerically fast with a reparametrization of the parameters. We briefly discuss an extension of the proposed copula model with the wrapped Cauchy marginals. Finally, an application of the proposed model is given to a dataset of trivariate dihedral angles of amino acids in bioinformatics.

Keywords

Circulas, directional statistics, protein structure, torus.

References

Kato, S. and Pewsey, A. (2015). A Möbius transformation-induced distribution on the torus. *Biometrika*, 102(2), 359–370.

Wehrly, T.E. and Johnson, R.A. (1980). Bivariate models for dependence of angular observations and a related Markov process. *Biometrika*, 67(1), 255–256.

Learning directed acyclic graphs for ligands and receptors based on spatially resolved transcriptomic analysis of ovarian cancer

Shrabanti Chowdhury[1], Sammy Ferri-Borgogno[2], Anna P Calinawan[1], Peng Yang[2], Wenyi Wang[2], Jie Peng[3], Samuel Mok[2], Pei Wang[1]

[1] *Icahn School of Medicine, Genetics and Genomic Sciences, USA*
[2] *MD Andersen Cancer Center, Bioinformatics and Computational Biology, USA*
[3] *UC Davis, Statistics, USA*

Abstract

In order to understand the immune activation and suppression mechanisms in tumors, a critical step is to identify transcriptional signals underlying cell-cell communication between tumor and immune/stromal cells in the tumor micro-environment. A major mechanism enabling cell-cell communication is interactions between secreted ligands and cell-surface receptors. These interactions create a highly connected signaling network of ligands and receptors between cells. The latest advances in *in situ* -omics profiling, such as the spatial transcriptomic (ST) technology, provide unique opportunities to directly characterize ligand-receptor signaling networks that powers cell-cell communication. In this paper, we propose a novel statistical method, `LRSTNet`, to characterize the ligand-receptor interaction networks between adjacent tumor and stroma cells based on ST data, and apply it to a high grade serous ovarian cancer (HGSC) study. `LRSTNet` utilizes a directed acyclic graph (DAG) model with a novel approach to handle the zero-inflated distribution observed in the ST data. It also leverages existing ligand-receptor regulation databases as prior information, and employs a bootstrap aggregation strategy to achieve robust network estimation. We applied `LRSTNet` to the ST data of four HGSC patient tumor samples, and identified common and distinct ligand-receptor

regulations in different tumors. Some of the these interactions were also validated on the MERFISH data of an independent set of ovarian cancer patients. These results cast light on biological processes relating to the communication between tumor and immune/stromal cells in these ovarian tumors. An open-source R package of `LRSTNet` is available at github. We have also developed an interactive web application for the exploration of the HGSC ST data and the analysis results.

Keywords

Spatial transcriptomics data, ligand-receptor network, Hill climbing, Bootstrap aggregation, Prior.

Concentration of measure bounds for matrix-variate data with missing values

Shuheng Zhou[1]

[1] *University of California, Riverside, Department of Statistics, USA, szhou@ucr.edu*

Abstract

We consider the following data perturbation model, where the covariates incur multiplicative errors. For two $n \times m$ random matrices U, X, we denote by $U \circ X$ the Hadamard or Schur product, which is defined as $(U \circ X)_{i,j} = (U_{i,j})(X_{ij})$. In this paper, we study the subgaussian matrix variate model, where we observe the matrix variate data X through a random mask U:

$$\mathcal{X} = U \circ X,$$

where

$$X = B^{1/2} Z A^{1/2},$$

where Z is a random matrix with independent subgaussian entries, and U is a mask matrix with either zero or positive entries, where $EU_{ij} \in [0, 1]$ and all entries are mutually independent. Under the assumption of independence between X and U, we introduce componentwise unbiased estimators for estimating covariance A and B, and prove the concentration of measure bounds in the sense of guaranteeing the restricted eigenvalue (RE) conditions to hold on the unbiased estimator for B, when columns of data matrix X are sampled with different rates. We further develop multiple regression methods for estimating the inverse of B and show statistical rate of convergence. Our results provide insight for sparse recovery for relationships among entities (samples, locations, items) when features (variables, time points, user ratings) are present in the observed data matrix X with heterogeneous rates. Our proof techniques can certainly be extended to other

scenarios. We provide simulation evidence illuminating the theoretical predictions.

Keywords

Covariance estimation; inverse covariance estimation; matrix variate data; missing values; space-time model; sparse Hanson-Wright inequality; sparse quadratic forms.

Strategies for high-dimensional empirical Bayes problems

Sihai Dave Zhao[1]

[1] *University of Illinois Urbana-Champaign, Department of Statistics, United States, sdzhao@illinois.edu*

Abstract

Many modern data analysis problems require simultaneous estimation and/or inference for a large number of features. These problems are amenable to empirical Bayes approaches, which share information across the features. Nonparametric empirical Bayes methods are especially useful because they can automatically identify the optimal way to share that information in a given dataset. However, when the parameters of interest are of moderate or high dimension, nonparametric methods suffer from the curse of dimensionality and become extremely inaccurate. There are few practical solutions. This talk will introduce some motivating problems, from the fields of metabolomics and spatial transcriptomics, and will then explore new strategies for overcoming dimensionality when using nonparametric empirical Bayes methods.

Keywords

Compound decision, genomics, nonparametric maximum likelihood.

Adaptive class embedding for classification with a large number of classes

Yuanhao Liu[1], **Sijian Wang**[2]

[1] *Rutgers University, Department of Statistics, USA, yl1398@stat.rutgers.edu*
[2] *Rutgers University, Department of Statistics, USA, sijian.wang@stat.rutgers.edu*

Abstract

In the realm of digital economy, challenges often arise in tackling multi-class classification problems with a large number of classes, commonly found in tasks like product categorization and user behavior analysis. Traditional classification methods struggle to deliver satisfactory results on these large-scale problems due to their computational demands and subpar performance. This talk focuses on viewing classification from a class-embedding perspective, where classification is seen as balancing the forces acting upon the class-embedding vector from both correctly and incorrectly classified data points. From this viewpoint, we propose a framework that introduces an adaptively-weighted loss function designed to handle such complex classification tasks more efficiently. We also utilize sampling techniques to expedite computation. The utility and effectiveness of these methods are showcased through simulation studies and real-world data analysis.

Keywords

Class embedding, adaptively-weighted loss function, sampling techniques.

The Fundamental Limits of Structure-Agnostic Functional Estimation

Sivaraman Balakrishnan[1], **Edward Kennedy**[1], **Larry Wasserman**[1]

[1] *Carnegie Mellon University, Department of Statistics and Data Science, Machine Learning Department, USA*

Abstract

Many recent developments in causal inference, and functional estimation problems more generally, have been motivated by the fact that classical one-step (first-order) debiasing methods, or their more recent sample-split double machine-learning avatars, can outperform plugin estimators under surprisingly weak conditions. These first-order corrections improve on plugin estimators in a black-box fashion, and consequently are often used in conjunction with powerful off-the-shelf estimation methods. In this talk we will discuss the fundamental limits of structure-agnostic functional estimation, where relatively weak conditions are placed on the underlying nuisance functions. We show that there is a strong sense in which existing first-order methods are optimal. We provide a formalization of the problem of functional estimation with black-box nuisance function estimates, and deriving minimax lower bounds for this problem. Our results highlight some clear tradeoffs in functional estimation – if we wish to remain agnostic to the underlying nuisance function spaces, impose only high-level rate conditions, and maintain compatibility with black-box nuisance estimators then first-order methods are optimal. When we have an understanding of the structure of the underlying nuisance functions then carefully constructed higher-order estimators can outperform first-order estimators.

Keywords

Robust Functional Estimation, Causal Inference, Double Machine Learning.

Selective inference using randomized group lasso estimators for general models

Yiling Huang[1], **Sarah Pirenne**[2], **Snigdha Panigrahi**[1], **Gerda Claeskens**[2]

[1] *University of Michigan, Department of Statistics, U.S.*
[2] *KU Leuven, Orstat and Leuven Statistics Research Center, Belgium*

Abstract

Our work is motivated by the need for inference after regularized estimation with high dimensional datasets that contain grouped covariates. As an example, consider applying a logistic Group LASSO to a dataset with a binary outcome and categorical predictors. How do we conduct selective inference in the estimated sparse model? This problem is challenging due to two reasons: (1) existing approaches for a polyhedral selection method do not apply to the Group LASSO because there is no easy description of the selection event; (2) our data is no longer normal. To solve this problem, we construct an asymptotic selective likelihood that uses extra randomization to obtain an easy to describe selection event. Our new approach provides selective inference using randomized Group LASSO estimators in likelihood models including generalized linear models, and in other general forms of estimation, such as quasi-likelihood estimation to include possible overdispersion, for example.

Keywords

Group-regularized estimation; M-estimation; post-selection inference; randomization; selective inference.

Estimating the Complexity of Graph Limits

Sofia Olhede[1], **Anda Skeja**[1]

[1] *EPFL, Mathematics Institute, Switzerland, sofia.olhede@epfl.ch*

Abstract

Model complexity remains a key feature of any proposed data generating mechanism. Measures of complexity can be extended to complex patterns such as signals in time and space. In this paper, we study exchangeable graphs and their complexity. Exchangeability for graphs implies a distributional invariance under node permutation and is a suitable default model that can widely be used for network data. For this well-studied class of graphs, we make a choice to quantify model complexity based on the (Shannon) entropy, resulting in graphon entropy. We estimate the entropy of the generating mechanism of a given graph, instead of choosing a specific graph descriptor suitable only for one graph generating mechanism. In this manner, we naturally consider the global properties of a graph and capture its important graph-theoretic and topological properties. Under an increasingly complex set of generating mechanisms, we propose a set of estimators of graphon entropy as measures of complexity for real-world graphs. We determine the large–sample properties of such estimators and discuss their usage for characterizing evolving real-world graphs.

Keywords

Network complexity, Network data, Graph limit function.

A geostatistical mixture model to deal with both extra zeros and extreme values: a case study of sardine egg density in Portugal

Soraia Pereira[1], **Raquel Menezes**[2], **Maria Manuel Angélico**[3], **Tiago Marques**[4] , **Guido Moreira**[5]

[1] *CEAUL and FCUL, University of Lisbon, Portugal, sapereira@fc.ul.pt*
[2] *CEAUL and CMAT, University of Minho, Portugal*
[3] *IPMA, Portugal*
[4] *CREEM, University of St Andrews, UK; CEAUL, FCUL, University of Lisbon, Portugal*
[5] *Bavarian Nordic, Denmark*

Abstract

In our research on understanding the spatial distribution of sardine egg density for effective population management, we address the challenging data structure characterized by an excess of zeros and extreme values. To tackle this issue, we propose a geostatistical mixture model using a hierarchical Bayesian approach that includes a point mass at zero, a Gamma bulk, and a Generalized Pareto Distribution (GPD) tail. In the context of geostatistical models, Markov Chain Monte Carlo (MCMC) methods can be computationally intensive for inference. Thus, we formulate the statistical model and discuss its implementation within the integrated nested Laplace approximation (INLA) approach.

Keywords

Geostatistical models, mixture models, Bayesian hierarchical modeling, INLA, sardine egg density, zero-inflated models, extremes.

Frequency recovery from sketched data: a novel approach bridging Bayesian and frequentist views

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

[1] *University of Torino, Italy, mario.beraha@unito.it*
[2] *University of Torino and Collegio Carlo Alberto, Italy, stefano.favaro@unito.it*
[3] *University of Southern California, United States, sesia@marshall.usc.edu*

Abstract

We study how to recover the frequency of a symbol in a large discrete data set, using only a lossy-compressed representation, or sketch, of those data obtained via random hashing. This is a classical problem at the crossroad of computer science and information theory, with various algorithms available, such as the count-min sketch. However, these algorithms often assume that the data are fixed, leading to overly conservative and potentially inaccurate estimates when dealing with randomly sampled data. Here, we consider the sketched data as a random sample from an unknown distribution, and then we introduce novel estimators that improve upon existing approaches. Our method combines Bayesian nonparametric and classical (frequentist) perspectives, addressing their unique limitations to provide a principled and practical solution.

Keywords

Frequency recovery; nonparametric estimation; minimax analysis; normalized random measures; random hashing; smoothed estimation; worst-case analysis.

Nonasymptotic analysis of the empirical angular measure for multivariate extremes, with applications to classification and minimum volume set estimation

Stéphan Clémençon[1], **Hamid Jalalzai**[1], **Anne Sabourin**[1], **Johan Segers**[2]

[1] *LTCI, Télécom Paris, Institut polytechnique de Paris, France*
[2] *Université Catholique de Louvain, Belgium*

Abstract

In multivariate extreme value theory, the angular measure characterizes the first order dependence structure of multivariate heavy-tailed variables. In the case where the components have different tail indices, standardization using the rank-transformation (empirical distribution function) is a common practice. We provide a nonasymptotic bound for the uniform deviations of the empirical angular measure evaluated on rectangles of the unit sphere. Our bound scales as the squared root of the number of observations used for inference $\log(k)/\sqrt{k}$, up to a logarithmic factor. This nonasymptotic study is, to the best of our knowledge, the first of its kind in this domain. In addition we propose a modification of the classical empirical estimator based on the rank-transformed sample, based on intermediate data, *i.e.* upon data whose norm rank among the largest of the observed sample, but not among the very largest. In other words we discard the very largest data. Our error bound for this modified estimator does not suffer from a logarithmic factor, but includes a multiplicative term depending on the truncation level. The relative merits of both versions of the empirical measure are illustrated by numerical experiments. As an application, we provide finite sample guarantees for classification in extreme regions and anomaly detection *via* minimum-volume sets estimation on the sphere.

Assessing COVID-19 Prevalence in Austria with Infection Surveys and Case Count Data as Auxiliary Information

Stéphane Guerrier[1], **Christoph Kuzmics**[2], **Maria-Pia Victoria-Feser**[3]

[1] *University of Geneva, Geneva School of Economics and Management & Faculty of Science, Switzerland, Stephane.Guerrier@unige.ch*

[2] *University of Graz, Department of Economics, Austria, christoph.kuzmics@uni-graz.at*

[3] *University of Geneva, Geneva School of Economics and Management, Switzerland, Maria-Pia.VictoriaFeser@unige.ch*

Abstract

Countries officially record the number of COVID-19 cases based on medical tests of a subset of the population. These case count data obviously suffer from participation bias, and for prevalence estimation, these data are typically discarded in favour of infection surveys, or possibly also completed with auxiliary information. One exception is the series of infection surveys recorded by the Statistics Austria Federal Institute to study the prevalence of COVID-19 in Austria in April, May and November 2020. In these infection surveys, participants were additionally asked if they were simultaneously recorded as COVID-19 positive in the case count data. In this paper, we analyze the benefits of properly combining the outcomes from the infection survey with the case count data, to analyze the prevalence of COVID-19 in Austria in 2020, from which the case ascertainment rate can be deduced. The results show that our approach leads to a significant efficiency gain. Indeed, considerably smaller infection survey samples suffice to obtain the same level of estimation accuracy. Our estimation method can also handle measurement errors due to the sensitivity and specificity of medical testing devices and to the nonrandom sample weighting

scheme of the infection survey. The proposed estimators and associated confidence intervals are implemented in the companion open source R package pempi available on the Comprehensive R Archive Network (CRAN).

Keywords

Maximum likelihood estimation, (generalized) method of moments, sample proportion, case ascertainment rate, stratified sampling, infectious disease, Clopper-Pearson confidence interval, measurement error.

Unblinded Sample Size Re-estimation for the Wilcoxon-Mann-Whitney and Brunner-Munzel test

Stephen Schüürhuis[1], **Georg Zimmermann**[2], **Tobias Mütze**[3], **Frank Konietschke**[1]

[1]*Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany*
[2]*Statistical Methodology, Novartis Pharma AG, Basel, Switzerland*
[3]*Team Biostatistics and Big Medical Data, IDA Lab Salzburg, Paracelsus Medical University, Salzburg, Austria*

Abstract

Proper sample size determination throughout the planning stage of a randomized controlled trial is crucial. The sample size is usually determined as the (minimum) sample size to detect an alternative δ, say, with target power of $1-\beta$ at significance level α. In general, however, the sample size computed does not only depend on the aforementioned parameters, but also on nuisance parameters (e.g. variance). Hence, the appropriateness of the resulting sample size particularly depends on the validity of the specified input assumption on the effect. In practice, a-priori knowledge about parameter values might be scarce, e.g. in novel indications or rare diseases. Accordingly, the assumptions on the effect (and its variability) can be highly uncertain. Therefore, allowing for modifications of the preplanned sample size during the trial based on updated knowledge about this effect might be an attractive alternative option. Interim sample size re-estimation is regarded as a particular class of interim adaptions within the general framework of adaptive designs (see, e.g., [1]). In (unblinded) sample size re-estimation designs, a first-stage cohort of patient data can be used in order to adaptively de- or increase the overall sample size based on

the interim effect estimate as compared to the initially planned overall sample size. While extensive theory has been developed for binary, continuous and survival endpoints (e.g. [1, 2]), there has been comparatively little discussion in the adaptive design literature on how to perform sample size re-estimation if the underlying statistical procedure is nonparametric, e.g. if the analysis should be done using the Wilcoxon-Mann-Whitney or the Brunner-Munzel test. In disease areas such as amyotrophic lateral sclerosis (ALS), however, rank-based methods are commonly used, see e.g. [3], and they are considered more robust if distributional assumptions or asymptotics do not hold. The effect size of those tests is the so-called relative effect $p = \mathbb{P}(X < Y) + 0.5\mathbb{P}(X = Y)$ for $X \sim F_X$, $Y \sim F_Y$, where $p > 0.5$ means that Y stochastically tends to larger values than X. In this talk, we will present unblinded sample size re-estimation procedures for the Wilcoxon-Mann-Whitney and the Brunner-Munzel test. In particular, we will focus on interim sample size adaptations based on estimates of the relative effect p utilizing the conditional power of those tests, i.e. the probability to obtain a significant result given the already observed interim data. Moreover, we investigate the designs with respect to type I error rate control by means of simulation and compare the adaptive approach to the corresponding non-adaptive group-sequential design (see [4]) regarding performance characteristics.

[1] Wassmer, Gernot, and Werner Brannath. Group sequential and confirmatory adaptive designs in clinical trials. Vol. 301. Cham: Springer International Publishing, 2016.

[2] Chuang-Stein, Christy, et al. "Sample size reestimation: a review and recommendations." *Drug information journal: DIJ/Drug Information Association* 40 (2006): 475-484.

[3] Berry, James D., et al. "The Combined Assessment of Function and Survival (CAFS): a new endpoint for ALS clinical trials." *Amyotrophic lateral sclerosis and frontotemporal degeneration* 14.3 (2013): 162-168.

[4] Nowak, Claus P., Tobias Mütze, and Frank Konietschke. "Group sequential methods for the Mann-Whitney parameter." *Statistical Methods in Medical Research* 31.10 (2022): 2004-2020.

Keywords

Sample size re-estimation, Wilcoxon-Mann-Whitney test, Brunner-Munzel test, conditional power.

Inverse Leverage Effect for Cryptocurrencies and Meme Stocks: a Comprehensive Framework

Lendie Follett[1], **Steven Kou**[2], **Matthew Stuart**[3], **Cindy Yu**[4]

[1] *College of Business and Public Administration, Drake University, USA, lendie.follett@drake.edu*
[2] *Department of Finance, Boston University, USA, kou@bu.edu*
[3] *Department of Mathematics and Statistics, Loyola University, USA, mstuart1@luc.edu*
[4] *Department of Statistics, Iowa State University, USA, cindyyu@iastate.edu*

Abstract

Although the leverage effect, i.e., a negative correlation between the return and volatility, and the inverse leverage effect have been suggested for equities and commodities, respectively, the existing studies suffer from an identification problem because they only model one asset. By using a comprehensive multivariate model with jumps and heavy tail distribution for both an equity index and the asset, we find inverse leverage and volatility-varying leverage effects for cryptocurrencies and meme stocks. Network effects cannot explain this finding. To handle over 18,000 latent variables, a particle Gibbs with an ancestor sampling algorithm is extended to estimate parameters efficiently.

Keywords

Cryptocurrency, Jump-Diffusion Model, Bayesian Analysis, Asymmetric Laplace Distribution.

Orthogonal prediction of counterfactual outcomes

Stijn Vansteelandt[1], **Pawel Morzywolek**[1,2]

[1] *Ghent University, Department of Applied Mathematics, Computer Science and Statistics, Belgium, stijn.vansteelandt@ugent.be*
[2] *University of Washington, Department of Statistics, USA, pawel.morzywolek@ugent.be*

Abstract

Orthogonal meta-learners, such as DR-learner, R-learner and IF-learner, are increasingly used to estimate conditional average treatment effects. They improve convergence rates relative to naïve meta-learners (e.g., T-, S- and X-learner) through de-biasing procedures that involve applying standard learners to specifically transformed outcome data. This leads them to disregard the possibly constrained outcome space, which can be particularly problematic for dichotomous outcomes: these typically get transformed to values that are no longer constrained to the unit interval, making it difficult for standard learners to guarantee predictions within the unit interval. To address this, we construct orthogonal meta-learners for the prediction of counterfactual outcomes which respect the outcome space. As such, the obtained i-learner or imputation-learner is more generally expected to outperform existing learners, even when the outcome is unconstrained, as we confirm empirically in simulation studies and an analysis of critical care data. Our development also sheds broader light onto the construction of orthogonal learners for other estimands.

Keywords

Causal machine learning; Causal prediction; Conditional average treatment effect; Debiasing; Effect heterogeneity.

High dimensional tensor methods for multi-modal single cell genomics data

Kwangmoon Park[1], **Sunduz Keles**[2]

[1] *University of Wisconsin Madison, Department of Statistics, USA, kpark243@wisc.edu*

[2] *University of Wisconsin Madison, Department of Statistics and of Biostatistics and Medical Informatics, USA, keles@wisc.edu*

Abstract

Emerging single cell technologies that simultaneously capture long-range interactions of genomic loci together with their DNA methylation levels are advancing our understanding of three-dimensional genome structure and its interplay with the epigenome at the single cell level. While methods to analyze data from single cell high throughput chromatin conformation capture (scHi-C) experiments are maturing, methods that can jointly analyze multiple single cell modalities with scHi-C data are lacking. Here, we introduce Muscle, a semi-nonnegative joint decomposition of Multiple single cell tensors, to jointly analyze 3D conformation and DNA methylation data at the single cell level. Muscle takes advantage of the inherent tensor structure of the scHi-C data, and integrates this modality with DNA methylation. We developed an alternating least squares algorithm for estimating Muscle parameters and es- tablished its optimality properties. Parameters estimated by Muscle directly align with the key components of the downstream analysis of scHi-C data in a cell type specific manner. Evaluations with data-driven experiments and simulations demonstrate the advantages of the joint modeling framework of Muscle over single modality modeling or a baseline multi modality modeling for cell type delineation and elucidating associations between modalities.

Keywords

Single cell 3D genome, Single cell DNA methylation, Tensor decomposition, Block term tensor decomposition.

Testing of Hypotheses in Cancer Research: A Ranked Set Approach for Achieving Higher Efficiency

Sunil Mathur[1], **Ethan Burns**[1], **Shreya Mathur**[2], **Ravi Pingali**[1], **Jenny Chang**[1]

[1] *Houston Methodist Neal Cancer Center*
smathur2@houstonmethodist.org,eaburns@houstonmethodist.org, spingali@houstonmethodist.org, jccheng@houstonmethodist.org
[2] *Department of Internal Medicine, Boston Medical Center, Boston, MA, shreya.mathur@bmc.org*

Abstract

Triple-negative breast cancer (TNBC) has a higher rate of recurrence of 6.7–10.5 compared with an overall rate of 2.1–6.4 amongst all breast cancer, shorter times to recurrence, and intrinsic aggressive clinical course associated with worse prognosis and survival rate lower than non-TNBC patients. Testing drug efficacy for treating TNBC is one of the most challenging tasks in cancer research which can help in making informed medical decisions, support healthcare planning, and help patients in making informed healthcare decisions. That may lead to increased quality of care, avoidance of errors and adverse events, improved efficiency, increased cost-benefit, and higher provider and patient satisfaction. Given the significance of medicine with better efficacy, we propose a rank test using an empirical distribution function to detect the differences between the two drugs. It will also save time and cost in clinical trials. The test statistic is constructed as a power divergence between empirical distribution functions obtained from the two independent samples making it more powerful than its competitors under heavy-tailed and light-tailed distributions. The permutation principle is used to implement the test. Using the Monte Carlo method we computed empirical power which shows that our test performs better than its competitors under heavy-tailed, light-tailed, and

even elliptically asymmetric population distribution. Overall, the proposed test provides better power than its competitors considered here irrespective of the nature of the population.

Keywords

Power, divergence, cancer, rank-order, non-parametric test.

Predicting Patient Survival With Multi-Block Partial Lease Squares using Multi-Omics Data

Runzhi Zhang[1], **Susmita Datta**[2]

[1] *University of Florida, Biostatistics, USA, runzhi.zhang@ufl.edu*
[2] *University of Florida, University of Florida, USA, susmita.datta@ufl.edu*

Abstract

As high-throughput studies advance, more and more high-dimensional multi-omics data are available and collected from the same patient cohort. Using multi-omics data as predictors to predict the survival outcomes is critical but challenging due to the complex structure of such data. In this presentation, we introduce an adaptive sparse multi-block partial least square (asmbPLS) regression method by assigning different penalty factors to different blocks in different PLS components for feature selection and prediction. We compare the proposed method with several competitive algorithms in many aspects. The performance and the efficiency of our method are demonstrated using both the simulated and a real data which includes progression-free survival (PFS) interval/status, clinical variables, and various types of omics data collected for 167 melanoma patients. In addition, an R package called asmbPLS implementing this method is made publicly available on Github.

Keywords

PLS, RNA-seq, proteomics, survival.

One-Shot Federated Conformal Prediction

Pierre Humbert[1], **Batiste Le Bars**[2], **Aurélien Bellet**[2], **Sylvain Arlot**[1,3]

[1] *Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France.*
[2] *Université Lille, Inria, CNRS, Centrale Lille, UMR 9189, CRIStAL, F-59000 Lille*
[3] *Institut Universitaire de France (IUF)*

Abstract

In this work, we introduce a conformal prediction method to construct prediction sets in a oneshot federated learning setting. More specifically, we define a quantile-of-quantiles estimator and prove that for any distribution, it is possible to output prediction sets with desired coverage in only one round of communication. To mitigate privacy issues, we also describe a locally differentially private version of our estimator. Finally, over a wide range of experiments, we show that our method returns prediction sets with coverage and length very similar to those obtained in a centralized setting. Overall, these results demonstrate that our method is particularly well-suited to perform conformal predictions in a one-shot federated learning setting.

Keywords

Prediction set, conformal prediction, federated learning.

Reference

Pierre Humbert, Batiste Le Bars, Aurélien Bellet, and Sylvain Arlot (2023). One-Shot Federated Conformal Prediction. *ICML 2023 - Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. https://proceedings.mlr.press/v202/humbert23a.html

An RKHS Approach for Variable Selection in High-dimensional Functional Linear Models

Xingche Guo[1], **Yehua Li**[2], **Tailen Hsing**[3]

[1] *Columbia University, Department of Biostatistics, USA, xguo@iastate.edu*
[2] *University of California at Riverside, Department of Statistics, USA, yehuali@ucr.edu*
[3] *University of Michigan, Departments of Statistics, USA, thsing@umich.edu*

Abstract High-dimensional functional data has become increasingly prevalent in modern applications such as high-frequency financial data and neuroimaging data analysis. We investigate a class of high-dimensional linear regression models, where each predictor is a random element in an infinite dimensional function space, and the number of functional predictors p can potentially be much greater than the sample size n. Assuming that each of the unknown coefficient functions belongs to some reproducing kernel Hilbert space (RKHS), we regularized the fitting of the model by imposing a group elastic-net type of penalty on the RKHS norms of the coefficient functions. We show that our loss function is Gateaux sub-differentiable, and our functional elastic-net estimator exists uniquely in the product RKHS. Under suitable sparsity assumptions and a functional version of the irrepresentible condition, we establish the variable selection consistency property of our approach.

Keywords

Functional linear regression; Elastic-net penalty; Reproducing kernel Hilbert space; Model selection consistency; Sparsity.

Variational Inference Aided Variable Selection For Spatially Structured High Dimensional Covariates

S. Nandy[1], **M. Kim**[2], **S. Bhattacharya**[3], **T. Maiti**[3]

[1] *Case Western Reserve University, Department of Mathematics, Applied Mathematics and Statistics, USA*
[2] *Pusan National University, Department of Statistics, South Korea*
[3] *Michigan State University, Department of Statistics and Probability, USA*

Abstract

We consider the problem of Bayesian high dimensional variable selection under linear regression when a spatial structure exists among the covariates. We use an Ising prior to model the structural connectivity of the covariates with an undirected graph and the connectivity strength with Ising distribution parameters. Ising models which originated in statistical physics, is widely used in computer vision and spatial data modeling. Although a Gibbs solution to this problem exists, the solution involves the computation of determinants and inverses of high dimensional matrices rendering it unscalable to higher dimensions. Further, the lack of theoretical support limits this important tool's use for the broader community. This paper proposes a variational inference-aided Gibbs approach that enjoys the same variable recovery power as the standard Gibbs solution all the while being computationally scalable to higher dimensions. We establish strong selection consistency of our proposed approach together with its competitive numerical performance under varying simulation scenarios.

Keywords

Bayesian Variable Selection, Structured Covariates, Variational Bayes, Ising Distribution.

Iterative regularisation in ill-posed generalised linear models

Tatyana Krivobokova[1], **Gianluca Finocchio**[2]

[1] *University of Vienna, Department of Statistics and Operations Research, Austria, tatyana.krivobokova@univie.ac.at*
[2] *University of Vienna, Department of Statistics and Operations Research, Austria, gianluca.finocchio@univie.ac.at*

Abstract

We study the problem of regularized maximum-likelihood optimization in ill-posed generalized linear models with covariates that include subsets that are relevant and that are irrelevant for the response. It is assumed that the source of ill-posedness is a joint low dimensionality of the response and a subset of the relevant covariates in the sense of a latent factor generalized linear model (GLM). In particular, we propose a novel iteratively-reweighted-partial-least-squares (IRPLS) algorithm and show that it is better than any other projection or penalization-based regularisation algorithm. Under regularity assumptions on the latent factor GLM we show that the convergence rate of the IRPLS estimator with high probability is the same as that of the maximum likelihood estimator in our latent factor GLM, which is an oracle achieving an optimal parametric rate. Our findings are confirmed by numerical studies.

Keywords

Dimensionality reduction, latent factor models, partial least squares.

From Puzzle Pieces to Masterpiece: Connecting strengths between Risk Analysis, Incomplete Block Designs, Data Science, and Artificial Intelligence

Teresa A. Oliveira

Department of Sciences and Technology, Universidade Aberta, and Center of Statistics and it Applications, Portugal,
teresa.oliveira@uab.pt

Abstract

In today's rapidly advancing world, understanding complex systems and making informed decisions is crucial for ensuring organizational success. Traditional methodologies often fall short in capturing the intricacies of these systems, and thus, utilizing a combination of multiple methodologies has emerged as a powerful approach. This presentation highlights the importance and benefits of employing risk analysis, incomplete block designs, data science, and artificial intelligence (AI) in unison. Synergies arise from leveraging each methodology's unique strengths, compensating for their respective weaknesses. Traditional risk analysis approaches often oversimplify complex systems, overlooking crucial factors that influence outcomes and providing a framework for uncertainty quantification. Similarly, incomplete block designs have enabled efficient experimentation but may fail to fully account for the inherent variability in real-world phenomena. The Data Science assists in uncovering insights, while AI techniques augment predictive capabilities and, both areas, have emerged to address some of the previously mentioned limitations, harnessing the power of Big Data and Machine Learning Algorithms, among others. Connecting dots between these four methodologies may certainly lead the decision-makers to obtain a more comprehensive understanding of complex systems. This multidimensional analysis provides greater

confidence in decision-making, as it accounts for the interplay of multiple factors, uncertainties, and interactions. The result is an analytical masterpiece, where disparate puzzle pieces combine to reveal a holistic picture of the system at hand, and of eventual emerging ones.

Quantization based clustering: An iterative approach

Thomas Laloë[1]

[1] *Laboratoire J.A. Dieudonné, UMR CNRS 7351, Université Côte d'Azur, Nice, France*

Abstract

Clustering consists in partitioning a set of unlabeled objects into homogeneous groups (or clusters), so that the data in each subset share some common trait (see [1] for a thorough introduction to the subject) . Over the years, many methods have been proposed to deal with clustering : density based clustering, Hierarchical clustering, and partitioning clustering... We focus in this paper on the last one. More precisely we use a method coming from the signal compression theory : the quantization [4].
The proximity notion is crucial in the definition of what is a "good clustering". We propose here to rely on the method proposed in [2] which is based on a L_1 (or Manhattan) distance. The algorithm (called Alter) proposed to perform the clustering is proved to be consistent but suffers from a high complexity. A first alternative has been proposed in [3] to lower the complexity, adapting the X-means approach proposed in [5].
The purpose of this paper is to propose a new alternative to lower the complexity of the Alter algorithm (with respect to the number of clusters), best preserving its ability to converge to the global optimum.

Keywords

Quantization, Clustering, functional data.

References

[1] Kaufman, L. and Rousseeuw, P.J. (1990). Finding Groups in Data: an Introduction to Cluster Analysis. *John Wiley & Sons.*

[2] Laloë, T (2010). L_1 quantizationand clustering in Banach spaces. *Mathematical Methods of Statistics.*

[3] Laloë, Thomas and Servien, Rémi (2013). The X-Alter algorithm : a parameter-free method to perform unsupervised clustering. *Journal of Modern Applied Statistical Methods.*

[4] Linder, T. (2002). Learning-theoretic methods in vector quantization. *Principles of Nonparametric Learning.*

[5] Pelleg, D. and Moore, A. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *Proceedings of the Seventeenth International Conference on Machine Learning.*

Generalizing Conditional Independence: Nested Markov Models

Thomas S. Richardson[1], **Robin J. Evans**[2], **James M. Robins**[3], **Ilya Shpitser**[4]

[1] *University of Washington, Department of Statistics, USA, thomasr@uw.edu*
[2] *Oxford University, Department of Statistics, UK, evans@stats.ox.ac.uk*
[3] *Harvard School of Public Health, Department of Epidemiology, USA, robins@hsph.harvard.edu*
[4] *Johns Hopkins University, Department of Computer Science, USA, ilyas@cs.jhu.edu*

Abstract

It has been understood for more than 30 years that directed acyclic graph (DAG) models with hidden variables give rise to non-parametric ("Verma") constraints that generalize conditional independence. The nested Markov model is a class of graphical models associated with acyclic graphs containing directed ($\rightarrow$) and bidirected ($\leftrightarrow$) edges that encode all of the non-parametric equality constraints implied by DAGs with latent variables.

In this talk I will first motivated and define the nested Markov model and the associated global property. I will then describe simple rules for reasoning with such constraints. Finally, I will define a local property, outlining why this construction is harder than for ordinary independence models.

Keywords

Graphical models, directed acylic graph models, hidden variable models, conditional independence, causal inference.

Lower complexity adaptation of empirical optimal transport

Thomas Staudt[1], **Shayan Hundrieser**[2], **Axel Munk**[3]

[1] *University of Göttingen, Institute for Mathematical Stochastics, Germany, thomas.staudt@uni-goettingen.de*
[2] *University of Göttingen, Institute for Mathematical Stochastics, Germany, s.hundrieser@math.uni-goettingen.de*
[3] *University of Göttingen, Institute for Mathematical Stochastics, Germany, munk@math.uni-goettingen.de*

Abstract

This talk summarizes a number of recent findings regarding the rate of convergence of the empirical optimal transport cost to its population counterpart. It discusses the roles that (a) the regularity properties of the ground cost function, (b) the intrinsic dimensionality of the involved population measures, and (c) their tail behavior (i.e., moment conditions) each play in governing the actual convergence rate. Two areas of particular focus will be the property of lower complexity adaptation, as well as the extension of convergence results from compact spaces to unbounded domains. The former complements our understanding of (b) in an important way, as it describes that the statistical performance of empirical optimal transport costs between two different measures, which might be supported on different spaces, is determined by the "simpler" measure, no matter which of the two distributions is sampled from. The latter answers (c) in great generality by providing (nearly) sharp moment conditions that stem from a generic decomposition approach in the primal formulation of optimal transport.

Keywords

Optimal transport, Wasserstein distance, curse of dimensionality, moment conditions.

Modeling Time-Varying Effects of Mobile Health Interventions Using Longitudinal Functional Data from HeartSteps Micro-Randomized Trial

Jiaxin Yu[1], **Predrag Klasnja**[2], **Susan A. Murphy**[3], **Tianchen Qian**[4]

[1] *University of California, Irvine, Department of Statistics, USA, jiaxiny4@uci.edu*
[2] *University of Michigan, School of Information, USA, klasnja@umich.edu*
[3] *Harvard University, Department of Statistics and Department of Computer Science, USA, samurphy@fas.harvard.edu*
[4] *University of California, Irvine, Department of Statistics, USA, t.qian@uci.edu*

Abstract

Understanding how the effect of a mobile health intervention varies over time and with contextual information is critical for both optimizing the intervention and advancing domain knowledge. This analysis aims to assess how a push notification suggesting physical activity influences individuals' step count and how such influence varies over time, using data from the HeartSteps micro-randomized trial (MRT). The statistical challenges include the time-varying treatments and the longitudinal functional step count measurements. We propose the first semiparametric causal excursion effect model with varying coefficients to model the time-varying effects within a decision point and across decision points in an MRT. The proposed model incorporates double time indices to accommodate the longitudinal functional outcome, and it enables the assessment of time-varying effect moderation by contextual variables. We propose a two-stage causal effect estimator that is robust against a misspecified high-dimensional outcome regression model. We establish asymptotic theory and conduct simulation studies

to validate the proposed estimator. Our analysis provides new insights into individuals' change in response profiles (such as how soon a response occurs) due to the activity suggestions, how such changes differ by the type of suggestion they receive, and how such changes depend on other contextual information such as real-time location.

Keywords

Causal inference, longitudinal data, varying coefficient model.

Semi-supervised Triply Robust Inductive Transfer Learning

Tianxi Cai[1], **Mengyan Li**[2], **Molei Liu**[3]

[1] *Harvard University, Department of Biostatistics, USA,*
tcai@hsph.harvard.edu
[2] *Bentley University, Department of Mathematical Sciences, USA,*
mengyanli@bentley.edu
[3] *Columbia University, Department of Biostatistics, USA,*
ml4890@cumc.columbia.edu

Abstract

In this work, we propose a **S**emi-supervised **T**riply **R**obust **I**nductive trans**F**er **LE**arning (STRIFLE) approach, which integrates heterogeneous data from a label-rich source population and a label-scarce target population and utilizes a large amount of unlabeled data simultaneously to improve the learning accuracy in the target population. Specifically, we consider a high dimensional covariate shift setting and employ two nuisance models, a density ratio model and an imputation model, to combine transfer learning and surrogate-assisted semi-supervised learning strategies organically and achieve triple robustness. While the STRIFLE approach requires the target and source populations to share the same conditional distribution of outcome Y given both the surrogate features $\mathbf{S}$ and predictors $\mathbf{X}$, it allows the true underlying model of $Y \mid \mathbf{X}$ to differ between the two populations due to the potential covariate shift in $\mathbf{S}$ and $\mathbf{X}$. Different from double robustness, even if both nuisance models are misspecified or the distribution of $Y \mid \mathbf{S}, \mathbf{X}$ is not the same between the two populations, when the transferred source population and the target population share enough similarities, the triply robust STRIFLE estimator can still partially utilize the source population, and it is guaranteed to be no worse than the target-only surrogate-assisted semi-supervised estimator with negligible errors. These desirable properties of our

estimator are established theoretically and verified in finite samples via extensive simulation studies. We utilize the STRIFLE estimator to train a Type II diabetes polygenic risk prediction model for the African American target population by transferring knowledge from electronic health records linked genomic data observed in a larger European source population.

Keywords

Covariate shift, surrogate-assisted semi-supervised learning, high dimensional data, inductive transfer learning, model misspecification, triple robustness.

MDI+: A Flexible Random Forest-Based Feature Importance Framework

Abhineet Agarwal[1], **Ana M. Kenney**[2], **Yan Shuo Tan**[3], **Tiffany M. Tang**[4], **Bin Yu**[5]

[1] *University of California, Berkeley, Department of Statistics, USA, aa3797@berkeley.edu*

[2] *University of California, Irvine, Department of Statistics, USA, akenney1@uci.edu*

[3] *National University of Singapore, Department of Statistics and Data Science, Singapore, yanshuo@nus.edu.sg*

[4] *University of Michigan, Department of Statistics, USA, tmtang@umich.edu*

[5] *University of California, Berkeley, Department of Statistics, EECS, CCB, USA, binyu@berkeley.edu*

Abstract

Mean decrease in impurity (MDI) is a popular feature importance measure for random forests (RFs). We show that the MDI for a feature X_k in each tree in an RF is equivalent to the unnormalized R^2 value in a linear regression of the response on the collection of decision stumps that split on X_k. We use this interpretation to propose a flexible feature importance framework called MDI+. Specifically, MDI+ generalizes MDI by allowing the analyst to replace the linear regression model and R^2 metric with regularized generalized linear models (GLMs) and metrics better suited for the given data structure. Moreover, MDI+ incorporates additional features to mitigate known biases of decision trees against additive or smooth models. Extensive data-inspired simulations show that MDI+ significantly outperforms popular feature importance measures in identifying signal features. We also apply MDI+ to two real-world case studies on drug response prediction and breast cancer subtype classification. We show that MDI+ extracts well-established predictive genes with significantly greater stability compared to existing feature importance measures.

Keywords

Interpretable machine learning, explainable AI, decision trees, ensembles, non-parametrics.

Analysis and sample-size determination for 2^K audit experiments with binary response and application to identification of effect of racial discrimination on access to justice

Nicole Pashley[1], **Brian Libgober**[2], **Tirthankar Dasgupta**[3]

[1] *Rutgers University, Department of Statistics, USA,*
np755@stat.rutgers.edu
[2] *Northwestern University, Department of Political Science, USA,*
brian.libgober@northwestern.edu
[3] *Rutgers University, Department of Statistics, USA,*
td370@stat.rutgers.edu

Abstract

Social scientists have increasingly turned to audit experiments to investigate discrimination in the market for jobs, loans, housing and other opportunities. In a typical audit experiment, researchers assign "signals" (the treatment) to subjects at random and compare success rates across treatment conditions. In the recent past there has been an increased interest in using randomized multifactor designs for audit experiments, popularly called factorial experiments, in which combinations of multiple signals are assigned to subjects. However although social scientists (e.g., Libgober 2020) have manipulated multiple factors like race, gender and income, the analyses have been mostly exploratory in nature. In this paper we lay out a comprehensive methodology for analysis of 2^K factorial designs using binary response using the model-free, randomization-based Neymanian inference and demonstrate its application by analyzing the experiment reported in Libgober (2020). Specifically, we extend the results on 2^1 (two-armed) and 2^2 experiments with binary response to the case of 2^K experiments for any integer K, adding the following new elements to the

existing framework: (i) proposing multiple hypotheses testing procedures consistent with those in existing literature on model-based inference of factorial designs, (ii) developing a methodology for determining the sample size based on the level of power desired to identify active causal effects for future studies and (iii) defining non-linear factorial effects for binary responses, that can be considered as generalizations of the risk ratio and risk odds ratio, and proposing methods for their asymptotic inference.

Keywords

Causal Inference, Factorial designs, Neyman, Model-free, Design-based inference, Finite-population asymptotics.

Optimal Federated Learning for Nonparametric Function Estimation

T. Tony Cai, Abhinav Chakraborty, and Lasse Vuursteen

University of Pennsylvania, Department of Statistics and Data Science, USA

Abstract

Federated learning is a machine learning paradigm designed to tackle the challenges of data governance and privacy. It enables organizations (e.g., hospitals) to collaboratively train and enhance a shared global statistical model without sharing raw data externally. Instead, the learning process occurs locally at each participating entity, and only model characteristics, such as parameters and gradients, are exchanged, while preserving privacy. In this talk, we consider statistical optimality for federated learning in the context of nonparametric regression. The setting we study is heterogeneous, encompassing varying sample sizes and differential privacy constraints across different servers. Minimax optimal rates of convergence, up to logarithmic factors, are established for both global and pointwise estimation of the regression function. We propose distributed privacy-preserving estimation procedures and analyze their theoretical properties. The findings shed light on the delicate balance between accuracy and privacy preservation. In particular, we characterize the compromise not only in terms of the privacy budget but also concerning the loss incurred by distributing data within the privacy framework as a whole.

Keywords

Distributed Computation, Differential Privacy, Nonparametric Regression, Function Estimation.

Central Limit Theorems for Smooth Optimal Transport Maps

Tudor Manole[1], **Sivaraman Balakrishnan**[1,2], **Jonathan Niles-Weed**[3,4], **Larry Wasserman**[1,2]

[1] *Carnegie Mellon University, Department of Statistics and Data Science*
[2] *Carnegie Mellon University, Machine Learning Department*
[3] *New York University, Courant Institute of Mathematical Sciences*
[4] *New York University, Center for Data Science*
{tmanole,siva,larry}@stat.cmu.edu, jnw@cims.nyu.edu

Abstract

One of the central objects in the optimal transport framework is the optimal transport map: the transformation which pushes forward a source distribution onto a target distribution with minimal cost. Several recent works have analyzed the L^2 risk of plugin estimators of optimal transport maps, which are defined as the unique optimal transport map between density estimates of the underlying distributions. In this work, we show that such estimators enjoy pointwise central limit theorems. These results provide a first step toward performing statistical inference for smooth optimal transport maps in general dimension. We also derive a negative result, showing that these estimators cannot satisfy uniform central limit theorems when the dimension is sufficiently large. As a byproduct of our study, we show that the pointwise risk of our estimators is minimax optimal. Our proofs hinge upon a linearization of the Monge-Ampère equation, which allows us to reduce our problem to deriving limit laws for the solution of a uniformly elliptic partial differential equation with random coefficients.

Keywords

Optimal Transport, Nonparametric Inference, Pointwise Central Limit Theorem, Elliptic Partial Differential Equation.

The underlap coefficient: the concept and its need and Bayesian estimators

Zhaoxi Zhang[1], **Vanda Inácio**[2], **Miguel de Carvalho**[3],

[1] *University of Edinburgh, School of Mathematics, UK,*
z.zhang-156@sms.ed.ac.uk
[2] *University of Edinburgh, School of Mathematics, UK,*
vanda.inacio@ed.ac.uk
[3] *University of Edinburgh, School of Mathematics, UK,*
miguel.decarvalho@ed.ac.uk

Abstract

The first step when evaluating a diagnostic test is to determine the variation in its values across different disease groups. In the three-class disease setting, the volume under the receiver characteristic surface (VUS) and the three-class Youden index (YI) are the commonly used summary measures of a test's discriminatory ability. However, these measures are only appropriate under a stochastic ordering assumption for the distributions of test outcomes in the three groups. This assumption is stringent, not always plausible, and its violation can lead to incorrect conclusions about a test's performance to distinguish between the three classes. To address this, we propose the underlap coefficient, study its properties, as well as its relationship with the VUS and YI when a stochastic order is enforced. We further propose Bayesian nonparametric estimators for both the unconditional underlap coefficient and for its covariate-specific version. A simulation study reveals a good performance of the proposed estimators across a range of conceivable scenarios. We have applied our methods to an Alzheimer's disease (AD) dataset to assess how different potential AD biomarkers distinguish between individuals with normal cognition, mild impairment, and dementia, and how age and gender impact this discriminatory ability.

Keywords

Bayesian nonparametrics, covariate-adjustment, diagnostic test, (dependent) Dirichlet process mixtures, underlap coefficient.

Adaptive Bayesian Predictive Inference

Veronika Ročková[1]

[1] *University of Chicago, Booth School of Business, USA,*
Veronika.Rockova@ChicagoBooth.edu

Abstract

Bayesian predictive inference provides a coherent description of entire predictive uncertainty through predictive distributions. We examine several widely used sparsity priors from the predictive (as opposed to estimation) inference viewpoint. Our context is estimating a predictive distribution of a high-dimensional Gaussian observation with a known variance but an unknown sparse mean under the Kullback-Leibler loss. First, we show that LASSO (Laplace) priors are incapable of achieving rate-optimal performance. This new result contributes to the literature on negative findings about Bayesian LASSO posteriors. However, deploying the Laplace prior inside the Spike-and-Slab framework (for example with the Spike-and-Slab LASSO prior), rate-minimax performance can be attained with properly tuned parameters (depending on the sparsity level s_n). We highlight the discrepancy between prior calibration for the purpose of prediction and estimation. Going further, we investigate popular hierarchical priors which are known to attain adaptive rate-minimax performance for estimation. Whether or not they are rate-minimax also for predictive inference has, until now, been unclear. We answer affirmatively by showing that hierarchical Spike-and-Slab priors are adaptive and attain the minimax rate without the knowledge of s_n. This is the first rate-adaptive result in the literature on predictive density estimation in sparse setups. This finding celebrates benefits of a fully Bayesian inference.

Keywords

Asymptotic Minimaxity, Kullback-Leibler Loss, Predictive Densities, Sparse Normal Means.

Distributional Regression and Autoregression via Optimal Transport

Victor Panaretos[1] **and Laya Ghodrati**[2]

[1] *Institute of Mathematics, EPFL, Switzerland,*
victor.panaretos@epfl.ch
[2] *Research Center for Statistics, University of Geneva, Switzerland,*
laya.ghodrati@unige.ch

Abstract

We present a framework for performing regression when both covariate and response are probability distributions. Our framework is based on the theory of optimal transportation and links the conditional Fréchet mean of the response to the covariate via an optimal transport map. In the simplest context of distributions on the real line, we show how one can define and compute a Fréchet least squares estimator of the regression map, which attains the minimax convergence rate under minimal assumptions. We also discuss possible extensions to higher dimensional laws and to autoregressive settings, and how these seem to require additional smoothness conditions for consistent estimation.

Keywords

Fréchet regression; Wasserstein space.

Barycenters in metric spaces with non-positive curvature

Victor-Emmanuel Brunel[1], **Jordan Serres**[2]

[1] *ENSAE-CREST, Palaiseau, France,*
victor.emmanuel.brunel@ensae.fr
[2] *ENSAE-CREST, Palaiseau, France, jordan.serres@ensae.fr*

Abstract

Barycenters, aka Fréchet means, offer a natural extension of averages from linear spaces to general metric spaces. Just as in Euclidean spaces, limit theorems (such as laws of large numbers and central limit theorems) are well known under fairly general assumptions. In this talk, after a brief introduction of barycenters in metric spaces, I will present a framework - that of geodesic spaces with non-positive curvature - where non-asymptotic guarantees can be proven. This talk is inspired from a recent work in collaboration with Jordan Serres (ENSAE).

Keywords

Barycenters, Metric spaces, Geodesics, Alexandrov's curvature, NPC spaces, Concentration inequalities.

Boosting Census Data Privacy via Gaussian Differential Privacy, for FREE!

Weijie Su[1]

[1] *University of Pennsylvania, Department of Statistics and Data Science, USA, suw@wharton.upenn.edu*

Abstract

In the 2020 Decennial Census, differential privacy (DP) was used to protect the privacy of sensitive census data. While being a mathematically rigorous approach to privacy-preserving data analysis, DP often results in a poor tradeoff between privacy guarantees and statistical efficiency, particularly under composition. In this talk, we demonstrate how to achieve a better tradeoff using f-DP, through an application to decennial census data. Our results are based on a refined analysis of privacy losses under composition using Edgeworth expansions. Experimental results indicate that our new approach can achieve a roughly 10% decrease in the MSE of census data, while maintaining the same privacy guarantees as the method employed in the 2020 Decennial Census.

Keywords

Differential privacy, 2020 Decennial Census, Edgeworth expansions.

Bayesian biclustering and its application in education data analysis

Weining Shen[1]

[1] *University of California, Irvine, Department of Statistics, USA, weinings@uci.edu*

Abstract

We propose a novel nonparametric Bayesian IRT model that estimates clusters at the question level, while simultaneously allowing for heterogeneity at the examinee level under each question cluster, characterized by the mixture of Binomial distributions. The main contribution of this work is threefold. First, we present our new model and demonstrate that it is identifiable under a set of conditions. Second, we show that our model can correctly identify question-level clusters asymptotically, and the parameters of interest that measure the proficiency of examinees in solving certain questions can be estimated at a $\sqrt{n}$ rate (up to a log term). Third, we present a tractable sampling algorithm to obtain valid posterior samples from our proposed model. Compared to the existing methods, our model manages to reveal the multi-dimensionality of the examinees' proficiency level in handling different types of questions parsimoniously by imposing a nested clustering structure. The proposed model is evaluated via a series of simulations as well as apply it to an English proficiency assessment data set. This data analysis example nicely illustrates how our model can be used by test makers to distinguish different types of students and aid in the design of future tests.

Keywords

Model averaging, Nonparametric Bayes, Posterior contraction rate, Rasch model.

Counting the unseen: Estimation of susceptibility proportions in zero-inflated models using a conditional likelihood approach

Wen-Han Hwang[1], **Lu-Fang Chen**[2], **Jakub Stoklosa**[3]

[1] *National Tsing Hua University, Taiwan, wenhan@stat.nthu.edu.tw*
[2] *Minghsin University of Science and Technology, Taiwan*
[3] *The University of New South Wales, Australia*

Abstract

Zero-inflated count data models are widely used in various fields such as ecology, epidemiology, and transportation, where count data with a large proportion of zeros is prevalent. Despite their widespread use, their theoretical properties have not been extensively studied. This study aims to investigate the impact of ignoring heterogeneity in event count intensity and susceptibility probability on zero-inflated count data analysis within the zero-inflated Poisson framework. To address this issue, we propose a novel conditional likelihood approach that utilizes positive count data only to estimate event count intensity parameters and develop a consistent estimator for estimating the average susceptibility probability. Our approach is compared with the maximum likelihood approach, and we demonstrate our findings through a comprehensive simulation study and real data analysis. The results can also be extended to zero-inflated binomial, geometric, and negative binomial models with similar conclusions. These findings contribute to the understanding of the theoretical properties of zero-inflated count data models and provide a practical approach to handling heterogeneity in such models.

Keywords

Conditional likelihood, Intensity heterogeneity, Susceptibility probability, Zero-inflated Poisson distribution.

Approximate Inferences for Bayesian Hierarchical Generalized Linear Regression Models

Brandon Berman[1], **Wesley Johnson**[2], **Weining Shen**[3]

[1] *Sandia Laboratories, USA, bjberma@sandia.gov*
[2] *UC Irvine, Statistics, USA, wjohnson@uci.edu*
[3] *UC Irvine, Statistics, USA, wjohnson@uci.edu*

Abstract

Generalized linear mixed regression models are fundamental in Statistics. There are many methods and statistical packages available for analyzing data using these models; most require some form of numerical or analytic approximation since the likelihood function generally involves intractable integrals over the latents. The Bayesian approach avoids this issue by iteratively sampling the full conditional distributions for various blocks of parameters and latents. Depending on the choice of prior, some full conditionals are recognizable while others are not. In this paper we develop a novel normal approximation for the random effects full conditional, establish its asymptotic correctness and evaluate how well it performs. We focus on several families of GLMMs, and also develop a sufficient reduction (SR) approach to the MCMC algorithm. We illustrate with a three level hierarchical binomial regression model for data on health outcomes for patients who are clustered within physicians who are clustered within particular hospitals or hospital systems.

Keywords

Markov chain Monte Carlo, Asymptotic approximation, Gibbs Sampling.

Statistical modelling of irregularly observed astronomical time series data

Susana Eyheramendy[1,3], **Felipe Elorrieta**[2,3], **Wilfredo Palma**[3]

[1] *Universidad Adolfo Ibañez, Faculty of Engineering and Science, susana.eyheramendy@uai.cl*
[2] *Universidad de Santiago, Department of Mathematics and Computer Science, Chile, felipe.elorrieta@usach.cl*
[3] *Millennium Institute of Astrophysics, MAS, Chile, wilfredo.palma@gmail.com*

Abstract

Most time series methods assume that the data are observed regularly. However, this assumption does not hold in many diverse scientific fields such as astronomy, finance, climatology, among others. There are some techniques to fit unequally spaced time series, including for example the continuous-time autoregressive moving average (CARMA) processes. In this talk, we discuss a number of discrete-time processes that allow for the handling of irregularly observed time series data, including the irregular autoregressive (IAR) model, the complex irregular autoregressive (CIAR) model which allows the estimation of both positive and negative autocorrelations, the iARMA processes and their multivariate counterparts. We show that the models proposed are weakly stationary and that they can be represented by state-space systems. This allows for the efficient implementation of maximum likelihood estimation (MLE) procedures based on the Kalman recursions that are stable under some conventional assumptions. Furthermore, we show via Monte Carlo simulations that the finite sample performance of the parameter estimation is accurate. Applications of these methodologies are illustrated by modelling astronomical data consisting of residuals from the harmonic filtering of light curves of variable stars with positive and negative serial dependencies along with other real-life time series data.

Keywords

Autoregressive models, Time series, Light curves.

Randomization Tests and Causal Inference for Randomized Clinical Trials

William Fisher Rosenberger[1]

[1] *Department of Statistics, George Mason University, Fairfax, VA, USA*

Abstract

Any inference procedure which assumes random sampling from a population ignores Fisherian principles regarding the analysis of designed experiments. And clinical trials are the quintessential designed experiment. While we hear quite often about preservation of type I error rates and, more recently, about causal inference, these are natural elements of a randomization test. We discuss these issues and demonstrate that randomization tests can be used for more complex settings, such as multiple (> 2) treatment comparisons, analyses with missing outcome data, and subgroup analyses. It is interesting to note that the only cohort of statisticians NOT excited about randomization tests in this age of causal inference are the designers and conductors of randomized clinical trials!

Keywords

Estimands, Missing Data, Preservation of type I error rates, Randomization as a basis for inference.

Inference for Topological Data Analysis

Bejamin Roycraft[1], **Johannes Krebs**[2], **Wolfgang Polonik**[3]

[1] *University of California, Davis, Department of Statistics, USA, btroycraft@ucdavis.edu*
[2] *Catholic University of Eichstätt-Ingolstadt, Department of Mathematics, Germany, johannes.krebs@ku.de*
[3] *University of California, Davis, Department of Statistics, USA, wpolonik@ucdavis.edu*

Abstract

This talk presents some novel contributions to statistical inference for Topological Data Analysis (TDA). The presented inference methods consist of bootstrap based confidence regions for (persistent) Betti numbers and Euler characteristic curves. In contrast to most of the other existing inference methods for TDA, our methods are based on one data set of size n, and large sample guarantees are thus established for n tending to infinity. The presented results provide insights into how the sampling distribution affects the persistence diagram.

Keywords

stabilization, rate of normal approximation.

An ODE Model for Dynamic Matching in Heterogeneous Networks

Xiaowu Dai[1] **and Hengzhi He**[2]

[1] *UCLA, Department of Statistics and Data Science, and Department of Biostatistics, USA, e-mail: dai@stat.ucla.edu*
[2] *UCLA, Department of Statistics and Data Science, USA, e-mail: hengzhihe2022@163.com*

Abstract

We study the problem of dynamic matching in heterogeneous networks, where agents are subject to compatibility restrictions and stochastic arrival and departure times. In particular, we consider networks with one type of easy-to-match agents and multiple types of hard-to-match agents, each subject to its own compatibility constraints. Such a setting arises in many real-world applications, including kidney exchange programs and carpooling platforms. We introduce a novel approach to modeling dynamic matching by establishing the ordinary differential equation (ODE) model, which offers a new perspective for evaluating various matching algorithms. We study two algorithms, namely the Greedy and Patient Algorithms, where both algorithms prioritize matching compatible hard-to-match agents over easy-to-match agents in heterogeneous networks. Our results demonstrate the trade-off between the conflicting goals of matching agents quickly and optimally, offering insights into the design of real-world dynamic matching systems. We provide simulations and a real-world case study using data from the Organ Procurement and Transplantation Network to validate theoretical predictions.

Keywords

Dynamic matching; Heterogeneous networks; Greedy Algorithm; Kidney exchange.

Ensemble methods for testing a global null

Yaowu Liu[1], Zhonghua Liu[2], Xihong Lin[3]

[1]*Southwestern University of Finance and Economics, Department of Statistics, China, yaowuliu615@gmail.com*
[2] *Columbia University, Department of Biostatistics, USA, zl2509@cumc.columbia.edu*
[3] *Department of Biostistics and Department of Statistics, Harvard University, USA, xlin@hsph.harvard.edu*

Abstract

Testing a global null is a canonical problem in statistics and has a wide range of applications. In view of the fact that no uniformly most powerful test exists, prior and/or domain knowledge are commonly used to focus on a certain class of alternatives to improve the testing power. However, it is generally challenging to develop tests that are particularly powerful against a certain class of alternatives. In this paper, motivated by the success of ensemble learning methods for prediction or classification, we propose an ensemble framework for testing that mimics the spirit of random forests to deal with the challenges. Our ensemble testing framework aggregates a collection of weak base tests to form a final ensemble test that maintains strong and robust power for global nulls. We apply the framework to four problems about global testing in different classes of alternatives arising from Whole Genome Sequencing (WGS) association studies. Specific ensemble tests are proposed for each of these problems, and their theoretical optimality is established in terms of Bahadur efficiency. Extensive simulations and an analysis of a real WGS dataset are conducted to demonstrate the type I error control and/or power gain of the proposed ensemble tests.

Keywords

Bahadur efficiency; Cauchy P-value combination methods; Random weights; Robust test; Whole genome sequencing studies.

A Tree-based Bayesian Accelerated Failure Time Cure Model for Estimating Heterogeneous Treatment Effect

Rongqian Sun[1], **Xinyuan Song**[2]

[1] *The Chinese University of Hong Kong, Department of Statistics, Hong Kong, sunrq@link.cuhk.edu.hk*
[2] *The Chinese University of Hong Kong, Department of Statistics, Hong Kong, xysong@sta.cuhk.edu.hk*

Abstract

Estimating heterogeneous treatment effects has drawn increasing attention in medical studies, considering that patients with divergent features can undergo a different progression of disease even with identical treatment. Such heterogeneity can co-occur with a cured fraction for biomedical studies with a time-to-event outcome and further complicates the quantification of treatment effects. This study considers a joint framework of Bayesian causal forest and accelerated failure time cure model to capture the cured proportion and treatment effect heterogeneity through three separate Bayesian additive regression trees. Under the potential outcomes framework, conditional and sample average treatment effects within the uncured subgroup are derived on the scale of log survival time subject to right-censoring, and treatment effects on the scale of survival probability are derived for each individual. Bayesian backfitting Markov chain Monte Carlo algorithm with the Gibbs sampler is conducted to estimate the causal effects. Simulation studies show the satisfactory performance of the proposed method. The proposed model is then applied to a breast cancer dataset extracted from the SEER database to demonstrate its usage in detecting heterogeneous treatment effects and cured subgroups. Combined with popular mitigation strategies, the proposed method can also alleviate confounding induced by immortal time bias.

Keywords

Bayesian additive regression trees, cured subgroup, heterogeneous treatment effect, nonparametric methods, survival outcome.

So Many Jumps, So Few News

Yacine Aït-Sahalia[1], **Chen Xu Li**[2], **Chenxu Li**[3]

[1] *Princeton University, Department of Economics and Bendheim Center for Finance, USA, yacine@princeton.edu*
[2] *Renmin University of China, School of Business, P.R. China, lichenxu@rmbs.ruc.edu.cn*
[3]*Peking University, Guanghua School of Management, P.R. China, cxli@gsm.pku.edu.cn*

Abstract

This paper relates jumps in high frequency stock prices to firm-level, industry and macroeconomic news, in the form of machine-readable releases from Thomson Reuters News Analytics. We begin by examining the relationship from news to price jumps. We find that relevant news, both idiosyncratic and systematic, gets incorporated quickly into prices, as market efficiency suggests. However, in the reverse direction, the situation is different: the vast majority of price jumps do not have identifiable public news that can explain them. We then analyze the various market microstructure features that lead to jumps without news.

Keywords

Machine-readable news, stock price jumps, market efficiency, excess jumps, liquidity, price impact.

Reidentification Risk in Panel Data: Protecting for k-Anonymity

Shaobo Li[1], **Matthew Schneider**[2], **Yan Yu**[3], **Sachin Gupta**[4]

[1] *University of Kansas, School of Business, USA, shaobo.li@ku.edu*
[2] *Drexel University, LeBow College of Business, USA, mjs624@drexel.edu*
[3] *University of Cincinnati, Lindner College of Business,USA,yan.yu@uc.edu*
[4] *Cornell University, Johnson College of Business, USA, sg248@cornell.edu*

Abstract

We consider the risk of re-identification of panelists in marketing research data that are widely used to obtain insights into buyer behavior and to develop marketing strategy. We find that 17% to 94% of the panelists in 15 frequently bought consumer goods categories are subject to high risk of reidentification through a potential record linkage attack based on their unique purchasing histories, even when their identities have been anonymized. We first demonstrate that the risk of reidentification is vastly understated by unicity, the conventional measure. Instead, we propose a new measure of reidentification risk, termed sno-unicity, that accounts for the longitudinal nature of panel data and show that it is much larger than unicity. To protect the privacy of panelists we consider the well-known privacy notion of k-anonymity, and develop a new approach called graph-based minimum movement k-anonymization (k-MM) that is designed especially for panel data. The proposed k-MM approach can be formulated as an optimization problem where the objective is to minimally distort variables in the original data based on weights which users pre-specify corresponding to their use case. We further show how our approach can be extended to achieve l-diversity. We apply the k-MM approach

to two different panel datasets that are widely used in marketing research. To achieve a given privacy level, compared to several benchmark protection methods, the protected data from our method result in the least distortion in inferences about key marketing metrics such as brand market shares, share of category requirements, brand switching rates, and marketing-mix parameters estimated from a hierarchical Bayesian brand choice model.

Keywords

Brand choice, data privacy, data sharing, optimization.

ML-Powered Outlier Detection: False Discovery Rate Control and Derandomization

Yaniv Romano

Technion – Israel Institute of Technology
Departments of ECE and CS
Israel
yromano@technion.ac.il

Abstract

Outlier detection, also known as out-of-distribution or anomaly detection, stands as a significant challenge in statistics and machine learning. Given a collection of test observations, the primary objective is to identify outlier samples that diverge from the distribution of a reference inlier dataset. Powerful machine learning algorithms are available for this task, with key applications such as fraud detection and epileptic seizure detection. However, these algorithms often lack error-controlling guarantees. A high and uncontrolled error rate can lead to unnecessary investigations of legitimate transactions, degraded user experiences, and, in medical applications, can expose patients to unnecessary procedures. In this talk, we will delve into recent advancements in this field, highlighting how conformal inference plays a pivotal role in creating outlier detection algorithms that control the false discovery rate. After outlining the advantages of using conformal p-values for this task, we will address an inherent limitation of this approach: its randomized nature. Such randomness often leads to different outcomes when analyzing the same test data, complicating the interpretation of findings. To alleviate this issue, we will present a principled solution to make conformal inferences more stable by leveraging suitable conformal e-values instead of p-values to quantify statistical significance. This talk navigates the landscape of machine learning and multiple hypothesis testing to ensure that conclusions, extracted

from any complex outlier detection model, are reliable, stable, and reproducible.

Link to relevant papers: paper #1, paper #2.

Keywords

Outlier Detection, Conformal Prediction, False Discovery Rate.

Robust Estimation in Exponential Families

Yannick Baraud[1], **Juntong Chen**[2]

[1] *Luxembourg University, DMATH, Luxembourg, yannick.baraud@uni.lu*
[2] *University of Twente, Netherlands, juntong.chen@utwente.nl*

Abstract

We observe a finite number of pairs of random variables that are presumed to be i.i.d. and we consider the problem of estimating the conditional distribution of the second coordinate given the first. We model this conditional distribution as an element of a given single-parameter exponential family for which the value of the parameter is an unknown function of the first coordinate of the pair. We provide an estimator of the conditional distribution based on our observations and analyse its performance not only when the statistical model is exact, as commonly done in statistics, but also when it is possibly misspecified (the pairs are independent but not exactly i.i.d., the true conditional distribution does not belong to the chosen exponential family, etc). We establish non-asymptotic risk bounds and show that our estimator is robust to a possible departure from the hypotheses we started from. Finally we provide an algorithm to compute the estimator in low or medium dimensions and compare its performance to that of the celebrated maximum likelihood estimator.

Keywords

Robust Estimation, Regression Function, Logistic Regression, Exponential Family.

Beyond MLE: Monotone Variational Inequality for Statistical Estimation

Yao Xie[1]

[1] *Georgia Institute of Technology, H. Milton Stewart School of Industrial and Systems Engineering, USA, yao.xie@isye.gatech.edu*

Abstract

We present a new statistical model estimation method based on solving monotone operator Variational Inequality (VI) and converting certain non-convex optimization problems into convex problems and achieve better statistical performance and fast algorithm convergence, inspired by a seminal work of (Juditsky & Nemirovsky, 2019). The method can be applied to nonlinear regression, generalized linear models (GLM), and single-index models, utilizing the problem structures. We demonstrate the algorithm's performance using numerical simulations and a real data examples.

Keywords

Statistical Estimation, Monotone Variational Inequality, Convex Optimization.

Geometric Exploration of Random Objects Through Optimal Transport

Paromita Dubey[1*], **Yaqing Chen**[2*], **Hans-Georg Müller**[3]

[1] *Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, United States, paromita@marshall.usc.edu*

[2] *Department of Statistics, Rutgers University, United States, yqchen@stat.rutgers.edu*

[3] *Department of Statistics, University of California, Davis, United States, hgmueller@ucdavis.edu*

Abstract

We propose new tools for the geometric exploration of data objects taking values in a general separable metric space. For a random object, we first introduce the concept of depth profiles. Specifically, the depth profile of a point in a metric space is the distribution of distances between the very point and the random object. Depth profiles can be harnessed to define transport ranks based on optimal transport, which capture the centrality and outlyingness of each element in the metric space with respect to the probability measure induced by the random object. We study the properties of transport ranks and show that they provide an effective device for detecting and visualizing patterns in samples of random objects. In particular, we establish the theoretical guarantees for the estimation of the depth profiles and the transport ranks for a wide class of metric spaces, followed by practical illustrations.

Keywords

Metric space valued data; transport rank; center-outward ordering; visualization of random objects.

[*]Contributed equally.

Sparse topic modeling via spectral decomposition and thresholding

Huy D Tran[1], **Yating Liu**[2], **Claire Donnat**[3]

[1] *University of Chicago, Department of Statistics, United States, huydtran@uchicago.edu*
[2] *University of Chicago, Department of Statistics, United States, yatingliu@uchicago.edu*
[3] *University of Chicago, Department of Statistics, United States, cdonnat@uchicago.edu*

Abstract

By modeling documents as mixtures of topics, Topic Modeling allows the discovery of latent thematic structures within large text corpora, and has played an important role in natural language processing over the past decades. Beyond text data, topic modeling has proven itself central to the analysis of microbiome data, population genetics, or, more recently, single-cell spatial transcriptomics. Given the model's extensive use, the development of estimators — particularly those capable of leveraging known structure in the data — presents a compelling challenge.

In this talk, we focus more specifically on the probabilistic Latent Semantic Indexing model, which assumes that the expectation of the corpus matrix is low-rank and can be written as the product of a topic-word matrix and a word-document matrix. Although various estimators of the topic matrix have recently been proposed, their error bounds highlight a number of data regimes in which the error can grow substantially — particularly in the case where the size of the dictionary p is large. In this talk, we propose studying the estimation of the topic-word matrix under the assumption that the ordered entries of its columns rapidly decay to zero. This sparsity assumption is motivated by the empirical observation that the word frequencies in a text often adhere to Zipf's law. We introduce a new spectral procedure for estimating the topic-word matrix that thresholds words based on their

corpus frequencies, and show that its ℓ_1-error rate under our sparsity assumption depends on the vocabulary size p only via a logarithmic term. Our error bound is valid for all parameter regimes and in particular for the setting where p is extremely large; Our procedure also empirically performs well relative to well-established methods when applied to a large corpus of research paper abstracts, as well as the analysis of single-cell and microbiome data where the same statistical model is relevant but the parameter regimes are vastly different.

Keywords

Topic Models; Sparsity; High-dimensional Statistics.

Quantum Machine Learning

Yazhen Wang[1]

[1] *University of Wisconsin-Madison, Department of Statistics, USA,*
yzwang@stat.wisc.edu

Abstract

Quantum computation and quantum information are of great current interest across various fields, including computer science, mathematics and statistics, physical sciences and engineering. As the theory of quantum physics is fundamentally stochastic, quantum computation and quantum information are inherently infused with elements of randomness and uncertainty. Consequently quantum algorithms are random in nature. This highlights the important role for statistics to play in the realm of quantum computation, which in turn offers great potential to revolutionize machine learning and computational statistics. In this talk, I will provide an overview of quantum computation and quantum information, covering the fundamental concepts and exploring quantum advantage along with the role of statistics and machine learning and the implications for machine learning and statistics.

Keywords

Machine learning, quantum, statistics.

Partially-Global Fréchet Regression

Danielle C. Tucker[1], **Yichao Wu**[2]

[1] *University of Illinois at Chicago, Department of Mathematics, Statistics and Computer Science, USA, dtucke5@uic.edu*
[2] *University of Illinois at Chicago, Department of Mathematics, Statistics and Computer Science, USA, yichaowu@uic.edu*

Abstract

We propose a partially-global Fréchet regression model by extending the profiling technique for the partially linear regression model (Severini and Wong, 1992). This extension allows for the response to come from a generic metric space and can incorporate a combination of Euclidean predictors and a predictor which comes from another generic metric space. By melding together the local and global Fréchet regression models proposed by Petersen and Müller (2019), we gain a model that is more flexible than global Fréchet regression and more accurate than local Fréchet regression when the data generating process relies on a non-Euclidean predictor or is truly "global (linear)" for some scalar predictors. In this paper, we provide theoretical support for partially-global Fréchet regression and demonstrate its competitive finite-sample performance when applied to both simulated data and to real data which is too complex for traditional statistical methods.

Keywords

Conditional Fréchet mean, metric space, partially linear.

Empirical Bayes estimation: When does g-modeling beat f-modeling in theory (and in practice)?

Yandi Shen[1], **Yihong Wu**[2]

[1] *Department of Statistics and Data Science, Yale University, New Haven, CT 06511, USA, yandi.shen@yale.edu*
[2] *Department of Statistics and Data Science, Yale University, New Haven, CT 06511, USA, yihong.wu@yale.edu*

Abstract

Empirical Bayes (EB) is a popular framework for large-scale inference that aims to find data-driven estimators to compete with the Bayesian oracle who knows the true prior. Two principled approaches to EB estimation have emerged over the years: $f-$*modeling*, which constructs an approximate Bayes rule by estimating the marginal distribution of the data, and g-modeling, which estimates the prior from data and then applies the learned Bayes rule. For the Poisson model, the prototypical examples are the celebrated Robbins estimator and the nonparametric MLE (NPMLE), respectively. It has long been recognized in practice that the Robbins estimator, while being conceptually appealing and computationally simple, lacks robustness and can be easily derailed by "outliers" (data points that were rarely observed before), unlike the NPMLE which provides more stable and interpretable fit thanks to its Bayes form. On the other hand, not only do the existing theories shed little light on this phenomenon, but they all point to the opposite, as both methods have recently been shown optimal in terms of the regret (excess over the Bayes risk) for compactly supported and subexponential priors with exact logarithmic factors. In this paper we provide a theoretical justification for the superiority of NPMLE over Robbins for heavy-tailed data by considering priors with bounded pth moment previously studied for the Gaussian model.

For the Poisson model with sample size n, assuming $p > 1$ (for otherwise triviality arises), we show that the NPMLE with appropriate regularization and truncation achieves a total regret $\tilde{\Theta}(n^{\frac{3}{2p+1}})$, which is minimax optimal within logarithmic factors. In contrast, the total regret of Robbins estimator (with similar truncation) is $\tilde{\Theta}(n^{\frac{3}{p+2}})$ and hence suboptimal by a polynomial factor. As a by-product of our regret analysis, we obtain the minimax Hellinger rate of estimating Poisson mixture density over the moment class, which may be of independent interest.

Keywords

Empirical Bayes, mixture model, regret, nonparametric MLE, Robbins.

A Double Projection Approach for Safe and Efficient Semi-Supervised Data-Fusion

Yiming Li[1], Xuehan Yang[2], Ying Wei[3] and Molei Li[4]

[1] *Columbia University, Department of Biotatistics, U.S.A., e-mail*
[2] *Columbia University, Department of Biotatistics, U.S.A.,*
[3] *Columbia University, Department of Biotatistics, U.S.A., e-mail*
[4] *Columbia University, Department of Biotatistics, U.S.A., e-mail*

Abstract

Advances in data collection and transmission technologies have made larger amounts of data readily available. However, there are differences in the data collection capabilities of different data centers, or there are inevitable data missing. Many previous approaches to handling missing information have solely focused on either missing predictors or missing responses. We consider both types of missing and incorporate more information by projecting score functions into subsets, thus proposing algorithms that have ensured efficiency relative to the complete-case analysis. By generalizing the algorithm of this paper, it is promising to be able to handle more complex missing data structures in the future. This is joint work with Yiming Li, Xuehan Yang and Molei Liu.

Keywords:

Semi-Supervised Learning, Multisource Data Intergration, Missing data.

Diagnosing the role of observable distribution shift in scientific replications

Ying Jin[1], **Kevin Guo**[2], **Dominik Rothenhäusler**[3]

[1] *Stanford University, Department of Statistics, United States of America, ying531@stanford.edu*
[2] *Stanford University, Department of Statistics, United States of America, kxguo@alumni.stanford.du*
[3] *Stanford University, Department of Statistics, United States of America, rdominik@stanford.edu*

Abstract

Many researchers have identified distribution shift as a likely contributor to the reproducibility crisis in behavioral and biomedical sciences. The idea is that if treatment effects vary across individual characteristics and experimental contexts, then studies conducted in different populations will estimate different average effects. This paper uses "generalizability" methods to quantify how much of the effect size discrepancy between an original study and its replication can be explained by distribution shift on observed unit-level characteristics. More specifically, we decompose this discrepancy into "components" attributable to sampling variability (including publication bias), observable distribution shifts, and residual factors. We compute this decomposition for several directly-replicated behavioral science experiments and find little evidence that observable distribution shifts contribute appreciably to non-replicability. In some cases, this is because there is too much statistical noise. In other cases, there is strong evidence that controlling for additional moderators is necessary for reliable replication.

Keywords

Replicability, Distribution shift, Publication bias, Generalizability.

Spatio-temporal DeepKriging for Interpolation and Probabilistic Forecasting

Pratik Nag[1], **Ying Sun**[1], **Brian Reich**[2]

[1] *CEMSE Division, Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia, ying.sun@kaust.edu.sa*

[2] *Department of Statistics, North Carolina State University, Raleigh, USA*

Abstract

Gaussian processes (GP) and Kriging are widely used in traditional spatio-temporal modeling and prediction. These techniques typically presuppose that the data are observed from a stationary GP with a parametric covariance structure. However, processes in real-world applications often exhibit non-Gaussianity and nonstationarity. Moreover, likelihood-based inference for GPs is computationally expensive and thus prohibitive for large datasets. In this paper, we propose a deep neural network (DNN) based two-stage model for spatio-temporal interpolation and forecasting. Interpolation is performed in the first step, which utilizes a dependent DNN with the embedding layer constructed with spatio-temporal basis functions. For the second stage, we use Long Short-Term Memory (LSTM) and convolutional LSTM to forecast future observations at a given location. We adopt the quantile-based loss function in the DNN to provide probabilistic forecasting. Compared to Kriging, the proposed method does not require specifying covariance functions or making stationarity assumptions and is computationally efficient. Therefore, it is suitable for large-scale prediction of complex spatio-temporal processes. We apply our method to monthly PM2.5 data at more than 200,000 space-time locations from January 1999 to December 2022 for fast imputation of missing values and forecasts with uncertainties.

Keywords

Deep learning, Feature embedding, Long short-term memory forecasting, Quantile machine learning, Radial basis function, Spatio-temporal modeling.

ARK: Robust Knockoffs Inference with Coupling

Yingying Fan[1], **Lan Gao**[2], **Jinchi Lv**[3],

[1] *University of Southern California, USA, fanyingy@marshall.usc.edu*
[2] *University of Tennessee, USA, lgao13@utk.edu*
[3] *University of Southern California, USA, jinchilv@marshall.usc.edu*

Abstract

We investigate the robustness of the model-X knockoffs framework with respect to the misspecified or estimated feature distribution. We achieve such a goal by theoretically studying the feature selection performance of a practically implemented knockoffs algorithm, which we name as the approximate knockoffs (ARK) procedure, under the measures of the false discovery rate (FDR) and family wise error rate (FWER). The approximate knockoffs procedure differs from the model-X knockoffs procedure only in that the former uses the misspecified or estimated feature distribution. A key technique in our theoretical analyses is to couple the approximate knockoffs procedure with the model-X knockoffs procedure so that random variables in these two procedures can be close in realizations. We prove that if such coupled model-X knockoffs procedure exists, the approximate knockoffs procedure can achieve the asymptotic FDR or FWER control at the target level. We showcase three specific constructions of such coupled model-X knockoff variables, verifying their existence and justifying the robustness of the model-X knockoffs framework.

Keywords

Knockoffs inference; High dimensionality; Feature selection; False discovery rate control; Family-wise error rate control; Coupling; Robustness.

Transportation-Based Functional ANOVA and PCA for Covariance Operators

Valentina Masarotto[1], Victor Panaretos[2], Yoav Zemel[3]

[1] *Universiteit Leiden, Mathematisch Instituut, The Netherlands, v.masarotto@math.leideuniv.nl*
[2] *École polytechnique fédérale de Lausanne, Institut de Mathématiques, Switzerland, victor.panaretos@epfl.ch*
[3] *École polytechnique fédérale de Lausanne, Institut de Mathématiques, Switzerland, yoav.zemel@epfl.ch*

Abstract

We consider the problem of comparing several samples of stochastic processes with respect to their second-order structure, and describing the main modes of variation in this second order structure, if present. These tasks can be seen as an Analysis of Variance (ANOVA) and a Principal Component Analysis (PCA) of covariance operators, respectively. They arise naturally in functional data analysis, where several populations are to be contrasted relative to the nature of their dispersion around their means, rather than relative to their means themselves. We contribute a novel approach based on optimal (multi)transport, where each covariance can be identified with a a centred Gaussian process of corresponding covariance. By means of constructing the optimal simultaneous coupling of these Gaussian processes, we contrast the (linear) maps that achieve it with the identity with respect to a norm-induced distance. The resulting test statistic, calibrated by permutation, is seen to distinctly outperform the state-of-the-art, and to furnish considerable power even under local alternatives. This effect is seen to be genuinely functional, and is related to the potential for perfect discrimination in infinite dimensions. In the event of a rejection of the null hypothesis stipulating equality, a geometric interpretation of the transport maps allows us to construct a (tangent space) PCA revealing the main modes of variation.

As a necessary step to developing our methodology, we prove results on the existence and boundedness of optimal multitransport maps. These are of independent interest in the theory of transport of Gaussian processes. The transportation ANOVA and PCA are illustrated on a variety of simulated and real examples.

Keywords

Covariance Operator, K-sample Testing, Optimal Transport, Permutation Test.

Accommodating Time-Varying Heterogeneity in Risk Estimation under the Cox Model: A Transfer Learning Approach

Ziyi Li, Yu Shen[1] & Jing Ning

[1] *Department of Biostatistics, UT MD Anderson Cancer Center Houston, Texas*

Abstract

Cancer registries have been widely used in clinical research because of their easy accessibility and large sample size. To use cancer registry data as a complement to improve the estimation precision of individual risks of death for inflammatory breast cancer (IBC) patients at The University of Texas MD Anderson Cancer Center, we proposed to use transfer learning method for adaptive information borrowing. When transferring information for risk estimation based on the cancer registries (i.e., source cohort) to a single cancer center (i.e., target cohort), time-varying population heterogeneity needs to be appropriately acknowledged. However, there is no literature on how to adaptively transfer knowledge on risk estimation with time-to-event data from the source cohort to the target cohort while adjusting for time-varying differences in event risks between the two sources. Our goal is to address this statistical challenge by developing a transfer learning approach under the Cox proportional hazards model. To allow data-adaptive levels of information borrowing, we impose Lasso penalties on the discrepancies in regression coefficients and baseline hazard functions between the two cohorts, which are jointly solved in the proposed transfer learning algorithm. We develop a more accurate risk estimation model for the MD Anderson IBC cohort given various treatment and baseline covariates, while adaptively borrowing information from the National Cancer Database to improve risk assessment. Supplementary materials for this article are available online

De-confounding causal inference using latent multiple-mediator pathways

Yubai Yuan[1], **Annie Qu**[2]

[1] *The Pennsylvania State University, Department of Statistics, USA, yvy5509@psu.edu*
[2] *University of California Irvine, Department of Statistics, USA, aqu2@uci.edu*

Abstract

Causal effect estimation from observational data is one of the essential problems in causal inference. However, most estimation methods rely on the strong assumption that all confounders are observed, which is impractical and untestable in the real world. We develop a mediation analysis framework inferring the latent confounder for debiasing both direct and indirect causal effects. Specifically, we introduce generalized structural equation modeling that incorporates structured latent factors to improve the goodness-of-fit of the model to observed data, and deconfound the mediators and outcome simultaneously. One major advantage of the proposed framework is that it utilizes the causal pathway structure from cause to outcome via multiple mediators to debias the causal effect without requiring external information on latent confounders. In addition, the proposed framework is flexible in terms of integrating powerful nonparametric prediction algorithms while retaining interpretable mediation effects. In theory, we establish the nonparametric identification of both causal and mediation effects based on the proposed deconfounding method. Numerical experiments on both simulation settings and a normative aging study indicate that the proposed approach reduces the estimation bias of both causal and mediation effects.

Keywords

Causal identification, Generalized additive model, Latent factor modeling, Mediation analysis, Sequential ignorability.

Statistical Significance of Clustering for High Dimensional Data

Yufeng Liu[1]

[1]*University of North Carolina at Chapel Hill, yfliu@email.unc.edu*

Abstract

Clustering serves as a fundamental tool for exploratory data analysis, but a key challenge lies in determining the reliability of the clusters identified by these methods, differentiating them from artifacts resulting from natural sampling variations. In this talk, I will present statistical significance of clustering (SigClust) as a cluster evaluation tool for high dimensional data. To begin, we define a cluster as data originating from a single Gaussian distribution and frame the assessment of statistical significance of clustering as a formal testing procedure. Addressing the challenge of high-dimensional covariance estimation in SigClust, we employ a combination of invariance principles and a factor analysis model. I'll also discuss an enhanced SigClust using multidimensional scaling (MDS) on dissimilarity matrices. SigClust for hierarchical clustering will be presented as well. Simulations and real data, including cancer subtype analysis, validate SigClust's effectiveness in assessing clustering significance.

Keywords

Gaussian, Hierarchal clustering, MDS, PCA, Unsupervised learning.

Test for the existence of the residual spectrum with application to brain functional connectivity detection

Yuichi Goto[1], **Xuze Zhang**[2], **Benjamin Kedem**[3], **Shuo Chen**[4]

[1] *Kyushu University, Faculty of Mathematics, Japan, yuichi.goto@math.kyushu-u.ac.jp*
[2] *University of Maryland, Department of Mathematics and Institute for Systems Research, United States of America, xzhang51@umd.edu*
[3] *University of Maryland, Department of Mathematics and Institute for Systems Research, United States of America, bnk@umd.edu*
[4] *University of Maryland, Maryland Psychiatric Research Center, School of Medicine, United States of America, shuochen@umd.edu*

Abstract

Coherence is a similarity measure between two time series and takes the form of the time series extension of Pearson's correlation. However, only a linear relationship between two time series can be measured by coherence. In this talk, we introduce a residual spectrum in order to measure non-linear relationships and show the existence and uniqueness of the residual spectrum by decomposing the regression model we consider into orthogonal processes. Moreover, we propose a test for the existence of the residual spectrum and show that our proposed test has asymptotically correct size and is consistent. Finally, we highlight the utility of the residual spectrum in brain functional connectivity detection.

Keywords

Coherence, Spectral density, Time series, Frequency domain, fMRI.

Logistic Regression and Classification with non-Euclidean Covariates

Yinan Lin[1], **Zhenhua Lin**[2],

[1] *National University of Singapore, Department of Statistics and Data Science, Singapore, stayina@nus.edu.sg*
[2] *National University of Singapore, Department of Statistics and Data Science, Singapore, linz@nus.edu.sg*

Abstract

We introduce a logistic regression model for data pairs consisting of a binary response and a covariate residing in a non-Euclidean metric space without vector structures. Based on the proposed model we also develop a binary classifier for non-Euclidean objects. We propose a maximum likelihood estimator for the non-Euclidean regression coefficient in the model, and provide upper bounds on the estimation error under various metric entropy conditions that quantify complexity of the underlying metric space. Matching lower bounds are derived for the important metric spaces commonly seen in statistics, establishing optimality of the proposed estimator in such spaces. Similarly, an upper bound on the excess risk of the developed classifier is provided for general metric spaces. A finer upper bound and a matching lower bound, and thus optimality of the proposed classifier, are established for Riemannian manifolds. We investigate the numerical performance of the proposed estimator and classifier via simulation studies, and illustrate their practical merits via an application to task-related fMRI data.

Keywords

Excess risk, manifold, metric entropy, metric space, minimax.

Policy learning "without" overlap: Pessimism and generalized empirical Bernstein's inequality

Ying Jin[1], **Zhimei Ren**[2], **Zhuoran Yang**[3], **Zhaoran Wang**[4]

[1] *Stanford University, Statistics Department, USA, ying531@stanford.edu*
[2] *University of Pennsylvania, Department of Statistics and Data Science, USA, zren@wharton.upenn.edu*
[3] *Yale University, Department of Statistics and Data Science, USA, zhuoran.yang@yale.edu*
[4] *Northwestern University, Department of Industrial Engineering & Management Sciences, USA, zhaoranwang@gmail.com*

Abstract

Offline policy learning aims at utilizing observations collected a priori (from either fixed or adaptively evolving behavior policies) to learn the optimal individualized decision rule in a given class. Existing policy learning methods rely on a uniform overlap assumption, i.e., the propensities of exploring *all* actions for *all* individual characteristics are lower bounded in the offline dataset. As one has no control over the data collection process, this assumption can be unrealistic in many situations, especially when the behavior policies are allowed to evolve over time with diminishing propensities. In this work, we propose a new algorithm that optimizes lower confidence bounds (LCBs) — instead of point estimates — of the policy values. The LCBs are constructed by quantifying the estimation uncertainty of the augmented-inverse-propensity-weighted (AIPW)-type estimators using knowledge of the behavior policies for collecting the offline data. Without assuming any uniform overlap condition, we establish a data-dependent upper bound for the suboptimality of our algorithm, which depends only on (i) the overlap for the *optimal* policy, and (ii) the complexity

of the policy class. As an implication, for adaptively collected data, we ensure efficient policy learning as long as the propensities for optimal actions are lower bounded over time, while those for suboptimal ones are allowed to diminish arbitrarily fast. In our theoretical analysis, we develop a new self-normalized concentration inequality for IPW estimators, generalizing the well-known empirical Bernstein's inequality to unbounded and non-i.i.d. data.

Keywords

Contextual bandit, Offline policy learning, Pessimism, Self-normalization.

Statistical Inference for Maximin Effects: Identifying Stable Associations across Multiple Studies

Zijian Guo[1]

[1] *Rutgers, Department of Statistics, USA, zijguo@stat.rutgers.edu*

Abstract

Integrative analysis of data from multiple sources is critical to making generalizable discoveries. Associations consistently observed across multiple source populations are more likely to be generalized to target populations with possible distributional shifts. In this paper, we model the heterogeneous multi-source data with multiple high-dimensional regressions and make inferences for the maximin effect (Meinshausen, Bühlmann, AoS, 43(4), 1801–1830). The maximin effect provides a measure of stable associations across multi-source data. A significant maximin effect indicates that a variable has commonly shared effects across multiple source populations, and these shared effects may be generalized to a broader set of target populations. There are challenges associated with inferring maximin effects because its point estimator can have a non-standard limiting distribution. We devise a novel sampling method to construct valid confidence intervals for maximin effects. The proposed confidence interval attains a parametric length. This sampling procedure and the related theoretical analysis are of independent interest for solving other non-standard inference problems. Using genetic data on yeast growth in multiple environments, we demonstrate that the genetic variants with significant maximin effects have generalizable effects under new environments. The proposed method is implemented in the R package `MaximinInfer` available from CRAN.

Keywords

Heterogeneous multi-source data; Distributionally robust optimization; Non-standard inference; High-dimensional Inference; Distributional shifts.

Flexible spatial dependence modeling using a shrinkage process prior

Veronica Berrocal[1], **Hwanggwan Gwon**[1], **Romain Drai**[1], **Francesco Denti**[2], **Angela Rigden**[3]

[1] *University of California, Irvine, Department of Statistics, USA*
[2] *Universitá Cattolica del Sacro Cuore, Milano, Department of Statistics, Italy*
[3] *University of California, Irvine, Department of Earth System Sciences, USA*

Abstract

Any spatial statistical analysis often starts with decisions regarding how to model the spatial dependence structure. It is very common, at the initial stage, to have to determine whether the spatial process can be thought as stationary or non-stationary. In the latter case, a modeling choice could be to envision the process as globally non-stationary, but locally stationary. A drawback of this choice lies in the fact that identifying regions of local stationarity remains still challenging, at least from a computationally point of view. To address this issue, and to propose a unified modeling framework that can accommodate both stationary and global non-stationary, but locally stationary processes, we introduce the CUSP-MRA prior. The prior joins two ideas, the CUmulative Shrinkage Prior of Legramanti et al. (2020) and the Multi-Resolution Approximation (MRA) of Katzfuss (2017). It leverages the MRA to represent the spatial process of interest through a basis function expansion that involves, a priori, an infinite number of multi-resolution basis functions. The basis function weights are in turn provided with a CUSP prior that provides increasing shrinkage as the spatial resolution of the basis function increases. Inference on the basis function weights, the number of levels of resolutions needed, and the spatial variability in the number of levels needed provide information on whether the process can be considered stationary or not.

It also allows to determine regions of local stationarity. We showcase the ability of our model to correctly capture regions of local stationarity through simulation experiments. Finally, we apply our model to generate daily, point-level estimates of soil moisture, an important parameter on the wetness of the soil that plays a critical role in the global water and energy cycle.

Keywords:

Spatial process, multi-resolution, basis function expansion, shrinkage, non-stationarity.

Chapter 3

Oral Contributed Talks

Data Mining in Higher Education Institutions and Future Directions

Abdel-Salam G. Abdel-Salam
Qatar University, College of Arts and Sciences, Department of Mathematics and Statistics, Doha, Qatar.
e-mail: abdo@qu.edu.qa

Abstract

Data mining techniques have been increasingly applied in higher education institutions to uncover valuable insights from student data and improve teaching and learning outcomes. This study discusses the applications and impacts of data mining in higher education. Data mining techniques have been employed to identify at-risk students, optimize teaching strategies, evaluate courses, and enhance curriculum development. Moreover, predictive modeling forecasts student performance and identifies those needing additional academic support. Despite the potential benefits, data mining in higher education faces several challenges, including data quality issues, ethical considerations, lack of domain knowledge, potential bias and stereotypes, and implementation challenges. This talk highlights the importance of addressing these challenges to ensure the effectiveness and ethical use of data mining in higher education while emphasizing the potential of these techniques to improve academic performance and contribute to more efficient and data-driven decision-making processes.

Keywords

Educational Data mining, Higher education, Operational efficiency, Data mining techniques.

Almost infinite sites model

Alejandra Avalos-Pacheco[1,2], **Mathias C. Cronjäger**[3], **Jotun Hein**[3],**Paul A. Jenkins**[4,5,6]

[1] *TU Wien, Research Unit of Applied Statistics, Vienna, Austria*
[2] *Harvard Medical School, Harvard-MIT Center for Regulatory Science, Boston, MA*
[3] *University of Oxford, Department of Statistics, Oxford, UK*
[4] *University of Warwick, Department of Statistics, Coventry, UK*
[5] *University of Warwick, Department of Computer Science, Coventry, UK*
[6] *The Alan Turing Institute, London, UK*

Abstract

A main challenge in molecular evolution is to provide computationally efficient mutation models with flexible assumptions that properly reflect genetic variation. The infinite sites model assumes that each mutation event occurs at a site never previously mutant, i.e. it does not allow recurrent mutations. This is reasonable for low mutation rates and makes statistical inference much more tractable. However, recurrent mutations are common enough to be observable from genetic variation data, even in species with low per-site mutation rates such as humans. The finite sites model on the other hand allows for recurrent mutations but is computationally unfeasible to work with in most cases. In this work, we bridge these two approaches by developing a novel molecular evolution model, the almost infinite sites model, that both admits recurrent mutations and is tractable. We provide a recursive characterisation of the likelihood of our proposed model and outline a parsimonious approximation scheme for computing it. We show the usefulness of our model in simulated and human mitochondrial data.

Keywords

Coalescent, molecular evolution, infinite sites, finite sites, population genetics.

Bayesian Causal Discovery from Unknown General Interventions

Alessandro Mascaro[1], **Federico Castelletti**[2]

[1] *University of Milano-Bicocca, Department of Economics, Management and Statistics, Italy, a.mascaro3@campus.unimib.it*
[2] *Università Cattolica del Sacro Cuore, Department of Statistical Sciences, Italy, federico.castelletti@unicatt.it*

Abstract

Directed Acyclic Graphs (DAGs) are often used to represent causal relationships between variables. In this setting, the process of identifying the DAG structure from data is referred to as causal discovery. If only observational data are available, the DAG is identifiable only up to its Markov equivalence class. However, if in addition one uses experimental data, i.e. data in which the generating process has been altered by an external intervention, then it is possible to identify smaller subclasses of DAGs, known as I-Markov Equivalence Classes (I-MECs). Different types of interventions modify the causal structures in different ways and, accordingly, imply distinct definitions of I-MECs. Current causal discovery algorithms from experimental data assume that interventions do not modify the parents of the intervened nodes in the DAG, even when the targets of interventions are unknown. We relax this assumption by proposing a Bayesian methodology for causal discovery from experimental data arising from unknown general interventions. Our contribution includes (i) providing definitions and graphical characterizations of general I-MECs; (ii) developing priors which guarantee score equivalence of DAGs within the same I-MECs and (iii) devising suitable MCMC schemes to sample from the posterior distribution over DAGs and unknown interventions.

Keywords

Causal discovery, Bayesian model selection, Graphical models.

Clustering approaches for mixed-type data: A comparative study

Badih Ghattas[1], **Alvaro Sanchez San Benito**[2]

[1] *Aix-Marseille University, Aix-Marseille School of Economics, France, badih.ghattas@univ-amu.fr*
[2] *Aix-Marseille University, Airbus Helicopters, France, alvaro.sanchez-san-benito@airbus.com*

Abstract

Clustering is widely used for data analysis in an unsupervised framework to find homogeneous groups of data points within a dataset. However, clustering mixed-type data remains a challenge, as only a limited number of existing methods are suited for this task. This study presents the state-of-the-art of these approaches and compares them across three main simulation models: Gaussian, exponential-multinomial and Bayesian network. The methods include KAMILA, k-prototypes, PDC, convex k-means and TAN Bayesian network classifier. The primary objective is to provide insight into the behaviour of different methods across a wide range of scenarios. The results suggest that all the analyzed clustering methods perform better when dealing with data following Gaussian distributions compared to data following an exponential-multinomial distribution. In cases where datasets have a dependence data structure, only a limited number of methods are capable of achieving satisfactory results. Furthermore, various experimental factors, including the number of clusters, the cluster overlap, the proportion of continuous variables in the dataset, and the balance of cluster sizes, have a significant influence on their performance.

Keywords

Clustering, Mixed-type data, KAMILA, Bayesian networks.

References

Foss, A.H., Markatou, M. and Ray, B. and Heching, A. (2016). A semi-parametric method for clustering mixed data. *Machine Learning 105*, 419–458.

Huang, Z. (1998). Extension to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery2* 283–304.

Iyigun, I. (2007). Probabilistic Distance Clustering. *PhD Thesis*, New Brunswick Rutgers, The State University of New Jersey.

Modha, D.S. and Spangler, W.S. (2003). Feature Weighting in k-Means Clustering. *Machine Learning* v.52, n.3, 217–237.

Truong Pham, D. and A.Ruz, G. (2009). Unsupervised training of Bayesian networks for data clustering. *Proc. R. Soc. 465*, 2927–2948.

Temperature-Mortality Association: Portuguese Extreme Weather Event Early Warning System

André Brito[1,2,4], **Baltazar Nunes**[2,3], **Susana Silva**[2], **Regina Bispo**[1,4]

[1] *NOVA School of Science and Technology, Universidade NOVA de Lisboa, Portugal,* anm.brito@campus.fct.unl.pt
[2] *Department of Epidemiology, Instituto Nacional de Saúde Dr Ricardo Jorge, Lisboa, Portugal*
[3] *Centro de Investigação em Saúde Pública, Escola Nacional de Saúde Pública, Universidade NOVA de Lisboa, Lisboa, Portugal*
[4] *Center for Mathematics and Applications (NOVA Math)*

Abstract

Portugal is among the European countries with higher excess of mortality during winter, even though winters are considered to be relatively mild. There is also an excess of mortality during the summer. This excess mortality might be associated with a larger vulnerability of the Portuguese population to non-optimal temperature exposure caused by poor housing conditions and an ageing population. In this study, supported by FCT (RELIABLE: DSAIPA/DS/0111/2019), an update of the heat and cold health early warning systems is proposed for use in Mainland Portugal. The aim was to develop a risk indicator, active throughout the whole year, and easily understood by the entire population, with the highest possible spatial resolution. Daily data of all-causes mortality, maximum and minimum temperatures were gathered from public data sources for the 1995-2020 time period. District-specific temperature-mortality associations were estimated using quasi-Poisson regression. Linear threshold Distributed Lag Models (DLM) were proposed and estimated for cold and warm semesters, where minimum temperatures were considered in autumn/winter and

maximum temperatures in spring/summer, to identify worst case exposure scenarios. Influenza incidence was also included in the models to improve predictive performance. Model selection was based on goodness-of-fit criteria.

Models proposed here could serve as updates for heat and cold health early warning systems, as they provide the results to maintain a risk indicator with the aforementioned properties. Differences between the optimum district-specific models were found and completely justify the need for region-specific warnings. Optimum cold thresholds were found to be relatively mild temperatures when compared to optimum heat thresholds, suggesting the effects of cold temperatures on mortality start at fairly milder temperatures. Evidence was found supporting the hypothesis of acclimatisation of the population to their own specific climates.

Keywords

Time Series Regression, GLM , DLM, Temperature, Mortality.

References

Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L. & Armstrong, B. Time series regression studies in environmental epidemiology. *International Journal Of Epidemiology.* **42**, 1187-1195 (2013)

Guo, Y., Gasparrini, A., Armstrong, B., Tawatsupa, B., Tobias, A., Lavigne, E., De Sousa Zanotti Stagliorio Coelho, M., Pan, X., Kim, H., Hashizume, M., Honda, Y., Leon Guo, Y., Wu, C., Zanobetti, A., Schwartz, J., Bell, M., Scortichini, M., Michelozzi, P., Punnasiri, K., Li, S., Tian, L., Garcia, S., Seposo, X., Overcenco, A., Zeka, A., Goodman, P., Dang, T., Van Dung, D., Mayvaneh, F., Saldiva, P., Williams, G. & Tong, S. Heat wave and mortality: A multicountry, multicommunity study. *Environmental Health Perspectives.* **125**, 1-11 (2017)

Healy, J. Excess winter mortality in Europe: A cross country analysis identifying key risk factors. *Journal Of Epidemiology And Community Health.* **57**, 784-789 (2003)

Gasparrini, A., Masselot, P., Scortichini, M., Schneider, R., Mistry, M., Sera, F., Macintyre, H., Phalkey, R. & Vicedo-Cabrera, A. Small-area assessment of temperature-related mortality risks in England and Wales: a case time series analysis. *The Lancet Planetary Health.* **6**, e557-e564 (2022), http://dx.doi.org/10.1016/S2542-5196(22)00138-3

Gasparrini, A. Modeling exposure-lag-response associations with distributed lag non-linear models. *Statistics In Medicine.* **33**, 881-899 (2014)

Interacting innovation processes

Giacomo Aletti[1], **Irene Crimaldi**[2], **Andrea Ghiglietti**[3]

[1] *Università degli Studi di Milano, Milan, Italy,*
giacomo.aletti@unimi.it
[2] *IMT School for Advanced Studies Lucca, Lucca, Italy,*
irene.crimaldi@imtlucca.it
[3] *Università degli Studi di Milano-Bicocca, Milan, Italy,*
andrea.ghiglietti@unimib.it

Abstract

We propose a general model that can incorporate various sources of interaction among multiple innovation processes. These processes typically represent the mechanisms through which novelties emerge and trigger further novelties, such as first-time occurrences of specific events. The stochastic dynamics of innovation processes are commonly studied in the literature using urn schemes. In these schemes, the colors extracted from the urn at each time indicate whether a novelty was discovered or if an old item was selected. Therefore, the urn proportions represented the probability that further novelties could arise or the probability of selecting a specific old item. Our model is based on a system of ternary urns with triggering known as Poisson-Dirichlet processes, which are an extension of the classical Dirichlet processes with an additional parameter that allows a reinforcement in the probability of observing a novelty, i.e. the occurrence of a novelty increase the probability of observing further novelties in the future. Each innovation process will be thought as located on the nodes of a finite directed graph, with different adjacency matrices to model multiple types of interactions. In particular, these interactions consist in the fact that, for each node, the probability of observing a new or an old item depends, not only on the items observed for that node, but also on the items observed for the other nodes. More precisely, our model is able to implement the following two types of interactions:

(i) the probability of *exploitation* of an old item c by node h, has an increasing dependence not only on the number of times c has been observed in node h itself, but also on the number of times c has been observed in each of the other nodes;

(ii) the probability of *production* (or *exploration*) of a novelty for the entire system by node h (item never seen before from any node of the graph) has an increasing dependence not only on the number of novelties produced by h itself in the past, but also on the number of novelties produced by each of the other nodes in the past.

Under the assumption of strongly connected adjacency matrices, we are able to prove analytically several first-order asymptotic results on the main quantities that rule the innovation dynamics. In particular, we show that the number of novelties arose in any node of the system grows with the same Heaps' exponent $\gamma^* \in (0, 1)$. Moreover, the ratio between the number of novelties arose in two different nodes converges almost surely to a constant that can be easily determined by the leading Perron eigenvector of one of the adjacency matrices. In addition, we also show that, asymptotically, the number of times each item has been adopted in the whole system are uniformly distributed among the nodes. Finally, the model has been applied on two real data sets: one taken from the social content aggregation website *Reddit*, and one from the on-line library *Project Gutenberg*. We show that both data sets exhibit empirical behaviors that are in accordance with those predicted by the proven theoretical results.

Keywords

Innovation processes, urn models, interacting reinforcement.

References

G. Aletti, I. Crimaldi, and A. Ghiglietti (2023). Interacting innovation processes. *Sci. Rep. 13, 17187.*

Spectral CLTs with long memory and aging for large language and large multimodal models

Andrej Srakar[1]

[1] *Institute for Economic Research and University of Ljubljana, Slovenia, andrej.srakar@ier.si*

Abstract

Since the pioneering works from the 1980s by Breuer, Dobrushin, Major, Rosenblatt, Taqqu and others, central and noncentral limit theorems for Y_t have been constantly refined, extended and applied to an increasing number of diverse situations. In recent years, fourth moment theorem CLTs, quantitative CLTs, Breuer-Major and Dobrushin-Major CLTs, de Jong CLTs, functional CLTs and others have been developed. Recently, Maini and Nourdin (2023) extended this to spectral central limit theorems valid for additive functionals of isotropic and stationary Gaussian fields. Their work uses Malliavin-Stein method and Fourier analysis techniques to situations where Y_t admits Gaussian fluctuations in a long memory context. In another recent article, Wang et al. (2023) augmented existing language models with long-term memory. Namely, existing large language models (LLMs) can only afford fix-sized inputs due to the input length limit, preventing them from utilizing rich long-context information from past inputs. They proposed a framework of Language Models Augmented with Long-Term Memory, which enables LLMs to memorize long history. In our article we develop spectral central limit theorems in a context of augmented large language models of Wang and coauthors, as well as to present extensions of LLM labeled large multimodal models (for an overview see Yang et al., 2023). We develop spectral central limit theorems based on extensions of Gaussian fields with aging in a Bouchard trap model context (Croydon and Muirhead, 2014). Our main stochastic calculus tools derive from Malliavin-Stein method, Fourier analysis and free probability. Applications and extensions of our work are

possible in multiple areas in probability theory, statistics, data science and econometrics, such as stochastic geometry, spherical random fields, deep neural networks and graph neural networks, causal AI and functional data analysis. We present applications on datasets from finance and medical imaging. In conclusion we discuss possible Bayesian extensions.

Keywords

Spectral central limit theorem, long memory, large language model, large multimodal model, Bouchard trap model, free probability.

References

Croydon, D.; Muirhead, S. (2015). Functional limit theorems for the Bouchaud trap model with slowly varying traps. *Stochastic Processes and their Applications*, v.125, n.5, 1980–2009.

Maini, L.; Nourdin, I. (2023). Spectral central limit theorem for additive functionals of isotropic and stationary Gaussian fields. *arXiv:2206.14458*.

Wang, W., Dong, L., Cheng, H., Liu, X., Yan, X., Gao, J., and Wei, F. (2023). Augmenting Language Models with Long-Term Memory. *arXiv:2306.07174*.

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., and Wang, L. (2023). The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv:2309.17421v2*.

Estimation of timing of past events in cancer based on DNA sequencing data

Andrew Koval[1], **Khanh Dinh**[2], **Emmanuel Asante**[1], **Simon Tavaré**[2], **Marek Kimmel**[1]

[1] *Rice University, Department of Statistics, USA, alk3@rice.edu*
[2] *Columbia University, Department of Statistics, USA, knd2127@columbia.edu*

Abstract

A malignant tumor is often composed of clusters (i.e., clones) of cells that compete for scarce resources in the tumor microenvironment. Estimating the growth rates, mutation rates, and birth times (i.e., the evolutionary parameters) of these clones can be used to detect whether newer, more malignant tumor clusters arrive early or late in the lifetimes of tumors for a given cancer type. However, estimating these parameters from typically noisy single-cell sequencing data of single-nucleotide variants (SNVs) remains challenging, even as the high-resolution data provides opportunities for more precise estimation. We use theory from population genetics, a set of neutral mutation equations, and the site frequency spectrum (SFS) of single-cell DNA sequencing data to estimate the evolutionary parameters of each clone. Our model generates numerically unbiased estimates of the evolutionary parameters of each clone. Furthermore, we demonstrate that our model is robust to varying the true underlying evolutionary parameters, to increases in sequencing errors, and to moderate misclassification of cells to clones. Future work is needed to better quantify and reduce the uncertainty of our estimates. However, our model can be used to provide guidance when making decisions about screening times for a given cancer type by potentially detecting patterns of the birth times of faster growing clones.

Keywords

Cancer evolution, site frequency spectrum, tumor heterogeneity, clonal selection, single-cell sequencing.

Incorporating Novel Input Variable Selection for Improved Precipitation Forecasting in the Different Water Basins of Thailand

Angkool Wangwongchai[1], **Muhammad Waqas**[2,6], **Usa Wannasingha Humphries**[3], **Porntip Dechpichai**[4], **Phyo Thandar Hlaing**[5,6]

[1] *Department of Mathematics, King Mongkut's University of Technology Thonburi , THAILAND, angkool.wan@kmutt.ac.th*
[2] *The Joint Graduate School of Energy and Environment (JGSEE), King Mongkut's University of Technology Thonburi , THAILAND, muhammad.waqa@kmutt.ac.th*
[3] *Department of Mathematics, King Mongkut's University of Technology Thonburi , THAILAND, usa.wan@kmutt.ac.th*
[4] *Department of Mathematics, King Mongkut's University of Technology Thonburi , THAILAND, porntip.dec@kmutt.ac.th*
[5] *The Joint Graduate School of Energy and Environment (JGSEE), King Mongkut's University of Technology Thonburi , THAILAND, phyothandar.hlai@kmutt.ac.the-mail*
[6] *Center of Excellence on Energy Technology and Environment (CEE), Ministry of Higher Education, Science, Research and Innovation, Bangkok, Thailand*

Abstract

Precipitation forecasting is essential in water resource planning and management, particularly in tropical countries like Thailand. Selecting appropriate input variables for developing prediction models for rainfall is a significant difficulty. Recent studies in various disciplines have highlighted the utility of artificial intelligence-based techniques for determining explanatory variables for use in non-linear scenarios, which remain largely unexplored in rainfall forecasting. The present study was carried out to fill this knowledge gap. Two river basins in the northern region of Thailand were selected as a study area. Monthly

observation and large-scale climatic variables (LCVs) at both River basins from 1993 to 2022 were used for model development. This study proposed a novel hybrid bootstrapped long short-term recurrent neural network (BTSP-LSTM-RNN) for input variables selection (IVS) for monthly precipitation forecasting. A novel BTSP-LSTM-RNN model was compared with the support vector regression with recursive feature elimination (SVR-RFE) and gradient boosting (GB). For the evaluation of these models, statistical metrics such as coefficient of determination (R2), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) are used. Remarkably, the proposed BTSP-LSTM-RNN hybrid model outperformed other models, achieving a notably higher R2 value of 0.85 compared to SVR-RFE (0.68) and GB (0.72). BTSP-LSTM-RNN exhibited a remarkably low MAE of 0.0094, underscoring its accuracy in IVS, whereas SVR-RFE and GB achieved 120.92 and 123.7, respectively. The RMSE value of 116.62, although higher than the MAE, still indicates a relatively low error compared to SVR-RFE and GB. Most notably, the BTSP-LSTM-RNN model demonstrated the lowest MAPE among the three models, standing at 27.88

Keywords

Artificial Intelligence, Bootstrapping, Forecasting, Input Variable Selection, Recurrent Neural Networks, Long Short Term Memory.

Concentration of Aggregated Adjacency and Laplacian Matrices for Lazy Network-Valued Stochastic Processes with Applications

Sayak Chatterjee[1], Shirshendu Chatterjee[2], Soumendu Sundar Mukherjee[3], Anirban Nath[4], Shamodeep Bhattacharya[5]

[1] *University of Pennsylvania, Department of Statistics and Data Science, USA, sayakc@wharton.upenn.edu*
[2] *City College and Graduate Center of the City University of New York, Department of Mathematics, USA, shirshendu@ccny.cuny.edu*
[3] *Indian Statistical Institute Kolkata, Statistics and Mathematics Unit, India, ssmukherjee@isical.ac.in*
[4] *Columbia University in the City of New York, Department of Statistics, USA, an3145@columbia.edu*
[5] *Oregon State University, Department of Statistics, USA, bhattash@science.oregonstate.edu*

Abstract

Network-valued time series are currently a common form of network data. However, the study of the aggregate behavior of network sequences generated from network-valued stochastic processes is relatively rare. Most of the existing research focuses on the simple setup where the networks are independent (or conditionally independent) across time, and all edges are updated synchronously at each time step. In this paper, we study the concentration properties of the aggregated adjacency matrix and the corresponding Laplacian matrix associated with network sequences generated from lazy network-valued stochastic processes, where edges update asynchronously, and each edge follows a lazy stochastic process for its updates independent of the other edges. We demonstrate the usefulness of these concentration results in proving consistency of standard estimators in community estimation and changepoint estimation problems. We also conduct a simulation study

to demonstrate the effect of the laziness parameter, which controls the extent of temporal correlation, on the accuracy of community and changepoint estimation.

Keywords

Dependent NetworksC, Concentration Inequalities, Spectral Clustering, Community Estimation, Changepoint Estimation.

Regularized AMMI Model for Multi-Environment Agricultural Trials

Aniruddha Pathak[1], **Somak Dutta**[2]

[1] *Iowa State University, Department of Statistics, USA, anipath@iastate.edu*
[2] *Iowa State University, Department of Statistics, USA, somakd@iastate.edu*

Abstract

The additive main effects and multiplicative interaction (AMMI) model is widely used for studying yield stability in multi-environment field trials. However, the ordinary AMMI model does not allow predictions for untested genotypes. In this work, A novel hierarchical AMMI mixed-effects model for likelihood-based inference is developed that regularizes the genomic main effects and factor scores using kinship information among the genotypes and accommodates the missing data. The kinship information also allows genomic prediction of untested genotypes. A scalable stochastic expectation-maximization algorithm is developed for large multi-environment trial datasets and is further accelerated by the squared extrapolation method. Simulation studies and maize data from the Genomes to Fields Initiative illustrate improvements in yield prediction and the detection of nonlinear genotype-by-environment interactions.

Keywords

Factor analysis, Finlay-Wilkinson model, Genomic prediction, Prediction interval.

Bayesian Variable Selection and Sparse Estimation for High-Dimensional Graphical Models

Anwesha Chakravarti[1], **Naveen Narisetty**[2], **Feng Liang**[3]

[1] *University of Illinois Urbana Champaign, Department of Statistics, United States, anwesha5@illinois.edu*
[2] *University of Illinois Urbana Champaign, Department of Statistics, United States, naveen@illinois.edu*
[3] *University of Illinois Urbana Champaign, Department of Statistics, United States, liangf@illinois.edu*

Abstract

We introduce a novel Bayesian approach that can perform both covariate selection and sparse precision matrix estimation in the context of high-dimensional Gaussian graphical models involving multiple responses and covariates. Our method involves introducing covariate-level sparsity in the precision matrix between the multiple responses and the covariates which induces column-wise group sparsity into our regression coefficient matrix estimates. This is achieved by using a Bayesian conditional random field model with an appropriately chosen hierarchical spike and slab prior setup. A distinctive feature of our method is that it can provide a sparse estimation of the three distinct sparsity structures: the conditional dependency structure among the responses, the conditional dependency structure between the responses and the covariates, and the regression coefficient matrix, unlike existing methods which typically concentrate on any two of these structures, but seldom achieve simultaneous sparse estimation for all three. Despite the non-convex nature of the problem, we establish statistical accuracy for all points in the high posterior density region, including the maximum-a-posteriori (MAP) estimator. We also present an efficient Expectation-Maximization (EM) algorithm to compute the estimators. Through simulation experiments, we demonstrate the competitive performance of our method, particularly in scenarios where

the signal strength in the precision matrices is low. Finally, we apply our method to a bikeshare data set, showcasing its predictive performance.

Keywords

Bayesian variable selection, Gaussian conditional random field, Bayesian regularization, Spike and slab Lasso prior, Graphical models.

Polyhedral sets in the Wasserstein space and algorithms for mean-field variational inference

Yiheng Jiang[1], **Sinho Chewi**[2], **Aram-Alexandre Pooladian**[3]

[1] *New York University, Courant Institute of Mathematical Sciences, yj2070@nyu.edu*
[2] *School of Mathematics, Institute for Advanced Study, schewi@ias.edu*
[3] *New York University, Center for Data Science, aram-alexandre.pooladian@nyu.edu*

Abstract

We develop a theory of finite-dimensional polyhedral subsets over the Wasserstein space and optimization of functionals over them via first-order methods. Our main application is to the problem of mean-field variational inference, which seeks to approximate a distribution π over $\mathbb{R}^d$ by a product measure $\hat{\pi}$. When π is strongly log-concave and log-smooth, we provide (1) approximation rates certifying that $\hat{\pi}$ is close to a polyhedral set $\mathcal{P}_\diamond$, and (2) an algorithm for minimizing $\mathrm{KL}(\cdot\|\pi)$ over $\mathcal{P}_\diamond$ with accelerated complexity $O(\sqrt{\kappa}\log(\kappa d/\varepsilon^2))$, where κ is the condition number of π.

Keywords

Variational inference, optimal transport, accelerated gradient descent.

Model selection over partially ordered sets

Armeen Taeb[1], **Peter Buhlmann**[2], **Venkat Chandrasekaran**[3]

[1] *University of Washington, Department of Statistics, USA,*
ataeb@uw.edu
[2] *ETH Zurich, Seminar for Statistics, Switzerland,*
peter.buehlmann@stat.math.ethz.ch
[3] *California Institute of Technology, Departments of Computing and Mathematical Sciences and of Electrical Engineering, USA,*
venkatc@caltech.edu

Abstract

In problems such as variable selection and graph estimation, models are characterized by Boolean logical structure such as presence or absence of a variable or an edge. Consequently, false positive and true negative errors can be specified as the number of variables or edges that are incorrectly included/excluded in an estimated model. However, there are several other problems such as ranking, clustering, and causal inference in which the associated model classes do not admit transparent notions of false positive and true negative errors due to the lack of an underlying Boolean logical structure. In this paper, we present a generic approach to endow a collection of models with partial order structure, which leads to a hierarchical organization of model classes as well as natural analogs of false positive and true negative errors. We describe model selection procedures that provide false positive error control in our general setting and we illustrate their utility with numerical experiments.

Keywords

Combinatorics, greedy algorithms, multiple testing, stability.

Hybrid of node and link communities for graphon estimation

Arthur Verdeyme[1], **Sofia C. Olhede**[2]

[1] *Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne, Switzerland, arthur.verdeyme@epfl.ch*
[2] *Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne, Switzerland, sofia.olhede@epfl.ch*

Abstract

Networks serve as a ubiquitous tool for examining the large-scale connectivity framework within complex systems. Modelling their generative mechanism nonparametrically is often based on step-functions, such as stochastic block models. These models are capable of addressing two prominent topics in network science: link prediction and community detection. However, such methods often have a resolution limit, making it difficult to separate actual structure from noise. As such, we propose a different estimation method using a multiscale model, which we call the *stochastic shape model*. Typically, methods model node or link communities. In contrast, we take a hybrid approach, bridging the two notions of community. Consequently, we obtain a more parsimonious representation, enabling a more interpretable summary of the network structure. By considering multiple resolutions, we trade bias and variance to ensure that our estimator is rate-optimal. We also examine the performance of our model through simulations and applications to real network data.

Keywords

Networks, graphons, community detection.

Optimal allocation of sample size for randomization-based inference from 2^K factorial designs

Arun Ravichandran[1], **Nicole E. Pashley**[2], **Brian Libgober**[3], **Tirthankar Dasgupta**[4]

[1] *Rutgers University, Department of Statistics, United States, arun.ravi@rutgers.edu*
[2] *Rutgers University, Department of Statistics, United States, np755@stat.rutgers.edu*
[3] *Northwestern University, Department of Political Science and Law, United States, brian.libgober@northwestern.edu*
[4] *Rutgers University, Department of Statistics, United States, td370@stat.rutgers.edu*

Abstract

Optimizing the allocation of units into treatment groups can help researchers improve the precision of causal estimators and decrease costs when running factorial experiments. However, existing optimal allocation results typically assume a super-population model and that the outcome data comes from a known family of distributions. Instead, we focus on randomization-based causal inference for the finite-population setting, which does not require model specifications for the data or sampling assumptions. We propose exact theoretical solutions for optimal allocation in 2^K factorial experiments under complete randomization with A-, D- and E-optimality criteria. We then extend this work to factorial designs with block randomization. We also derive results for optimal allocations when using cost-based constraints. To connect our theory to practice, we provide convenient integer-constrained programming solutions using a greedy optimization approach to find integer optimal allocation solutions for both complete and block randomization. The proposed methods are demonstrated

using two real-life factorial experiments conducted by social scientists.

Keywords

Randomization-based inference, Causal Inference, Optimal designs, 2^K Factorial designs, A-optimality, D-optimality, E-optimality, Neyman allocation, Cost-based allocation.

Nonparametric Causal Additive Models with Smooth Backfitting

Asger B. Morville[1], **Byeong U. Park**[2]

Seoul National University, Department of Statistics, South Korea,
[1]*asgermorville@snu.ac.kr,* [2]*bupark@snu.ac.kr*

Abstract

We propose a fully nonparametric approach to learning causal structures in observational data. The method is outlined in the setting of additive structural equation models, and we describe the link to causal inference. The estimation procedure of the additive structural equation functions is based on a smooth backfitting approach. The flexibility of the nonparametric procedure results in strong theoretical properties in the estimation of the variable ordering. We show that under mild conditions the ordering estimate is consistent. Through simulations we demonstrate that our method is competitive with the state of the art approaches to causal learning. In particular, the smooth backfitting approach shows robustness when the noise is heteroscedastic.

Keywords

Smooth backfitting, causal models, structural equation models.

Parameter and State Estimation in Queues

Azam Asanjarani[1], **Yoni Nazarathy**[2], **Peter Taylor**[3]

[1] *The University of Auckland, Department of Statistics, New Zealand, azam.asanjarani@auckland.ac.nz*
[2] *The University of Queensland, School of Mathematics and Physics, Australia, y.nazarathy@uq.edu.au*
[3] *The University of Melbourne, School of Mathematics and Statistics, Australia, taylorpg@unimelb.edu.au*

Abstract

We present a broad literature survey of parameter and state estimation for queueing systems. Our approach is based on various inference activities, queueing models, observations schemes, and statistical methods. We categorize these into branches of research that we call estimation paradigms. These include: the classical sampling approach, inverse problems, inference for non-interacting systems, inference with discrete sampling, inference with queueing fundamentals, queue inference engine problems, Bayesian approaches, online prediction, implicit models, and control, design, and uncertainty quantification. For each of these estimation paradigms, we outline the principles and ideas, while surveying key references. We also present various simple numerical experiments. In addition to some key references mentioned here, a periodically updated comprehensive list of references dealing with parameter and state estimation of queues will be kept in an accompanying annotated bibliography.

Keywords

Parameter estimation, State estimation, Queueing systems, Statistical methods, Observations schemes.

Approximating Markov Chains via Weak Perturbation Theory

Badredine Issaadi

[1] *Laboratoire LITAN École supérieure en Sciences et Technologie de l'Informatique et du Numérique RN 75, Amizour 06300, Bejaia, Algérie. issaadi@estin.dz*

Abstract

The calculation of the stationary distribution for a stochastic infinite matrix is generally difficult and does not have closed form solutions, it is desirable to have simple approximations converging rapidly to this distribution. Let P be the transition matrix of discrete Markov chain on $\mathcal{S}$, with invariant distribution π. Let $\mathcal{S}_k$ be a sequence of $\mathcal{S}$, we are interested in procedures for approximating π using P_k, where P_k is derived from the linear augmentation of the $\mathcal{S}_k \times \mathcal{S}_k$ "northwest truncation" of P. Let $V : \mathcal{S} \to [1, +\infty)$. In this paper, an explicit bounds are obtained for the distance between $\widehat{\pi}_k$ and π. Such computable bounds are derived in terms of standard drift conditions.

Keywords

Truncation, Queuing System, Weak Stability, Markov chain, Algorithm.

Statistical Considerations in Identifying Biomarkers for Diagnosing Myofascial Pain Syndrome

Yu-lin Hsu[1], **Ben Seiyon Lee**[2], **and William F. Rosenberger**[3]

[1] *George Mason University, Department of Statistics, USA, yhsu7@gmu.edu*
[2] *George Mason University, Department of Statistics, USA, slee287@gmu.edu*
[3] *George Mason University, Department of Statistics, USA, wrosenbe@gmu.edu*

Abstract

Myofascial pain syndrome (MPS) is a condition characterized by pain associated with inflammation or irritation of the musculoskeletal soft tissue. Myofascial pain is widespread, affecting approximately 85% of the population at some point in their lives. The lack of quantitative biomarkers, capable of pinpointing the root causes of myofascial dysfunction, has posed a significant challenge to the development of effective diagnostic and therapeutic strategies for MPS. Clinical exams have been the primary diagnostic tool for diagnosing MPS; however, these examinations are both costly and have been inconsistent with poor inter-examiner reproducibility. We propose a novel classification method for MPS diagnosis based on a composite quantitative biomarker. The biomarker is constructed using quantitative measures acquired through state-of-the-art modalities including shear anisotropy ratio, bioimpedance spectroscopy, EMG ratios, and ultrasound gliding analysis. To address the sheer volume of imaging data, we employ dimension-reduction techniques as well as statistical learning methods for classification. Finally, we demonstrate and validate our approach on data collected from an ongoing cohort study.

Keywords

Statistical image analysis, statistical learning, functional biomarker-based classification, biomedical applications, high-dimensional data.

Minimax Risks and Optimal Procedures for Estimation under Functional Local Differential Privacy

Bonwoo Lee[1], **Jeongyoun Ahn**[2], **Cheolwoo Park**[3]

[1] *Korea Advanced Institute of Science and Technology, Mathematical science, Korea, righthim@kaist.ac.kr*
[2] *Korea Advanced Institute of Science and Technology, Industrial & Systems Engineering, Korea, jyahn@kaist.ac.kr*
[3] *Korea Advanced Institute of Science and Technology, Mathematical science, Korea, parkcw2021@kaist.ac.kr*

Abstract

As concerns about data privacy continue to grow, differential privacy (DP) has emerged as a fundamental concept that aims to guarantee privacy by ensuring individuals' indistinguishability in data analysis. Local differential privacy (LDP) is a rigorous type of DP that requires individual data to be privatized before being sent to the collector, thus removing the need for a trusted third party to collect data. Among the numerous (L)DP-based approaches, functional DP has gained considerable attention in the DP community because it connects DP to statistical decision-making by formulating it as a hypothesis-testing problem and also exhibits Gaussian-related properties. However, the utility of privatized data is generally lower than that of non-private data, prompting research into optimal mechanisms that maximize the statistical utility for given privacy constraints. In this study, we investigate how functional LDP preserves the statistical utility by analyzing minimax risks of univariate mean estimation as well as nonparametric density estimation. Our theoretical study reveals that it is possible to establish an interpretable, continuous balance between the statistical utility and privacy level, which has not been achieved under the ϵ-LDP framework. Furthermore, we suggest

minimax optimal mechanisms based on Gaussian LDP (a type of functional LDP) that achieve the minimax upper bounds and show via a numerical study that they are superior to the counterparts derived under ϵ-LDP. The theoretical and empirical findings of this work suggest that Gaussian LDP should be considered a reliable standard for LDP.

Keywords

Data privacy, Functional local differential privacy, Gaussian mechanism, Minimax risks, Statistical utility.

Improving Diagnostic Models for Temporomandibular Disease Using Cost-Effective Variables: An Analysis of the Dimitroulis Classification

Carlos Brás-Geraldes[1], Ricardo São João[2], Henrique José Cardoso[3], David Faustino Ângelo[4]

[1] *ISEL - Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa & Centro de Estatística e Aplicações Universidade de Lisboa , Portugal, carlos.geraldes@isel.pt*
[2] *Escola Superior de Gestão e Tecnologia, Instituto Politécnico de Santarém & Centro de Estatística e Aplicações Universidade de Lisboa, Portugal, ricardo.sjoao@esg.ipsantarem.pt*
[3] *Instituto Português da Face, Portugal, henrique.cardoso@ipface.pt*
[4] *Instituto Português da Face & Faculdade de Medicina, Universidade de Lisboa, Portugal, david.angelo@ipface.pt*

Abstract

Background: Temporomandibular disorders (TMD) are a class of degenerative musculoskeletal and neuromuscular conditions involving the temporomandibular joint (TMJ) complex and surrounding musculature. The etiology of TMD is multifactorial, including biological, environmental, social, emotional, and cognitive triggers. Due to the complexity of the disease's signs and symptoms, the diagnosis and correct treatment of TMD remain a challenge. The Dimitroulis classification (DC) divides TMD into five categories (DC1, DC2, ..., DC5) based on the degree of disease severity with an indication for treatment. The classification is based on history and physical examination and diagnostic imaging is used to access intra-articular derangements. This process presented some subjectivity in the analysis and, has significant associated costs. The present study aims to identify variables based on patient complaints with lower associated costs and more objective, prompt, and less burdensome classification.

Methods and Results: 535 patients were included using an online database: EUROTMJ. The main complaints and final diagnosis were accessed. DC was recorded as the response variable and considered the gold standard classification; complaints were considered explanatory variables. The DC distribution (absolute frequency) for each severity category is: DC1-116; DC2-133; DC3-71; DC4-54; DC5-0. The sample was split into two parts: training with 70% of the observations and testing with the remaining 30%. Initially, the severity categories from the Dimitroulis classification were considered. However, due to the multicategory response variable and the binary nature of the explanatory variables, multinomial logistic regression was determined to be the appropriate statistical method. A variable selection process was then conducted using the bidirectional stepwise method, resulting in a multivariable model with the explanatory variables being TMJ locking, tinnitus, cervical muscle tension, and limitation of mouth opening. Despite the application of the multinomial logistic model, the achieved accuracy rate was only 42.9%, indicating poor performance. Consequently, the analysis was simplified to consider only two categories: not having a TMJ disorder (corresponding to Dimitroulis category 1) and having a TMJ disorder (corresponding to Dimitroulis categories 2, 3, and 4). This led to the use of a generalized linear model with a logistic link function for further analysis. The probabilities of having TMD were obtained as the corresponding classifications, determined by the cut-point maximizing sensitivity and 1-specificity, measured on the Receiver Operating Characteristic (ROC) curve. Preliminary results indicate a model with improved accuracy $(68,9\%)$ and satisfactory discriminatory power, measured by the Area Under the Curve of the Receiver Operating Characteristic curve (AUC), of 0.76 in the model based on the test sample.

Conclusion: This study has enabled the identification of some of the most relevant explanatory variables in the diagnosis of TMD, resulting in a model that can make a classification of having the disease based on these variables. The measurements in this set of variables are easily obtainable and incur no cost. However, the authors suggest that shortly, to increase accuracy, the model could be improved by including more observations in all categories of Dimitroulis classification, particularly in higher severities categories.

Keywords

Dimitroulis Scale, Temporomandibular disorders, Logistic Regression, Receiver Operating Characteristic curve, bidirectional stepwise method.

References

Ângelo, D. F., Mota, B., João, R. S., Sanz, D., Cardoso, H. J. (2023). Prevalence of Clinical Signs and Symptoms of Temporomandibular Joint Disorders Registered in the EUROTMJ Database: A Prospective Study in a Portuguese Center. *Journal of Clinical Medicine*, 12(10), 3553. MDPI AG.
http://dx.doi.org/10.3390/jcm12103553.

Dimitroulis, G. (2013). A new surgical classification for temporomandibular joint disorders.*Int J Oral Maxillofac Surg.*, 42(2), 218–222.
https://doi.org/10.1016/j.ijom.2012.11.004.

Fawcett, Tom (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters.* 27 (8): 861–874.
https://doi:10.1016/j.patrec.2005.10.010.

Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013). Applied Logistic Regression. 3rd Edition, John Wiley & Sons, Hoboken, NJ.
https://doi.org/10.1002/9781118548387.

Harrell, F. E. (2001). Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis. Springer-Verlag, New York.

Development, validation and use of imputed data in precision medicine

Cecilia Balocchi[1], **Massimiliano Russo**[2], **Stefano Favaro**[3], **Steffen Ventz**[4], **Lorenzo Trippa**[5]

[1] *University of Edinburgh, School of Mathematics, UK, cecilia.balocchi@ed.ac.uk*
[2] *Harvard University, Division of Pharmacoepidemiology and Pharmacoeconomics, USA, mrusso@bwh.harvard.edu*
[3] *University of Torino, Department of Economics Social Studies Applied Mathematics and Statistics, Italy*
[4] *University of Minnesota, Division of Biostatistics, USA*
[5] *Harvard University, School of Public Health and Dana Faber Institute, USA.*

Abstract

Data from past clinical trials play a crucial role in precision medicine by improving the efficiency of various methods. Despite the importance of these data is widely recognised and there is an increasing availability of methods that use them, they are still difficult to obtain in a reasonable time. In this paper, we create pseudo patient-level data leveraging trial summaries available in publications and external data from similar disease settings. Specifically, we consider several algorithms that generate data from the distribution of the available external data, restricted to match the published summary statistics. This is achieved by considering a loss-based Bayesian approach, combined with either simple and interpretable bootstrap algorithms or flexible black-box optimisation procedures. We demonstrate that combining the produced imputed data with the available external data improves inference in different real examples.

Keywords

External data; data imputation; meta analysis; precision medicine.

Predictive Accuracy of Stroke Risk Prediction Models Across Black and White Race, Sex, and Age Groups

Chuan Hong[1], **Michael J. Pencina**[1], **Daniel M. Wojdyla**[1], **Jennifer L. Hall**[2], **Suzanne E. Judd**[3], **Michael Cary**[1], **Matthew M. Engelhard**[1],**Samuel Berchuck**[1],**Ying Xian**[4],**Ralph D'Agostino Sr.**[5],**George Howard**[3],**Brett Kissela**[6],**Ricardo Henao**[1]

[1] *Duke AI Health, Department of Biostatistics & Bioinformatics, Duke University School of Medicine, Durham, NC*
[2] *American Heart Association, Dallas, TX*
[3] *School of Public Health, University of Alabama at Birmingham, Birmingham, AL*
[4] *Department of Neurology, University of Texas Southwestern Medical Center, Dallas, TX*
[5] *Department of Mathematics & Statistics, Boston University Arts & Sciences, Boston, MA*
[6] *College of Medicine, University of Cincinnati, Cincinnati, OH*

Abstract

Importance: Stroke is the fifth highest cause of death in the US and a leading cause of serious long-term disability with particularly high risk in Black individuals. Quality risk prediction algorithms, free of bias, are key for comprehensive prevention strategies.
Objective: To compare the performance of stroke-specific algorithms with Pooled Cohort Equations (PCE) developed for atherosclerotic cardiovascular disease (ASCVD) for the prediction of new onset stroke across different subgroups (race, sex, and age) and to determine the added value of novel machine learning techniques.
Design, Setting, and Participants: Retrospective Cohort Study on combined and harmonized data from Black and White participants of the Framingham Offspring, ARIC, MESA and REGARDS studies

(1983-2019) conducted in the US. Participants (N=62,482) who were 45 years or older and free of stroke or transient ischemic attack (TIA) at baseline were included.
Exposures: Published stroke-specific algorithms from Framingham and REGARDS (based on self-reported risk factors) as well as PCE for ASCVD plus two newly developed machine learning algorithms.
Main Outcomes and Measures: Models designed to estimate the 10-year risk of new onset stroke (ischemic or hemorrhagic). Discrimination C-index and calibration ratios of expected versus observed event rates at 10 years. Analyses conducted by race, sex, and age groups.
Results: The combined study sample included $62,482$ participants (median age 61, 54% women and 29% Black individuals). Discrimination C-indices were not significantly different for the two stroke-specific models (Framingham, 0.72; REGARDS self-report, 0.73) versus the PCE (0.72): differences ≤ 0.01 (P-values< 0.05) in the combined sample. Significant differences in discrimination were observed by race: the C-indices were 0.76 for all three models in White versus 0.69 in Black women (all P-values< 0.001) and $0.71 - 0.72$ in White and 0.64-0.66 in Black men (all P-values< 0.01). The ratios of observed to expected 10-year stroke rates were closest to 1 for the REGARDS self-report model (1.04, 95% confidence interval: $1.00, 1.09$) and indicated risk overestimation for Framingham Stroke (0.86, 95% confidence interval: $0.82, 0.89$) and PCE (0.74, 95% confidence interval: $0.71, 0.77$). Performance did not significantly improve when novel machine learning algorithms were applied.
Conclusions and Relevance: In this analysis of Black and White individuals without stroke or TIA among four US cohorts, existing stroke-specific risk prediction models and novel machine learning techniques did not significantly improve discriminative accuracy for new onset stroke compared with the PCE, and the REGARDS self-report model had the best calibration. All algorithms exhibited worse discrimination in Black compared with White individuals, indicating the need to expand the pool of risk factors and improve modeling techniques to address observed racial disparities and improve model performance.

Keywords

Stroke, Risk prediction, Retrospective cohort study, Demographic disparity.

Matrix-Variate Canonical Correlation Analysis

Daniel Kessler[1], **Elizaveta Levina**[2]

[1] *University of Washington, Department of Statistics, United States, dakess@uw.edu*
[2] *University of Michigan, Department of Statistics, United States, elevina@umich.edu*

Abstract

We consider the extension of Canonical Correlation Analysis (CCA) to the matrix-variate setting, where one or both of the random vectors of classical CCA is replaced by random matrices. The goal remains the identification of pairs of linear functions that transform the data into maximally correlated canonical variates. We exploit matrix-specific structure by seeking low-rank representations through the use of a nuclear norm penalty. Although generally applicable to matrix-variate data, this approach is motivated by applications in network neuroscience, where the matrix-variate data is a participant-specific connectivity matrix of spatial correlations. When applied to network data, these low-rank canonical directions can be understood as seeking latent network structure. We show in synthetic data that our approach is effective at recovering low rank signals even in noisy cases with relatively few observations, and we apply the method to human neuroimaging data.

Keywords

Covariance matrix, matrix computations, network data, penalization, neuroimaging.

On the connectivity of community affiliation graph

Daumilas Ardickas[1], **Mindaugas Bloznelis**[2]

[1] *Vilnius University, Institute of computer science, Lithuania, daumilas.ardickas@mif.vu.lt*
[2] *Vilnius University, Institute of computer science, Lithuania,, mindaugas.bloznelis@mif.vu.lt*

Abstract

Community affiliation graph is a random graph model of an overlapping community network, where communities are represented by independent binomial random graphs of various sizes and densities [Yang, J. and Leskovec, J., ACM Trans. Intell. Syst. Technol. 2014]. We are interested in the parametric regime where the community affiliation graph becomes connected and establish the connectivity threshold under the optimal moment conditions on the (limiting) size/density distribution of the contributing binomial random graphs.

Keywords

Community affiliation graph, overlapping community network, connectivity threshold.

Confidence in Causal Inference under Structure Uncertainty

David Strieder[1], **Mathias Drton**[2]

[1] *School of Computation, Information and Technology, Munich Center for Machine Learning, Technical University of Munich, Germany, david.strieder@tum.de*
[2] *School of Computation, Information and Technology, Munich Center for Machine Learning, Technical University of Munich, Germany, mathias.drton@tum.de*

Abstract

Inferring the effect of interventions within complex systems is a fundamental problem of statistics. A widely studied approach employs structural causal models that postulate noisy functional relations among a set of interacting variables. The underlying causal structure is then naturally represented by a directed graph whose edges indicate direct causal dependencies. In a recent line of work, additional assumptions on the causal models have been shown to render this causal graph identifiable from observational data alone. One example is the assumption of linear causal relations with homoscedastic errors that we will take up in this work. When the graph structure is known, classical methods may be used for calculating estimates and confidence intervals for causal effects. However, in many applications, expert knowledge that provides an a priori valid causal structure is not available. Lacking alternatives, a commonly used two-step approach first learns a graph and then treats the graph as known in inference. This, however, yields confidence intervals that are overly optimistic and fail to account for the data-driven model choice. We argue that to draw reliable conclusions, it is necessary to incorporate the remaining uncertainty about the underlying causal structure in confidence statements about causal effects. To address this issue, we present a framework based on test inversion that allows us to give confidence regions for

total causal effects that capture both sources of uncertainty: causal structure and numerical size of nonzero effects.

Keywords

Confidence intervals, causal effects, linear structural equation models, homoscedasticity, graphical models.

Probabilistic Guarantees on Sensitivities of Bayesian Neural Network

Diptarka Saha[1], **Zihe Liu**[2], **Feng Liang**[3]

[1,2,3] *University of Illinois at Urbana-Champaign* *

Abstract

The study of theoretical properties of wide and deep neural networks is a growing body of research that complements their empirical success. In this paper, we study the partial derivatives of a random, wide, fully connected Bayesian neural network w.r.t. each individual feature, which we refer to as the feature sensitivities of that neural network. Under a set of general conditions, as the network widens, we show that these sensitivities are consistent around their mean. Moreover, we show that these sensitivities, as a random function of the features, converge in distribution to Gaussian processes under proper scaling. We discuss the ramifications of such behavior and how this can be leveraged to obtain robust estimates of feature importance and pruning strategies. We employ these strategies to prune larger models to obtain more concise models with equivalent predictive power.

Keywords

Bayesian Neural Network (BNN), Feature sensitivities, Gaussian processes convergence, Feature relevance, Pruning strategies.

*DS is the corresponding author, please direct any questions to saha12@illinois.edu

A Workflow for Statistical Inference in Stochastic Gradient Descent

Rahul Singh[1], **Abhinek Shukla**[2], **Dootika Vats**[3],

[1] *University of Haifa, Department of Statistics, Israel, wrahulsingh@gmail.com*
[2] *Indian Institute of Technology Kanpur, Department of Mathematics and Statistics, India, abhinek@iitk.ac.in*
[3] *Indian Institute of Technology Kanpur, Department of Mathematics and Statistics, India, dootika@iitk.ac.in*

Abstract

The stochastic gradient descent (SGD) algorithm is used for parameter estimation, particularly for massive datasets and online learning. Inference in SGD has been a generally neglected problem and has only recently started to get some attention. I will first introduce SGD for relatively simple statistical models and explain the limiting behavior of Averaged SGD. Then, I will present a memory-reduced batch-means estimator of the limiting covariance matrix that is both consistent and amenable to finite-sample corrections. Further, I will discuss the practical usability of error covariance matrices for problems where SGD is relevant, and present ongoing challenges in this area.

Keywords

Stochastic gradient descent, inference, covariance matrix, batch-means estimator.

Smoothing Method for Unit Quaternion Time Series: An application to Multiple Sclerosis motion data

Elena Ballante[1,2], **Lise Bellanger**[3], **Pierre Drouin**[3,4], **Silvia Figini**[1], **Aymeric Stamm**[3]

[1] *Department of Political and Social Sciences, University of Pavia, Italy*
[2] *BioData Science Unit, IRCCS Mondino Foundation, Italy*
[3] *Department of Mathematics Jean Leray, UMR CNRS 6629, Nantes University, France*
[4] *Department of Research and Development, UmanIT, France*

Abstract

Introduction and Objective: Smoothing orientation data is a fundamental issue in different research fields where time series are involved. Unit Quaternion Time Series are powerful objects to describe the rotation of a rigid body in the 3D space. Different methods to smooth time series in quaternion algebra were described in literature, but their application in real world problems is still an open point. The aim of this work is to present an effective and easy to apply method for quaternion time series smoothing to classify a dataset of patients affected by Multiple Sclerosis.
Methods and Result: We proposed a new method based on deploying logarithm function to transform the quaternion time series in a real three-dimensional time series that can be smoothed with classical methods. In literature, a version of this algorithm using angular velocity transformation already exists, but logarithm function is smoother and intrinsically different from angular velocity. Usually, this leads to better classification performances.
Our method is applied and compared with the literature one on a dataset of Multiple Sclerosis patients. The 27 subjects are assigned to

three different levels of walking impairments. Smoothing methods performances are compared in terms of accuracy, AUC, ARI and another literature index to assess ordinal classification (OrdInd). We select the methods that perform better than the classification of original curves. Even if the angular velocity method shows higher accuracy, the logarithm method performs better in the other three indices, see Table 3.1.

smoothing	transformation	Accuracy	AUC	ARI	OrdInd
None	none	0.627	0.571	0.451	0.428
Fourier	ang vel	**0.720**	0.683	0.454	0.224
Fourier	logarithm	0.633	0.696	**0.505**	0.200
Spline	logarithm	0.633	0.688	**0.505**	0.209
Wavelet	ang vel	**0.720**	0.700	0.454	0.224
Wavelet	logarithm	0.673	**0.713**	0.471	**0.179**

Table 3.1: Performances of the different methods compared.

Conclusions Our results confirm the need for applying smoothing techniques in the preprocessing stage of classification of unit quaternion time series. Especially when the data are noisy, the suitability of deploying the proposed method emerges to obtain overall better results than the literature method.

Keywords

Signal Smoothing, Quaternion time series, Multiple Sclerosis.

Regression Discontinuity Designs Under Interference

Elena Dal Torrione[1], Tiziano Arduini[2], Laura Forastiere[3]

[1] *Tor Vergata University of Rome, Department of Economics and Finance, Italy, elena.daltorrione@students.uniroma2.eu*
[2] *Tor Vergata University of Rome, Department of Economics and Finance, Italy, tiziano.arduini@uniroma2.it*
[3] *University of Yale, Department of Biostatistics, U.S., laura.forastiere@yale.edu*

Abstract

Interference takes place whenever a "treatment" on one unit affects the outcome of another unit, and such a phenomenon can occur in regression discontinuity designs (RDDs). For instance, in conditional cash transfer programs for education, eligible children's schooling choices may affect the schooling choices of their ineligible peers. We propose an extension of the continuity-based framework to RDD to identify and estimate a set of causal estimands in the presence of interference. In this setting, assignment to effective treatment is determined by a unit's score and the scores of other units—for example, their neighbors. Unlike the standard RDD, embedding the exposure mapping function as a summary of the treatment of other units can lead to complex, multidimensional frontiers. We provide a method to characterize such frontiers for a broad class of exposure mapping functions and derive generalized continuity assumptions to identify the proposed es- timands. Next, we develop a distance-based estimation method capable of handling multidimensional, and potentially heterogeneous, score spaces, and evaluate its empirical performance in a simulation study. Finally, we apply the presented methodology to the PROGRESA/Oportunidades data to es- timate the spillover effects of financial aid to families on children's school attendance.

Keywords

Causal inference, regression discontinuity, interference, SUTVA, local polynomials.

Longitudinal Structural Equation Modelling Assessment of Factors influencing Learning Mathematics in a Bayesian Framework

Elizabeth Stojanovski[1]

[1] *The University of Newcastle, School of Information and Physical Sciences, Australia, elizabeth.stojanovski@newcastle.edu.au*

Abstract

Predictive associations between self-efficacy in mathematics during secondry school and the study of higher level mathematics is examined using a longitudinal structural equation modelling framework with Bayesian extensions to demonstrate the versatility of incorporating different aspects of uncertainty within this modelling framework.

Keywords

Longitudinal, multivariate, Bayesian.

Sandwich Boosting for semiparametric estimation with grouped data

Elliot H. Young[1], **Rajen D. Shah**[1]

[1] *University of Cambridge, Statistics Laboratory, UK,*
ey244@cam.ac.uk

Abstract

We study partially linear models in settings where observations are arranged in independent groups but may exhibit within-group dependence. Existing approaches estimate linear model parameters through weighted least squares, with optimal weights (given by the inverse covariance of the response, conditional on the covariates) typically estimated by maximising a (restricted) likelihood from random effects modelling or by using generalised estimating equations. We introduce a new 'sandwich loss' whose population minimiser coincides with the weights of these approaches when the parametric forms for the conditional covariance are well-specified, but can yield arbitrarily large improvements in linear parameter estimation accuracy when they are not. Under relatively mild conditions, our estimated coefficients are asymptotically Gaussian and enjoy minimal variance among estimators with weights restricted to a given class of functions, when user-chosen regression methods are used to estimate nuisance functions. We further expand the class of functional forms for the weights that may be fitted beyond parametric models by leveraging the flexibility of modern machine learning methods within a new gradient boosting scheme for minimising the sandwich loss. We demonstrate the effectiveness of both the sandwich loss and what we call 'sandwich boosting' in a variety of settings with simulated and real-world data.

Keywords

Semparametric statistics, dependent data, partially linear models, boosting.

Localizing differences in graphical models

Erika Banzato[1], **Monica Chiogna**[2], **Vera Djordjilović**[3], **Davide Risso**[1]

[1] *University of Padua, Department of Statistical Sciences, Italy*
[2] *University of Bologna, Department of Statistical Sciences, Italy*
[3] *Ca' Foscari University of Venice, Department of Economics, Italy*

Abstract

The use of graphs for describing and analyzing networks is increasing in relevance in many fields of applications. This work deals with statistical inferences in differential network, where the focus is on whether and how a specific network changes between two conditions, with focus on the application to genomic data: microarray and single-cell RNAseq data. In decomposable Gaussian graphical models, the problem of testing the equality of two models breaks down into a sequence of problems defined on smaller sets of variables (cliques). This renders tractable inference in the setting of large-scale graphical models, where the dimension p is higher than the available sample size n. Using the information on the graphical structure allows us both to improve the power of detecting a difference between the two distributions under study and to localize that difference (Djordjilović and Chiogna, 2022). However, the decomposition leads to tackling tests of different dimensions at the same time. In this setting, using the corrected statistic in Banzato et al. (2022) leads to valid inference at different dimensionality regimes and overcomes some weaknesses that occur at small sample sizes and in particular when the dimension p is close to the sample size n. When dealing with count data, the extension of the method above is not straightforward, as data transformation is not a valid choice. In this scenario, it is convenient to assume a Poisson graphical model, where all node-conditional distributions are Poisson. Our proposal is to estimate, for each node, two models: one under the null hypothesis (reduced model) and the other by including the

interaction with a group membership variable (full model). By testing the equality of the two models, one can obtain a list of differing nodes, and localize the differences by testing the significance of the interaction terms in the full model.

Keywords

Graphical models, Likelihood ratio test, Hypothesis test.

References

Banzato, E.; Chiogna, M.; Djordjilović, V. and Risso, D. (2023). A Bartlett-type correction for likelihood ratio tests with application to testing equality of Gaussian graphical models. *Statistics & Probability Letters*, *193*, 109732.

Djordjilović, V. and Chiogna, M. (2022) Searching for a source of difference in graphical models. *Journal of Multivariate Analysis*, *190*, 1–13.

Testing Markov Random Field Models for Binary Spatial Data

Eva Biswas[1], **Andee Kaplan**[2], **Dan Nordman**[3]

[1] *Iowa State University, Statistics, USA, ebiswas@iastate.edu*
[2] *Colorado State University, Statistics, USA, andee.kaplan@colostate.edu*
[3] *Iowa State University, Statistics, USA, dnordman@iastate.edu*

Abstract

Binary spatial observations often arise in environmental and ecological studies (e.g. disease in plants, presence of a species in an area, image analysis), where the Markov random field (MRF) models are commonly applied for such data. Despite the prevalence and the long history of MRF models for spatial binary data, no appropriate formal diagnostics have existed for these models, which is compounded by the fact that the MRF involves the specification of "neighborhoods" that are difficult to assess. In this work, we propose a simple but rigorous goodness-of-fit test for assessing an MRF model for spatial binary values. The test statistic is a type of conditional Moran's I based on the fitted conditional probabilities, which intends to detect departures in general model form, including the neighborhood structure. An effective bootstrap procedure is used to calibrate the spatial test. Numerical studies show that the goodness-of-fit test performs well in both size control and power, even when the deviations from the null model lie purely in the neighborhood specification. The test then provides an attractive alternative to other diagnostics, which are either narrow and ad hoc or unstable with low power for binary data. We illustrate the spatial test with an application to studying the breeding pattern (and decline) of grasshopper sparrows across Iowa.

Keywords

Markov Random field, Bootstrap, Goodness of fit.

Variance-Reduced Stochastic Optimization for Efficient Inference of Hidden Markov Models

Evan Sidrow[1], **Nancy Heckman**[2], **Alexandre Bouchard-Côté**[2], **Sarah M. E. Fortune**[3], **Andrew W. Trites**[4,5], **Marie Auger-Méthé**[2,4]

[1] *University of British Columbia, Department of Statistics, Canada, evan.sidrow@stat.ubc.ca*
[2] *University of British Columbia, Department of Statistics, Canada*
[3] *Dalhousie University, Department of Oceanography, Canada*
[4] *University of British Columbia, Institute for the Oceans and Fisheries, Canada*
[5] *University of British Columbia, Department of Zoology, Canada*

Abstract

Hidden Markov models (HMMs) are popular models to identify a finite number of latent states from sequential data. However, fitting them to large data sets can be computationally demanding because most likelihood maximization techniques require iterating through the entire underlying data set for every parameter update. We propose a novel optimization algorithm that updates the parameters of an HMM without iterating through the entire data set. Namely, we combine a partial E step with variance-reduced stochastic optimization within the M step. We prove the algorithm converges under certain regularity conditions. We test our algorithm empirically using a simulation study as well as a case study of kinematic data collected using suction-cup attached biologgers from eight northern resident killer whales (*Orcinus orca*) off the western coast of Canada. In both, our algorithm converges in fewer epochs and to regions of higher likelihood compared to standard numerical optimization techniques. Our algorithm allows practitioners to fit complicated HMMs to large time-series data sets more efficiently than existing baselines.

Keywords

Expectation-maximization algorithm, Maximum likelihood estimation, State space model, Statistical ecology, Stochastic gradient descent.

Sequential pointwise Monte-Carlo approximation of data depth with statistical guarantees

Felix Gnettner[1], **Claudia Kirch**[2], **Alicia Nieto-Reyes**[3]

[1] *Otto-von-Guericke-Universität Magdeburg, Germany, felix.gnettner@ovgu.de*
[2] *Otto-von-Guericke-Universität Magdeburg, Germany, claudia.kirch@ovgu.de*
[3] *Universidad de Cantabria, Santander, Spain, alicia.nieto@unican.es*

Abstract

Evaluating population depth functions which have a representation as a sum of indicator variables can be costly in higher dimensions, respectively big data settings. For instance, the simplicial depth with respect to n iid d-variate observations suffers from this problem for $d \geq 5$. Instead of computing the exact value, we propose a sequential Monte-Carlo procedure that helps to decide whether or not the target value lies in an element of a pre-specified finite family of intervals. The procedure terminates according to a criterion such that wrong decisions are only made with a pre-specified error probability of γ. For a more general class of sum-based depth functions a similar method, that satisfies the precision guarantee asymptotically, is introduced. The latter algorithm can be applied in binary classification. An empirical data study demonstrates the performance of the proposed methods.

Keywords

Data depth, Monte-Carlo approximation, sequential testing.

Achieving fairness with a simple ridge penalty

Marco Scutari[1], **Francesca Panero**[2], **Manuel Proissl**[3]

[1] *Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Switzerland, scutari@bnlearn.com*
[2] *London School of Economics and Political Science, United Kingdom, f.panero@lse.ac.uk*
[3] *IBM, Switzerland, mproissl@gmail.com*

Abstract

This work offers a general framework for estimating regression models subject to a user-defined level of fairness. We enforce fairness as a model selection step in which we choose the value of a ridge penalty to control the effect of sensitive attributes. We will show how it performs on benchmark dataset against other models in the literature. Our proposal can accommodate multiple definitions of fairness like statistical parity, equality of opportunity and individual fairness. It deals with multiple sensitive attributes (continuous or discrete), is mathematically simple, with a solution that is partly in closed form and produces estimates of the regression coefficients that are intuitive to interpret as a function of the level of fairness. Furthermore, it is easily extended to generalised linear models, kernelised regression models and other penalties.

Keywords

Fairness, linear regression, generalised linear models, ridge regression.

A note on Social Learning in nonatomic Routing Games

Francesco Giordano[1]

[1] *HEC Paris, Department of Economics and Decision Sciences, France, francesco.giordano@hec.edu*

Abstract

This work expands on the existing body of research about Social Learning in Dynamic Non-atomic Routing Games to include infinite capacity multi-commodity instances and capacitated instances with one or more commodities. We define when a network state is identifiable, and show the network conditions sufficient to achieve almost sure learning. In capacitated instances, a sufficient condition to achieve learning is that the network respects a condition called *weak capacity conservation*, that is independent on the network topology. In infinite capacity instances, a sufficient condition is that the network is *series-parallel* in each commodity sub-network.

Keywords

Routing Games, Wardrop Equilibrium, Incomplete Information, Social Learning, Congestion Games, Learning Failures, SP Networks, Capacity Conservation.

Backward Filtering Forward Guiding for Markov processes

Frank van der Meulen[1], **Moritz Schauer**[2]

[1] *Vrije Universiteit Amsterdam, Department of mathematics, The Netherlands, f.h.van.der.meulen@vu.nl*
[2] *Chalmers University of Technology and University of Gothenburg, Department of Mathematical Sciences, Sweden, smoritz@chalmers.se*

Abstract

We consider the setting where a continuous time stochastic process evolves on a directed tree. Examples include diffusions generated by a stochastic differential equation or continuous time Markov chains. The process is only observed partially at the leaf vertices of the graph and we wish to reconstruct the process probabilistically, a problem known as the *smoothing problem*. This setting incorporates state space models as a special case, but the more general setting is of fundamental importance for example in phylogenetics. The smoothing problem is only tractable in very specific settings, for example when the stochastic process is Brownian Motion. Here, we will discuss a structured approach to deal with this problem in general. Our approach is based upon

1. the backward information filter (also known in phylogenetics as Felsenstein's algorithm);
2. conditioning a continuous time stochastic process by exponential change of measure;
3. approximating the change of measure and correcting for the approximation error by stochastic simulation.

This yields an elegant solution that can be incorporated in MCMC, SMC or variational inference. We illustrate the approach in numerical examples.

Keywords

Backward information filter, Bayesian network, Branching diffusion process, Directed acyclic graph, Doob's h-transform, Twisted measure, Guided process.

Extreme Value Index estimation with Probability Weighted Moments

Frederico Caeiro[1], **M. Ivette Gomes**[2]

[1] *Center for Mathematics and Applications (CMA), NOVA University Lisbon, Portugal, fac@fct.unl.pt*
[2] *CEAUL, University of Lisbon, Portugal, migomes@ciencias.ulisboa.pt*

Abstract

In hydrology and other applied fields, Probability Weighted Moments (PWMs) are a common tool for estimating distribution parameters (Greenwood *et al.*, 1979). In this study, we focus on the PWM estimator of the Extreme Value Index (EVI) of a Pareto type model (Caeiro and Gomes, 2011), based on the observation above a high threshold. Due to the estimators specific properties, a direct estimation of an "optimal" threshold is not straightforward (Caeiro *et al.*, 2014, Caeiro and Gomes, 2022). In this work, we study an adaptive approach to choose the number of order statistics to be used in the estimation. Furthermore, we apply the introduced methodology to a dataset in the insurance field.

Keywords

Extreme value index, probability weighted moments, statistics of extremes, threshold.

Acknowledgement

Research partially supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under the projects UIDB/00297/2020 (CMA/UNL) and UIDB/00006/2020 (CEAUL).

References

Caeiro, F. and Gomes, M. I. (2011). Semi-parametric tail inference through probability weighted moments. *Journal of Statistical Planning and Inference*, *141*(2), 937–950.

Caeiro, F., Gomes, M. I., and Vandewalle, B. (2014). Semi-parametric probability-weighted moments estimation revisited. *Methodology and Computing in Applied Probability*, *16*(1), 1–29.

Caeiro, F., Gomes, M.I. (2022). Computational Study of the Adaptive Estimation of the Extreme Value Index with Probability Weighted Moments. In: Bispo, R., Henriques-Rodrigues, L., Alpizar-Jara, R., de Carvalho, M. (eds) *Recent Developments in Statistics and Data Science. SPE 2021. Springer Proceedings in Mathematics & Statistics*, *398*, 29–39, Springer, Cham.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979). Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, *15*(5), 1049–1054.

Multiply Robust Estimation of Heterogeneous Direct and Indirect Policy Exposures

Gary Hettinger[1*], **Youjin Lee**[2], **Nandita Mitra**[1]

[1] *University of Pennsylvania, Department of Biostatistics, Epidemiology, and Informatics, U.S.A*
[2] *Brown University, Department of Biostatistics, U.S.A*
[*] *ghetting@pennmedicine.upenn.edu*

Abstract

Public policy interventions are often evaluated with the difference-in-differences (DiD) approach, which does not directly account for exposure heterogeneity or spillover effects common with such policies. For example, an excise tax on sweetened beverages in Philadelphia, Pennsylvania (PHL) was associated with substantial decreases in volume sales of taxed beverages in PHL as well as increases in beverage sales of nontaxed bordering counties. The latter association may be explained by cross-border shopping behaviors of PHL residents, which vary by residents' accessibility to an untaxed region, among other population dynamics. Because such effects can substantially offset the total effect of such interventions, understanding effect dynamics is essential to holistically evaluate public policies. Further, such insights may help predict policy effects under diverse implementation strategies. To address these concerns, we build upon efficient semiparametric theory to extend DiD methodology to robustly identify the causal effects of policy interventions under cross-border shopping and exposure heterogeneity. Our estimators relax standard assumptions regarding confounding, binary exposures, spillover, and model specification. Here, we present initial work demonstrating our framework with an evaluation of the PHL Beverage Tax policy.

Keywords

Causal Inference, Difference-in-Differences, Dose Response Curve, Spillover.

Deep Learning for Spatial Statistics

Ghulam A. Qadir[1], **Tilmann Gneiting**[2]

[1] *Heidelberg Institute for Theoretical Studies, Computational Statistics, Germany, ghulam.qadir@h-its.org*
[2] *Heidelberg Institute for Theoretical Studies, Computational Statistics, Germany, tilmann.gneiting@h-its.org*

Abstract

The traditional spatial statistics has been mostly centralized on stochastic process models, typically, the Gaussian process models, which are well studied and understood in terms of inference and predictions. While so far the stochastic process models have been the leading focus in spatial statistics, the deep learning methods are now also receiving notable heed from spatial statisticians. In this paper, we explore the use of deep learning for spatial statistics by using the theoretical framework of the Neural Tangent Kernel (NTK) which investigates the training dynamics of deep learning models in the infinite-width limit. In our work, we study some recently developed deep learning variants for spatial statistics against the Gaussian process regression through multiple simulation studies and data applications. Through our case studies, we provide some crucial guidelines for developing the appropriate deep learning models in the context of spatial data modeling.

Keywords

Deep Learning, Neural Networks, Wide-Networks, Spatial Statistics, Gaussian Process.

Manifold learning with sparse regularised optimal transport

Stephen Zhang[1], **Gilles Mordant**[2], **Tetsuya Matsumoto**[3], **Geoffrey Schiebinger**[4]

[1] *University of Melbourne, School of Mathematics and Statistics, Australia, stephenz@student.unimelb.edu.au*
[2] *Universität Göttingen, Institut für Mathematische Stochastik, Germany, gilles.mordant@uni-goettingen.de*
[3] *University of British Columbia, Department of Mathematics, Canada*
[4] *University of British Columbia, Department of Mathematics, Canada, geoff@math.ubc.ca*

Abstract

Manifold learning is a central task in modern statistics and data science. Many datasets (cells, documents, images, molecules) can be represented as point clouds embedded in a high dimensional ambient space, however the degrees of freedom intrinsic to the data are usually far fewer than the number of ambient dimensions. The task of detecting a latent manifold along which the data are embedded is a prerequisite for a wide family of downstream analyses. Real-world datasets are subject to noisy observations and sampling, so that distilling information about the underlying manifold is a major challenge. We propose a method for manifold learning that utilises a symmetric version of optimal transport with a quadratic regularisation that constructs a *sparse* and *adaptive* affinity matrix, that can be interpreted as a generalisation of the bistochastic kernel normalisation. We prove that the resulting kernel is consistent with a Laplace-type operator in the continuous limit, establish robustness to heteroskedastic noise and exhibit these results in simulations. We identify a highly efficient computational scheme for computing this optimal transport for discrete data and demonstrate that it outperforms competing methods in a set of examples.

Keywords

Regularised optimal transport, manifold learning, single cell analysis.

Poisson Kernel-Based Tests for Uniformity on the d-dimensional Sphere with the `QuadratiK` package

Giovanni Saraceno[1], **Marianthi Markatou**[1], **Yuxin Ding**[2]

[1] *University at Buffalo, Department of Biostatistics, USA*
[2] *Eli Lilly Corporation, USA*

Abstract

Directional statistics, encompassing data represented as unit vectors in high-dimensional space or as points on the hyper-sphere, is increasingly significant in contemporary applications, and the unit hyper-sphere $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d : ||x||^2 = 1\}$ represents a pivotal framework in this domain. Additionally, many non-directional datasets can be usefully re-expressed in the form of directions, for example standardized gene expressions in genetic analysis, with zero mean and unit variance. Existing literature, however, is limited reference the performance evaluation of uniformity tests with multiple modes in the data. We introduce advanced tests for uniformity formulated upon a family of kernel-based quadratic distances, that can handle well multi-modal spherical data. We discuss a special class of distance kernels, namely diffusion kernels, characterized by their tunability and computational simplicity. Among these diffusion kernels, the Poisson kernel offers a valuable tool for our analysis. As the Brownian motion generates the Normal distribution in the Euclidean space, the Poisson kernel stands as its natural generalization to the sample space $\mathcal{S}^{d-1}$. The d-dimensional Poisson kernel is explored in detail: eigendecomposition, centered version and degrees of freedom. Utilizing the centered kernel, we establish uniformity tests, characterized as U- and V-statistic estimates of quadratic distances. The intrinsic connection between Poisson kernel-based tests and the Sobolev test class is elucidated. A comparative analysis against the Rayleigh, Bingham, Giné, and Ajne

tests using simulations underscores the enhanced efficacy of our proposed tests. We then present the novel `QuadratiK` package, a curated compendium of methods anchored on kernel-based quadratic distances and implemented in `R` and `Python`. Embedded within this package is the `pk.test` function, which is designed to perform the Poisson kernel-based tests. Additionally, the package provides the algorithm for automated selection of the tuning parameter. The implementation, efficiently using `C++` code and parallel computing, ensures rapid and robust test statistic calculations. The function's usage and effectiveness are further demonstrated via real data examples.

Keywords

Kernel-based Quadratic Distance, Poisson Kernel, Poisson Kernel-Based Density, Spherical Data, Tests for Uniformity.

Identifying Differential Item Functioning in Diagnostic Classification Models

Guaner Rojas[1]

[1] *Universidad de Costa Rica, School of Statistics, Costa Rica,*
guaner.rojas@ucr.ac.cr

Abstract

Diagnostic classification models (DCMs) are defined as a statistical family of confirmatory multidimensional latent-variable models with categorical latent variables. In most cases, DCMs are a type of statistical model used in educational and psychological contexts to measure categorical attributes that describe latent variables or constructs present in cognitive tests and survey questionnaires. When measuring constructs within population groups, the functioning of items or questions may be influenced by extraneous factors unrelated to the construct. This phenomenon is identified as differential item functioning (DIF) and occurs when individuals from different population groups exhibit distinct response probabilities to an item even though those individuals manifest the same level of ability or latent class. This talk will introduce a method for detecting DIF from a simpler to more complex model perspective. A simulation study is used to evaluate the method, which involves comparing probabilities using the *deterministic input noisy 'and' gate* (DINA) and the generalized DINA (GDINA) models to identify items that function differentially across groups.

Keywords

Classification, Diagnostic, Construct, DIF, Differences.

Effective sample size estimation based on the concordance between p-value and posterior probability of the null hypothesis

Han Wang[1], **Yan Zhang**[2], **Guosheng Yin**[3]

[1] *University of Hong Kong, Department of Statistics and Actuarial Science, Hong Kong, whanstat@hku.hk*
[2] *University of Hong Kong, Department of Statistics and Actuarial Science, Hong Kong, doraz@hku.hk*
[3] *Imperial College London, Department of Mathematics, UK, guosheng.yin@imperial.ac.uk*

Abstract

Estimating the effective sample size (ESS) of a prior distribution is an age-old yet pivotal challenge, with great implications for clinical trials and various biomedical applications. Although numerous endeavors have been dedicated to this pursuit, most of them neglect the likelihood context in which the prior is embedded, thereby considering all priors as "beneficial". In the limited studies of addressing harmful priors, specifying a baseline prior remains an indispensable step. In this paper, by means of the elegant bridge between the p-value and the posterior probability under the null hypothesis, we propose a new p-value based ESS estimation method in the framework of hypothesis testing, expanding the scope of existing ESS estimation methods in three key aspects: (I) We address the specific likelihood context of the prior, enabling the possibility of negative ESS values in case of prior-likelihood disconcordance. (II) By leveraging the well-established bridge between the frequentist and Bayesian configurations under noninformative priors, there is no need to specify a baseline prior which incurs another criticism of subjectivity. (III) By incorporating ESS into the hypothesis testing framework, our p-value based method transcends the conventional one-ESS-one-prior paradigm and accommodates one-ESS-multiple-priors paradigm, where the sole ESS may

reflect the collaborative impact of multiple priors in diverse contexts. Through comprehensive simulation analyses, we demonstrate the superior performance of the p-value based ESS estimation method in comparison with existing approaches. Furthermore, by applying this approach to an expression quantitative trait loci (eQTL) data analysis, we show the effectiveness of informative priors in uncovering gene eQTL loci.

Keywords

Informative prior, prior distribution, prior-likelihood disconcordance, p-value.

On universal inference in normal mixture models

Hongjian Shi[1] **and Mathias Drton**[2]

[1] *Department of Mathematics, TUM School of Computation, Information and Technology, Technical University of Munich, 85748 Garching bei München, Germany,* hongjian.shi@tum.de
[2] *Department of Mathematics, TUM School of Computation, Information and Technology, Technical University of Munich, 85748 Garching bei München, Germany,* mathias.drton@tum.de

Abstract

Recent work on game-theoretic statistics and safe anytime-valid inference (SAVI) provides new tools for statistical inference without assuming any regularity conditions. In particular, the framework of universal inference proposed by Wasserman, Ramdas, and Balakrishnan (2020) offers new solutions by modifying the likelihood ratio test in a data-splitting scheme (split likelihood ratio test). In this paper, we study the performance of the split likelihood ratio test under normal mixture models, which are canonical examples of models in which classical regularity conditions fail to hold. We first establish that under the null hypothesis, the split likelihood ratio test statistic is asymptotically normal with increasing mean and variance. Moreover, contradicting the usual belief that the flexibility of SAVI and universal methods comes at the price of a loss of power, we are able to prove that universal inference surprisingly achieves the same detection rate $(n^{-1} \log\log n)^{1/2}$ as the classical likelihood ratio test.

Keywords

E-value, universal inference, singularity, mixture model, likelihood ratio test.

References

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proc. Natl. Acad. Sci. USA*, 117(29):16880–16890.

Distribution-uniform anytime-valid inference

Ian Waudby-Smith[1], **Aaditya Ramdas**[2]

[1] *Carnegie Mellon University, Dept. of Statistics & Data Science, USA, ianws@cmu.edu*
[2] *Carnegie Mellon University, Dept. of Statistics & Data Science, USA, aramdas@cmu.edu*

Abstract

Are asymptotic confidence sequences and anytime p-values uniformly valid for a nontrivial class of distributions $\mathcal{P}$? We give a positive answer to this question by deriving *distribution-uniform* anytime-valid inference procedures. Historically, anytime-valid methods — including confidence sequences, testing by betting, universal inference, and other frameworks that enable inference at stopping times — have been justified nonasymptotically. Nevertheless, asymptotic procedures — such as those based on the central limit theorem (CLT) — occupy an important part of statistical toolbox due to their simplicity, universality, and weak assumptions. While recent work has derived asymptotic analogues of anytime-valid methods with the aforementioned benefits, these were not shown to be $\mathcal{P}$*-uniform* (meaning their asymptotics are not uniformly valid in a class of distributions $\mathcal{P}$). Indeed, the anytime-valid inference literature currently has no central limit theory to draw from that is both uniform in $\mathcal{P}$ and in time. This paper fills that gap by deriving a novel $\mathcal{P}$-uniform strong Gaussian coupling inequality, enabling $\mathcal{P}$-uniform anytime-valid inference for the first time.

Keywords

Strong invariance principles, sequential testing, confidence sequences, honest inference.

Exhaustive Nested Cross-Validation for High-dimensional Testing

Iris Ivy Gauran[1], **Hernando Ombao**[2], **Zhaoxia Yu**[3]

[1] *Biostatistics Group, King Abdullah University of Science and Technology (KAUST), Kingdom of Saudi Arabia, irisivy.gauran@kaust.edu.sa*

[2] *Biostatistics Group, King Abdullah University of Science and Technology (KAUST), Kingdom of Saudi Arabia, hernando.ombao@kaust.edu.sa*

[3] *Department of Statistics, University of California, Irvine, United States of America, zhaoxia@ics.uci.edu*

Abstract

Cross-validation is a widely utilized algorithmic technique for tasks such as estimating prediction error, tuning regularization parameters, and selecting the best predictive models. Nevertheless, its behavior is intricate due to various complex factors at play. In this study, we introduce a novel high-dimensional test based on the exhaustive nested cross-validation procedure. This method is not only straightforward to apply but also operates almost automatically in numerous scenarios, with minimal assumptions about the underlying data distribution. Furthermore, our proposed approach can establish valid confidence intervals for comparing prediction error differences between two model-fitting algorithms. To address concerns about computational complexity, we have derived a highly efficient expression for the cross-validation estimator. Our research also delves into strategies for enhancing statistical power in high-dimensional scenarios while preserving the Type I error rate. Lastly, we showcase the application of our method to an RNA sequencing study and biological data, illustrating its practical utility in real-world scenarios.

Keywords

Cross-validation, High-dimensional testing, Epigenetics.

Censored Multivariate Linear Regression Models with Autocorrelated Errors — A Classical and Bayesian Approach

Rodney Sousa[1], **Isabel Pereira**[2], **M. Eduarda Silva**[3]

[1] *Universidade de Aveiro*
CIDMA, Portugal, rodney@ua.pt
[2] *Universidade de Aveiro*
CIDMA, Departamento de Matemática, Portugal,
isabel.pereira@ua.pt
[3] *Universidade do Porto*
LIADD-INESCTEC, Faculdade de Economia, Portugal,
mesilva@fep.up.pt

Abstract

Censoring data occurs when measurements of the variable of interest can only be obtained within a certain interval, and observations whose values fall outside this interval are recorded as having values at the detection limits. Therefore, it is only known that the values of censored observations are below (or above) the detection limits. Censoring renders the observed data incomplete, and as a result, direct analysis of such data using traditional methods for estimating linear regression models, such as Ordinary Least Squares (OLS) or Generalized Least Squares (GLS), is not recommended. This is because it can lead to inconsistent estimates. Estimating multivariate linear regression models for censored and autocorrelated data (CMLR-AR) can pose a substantial challenge. Despite the proposal of numerous methods for parameter estimation in linear regression models with censored and autocorrelated data in the context of univariate data, extending these methods to the multivariate context is not straightforward. Many of the approaches found in the existing literature are applicable only to some specific cases of this model or prove to be impractical, particularly when dealing with a high percentage of censored

observations. In light of these limitations, the objective of this study was to develop practical and effective methods for the estimation of CMLR-AR models, employing both classical and Bayesian approaches. One of the major challenges of this work is the need to model two correlation matrices, namely, cross-correlation and temporal correlation. The two proposed methods, namely the Feasible Generalized Least Squares estimation and Gibbs sampler are based on the data augmentation strategy. To assess their accuracy, a simulation study was conducted, revealing that, in general, these methods produce good parameter estimates under different scenarios of censoring.

Keywords

Autoregressive models, censored data, data augmentation, Gibbs sampler.

Asymptotic distribution of low-dimensional patterns by regularizors with convex non-differentiable penalties

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

[1] *Lund University, Department of Statistics, Sweden,*
ivan.hejny@stat.lu.se
[2] *University of Wroclaw, Department of Mathematics, Poland*

Abstract

In this paper, we explore the asymptotic distribution of the patterns generated by regularizers with non-differentiable penalties. These patterns depend on the penalty through its subdifferential and can take various forms, such as the sign vector of regression coefficients for LASSO, or the more refined SLOPE pattern, which also identifies clusters of coefficients with the same absolute values. We focus on the classical asymptotics, where the sample size approaches infinity while the number of regressors remains fixed. We derive the asymptotic distribution of the $\sqrt{n}$ scaled estimation error and its pattern for a broad class of regularizers. To achieve the pattern convergence, we utilize the Hausdorff distance, which provides a suitable notion of convergence for the penalty subdifferentials. Our framework encompasses various regularizers, including Generalized LASSO, SLOPE or Elastic net. Importantly, it extends beyond ordinary least squares to the robust Huber and Quantile loss functions. For SLOPE, we also establish asymptotic control of the false discovery rate in the context of an asymptotically orthogonal design of the regressors. Additionaly, sampling from the asymptotic error distribution facilitates comparisons between different regularizers. We provide a short simulation study showcasing an illustrative comparison between the asymptotic properties of LASSO, fused LASSO and SLOPE.

Keywords

Regularization, pattern recovery, low-dimensional asymptotics, robust regression, SLOPE.

Spectral estimation for spatial point processes and random fields

Jake P. Grainger[1], **Tuomas A. Rajala**[2], **David J. Murrell**[3], **Sofia C. Olhede**[4]

[1] *EPFL, Institute of Mathematics, Switzerland, jake.grainger@epfl.ch*
[2] *Natural Resources Institute Finland, Finland, tuomas.rajala@luke.fi*
[3] *University College London, Evolution and Environment, Centre for Biodiversity and Environment Research, UK, d.murrell@ucl.ac.uk*
[4] *EPFL, Institute of Mathematics, Switzerland, sofia.olhede@epfl.ch*

Abstract

Spectral analysis has proven to be a powerful tool for understanding time series and random fields. However, the extension to other kinds of spatial processes is not straightforward. In this talk, we introduce a technique for the multivariate spectral analysis of mixtures of random fields and potentially marked point processes, based on multitapering. This enables us to estimate measures of dependence between a wide variety of complex multidimensional processes. Such dependence, both inter- and intra-process, are the subject of interest in a range of different applications in data science. For example, say we collect the location and size of individual tree species within a forest, alongside other variables such as soil chemistry or topology. We then may be interested in understanding the influence which different tree species and/or chemicals have on each other. One way to achieve such understanding is through spectral analysis, the practical and theoretical details of which we outline in this talk.

Keywords

Spatial point processes, Random fields, Multitapering, Spectra, Multivariate.

Causal-DRF: Conditional Kernel Treatment Effect using Distributional Random Forest

Jeffrey Näf[1], **Herb Susmann**[2], **Julie Josse**[3]

[1] *Inria, PreMeDICaL Team, University of Montpellier, France, jeffrey.naf@inria.fr*
[2] *CEREMADE (UMR 7534), Université Paris-Dauphine PSL, France, herbert.susmann@dauphine.psl.eu;*
[3] *Inria, PreMeDICaL Team, University of Montpellier, France, julie.josse@inria.fr*

Abstract

A frequent measure to assess the effectiveness of a treatment given some covariates is the conditional average treatment effect (CATE), the difference in conditional expectation of the counterfactuals. However, there is increased interest in the effects of treatments beyond the mean. Inspired by the causal-forest for CATE estimation, we develop a forest-based method to estimate the conditional kernel treatment effect (CKTE), based on the recently introduced Distributional Random Forest (DRF) algorithm. Adapting the splitting criterion of DRF, we show how one forest fit can be used to obtain a consistent estimator, as well as an approximation of its sampling distribution. This allows, in particular, to construct a conditional kernel-based test for distributional effects with provably valid type-I error. We illustrate the effectiveness of the new approach on several simulated and real data examples.

Keywords

Bootstrap, causality, conditional distributional treatment effect, conditional kernel treatment effect, distributional regression, ensemble methods, two-sample testing.

A Meta-Learning Method for Estimation of Causal Excursion Effects to Assess Time-Varying Moderation

Jieru Shi[1], **Walter Dempsey**[2]

[1] *University of Cambridge, Department of Pure Mathematics and Mathematical Statistics, UK, js2882@cam.ac.uk*
[2] *University of Michigan, Department of Biostatistics, USA, wdem@umich.edu*

Abstract

Twin revolutions in wearable technologies and smartphone-delivered health interventions have greatly increased the accessibility of mobile health (mHealth) interventions. Micro-randomized trials (MRTs) are designed to assess mHealth intervention effectiveness and introduce a novel class of causal estimands called "causal excursion effects." These estimands enable the evaluation of how intervention effects change over time and are influenced by individual characteristics or context. However, existing analysis methods for causal excursion effects require pre-specified features of the observed high-dimensional history to build a working model for a critical nuisance parameter. Machine learning appears ideal for automatic feature construction, but their naive application can lead to bias under model misspecification. To address this issue, this paper revisits the estimation of causal excursion effects from a meta-learner perspective, where the analyst remains agnostic to the supervised learning algorithms used to estimate nuisance parameters. We present the asymptotic properties of proposed estimators and compare them both theoretically and through extensive simulations. The results show relative efficiency gains and support the suggestion of a doubly robust alternative to existing methods. Finally, the proposed methods' practical utility is demonstrated by analyzing data from a multi-institution cohort of first-year medical residents in the United States (NeCamp et al., 2020).

Keywords

Debiased/Orthogonal Estimation, Machine Learning, Double Robustness, Causal Excursion Effect, Mobile Health, Time-Varying Treatment.

Exploring Encoder-Decoder Frameworks for Learning Latent Representations of High-Frequency Wearable Device Data

Howon Ryu[1], **Jingjing Zou**[2]

[1] *University of California, San Diego, horyu@ucsd.edu*
[2] *University of California, San Diego, j2zou@ucsd.edu*

Abstract

The use of wearable devices has become increasingly prevalent in recent years, enabling the collection of high-frequency data that provides valuable insights into individuals' health behavior patterns. With the rapid advancement of machine learning techniques, there has been a growing interest in leveraging these methods to extract meaningful latent representations from wearable device data. This study focuses on exploring the application of state-of-art encoder-decoder frameworks in learning latent representations of high-frequency wearable device data, with a specific emphasis on tracker-measured physical activity, and evaluates their performance in capturing key features and patterns as well as in predicting posture types. The findings contribute to advancing personalized health monitoring and behavior analysis using wearable device data.

Keywords

Wearable device data, physical activity, decoder-encoder, deep learning.

Cotinine: Exploring the Impact of Smoking Habits on Periodontal Disease

João Onofre[1], **Luzia Mendes**[1], **Pereira J.A.**[1,2]

[1] *Faculty of Dental Medicine of University of Porto, Portugal, up201907471@edu.fmd.up.pt*
[2] *Centro de Estatística Aplicações da Universidade de Lisboal, Portugal*

Abstract

Background: On July 31, 2023, the World Health Organization reported on the global tobacco epidemic, highlighting that smoking leads to over 8 million deaths annually. Of the world's 1.3 billion smokers, 80% live in low- and middle-income countries. Nicotine, a primary constituent of tobacco, causes addiction due to its psychoactive properties. While nicotine is short-lived in the bloodstream, its main metabolite, cotinine, remains longer, serving as a reliable marker for nicotine exposure.
Objective: The study aimed to determine if the depth of an individual's periodontal pockets is associated with serum cotinine levels, age of smokers, gender, and duration of smoking.
Materials and Methods: Data from the National Health and Nutrition Examination Survey was analyzed using GMLSS models within the R software. The variables considered in this study were probing pocket depth, age, cotinine levels, gender, and years of smoking.
Results: The best-fitted model suggests that larger probing depths are significantly associated ($p < 0.05$) with age ($\beta = 0.050$), cotinine levels ($\beta = 0.001$), and gender (female) ($\beta = -0.319$), with females being associated with a healthier periodontium.
Conclusions: Higher cotinine levels are significantly associated with increased periodontal destruction. These findings contrast with results from previous studies that did not find such association.

Martingale Testing with the Smoothed Bicausal Wasserstein Distance

Jose Blanchet[1], **Johannes Wiesel**[2], **Erica Zhang**[3], **Zhenyuan Zhang**[4]

[1] *Department of Management Science and Engineering Stanford University, jose.blanchet@stanford.edu*
[2] *Department of Mathematics, Carnegie Mellon University, wiesel@cmu.edu*
[3] *Department of Statistics, Columbia University, yz4232@columbia.edu*
[4] *Department of Mathematics, Stanford University, zzy@stanford.edu*

Abstract

Martingales play a key role in machine learning, in particular in the non-parametric estimation of regression functions. Martingales also take center stage in the theory of stochastic calculus and mathematical finance. Motivated by these applications we focus on the problem of projecting a sample of vectors to the space of martingales using the bicausal/nested Wasserstein distance. If the data-generating process forms a martingale, we characterise the asymptotic distribution of this projection and give finite sample guarantees. This can be used to build non-parametric tests for the martingale property. If the data-generating process fails to form a martingale, the projection can also be used to gain insight into exactly how the data-generating distribution differs from a martingale by linking a certain subspace with the growth of the normalized projection statistic. This can be used to improve the power of a martingale test as a function of design parameters.

Keywords

Optimal transport, (nested) Wasserstein distance, Martingale test.

Bayesian image analysis in Fourier space for neuroimaging

John Kornak[1], **Karl Young**[1], **Eric Friedman**[2], **Konstantinos Bakas**[3], **Hernando Ombao**[3]

[1] *University of California, San Francisco, USA, john.kornak@ucsf.edu*
[2] *International Computer Science Institute, Berkeley, USA*
[3] *King Abdullah Institute of Science and Technology, Saudi Arabia*

Abstract

For more than 30 years now, Bayesian image analysis has been a leading approach to image reconstruction and enhancement. The idea of the approach is to balance a priori expectations of image characteristics (the prior) with a model for the image degradation process (the likelihood). The conventional Bayesian modeling approach as defined in image space, implements priors that describe inter-dependence between spatial locations on the image lattice (commonly through Markov random field, MRF, models) and can therefore be difficult to model and compute. Bayesian image analysis in Fourier space (BIFS) provides for an alternate approach that can generate a wide range of models, including ones with similar properties to conventional models, but with reduced computational burden; the originally complex high-dimensional estimation problem in image space can be similarly modeled as a series of (trivially parallelizable) independent one-dimensional problems in Fourier space. We will examine development of different prior models in Fourier space and illustrate with neuroimaging applications of BIFS.

Keywords

Bayesian image analysis, Fourier-space, Image priors, Markov random fields, Statistical image analysis.

Combining Stochastic Tendency and Distribution Overlap Towards Improved Nonparametric Inference for K-Samples

Jonas Beck[1], **Patrick B. Langthaler**[2], **Arne C. Bathke**[3]

[1] *Paris Lodron University of Salzburg, Department of Artificial Intelligence and Human Interfaces, Austria, jonas.beck@plus.ac.at*
[2] *Paris Lodron University of Salzburg, Department of Artificial Intelligence and Human Interfaces, Austria, patrickbenjamin.langthaler@stud.plus.ac.at*
[3] *Paris Lodron University of Salzburg, Department of Artificial Intelligence and Human Interfaces, Austria, arne.bathke@plus.ac.at*

Abstract

Statistical functionals are omnipresent in the nonparametric comparison of k independent samples. One of the most common ones is the nonparametric relative effect, e.g. used for the Kruskal-Wallis Test. The main drawback of using this functional is its inability of capturing scale differences. Therefore we incorporate an extension of the so-called distribution overlap index, which was introduced in the two-sample case to quantify the overlap of ecological niches for different species. We derive the joint asymptotic distribution of the respective estimators of both functionals and construct confidence regions. Extending the Kruskal-Wallis test, we introduce a new test based on the joint use of these functionals. As in an all pairwise comparison in some cases intransitivity may occur, we use a mean distribution function as a reference distribution.

Keywords

Nonparametric Statistics, Pseudoranks, Rank Statistics, Resampling.

Selecting Prior Information for Generalized Maximum Entropy Estimation

Jorge Cabral[1], **Pedro Macedo**[2], **Vera Afreixo**[3]

[1] *Center for Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal, jorgecabral@ua.pt*
[2] *CIDMA, University of Aveiro, Portugal, pmacedo@ua.pt*
[3] *CIDMA, University of Aveiro, Portugal, vera@ua.pt*

Abstract

The most widely used technique to describe and predict the effect of k independent variables of N independent subjects, defining a $(N \times k)$ matrix X, on a dependent variable y: $y = X\beta + e$, where $\beta = (\beta_1, \beta_2, \ldots, \beta_k)'$ is a vector of coefficients and $e = (e_1, e_2, \ldots, e_N)'$ is the error. The most common coefficient estimation approach is the ordinary least squares (OLS). Although unbiased it lacks stability under collinearity scenarios, whichcould be improved by the Ridge regression. The presence of extreme y_i values can affect estimation and the M estimation is one possible approach. The Generalized Maximum Entropy (GME) formalism also provides a possible approach for solving problems with poor specification and data that are partial or incomplete. Its implementation starts by choosing sets of discrete points (support spaces) based on a priori information of the coefficients which is in many times unknown. We propose choosing the range of the support spaces based on a prior GME estimation of the coefficients on standardized data which will serve as reference for the extremes of a zero centred symmetric support space, specific for each coefficient. We then re-estimate the coefficients t times for support spaces symmetrically M times divided around zero with decreasing ranges $2 \times |\hat{\beta}_k| \times r_i$, $r_i > 1$, $i > j \Rightarrow r_j < r_i$, $1 \leq i \leq t$ and $1 \leq j \leq t$. Finally, we do a selection of supports by determining the k-fold cross-validation root mean square error (CV-RMSE) for each t iteration keeping the one

that produces the lowest CV-RMSE. We implemented the algorithm in *R*, provided a Shiny web-application and compared the performance of our approach in a simulation study. Our proposal generally returned the lowest median 5-fold CV-RMSE. As k increased, the ratio between the median 5-fold CV-RMSE for OLS or M estimation and GME also increased. As the standard deviation of the error increases the range of the support spaces decreases.

Keywords

Generalized maximum entropy, prior information, support spaces.

Acknowledgements

The authors are supported by the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT), project UIDB/04106/2020. Cabral is grateful to the PhD fellowship at CIDMA-DMat-UA, reference UIDP/04106/2020.

References

Golan, A.; Judge, G.G. and Miller, D. (1996). Maximum entropy econometrics: robust estimation with limited data, Wiley, Chichester [England], New York.

Assessing Dental Symmetry: Introduction of the Symmetry Measure Score (SMS) in Periodontal Disease Analysis

Pereira J. A.[1,4,5], **Anuj Mubayi**[2], **Davide Carvalho**[3], **Teresa A. Oliveira**[1,4]

[1] *Universidade Aberta, Portugal; up241045@up.pt*
[2] *Illinois State University, USA*
[3] *Faculty of Medicine Universidade of Porto, Portugal*
[4] *Center of Statistics and Applications of Universidade de Lisboa*
[5] *Faculty of Dental Medicine of University of Porto, Portugal*

Abstract

Assuming symmetry in the human mouth in terms of shape and clinical signs, quantifying the symmetry of periodontal disease is relevant in two distinct contexts: epidemiological and clinical. From the epidemiological standpoint, symmetry quantification can enhance the estimation of population disease parameters, particularly in studies employing half-mouth evaluations. The clinical significance of symmetry is highlighted by the idea that asymmetrical values of periodontal disease indicators might be affected by asymmetrical factors influencing both the disease's onset and progression. Building on Zadeh [1] concept of fuzzy symmetry, we introduced a Symmetry Measure Score (SMS) to categorize the symmetry grade of two contralateral observations on periodontal disease indicators. This categorization considers both the differential and their mean value. The SMS function embodies the characteristics of a membership function within the fuzzy symmetry framework and offers insights into symmetry, relevant to both epidemiology and clinical periodontology. We tested the function using data from NHANES 2011-2012. The SMS values fluctuated based on the type of tooth, with values ranging from 0.8 to 1, indicating a high symmetry grade. We evaluated the results' validity in two ways: using GAMLSS models and graphically. The GAMLSS

models helped determine the magnitude and statistical significance of the side location effect size. Graphically, we compared the empirical probability density function of contralateral variables. Both sets of results aligned with the SMS findings. The SMS function proved to be a dependable measure of symmetry, consistent with the methods we employed for its testing.

Keywords

Symmetry, Fuzzy Symmetry, Symmetry Measure Score, Periodontal Disease.

References

Zadeh, L. A. (1965). Fuzzy sets. *Information and control.* 8(3), 338-353.

An introduction to (and an application of) the random projection method

Juan A. Cuesta-Albertos

Universidad de Cantabria, Department of Matemáticas, Estadística y Computación, Spain, cuestaj@unican.es

Abstract

In the first part of the talk, I will describe the reasons that make it possible to test p-dimensional hypotheses by replacing the hypothesis to be tested by its "projected" counterpart on just one randomly chosen 1-dimensional subspace. Obviously, this procedure implies a loss of power. I will comment on the two proposed ways to alleviate this problem: 1) Use several 1-dimensional projections. 2) If the dimension is low, it is also possible to integrate over all possible projections. In the second part of the talk, I will comment on the use of the latter idea to obtain a family of uniformity tests on the p-dimensional sphere. This family includes some well-known uniformity tests, but it also allows to extend some circular tests to higher dimensions as well as to introduce some new ones. Results in the second part have been obtained in cooperation with E. García-Portugués (U. Carlos III, Spain) and P. Navarro-Esteban (U. de Cantabria, Spain).

Keywords

Random projections, Goodness of fit tests, Uniformity, p-dimensional sphere.

Sharp global convergence guarantees for iterative nonconvex optimization with random data

Kabir Aladin Verchand[1], **Ashwin Pananjady**[2], **Christos Thrampoulidis**[3]

[1] *University of Cambridge, Statistical Laboratory, UK, kav29@cam.ac.uk*
[2] *Georgia Institute of Technology, Schools of Industrial & Systems Engineering and Electrical & Computer Engineering, USA, ashwinpm@gatech.edu*
[3] *University of British Columbia, Department of Electrical & Computer Engineering, Canada, cthrampo@ece.ubc.ca*

Abstract

We consider a general class of regression models with normally distributed covariates, and the associated nonconvex problem of fitting these models from data. We develop a general recipe for analyzing the convergence of iterative algorithms for this task from a random initialization. In particular, provided each iteration can be written as the solution to a convex optimization problem satisfying some natural conditions, we leverage Gaussian comparison theorems to derive a deterministic sequence that provides sharp upper and lower bounds on the error of the algorithm with sample splitting. Crucially, this deterministic sequence accurately captures both the convergence rate of the algorithm and the eventual error floor in the finite-sample regime, and is distinct from the commonly used "population" sequence that results from taking the infinite-sample limit. We apply our general framework to derive several concrete consequences for parameter estimation in popular statistical models including phase retrieval and mixtures of regressions. Provided the sample size scales near linearly in the dimension, we show sharp global convergence rates for both higher-order algorithms based on alternating updates and first-order

algorithms based on subgradient descent. These corollaries, in turn, reveal multiple nonstandard phenomena that are then corroborated by numerical experiments.

Keywords

Nonconvex optimization, convergence rate, precise iterate-by-iterate prediction.

A general framework for causal learning algorithms

Kai Teh[1], **Kayvan Sadeghi**[2], **Terry Soo**[3]

University College London, Department of Statistical Science, UK,[1]*kai.teh.21@ucl.ac.uk,*[2]*k.sadeghi@ucl.ac.uk,*[3]*t.soo@ucl.ac.uk*

Abstract

We present a general framework for constructing all exact causal learning algorithms, along with corresponding sufficient conditions for consistency. By appropriate substitution, our framework includes previous work in exact causal learning including conventional SGS, sparsest permutation algorithm, and more recent ones such as natural structure learning using ordered stabilities. We then apply the framework to obtain a relaxed version of the natural structure learning algorithms, along with sufficient and necessary conditions for consistency. We then present the construction of such an algorithm, which we call (Me)-LoNS, and compare the conditions for consistency of our approach to some existing causal learning approaches.

Keywords

Causal learning, graphical modelling, causal order.

References

K. Sadeghi, and T. Soo (2022). Conditions and Assumptions for Constraint-based Causal Structure Learning. *Journal of Machine Learning Research*

G. Raskutti, and C. Uhler (2012). Learning directed acyclic graphs based on sparsest permutations. *The ISI's Journal for the Rapid Dissemination of Statistics Research*

The Completion of Covariance Kernels

Kartik G. Waghmare[1], **Victor M. Panaretos**[2]

[1] *Ecole Polytechnique Fédérale de Lausanne (EPFL), Institute of Mathematics, Switzerland, kartik.waghmare@epfl.ch*
[2] *Ecole Polytechnique Fédérale de Lausanne (EPFL), Institute of Mathematics, Switzerland, victor.panaretos@epfl.ch*

Abstract

We consider the problem of positive-semidefinite continuation: extending a partially specified covariance kernel from a subdomain Ω of a rectangular domain $I \times I$ to a covariance kernel on the entire domain $I \times I$. For a broad class of domains Ω called *serrated domains*, we are able to present a complete theory. Namely, we demonstrate that a canonical completion always exists and can be explicitly constructed. We characterise all possible completions as suitable perturbations of the canonical completion, and determine necessary and sufficient conditions for a unique completion to exist. We interpret the canonical completion via the graphical model structure it induces on the associated Gaussian process. Furthermore, we show how the estimation of the canonical completion reduces to the solution of a system of linear statistical inverse problems in the space of Hilbert-Schmidt operators, and derive rates of convergence. We conclude by providing extensions of our theory to more general forms of domains, and by demonstrating how our results can be used to construct covariance estimators from sample path fragments of the associated stochastic process. Our results are illustrated numerically by way of a simulation study and a real example.

Keywords

Positive-definite continuation, functional data analysis, graphical model, identifiability, inverse problem, fragments.

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

Kasper Bågmark[1], **Adam Andersson**[2], **Stig Larsson**[3]

[1] *Chalmers University of Technology and University of Gothenburg, Mathematical Sciences, Sweden, bagmark@chalmers.se*
[2] *Chalmers University of Technology and University of Gothenburg, Mathematical Sciences, & Saab AB Radar Solutions, Gothenburg, Sweden, adam.andersson@chalmers.se*
[3] *Chalmers University of Technology and University of Gothenburg, Mathematical Sciences, Sweden, stig@chalmers.se*

Abstract

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.

Keywords

Filtering problem, Zakai equation, stochastic partial differential equation, deep learning, energy-based method.

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. (2023). An energy-based deep splitting method for the nonlinear filtering problem. *Partial Differ. Equ. Appl.* 4, 14 (2023)

[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

A scalable clustering algorithm to approximate graph cuts

L. Suchan[1], **H. Li**[2,4], **A. Munk**[3,4]

[1] *Institute of Mathematical Stochastics, Georg August University of Göttingen, Germany, leo.suchan@uni-goettingen.de*
[2] *Institute of Mathematical Stochastics, Georg August University of Göttingen, Germany, housen.li@mathematik.uni-goettingen.de*
[3] *Institute of Mathematical Stochastics, Georg August University of Göttingen, Germany, munk@math.uni-goettingen.de*
[4] *Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC)*

Abstract

Graph cuts such as e.g. Ratio, Normalized or Cheeger Cut realize the fundamental principle of clustering, that is to separate data points in different groups according to their similarities. However, due to their inherent NP-hardness to compute, certain forms of convex relaxation are often exploited in data analysis, which unfortunately sacrifice the underlying combinatorial geometry of the data. As a remedy, we propose the *Xist* algorithm. It minimizes a *combinatorial surrogate* of the graph cut optimization which preserves the intrinsic geometry of the original cuts while allowing for efficient computation (linear runtime in the number of vertices and quadratic in the number of edges), even for large-scale datasets. The Xist algorithm is built on the selection strategy of vertices that restricts the optimization to tightly connected clusters, the efficient computability of *st*-MinCuts, and the intrinsic properties of Gomory-Hu trees. The performance of Xist is analyzed in theory and investigated empirically on various datasets. In particular, Xist is capable of identifying clusters at various scales, and it reports better cut values compared to convex relations such as spectral clustering and related algorithms.

Keywords

Graph partition, classification, clustering, image segmentation.

Refined methods for trial sequential analyses for living systematic reviews

Yipeng Wang[1], **Lifeng Lin**[2]

[1] *Department of Biostatistics, University of Florida, USA, yipeng.wang1@ufl.edu*
[2] *Department of Epidemiology and Biostatistics, University of Arizona, USA, lifenglin@arizona.edu*

Abstract

A living systematic review (LSR) is an evolving approach that aims to provide continuous updates and real-time synthesis of evidence. Unlike traditional systematic reviews conducted at a specific time, LSRs incorporate new studies as they become available. Compared to the study collection and qualitative assessment in LSRs, few works are devoted to developing quantitative methods. The trial sequential analysis (TSA) is a well-known procedure to assess the adequacy of the available evidence based on the collected studies in an LSR. It uses trial sequential monitoring boundaries for assessing the efficacy of an intervention and futility boundaries for evaluating whether the intervention does not differ significantly from the control. Although TSAs have recently gained popularity, existing TSA methods have limitations stemming from their heavy reliance on interim analyses of randomized controlled trials, where individuals are often more homogeneous than those in meta-analyses. For random-effects meta-analyses, the normality-based methods can perform poorly when the number of studies is small. In such cases, the Hartung–Knapp–Sidik–Jonkman (HKSJ) method based on the t-statistic is more appropriate. In this talk, we introduce novel trial sequential methods based on the t-statistic of the cumulative meta-analysis. The proposed methods can avoid LSRs being terminated prematurely, allowing for more robust evidence syntheses. Numerical studies show that the proposed methods are more reliable than the existing methods.

Keywords

Meta-analysis, living systematic review, trial sequential analysis.

Optimal Transport for Single-Cell Heterogeneous Data Analysis

Lin Wan[1]

[1] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, lwan@amss.ac.cn*

Abstract

Advances in single-cell technologies enable comprehensive studies of heterogeneous cell populations that make up tissues, the dynamics of developmental processes, and the underlying regulatory mechanisms that control cellular functions. The computational integration of single-cell datasets is drawing heavy attention toward making advancements in machine learning and data science. Optimal transport (OT) is a powerful tool in the analysis of complex data, as it learns an optimal cost-effective mapping between data distributions. In this talk, I will report our recent work on developing OT-based data analysis methods for single-cell multi-omics integration and dynamic inference of time series single-cell data.

Keywords

Optimal transport, dynamic inference, Gromov-Wasserstein distance, time series single-cell data, nonlinear Fokker-Planck equation, multi-omics. integration

Bayesian inference for aggregated Hawkes processes

Lingxiao Zhou[1] **and Georgia Papadogeorgou**[2]**.**

[1] *University of Florida, Department of Statistics, USA,*
zhou.l@ufl.edu
[2] *University of Florida, Department of Statistics, USA,*
gpapadogeorgou@ufl.edu

Abstract

The Hawkes process, a self-exciting point process, has a wide range of applications in modeling earthquakes, social networks and stock markets. The established estimation process requires that researchers have access to the exact time stamps and spatial information. However, available data are often rounded or aggregated. We develop a Bayesian estimation procedure for the parameters of a Hawkes process based on aggregated data. Our approach is developed for temporal, spatio-temporal, and mutually exciting Hawkes processes where data are available over discrete time periods and regions. We show theoretically that the parameters of the Hawkes process under different specifications are identifiable from aggregated data, and demonstrate the method on simulated temporal and spatio-temporal data with various model specifications in the presence of one or more interacting processes, and under varying coarseness of data aggregation. Finally, we analyze spatio-temporal point pattern data of insurgent attacks in Iraq from October to December 2006, and we find coherent results across different time and space aggregations.

Keywords

Hawkes process, insurgent violence, mutually-exciting, point process, spatio-temporal data.

An extreme value support measure machine for group anomaly detection

Lixuan An[1], **Bernard De Baets**[2], **Stijn Luca**[3]

[1] *Ghent University, Department of Data Analysis and Mathematical Modelling, Belgium, lixuan.an@ugent.be*
[2] *Ghent University, Department of Data Analysis and Mathematical Modelling, Belgium, bernard.debaets@ugent.be*
[3] *Ghent University, Department of Data Analysis and Mathematical Modelling, Belgium, stijn.luca@ugent.be*

Abstract

Group anomaly detection is a subfield of pattern recognition that aims at detecting anomalous groups, or anomalous realizations of point process patterns rather than individual anomalous points. Existing approaches, however, are mainly focusing on modelling unusual aggregates of points in the bulk of the data, *i.e.*, in high-density regions. In this way, unusual group behaviour with a number of points located in low-density regions is not fully detected. In this study, we introduce a probabilistic model for group anomaly detection, which comprehensively detects group anomalous behaviour at both the point level and the distribution level. At the point level, we prove an analytical result where extreme value theory, a field of statistics that is well suited to model tails of distributions, is combined with point process models for generating a probabilistic point-based anomaly score. This effectively addresses the multiple hypothesis testing issue that often arises in conventional group anomaly detection methods. At the distribution level, we extend a stable discriminative model one-class support measure machine with a sigmoid probabilistic calibration technique to define a probabilistic distribution-based anomaly score. We employ a well-known uninorm aggregation function to aggregate the point-based and distribution-based anomaly scores, leading to an overall score that is more sensitive and achieves a higher accuracy compared to existing

group anomaly detection models across synthetic as well as real-world datasets.

Keywords

Group anomaly detection, Extreme value theory, Point processes, One-class support measure machine, Uninorm.

Randomization-based Inference in Nonparametric Repeated Measure Models with Missing Data

Lubna Amro[1], **Dennis Dobler**[2], **Jörg-Tobias Kuhn**[3]

[1] *Department of Statistics, Technical University of Dortmund, Germany, lubna.amro@tu-dortmund.de*
[2] *Department of Statistics, Technical University of Dortmund, Germany, dennis.dobler@tu-dortmund.de*
[3] *Department of Methods in Empirical Educational Research, Technical University of Dortmund, Germany, tobias.kuhn@tu-dortmund.de*

Abstract

Relative effects enjoy great popularity across various field of research. Also in statistical methodology research, extensions of this method have been developed in many different directions. In this talk, we will focus on repeated measures designs with randomly missing data. Here, relative effects can be used to find a time or other effects on the outcomes. In a previous SMMR-paper by Rubarth, Pauly, and Konietschke (2022), relevant theory has been developed for tests based on quadratic forms in this context. They developed Wald- and ANOVA-type tests that are based on approximations using estimated chi-squared and F-distributions. In this talk, we re-visit that testing problem by means of a randomization procedure which will give rise to asymptotically exact inference procedures. Simulations demonstrate the small sample performance and a real data analysis illustrates several aspects of our method.

Keywords

Missing Values, Repeated measures, Randomization procedure, Rank tests.

Deviance Matrix Factorization

Luis Carvalho[1], **Liang Wang**[2]

[1] *Boston University, Dept of Mathematics and Statistics, USA, lecarval@bu.edu*
[2] *Boston University, Dept of Mathematics and Statistics, USA, leonwang@bu.edu*

Abstract

The singular value decomposition can be used to find a low-rank representation of a matrix under the Frobenius norm (entrywise square-error loss) and, for this reason, it enjoys an ubiquitous presence in many areas, including in Statistics with principal component and factor analyses. In this talk, we discuss a generalization of this matrix factorization, the deviance matrix factorization (DMF), that assumes broader deviance losses and thus allows for more meaningful and representative decompositions under different data domains and variance assumptions. We provide an efficient algorithm for the DMF and discuss using entrywise weights to represent missing data. We propose two tests to identify suitable decomposition ranks and data distributions and prove a few theoretical guarantees such as consistency. To showcase the practical performance of the proposed decomposition, we present a number of case studies in genetics, network analysis, and image classification. Finally, we offer a few directions for future work.

Keywords

Non-negative matrix factorization, factor models, principal component analysis.

On the use of graph theory and machine learning algorithms in anti-money laudering systems

Mafalda Sá Ferreira[1], **Regina Bispo**[1,2]

[1] *NOVA School of Science and Technology, NOVA University of Lisbon, Mathematics, Portugal, msm.ferreira@campus.fct.unl.pt*
[2] *Center for Mathematics and Applications (NOVA Math) and Department of Mathematics, NOVA School of Science and Technology, NOVA University of Lisbon, Mathematics, Portugal, r.bispo@fct.unl.pt*

Abstract

Anti-money laundering consists of a set of actions aiming at preventing the movement of illegally obtained money through the financial system. The procedures adopted by financial institutions to detect suspicious activities are typically rule-based, resulting in high false positive rates. To improve the effectiveness of these systems, we propose a graph-based approach that incorporates the transactional relationships between clients. New features are computed via random walks and used as input in machine learning methods. To help analyze the graph information in a financial context, we aim at building an R package with synthetic data sets that characterizes real relationships and behaviour between clients. We describe all the concepts and methods needed for this study and detail the content of the built R package. We also describe and provide an exploratory analysis of the data sets used in the present work. We show that the new computed features are the most important for the tested models and we can detect up to 98% of the report transactions, confirming that our approach is adequate to this problem.

Keywords

Anti-money laundering, graph theory, machine learning.

missKnockoff: Controlled Variable Selection with Missing Values

D. Nowakowski[1], **J. Josse**[2], **S. Majewski**[3], **A. Weinstein**[4], **M. Bogdan**[5]

[1] *Medical University of Białystok, Department of Biostatistics and Medical Informatics, Poland, dominik.nowakowski@umb.edu.pl*
[2] *Inria-Inserm Centre in Montpellier, PreMeDICaL team, France, julie.josse@inria.fr*
[3] *University of Warsaw, Faculty of Mathematics, Informatics and Mechanics, Poland, sjm.majewski@gmail.com*
[4] *Hebrew University of Jerusalem, Department of Statistics and Data Science, Israel, asaf.weinstein@mail.huji.ac.il*
[5] *Lund University, Department of Statistics, Sweden and University of Wroclaw, Department of Mathematics, Poland, malgorzata.bogdan@stat.lu.se*

Abstract

Model selection with high-dimensional data has captured considerable attention in the past two decades. While plenty of methods have been proposed and analyzed, few are suitable for handling missing data. We propose *missKnockoffs*, a platform for controlled variable selection that extends Model-X knockoffs to datasets with missing values. Our method first uses one of the well known strategies for imputing the missing values, and proceeds with performing model-X knockoffs on the imputed data set to achieve FDR control. In order to account for the internal randomness, multiple imputation is easily incorporated by generating several knockoff copies. We discuss different ways to aggregate the results and propose a novel solution, which we support by a theoretical bound on the expected value of the corresponding FDR estimator. We study the performance of various aggregation schemes in terms of power and FDR through extensive simulations, demonstrating superior performance of our proposed aggregation scheme. We also

demonstrate the usefulness of missKnockoffs in practice by analyzing a real gene expression data set.

Keywords

Incomplete data, FDR control, multiple knockoffs, multiple imputation.

Functional regression models with functional response

Manuel Oviedo-de la Fuente[1], **Manuel Febrero-Bande**[2], **Morteza Amini**[3], **Mohammad Darbalaei**[2,3]

[1] *Universidade da Coruña, CITIC, Spain, manuel.oviedo@udc.es*
[2] *Universidade de Santiago de Compostela, CITIC, Spain*
[3] *University of Tehran, Iran*

Abstract

This paper proposes three new approaches for additive functional regression models with functional responses. The first one is a reformulation of the linear regression model, and the last two are on the yet scarce case of additive nonlinear functional regression models. One of the nonlinear models is based on constructing a Additive Model where the representation of the covariates in a $\mathcal{L}_2$ basis is restricted (by construction) to Hilbertian spaces. The other one extends the kernel estimator, and it can be applied to general metric spaces since it is only based on distances. We include our new approaches as well as real data-sets in an R package. The performances of the new proposals are compared with previous ones ([2], [3]), which are reviewed from the theoretical and practical point of view. The simulation results show the advantages of the nonlinear proposals and the small loss of efficiency when the simulation scenario is truly linear. Finally, a visualization tool is also provided for checking the linearity of the relationship between a single covariate and the response. Simulation codes, packages and real datasets are included in a GitHub repository ([1]).

Keywords

Functional regression, Functional response, Nonlinear models.

Acknowledgements

The research by Manuel Febrero-Bande and Manuel Oviedo-de la Fuente has been partially supported by the Spanish Grant PID2020-116587GB-I00 and PID2020-113578RB-I00 respectively. The research

by Mohammad Darbalaei and Morteza Amini is based upon research funded by Iran National Science Foundation, project No. 99014748.

References

[**1**] M. Febrero-Bande, M. Oviedo-de la Fuente. *FRMFR GitHub repository* (2023). `https://github.com/moviedo5/FRMFR/`

[**2**] J. Goldsmith, F. Scheipl, L. Huang, J. Wrobel, Ch. Di, J. Gellar, J. Harezlak, M.W. McLean, B. Swihart, L. Xiao, C.M. Crainiceanu, Ph. Reiss. *refund: regression with functional data (2023).* R package version 0.1-30. `https://cran.r-project.org/web/packages/refund/index.html`

[**3**] L. Ruiyan, X. Qi. *FRegSigCom: functional regression using signal compression approach (2019).* R package version 0.3.0. `https://cran.r-project.org/web/packages/FRegSigCom/index.html`

Model-Free Conditional Conformal Depth Measures Algorithm for Uncertainty Quantification in Complex Functional Regression Models

Marcos Matabuena[1], Rahul Ghosal[2], Pavlo Mozharovskyi[3], Oscar Hernán Padilla[4], Jukka-Pekka Onnela[1]

[1] *Harvard University, mmatabuena@hsph.harvard.edu*
[2] *University of South Carolina*
[3] *LTCI, Telecom Paris, Institut Polytechnique de Paris*[4] *UCLA*

Abstract

Depth measures have gained popularity in the statistical literature for defining level sets in the context of multivariate and more complex data structures such as functional data objects and graphs. However, their application in regression modelling for providing prediction regions is currently limited. We propose a novel conditional depth measure based on conditional kernel mean embeddings to address this research gap. The new measure has the potential to introduce prediction regions in regression models for complex statistical responses and predictors that take values in separable Hilbert spaces. To enhance the practicality of our approach, we incorporate a conformal inference algorithm into the conditional depth measure. Our algorithm has the potential to offer non-asymptotic guarantees for constructing prediction regions. Moreover, we introduce conditional and unconditional consistency results for the derived prediction regions. In order to evaluate the performance of our approach across different scenarios with finite samples, we conducted an extensive simulation study. This study encompassed various response types, including Euclidean as well as complex statistical data types such as graphs and probability distributions. Through these simulations, we demonstrate the versatility and robustness of our method on finite samples.

Keywords

Conformal prediction; Depth functions; Functional data analysis; Regression models.

A goodness-of-fit test for the latency in a mixture cure model with covariates

Wenceslao González-Manteiga[1], **María Dolores Martínez-Miranda**[2], **Ingrid Van Keilegom**[3]

[1] *University of Santiago de Compostela, Spain, wenceslao.gonzalez@usc.es*
[2] *University of Granada, Spain, mmiranda@ugr.es*
[3] *KU Leuven, Belgium, ingrid.vankeilegom@kuleuven.be*

Abstract

In this talk we present a general goodness-of-fit test for the latency in a mixture cure model. In the presence of right censoring and a cure fraction a formal test is constructed to check the validity of three common models for the latency: a fully parametric model, a semiparametric Cox model and an accelerated failure time model. Two test statistics, the Cramér-von Misses and the Kolmogorov-Smirnov distances, are proposed. Both cases involve estimation under the null and the alternative hypotheses. Under the alternative we assume and compute a nonparametric estimator for the latency. To calibrate the test in practice it is suggested a Bootstrap method.

Keywords

Mixture cure model, Goodness-of-fit, semiparametric, bootstrap.

Inverse problem for parameters identification in a modified SIRD epidemic model using ensemble neural networks

Marian Petrica[1,2], **Ionel Popescu**[2,3]

[1] *Gheorghe Mihoc - Caius Iacob Institute of Mathematical Statistics and Applied Mathematics*
[2] *Faculty of Mathematics and Computer Science, University of Bucharest*
[3] *Institute of Mathematics of the Romanian Academy*

Abstract

In this talk, we present a parameter identification methodology of the SIRD model, an extension of the classical SIR model, which considers the deceased as a separate category. In addition, our model includes one parameter which is the ratio between the real total number of infected and the number of infected that were documented in the official statistics. Due to many factors, like governmental decisions, several variants circulating, opening and closing of schools, the typical assumption that the parameters of the model stay constant for long periods of time is not realistic. Thus, our objective is to create a method which works for short periods of time. In this scope, we approach the estimation relying on the previous 7 days of data and then use the identified parameters to make predictions. To perform the estimation of the parameters, we propose the average of an ensemble of neural networks. Each neural network is constructed based on a database built by solving the SIRD for 7 days, with random parameters. In this way, the networks learn the parameters from the solution of the SIRD model. Lastly we use the ensemble to get estimates of the parameters from the real data of Covid19 in Romania and then we illustrate the predictions for different periods of time, from 10 up to 45 days, for the number of deaths. The main goal was to apply this approach on the analysis of COVID-19 evolution in Romania, but

this was also exemplified in other countries like Hungary, the Czech Republic and Poland with similar results. The results are backed by a theorem which guarantees that we can recover the parameters of the model from the reported data. We believe this methodology can be used as a general tool for dealing with short term predictions of infectious diseases or in other compartmental models.

Keywords

SIRD Model, Neural Networks, Compartmental models, Data Science, Parameter identifiability.

A latent causal inference framework for ordinal variables

M. Scauda[1,2], **J. Kuipers**[3], **G. Moffa**[1,4]

[1] *University of Basel, Dep. of Mathematics and Computer Science, Switzerland*
[2] *University of Cambridge, Statistical Laboratory, UK, ms2985@cam.ac.uk*
[3] *ETH Zurich, Dep. of Biosystems Science and Engineering, Switzerland, jack.kuipers@bsse.ethz.ch*
[4] *University College London, Division of Psychiatry, UK, giusi.moffa@unibas.ch*

Abstract

Ordinal data, including Likert scales, economic status, and education levels are commonly encountered in applied research. Yet, existing causal methods often fail to account for the inherent order among categories, as they are primarily developed either for nominal data or for continuous data where relative magnitudes are well-defined. Hence, there is a pressing need for an order-preserving methodology to compute interventional effects between ordinal variables. Presuming a latent Gaussian Directed Acyclic Graph (DAG) model as data-generating mechanism provides one possible solution [1]. Precisely, the model assumes that ordinal variables originate from marginally discretizing at given thresholds a set of Gaussian variables, whose latent covariance matrix is constrained to satisfy the conditional independencies inherent in a DAG. Conditionally on a given latent covariance matrix and thresholds, this model leads to a closed-form function for ordinal causal effects in terms of interventional distributions in the latent space. For binary variables, this approach reduces to classical methods for causal effect estimation. When the underlying DAG is unknown, one can use the Ordinal Structural EM (OSEM) algorithm [2] to learn both a plausible latent DAG, up to an equivalence class, and

the model's parameters from observational data. Simulations demonstrate the performance of the proposed approach in estimating ordinal causal effects both for known and unknown structures of the latent graph. As an illustration of a real world use case, the method is applied to survey data of 408 patients from a study [3] on the functional relationships between symptoms of obsessive-compulsive disorder and depression.

Keywords

Causal inference, Causal diagrams, Latent graphical models, Directed acyclic graph-probit, Ordinal data.

References

[1] Silva, R., Ghahramani, Z. (2009). The Hidden Life of Latent Variables: Bayesian Learning with Mixed Graph Models. *Journal of Machine Learning Research*, v. 10, n. 41, 1187–1238.

[2] Luo, X. G., Moffa, G., and Kuipers, J. (2021). Learning Bayesian Networks from Ordinal Data. *Journal of Machine Learning Research*, v. 22, n. 266, 1–44.

[3] McNally, R. J., Mair, P., Mugno, B. L., and Riemann, B. C. (2017). Co-morbid obsessive-compulsive disorder and depression: a Bayesian network approach. *Psychological Medicine*, v. 47, n. 7, 1204–1214.

Bounds and inference on optimal decision rules

Mats Stensrud[1], **Julien Laurendeau**[2], **Aaron Sarvet**[3]

[1] *Ecole Polytechnique Federale de Lausanne, Department of Mathematics, Switzerland, mats.stensrud@epfl.ch*
[2] *Ecole Polytechnique Federale de Lausanne, Department of Mathematics, Switzerland, julien.laurendeau@epfl.ch*
[3] *Ecole Polytechnique Federale de Lausanne, Department of Mathematics, Switzerland, aaron.sarvet@epfl.ch*

Abstract

We present new results on average causal effects in settings with unmeasured exposure-outcome confounding. Our results are motivated by a class of estimands, frequently of interest in medicine and public health, that are currently not targeted by standard approaches for average causal effects. These estimands correspond to queries about the average causal effect of an *intervening* variable. We anchor our introduction of the methodological in an investigation of the role of chronic pain and opioid prescription patterns, and illustrate how conventional approaches will lead to unreplicable estimates with ambiguous policy implications. We argue that effects of intervening variables are replicable and have policy implications. Furthermore, we show that they are non-parametrically identified by the classical frontdoor formula. As an independent contribution, we derive a new semiparametric efficient estimator of the frontdoor formula with a uniform sample boundedness guarantee. This property is unique among previously described estimators in its class, and we demonstrate superior performance in finite-sample settings. The theoretical results are applied to data from the American National Health and Nutrition Examination Survey.

Keywords

Causal inference, Optimal regimes, Semiparametric inference, Unmeasured confounding.

On spatial point processes with composition-valued marks

Matthias Eckardt[1], **Sonja Greven**[2] **& Mari Myllymäki**[3]

[1] *Humboldt-Universität zu Berlin, Chair of Statistics, Germany,*
m.eckardt@hu-berlin.de
[2] *Humboldt-Universität zu Berlin, Chair of Statistics, Germany,*
Sonja.Greven@hu-berlin.de
[3] *Natural Resources Institute Finland (Luke), Helsinki, Finland,*
mari.myllymaki@luke.fi

Abstract

Marked spatial point processes have become a highly attractive field of methodological and applied research. While the impressive progress in data collection and storage capacities yielded an immense range of spatial point process data with highly challenging non-scalar marks, the problem of analysing such more complex spatial point processes remains elusive. In par- ticular, there are no methods for composition-valued marks, where at each point location the mark is a composition of D parts summing to a constant. Prompted by the need for a methodological framework, we extend the present methodological framework to spatial point processes with composition-valued marks and adapt common mark characteristics to the present context. The proposed methods are applied to analyse correlations in data on tree crown- to-trunk ratios and business sector compositions.

Keywords

Business sector composition, Constrained multivariate marks, Crown-to-trunk ratios, Mark correlation function, Mark variogram.

Symbolic Mathematics in R for Statistics and Data Science

Mikkel Meyer Andersen[1] **and Søren Højsgaard**[2]

[1] *Department of Mathematical Sciences, Aalborg University, Denmark, mikl@math.aau.dk*
[2] *Department of Mathematical Sciences, Aalborg University, Denmark, sorenh@math.aau.dk*

Abstract

There are many tasks in statistics and data science that involve symbolic mathematics, e.g. inversion of symbolic matrices, limits and solving non-linear equations. Earlier, users had to resort to other computer algebra systems (CAS) than R for such tasks as the ability to do symbolic mathematics in R is largely restricted to finding derivatives. With the R packages `caracas` and `Ryacas`, symbolic mathematics is now available from within R using R syntax. This includes going from mathematical (symbolic) objects to numerical evaluations. There are also other indirect use-cases of symbolic mathematics in R that can exploit other strengths of R, including literate programming in Rmarkdown and Quarto and Shiny apps with auto-generated mathematics exercises. In this talk we will discuss the two packages and demonstrate various use-cases including uses that help understanding statistical models and Shiny apps with auto-generated mathematics exercises.

Keywords

Computer Algebra System, CAS, Numerical Evaluation, Teaching, Literate Programming.

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

Myrto Limnios[1], **Niels R. Hansen**[1]

[1]*Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark,* {*myli,niels.r.hansen*}*@math.ku.dk*

Abstract

In the context of disease progression analysis, estimating the causal effect of a time-continuous treatment assigned to a given population is an important problem. This is particularly relevant for understanding (the causal) underlying phenomena in many applications gathering massive high-dimensional and time-dependent data structures, ranging from biomedicine to financial markets. In practice, existing learning algorithms inferring the underlying causal graph consider the progression of the recorded markers as time-continuous event processes, for which it is required to specify a model, see e.g., algorithms for directed mixed graphs Mogensen *et al.* (2018), and the Causal Analysis algorithm Meek (2014). We propose, in this work, a nonparametric model for testing if, a process directly influences another when conditioned on the history of others. Known as the (asymmetric) conditional local independence test, we use a version of the Local Covariance Measure from Christgau *et al.* (2022), where we model both the unknown intensity process and the test statistic using their finite order Volterra expansion. This results in a linear combination of tensor products of kernel functions composed of stochastic integrals w.r.t. the event processes observed up to that time. Under some assumptions of sparse decomposition and adequate regularity, the optimal parameters involved in the Volterra expansions are solution of a LASSO penalized method. Finite-sample concentration bounds for the estimation and prediction errors are derived, yielding data-driven optimal weights for the LASSO

penalty term, that allow for sparse and heterodastic expansions. These results are obtained by investigating nonasymptotic probabilistic Bernstein bounds for time-dependent martingales, following the works of Bacry *et al.* (2020), Hansen *et al.* (2015). Once the optimal parameters are estimated for a particular node of the graph, a practical implementation of the statistical test based on cross-validation and sample splitting is proposed and adapted from Christgau *et al.* (2022), that can be plugged in the aforementioned learning graphs algorithms.

Keywords

Lasso procedures, point processes, concentration inequalities, conditional independence hypothesis testing, causal inference.

References

Christgau A. M., Petersen L., Hansen N. R. (2022). Nonparametric Conditional Local Independence Testing. *arxiv.2203.13559*.

Bacry E., Bompaire M., Gaiffas S., Muzy J.-F. (2020). Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, v.21, n.50, 1-32.

Hansen N. R., Reynaud-Bouret P., Rivoirard V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, v.21, n.1, 83-143.

Mogensen S. W., Malinsky D., Hansen N. R. (2018). Causal learning for partially observed stochastic dynamical systems. *Proceedings of the 34th conference on Uncertainty in Artificial Intelligence*, 350-360.

Meek C. (2014). Toward learning graphical and causal process models. *Proceedings of the UAI 2014 workshop causal inference: Learning and prediction*, 43-48.

Improved estimation and prediction of COVID-19 patient-occupied intensive care unit beds with random regression coefficient Poisson models

N. Diz-Rosales[1], **M.J. Lombardía**[2], **and D. Morales**[3].

[1] *Universidade da Coruña, CITIC, Spain, naomi.diz.rosales@udc.es*
[2] *Universidade da Coruña, CITIC, Spain, maria.jose.lombardia@udc.es*
[3] *Universidad Miguel Hernández de Elche, IUICIO, Spain, d.morales@umh.es*

Abstract

The COVID-19 pandemic has had far-reaching consequences, highlighting the urgency for explanatory and predictive tools to track infection rates and care burden over time and space. In response to the challenges posed by data scarcity, uncertainty about the virus and territorial disparities, the application of mixed models in Small Area Estimation shows promise. This methodology has the potential to contribute significantly to health planning, resource allocation and intervention strategies. In this research we develop a robust framework for predicting occupied Intensive Care Unit beds by presenting an innovative Small Area Estimation methodology based on the definition of generalised linear mixed models with random regression coefficients and introducing bootstrap estimators. We employ a Laplace approximation algorithm to compute maximum likelihood estimates of model parameters and random effects mode predictors. Through extensive simulation experiments, we evaluate the performance of the methodology and then apply it to estimate and predict daily Intensive Care Unit beds occupancy by COVID-19 in Spanish health areas from November 2020 to March 2022. Our results demonstrate the predictive power of these models, which makes them very valuable for health planning

and resource allocation, regardless of factors such as SARS-CoV-2 variants, vaccination levels or restrictions. In addition, we are currently in the process of developing an interactive Shiny application to facilitate user-level health resource management, positioning our approach as a tool with potential for future pandemic preparedness.

Keywords

COVID-19, Intensive Care Unit occupancy, Random regression coefficient mixed models, Small Area Estimation.

Acknowledgments

This research is part of the grant PID2020-113578RB-I00, funded by MCIN/AEI/10.13039/501100011033/. It has also been supported by the Spanish grant PID2022-136878NB-I00, the Valencian grant Prometeo/2021/063, by the Xunta de Galicia (Competitive Reference ED431C-2020/14) and by CITIC that is supported by Xunta de Galicia, collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Sistema Universitario de Galicia (CIGUS). The first author was also sponsored by the Spanish Grant for Predoctoral Research Trainees RD 103/2019 being this work part of grant PRE2021-100857, funded by MCIN/AEI/10.13039/501100011033/ and ESF+. In addition, we thank to the Galicia Supercomputing Center (CESGA) for providing their services for part of the simulations in this work.

Clustering performance analysis using a new correlation-based cluster validity index with an R package

Nathakhun Wiroonsri[1]

[1] *King Mongkut's University of Technology Thonburi, Department of Mathematics, Thailand, nathakhun.wir@kmutt.ac.th*

Abstract

There are various cluster validity indices used for evaluating clustering results. One of the main objectives of using these indices is to seek the optimal unknown number of clusters. Some indices work well for clusters with different densities, sizes, and shapes. Yet, one shared weakness of those validity indices is that they often provide only one optimal number of clusters. That number is unknown in real-world problems, and there might be more than one possible option. We develop a new cluster validity index based on a correlation between an actual distance between a pair of data points and a centroid distance of clusters that the two points occupy. Our proposed index constantly yields several local peaks and overcomes the previously stated weakness. Several experiments in different scenarios, including UCI real-world data sets, have been conducted to compare the proposed validity index with several well-known ones. The R package used in this work is available at https://github.com/nwiroonsri/NCvalid.

Keywords

CVI, k-means, hierarchical clustering, R package, sub-optimal.

References

Wiroonsri, N. (2023). Clustering performance analysis using a new correlation-based cluster validity index. *Pattern Recognition*, 109910.

Optimising interval PLS via History Matching

Yoonsun Choi[1], **Nicolás Hernández**[2], **Tom Fearn**[3]

[1] *Dept. of Statistical Science, University College London, UK, yoonsun.choi.22@ucl.ac.uk*
[2] *Dept. of Statistical Science, University College London, UK, n.hernandez@ucl.ac.uk*
[3] *Dept. of Statistical Science, University College London, UK, t.fearn@ucl.ac.uk*

Abstract

Interval Partial Least-Squares Regression (iPLS) is an adaptation of the Partial Least-Squares Regression (PLS) tailored for high-dimensional spectral data, such as Near-infrared spectra. Spectometric data is expressed over a continous domain, therefore interval selection is a more viable alternative for feature extraction than variable selection. Despite its potential, a primary challenge in iPLS remains in the selection of optimal intervals. Although traditional approaches, such as forward and backward selection methods, have practical benefits, they have crucial limitations of heavy reliance on heuristic approaches. This project aims to propose a novel approach to interval selection in iPLS via history matching, a statistical method for calibrating complex computer models, and uncertainty quantification techniques. Gaussian Process Regression is used, emphasising its ability for flexible modelling and its provision of uncertainty estimates. This integration aims to optimise the accuracy of interval selection by utilising implausibility measures to highlight discrepancies between model predictions and observations. This work will contribute to the evolving dialogue on improving spectral data analysis techniques in the iPLS domain, with an application to the Spectrometric field.

Keywords

Interval PLS, Gaussian Processes, History Matching, NIR spectra.

Metabolic cost of load carriage in a Portuguese Army special forces team – A non-parametric approach

Nuno Almeida[1], **Rui Lucena**[2], **Paula Simões**[3]

[1] *Military Academy Research Center - Military University Institute (CINAMIL), Military Readiness Lab (MRLab), Military Academy and Interdisciplinary Centre for the Study of Human Performance (CIPER) – Faculty of Human Kinetics, University of Lisbon, Portugal, almeida.nrc@academiamilitar.pt*

[2] *Military Academy Research Center - Military University Institute (CINAMIL)and Military Readiness Lab (MRLab), Military Academy, Portugal, rui.lucena@academiamilitar.pt*

[3] *Military Academy Research Center - Military University Institute (CINAMIL), Military Readiness Lab (MRLab), Military Academy, Portugal and NOVA MATH - Center for Mathematics and Applications, NOVA University of Lisbon, paula.simoes@academiamilitar.pt*

Abstract

Military missions are often comprised of heavy load carriage whilst marching in difficult terrains. Furthermore, the ability of carrying a load is a fundamental requirement for successful missions and safety in these populations [2,6,7]. The aim of the present study was to quantify the metabolic cost of load carriage in a Portuguese special forces team. Ten active-duty military personnel of a special forces team (age;24,7 ± 3,47years; weight: 77,34±6,68 kg; height: 175±6,21cm) performed a graded protocol of until exhaustion on a treadmill in a thermoneutral environment in two different conditions: Unloaded and Fully loaded for combat [5]. The protocol consisted of a maximal continuous incremental exercise of two-minute steps the gradient was increased every 2nd minute by 2 ° up to a gradient of 8 °. Thereafter, the gradient of 8 ° was maintained and the walking speed was increased by

1 km/h every 2nd min until the subjects were not able or willing to continue the test. Peak oxygen uptake (VO2peak) and maximal aerobic velocity (MAV) were determined. Ventilatory threshold (VT) and respiratory compensation point (RCP) with corresponding heart rate and power output were also determined. All the participants were free from any injury or pain that would prevent maximal effort during performance testing. After receiving a thorough explanation of the protocol, all gave their written informed consent to the study. Data were analysed using Excel and IBM SPSS statistical software [5]. Using non-parametric statistical methods for hypothesis testing development, considering the Wilcoxon Test to compare two groups Kruskal-Wallis test for analysing the differences by step in the different variables, both with multiple pairwise comparisons, the study is performed in two main phases [1,3]. First, considering the Unloaded and Fully loaded groups, it was analysed if them provide different results of VO2peak, heart rate . Second, taking into account that higher differences were detected, we are interested in conducting an analyses which enables the definition of the lactate/ventilatory threshold, respiratory compensation point and peak oxygen uptake for each condition. The fully loaded for combat condition showed higher values of oxygen consumption and Heart Rate in all steps of the exercise when compared to the U condition. VO2peak and heart rate were significantly different between the two conditions. This study enables to enhance our understanding of the metabolic cost of load carriage, in a special forces team, and its impact in military readiness.

Keywords

Metabolic Cost, Military, Load Carriage, Non-parametric tests, Multiple Pairwise Comparisons.

Acknowledgements

This work is funded by national funds through the project SmartVest, an Army research and development project of Military Academy Research Center and Military Readiness Lab.

References

1 Casella, G. and Berger, R. (2002). Statistical inference. duxbury. Pacic Grove, CA *duxbury. Pacic Grove, CA*

2 Gerhart, H. D., Pressl, R., Storti, K. L., Bayles, M. P., Seo, Y. (2020). The effects of a loaded rucksack and weighted vest on metabolic cost and stride frequency in female adults. *Ergonomics*, 63(2), 145–151.

3 Jiang, J. (2021). Nonparametric statistics. *In Large Sample Techniques for Statistics*, 379–415, Springer International Publishing.

4 Louhevaara, V., Ilmarinen, R., Griefahn, B., Künemund, C., Mäkinen, H. (1995). Maximal physical work performance with European standard based fire-protective clothing system and equipment in relation to individual characteristics. *European journal of applied physiology and occupational physiology*, 71, 223–229.

5 Maroco, J. (2018) Análise Estatística com o SPSS Statistics.: 7ª Edição. *ReportNumber, Lda.*

6 Ricciardi, R., Deuster, P. A., Talbot, L. A. (2008). Metabolic demands of body armor on physical performance in simulated conditions. *Military medicine*, 173(9), 817–824.

7 Taylor, N. A., Peoples, G. E., Petersen, S. R. (2016). Load carriage, human performance, and employment standards. *Applied physiology, nutrition, and metabolism*, 41(6), S131–S147.

On the Bayesian Modeling of Suspended Solids in Oyo State Reservoirs

Oladapo Muyiwa Oladoja[1], **Taiwo Mobolaji Adegoke**[2]

[1] *Department of Mathematics and Statistics, First Technical University, Ibadan, Nigeria, oladapo.oladoja@tech-u.edu.ng*
[2] *Department of Mathematics and Statistics, First Technical University, Ibadan, Nigeria, taiwo.adegoke@tech-u.edu.ng*

Abstract

Aquatic life and water quality can be negatively impacted by suspended particles. They may lessen water clarity and obstruct sunlight, which may prevent aquatic plants from photosynthesis. A risk to human health if consumed, they can also transport nutrients and toxins like germs and heavy metals that can affect aquatic life. Thus minimizing suspended particles is a crucial part of managing water quality. This study is aimed at modeling suspended solids in the two major reservoirs in Oyo state (Asejire and Eleyele reservoirs). Bayesian inference, because of its incorporation of prior information, flexibility, probability estimations and prediction was used for this study. Suspended solids in both reservoirs observed over a period of 200 months from 2003 to 2019 assumes Normal Distribution. A conjugate prior for normal density was used to give an update of knowledge about suspended solids inform of the posterior distribution with unknown mean and known precision. The posterior mean and precision for Asejire Reservoir and that of Eleyele Reservoir was obtained. In addition, the 95% credible interval was obtained for the two reservoirs. It is likely that the true (unknown) Bayesian estimate of suspended solids in Asejire Reservoir and Eleyele Reservoir would lie within a particular interval. Using the knowledge of the posterior distribution, the posterior predictive distribution of future observation was determined. There is an update of belief about the posterior mean of suspended solids in both reservoirs in Oyo state. Also the posterior median and

standard deviation were evaluated if rounding was ignored. Concerned authorities should be informed that Eleyele reservoir is more polluted with suspended solids than Asejire reservoir.

Keywords

Normal Distribution, Conjugate Prior, Credible Interval, Posterior Distribution.

A Semiparametric Instrumented Difference-in-Differences Approach to Policy Learning

Pan Zhao[1], **Yifan Cui**[2]

[1] *Inria & Université de Montpellier, PreMeDICaL, France,*
pan.zhao@inria.fr
[2] *Zhejiang University, Center for Data Science, China,*
cuiyf@zju.edu.cn

Abstract

Recently, there has been a surge in methodological development for the difference-in-differences (DiD) approach to evaluate causal effects. Standard methods in the literature rely on the parallel trends assumption to identify the average treatment effect on the treated. However, the parallel trends assumption may be violated in the presence of unmeasured confounding, and the average treatment effect on the treated may not be useful in learning a treatment assignment policy for the entire population. In this article, we propose a general instrumented DiD approach for learning the optimal treatment policy. Specifically, we establish identification results using a binary instrumental variable (IV) when the parallel trends assumption fails to hold. Additionally, we construct a Wald estimator, novel inverse probability weighting (IPW) estimators, and a class of semiparametric efficient and multiply robust estimators, with theoretical guarantees on consistency and asymptotic normality, even when relying on flexible machine learning algorithms for nuisance parameters estimation. Furthermore, we extend the instrumented DiD to the panel data setting. We evaluate our methods in extensive simulations and a real data application.

Keywords

Individualized treatment rule, instrumental variable, multiple robustness, semiparametric efficiency, unmeasured confounding.

Model Selection for Sequential Inference and Optimization

Parnian Kassraie[1], **Nicolas Emmenegger**[1]
Andreas Krause[1], **Aldo Pacchiano**[2,3]

[1] *ETH Zurich,* [2] *Broad Institute of MIT and Harvard,* [3] *Boston University*

Abstract

We consider the problem of online inference and optimization, when the target function is unknown and is costly to sample from. This setting formalizes applications such as molecular design, personalized mHealth, scheduled clinical trials, and environmental monitoring, to name a few. Sequential decision-making and Bandits address such problems through algorithms that iteratively interact with the environment by drawing samples that are expected to be informative, or yield a high target value. To this end, such algorithms maintain an adaptive estimate of the target function, and use it for choosing the next sample. Therefore, the statistical modeling of the target function plays a crucial role. It is not known a priori which model is going to yield the most sample efficient algorithm, and we can only select the right model as we gather empirical evidence. This leads us to ask, can we perform online model selection, while simultaneously optimizing for the target function? This talk details the problem of online model selection and its challenges, e.g., handling non-i.i.d. and non-diverse data. We recover a scenario under which simultaneous model selection and optimzation is possible, and propose ALExp, an exponential weighting algorithm for probabilistic model aggregation. ALExp can be stopped at any time with valid *regret* guarantees, and its regret has an exponentially improved dependence ($\log M$) on the number of models M. Our approach utilizes a novel time-uniform analysis of the Lasso and establishes a new connection between online learning and high-dimensional statistics. This result is presented in Kassraie et al. (2023).

Keywords

Sequential Inference, Online Model Selection, Anytime Valid Confidence Sequences, Anytime Martingale Bounds, Optimization with Expert Advice.

References

Kassraie, P., N. Emmenegger, A. Krause, and A. Pacchiano (2023). Anytime Model Selection in Linear Bandits. *Proc. Advances in Neural Information Processing Systems*.

Analysing the weight carried by a soldier, according to his function, for the development of exoskeletons

Paulo Fernandes[1], **Luís Quinto**[2], **Paula Simões**[3]

[1] *Military Academy Research Center - Military University Institute (CINAMIL), Portugal, fernandes.pjl@exercito.pt*
[2] *Military Academy Research Center - Military University Institute (CINAMIL) and Mechanical Engineering Institute (IDMEC), Instituto Superior Técnico, University of Lisbon, Portugal, luis.quinto@academiamilitar.pt*
[3] *Military Academy Research Center - Military University Institute (CINAMIL), Military Readiness Lab (MRLab), Military Academy, Portugal and NOVA MATH - Center for Mathematics and Applications, NOVA University of Lisbon, paula.simoes@academiamilitar.pt*

Abstract

Military personnel are subject to carrying loads that have physiological impacts on their bodies, which may compromise their ability to perform their tasks effectively during operations [1,3,4,6,8]. Challenges in the development of exoskeletons include the specific nature of the equipment carried by soldiers, the variety of associated movements and the physical environment in which these systems must be able to operate. In order to be applied in military operations, an exoskeleton must, among other things, help carry loads to mitigate injuries associated with the weight of the equipment and reduce the level of fatigue [8]. This work is part of the ELITE - Enhancement LITe Exoskeleton project, which aims to develop a passive exoskeleton, i.e. that does not use external energy sources, to reduce the risk of injury to soldiers and increase their readiness level. The following research aims to analyse and characterise the effort of transporting

military loads within the infantry section in an operational environment [6,7]. Various studies were conducted to understand the impact of load on the effort to which military personnel are exposed. Initially, a descriptive analysis of the weight of equipment carried by soldiers in each role was performed. Subsequently, an analysis of this weight regarding the body mass of military personnel was carried out. A study was also performed to understand the proportion of soldiers under strain based on the weight they carry. It is possible to discern that the soldier's combat system's primary influence on military personnel is mobility. It should be noted that the soldier's combat system in infantry personnel varies depending on their roles within the infantry section, as individual equipment and assigned missions differ for each military role. Considering the weight associated with equipment and weaponry, by functions performed during military operations, the aim is to analyse and compare the proportion of weight carried, compared to the weight of the soldier, as well as whether there are significant differences between the different functions, characterising the associated level of overload. Such an approach takes into account that carrying backpacks with excessive weight significantly reduces the efficiency and speed of the soldier in the performance of his work, increasing the risk of injury [3,6,7,8]. The sample consisted of 181 soldiers from the Portuguese Army, belonging to the 6th to 12th National Forces detached to the Central African Republic deployed from 2020 to 2023. Military personnel were grouped according to their roles within the Infantry section. The data was analysed using various statistical methods, implemented using IBM SPSS Statistics software, namely descriptive statistical techniques, essentially in the information summary phase, and statistical inference, to evaluate various hypotheses given the enunciated problem [2,5]. Resorting to parametric statistical methods for hypothesis testing development, considering the One Way Analysis of Variance (ANOVA) test and Kruskal-Wallis test, both with multiple pairwise comparisons, the study is performed in three main phases. First, considering the average percentage of weight carried by soldiers, according to their role, during peacekeeping operations in an International environment, second a comparison between these different functions follows [5]. Then, the overall characterisation of the force is also considered to understand the actual proportion of military personnel under stress/ workload, with a threshold of 40% of their weight in weight carried. Therefore, the development of lower limb exoskele-

tons that can enhance the capabilities of military personnel to meet all operational needs becomes evident. Future research should analyse other National Contingents detached abroad since the operational environment and mission typologies significantly vary and may impact exoskeleton's performance.

Keywords

Exoskeleton , Military Load, Soldier's Combat System, Military Requirements, Hypothesis Testing.

Acknowledgements

This work is funded by national funds through the project ELITE 2 - Enhancement LITe Exoskeleton, an Army research and development project of the Military Academy Research Center.

References

1 Andersen, K. A., Grimshaw , P. N., Kelso , R. M., Bentley , D. J. (2016). Musculoskeletal Lower Limb Injury Risk in Army Populations. *Sports Medicine Open*, 2(1), 22. https://doi.org/10.1186/s40798 016 0046 z

2 Casella, G. and Berger, R. (2002). Statistical inference. duxbury. Pacic Grove, CA *duxbury. Pacic Grove, CA*

3 Fish, L., Scharre, P. (2018). The Soldier's Heavy Load. *Center for a New American Security*, 0–20.

4 Jaworski , R. L., Jensen , A., Niederberger , B., Congalton , R., Kelly, K. R. (2015). Changes in combat task performance under increasing loads in active duty marines. *Military Medicine* , 180 (3), 179–186.

5 Maroco, J. (2018) Análise Estatística com o SPSS Statistics.: 7ª Edição. *ReportNumber, Lda.*

6 Quinto, L., Pinheiro, P., Gonçalves, S., Roupa , I., Silva, M. T. (2022). Analysis of a Passive Ankle Exoskeleton for the Reduction of the Metabolic Costs During walking A Preliminary Study. *Biosystems and Biorobotics* , 27, 541– 544.

7 Quinto, L., Pinheiro, P., Gonçalves, S., Ferreira, R., Roupa, I., da Silva, M. T. (2022). Development and Functional Evaluation of a Passive Ankle Exoskeleton to Support Military Locomotion. *Advances in Military Technology*, 17(1), 79–95.

8 Reese , M. C. T. (2010). Exoskeleton Enhancements for Marines : Tactical level Technology for an Operational Consequence . *In United States Marine Corps School of Advanced Warfighting*

The use of aggregate time series for testing conditional heteroscedasticity

Paulo Teles[1], **Wai Sum Chan**[2]

[1] *School of Economics, University of Porto, and LIAAD-INESC Porto LA, Portugal, pteles@fep.up.pt*
[2] *School of Decision Sciences, The Hang Seng University of Hong Kong, Hong Kong, PR China, wschan@hsu.edu.hk*

Many time series exhibit conditional heteroscedasticity such as stock prices or returns, interest rates or exchange rates. Time series used in empirical analysis are often temporal aggregates. We study the effects of using temporally aggregated time series in testing for heteroscedasticity. The distribution of the test statistics is affected by aggregation which causes a severe power loss that worsens with the order of aggregation. Thus, the tests often fail to detect the heteroscedastic nature of the data which is a misleading outcome and can entail wrong decisions. Our conclusions are illustrated by an empirical application.

Keywords

Conditional heteroscedasticity, Lagrange multiplier test, portmanteau test, power loss, temporal aggregation.

A Bayesian shared-parameter approach to jointly model multiple (non-)Gaussian longitudinal markers with recurrent and competing event times

Pedro Miranda Afonso[1], **Dimitris Rizopoulos**[1], **Anushka Palipana**[2], **Rhonda D. Szczesniak**[2], **Eleni-Rosalina Andrinopoulou**[1]

[1] *Erasmus Medical Center, Department of Biostatistics, The Netherlands*
[2] *Cincinnati Children's Hospital Medical Center, Division of Biostatistics and Epidemiology, U.S.A.*

Abstract

Motivated by a clinical study on cystic fibrosis (CF), we propose a Bayesian shared-parameter joint model that accommodates multiple longitudinal markers following different distributions, a recurrent event process, and multiple competing risks. The model links time-to-event and longitudinal processes with various functional forms (e.g., slope and cumulative effect) and accounts for discontinuous risk intervals and both gap and calendar timescales. We analyze the US CF Foundation Patient Registry (23,543 individuals with 266,345 years of cumulative follow-up) to study the associations between lung function decline (ppFEV_1), changes in BMI, and the risk of recurrent pulmonary exacerbations (PEx), while accounting for the competing risks of death and lung transplantation. Acknowledging ppFEV_1 as a bounded marker, we use the beta distribution to prevent biologically implausible values, whereas BMI is modeled using the Gaussian distribution. Our parallel and C++ implementation of the posterior sampling algorithms allows fast model fitting despite its complexity and large sample size. A highlight of our results is the finding that a ten-unit increase in the rate of ppFEV_1 decline increases the risk of PEx by 14.69%. The incidence of PEx is positively associated with transplantation and death,

with a one-standard-deviation increase in the frailty term increasing the hazard by 290.74% and 229.95%, respectively. Our comprehensive approach provides new insights into CF progression. The model has been added to the CRAN R package `JMbayes2`.

Keywords

Bounded outcomes, competing risks, joint model, multivariate longitudinal data, recurrent events.

References

Rizopoulos, D.; Miranda Afonso, P.; Papageorgio, G. (2023). JMbayes2: Extended joint models for longitudinal and time-to-event data.

A dynamically rational framework of probability aggregation

Polina Gordienko[1]

[1] *Ludwig Maximilian University of Munich, Department of Statistics, Germany, polina.m.gordienko@gmail.com*

Abstract

Advocating for a paradigm shift towards a dynamically rational theory of collective decision-making and addressing an impossibility theorem of dynamically rational judgement aggregation, I present a new decision theoretic framework based on propositional probability logic. The model of probability aggregation offers the possibility of non-degenerate and dynamically rational aggregation of individual attitudes under minimal constraints on the aggregation function. I produce a characterization result for a linear probabilistic aggregation rule and show that an aggregation function meeting the conditions of universal domain, systematicity, monotonicity and collective rationality is non-dictatorial. In a dynamic setting, I illustrate that a linear expert rule that satisfies universal domain, collective rationality and non-dictatorship is dynamically rational. I explore applications of the models of judgement and probability aggregation in statistics, particularly in probabilistic classification as well as in ensemble learning.

Keywords

Probability aggregation, judgement aggregation, group decisions, external Bayesianity.

Imputation of Missing Daily Rainfall Data; A Comparison Between Artificial Intelligence and Statistical Techniques

Porntip Dechpichai[1], Usa Wannasingha Humphries[2], Muhammad Waqas[3,6], Angkool Wangwongchai[4], Phyo Thandar Hlaing[5,6]

[1] *Department of Mathematics, King Mongkut's University of Technology Thonburi , Thailand, porntip.dec@kmutt.ac.th*
[2] *Department of Mathematics, King Mongkut's University of Technology Thonburi , Thailand, usa.wan@kmutt.ac.th*
[3] *The Joint Graduate School of Energy and Environment (JGSEE), King Mongkut's University of Technology Thonburi , Thailand, muhammad.waqa@kmutt.ac.th*
[4] *Department of Mathematics, King Mongkut's University of Technology Thonburi , Thailand, angkool.wan@kmutt.ac.thl*
[5] *The Joint Graduate School of Energy and Environment (JGSEE), King Mongkut's University of Technology Thonburi , Thailand, phyothandar.hlai@kmutt.ac.th*
[6] *Center of Excellence on Energy Technology and Environment (CEE), Ministry of Higher Education, Science, Research and Innovation, Bangkok, Thailand*

Abstract

The acquisition of a comprehensive and extensive rainfall dataset is crucial in ensuring the effective completion of a hydrological study. This study examines different statistical and artificial intelligence-based techniques (AITs) for imputing missing daily rainfall data. This study evaluated daily rainfall data collected at twenty stations from northern Thailand. The evaluation of imputation methods was conducted through the mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R2), and correlation (r).

The experimental findings revealed that the MLR and M5-MT techniques exhibited promising performance. Overall, For the MLR model, the average MAE was approximately 0.98, the average RMSE was around 4.52, and the average R2 was about 79.6the average MAE was approximately 0.91, the average RMSE was about 4.52, and the average R2 value was around 79.8recommended approach due to its ability to deliver good estimation results while offering a transparent mechanism and not necessitating prior knowledge for model creation. This study used Statistical techniques (STs), including arithmetic averaging (AA), multiple linear regression (MLR), normal-ratio (NR), nonlinear iterative partial least squares (NIPALS) algorithm, and linear interpolation were used. STs results were compared with AITs, including long-short-term-memory recurrent neural network (LSTM-RNN), M5 model tree (M5-MT), multilayer perceptron neural networks (MLPNN), support vector regression with polynomial and radial basis function SVR-poly and SVR-RBF.

Keywords

Artificial Intelligence, Deep Learning, Machine Learning, Neural Networks, Rainfall, Imputation, Missing Data.

Second-Order Stochastic Differential Equations: Parameter Estimation and Applications to Greenland Ice Core Data

Predrag Pilipovic[1], **Adeline Samson**[2], **Susanne Ditlevsen**[3]

[1] *Department of Mathematical Sciences, University of Copenhagen, Denmark, predrag@math.ku.dk*
[2] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, adeline.leclercq-samson@univ-grenoble-alpes.fr*
[3] *Department of Mathematical Sciences, University of Copenhagen, Denmark, susanne@math.ku.dk*

Abstract

Second-order stochastic differential equations (SDEs) are encountered in diverse scientific disciplines, with applications ranging from physics to biology. This study presents a robust parameter estimation method for second-order SDEs by transforming them into a system of first-order SDEs, introducing an auxiliary velocity variable. This transformation yields a hypoelliptic first-order system since the noise is directly affecting only the velocity variables. Moreover, the model is only partially observed due to the artificially introduced velocity variables. In such a setup, the conventional estimator based on the Euler-Maruyama scheme is ill-conditioned. To address this challenge, we introduce two Strang splitting scheme estimators tailored for both complete and partially observed cases, demonstrating the consistency and asymptotic normality of both estimators. Namely, the complete-case estimator is efficient, while the estimator in the partial case exhibits a larger variance for diffusion parameters due to information loss during the approximation process. We illustrate these theoretical findings through a comprehensive simulation study employing the Duffing oscillator model. Additionally, we apply our methodology to

real-world data sourced from the Greenland ice core and fit the Duffing oscillator model to it. Our research advances our understanding of parameter estimation in complex SDE systems and offers valuable insights into practical applications such as environmental studies.

Keywords

Second-Order SDEs, Parameter Estimation, Hypoellipticity, Partially Observed Systems, Greenland Ice Core Data.

Jackknife Empirical Likelihood for Quantifying Variability of Infinite-order U-statistics

Qing Wang[1], **Yichuan Zhao**[2], **Ting Zhang**[3]

[1] *Wellesley College, Department of Mathematics, USA, qwang@wellesley.edu*
[2] *Gerogia State University, Department of Mathematics and Statistics, USA, yichuan@gsu.edu*
[3] *University of Gerogia, Department of Statistics, USA, tingzhang@uga.edu*

Abstract

Infinite-order U-statistics have an abundance of practical applications, such as subsampling-based ensemble methods. However, due to the dependence of the degree k on the sample size n, theories and results under the traditional fixed-degree U-statistic framework do not apply directly. In particular, there has not been a promising method to estimate the variance of an infinite-order U-statistic, especially when k is not of a much lower order of n. We employ the jackknife empirical likelihood technique to infinite-order U-statistics. We prove that under some mild regularity conditions the fundamental framework of jackknife empirical likelihood still holds. In the context of subsampling-based ensemble methods, we evaluate the performance of the proposed methodology to construct confidence intervals for ensemble predictions through simulation studies. The proposal yields superior results compared to existing methods across various k/n settings. In addition, it yields a coverage probability that is approaching the nominal level, as the number of trees used to build the ensemble increases.

Keywords

Confidence interval, ensemble methods, jackknife empirical likelihood, U-statistics.

Decreasing the human coding burden in randomized trials with text-based outcomes via model-assisted impact analysis

Reagan Mozer[1] **and Luke Miratrix**[2]

[1] *Bentley University, Department of Mathematical Sciences, USA, RMOZER@bentley.edu*
[2] *Harvard University, Graduate School of Education, USA, luke_miratrix@gse.harvard.edu*

Abstract

For randomized trials that use text as an outcome, traditional approaches for assessing treatment impact require that each document first be manually coded for constructs of interest by trained human raters. These hand-coded scores are then used as a measured outcome for an impact analysis, with the average scores of the treatment group compared to the control (possibly adjusting for demographic variables, other observed covariates, etc.). This process, the current standard, is both time-consuming and limiting: even the largest human coding efforts are typically constrained to measure only a small set of dimensions across a subsample of available texts. In this work, we present an inferential framework that can be used to increase the power of an impact assessment, given a fixed human-coding budget, by taking advantage of any "untapped" observations – those documents not manually scored due to time or resource constraints – as a supplementary resource. Our approach, a methodological combination of causal inference, survey sampling methods, and machine learning, has four steps: (1) select and code a sample of documents; (2) build a machine learning model to predict the human-coded outcomes from a set of automatically extracted text features; (3) generate machine-predicted scores for all documents and use these scores to estimate treatment impacts; and (4) adjust the final impact estimates using the

residual differences between human-coded and machine-predicted outcomes. As an extension to this approach, we also develop a strategy for identifying an optimal subset of documents to code in Step 1 in order to further enhance precision. Through an extensive simulation study based on data from a recent field trial in education, we show that our proposed approach can be used to reduce the scope of a human-coding effort while maintaining nominal power to detect a significant treatment impact.

Keywords

Text analysis, automated scoring, randomized controlled trial, causal inference, machine learning.

Using spatial point process models to define confidence service facilities sitting regions

Regina Bispo[1] **and Filipe J. Marques**[1]

[1]*NOVAMATH Center for Mathematics and Applications and Department of Mathematics, NOVA School of Science and Technology, Universidade NOVA da Lisboa, Caparica, Portugal*

Abstract

Traditional location problems include customers with known locations, facilities that need to be sited, a space occupied by both customers and facilities, and a metric that measures the cost allocation. The optimal location for facilities then typically involves minimizing a given cost function, e.g., time or distance. These approaches, although extensively used, ignore the potential inhomogeneous spatial distribution of customers and/or clients and fail to address the random nature of the process. To overcome these limitations, in this study we propose a method based on modelling customers as a realization of a spatial point process. Modelling the process allowed to infer the true underlying process and simulate independent patterns, which in turn enable to, instead of giving just one unique location as optimal, define a spatial distribution for locations and, consequently, define an optimal confidence region for siting facilities. As an application example, the method was used to reconfigure fire stations layout at Aveiro, Portugal.

Keywords

Location-allocation problems, spatial point processes, urban fires.

Robust Estimators of Two-Dimensional Sinusoidal Model Parameters

Debasis Kundu[1], **Rhythm Grover**[2]

[1] *Indian Institute of Technology Kanpur, Department of Mathematics and Statistics , India, kundu@iitk.ac.in*
[2] *Indian Institute of Technology Guwahati, Mehta Family School of Data Science and Artificial Intelligence, India, rhythmgrover@iitg.ac.in*

Abstract

In this work, we consider a two-dimensional (2-D) sinusoidal model. This particular model has several applications in statistical signal processing and texture analysis. Exten- sive work has been done in developing several efficient procedures and establishing their properties. The least squares estimators (LSEs) are known to be the most efficient esti- mators in presence of additive noise. But it is observed that the accuracy of the LSEs is easily affected by even small perturbations in the data. In this paper, we propose to use the weighted least squares estimators (WLSEs) of the unknown parameters in presence of additive white noise. It is observed that the WLSEs are more robust towards outliers than the least squares estimators (LSEs). Moreover, the WLSEs behave in the same manner as one of the most well-known robust estimators, the least absolute deviation estimators (LADEs). It is observed that developing the properties of the LADEs is not immediate. We derive the consistency and asymptotic normality properties of the WLSEs. Extensive simulations have been performed to show the effectiveness of the proposed method. One synthetic data set has been analyzed to illustrate how the proposed method can be used in practice.

Keywords

Two-dimensional sinusoidal model, least squares estimators, weighted least squares estimators, robust estimators, asymptotic properties.

Conditional sampling via block-triangular optimal transport maps

Ricardo Baptista[1], **Bamdad Hosseini**[2], **Nikola Kovachki**[3], **Youssef Marzouk**[4]

[1] *California Institute of Technology, Pasadena CA,* rsb@caltech.edu
[2] *University of Washington, Seattle WA,* bamdadh@uw.edu
[3] *NVIDIA Corporation, Santa Clara CA,* nkovachki@nvidia.com
[4] *Massachusetts Institute of Technology, Cambridge MA,* ymarz@mit.edu

Abstract

We present an optimal transport framework for conditional sampling of probability measures. Conditional sampling is a fundamental task of solving Bayesian inverse problems and generative modeling. Optimal transport provides a flexible methodology to sample target distributions appearing in these problems by constructing a deterministic coupling that maps samples from a reference distribution (e.g., a standard Gaussian) to the desired target. To extend these tools for conditional sampling, we first develop the theoretical foundations of block triangular transport in a Banach space setting by drawing connections between monotone triangular maps and optimal transport. To learn these block triangular maps, I will then present a computational approach, called monotone generative adversarial networks (MGANs). Our algorithm uses only samples from the underlying joint probability measure and is hence likelihood-free, making it applicable to inverse problems where likelihood evaluations are inaccessible or computationally prohibitive. We will demonstrate the accuracy of MGAN for sampling the posterior distribution in Bayesian inverse problems involving ordinary and partial differential equations and for probabilistic image in-painting.

Keywords

Optimal transport, conditional simulation, likelihood-free inference, generative models.

Harmonized Estimation of Subgroup-Specific Treatment Effects in Randomized Trials: The Use of External Control Data

Daniel Schwartz[1], **Riddhiman Saha**[2], **Steffen Ventz**[3], **Lorenzo Trippa**[4]

[1]*Dana-Farber Cancer Institute, Department of Data Science, United States, daniels@ds.dfci.harvard.edu*
[2]*Harvard T.H. Chan School of Public Health, Department of Biostatistics, United States, riddhimansaha@fas.harvard.edu*
[3]*School of Public Health, University of Minnesota, Division of Biostatistics, United States, ventz001@umn.edu*
[4]*Dana-Farber Cancer Institute, Department of Data Science, United States, ltrippa@jimmy.harvard.edu*

Abstract

Subgroup analyses of randomized controlled trials (RCTs) constitute an important component of the drug development process in precision medicine. In particular, subgroup analyses of early-stage trials often influence the design and eligibility criteria of subsequent confirmatory trials and ultimately impact which subpopulations will receive the treatment after regulatory approval. However, subgroup analyses are often complicated by small sample sizes, which leads to substantial uncertainty about subgroup-specific treatment effects. We explore the use of external control (EC) data to augment RCT subgroup analyses. We define and discuss *harmonized estimators* of subpopulation-specific treatment effects that leverage EC data. Our approach modifies subgroup-specific treatment effect estimates that are obtained by combining RCT and EC data through popular methods such as linear regression. We alter these subgroup-specific estimates to make them coherent with a robust estimate of the average effect in the enrolled population that uses RCT data only. The weighted average of the resulting subgroup-specific harmonized estimates matches the RCT-only

estimate of the overall effect in the enrolled population. We discuss the proposed harmonized estimators through analytic results and simulations, and investigate standard performance metrics. The method is illustrated in a case study of a glioblastoma RCT.

Keywords

Historical control, model misspecification, real world data, shrinkage estimation, subgroup analysis.

Stability and inference for semidiscrete OT maps

Ritwik Sadhu[1], **Ziv Goldfeld**[2], **Kengo Kato**[1]

[1] *Cornell University, Department of Statistics and Data Science, USA, rs2526@cornell.edu*
[2] *Cornell University, Department of Electrical and Computer Engineering, USA, goldfeld@cornell.edu*

Abstract

We study statistical inference for the optimal transport (OT) map (also known as the Brenier map) from a known absolutely continuous reference distribution onto an unknown finitely discrete target distribution. We derive limit distributions for the L^p-error with arbitrary $p \in [1, \infty)$ and for linear functionals of the empirical OT map, together with their moment convergence. The former has a non-Gaussian limit, whose explicit density is derived, while the latter attains asymptotic normality. For both cases, we also establish consistency of the nonparametric bootstrap. The derivation of our limit theorems relies on new stability estimates of functionals of the OT map with respect to the dual potential vector, which may be of independent interest. We also discuss applications of our limit theorems to the construction of confidence sets for the OT map and inference for a maximum tail correlation.

Keywords

Semi-discrete optimal transport, Brenier map, limit Theorems, bootstrap consistency, multivariate quantile.

Robust and Scalable Inference for Stochastic Processes

Roberto Molinari[1], **Stéphane Guerrier**[2], **Maria-Pia Victoria-Feser**[3], **Haotian Xu**[4]

[1] *Auburn University, Department of Mathematics and Statistics, USA, robmolinari@auburn.edu*
[2] *University of Geneva, Research Center for Statistics,Switzerland, stephane.guerrier@unige.ch*
[3] *University of Geneva, Research Center for Statistics,Switzerland, maria-pia.victoriafeser@unige.ch*
[4] *University of Warwick, Department of Statistics, United Kingdom, haotian.xu@warwick.ac.uk*

Abstract

This talk presents a general modelling framework that uses statistically informative quantities within the space of the wavelet transform of the data in a moment-matching fashion. This approach allows to deliver inferential solutions that scale well with the data size and are valid in the presence of complex features in the original data, including (i) non-stationary spatio-temporal dependence, (ii) missing/irregular observations and (iii) contaminated data points. We will firstly present theoretical findings specifically for the fields of time series analysis and signal processing where the computational efficiency and robust statistical properties of this approach allow to address various problems in different areas of research where large signals and different forms of data contamination make other statistical methods unusable or unreliable. We then discuss preliminary ideas and results for general stochastic processes with complex data features, going from random-effects to spatial models. The computational and practical advantages of this (distribution-free) approach will be highlighted through simulation studies and applications in fields such as engineering and economics.

Keywords

Wavelet Variance, Generalized Method of Wavelet Moments, Time Series, Spatial Data, Missing and Irregular Data, Outliers.

Generalization bounds for learning under graph-dependence

Rui-Ray Zhang[1], **Massih-Reza Amini**[2]

[1] *Barcelona Graduate School of Economics, Barcelona, Spain,*
rui.zhang@bse.eu
[2] *LIG/CNRS, University Grenoble Alpes, Grenoble, France,*
massih-reza.amini@imag.fr

Abstract

Traditional statistical learning theory relies on the assumption that data are identically and independently distributed (i.i.d.). The independently distributed assumption, on the other hand, fails to hold in many real applications. In this survey, we consider learning settings in which examples are dependent and their dependence relationship may be characterized by *dependency graphs*, a commonly used model in probability and combinatorics. We collect various graph-dependent concentration bounds, which are then used to derive Rademacher and stability generalization bounds for learning from graph-dependent data. We illustrate this paradigm with practical learning tasks and provide some research directions for future work.

Keywords

Generalization bounds, dependency graphs, uniform stability, Rademacher complexity, bipartite ranking.

Spatial data fusion adjusting for preferential sampling using INLA and SPDE

Ruiman Zhong[1], **André Victor Ribeiro Amaral**[2], **Paula Moraga**[2],

[1] *King Abdullah University of Science and Technology, Saudi Arabia, ruiman.zhong@kaust.edu.sa*
[2] *King Abdullah University of Science and Technology, Saudi Arabia*

Abstract

Spatially misaligned data can be fused by using a Bayesian melding model that assumes that underlying all observations there is a spatially continuous Gaussian random field process. This model can be used, for example, to predict air pollution levels by combining point data from monitoring stations and areal data from satellite imagery. However, if the data presents preferential sampling, that is, if the observed point locations are not independent of the underlying spatial process, the inference obtained from models that ignore such a dependence structure might not be valid. In this paper, we present a Bayesian spatial model for the fusion of point and areal data that takes into account preferential sampling. The model combines the Bayesian melding specification and a model for the stochastically dependent sampling and underlying spatial processes. Fast Bayesian inference is performed using the integrated nested Laplace approximation (INLA) and the stochastic partial differential equation (SPDE) approaches. The performance of the model is assessed using simulated data in a range of scenarios and sampling strategies that can appear in real settings. The model is also applied to predict air pollution in the USA.

Keywords

Spatial misalignment,Preferential Sampling , log Gaussian Cox process, Point patterns, INLA, SPDE.

Degree Heterogeneity in Higher-Order Networks: Inference in the Hypergraph β-Model

Sagnik Nandy[1], **Bhaswar B. Bhattacharya**[2]

[1] *University of Pennsylvania, Department of Statistics and Data Science, USA, sagnik@wharton.upenn.edu*
[2] *University of Pennsylvania, Department of Statistics and Data Science, USA, bhaswar@wharton.upenn.edu*

Abstract

The β-model for random graphs is commonly used for representing pairwise interactions in a network with degree heterogeneity. Going beyond pairwise interactions, Stasi et. al. (2014) introduced the hypergraph β-model for capturing degree heterogeneity in networks with higher-order (multi-way) interactions. In this paper we initiate the rigorous study of the hypergraph β-model with multiple layers, which allows for hyperedges of different sizes across the layers. To begin with, we derive the rates of convergence of the maximum likelihood (ML) estimate and establish their minimax rate optimality. We also derive the limiting distribution of the ML estimate and construct asymptotically valid confidence intervals for the model parameters. Next, we consider the goodness-of-fit problem in the hypergraph β-model. Specifically, we establish the asymptotic normality of the likelihood ratio (LR) test under the null hypothesis, derive its detection threshold, and also its limiting power at the threshold. Interestingly, the detection threshold of the LR test turns out to be minimax optimal, that is, all tests are asymptotically powerless below this threshold. The theoretical results are further validated in numerical experiments. In addition to developing the theoretical framework for estimation and inference for hypergraph β-models, the above results fill a number of gaps in the graph β-model literature, such as the minimax optimality of the ML

estimates and the non-null properties of the LR test, which, to the best of our knowledge, have not been studied before.

Keywords

Hypergraph, degree heterogeneity, maximum likelihood estimation, testing threshold, minimax rate.

A Kumaraswamy-Normal (Kw-N) Distribution Approach to the Basic Control Charts for Process Monitoring in Environmental Sciences

Saheed Abiodun, Afolabi*

* *Department of Mathematics, College of Computing and Mathematics, King Fahd University of Petroleum and Minerals, Kingdom of Saudi Arabia*
g202115770@kfupm.edu.sa

Abstract

For modeling purposes in various fields of research like engineering, medical sciences, biology studies, finance, economics, environmental sciences and many more; different probability distributions have been used over time. One such distribution is Kumaraswamy (Kw) distribution. Many researchers have worked on the combination of Kw and Beta (B) distributions using different approaches related to the existing literatures. This study proposes new control limits based on Kumaraswamy-Normal distribution, which belongs to Kumaraswamy-Generalized (Kw-G) family of distributions that are more flexible in controlling the skewness of data. Moreover, an algorithm is developed in R software for generating random numbers from Kumaraswamy Normal Distribution which was used to evaluate and compare the performance of some basic control charts under Kw-Normal environment. The performance of the proposed chart is evaluated through an extensive Monte Carlo simulation study. The numerical results showed that the new control chart outperforms the Kumaraswamy control chart in terms of run length analysis. Lastly, the practical application of the Kw-N control chart in modeling environmental data is demonstrated.

Keywords

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

BART for network-linked data

Sameer K. Deshpande[1]

[1] *University of Wisconsin–Madison, Department of Statistics, USA,*
sameer.deshpande@wisc.edu

Abstract

We consider regression with network-linked data in which (1) covariate-response pairs are observed at the vertices of a given network but (2) the regression relationship might be different vertex-to-vertex. We describe how to use the popular Bayesian Additive Regression Trees (BART) model for this problem in a way that does not require pre-specifying the functional form of the regression function or how the regression function varies across the network. Key to our proposal are several stochastic processes that randomly partition a network into two, possibly connected, components.

Keywords

Bayesian trees, spatial clustering, ensembles.

Large-width asymptotics for ReLu neural networks with α-stable initializations

S. Favaro[1], **Sandra Fortini**[2], **S. Peluchetti**[3]

[1] *University of Torino and Collegio Carlo Alberto, Department of Economics, Social Studies, Applied Mathematics and Statistics, Italy, stefano.favaro@unito.it*

[2] *Bocconi University, Department of Decision Sciences, Italy, sandra.fortini@unibocconi.it*

[3] *Cogent Labs, Japan, speluchetti@cogent.co.jp*

Abstract

There is a recent and growing literature on large-width asymptotic properties of Gaussian neural networks (NNs), namely NNs whose weights are initialized according to Gaussian distributions. In such a context, two popular problems are: i) the study of the large-width distributions of NNs, which characterizes the infinitely wide limit of a rescaled NN in terms of a Gaussian stochastic process; ii) the study of the large-width training dynamics of NNs, which characterizes the infinitely wide dynamics in terms of a deterministic kernel, referred to as the neural tangent kernel (NTK), and shows that, for a sufficiently large width, the gradient descent achieves zero training error at a linear rate. In this paper, we consider these problems for α-Stable NNs, namely NNs whose weights are initialized according to α-Stable distributions with $\alpha \in (0, 2]$, i.e. distributions with heavy-tails. First, for α-Stable NNs with a ReLU activation function, we show that if the NN's width goes to infinity then a rescaled NN converges weakly to an α-Stable stochastic process. As a difference with respect to the Gaussian setting, our result shows that the choice of the activation function affects the scaling of the NN. Then, we study the large-width training dynamics of α-Stable ReLU-NNs, characterizing the infinitely wide dynamics in terms of a random kernel, referred to as the α-Stable NTK, and showing that, for a sufficiently large width, the gradient descent achieves zero training error at a linear rate. The randomness of

the α-Stable NTK is a further difference with respect to the Gaussian setting, that is: within the α-Stable setting, the randomness of the NN at initialization does not vanish in the large-width regime of the training. An extension of our results to deep α-Stable NNs is discussed.

Keywords

Relu Neural network, α-Stable stochastic process, infinitely wide limit, large-width training dynamics, neural tangent kernel.

Knowing Unknowns in an Age of Incomplete Information

Saurabh Khanna[1]

[1] *University of Oxford, Pembroke College, United Kingdom, saurabh.khanna@pmb.ox.ac.uk*

Abstract

The technological revolution of the Internet has digitized the social, economic, political, and cultural activities of billions of humans. While researchers have been paying due attention to concerns of misinformation and bias, these obscure a much less researched and equally insidious problem – that of uncritically consuming 'incomplete information'. The problem of incomplete information consumption stems from the very nature of explicitly ranked information on digital platforms, where our limited mental capacities leave us with little choice but to consume the tip of a pre-ranked information iceberg. This study makes two chief contributions. First, I leverage the context of Internet search to propose a novel metric quantifying 'information completeness', i.e. how much of the information spectrum do we see, when browsing the Internet. I then validate this metric using 6.5 trillion search results extracted from daily search trends across 48 nations for one year. Second, I find causal evidence that awareness of information completeness while browsing the Internet reduces resistance to factual information, hence paving the way towards an open-minded and tolerant mindset.

Keywords

Information completeness, information retrieval, text embeddings, tolerance.

Varying coefficient regression: revisit and parametric help

Seung Hyun Moon[1], **Byeong U. Park**[2], **Young Kyung Lee**[3]

[1] *Seoul National University, Department of Statistics, South Korea, msh94kr@snu.ac.kr*
[2] *Seoul National University, Department of Statistics, South Korea, bupark@snu.ac.kr*
[3] *Kangwon National University, Department of Information Statistics, South Korea, youngklee@kangwon.ac.kr*

Abstract

This paper concerns the estimation of varying coefficient models, which is considered very useful in analyzing the regression relationship between variables. The purpose of this paper is threefold. Firstly, we introduce a new formulation of varying coefficient regression under which a structure-respecting constraint is developed for identifying the model components. Secondly, we develop a full account of locally linear kernel smoothing approach to estimating varying coefficient models which is largely missing in the literature. Thirdly, we address a bias reduction technique applied to locally linear varying coefficient regression which turns out to be successful under mild condition. We develop our methodology and theory for response variables taking values in a general Hilbert space. We discuss relevant theory for the associated projection operators, the convergence of an iterative backfitting algorithm, the error rates and asymptotic distributions of the estimators. We also include a brief simulation result demonstrating the success of the proposed approach.

Keywords

Varying coefficient model, Model identification, Local linear smoothing, Parametric help, Hilbertian response.

On Small Area Estimation Strategies using Data from Successive Surveys

Shakeel Ahmed[1]

[1] *School of Natural Sciences National University of Science and Technology, Pakistan, e-mail shakeel.ahmed@sns.nust.edu.pk*

Abstract

When information on similar characters are available on two or more surveys conducted on the same population and the estimates are likely to be stable over the period then the surveys can be combined to obtain reliable estimates at more granular level. In this article, we suggest four different strategies of obtaining small area estimates by combining the data from two successive surveys under direct, synthetic, and composite methods. The performance of the mean estimators proposed strategies is evaluated through a bootstrapped study using the demographic health surveys conducted by National Institute of Population Studies Pakistan in years 2017-18 and 2019. Strategy 2 (S2) and Strategy 3 (S3) outperform other strategies considered in this studies both in terms of mean squared error (MSE) and percentage contribution of bias (PCoB) in MSE. The suggested strategies are used to obtained estimates of the parameters (totals or means) on reproductive health characteristics in different geographical units of Pakistan. An R Package is established to obtain the estimated sample sizes, estimates of mean along with their root mean square error and 95% confidence intervals using the suggested strategies.

Keywords

Successive surveys, Super-population, Prediction, Small area.

A Unifying Approach to Distributional Limits for Empirical Optimal Transport

Shayan Hundrieser[1], **Marcel Klatt**[2], **Axel Munk**[3], and **Thomas Staudt**[4]

[1] *Institute for Mathematical Stochastics, University of Göttingen, Germany, s.hundrieser@math.uni-goettingen.de*
[2] *Institute for Mathematical Stochastics, University of Göttingen, Germany, mklatt@mathematik.uni-goettingen.de*
[3] *Institute for Mathematical Stochastics and Max Planck Institute for Multidisciplinary Sciences, University of Göttingen, munk@math.uni-goettingen.de*
[4] *Institute for Mathematical Stochastics, University of Göttingen, Germany, thomas.staudt@uni-goettingen.de*

Abstract

We provide a unifying approach to *central limit type theorems* for the empirical optimal transport (OT) cost. The limit distribution is characterized as a supremum of a Gaussian process and we explicitly characterize when it is centered normal or degenerates to a Dirac measure. Moreover, in contrast to recent contributions to distributional limit laws for empirical OT on Euclidean spaces which require centering around its expectation, the limits obtained here are centered around the *population* quantity, which is well-suited for statistical applications such as goodness-of-fit testing and randomized OT computation. At the heart of our theory is the Kantorovich duality representing OT as a supremum over a function class $\mathcal{F}_c$ for an underlying sufficiently regular cost function c, which may be unbounded. We give sufficient conditions depending on the dimension of the ground space, the underlying cost function and the probability measures under consideration to guarantee the Donsker property. Overall, our approach reveals a noteworthy trade-off inherent in central limit theorems for empirical OT: Kantorovich duality requires $\mathcal{F}_c$ to be sufficiently rich,

while the empirical processes only converges weakly if $\mathcal{F}_c$ is not too complex.

Keywords

Bootstrap, central limit theorem, empirical processes, optimal transport, Wasserstein distance.

Gaussian Approximation For Non-stationary Time Series with Optimal Rate and Explicit Construction

Soham Bonnerjee[1], **Sayar Karmakar**[2] **and Wei Biao Wu**[3]

[1] *Department of Statistics, University of Chicago,*
sohambonnerjee@uchicago.edu
[2] *Department of Statistics, University of Florida,*
sayarkarmakar@ufl.edu
[3] *Department of Statistics, University of Chicago,*
wbwu@uchicago.edu

Abstract

Inference problems for time series such as curve estimation for time-varying models or testing for existence of change-point have garnered significant attention. However, these works are restricted to the limiting assumption of independence and/or stationarity at its best. The main obstacle is that the existing optimal Gaussian approximation results for nonstationary processes only provides an existential proof and thus they are difficult to apply. In this paper, we provide a clear path to construct such a Gaussian approximation. Our proposed Gaussian approximation results encapsulates a very large class of non-stationary time series, obtains the optimal rate and yet has good applicability. Building on such a Gaussian approximation, we show theoretical results for change-point detection and simultaneous inference in presence of non-stationary errors. We substantiate our theoretical results with extensive simulation studies and some real data analyses.

Keywords

Gaussian approximation, Nonstationary time series,Simultaneous confidence band,change-point testing.

Bayesian variable selection with embedded screening

Somak Dutta[1], **Dongjin Li**[2], **Vivekananda Roy**[3], ...

[1] *Iowa State University, USA, somakd@iastate.edu*
[2] *Wells Fargo, USA, liyangxiaobei@gmail.com*
[3] *Iowa State University, USA, vroy@iastate.edu*

Abstract

During the last few decades, substantial research has been devoted to identifying the important covariates in an ultra-high dimensional linear regression where the number of covariates is in the lower exponential order of sample size. While the notion of variable screening focuses on identifying a smaller subset of covariates that includes the important ones with overwhelmingly large probability, the notion of variable selection indulges only on identifying the truly important ones. Typically, because variable selection is computationally costly, a screening step is performed to reduce the number of potential covariates. In this talk, we propose a scalable variable selection method that embeds variable screening in its algorithm, thus providing scalability and alleviating the need for a two-stage method. Our theoretical investigations relax some conditions for screening consistency and selection consistency under ultra-high dimensional setup. We illustrate our methods using a dataset with nearly half a million covariates. This talk is based on a paper by Li et al. (2022) and a method SVEN contained in R-package bravo: https://cran.r-project.org/package=bravo.

Keywords

GWAS, model averaging, prediction interval, stochastic shotgun search, SVEN.

References

Li, D., Dutta, S., and Roy, V. (2023). Model based screening embedded Bayesian variable selection for ultra-high dimensional settings. *Journal of Computational and Graphical Statistics 32*, 61–73.

On higher order approximation of Bayesian procedures through empirical Bayes

Sonia Petrone[1], **Stefano Rizzelli**[2], **Judith Rousseau**[3]

[1] *Bocconi University, Department of Decision Sciences, Italy,*
sonia.petrone@unibocconi.it
[2] *University of Padova, Department of Statistics, Italy,*
stefano.rizzelli@unipd.it
[3] *Université Paris Dauphine - PSL, CEREMADE, France,*
rousseau@ceremade.dauphine.fr

Abstract

Bayesian procedures are often optimal but analytically complex. Moreover, even when the prior law is carefully specified, it may be delicate to fix the prior hyperparameters, so that, to bypass further computational issues, it is tempting to fix them from the data, obtaining a so-called empirical Bayes posterior distribution. Although questionable, this is a common practice. We show that, in regular cases, the EB posterior distribution gives a fast approximation of an oracle Bayesian posterior (within the given class of priors). This is a faster than Bernstein-von Mises Gaussian approximations. We however also disentangle cases where asymptotic agreement fails.

Keywords

Asymptotics, Information, Marginal Maximum likelihood, Merging.

Compositional splines for approximation of bivariate densities

Stanislav Škorňa[1], **Jitka Machalová**[1], **Jana Burkotová**[1], **Karel Hron**[1], **Sonja Greven**[2]

[1] *Palacký University, Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Czech Republic, stanislav.skorna01@upol.cz*

[2] *Humboldt-Universität zu Berlin, Chair of Statistics, School of Business and Economics, Germany*

Abstract

Probability density functions represents a specific instance of distributional data resulting from massive data collection in many applications. They are used to analyze the association structure of studied phenomena and to further process them using methods of functional data analysis. For this very purpose, proper spline (continuous) representation of the input discrete data is crucial. Bayes Hilbert spaces methodology enables to capture specific properties of probability density functions and to construct so-called compositional splines which represents estimates of density functions and respect their decomposition into interactive and independent parts. Centered log-ratio, a key tool of this methodology, enables to express the original densities (and compositional splines) in the standard L^2 space by so-called ZB-spline representation. As a consequence of the transformation, the resulting spline functions fulfill zero-integral condition, which must be considered already when building the basis of the ZB-spline representation. This can be done using the standard B-spline basis with implemented zero-integral constraint or using the ZB-spline basis, which ensures integration to zero automatically and represents numerically more stable approach. In the contribution we focus on the latter case, introduce the idea of appropriate spline representation supported with a detailed simulation study and application on geochemical data.

Keywords

Functional Data Analysis, Probability Density Function, Spline Approximation.

Inverse Leverage Effect for Cryptocurrencies and Meme Stocks: a Comprehensive Framework

Lendie Follett[1], **Steven Kou**[2], **Matthew Stuart**[3], **Cindy Yu**[4]

[1] *College of Business and Public Administration, Drake University, USA, lendie.follett@drake.edu*
[2] *Department of Finance, Boston University, USA, kou@bu.edu*
[3] *Department of Mathematics and Statistics, Loyola University, USA, mstuart1@luc.edu*
[4] *Department of Statistics, Iowa State University, USA, cindyyu@iastate.edu*

Abstract

Although the leverage effect, i.e., a negative correlation between the return and volatility, and the inverse leverage effect have been suggested for equities and commodities, respectively, the existing studies suffer from an identification problem because they only model one asset. By using a comprehensive multivariate model with jumps and heavy tail distribution for both an equity index and the asset, we find inverse leverage and volatility-varying leverage effects for cryptocurrencies and meme stocks. Network effects cannot explain this finding. To handle over 18,000 latent variables, a particle Gibbs with an ancestor sampling algorithm is extended to estimate parameters efficiently.

Keywords

Cryptocurrency, Jump-Diffusion Model, Bayesian Analysis, Asymmetric Laplace Distribution.

Unguided structure learning of DAGs for count data

Thi Kim Hue Nguyen[1], **Monica Chiogna**[2], **Davide Risso**[3]

[1] *University of Padova, Department of Statistical Sciences, Italy, nguyen@stat.unipd.it*
[2] *University of Bologna, Department of Statistical Sciences "Paolo Fortunati", Italy, monica.chiogna2@unibo.it*
[3] *University of Padova, Department of Statistical Sciences, Italy, davide@stat.unipd.it*

Abstract

Directed acyclic graphs (DAGs) are considered as models for various networks for a number of reasons. First of all, when considering a network, the interest usually lies in the direction of influence making directed graphs preferred compared to the undirected graphs. When the prior information about underlying structure is available, such as, the ordering of variables is known, then the strategy of neighborhood recovery turns the problem of learning the structure of a DAG into a straight forward task. However, in many real situations, we might not know the topological ordering or it could be only unprecisely known. Here, we present a new algorithm, called learnDAG, for learning the structure of DAGs for count data without requiring prior knowledge of the ordering of variables. In particular, the proposed algorithm consists of three main steps: 1) preliminary neighbourhood selection; 2) estimation of candidate parent sets; and 3) pruning the resulting DAG. We experimentally compare learnDAG to a number of popular competitors in recovering the true structure of the graphs in situations where relatively moderate sample sizes are available. Furthermore, to make our algorithm is stronger, a validation of the algorithm is presented through the analysis of real data sets.

Keywords

Directed acyclic graphs, Graphical models, Structure learning, Count data.

Functional Graphical Lasso

Kartik Waghmare[1], **Tomas Masak**[1] **and Victor M. Panaretos**[1]

[1] *Institute of Mathematics, EPFL, Switzerland,*
firstname.lastname@epfl.ch

Abstract

We consider the problem of recovering conditional independence relationships between p jointly distributed Hilbertian random elements given n realizations thereof. We operate in the sparse high-dimensional regime, where $n \ll p$ and no element is related to more than $d \ll p$ other elements. In this context, we propose an infinite-dimensional generalization of the graphical lasso. We prove model selection consistency under natural assumptions and extend many classical results to infinite dimensions. In particular, we do not require finite truncation or additional structural restrictions. The plug-in nature of our method makes it applicable to any observational regime, whether sparse or dense, and indifferent to serial dependence. Importantly, our method can be understood as naturally arising from a coherent maximum likelihood philosophy.

Keywords

Gaussian graphical models, functional data analysis, correlation operator.

Pattern recovery by SLOPE

Tomasz Skalski[1]

[1]*Faculty of Pure and Applied Mathematics, Wrocław University of Science and Technology, Poland, tomasz.skalski@pwr.edu.pl*

Abstract

Sorted L-One Penalized Estimator (SLOPE), also known as the Ordered Weighted L-One regularization regression (OWL) is a convex regularization method for fitting high-dimensional regression models. While LASSO can eliminate redundant predictors by setting the corresponding regression coefficients to zero, SLOPE can also identify clusters of variables with the same absolute values of regression coefficients. In this talk I will discuss sufficient and necessary conditions for the proper identification of the SLOPE pattern, i.e. of the proper sign and of the proper ranking of the absolute values of individual regression coefficients, including a proper clustering. I will also mention the asymptotic results on the strong consistency of pattern recovery by SLOPE when the number of columns in the design matrix is fixed, but the sample size diverges to infinity. I will also present the geometric interpretation of the SLOPE estimator as an example of a penalized regression method with a polyhedral penalty function.

This research was supported by a French Government Scholarship.

Keywords

Linear regression, SLOPE, pattern recovery, irrepresentability condition.

References

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

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

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

Nonparametric classification with missing data

Torben Sell[1], **Thomas B. Berrett**[2], **Timothy I. Cannings**[3]

[1] *University of Edinburgh, School of Mathematics, UK,*
torben.sell@ed.ac.uk
[2] *University of Edinburgh, School of Mathematics, UK,*
timothy.cannings@ed.ac.uk
[3] *University of Warwick, Department of Statistics, UK,*
tom.berrett@warwick.ac.uk

Abstract

We introduce a new nonparametric framework for classification problems in the presence of missing data. The key aspect of our framework is that the regression function decomposes into an anova-type sum of orthogonal functions, of which some (or even many) may be zero. Working under a general missingness setting, which allows features to be missing not at random, our main goal is to derive the minimax rate for the excess risk in this problem. In addition to the decomposition property, the rate depends on parameters that control the tail behaviour of the marginal feature distributions, the smoothness of the regression function and a margin condition. The ambient data dimension does not appear in the minimax rate, which can therefore be faster than in the classical nonparametric setting. We further propose a new method, called the Hard-thresholding Anova Missing data (HAM) classifier, based on a careful combination of a k-nearest neighbour algorithm and a thresholding step. The HAM classifier attains the minimax rate up to polylogarithmic factors and numerical experiments further illustrate its utility.

Keywords

Missing data, classification, minimax.

Machine Learning-Based Modeling of Spatio-Temporally Varying Responses of Coffee Production to Climate Change: A Case Study of the Northern Region of Thailand

Usa Wannasingha Humphries[1], **Porntip Dechpichai**[2], **Muhammad Waqas**[3,6],**Angkool Wangwongchai**[4], **Phyo Thandar Hlaing**[5,6]

[1] *Department of Mathematics, King Mongkut's University of Technology Thonburi , THAILAND, usa.wan@kmutt.ac.th*
[2] *Department of Mathematics, King Mongkut's University of Technology Thonburi , THAILAND, porntip.dec@kmutt.ac.th*
[3] *The Joint Graduate School of Energy and Environment (JGSEE), King Mongkut's University of Technology Thonburi , THAILAND, muhammad.waqa@kmutt.ac.th*
[4] *Department of Mathematics, King Mongkut's University of Technology Thonburi , THAILAND, angkool.wan@kmutt.ac.th*
[5] *The Joint Graduate School of Energy and Environment (JGSEE), King Mongkut's University of Technology Thonburi , THAILAND, phyothandar.hlai@kmutt.ac.the-mail*
[6] *Center of Excellence on Energy Technology and Environment (CEE), Ministry of Higher Education, Science, Research and Innovation, Bangkok, Thailand*

Abstract

The Intergovernmental Panel on Climate Change (CC) reports indicate that CC will have detrimental effects on coffee production, leading to reduced global yields and a decrease in suitable land for coffee cultivation by 2050. Coffee holds economic significance as a cash crop in Thailand. Changes in rainfall patterns, rising temperatures, and other climatic variables can harm coffee plants. Therefore, it is essential to understand the relationship between climatic variables and coffee yield. Developing an in-depth grasp of the changes in coffee yield

is crucial for evaluating the vulnerability and adaptability of coffee production. This study involved a comprehensive data-driven analysis of the five coffee-producing provinces in Northern Thailand. The objective was to examine and model the impact of climate variability on rainfed coffee yield. Machine learning can potentially develop and understand these relations; we employed artificial neural network (ANN), support vector regression (SVR), and regular Ordinary Least Squares (OLS) regression model to estimate the correlation. Results revealed the predominant effects of climate average and extreme conditions on coffee yield. The OLS regression model demonstrated a good fit (R2=0.81). The climatic variables exhibited statistically significant relationships with the coffee yield. These variables have a substantial impact on the prediction of coffee yield. The findings highlight the significant influence of various climatic and environmental factors on the prediction of coffee yield. These results provide valuable insights into the relationship between these climatic variables and the coffee yield, contributing to our understanding of the factors influencing yield. Lastly, two machine learning algorithms (SVR and ANN) were employed to predict coffee yield using climate variables as predictors. The ANN model demonstrated superior performance (R2= 0.84) to the SVR model regarding in-season prediction skills. This study explored the relationship between climate variability and rainfed coffee yield in the northern region of Thailand, considering the potential impacts of climate change. The analysis revealed significant associations between climatic variables and coffee yield by employing OLS regression. Moreover, SVR and ANN were utilized to predict coffee yields based on climate variables. Thailand's geographical advantages and efficient coffee production processes establish it as a potential regional hub for coffee production, despite its lower output than neighboring ASEAN countries such as Vietnam and Indonesia.

Keywords

Artificial Intelligence, Machine Learning, Climate Change, Coffee, Neural Networks.

On the robustness of machine learning methods for genomic prediction

Vanda M. Lourenço[1], **Joseph O. Ogutu**[2], **Hans-Peter Piepho**[3]

[1] *NOVA School of Science and Technology, Department of Mathematics & CMA, Portugal, vmml@fct.unl.pt*
[2] *Biostatistics Unit, University of Hohenheim, Germany, jogutu2007@gmail.com*
[3] *Biostatistics Unit, University of Hohenheim, Germany, hans-peter.piepho@uni-hohenheim.de*

Abstract

The accurate prediction of genomic breeding values is central to genomic selection in both plant and animal breeding studies. Genomic prediction (GP) involves the use of thousands of molecular markers spanning the entire genome and therefore requires methods able to efficiently handle high dimensional data. Machine learning (ML) methods, which encompass different groups of supervised and unsupervised learning methods, are becoming widely advocated for and used in GP studies. Although several studies have compared the predictive performances of individual methods, studies comparing the predictive performance of different groups of methods are rare. This is also the case of studies that assess the predictive performance of methods when data are contaminated. However, such studies are crucial for (i) identifying groups of methods with superior predictive performance, and (ii) assessing the merits and demerits of such groups of methods relative to each other and to the established classical methods when the phenotypic data are and are not contaminated. Here, we comparatively evaluate in terms of predictive accuracy and prediction errors the genomic predictive performance and robustness of several groups of supervised ML methods. Specifically, regularized, ensemble, and instance-based methods, using one simulated dataset (animal breeding population; three distinct traits).

Keywords

Genomic prediction, predictive accuracy, SNPs, supervised ML methods.

Joined stochastic models for the evaluation of cancer progression from clinical data

Vincent Wieland[1], **Jan Hasenauer**[2]

[1] *University of Bonn, Life and Medical Science Institute, Germany, vincent.wieland@uni-bonn.de*
[2] *University of Bonn, Life and Medical Science Institute, Germany*

Abstract

One of today's foremost challenges in analyzing clinical data is the integration of different data modalities. Patients – especially in oncology – undergo a wide range of diagnostics which yield diverse measurement types including both discrete and continuous quantities, often in a longitudinal setting, as well as time-to-event data. In this project, we develop a computational approach for joining the different stochastic dynamics underlying such measurements into one mathematical model describing patients' disease progression, and demonstrate the use of Bayesian inference for the resulting set of model parameters. To illustrate this, we consider a simple example where one observes the size of a primary tumor, the number of different tissues affected by metastases, and the survival status of a patient. The first is growing continuously and modeled as an ordinary or stochastic differential equation, whereas the second and third are discrete events that are described as Poisson jump processes with intensities dependent on the other components. The resulting multidimensional process together with its discrete observations can be seen as a Hidden Markov Model whose parameters are to be estimated. We examine three different models from this class, which differ in the description of the continuous primary tumor growth, using simulated datasets. The first uses an exponential growth through an ODE, the second uses an Ornstein-Uhlenbeck-like SDE, whereas the third uses a version of a geometric Brownian Motion to represent the tumor size. For Bayesian inference, we leverage advanced Particle filter algorithms, a class of Markov chain

Monte Carlo methods targeting the posterior distribution by approximating the likelihood trough repeated sampling of a population of particles. This is, in the simplest ODE case, furthermore compared to maximum-likelihood estimates derived from the analytical likelihood function. For the SDE cases the results are compared to maximum-likelihood estimates obtained from a numerically approximated likelihood function. We showcase results obtained using an implementation in the Julia programming language. This newly proposed approach of joining different dynamics into one stochastic model creates valuable groundwork for clinical trajectory analysis, as it allows for a more holistic description of patient outcomes. Furthermore, it sheds light on the interplay of different stochastic processes joined into one single process and starts the development of efficient estimation methods for such classes of joined stochastic models. In future work, this can be further extended to account for different covariates, such as therapies, or mixed effects to better incorporate inter-individual variability.

Keywords

Stochastic modeling, Hidden Markov Models, Bayesian Inference, Cancer Modeling.

Coarse Personalization

Walter W. Zhang[1] **and Sanjog Misra**[2]

[1] *University of Chicago, Booth School of Business, USA,*
walterwzhang@chicagobooth.edu
[2] *University of Chicago, Booth School of Business, USA,*
sanjog.misra@chicagobooth.edu

Abstract

Advances in estimating heterogeneous treatment effects enable firms to personalize marketing mix elements and target individuals at an unmatched level of granularity, but feasibility constraints limit such personalization. In practice, firms choose which unique treatments to offer and which individuals to offer these treatments with the goal of maximizing profits: we call this the coarse personalization problem. We propose a two-step solution that makes segmentation and targeting decisions in concert. First, the firm personalizes by estimating conditional average treatment effects. Second, the firm discretizes by utilizing treatment effects to choose which unique treatments to offer and who to assign to these treatments. We show that a combination of available machine learning tools for estimating heterogeneous treatment effects and a novel application of optimal transport methods provides a viable and efficient solution. With data from a large-scale field experiment for promotions management, we find that our methodology outperforms extant approaches that segment on consumer characteristics or preferences and those that only search over a prespecified grid. Using our procedure, the firm recoups over 99.5% of its expected incremental profits under fully granular personalization while offering only five unique treatments. We conclude by discussing how coarse personalization arises in other domains.

Keywords

Optimal Transport, Machine Learning, Personalization, Targeting, Segmentation.

Feature Screening with Large Scale and High Dimensional Censored Data

Grace Y. Yi[1], **Wenqing He**[2], **Raymond Carroll**[3]

[1] *University of Western Ontario, Department of Statistical and Actuarial Sciences and Department of Computer Science, Canada, gyi5@uwo.ca*

[2] *University of Western Ontario, Department of Statistical and Actuarial Sciences, Canada, whe@stats.uwo.ca*

[3] *Texas A&M University, College Station, Texas, USA University of Technology Sydney, Broadway, Australia, carroll@stat.tamu.edu*

Abstract

Data with a huge size present great challenges in modeling, inferences, and computation. In handling big data, much attention has been directed to settings with "large p small n", and relatively less work has been done to address problems with p and n being both large, though data with such a feature have now become more accessible than before. To carry out valid statistical analysis, it is imperative to screen out noisy variables that have no predictive value for explaining the outcome variable. In this talk, we present a screening method for handling large-sized survival data, where the sample size n is large and the dimension p of covariates is of non-polynomial order of the sample size. We rigorously establish theoretical results for the proposed method and conduct numerical studies to assess its performance. Our research offers multiple extensions of existing work and enlarges the scope of high-dimensional data analysis. The proposed method capitalizes on the connections among useful regression settings and offers a computationally efficient screening procedure.

Keywords

Feature screening, High dimensional data,Survival analysis, Survival response.

Cox-Hawkes: doubly stochastic spatiotemporal Poisson processes

Xenia Miscouridou[1], **Samir Bhatt**[2], **George Mohler**[3], **Seth Flaxman** [†4], **Swapnil Mishra** [†5]

[1] *Imperial College London, Mathematics and I-X, UK, x.miscouridou@imperial.ac.uk*
[2] *University of Copenhagen, Department of Public Health, Denmark, s.bhatt@imperial.ac.uk*
[3] *Boston College, Department of Computer Science, USA, gmohler@iupui.edu*
[4] *University of Oxford, Department of Computer Science, UK, seth.flaxman@cs.ox.ac.uk*
[5] *National University of Singapore and National University Health System, Saw Swee Hock School of Public Health and Institute of Data Science, Singapore, swapnil.mishra@nus.edu.sg*
[†] *Equal Contribution*

Abstract

Hawkes processes are point process models that have been used to capture self-excitatory behaviour in social interactions, neural activity, earthquakes and viral epidemics. They can model the occurrence of the times and locations of events. Here we develop a new class of spatiotemporal Hawkes processes that can capture both triggering and clustering behaviour and we provide an efficient method for performing inference. We use a log-Gaussian Cox process (LGCP) as prior for the background rate of the Hawkes process which gives arbitrary flexibility to capture a wide range of underlying background effects (for infectious diseases these are called endemic effects). The Hawkes process and LGCP are computationally expensive due to the former having a likelihood with quadratic complexity in the number of observations and the latter involving inversion of the precision matrix which is cubic in observations. Here we propose a novel approach to perform

MCMC sampling for our Hawkes process with LGCP background, using pre-trained Gaussian Process generators which provide direct and cheap access to samples during inference. We show the efficacy and flexibility of our approach in experiments on simulated data and use our methods to uncover the trends in a dataset of reported crimes in the US.

Keywords

Gaussian process, self-excitation, Bayesian inference, space-time.

A Branching Process Model of Clonal Hematopoiesis

Xiaochen Long[1], **Marek Kimmel**[2]

[1] *Rice University, Department of Statistics, USA, xl81@rice.edu*
[2] *Rice University, Department of Statistics, USA, kimmel@rice.edu*

Abstract

We propose a hierarchical branching process model for clonal hematopoiesis. The model consists of a basic model that simulates clonal expansions and an observation process that represents the detection procedures. We consider two variants for the basic model, both based on Kendall's birth-death branching process: the first with Poisson migration, which models recurrent mutations from a fixed number of hematopoietic stem cells in spatially constrained niches, and the second with a single clone's expansion with a random starting time point. The latter variant assumes that a single mutation event gives rise to the observed mutant clones, which is appropriate when mutations only occur once. The observation process is a binomial sampling with the sequencing coverage as the total number of samples and the ratio of mutants from the basic model as the probability of detection. We also introduce multiple-timepoint observations and formulate the model as a Hidden Markov Model, which can be estimated using EM algorithm or Bayesian methods. Particularly, we derive the pmf of the two-timepoint model for estimation and the transition probability within the hidden layer. Our basic model, particularly Kendall's birth-death process with Poisson immigration, produces comparable results to Watson et al. (Science, 2020), which can be seen as an approximation of our model under a low mutation rate. Our model's predictions are consistent with the sequencing data from nearly 50,000 healthy individuals in various studies. Furthermore, by incorporating multiple-timepoint settings, our model enables the estimation and prediction of samples collected from a single individual at different times,

as well as the depiction of the entire clonal expansion history of the mutant clones from their emergence to dominance.

Keywords

Branching Process, Hidden Markov Models, Clonal Hematopoiesis.

Bayesian Analysis of Doubly Semiparametric Mixture Cure Models with Interval-censored Data

Xiaoyu Liu[1], **Liming Xiang**[2], **Shuangge Ma**[3]

[1] *Jinan University, Department of Statistics and data analysis, China, xyliu0075@jnu.edu.cn*
[2] *Nanyang Technological University, School of Physical and Mathematical Sciences, Singapore, lmxiang@ntu.edu.sg*
[3] *Yale University, School of Public health, USA, shuangge.ma@yale.edu*

Abstract

Interval-censored data with cure fraction are commonly encountered in medical studies, where the occurrence of the disease can only be recorded in intervals and the disease has a chance of being cured. The mixture cure model, assuming the Cox proportional hazards model as a latency component for the event time and logistic regression as an incidence component for the probability of uncured, is an important tool to identify risk factors for the probability of being cured and survival of uncured subjects. In the literature, linear predictors are typically incorporated in both components of the mixture cure model. In practice when some covariates are time-related, however, it is not realistic to assume the cure probability or hazard ratio of uncured subjects to be a known transformation of a linear combination of covariates. In this paper, we propose a class of doubly semiparametric mixture cure models for interval-censored data, allowing a combination of linear and nonlinear effects of covariates in both mixture components. We develop a computationally feasible Bayesian estimation procedure, which includes a two-stage data augmentation with Poisson latent variables for efficiently dealing with interval-censored data, monotone splines and polynomial splines for modelling the baseline cumulative hazard function and nonlinear terms in the model. Our

simulation results show the satisfactory performance of the proposed method in finite sample cases. The utility of the proposed method is demonstrated by the analysis of data from a hypobaric decompression sickness study.

Keywords

Bayesian analysis; Mixture cure model; Interval-censored data; Poisson data augmentation; Splines.

A High-dimensional Convergence Theorem for U-statistics with Applications to Kernel-based Testing

Kevin Han Huang[1], **Xing Liu**[2], **Andrew B. Duncan**[3], **Axel Gandy**[4]

[1] *Gatsby Unit, University College London, UK, han.huang.20@ucl.ac.uk*
[2] *Department of Mathematics, Imperial College London, UK, xing.liu16@imperial.ac.uk*
[3] *Department of Mathematics, Imperial College London and Alan Turing Institute, UK, a.duncan@imperial.ac.uk*
[4] *Department of Mathematics, Imperial College London, UK, a.gandy@imperial.ac.uk*

Abstract

We prove a convergence theorem for U-statistics of degree two, where the data dimension d is allowed to scale with sample size n. We find that the limiting distribution of a U-statistic undergoes a phase transition from the non-degenerate Gaussian limit to the degenerate limit, regardless of its degeneracy and depending only on a moment ratio. A surprising consequence is that a non-degenerate U-statistic in high dimensions can have a non-Gaussian limit with a larger variance and asymmetric distribution. Our bounds are valid for any finite n and d, independent of individual eigenvalues of the underlying function, and dimension-independent under a mild assumption. As an application, we apply our theory to two popular kernel-based distribution tests, MMD and KSD, whose high-dimensional performance has been challenging to study. In a simple empirical setting, our results correctly predict how the test power at a fixed threshold scales with d and the bandwidth.

Keywords

High-dimensional statistics, U-statistics, distribution testing, kernel method.

Exploratory Hidden Markov Factor Models for Longitudinal Mobile Health Data: Application to Adverse Posttraumatic Neuropsychiatric Sequelae

Lin Ge[1], **Xinming An**[2], **Rui Song**[3]

[1] *North Carolina State University, Department of Statistics, Institution,US, lge@ncsu.edu*
[2] *The University of North Carolina at Chapel Hill, Department of Anesthesiology, US, Xinming_An@med.unc.edu*
[3] *Amazon Inc, US, songray@gmail.com*

Abstract Adverse posttraumatic neuropsychiatric sequelae (APNS) are common among veterans and millions of Americans after traumatic exposures, resulting in substantial burdens for trauma survivors and society. Despite numerous studies conducted on APNS over the past decades, there has been limited progress in understanding the underlying neurobiological mechanisms due to several unique challenges. One of these challenges is the reliance on subjective self-report measures to assess APNS, which can easily result in measurement errors and biases (e.g., recall bias). To mitigate this issue, in this paper, we investigate the potential of leveraging the objective longitudinal mobile device data to identify homogeneous APNS states and study the dynamic transitions and potential risk factors of APNS after trauma exposure. To handle specific challenges posed by longitudinal mobile device data, we developed exploratory hidden Markov factor models and designed a Stabilized Expectation-Maximization algorithm for parameter estimation. Simulation studies were conducted to evaluate the performance of parameter estimation and model selection. Finally, to demonstrate the practical utility of the method, we applied it to mobile device data collected from the Advancing Understanding of RecOvery afteR traumA (AURORA) study.

Keywords

Hidden Markov Model, Mental Health, Digital Phenotyping, Time Series data.

Causality-oriented robustness: exploiting general additive interventions

Xinwei Shen[1], **Peter Bühlmann**[1], **Armeen Taeb**[2]

[1] *Seminar for Statistics, ETH Zürich, Switzerland,*
{xinwei.shen,buehlmann}@stat.math.ethz.ch
[2] *Department of Statistics, University of Washington, USA,*
ataeb@uw.edu

Abstract

Since distribution shifts are common in real-world applications, there is a pressing need for developing prediction models that are robust against such shifts. Existing frameworks, such as empirical risk minimization or distributionally robust optimization, either lack generalizability for unseen distributions or rely on postulated distance measures. Alternatively, causality offers a data-driven and structural perspective to robust predictions. However, the assumptions necessary for causal inference can be overly stringent, and the robustness offered by such causal models often lacks flexibility. In this paper, we focus on causality-oriented robustness and propose Distributional Robustness via Invariant Gradients (DRIG), a method that exploits general additive interventions in training data for robust predictions against unseen interventions, and naturally interpolates between in-distribution prediction and causality. In a linear setting, we prove that DRIG yields predictions that are robust among a data-dependent class of distribution shifts. Furthermore, we show that our framework includes anchor regression as a special case, and that it yields prediction models that protect against more diverse perturbations. We extend our approach to the semi-supervised domain adaptation setting to further improve prediction performance. Finally, we empirically validate our methods on synthetic simulations and on single-cell data.

Keywords

Distribution shifts, robust prediction, interventional data, structural causal models, invariance.

Exploring the causal role of the immune response to varicella-zoster virus on multiple traits: a phenome-wide Mendelian randomization study

Xinzhu Yu[1], **Artitaya Lophatananon**[2], **Krisztina Mekli**[2], **Kenneth R Muir**[2], **Hui Guo**[1]

[1] *Centre for Biostatistics, Division of Population Health, Health Services Research & Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom, xinzhu.yu@postgrad.manchester.ac.uk*

[2] *Centre for Integrated Genomic Medicine, Division of Population Health, Health Services Research & Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom*

Abstract

Background: The immune response to infections could be largely driven by the individual's genes, especially in the major histocompatibility complex (MHC) region. Varicella-zoster virus (VZV) is a highly communicable pathogen. In addition to infection, the reactivations of VZV can be a potential causal factor for multiple traits. Identification of VZV immune response-related health conditions can therefore help elucidate the aetiology of certain diseases.

Methods: A phenome-wide Mendelian randomization (MR) study of anti-VZV immunoglobulin G (IgG) levels with 1370 traits was conducted to explore the potential causal role of VZV-specific immune response on multiple traits using the UK Biobank cohort. For the robustness of the results, we performed MR analyses using five different methods. To investigate the impact of the MHC region on MR results, the analyses were conducted using instrumental variables (IVs) inside (IV_{mhc}) and outside ($IV_{no.mhc}$) the MHC region or all together (IV_{full}).

Results: Forty-nine single nucleotide polymorphisms (IV_{full}) were associated with anti-VZV IgG levels, of which five (IV_{mhc}) were located in the MHC region and 44 ($IV_{no.mhc}$) were not. Statistical evidence (false discovery rate ¡ 0.05 in at least three of the five MR methods) for a causal effect of anti-VZV IgG levels was found on 22 traits using IV_{mhc}, while no evidence was found when using $IV_{no.mhc}$ or IVfull. The reactivations of VZV increased the risk of Dupuytren disease, mononeuropathies of the upper limb, sarcoidosis, coeliac disease, teeth problems and earlier onset of allergic rhinitis, which evidence was concordant with the literature. Suggestive causal evidence (P ¡ 0.05 in at least three of five MR methods) using IV_{full}, IV_{mhc} and $IV_{no.mhc}$ was detected in 92, 194 and 56 traits, respectively. MR results from IVfull correlated with those from IV_{mhc} or $IV_{no.mhc}$. However, the results between IV_{mhc} and $IV_{no.mhc}$ were noticeably different, as evidenced by causal associations in opposite directions between anti-VZV IgG and ten traits.

Conclusions: In this exploratory study, anti-VZV IgG was causally associated with multiple traits. IVs in the MHC region might have a substantial impact on MR, and therefore, could be potentially considered in future MR studies.

Keywords

Varicella-zoster virus, anti-VZV IgG, Mendelian randomization, MR-PheWAS, MHC.

References

Hammer C, Begemann M, McLaren PJ, Bartha I, Michel A, Klose B, et al (2015). Amino Acid Variation in HLA Class II Proteins Is a Major Determinant of Humoral Response to Common Viruses. *Am J Hum Genet. 97*, 738–43.

Prediction-based statistical inference for multiple time series

Yan Liu

Waseda University, Department of Applied Mathematics, Japan,
liu@waseda.jp

Abstract

We introduce a novel minimum contrast estimator for multivariate time series in the frequency domain. While the minimum contrast estimator for univariate time series has been extensively explored, this extension remains relatively uncharted territory. The proposal in this paper is based on the prediction errors of parametric time series models. The properties of the proposed contrast estimation function are explained in detail. We also derive the asymptotic normality of the proposed estimator and compare the asymptotic variance with the existing results. The asymptotic efficiency of the proposed minimum contrast estimation is also considered. The theoretical results are illustrated by some numerical simulations.

Keywords

Multiple time series, prediction problem, minimum contrast estimation, asymptotic efficiency.

SAM: Self-adapting Mixture Prior to Dynamically Borrow Information from Historical Data in Clinical Trials

Ying Yuan[1], **Peng Yang**[2], **Yuansong Zhao**[3], **Lei Nie**[4], **Jonathon Vallejo**[4]

[1] *Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A., yyuan@mdanderson.org*
[2] *Department of Statistics, Rice University, Houston, TX, U.S.A.*
[3] *Department of Biostatistics, The University of Texas Health Science Center, Houston, TX, U.S.A.*
[4] *Center for Drug Evaluation and Research, Food and Drug Administration (FDA), Silver Spring, MD, U.S.A.*

Abstract

Mixture priors provide an intuitive way to incorporate historical data while accounting for potential prior-data conflict by combining an informative prior with a non-informative prior. However, pre-specifying the mixing weight for each component remains a crucial challenge. Ideally, the mixing weight should reflect the degree of prior-data conflict, which is often unknown beforehand, posing a significant obstacle to the application and acceptance of mixture priors. To address this challenge, we introduce self-adapting mixture (SAM) priors that determine the mixing weight using likelihood ratio test statistics. SAM priors are data-driven and self-adapting, favoring the informative (non-informative) prior component when there is little (substantial) evidence of prior-data conflict. Consequently, SAM priors achieve dynamic information borrowing. We demonstrate that SAM priors exhibit desirable properties in both finite and large samples and achieve information-borrowing consistency. Moreover, SAM priors are easy to compute, data-driven, and calibration-free, mitigating the risk of data dredging. Numerical studies show that SAM priors outperform existing methods in adopting prior-data conflicts effectively. We developed

R package "SAMprior" and web application that are freely available at www.trialdesign.org to facilitate the use of SAM priors.

Keywords

Adaptive design, Rare diseases; Dynamic information borrowing; Historical data; Mixture distribution; Real-world data.

A Frequentist Approach to Individual-Level Models for Modelling Epidemics

Patric Harrigan[1], **Mohamedou Ould Haye**[2], **Yiqiang Q. Zhao**[3]

[1] *Carleton University, Mathematics and Statistics, Canada,* *patricharrigan@cmail.carleton.ca*
[2] *Carleton University, Mathematics and Statistics, Canada,* *mohamedouhaye@cunet.carleton.ca*
[3] *Carleton University, Mathematics and Statistics, Canada,* *zhao@math.carleton.ca*

Abstract

Individual Level Models, or ILMs for short, have been studied and improved on in the last decade in order to model different large-scale epidemiological events. These models use covariate information in order to determine how probable an individual is to get sick. Under the traditional framework of this type of model, a Bayesian approach is used for estimating the parameters of the model. In this talk, we move away from this approach and look at the model through a frequentist lens. Using a reasonable asymptotic framework that we established, we show that in the univariate case, the maximum likelihood estimator is unbiased and has a normal limiting distribution. We use these results to construct confidence intervals for our parameter and determine its significance.

Keywords

Disease Modelling, Maximum Likelihood Estimators, Frequentist Approach.

Nonparametric Density Estimation for Toroidal Data

Danli Xu[1], **Yong Wang**[2]

[1] *University of Auckland, Department of Statistics, New Zealand, dxu452@aucklanduni.ac.nz*
[2] *University of Auckland, Department of Statistics, New Zealand, yongwang@auckland.ac.nz*

Abstract

Toroidal data is an extension of circular data on a torus and plays a critical part in various scientific fields. The density estimation of multivariate toroidal data based on semiparametric mixtures is studied. One of the major challenges of semiparametric mixture modelling in a multi-dimensional space is that one can not directly maximize the likelihood over the unrestricted component density as it will result in a degenerate estimate with an unbounded likelihood. To overcome this problem, it is proposed to fix the maximum of the component density, which subsequently bounds the maximum of the mixture and its likelihood function, hence providing a satisfactory density estimate. The product of univariate circular distributions are utilized to form multivariate toroidal densities as candidates for mixture components. Numerical studies show that the mixture-based density estimator is superior in general to the kernel density estimator.

Keywords

Toroidal data, density estimation, semiparametric mixture, bandwidth selection.

Adversarial Bayesian Simulation

Yuexi Wang[1], **Veronika Ročková**[2]

[1] *University of Illinois Urbana-Champaign, Department of Statistics, USA, yxwang99@illinois.edu*
[2] *University of Chicago, Booth School of Business, USA, veronika.rockova@chicagobooth.edu*

Abstract

In the absence of explicit or tractable likelihoods, Bayesians often resort to approximate Bayesian computation (ABC) for inference. Our work bridges ABC with deep neural implicit samplers based on generative adversarial networks (GANs) and adversarial variational Bayes. Both ABC and GANs compare aspects of observed and fake data to simulate from posteriors and likelihoods, respectively. We develop a Bayesian GAN (B-GAN) sampler that directly targets the posterior by solving an adversarial optimization problem. B-GAN is driven by a deterministic mapping learned on the ABC reference by *conditional* GANs. Once the mapping has been trained, iid posterior samples are obtained by filtering noise at a negligible additional cost. We propose two post-processing local refinements using (1) data-driven proposals with importance reweighting, and (2) variational Bayes. We support our findings with frequentist-Bayesian results, showing that the typical total variation distance between the true and approximate posteriors converges to zero for certain neural network generators and discriminators. Our findings on simulated data show highly competitive performance relative to some of the most recent likelihood-free posterior simulators.

Keywords

Approximate Bayesian Computation, Generative Adversarial Networks, Implicit Models, Likelihood-free Bayesian Inference, Variational Bayes.

Joint Mixed Membership Modeling of Multivariate Longitudinal and Survival Data for Learning the Individualized Disease Progression

Yuyang HE[1], **Kai Kang**[2], **Xinyuan Song**[3]

[1] *The Chinese University of Hong Kong, Department of Statistics, Hong Kong, China yuyanghe@link.cuhk.edu.hk*
[2] *Sun Yat-sen University, Department of Statistics, China kangk5@mail.sysu.edu.cn*
[3] *The Chinese University of Hong Kong, Department of Statistics, Hong Kong, China xysong@cuhk.edu.hk*

Abstract

Patients with Alzheimer's disease (AD) often exhibit substantial heterogeneity in disease progression due to multiple genetic causes for such a complex disease. Investigating diverse subtypes of neurodegeneration and individualized disease progression is essential for early diagnosis and precision medicine. In this article, we present a novel joint mixed membership model for multivariate longitudinal AD-related biomarkers and time of AD diagnosis. Unlike conventional finite mixture models that assign each subject a single subgroup membership, the proposed model assigns partial membership across subgroups, allowing subjects to lie between two or more subgroups. This flexible structure enables individualized disease progression and facilitates the identification of clinically meaningful neurological statuses often elusive in current mixed effects models. We employ a spline-based trajectory model to characterize complex and possibly nonlinear patterns of multiple longitudinal clinical markers. A Cox model is then used to examine the effects of time-variant risk factors on the hazard of developing AD. We develop a Bayesian method coupled with efficient Markov chain Monte Carlo sampling schemes to perform statistical

inference. The proposed approach is assessed through extensive simulation studies and an application to Alzheimer's Disease Neuroimaging Initiative study, showing a better performance in AD diagnosis than existing joint models.

Keywords

Mixed membership model, longitudinal data, MCMC methods, survival data.

Natural Experiment in Time Series with Bipartite Interference and Random Network

Zhaoyan Song[1], **Lucas Henneman**[2], **Georgia Papadogeorgou**[1]

[1] *Department of Statistics, University of Florida, United States, zhaoyan.song@ufl.edu*

[2] *Reva Dewberry Department of Civil, Environmental, and Infrastructure Engineering, George Mason University, United States*

[3] *Department of Statistics, University of Florida, United States, gpapadogeorgou@ufl.edu*

Abstract

In the presence of interference, a unit's outcome might be driven by the exposure level of multiple units. In some settings, the units that drive the exposure are different than the units on which the outcome is measured. Which units' exposure can drive which units' outcomes can be depicted in a bipartite graph which is often assumed to be known and fixed. In this manuscript, we consider the case of estimating causal effects in the presence of bipartite interference. We focus on the scenario where data are measured across multiple time points, and the bipartite interference network changes over time. We show that exogenous changes to the network lead to natural experiments locally in time, in that the exposure received by an outcome unit is as-if randomized within a temporal window. This allows us to estimate the causal effects of shifting the overall exposure for a specific outcome unit while controlling only for temporal trends. We propose matching algorithms and show that their bias for estimating the causal effect is bounded by algorithmic parameters and the form of temporal trends in the outcome model. We illustrate our approach with an extensive simulation study and an application on studying the effect of decreasing exposure from power plants on county-level mortality outcomes.

Keywords

Causal inference, time series, natural experiment, bipartite interference.

A Framework for Statistical Inference via Randomized Algorithms

Zhixiang Zhang[1], **Sokbae Lee**[2], **Edgar Dobriban**[3]

[1] *Department of Mathematics, University of Macau, Macao, China,* *zhixzhang@um.edu.mo*
[2] *Department of Economics, Columbia University, United States,* *sl3841@columbia.edu*
[3] *Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, United States,* *dobriban@wharton.upenn.edu*

Abstract

Randomized algorithms, such as randomized sketching or projections, are a promising approach to ease the computational burden in analyzing large datasets. However, randomized algorithms also produce non-deterministic outputs, leading to the problem of evaluating their accuracy. In this paper, we develop a statistical inference framework for quantifying the uncertainty of the outputs of randomized algorithms. We develop appropriate statistical methods—*sub-randomization*, *multi-run plug-in* and *multi-run aggregation* inference—by using multiple runs of the same randomized algorithm, or by estimating the unknown parameters of the limiting distribution. As an example, we develop methods for statistical inference for least squares parameters via random sketching using matrices with i.i.d. entries, or uniform partial orthogonal matrices. For this, we characterize the limiting distribution of estimators obtained via sketch-and-solve as well as partial sketching methods. The analysis of i.i.d. sketches uses a trigonometric interpolation argument to establish a differential equation for the limiting expected characteristic function and find the dependence on the kurtosis of the entries of the sketching matrix. The results are supported via a broad range of simulations.

Keywords

Randomized algorithms, Subsampling, Random Projections, Sketching, Random matrix theory, Least Squares.

Construction of an intelligent based CT-scan model to predict response of asthmatic patient

Marie-Félicia Béclin[1], **Pierre Lafaye de Micheaux**[2], **Nicolas Molinari**[3]

[1] *IDESP, Université de Montpellier, PreMeDICaL, Inria-Inserm, marie-felicia.beclin@umontpellier.fr*
[2] *IDESP, Université Paul Valéry Montpellier, PreMeDICaL, Inria-Inserm, pierre.lafaye-de-micheaux@univ-montp3.fr*
[3] *IDESP, Université de Montpellier, PreMeDICaL, Inria-Inserm, nicolas.molinari@inserm.fr*

Abstract

The objective is to ascertain the efficacy of Benralizumab, a medication to treat asthma. Practicians rely on specific biomarkers and clinical data. The challenge is to devise an informative feature derived from medical imaging to evaluate treatment response and predict patient's response. The images are thoracic scans in expiration and inspiration before and after one year of treatment. The hypothesis is that patients with improved conditions, will exhibit enhanced expiration scans after treatment. It is manifested by higher Hounsfield Unit values. This improvement is indicated by a shift to the right in the histograms between the pre-and post-treatment images.

We construct a model mimicking the classical linear regression model, based on the histograms. From histograms, quantiles are computed for the regression approach. The method introduced by Irpino and Verde* do not propose confidence interval and law of the estimators. So, we propose a way to define estimators by maximum of likelihood, law of the estimators and confidence interval. This approach has limitations, including the loss of spatial information and the assumption of linear relationships between voxel distributions. Investigation

is needed to develop a more general distribution-on-distribution regression method, such as the work by Chen and Ghodrati and the work of Panaretos. Another limitation is the inability to incorporate clinical covariates. Ongoing research aims to predict 2D-histograms post-treatment from scans in inspiration and expiration after registration, along with corresponding pre-treatment histograms, while including scalar covariates.

Keywords

Distribution on Distribution Regression, Imaging-derived biomarker, Treatment prediction, Histograms

Chapter 4

Posters

Revisiting the Jackson Exponentiality Test: An Investigation of its Properties and Performance

Ayana Mateus[1], **Frederico Caeiro**[2]

[1] *Center for Mathematics and Applications (NOVA Math) and Department of Mathematics, NOVA SST, Portugal, amf@fct.unl.pt*
[2] *Center for Mathematics and Applications (NOVA Math) and Department of Mathematics, NOVA SST, Portugal, fac@fct.unl.pt*

Abstract

The exponential distribution is a fundamental model applied in various fields, including queueing theory, reliability engineering, survival analysis, finance, telecommunications, quality control, machine learning, and artificial intelligence. Consequently, assessing the suitability of data for an exponential model against other alternatives has received in the last decades a lot of attention from different researchers. Possible alternative models are the gamma distribution, the Weibull distribution, the generalized Pareto distribution and the q-exponential distribution. In this work, we revisit the Jackson exponentiality test (against a general alternative). We review the exact and asymptotic properties of the statistic and compute the power of this test. We also provide a comparison between the power of the Jackson test and the power of the Lilliefors exponentiality test. Ultimately, this research contributes significantly to our understanding of the Jackson Exponentiality Test and its role within the broader context of statistical analysis.

Keywords

Exponential distribution, Jackson statistic, Monte Carlo simulation, Power of a statistical test.

References

Caeiro, F., Mateus, A. (2018). Empirical Power Study of the Jackson Exponentiality Test. *In: Skiadas, C. H., Skiadas, C. (eds) Demography and Health Issues. The Springer Series on Demographic Methods and Population Analysis*, *vol 46*. Springer, Cham.

Jackson, O.A.Y (1967). An analysis of departures from the exponential distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, *29*, 540–549.

Lilliefors, H. W. (1969). On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, *64(325)*, 387–389.

Application of Information Geometry in Incomplete Block Designs: Towards Statistical Efficiency

Carla Cardoso[1], **Amílcar Oliveira**[1,2], **Teresa A. Oliveira**[1,2]

[1] *Universidade Aberta*
[2] *Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa*

Abstract

Statistical Science along the time has been instrumental in crafting efficient experimental methodologies. A standout area in this regard is known as Design of Experiments (DoE), and the role of balanced incomplete block designs (BIBD) with or without block repetition is very well known. The concept of information geometry, coined by C.R. Rao, revolutionized the understanding of the structure of statistical parameter spaces. Rao (1945) introduced the concept of Riemannian spaces in statistics, establishing a field that remains fertile for research. Subsequently, Amari (1985) expanded these ideas, exploring the duality in parameter spaces and bringing new perspectives to information theory. In the context of incomplete block designs, Fisher (1935) emphasized the importance of efficiency in DoE and introduced fundamental concepts that remain the basis for many modern studies. The efficiency of incomplete block designs, as explored by Patterson and Bailey (1978), emphasizes the need for optimization in the selection and allocation of treatments within blocks. This work aims to explore the synergy between information geometry and incomplete block designs with repetition. The goal is not only theoretical understanding but also the practical applicability of these concepts, aiming to maximize the information extracted from each experiment with optimized resources, significantly contributing to the advancement of statistical methodology and opening new horizons for researchers and professionals in the field.

References

Rao, C.R.: Information and the accuracy attainable in the estimation of statistical parameters (1945).

Amari, S.: Differential-Geometrical Methods in Statistics (1985).

Fisher, R.A.: The Design of Experiments (1935).

Patterson, H.D., and Bailey, R.A.: Design and Analysis of Experiments in the Animal and Medical Sciences (1978).

Causal survival embeddings: non-parametric counterfactual inference under right-censoring

Carlos García-Meixide[1], **Marcos Matabuena**[2]

[1] *ETH Zürich, Switzerland, garciac@ethz.ch*
[2] *Harvard University, Biostatistics Department, US, mmatabuena@hsph.harvard.edu*

Abstract

Counterfactual inference at the distributional level presents new challenges with censored targets, especially in modern healthcare problems. To mitigate selection bias in this context, we exploit the intrinsic structure of reproducing kernel Hilbert spaces (RKHS) harnessing the notion of kernel mean embedding. This enables the development of a non-parametric estimator of counterfactual survival functions. We provide rigorous theoretical guarantees regarding consistency and convergence rates of our new estimator under general hypotheses related to smoothness of the underlying RKHS. We illustrate the practical viability of our methodology through extensive simulations and a relevant case study: the SPRINT trial. Our approach presents a distinct perspective compared to existing methods within the literature, which often rely on semi-parametric approaches and confront limitations in causal interpretations of model parameters.

Keywords

Causal Inference, Counterfactual Distributions, Survival Analysis, Right-censoring, RKHS.

Building and spatial analysis of a sustainable development index for several countries

Conceição Ribeiro[1,2], **Sílvia Pedro Rebouças**[1,34,5],**Paula Pereira**[1,6], **Mariana Corvo**[7]

[1] *Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal*
[2] *Universidade do Algarve, Instituto Superior de Engenharia, cribeiro@ualg.pt*
[3] *Instituto Superior Manuel Teixeira Gomes*
[4] *COPELabs, Universidade Lusófona*
[5] *Universidade do Algarve, Escola Superior de Gestão Hotelaria e Turismo, Portugal, smreboucas@ualg.pt*
[6] *Instituto Superior de Setúbal, Escola Superior de Tecnologia de Setúbal, Portugal, paula.pereira@estsetubal.ips.pt*
[7] *Philip Morris International, Portugal, marianacorvo0@gmail.com*

Abstract

The impact that the growth of economy had in on our planet in the last decades could not continue unnoticed, as questions started rising in the sense that development doesn't necessarily equal economic growth, and that this awareness lead to the concept of sustainable development. The Brundtland Report defines sustainable development as "the ability to meet the needs of the present without compromising the ability of future generations to meet their own needs". It also states that sustainable development is based on three dimensions: economic, social and environmental. To face these changes, various measures have been adopted at global level, including the establishment of the 17 Sustainable Development Goals (SDG). The construction of a global sustainable development index makes it possible to better understand where specific improvements should be made in order to achieve them. In this preliminary study, the aim is to analyse the existence of differences in terms of sustainable development in several

countries, according to the SDG. To this end, after constructing an index using confirmatory factor analysis, an exploratory spatial analysis was carried out in order to map the inequalities between the various countries under study.

Keywords

Sustainable development, spatial analysis.

References

Morrison, D.F. *Multivariate Statistical Methods.* 3rd edition, McGraw-Hill, N.Y, 1990.

Jolliffe, J.T. *Principal component analysis.* SpringerVerlag,New York, 1986.

Hair JF, Anderson RE, Tatham RL, Black WC. *Multivariate Data Analysis: With Readings.* London: Prentice Hall International, 1995.

Banerjee S, Carlin B, Gelfand A. *Hierarchical Modeling and Analysis for Spatial Data.* Boca Raton, FL: Chapman and Hall/CRC Press. 2nd ed., 2014.

Rebouças, S., Araripe-Silva, J., Ribeiro, C., Abreu, M. (2018). Building a sustainable development index and spacial. *Revista de Administração Pública*, 68, 257–270, 2014.

Rodríguez-Antón, J.M., Rubio-Andrada, L., Celemín-Pedroche, M.S. et al. From the circular economy to the sustainable development goals in the European Union: an empirical comparison. Int Environ Agreements 22, 67–95 (2022). https://doi.org/10.1007/s10784-021-09553-4

Lawson A.B. (2009). Bayesian Disease Mapping Hierarchical Modeling in Spatial Epidemiology. CRC Press, New York.

The Sustainable Development Goals Report 2023: Special Edition.(2023)

Increasing shrinkage in Bayesian nonparametric regression for differential expression analysis

Cristian Castiglione[1], **Nicolas Bianco**[2]

[1] *University of Padova, Department of Statistical Sciences, Italy, cristian.castiglione@unipd.it*
[2] *Universitat Pompeu Fabra, Department of Economics and Business, Spain, nicolas.bianco@upf.edu*

Abstract

Increasing shrinkage priors represent an appealing topic in both applied and theoretical Bayesian statistics. The latter have been largely used in the context of infinite factor models to estimate the number of components and learn the uncertainty around it. In these models, it is reasonable to assume that the information contained in the latent factors progressively declines, making increasing shrinkage priors a convenient choice. In this project, we exploit the increasing regularization property to introduce a new class of nonparametric Bayesian regression models. The proposed method relies on a convenient orthogonal basis expansion, wherein each component can be sorted *a priori* in descending order of importance. In particular, we implement increasing shrinkage priors for some relevant cases, such as splines, wavelets, and Fourier smoothing. An efficient Gibbs sampling algorithm for posterior inference is then discussed. Simulation results suggest that the Bayesian nonparametric regression with increasing shrinkage priors recovers the true underlying signal as well as alternative state-of-the-art approaches, while providing a more compact representation of the estimated function along with narrower credibility intervals for the regression coefficients. We then present a genomic application to the analysis of trajectory-based differential expression for single-cell sequencing data.

Keywords

Increasing shrinkage priors, nonparametric regression, Bayesian variable selection, orthogonal basis expansion, differential gene expression.

Generic Identifiability in LiNGAM models with correlated errors

Daniele Tramontano[1], **Mathias Drton**[2], **Jalal Etesami**[3], ...

[1] *Technical University of Munich, TUM School of Computation, Information and Technology, Germany, daniele.tramontano@tum.de*
[2] *Technical University of Munich, TUM School of Computation, Information and Technology, Germany, mathias.drton@tum.de*
[3] *Technical University of Munich, TUM School of Computation, Information and Technology, Germany,j.etesami@tum.de*

Abstract

For a given Directed Acyclic Mixed Graph, the Linear Non-Gaussian Acyclic Model (LiNGAM) postulates that each random variable is a linear function of its parents, with exogenous non-Gaussian error terms. In this context, we present both necessary and sufficient graphical criteria for identifying causal effects within a fixed graph. We also provide an algorithm for testing these criteria, which operates in polynomial time relative to the size of the graph. Furthermore, when the graphical criteria are met, we demonstrate that the model parameters can be determined as the solution to an optimization problem, which can be efficiently solved using gradient methods.

Keywords

LiNGAM, Generic Identifiability, Causal Effect Identification.

Enhancing Waterpipe Study Precision: Converting Pressure Drop Signals to Puffing Metrics with a Macro-Based Procedure

David Angeles[1], **Michael Pennell**[2], **Marielle Brinkman**[3]

[1] *The Ohio State University, Division of Biostatistics, United States, angeles.6@osu.edu*
[3]*The Ohio State University, Division of Biostatistics, United States, pennell.28@osu.edu*
[2] *The Ohio State University, Division of Epidemiology, United States, brinkman.224@osu.edu*

Abstract

Introduction: Commercially available waterpipes exhibit significant variations in their design and construction which affects the air infiltration rate when users puff on them. To estimate toxicant exposure from waterpipe usage, it is important to use a research-grade waterpipe (RWP) equipped with a puffing topography device to measure the puffs being taken on the RWP. Here a macro-based procedure was developed to convert pressure drop signals to flow rate and puffing topography metrics.
Methods: A single parameter quadratic regression model was proposed to provide a one-to-one correspondence between the pressure drop signals and flow rate. A smoothing technique was adapted to overcome the noisiness of the data due to bubbling of the water in the bowl of the RWP during puffing. Human puffing behavior data such as number of puffs, puff volume, and puff duration were extracted from the smooth flow rate data using a data driven response threshold value.
Results: Simulation results determined the optimality of a 15-point moving median method for smoothing puffing flow rate data when compared to a gaussian, locally weighted scatterplot smoothing, or moving average smoothing methods for a large range of moving window sizes.

The mean percentage error rate was 9.6% with a standard deviation of 8.1%. The data driven response threshold used for extracting puffing information was shown to work well for both machine-generated and human puffing data.
Conclusion: Our method provides a standardized procedure for processing raw analog voltage data from a pressure transducer monitoring the pressure drop through a fixed orifice pneumotach into meaningful puff data for each study participant. This method allows researchers to estimate the direct and indirect harm from additives in waterpipe tobacco based on human puffing behaviors.

Keywords

Pneumotach, regression, topography, transducer, waterpipe.

Asymptotic Consistency for Conditional Mode Estimator via Smoothed Quantile Regression

Eduardo Schirmer Finn[1], **Eduardo Horta**

[1] *Department of Statistics, Universidade Federal do Rio Grande do Sul (UFRGS),* eduardosfinn@outlook.com

Abstract

In situations of highly skewed or fat-tailed distributions, mean or median-based methods may be inadequate to capture the central tendencies in the data. This deficiency of traditional methods has fostered the emergence of conditional mode models as a valuable approach. However, estimating the conditional mode of a variable given its covariates presents challenges in nonparametric approaches due to the curse of dimensionality and slow convergence rates, whereas adopting a linear regression approach leads to non-convex optimization. Recent literature has addressed this issue by developing a computationally scalable way to estimate the conditional mode through the sample conditional quantile density. We propose an alternative approach, using a convolution-type smoothed version of the quantile regression instead of the canonical estimator, since the latter has jumps and, therefore $\hat{\beta}_\tau$ is not smooth in $\tau : (0, 1)$. Some Monte Carlo simulations showed that this smoothed approach outperforms mode estimator via standard quantile regression. The main goal of this work is to prove that the estimator of the conditional mode via smoothed quantile regression, $\hat{m}_h(x) = x^T \hat{\beta}_h(\hat{\tau}_{x,h})$, is consistent for the true conditional mode.

Keywords

Quantile Regression; Conditional Mode; Asymptotic Theory.

Work-life Conflict and Implementation Science: Evaluation of an Intervention Program Using a Mobile App

Gabriela Trombeta[1], **Elizabeth Joan Barham**[2]

[1] *Federal University of São Carlos (UFSCar), PostGraduate Program in Psychology, Brazil,* gabriela_trombeta@hotmail.com
[2] *Federal University of São Carlos (UFSCar), PostGraduate Program in Psychology, Brazil,* lisa@ufscar.br

Abstract

Many workers struggle to reconcile paid work demands with the rest of their lives, especially family demands, resulting in physical and psychological harm to the individuals and their families. In this study, an intervention program to reduce work-family conflict using a mobile app was developed and is currently being tested. The proposed intervention program is based on mindfulness training, aiming to improve emotional regulation to facilitate constructive reflections on how to manage difficulties related to work-family balance. Study participants will complete a pre-test protocol, weekly evaluations during the intervention period, and follow-up measures. Statistical analyses will be used to evaluate the effects of the intervention on participants' work-family conflict, management of work-family boundaries, awareness and attention, and psychological detachment.

Keywords

Work-Life Balance, Well-being, Workers, Economics, E-health.

Penalized estimation for finite mixture of multivariate regression models

Heeyeon Kang[1], **Sunyoung Shin**[2]

[1] *Pohang University of Science and Technology (POSTECH), Department of Mathematics, South Korea, heeyeonk@postech.ac.kr*
[2] *Pohang University of Science and Technology (POSTECH), Department of Mathematics, South Korea, sunyoungshin@postech.ac.kr*

Abstract

In the era of big data, it is more common to encounter high-dimensional and heterogeneous data that have dependencies among the variables. To deal with high-dimensional heterogeneous data with correlations, we consider a finite mixture of regressions model with multiple responses (mvFMR). mvFMR has the merit of capturing dependencies among the responses, accounting for multiple subpopulations within data and its own sparse regression relationships. We propose a penalized maximum likelihood approach which is an effective tool for variable selection. We develop the algorithm for obtaining a penalized maximum likelihood estimator that combines expectation-maximization algorithm with the alternating direction method of multipliers, named ADMM-EM algorithm. In our framework, the mixing proportions, the regression coefficients and the error covariances of each cluster are estimated. The optimal tuning parameter and the number of clusters are chosen using the Bayesian information criterion. Simulation studies show that our method performs successfully compared to alternatives approaches such as the unpenalized maximum likelihood estimation and the penalized estimation for finite mixture of regression models, which allows individualized clustering to each of the responses.

Keywords

Penalized maximum likelihood, Finite mixture model, Multivariate regression, ADMM, EM algorithm.

A linear mixed effects model-based permutation test to identify genes that have differentially expressed/spliced transcripts

Huining Kang[1], **Xichen Li**[2], **Li Luo**[3], **Scott A Ness**[4]

[1] *University of New Mexico, Department of Internal Medicine, USA, HuKang@salud.unm.edu*

[2] *University of New Mexico, Department of Math and Statistics, USA, jessieli@unm.edu*

[3] *University of New Mexico, Department of Internal Medicine, USA, LLuo@salud.unm.edu*

[4] *University of New Mexico, Department of Internal Medicine, USA, SNess@salud.unm.edu*

Abstract

RNA-sequencing technology has made it possible to reconstruct and quantify the alternative spliced isoform transcripts. We recently developed a mixed effects model approach to identifying genes with isoforms that are differentially expressed/spliced between different medical conditions (Luo et al, 2020, PLoS One 15(10):e0232646). The model considers correlation among the isoforms of the same gene, which have typically been ignored in other existing approaches to differential isoforms. In this presentation we further propose a permutation test for estimating the false discovery rates (FDR) which also takes into account the correlations among different genes. We evaluate and compare the performance of the permutation test to the conventional likelihood ratio test through the simulation and we applied the permutation test to an RNA-sequencing data set from a study of adenoid cystic carcinoma (ACC) to identify genes that have differentially expressed/spliced isoforms between tumor and normal tissues.

Keywords

Permutation test, likelihood ratio test, linear mixed effects model, differential gene/isoform expression, alternative splicing.

Risk factors for musculoskeletal disorders in farmers of Korea: based on survey on occupational diseases of farmers conducted by the rural development administration in 2020 and 2022

Jinheum Kim[1], **Jinwoo Park**[2]

[1] *University of Suwon, Department of Applied Statistics, South Korea, jkimdt65@gmail.com*
[2] *University of Suwon, Department of Data Science, South Korea, jwpark@suwon.ac.kr*

Abstract

According to the rural development administration (RDA)'s survey of occupational diseases of farmers, the prevalence of agricultural work-related musculoskeletal disorders (MSD) was 5.6 and 18.4% in 2020 and 2022, respectively, and the proportions of MSD in all diseases were 85.1 and 97.0%, respectively. They are very exposed to the risk of MSD. The purpose of this study is to find the risk factors for MSD in farmers and provide primary data for preventing and managing MSD in farmers. In this study, survey data on occupational diseases of farmers conducted by the RDA in 2020 and 2022 were used. The survey used a stratified multi-stage probability sampling design and adopted the face-to-face interview method. Pearson's chi-square test was performed to test whether the risk factors were marginally related to MSD. Since the number of subjects with MSD was very small compared to that of subjects who do not have MSD, resampling methods were employed to obtain artificial samples in which the number of subjects with MSD and the number of subjects who do not have MSD were almost the same, and logistic regression analysis was performed. In terms of recall, precision, and F1, the performance of both 'down' and 'down and up' resampling models were same, whereas the

former was greater than the latter in terms of Nagelkerke's R2. From the down sampling model, the odds of experiencing an MSD were 1.65 times higher for women than for men. The odds of developing MSD were 2.36 times higher in the 50-60s, 4.24 times higher in the 60-70s, and 7.40 times higher in the 70s or above than in the under 50s. For the types of farming, the odds of occurrence of MSD were 1.57 times and 1.58 times higher for greenhouse and livestock than for rice, while dry field and orchard were 0.74 and 0.84 times lower, respectively. Farmers who frequently use their necks, arms, waist, and knees were 1.14 times more likely to develop MSD than those who rarely use it, 1.35 times, 1.32 times, and 1.35 times, respectively. However, farmers who did not use their fingers or wrists for work had 1.22 times higher odds of developing MSD than farmers who used them frequently. Farmers who frequently lifted more than 20kg objects had 1.28 times higher odds of developing MSD than farmers who rarely lifted them, whereas farmers who rarely lifted objects weighing 10-20 kg were 1.39 times higher than those who rarely lifted them. In conclusion, when performing binary classification from data with imbalanced classes, it was found that analyzing with synthetic data resampled rather than original data can increase the accuracy of the model as well as parameter estimation. In addition, the major risk factors for MSD were age, gender, types of farming, use of pesticides, and income. Among body parts, cuff uses and lifting of objects weighing less than 20kg were risk factors.

Keywords

Down sampling, musculoskeletal disorders, Farmers, Accuracy.

Clustering Hidden Markov Model

Jinseong Bok[1], **Sunyoung Shin**[2]

[1] *Pohang University of Science Technology, Department of Mathematics, Republic of Korea, b36389@postech.ac.kr*
[2] *Pohang University of Science Technology, Department of Mathematics, Republic of Korea, sunyoungshin@postech.ac.kr*

Abstract

Hidden Markov Models (HMMs) are powerful tools for modeling sequential data across various domains, such as speech recognition, natural language processing, and bioinformatics. Statistical estimation and inference of HMMs have been investigated for a fixed number of hidden states in the literature, where the optimal number of hidden states may be estimated based on the data collected. In this study, we consider cases where the choice of hidden states of the HMM is not straightforward. A selected set of hidden states may include redundant states that share the same transition probabilities, which can be viewed as partitioning the hidden states. A novel framework we develop, named Clustered Hidden Markov Model (CHMM) transforms the hidden state space of an HMM into clusters, where these clusters collectively constitute a second-order HMM. CHMM keeps the partitioning of hidden states confidential, revealing only the original hidden state space. The exposition of the CHMM framework we provide explains how the CHMM is related with standard HMMs with second-order dependencies. CHMMs represent a new approach to hidden state modeling, reducing the number of hidden states while maintaining their information. It can be very useful when the choice of hidden states is ambiguous for modeling sequential data such that the number of hidden states we start with is large. To recover the hidden state partitioning, we consider penalized estimation in HMM, which maximizes the likelihood regularized by adaptive group fused lasso. We establish asymptotic properties of the penalized estimator. Simulations studies demonstrate the excellent performance of our penalized estimation.

Keywords

Hidden Markov Model, 2nd-order Hidden Markov Model, Adapted Group Fused Lasso, ADMM.

Low-rank, Orthogonally Decomposable Tensor Regression With Internal Variation Penalty

Jungmin Kwon[1], **Cheolwoo Park**[2], **Jeongyoun Ahn**[3]

[1] *KAIST, Department of Mathematical Sciences, South Korea, greenie1@kaist.ac.kr*
[2] *KAIST, Department of Mathematical Sciences, South Korea, parkcw2021@kaist.ac.kr*
[3] *KAIST, Department of Industrial & Systems Engineering, South Korea, jyahn@kaist.ac.kr*

Abstract

Multi-dimensional tensor data have gained increasing attention in the recent years. We consider the problem of fitting a generalized linear model with a three-dimensional image covariate, such as one obtained by functional magnetic resonance imaging (fMRI). Many of the classical penalized regression techniques do not account for the spatial structure in imaging data. We assume the parameter tensor is orthogonally decomposable, enabling us to penalize the tensor singular values and avoid a priori specification of the rank. Under this assumption, we propose to penalize internal variation of the parameter tensor. Our approach provides an effective method to reduce the dimensionality and control piecewise smoothness of imaging data. Effectiveness of our method is demonstrated on synthetic data and real MRI imaging data.

Keywords

Low-rank approximation, Nuclear norm, Internal variation, Tensor regression, ...

Local nonparametric linear estimation of regression functions based on random functional designs and correlated errors

Karim Benhenni[1], **Ali Hajj Hassan**[2]

[1] *Université Grenoble Alpes, Laboratoire Jean Kuntzmann, UMR 5224, CNRS, France, karim.benhenni@univ-grenoble-alpes.fr*
[2] *Université Grenoble Alpes, Laboratoire Jean Kuntzmann, UMR 5224, CNRS, France, ali.hajj-hassan@univ-grenoble-alpes.fr*

Abstract

This work considers the problem of nonparametric estimation of the regression operator r in a functional regression model $Y = r(X)+\varepsilon$ with a scalar response Y, a functional explanatory variable X, and a second order stationary error process ε. Under some specific criteria, we construct a local linear kernel estimator of r from functional random design with correlated errors. The exact rates of convergence in mean squared error of the constructed estimator are established for both short and long range dependent error processes. Simulation studies are conducted on the performance of the proposed simple local linear estimator and examples of time series data are considered.

Keywords

Local linear kernel estimation, functional random design data, nonparametric regression operator, short and Long memory error processes, OU and ARFIMA processes.

Transfer Learning With Efficient Estimators to Optimally Leverage Historical Data in Analysis of Randomized Trials

Lauren D. Liao[1], **Alan E. Hubbard**[2], **Alejandro Schuler**[3]

[1]*University of California, Division of Biostatistics, U.S.A., ldliao@berkeley.edu,*[2]*University of California, Division of Biostatistics, U.S.A., hubbard@berkeley.edu,*[3]*University of California, Division of Biostatistics, U.S.A., alejandro.schuler@berkeley.edu*

Abstract

Randomized controlled trials (RCTs) are a cornerstone of comparative effectiveness because they remove the confounding bias present in observational studies. However, RCTs are typically much smaller than observational studies because of financial and ethical considerations. Therefore it is of great interest to be able to incorporate plentiful observational data into the analysis of smaller RCTs. Previous estimators developed for this purpose rely on unrealistic additional assumptions without which the added data can bias the effect estimate. Recent work proposed an alternative method (prognostic adjustment) that imposes no additional assumption and increases efficiency in the analysis of RCTs. The idea is to use the observational data to learn a prognostic model: a regression of the outcome onto the covariates. The predictions from this model, generated from the RCT subjects' baseline variables, are used as a covariate in a linear model. In this work, we extend this framework to work when conducting inference with nonparametric efficient estimators in trial analysis. Using simulations, we find that this approach provides greater power (i.e., smaller standard errors) than without prognostic adjustment, especially when the trial is small. We also find that the method is robust to observed or unobserved shifts between the observational and trial populations and does not introduce bias. Lastly, we showcase this estimator leveraging real-world historical data on a randomized blood transfusion study of trauma patients.

Keywords

Causal inference, historical data, prognostic score, randomized trials.

Model-based clustering of pandemic trajectories with common historical change times

Riccardo Corradin[1], **Luca Danese**[2], **Wasiur KhudaBukhsh**[1], **Andrea Ongaro**[2]

[1] *University of Nottingham, School of Mathematical Sciences, UK*
[2] *University of Milano-Bicocca, DEMS, Italy*

Abstract

As the recent COVID-19 pandemic has shown, to understand the effect of policies such as lockdowns and mass immunization, it has become important to study the evolution of infectious diseases. In this context, we propose a novel modelling strategy to cluster different states according to the dynamic of the epidemic trajectory, where we assume as unique commonality that two states belong to the same group if structural changes in the evolution of their trajectories happen at the same times. We describe the dynamics of the epidemic with a mechanistic model of disease spread. Our object of interest is a latent random partition of the states generated by a discrete distribution whose weights are distributed as a Dirichlet distribution. To obtain a posterior estimate of the partition, we propose a collapsed Gibbs sampler based on a split-and-merge strategy. We validate the capability of the model to estimate the latent partition of the data through an intensive simulation study. Finally, our proposal is applied to real COVID-19 data where we cluster together different US states on the base of their structural changes in the epidemic spreading.

Keywords

Model-Based Clustering, Bayesian Statistics, SIR Model, Change Points.

Cross-Temporal Forecast Reconciliation at Digital Platforms with Machine Learning

Jeroen Rombouts[1], **Marie Ternes**[2], **Ines Wilms**[3]

[1] *Essec Business School, Information Systems, Decision Science and Statistics, France, rombouts@essec.edu*
[2] *Maastricht University, Quantitative Economics, Netherlands, m.ternes@maastrichtuniversity.nl*
[3] *Maastricht University, Quantitative Economics, Netherlands, i.wilms@maastrichtuniversity.nl*

Abstract

Time series to be forecast are oftentimes naturally organized in a hierarchical structure. In this paper, we consider on-demand delivery platforms as prime example as they require forecasts at different levels of cross-sectional and temporal aggregation. Indeed, their market place is typically split up in different geographical regions, so a natural *cross-sectional* aggregation scheme arises from many individual delivery areas over regions towards few cities. Moreover, a *temporal* aggregation scheme naturally arises since fast operational decisions (e.g., in terms of minutes) are needed to ensure the platform's service couriers are at the right time and location to serve consumer demand promptly, but strategic business decisions also require long-term planning since on and off-boarding of couriers is usually done at a lower temporal frequency (e.g., in terms of weeks). Accurate and coherent demand forecasts across all levels of the cross-sectional and temporal hierarchy are therefore key to the business' success and to support aligned decision making across different planning units. While there is a large literature on hierarchical forecast reconciliation to produce coherent forecasts across the hierarchy, only limited research has been done to reconcile forecasts in both the cross-sectional as well as the temporal direction, hence in a *cross-temporal framework*. We introduce a non-linear hierarchical forecast reconciliation method that produces cross-temporal reconciled forecasts in a direct and automated

way through the use of popular machine learning (ML) methods such as random forest and XGBoost. Our ML-based approach generalizes existing state-of-the-art linear forecast reconciliation methods for the cross-temporal framework. We empirically test our framework on a unique, large-scale demand data set from a leading on-demand delivery platform in Europe.

Keywords

Hierarchical time series, Forecast reconciliation, Machine learning.

Nonparametric Tests for Serial Independence in Linear Model against a Possible Autoregression of Error Terms

Jana Jurečková[1,2], Olcay Arslan[3], Yeşim Güney[3], Yetkin Tuaç[3], Jan Picek[4], Martin Schindler[4]

[1] *The Czech Academy of Sciences, Institute of Information Theory and Automation, Czech Republic*
[2] *Charles University, Czech Republic*
[3] *Department of Statistics, Faculty of Science, Ankara University, 06100, Ankara, Turkey*
[4] *Technical University in Liberec, Czech Republic, martin.schindler@tul.cz*

Abstract

In the linear regression model with possibly autoregressive errors, a family of nonparametric tests for autoregression of errors under a nuisance regression is proposed. The tests are based on regression rank scores under the null hypothesis. The asymptotic distribution of the test criterion under the null hypothesis, as well as under local autoregression alternatives is given. The behavior of the proposed test is illustrated and its computation is described. In the simulation study the power of the test is estimated under various setting of the parameters.

Keywords

Autoregression rank scores; Linear regression; Hypothesis testing; Rank test; Regression rank scores.

Testing Symmetry Around a Line or Subspace

S. Hudecová[1], M. Siman[2]

Charles University, KPMS MFF UK, Czechia,
hudecova@karlin.mff.cuni.cz[1]
he Czech Academy of Sciences, ÚTIA AV CR, v.v.i., Czechia[2]

Abstract

Suppose a d-dimensional random vector $\mathbf{Y}$ is accompanied with a random vector $\mathbf{Z}$. The conditional distribution $\mathcal{L}(\mathbf{Y}|\mathbf{Z})$ is said to be symmetric around a subspace generated by basic vectors $(\mathbf{u}_1, \ldots, \mathbf{u}_q)$, $1 \leq q < d$, up to a shift when $\mathcal{L}\{\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{Z})|\mathbf{Z}\} = \mathcal{L}\{\mathbb{R}(\mathbf{Y} - \mathbb{E}(\mathbf{Y}|\mathbf{Z}))|\mathbf{Z}\}$ for the reflection matrix $\mathbb{R} = 2\mathbf{u}_1\mathbf{u}_1' + \cdots + 2\mathbf{u}_q\mathbf{u}_{q-1}' - \mathbb{I}$, which is the rotational (orthonormal) matrix that satisfies $\mathbb{R}\mathbf{u} - i = \mathbf{u}_i$, $i = 1, \ldots, q$, and $\mathbb{R}\mathbf{v} = -\mathbf{v}$ for any vector $\mathbf{v}$ orthogonal to all vectors $\mathbf{u}_i, i = 1, \ldots, q$. The axial symmetry corresponds to $q = 1$, the hyperplane symmetry corresponds to $q = d - 1$ and the unconditional case can be obtained by omitting $\mathbf{Z}$. The presentation shows how to test these symmetries in various ways, e.g., by means of quantile regression coefficients, rank scores or integrated rank scores, canonical correlations or Kendall's rank correlations. It also deals with the properties and benefits of such tests.

Keywords

Axial symmetry, hyperplane symmetry, subspace symmetry, statistical test.

References

Hudecová, S; Siman, M. (2021). Testing axial symmetry by means of directional regression quantiles. *Electronic Journal of Statistics 15*, 2690–2715.

Hudecová, S; Siman, M. (2021). Testing symmetry around a subspace. *Statistical Papers 62*, 2491–2508.

Hudecová, S; Siman, M. (2023). Testing axial symmetry by means of integrated rank scores. *Journal of Nonparametric Statistics 35*, 474–490.

Siman, M. (2024). Testing axial symmetry by means of directional quantile regres- sion coefficients, in press.

Child Growth Curve in Sofala - Mozambique and its comparison with other contexts

Neto Pascoal[1], **Fernando Sequeira**[2], **Carlos Brás-Geraldes**[3]

[1] *Faculdade de Ciências, Universidade de Lisboa (FCUL) & Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal, polenepascoal@gmail.com*

[2] *Faculdade de Ciências, Universidade de Lisboa (FCUL) & Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal, fjsequeira@ciencias.ulisboa.pt*

[3] *ISEL - Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa & Centro de Estatística e Aplicações Universidade de Lisboa, Portugal, carlos.geraldes@isel.pt*

Abstract

The assessment of human growth is a crucial tool for understanding health, both at an individual and collective level. Carefully monitoring children from birth allows us to prevent and identify deviations from normal growth, warning of general health problems. The study involves estimating the child growth curve in the central region of Mozambique, using regression models, namely, Generalized Additive Model for Location, Scale, and Shape ($GAMLSS$), with a universal link function for longitudinal data. These estimates will be compared with other regions of Mozambique through the construction of a hypothesis test to evaluate the effect. The current references or standards for child growth curves used in Mozambique are international, recommended by the World Health Organization (WHO) in 2006 and 2007. These curves were constructed based on mixed multicenter and longitudinal studies, grouped in six selected countries: (i) Brazil (Pelotas), (ii) Ghana (Accra), (iii) India (New Delhi), (iv) Norway (Oslo), (v) Oman (Muscat) and (vi) United States (Davis). The main factors determining growth are intrinsic (genetics) and extrinsic (environmental, behavioral and dietary conditions). Due to social

and economic inequalities, it is almost impossible to adequately control extrinsic factors. However, these factors can influence population growth from region to region, calling into question the consistency of current WHO standards. Therefore, the six countries participating in the current growth curves may not necessarily represent the different characteristics of the populations. Therefore, there is a need to construct specific reference curves, essential for evaluating child development in particular contexts. The selection and recruitment of children for the study will be carried out in health units in the study locations, in order to provide a sample of children originating from the same study population. The eligibility criteria applied to mothers and children will be followed, in accordance with the recommendations in WHO multicenter studies. The sample was estimated to be at least 500 children. With this, it is expected to adopt reference curves specific to the Mozambican population, essential for understanding local growth and comparing them with other growth curves. Produce robust tools for constructing specific growth curves, using $GAMLSS$ Models that incorporate MLP as a universal link function component. This is expected to result in a significant improvement in model fit and prediction accuracy, due to MLP's ability to learn complex patterns.

Keywords

Growth curve; $GAMLSS$ method; Longitudinal data; RNA; Multilayer Perceptron.

References

Cole, T. J., Donaldson, M. D., & Ben-Shlomo, Y. (2010). SITAR–a useful instrument for growth curve analysis. International journal of epidemiology, 39(6), 1558–1566. https://doi.org/10.1093/ije/dyq115

Stasinopoulos, D. M., Rigby, R., Heller, G., Voudouris, V., & De Bastiani, F. (2017). Flexible regression and smoothing: using GAMLSS in R. (Chapman and Hall/CRC the R Series). Chapman & Hall/CRC. https://doi.org/10.1201/b21973

de Onis, M., Garza, C., Victora, C. G., Onyango, A. W., Frongillo, E. A., & Martines, J. (2004). The WHO Multicentre Growth Reference Study: planning, study design, and methodology. Food and nutrition bulletin, 25(1 Suppl), S15–S26. https://doi.org/10.1177/15648265040251S103

Computationally efficient segmentation for non-stationary time series

Nicolas Bianco[1], **Lorenzo Cappello**[1], **Eulalia Nualart**[1]

[1] *Universitat Pompeu Fabra, Department of Economics and Business, Spain*
e-mail: nicolas.bianco@upf.edu

Abstract

We propose a novel approach towards joint change-point detection and parameters' learning in non-stationary time series with oscillatory behavior. The latter are approximated by a piecewise model where each segment is a function defined as a mixture of periodic series with multiple unknown frequencies and amplitudes. In this setting, the detection of change-points helps to identify changes in the periodicity of the data. Bayesian inference for the amplitudes is straigthforward within each segment, while the estimation of the frequencies is not trivial due to non-linearities in the model. Here, we propose a simplification of the estimation problem that relies on the definition of a fixed grid of frequencies and implementation of fast Bayesian variable selection to identify the relevant ones within each segment. The price to be paid is the misspecification of the model, since the *true* frequencies may be not contained in the chosen grid. As a possible solution, we propose to alleviate this issue considering fractional posteriors for a more robust Bayesian inference. The segmentation of the series is conducted jointly with the parameters' learning and it exploits the *optimistic search* algorithm for binary segmentation, where we define a suitable information criteria to assess whether a proposed split is actually a change-point. We show through a simulation study that the proposed algorithm is computationally more efficient than state-of-the-art methods, while retaining a good identification of the change-point locations and frequencies.

Keywords

Bayesian variable selection, Change-point detection, Fractional posterior, Non-stationary time series, Optimistic search.

The concurrent effect of meteorological variables on the occurrence of extreme wildfires

P. de Zea Bermudez[1], **S. Pereira**[2], **Mafalda Sebastião**[3], **Carlos C. daCamara**[4] **and Célia M. Gouveia**[5]

[1] *Faculdade de Ciências da Universidade de Lisboa, Departmento de Estatística e Investigação Operacional, Portugal*
CEAUL, Faculdade de Ciências da Universidade de Lisboa, Portugal
pcbermudez@ciencias.ulisboa.pt
[2] *Faculdade de Ciências da Universidade de Lisboa, Departmento de Estatística e Investigação Operacional, Portugal*
CEAUL, Faculdade de Ciências da Universidade de Lisboa, Portugal
sapereira@ciencias.ulisboa.pt
[3] *Instituto Superior Técnico, Departamento de Engenharia de Minas e Georrecursos, Lisboa, Portugal*
mafaldacsebastiao@gmail.com
[4] *Faculdade de Ciências da Universidade de Lisboa, Departamento de Engenharia Geográfica, Geofísica e Energia, Portugal*
Instituto Dom Luiz (IDL), Portugal
cdcamara@ciencias.ulisboa.pt
[5] *Faculdade de Ciências da Universidade de Lisboa, Departamento de Engenharia Geográfica, Geofísica e Energia, Portugal*
Instituto Dom Luiz (IDL), Portugal
Instituto Português do Mar e da Atmosfera, Lisboa, Portugal
celia.gouveia@ipma.pt

Abstract

Compound extremes are a very important problem that is observed in various areas. Whenever several hazards occur, jointly or in cascade, their independent extreme effects may be unimportant, while their simultaneous impact(s) may be devastating. The occurrence of wildfires

in the Mediterranean basin countries is a major societal and environmental concern, which happens every year from June to September. The high temperatures which are observed in the late Spring/early Summer, intensified by low values of humidity, enhance the fire prone conditions. Additionally, the wind speed plays a very important role in fire spreading. Several recent wildfires were characterized by concurrent extreme high temperatures and high winds, namely the event which occurred in Portugal in October 2017 when fires concurred with the winds associated to the passage of the Ophelia storm. In this work, extreme value theory will be used to analyze the effect of some meteorological variables on the occurrence of large wildfires in Portugal.

Keywords

Extreme wildfires, Meteorological variables, Extreme value theory.

Sexual classification based on orthopantomography

João Alves[1], **Cristiana Palmela Pereira**[2], **Rui Santos**[3]

[1] *Master's student in Data Science at Polytechnic of Leiria, Portugal, email.do.alves@gmail.com*
[2] *Faculty of Dental Medicine, University of Lisbon, CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal, cpereira@campus.ul.pt*
[3] *School of Technology and Management, Polytechnic of Leiria, CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal, rui.santos@ipleiria.pt*

Abstract

Craniomandibular bone structures, as they are more resistant to taphonomy processes, are relevant in the sexual diagnosis of adult skeletons. This step is essential in the reconstruction of an unidentified corpse. Hence, based on a sample obtained by students of the Faculty of Dental Medicine of the University of Lisbon through a set of measurements carried out in orthopantomography (panoramic radiographs), this work assesses the performance of different sexual classification methodologies. Some of the evaluated methodologies are based on measurements taken, such as logistic regression, discriminant analysis, k-nearest neighbors (KNN), decision trees, naïve Bayes, support vector machines (SVM), and random forests. Moreover, pre-trained convolutional neural networks (CNN), such as VGG16, RESNET-50 and INCEPTION V-3, were apply directly on the orthopantomographies. The sample was randomly divided into 80% for estimating the parameters of each methodology (train) and the remaining 20% for performance evaluation (test). Performance comparison was based on the confusion matrix and associated measures (accuracy, sensitivity, specificity, predictive values and F-score), and on the area under the ROC curve (AUC).

Keywords

Classification performance, convolutional neural networks, forensic sciences, orthopantomography, sexual classification.

One Class Classification Using Bayesian Optimization

Inyoung Baek[1], **Jaeoh Kim**[2], **Seongil Jo**[3]

[1] *Inha University, Department of Statistics, South Korea,*
biy0322@inha.edu
[2] *Inha University, Department of Data Science, South Korea,*
jaeoh.k@inha.ac.kr
[3] *Inha University, Department of Statistics, South Korea,*
bstatsjo@gmail.com

Abstract

One-class classification (OCC) is a technique used to detect abnormal data by creating a decision boundary that defines normal data, especially in situations where there is an imbalance between normal and abnormal data. One-class support vector machine (OC-SVM) and deep support vector data description (Deep SVDD) are methodologies commonly employed for OCC. OC-SVM aims to find a hyperplane that separates the majority of normal data from the origin in the feature space. On the other hand, Deep SVDD seeks to find the smallest hyper-sphere encompassing the most normal data points. Both OC-SVM and Deep SVDD are sensitive to hyperparameters. The common methodologies for hyperparameter optimization include grid search, random search, and Bayesian optimization. In this paper, we compare the performance of these three hyperparameter optimization methodologies and demonstrate that Bayesian optimization outperforms grid search and random search.

Keywords

Bayesian optimization, One-class classification, One-class support vector machine, Deep support vector data description.

Prediction of felt age for SHARE survey data in COVID 19 waves

Sara Ribeiro Pires

Universidade Aberta, sararipires@uab.pt

Abstract

Elderly people face several challenges arising from physical and psychological changes associated with the natural aging process. With the expected increase in life expectancy, it is feared that the incidence of mental health problems will increase in the elderly and adults. SHARE (Health, Aging and Retirement Survey) is a European project that aims to respond to the call launched by the European Commission to promote research and study the impact of health, social, economic and environmental policies throughout the lives of European citizens. This study seeks to model the self-perception of age of adults aged 50 and over, taking a selection of physical and mental health variables, as well as daily activities. Data is from the special wave of SHARE administered in 2020 and 2021 in relation to COVID-19. Multiple regression is applied with different link functions, individual effects, and interactions included to examine whether a moderator, such as gender, for example, will change the strength of the relationship between the independent and dependent variables. Self-perception of health status, professional situation, as well as the feeling of nervousness in the last month, are among the variables that help predict perception of age in the period considered.

Keywords

Classification, multiple regression, felt age, SHARE project.

References

Ward, R. (2010). How old am I? Perceived age in middle and later life. The International Journal of Aging and Human Development, 71(3), 167–184. doi:10.2190/AG.71.3.a

Sugiyama, K.(2019). Generalized linear models. Springer.

Hierarchical Unbiased Estimation (HUE): Statistical accuracy/computational performance trade-offs with a weighted incomplete U-statistic

Thomas Schatz[1], **Louis Prévot**[2]

[1] *Aix Marseille Univ., CNRS, LIS, France, thomas.schatz@lis-lab.fr (corresponding author)*
[2] *Ecole Centrale Méditerranée, France, louis.prevot@centrale-marseille.fr*

Abstract

We consider the problem of estimating quantities which admit an unbiased estimator, as in the classical theory of U-statistics, but in the case where the data is modeled as a multi-level sample, with a hierarchical dependency structure of arbitrary complexity and with possibly heavy imbalances. This setting is motivated by applications to the evaluation of representation learning algorithms. These algorithms are often evaluated using probing stimuli from relatively large databases with heavily imbalanced multi-level annotations, such as databases of speech recordings (e.g. with phone-level transcriptions nested within phonetic context, speaker identity, speech register, topic, etc.) or images (e.g. with hierarchical annotations of the identity and relative positioning of body or object parts). Popular evaluation metrics for these algorithms sometimes take the form of U-statistics, as in the case of evaluation metrics measuring how well similarities between learned representations match a reference similarity structure (Kriegeskorte et al. 2008, Schatz 2016). Because of the multiple levels of hierarchy and the typically heavy imbalances in the setting considered, properly weighting contributions from various combinations of data subsets may substantially affect the statistical efficiency of estimators. Following this idea, we first obtain a weighted U-statistic whose variance is minimal among a linear set of unbiased estimators and

whose weights are the solution to a linear system involving a number of variance components, kernel sizes and sample sizes, which we carefully define. We then propose a practical scheme to approximate this weighted U-statistic within a fixed computational budget and when the variance components are unknown. Our scheme combines plug-in estimates of the variance components, incomplete computation of the weighted U-statistic's terms and partial optimization of the incomplete U-statistic's weights. We compare the statistical performance of our approach to that of simpler weighting schemes both theoretically and empirically. We also discuss the issue of estimating the variance of the proposed estimator to obtain simple asymptotic confidence intervals.

Keywords

U-statistics, multi-level, hierarchical, weighting scheme, efficient computation.

References

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*

Schatz, T. (2016). ABX-discriminability measures and applications *Doctoral dissertation, Université Paris 6 (UPMC).*

Characterizing Identifiability of Treatment Effects Under Presence of Unobserved Spatial Confounder

Tommy Tang[1], **Xinran Li**[2], **Bo Li**[3]

[1] *University of Illinois in Urbana-Champaign, Statistics, USA, tommymt2@illinois.edu*
[2] *University of Chicago, Statistics, USA, xinranli@uchicago.edu*
[3] *University of Illinois in Urbana-Champaign, Statistics, USA, libo@illinois.edu*

Abstract

Within the fields of both spatial statistics and causal inference, the problem of estimating treatment effects in the presence of unmeasured spatially varying confounding has been garnering increasing attention. One class of methods introduced to mitigate the issue of spatially varying confounding involves highly parametric models, which are used to construct unbiased estimators. A central area of investigation in these models is their identifiability. However, the conditions for uniquely identifiable parameters given an observable data distribution are not well-understood; often the conditions proposed are overly restrictive and unrealistic, or specific to certain explicit choices of covariance function and structures. In this work we aim to clarify properties for identifiability in parametric spatial confounding models of both areal and geostatistical data. In the case of discrete, areal data, we give highly permissive conditions for identifiability under a conditional autoregressive covariance structure. In the case of continuous, geostatistical data, we exhibit general conditions on selection of spatial kernel function families that guarantee model identifiability, and show that these conditions permit commonly used covariance functions such as the exponential, Gaussian, spherical and restricted Matern covariance functions. Finally we apply these results to both areal and geostatistical simulated data.

Keywords

Causal inference, spatial statistics, spatial confounding, identifiability.

Why There Are So Many Contradicted or Exaggerated Findings in Highly-Cited Clinical Research?

Suyu Liu[1], **Mengyi Lu**[1], **Ying Yuan**[1]

[1] *The UT MD Anderson Cancer Center, Department of Biostatistics, USA*

Abstract

It is not uncommon that clinical studies of the same intervention contradicted with each other, e.g., one study produced positive results, while the other produced negative results. In Ioanndis' report, 32% of the highly-cited clinical research studies were contradicted in subsequent large-scale studies or were shown to have potentially overestimated the efficacy of the experimental intervention. We perform Bayesian analysis of these highly-cited clinical studies based on Bayesian factor. We identified one cause of the issue: p values strongly overstated the experimental evidence. For the highly-cited studies, when the p value was 0.05, there was a 74.4% percentage chance that the null hypothesis was true. The use of a p value of 0.05 as the criterion for significance caused many researchers to mistakenly draw conclusions of positive findings, which were then contradicted by subsequent large-scale studies.

Keywords

P value, Bayesian, Significance.

Joined stochastic models for the evaluation of cancer progression from clinical data

Vincent Wieland[1], **Jan Hasenauer**[2]

[1] *University of Bonn, Life and Medical Science Institute, Germany, vwieland@uni-bonn.de*
[2] *University of Bonn, Life and Medical Science Institute, Germany*

Abstract

One of today's foremost challenges in analyzing clinical data is the integration of different data modalities. Patients – especially in oncology – undergo a wide range of diagnostics which yield diverse measurement types including both discrete and continuous quantities, often in a longitudinal setting, as well as time-to-event data. In this project, we develop a computational approach for joining the different stochastic dynamics underlying such measurements into one mathematical model describing patients' disease progression, and demonstrate the use of Bayesian inference for the resulting set of model parameters. To illustrate this, we consider a simple example where one observes the size of a primary tumor, the number of different tissues affected by metastases, and the survival status of a patient. The first is growing continuously and modeled as an ordinary or stochastic differential equation, whereas the second and third are discrete events that are described as Poisson jump processes with intensities dependent on the other components. The resulting multidimensional process together with its discrete observations can be seen as a Hidden Markov Model whose parameters are to be estimated. We examine three different models from this class, which differ in the description of the continuous primary tumor growth, using simulated datasets. For the parameter estimation task we showcase results obtained by the use of maximum-likelihood estimation and Particle Filter algorithms implemented in the Julia programming language. This newly proposed approach of joining different dynamics into one stochastic model creates valuable

groundwork for clinical trajectory analysis, as it allows for a more holistic description of patient outcomes.

Keywords

Stochastic modeling, Hidden Markov Models, Bayesian Inference, Cancer Modeling.

Deep Compositional Models for Nonstationary Extremal Dependence

Xuanjie Shao[1], **Jordan Richards**[2], **Raphaël Huser**[3]

[1] *King Abdullah University of Science and Technology, CEMSE Division, Saudi Arabia, xuanjie.shao@kaust.edu.sa*
[2] *King Abdullah University of Science and Technology, CEMSE Division, Saudi Arabia, jordan.richards@kaust.edu.sa*
[3] *King Abdullah University of Science and Technology, CEMSE Division, Saudi Arabia, raphael.huser@kaust.edu.sa*

Abstract

Modeling the nonstationarity and anisotropy that often prevails in the extremal dependence of spatial data can be challenging. Inference for stationary, and isotropic models, is considerably easier, but the assumptions that underpin these models are not typically met by data observed over large, or topographically-complex, domains. A simple approach to accommodating spatial non-stationarity in Gaussian processes, proposed by Sampson and Guttorp (1992), is to warp the original spatial domain to a latent space where stationarity and isotropy can be reasonably assumed. However, estimation of the warping function can be computationally expensive and the transformation is not guaranteed to be injective, which can lead to physically-unrealistic transformations. Zammit-Mangion et al. (2022) overcame these issues by exploiting deep Gaussian processes, where the transformation is constructed using a deep composition of injective mappings. We present an extension of this methodology to model non-stationarity in extremal dependence of data, by leveraging popularly-applied parametric models for spatial extremal processes. Also, Temporal varying pattern of the dependence is investigated.

Keywords

Nonstationary extremal dependence, spatial extremes, deformation, compositional model, max-stable processes, r-Pareto processes.

References

Sampson, P. D. and P. Guttorp (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association 87*(417), 108–119.

Zammit-Mangion, A., T. L. J. Ng, Q. Vu, and M. Filippone (2022). Deep compositional spatial models. *Journal of the American Statistical Association 117*(540), 1787–1808.

Measures of Uncertainty in Machine Learning: What are they actually quantifying?

Yusuf Sale[1,2], **Paul Hofman**[1,2], **Eyke Hüllermeier**[1,2]

[1] *Institute of Informatics, LMU Munich, Germany*
[2] *Munich Center for Machine Learning (MCML), Germany*

Abstract

Uncertainty quantification has a rich history within traditional statistics and probability theory, yet it has only recently seen an increase of attention in the machine learning community. This work explores and emphasizes the increasing relevance and multifaceted nature of uncertainty quantification within machine learning. As applications of machine learning permeate into critical domains, an in-depth understanding of various methods for quantifying uncertainty becomes vital. We raise the following important question: What do uncertainty measures actually quantify? Beyond distinguishing between aleatoric and epistemic uncertainty, our work extends to identifying various forms of uncertainty assessment, focusing on outcome uncertainty (uncertainty about the actual outcome), decision uncertainty (uncertainty about a decision to be made) and predictive uncertainty (uncertainty about the underlying data generating process itself). Furthermore, it is essential to note that these forms of uncertainty are intricately entangled and can not be assessed in isolation. We examine prevalent uncertainty quantification methods including formalism such as credal sets, Bayesian techniques, and conformal prediction and evaluate which forms of uncertainty are considered. Moreover, we emphasize the importance of specifying the exact form of uncertainty under consideration, as differing techniques cater to diverse forms of uncertainty. Through a structured overview of various forms of uncertainty, this work aims to provide clarity and guide future research, thereby fostering precision and rigor in the application and interpretation of machine learning models.

Keywords

Uncertainty representation, uncertainty quantification, form of uncertainty, decision-making under uncertainty, aleatoric and epistemic uncertainty.

Preliminary Research Results in Application of a Tree Distribution to Bayesian Offline Change Point Detection and Segmentation

Yuta Nakahara[1]

[1] *Waseda University, Center for Data Science, Japan,*
y.nakahara@waseda.jp

Abstract

We recently proposed a probability distribution on tree structures in (Nakahara *et al.*, 2022). It was a conjugate prior distribution for likelihood functions satisfying some conditions. We also proposed its exact and efficient Bayesian estimation algorithms. In this research, we apply them to Bayesian offline change point detection and segmentation. In previous studies of offline change point detection, tree structures are already used to represent recursive binary segmentation, see e.g., (Eckley *et al.*, 2011). However, no prior distribution is assumed on them. In this research, we assume the aforementioned prior distribution on the tree structure representing the recursive binary segmentation. It enables us to estimate change points based on a posterior tree distribution and output a posterior distribution of change points. Our Bayesian tree estimation algorithm works as a subroutine of variational Bayesian methods. In this poster presentation, we show preliminary research results on our Bayesian offline change point detection and segmentation.

Keywords

Tree distribution, change point detection, variational Bayesian method.

References

Nakahara, Y., Saito, S., Kamatsuka, A., and Matsushima, T. (2022). Probability Distribution on Full Rooted Trees. *Entropy*, 24(3), 328. `https://doi.org/10.3390/e24030328`

Eckley, I., Fearnhead, P., and Killick, R. (2011). Analysis of changepoint models. In D. Barber, A. Cemgil, and S. Chiappa (Eds.), *Bayesian Time Series Models* (pp. 205-224). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511984679.011

Chapter 5

Student Awards' Papers

High-dimensional variable clustering based on sub-asymptotic maxima of a weakly dependent random process

Alexis Boulin[1,3], **Elena Di Bernardino**[1], **Thomas Lalo**[1], **Gwladys Toulemonde**[2,3]

[1] *Université Côte d'Azur, CNRS, LJAD, Nice, France*
[2] *Univ Montpellier, CNRS, Montpellier, France*
[3] *INRIA, Lemon*

Abstract

Our work presents a contribution that encompasses both novel modeling techniques and practical algorithmic applications. In the first aspect of our contribution, we introduce a new class of models, called Asymptotic Independent block (AI-block) models. These models are made upon a model-based approach to variable clustering, wherein clusters at the population level are delineated based on the independence of extremes between these clusters. The second facet of our contribution revolves around the development and rigorous evaluation of an algorithm specifically tailored for AI-block models. Moreover, our paper advances the statistical performance of our method. We situate our work within the context of multivariate stationary mixing random processes, aligning with established statistical applications. This framework holds relevance across various domains, including but not limited to finance and climate analysis where temporal dependence is a crucial consideration. To substantiate the practical utility of our proposed AI-block models and associated algorithm, we present two compelling data analyses. These analyses are conducted on real-world data sets from the domains of neuroscience and environmental sciences. Our results underscore the interpretability and scientific relevance of both our clustering model and the algorithmic model, demonstrating their capacity to yield meaningful insight in complex, applied settings.

Keywords

Asymptotic independence, Consistent estimation, Extreme value theory, High dimensional models, Variable clustering.

Optimal designs for testing pairwise differences: a graph based game theoretic approach

Arpan Singh[1], **Satya Prakash Singh**[2], **Ori Davidov**[3]

[1]*Indian Institute of Technology, Hyderabad, India,*
ma20resch01002@iith.ac.in
[2]*Indian Institute of Technology, Kanpur, India,*
singhsp@iitk.ac.in
[3] *University of Haifa, Israel,*
davidov@stat.haifa.ac.il

Abstract

In a variety of experimental setting there is an interest in comparing a subset of pairs–of–treatments. Such experiments usually address one of the following two scientific questions: (1) is there a difference within *any* of the selected pairs of treatments? or, (2) is there a difference within *all* of the selected pairs of treatments? In this article we propose max–min optimal designs for testing the above hypotheses using a graph based game theoretic approach. Some of the max–min designs obtained are well known, but not recognized as optimal, others are novel and provide a substantial improvement over naive designs.

Keywords

Intersection Union Test, Least Favourable Configuration, Multiple Comparison, Max–Min design, Power, Union Intersection Test.

Rank tests for outlier detection

Chiara Gaia Magnani[1], **Aldo Solari**[2]

[1] *University of Milano-Bicocca, Department of Economics, Management and Statistics, Italy, c.magnani9@campus.unimib.it*
[2] *University of Milano-Bicocca, Department of Economics, Management and Statistics, Italy, aldo.solari@unimib.it*

Abstract

In novelty detection, the objective is to determine whether the test sample contains any outliers, using a sample of controls (inliers). This involves many-to-one comparisons of individual test points against the control sample. A recent approach applies the Benjamini-Hochberg procedure to the conformal p-values resulting from these comparisons, ensuring false discovery rate control. In this paper, we suggest using Wilcoxon-Mann-Whitney tests for the comparisons and subsequently applying the closed testing principle to derive post-hoc confidence bounds for the number of outliers in any subset of the test sample. We revisit an elegant result that under a nonparametric alternative known as Lehmann's alternative, Wilcoxon-Mann-Whitney is locally most powerful among rank tests. By combining this result with a simple observation, we demonstrate that the proposed procedure is more powerful for the null hypothesis of no outliers than the Benjamini-Hochberg procedure applied to conformal p-values.

Keywords

Closed testing; Conformal p-values; False discovery rate; Lehmann's Alternative; Wilcoxon-Mann-Whitney test.

Optimal high-dimensional and nonparametric distributed testing under communication constraints

Botond Szabó[1], **Lasse Vuursteen**[2], **Harry van Zanten**[3]

[1] *Department of Decision Sciences, Bocconi University, Bocconi Institute for Data Science and Analytics (BIDSA), Italy, botond.szabo@unibocconi.it*

[2] *Delft Institute of Applied Mathematics (DIAM), Delft University of Technology, The Netherlands, l.vuursteen@tudelft.nl*

[3] *Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands, j.h.van.zanten@vu.nl*

Abstract

We derive minimax testing errors in a distributed framework where the data is split over multiple machines and their communication to a central machine is limited to b bits. We investigate both the d- and infinite-dimensional signal detection problem under Gaussian white noise. We also derive distributed testing algorithms reaching the theoretical lower bounds. Our results show that distributed testing is subject to fundamentally different phenomena that are not observed in distributed estimation. Among our findings, we show that testing protocols that have access to shared randomness can perform strictly better in some regimes than those that do not. We also observe that consistent nonparametric distributed testing is always possible, even with as little as 1-bit of communication and the corresponding test outperforms the best local test using only the information available at a single local machine. Furthermore, we also derive adaptive nonparametric distributed testing strategies and the corresponding theoretical lower bounds.

Keywords

Distributed methods, Nonparametric, Hypothesis testing, Minimax optimal.

Isotonic subgroup selection

Manuel M. Müller[1], **Henry W. J. Reeve**[2], **Timothy I. Cannings**[3], **Richard J. Samworth**[4]

[1] *Statistical Laboratory, University of Cambridge, United Kingdom, mm2559@cam.ac.uk*

[2] *School of Mathematics, University of Bristol, United Kingdom, henry.reeve@bristol.ac.uk*

[3] *School of Mathematics, University of Edinburgh, United Kingdom, timothy.cannings@ed.ac.uk*

[4] *Statistical Laboratory, University of Cambridge, United Kingdom, r.samworth@statslab.cam.ac.uk*

Abstract

Given a sample of covariate-response pairs, we consider the subgroup selection problem of identifying a subset of the covariate domain where the regression function exceeds a pre-determined threshold. We introduce a computationally-feasible approach for subgroup selection in the context of multivariate isotonic regression based on martingale tests and multiple testing procedures for logically-structured hypotheses. Our proposed procedure satisfies a non-asymptotic, uniform Type I error rate guarantee with power that attains the minimax optimal rate up to poly-logarithmic factors. Extensions cover classification, isotonic quantile regression and heterogeneous treatment effect settings. Numerical studies confirm the practical effectiveness of our proposal, which is implemented in the `R` package `ISS`.

Keywords

Subgroup analysis, isotonic regression, post-selection inference, heterogenous treatment effects, superlevel set estimation.

Autoregressive Models for Time Series of Random Objects

Matthieu Bulté[1], **Helle Sørensen**[2]

[1] *University of Copenhagen, Department of Mathematical Sciences, Denmark, mb@math.ku.dk*
[2] *University of Copenhagen, Department of Mathematical Sciences, Denmark, helle@math.ku.dk*

Abstract

Random variables in metric spaces indexed by time and observed at equally spaced time points are receiving increased attention due to their broad applicability. However, the absence of inherent structure in metric spaces has resulted in a literature that is predominantly non-parametric and model-free. To address this gap in models for time series of random objects, we introduce an adaptation of the classical autoregressive model tailored for data lying in a Hadamard space. The parameters of interest in this model are the Fréchet mean and an autocorrelation parameter, both of which we prove can be consistently estimated from data. Additionally, we propose a test statistic and establish its asymptotic normality, thereby enabling hypothesis testing for the absence of autocorrelation. Finally, we introduce a bootstrap procedure to obtain critical values for the test statistic under the null hypothesis. Our theoretical findings are illustrated by numerical studies.

Keywords

Least squares regression, Time Series, Random Objects, Autoregressive model, Metric space.

Lower Complexity Adaptation for Empirical Entropic Optimal Transport

Michel Groppe[1], **Shayan Hundrieser**[2]

[1] *University of Göttingen, Institute for Mathematical Stochastics, Germany, michel.groppe@uni-goettingen.de*
[2] *University of Göttingen, Institute for Mathematical Stochastics, Germany, s.hundrieser@math.uni-goettingen.de*

Abstract

Entropic optimal transport (EOT) presents an effective and computationally viable alternative to unregularized optimal transport (OT), offering diverse applications for large-scale data analysis. In this work, we derive novel statistical bounds for empirical plug-in estimators of the EOT cost and show that their statistical performance in the entropy regularization parameter ε and the sample size n only depends on the simpler of the two probability measures. For instance, under sufficiently smooth costs this yields the parametric rate $n^{-1/2}$ with factor $\varepsilon^{-d/2}$, where d is the minimum dimension of the two population measures. This confirms that empirical EOT also adheres to the *lower complexity adaptation* principle, a hallmark feature only recently identified for unregularized OT. Additionally, we complement our findings with Monte Carlo simulations. Our techniques employ empirical process theory and rely on a dual formulation of EOT over a single function class. Crucial to our analysis is the observation that the entropic cost-transformation of a function class does not increase its uniform metric entropy by much.

Keywords

Optimal transport, convergence rate, metric entropy, curse of dimensionality.

Performance Guaranteed Confidence Sets of Ranks

Onrina Chandra[1], **Tirthankar Dasgupta**[2], **Min-ge Xie**[3]

[1] *Rutgers University,Department of Statistics, USA,*
E-mail: oc152@stat.rutgers.edu
[2] *Rutgers University, Department of Statistics, USA,*
E-mail: tirthankar.dasgupta@stat.rutgers.edu
[3] *Rutgers University, Department of Statistics, USA,*
E-mail: mxie@stat.rutgers.edu

Abstract

Ranks of institutes are often estimated based on estimates of certain latent features of the institutes, and due to sample randomness it is of interest to quantify the uncertainty associated with the estimated ranks. This task is especially important in often-seen "near tie" situations in which the estimated latent features are not well separated among some of the institutes resulting in a nonignorable portion of wrongly ordered estimated ranks. Uncertainty quantification can help mitigate some of the issues and give us a fuller picture, but the task is very challenging because the ranks are discrete parameters and the standard inference methods developed under regularity conditions do not apply. Bayesian methods are sensitive to prior choices while large sample-based methods do not work since the central limit theorem fail to hold for the estimated ranks. In this article, we propose a repro Samples Method to address this nontrivial irregular inference problem by developing a confidence set for the true rank of the institutes. The confidence set obtained has finite sample coverage guarantee and the method can handle difficult near tie cases. The effectiveness of the proposed development is illustrated using simulation studies and a real data example of ranking the performance of 78 US VHA facilities in their service to their diabetic patients.

Keywords

Inference on discrete parameter space; Finite-sample performance guarantee; Repro sample method; Latent model; Rank of performance and causal effects.

Support recovery with knowledge on sparsity structure and non-exchangeable regularization

Paul Rognon Vael[1], **David Rossell**[2], **Piotr Zwiernik**[3]

[1] *Universitat Pompeu Fabra, Department of Business and Economics, Spain, paul.rognon@gmail.com*
[2] *Universitat Pompeu Fabra, Department of Business and Economics, Spain, rosselldavid@gmail.com*
[3] *University of Toronto, Department of Statistical Sciences, Canada, piotr.zwiernik@utoronto.ca*

Abstract

Support recovery refers to the problem of correctly estimating the support set of a vector of parameters $\beta \in \mathbb{R}^p$. It arises for example in model selection for linear regression. In that setting, theoretical results have consistently shown that support recovery is constrained by the relationship between, on one side, the size of nonzero parameters and, on the other side, the dimension p and the size of the support. We observe that in many practical situation there is knowledge that this relationship changes across groups of covariates, creating structure in the sparsity of β. We study informed L_0 and L_1 selectors that incorporate that knowledge by varying the regularization across those groups. We show that informed selectors push the limits on support recovery in linear regression, outperforming their non-informed counterparts under both orthogonal and correlated designs of low and high dimension.

Keywords

high-dimensional statistics, model selection, regularization, sparse regression.

Guided Adversarial Robust Transfer Learning with Source Mixing

Xin Xiong[1], **Zijian Guo**[2], **Tianxi Cai**[1,3]

[1] *Department of Biostatistics, Harvard School of Public Health, USA*
[2] *Department of Statistics, Rutgers University, USA*
[3] *Department of Biomedical Informatics, Harvard Medical School, USA*

Abstract

Many existing transfer learning methods rely on leveraging information from source data that closely resembles the target data. However, this approach often overlooks valuable knowledge that may be present in different yet potentially related auxiliary samples. When dealing with a limited amount of target data and a diverse range of source models, our paper introduces a novel approach, Guided Adversarial Robust Transfer (GART) Learning, that breaks free from strict similarity constraints. GART is designed to optimize the most adversarial loss within an uncertainty set, defined as a collection of target populations generated as a convex combination of source distributions that guarantee excellent prediction performances for the target data. GART effectively bridges the realms of transfer learning and distributional robustness prediction models. We establish the identifiability of GART and its interpretation as a weighted average of source models closest to the baseline model. We also show that GART achieves a faster convergence rate than the model fitted with the target data. Our comprehensive numerical studies and analysis of multi-institutional electronic health records data using GART further substantiate the robustness and accuracy of GART, highlighting its potential as a powerful tool in transfer learning applications.

Keywords

Transfer learning, group distributional robustness, multi-source learning.

Conformalized Matrix Completion

Yu Gui[1], **Rina Foygel Barber**[1], **Cong Ma**[1]

[1] *Department of Statistics, University of Chicago, U.S.A.*

Abstract

Matrix completion aims to estimate missing entries in a data matrix, us- ing the assumption of a low-complexity structure (e.g., low rank) so that imputation is possible. While many effective estimation algorithms exist in the literature, uncertainty quantification for this problem has proved to be challenging, and existing methods are extremely sensitive to model misspec- ification. In this work, we propose a distribution-free method for predictive inference in the matrix completion problem. Our method adapts the frame- work of conformal prediction, which provides confidence intervals with guar- anteed distribution-free validity in the setting of regression, to the problem of matrix completion. Our resulting method, conformalized matrix completion (cmc), offers provable predictive coverage regardless of the accuracy of the low-rank model. Empirical results on simulated and real data demonstrate that cmc is robust to model misspecification while matching the performance of existing model-based methods when the model is correct.

Keywords

Uncertainty quantification, matrix completion, conformal predictive infer- ence, ...

Just Identified Indirect Inference Estimator: Accurate Inference through Bias Correction

Yuming Zhang[1], Yanyuan Ma[2], Samuel Orso[3], Mucyo Karemera[4], Maria-Pia Victoria-Feser[5], Stéphane Guerrier[6]

[1] *University of Geneva, Geneva School of Economics and Management, Switzerland, Yuming.Zhang@unige.ch*
[2] *Pennsylvania State University, Department of Statistics, United States, yzm63@psu.edu*
[3] *University of Geneva, Geneva School of Economics and Management, Switzerland, Samuel.Orso@unige.ch*
[4] *University of Geneva, Geneva School of Economics and Management, Switzerland, Mucyo.Karemera@unige.ch*
[5] *University of Geneva, Geneva School of Economics and Management, Switzerland, Maria-Pia.VictoriaFeser@unige.ch*
[6] *University of Geneva, Geneva School of Economics and Management & Faculty of Science, Switzerland, Stephane.Guerrier@unige.ch*

Abstract

An important challenge in statistical analysis lies in controlling the estimation bias when handling the ever-increasing data size and model complexity of modern data settings. In this paper, we propose a reliable estimation and inference approach for parametric models based on the Just Identified iNdirect Inference estimator (JINI). The key advantage of our approach is that it allows to construct a consistent estimator in a simple manner, while providing strong bias correction guarantees that lead to accurate inference. Our approach is particularly useful for complex parametric models, as it allows to bypass the analytical and computational difficulties (e.g., due to intractable estimating equation) typically encountered in standard procedures. The properties of JINI (including consistency, asymptotic normality, and

its bias correction property) are also studied when the parameter dimension is allowed to diverge, which provide the theoretical foundation to explain the advantageous performance of JINI in increasing dimensional covariates settings. Our simulations and an alcohol consumption data analysis highlight the practical usefulness and excellent performance of JINI when data present features (e.g., misclassification, rounding) as well as in robust estimation.

Keywords

Bias reduction, indirect inference, intractable likelihood function, misclassified logistic regression, weighted maximum likelihood estimator.

Index