2026 ISNPS THESSALONIKI

International Symposium on Nonparametric Statistics

22 - 26 June, 2026 | THESSALONIKI, GREECE

www.easyconferences.eu/isnps2026

Peter Hall Lecture

9:30 - 10:30

Classification and diffusion-induced neural density estimators and simulators for generative AI

J. Fan[1]

[1]Princeton University, United States

Neural network-based methods for conditional density estimation have recently gained substantial attention, as various neural density estimators have outperformed classical approaches in real-data experiments. Despite these empirical successes, implementation can be challenging due to the need to ensure non-negativity and unit-mass constraints, and theoretical understanding remains limited. In particular, it is unclear whether such estimators can adaptively achieve faster convergence rates when the underlying density exhibits a low-dimensional structure. This paper addresses these gaps by proposing a structure-agnostic neural density estimator, called the classification-induced neural density estimator and simulator (CINDES) that is straightforward to implement and provably adaptive, attaining faster rates when the true density admits a low-dimensional composition structure. Another key contribution of our work is to show that the proposed estimator integrates naturally into generative sampling pipelines, most notably score-based diffusion models, where it achieves provably faster convergence when the underlying density is structured. We validate its performance through extensive simulations and a real-data application. We also prove the optimality of score-based diffusion models for density estimation when the target density admits a factorizable, low-dimensional, nonparametric structure in a separate work. The main challenge is that the low-dimensional, factorizable structure no longer holds for most diffused timesteps, and it is very difficult to show that these diffused score functions can be well approximated without a significant increase in the number of network parameters. (Join works with Yihong Gu, Dehao Dai, Mukherjee, and Ximing Li)

Bayesian nonparametrics methods and applications

11:00 - 11:30

Using Bayesian Adaptive Regression Tress for fully nonparametric density regression

J. Griffin[1], M. Kalli[2]

[1]University College London, United Kingdom
[2]King's College London, United Kingdom

In density regression, we wish to model the distribution of a response indexed by covariates. Infinite mixture models have proved an attractive Bayesian nonparametric approach but full support is only possible if there is covariate-dependence in the weights and the parameters of the mixture components. It is challenging to build such models. In this talk, I will discuss how identifiability, variable selection and overfitting can be addressed using Bayesian Adaptive Regression Trees (BART) with well-defined shrinkage priors to provide interpretable methods. I will illustrate their use in a range of applications.

Estimation of cluster-specific causal effects on spatially associated survival data using SoftBART

d. sinha[1], I. Bhattacharya[1]

[1]FLORIDA STATE UNIVERSITY, United States

We propose a novel Bayesian approach to estimate causal effects in spatially clustered survival data. Using Soft Bayesian Additive Regression Trees (SBART), we introduce a nonparametric regression for a log-Normal survival model that accommodates spatial associations among unknown cluster effects through a Directed Acyclic Graph Auto-Regressive (DAGAR) model. We employ a two stage approach, which entails estimating the propensity score in the first step, and incorporating it as a confounder of the outcome model in the second step.

In our simulation study, we compare our method with existing approaches under various simulation scenarios, including both correctly specified and misspecified outcome models, to demonstrate the superior performance of our method. We then apply our method to analyze the causal effect of Treatment Delay (TD) on post-treatment survival of breast cancer patients from the Florida Cancer Registry (FCR).
Our analysis produces the county-specific as well as state-wide assessment of the causal effects while accommodating spatial association among counties.

Complexity bounds for Dirichlet process slice samplers

B. Franzolini[1], F. Gaffi[2]

[1]Bicocca University, Italy
[2]Bergamo University, Italy

Slice sampling is a standard Monte Carlo technique for Dirichlet process (DP)-based models, widely used in posterior simulation. However, formal assessments of the scalability of posterior slice samplers have remained largely unexplored, primarily because the computational cost of a slice-sampling iteration is random and potentially unbounded.
In this work, we obtain high-probability bounds on the computational complexity of DP slice samplers. Our main results show that, uniformly across posterior cluster-growth regimes, the overhead induced by slice variables, relative to the number of clusters supported by the posterior, is Op(log n). As a consequence, even in worst-case configurations, superlinear blow-ups in per-iteration computational cost occur with vanishing probability.
Our analysis applies broadly to DP-based models without any likelihood-specific assumptions, still providing complexity guarantees for posterior sampling on arbitrary datasets.
These results establish a theoretical foundation for assessing the practical scalability of slice sampling in DP-based models.

Adaptively Slice Sampling Mixture Models

A. Rumiancev[1], M. Kalli[1]

[1]King's College London, United Kingdom

Slice-efficient samplers of Kalli, Griffin and Walker (2011) are Markov chain Monte Carlo algorithms used for fitting infinite mixture models with a wide range of stick-breaking priors. These methods rely on the automatic truncation of the infinite mixture through the use of latent uniform random variables that determine an active set of mixture components, required to proceed with the chain. The general representation of the slice-efficient scheme allows for the upper bound of the uniform distribution to be determined by a positive sequence, the choice and parametrization of which is a non-trivial task, as it governs the delicate balance between efficiency and computational time. In order to find this balance, we propose using Bayesian optimization as a method to maximize the relative expected squared jumped distance, an objective function that captures the relationship between the mixing time of the Markov chain and the running cost of the algorithm. Additionally, we present an alternative sampling scheme that uses a single uniform random variate to perform inference and, similarly to the general slice-efficient approach, permits the user to remove the dependence between the stick-breaking weights and the slice variable.

Time series analysis for modern data

11:00 - 11:30

How Weak are Weak Factors? Uniform Inference for Signal Strength in Signal Plus Noise Models

A. Bykhovskaya[1], V. Gorin[2], S. Sodin[3, 4]

[1]Duke University, United States
[2]UC Berkeley, United States
[3]Hebrew University, Israel
[4]Queen Mary University, United Kingdom

In high-dimensional data analysis separating meaningful structure from noise is a central challenge. I will discuss four classical signal-plus-noise models: factor models, spiked covariance matrices, Wigner matrices with low-rank perturbations, and canonical correlation analysis, with a focus on measuring the strength of the signals. Traditional Gaussian approximations provide reliable inference only when signals are strong enough, but they break down near the critical threshold where signal and noise become indistinguishable. I will present a new framework for constructing confidence intervals that remain valid uniformly across strong, weak, and critical regimes. The method is based on a universal transitional distribution and offers a practical tool for applied work. I will illustrate its performance with applications to macroeconomic and financial data.

Uncovering Dynamics in Sparsely Observed Stochastic Processes

A. Aue[1], S. Hörmann[2], M. Ofner[2]

[1]University of California, Davis, United States
[2]Graz University of Technology, Austria

In this talk, we present a statistical framework for modeling the dynamics of stochastic processes based on stochastic differential equations. The model parameters, a drift and a diffusion function, are estimated from sparsely observed replicates of the process. To achieve this, we propose a maximum likelihood procedure for Gauss-Markov processes utilizing a truncated basis expansion. We discuss the asymptotic properties of this approach and demonstrate its practical application on real data. The methodology interfaces with time series and functional data analysis.

Nonparametric Estimation of Predictive Power

D. Strenger-Galvis[1], S. Hörmann[1]

[1]Graz University of Technology, Austria

Quantifying the predictive power of covariates is a fundamental task in statistics, traditionally addressed in linear models using the coefficient of determination. Pearson's correlation ratio offers a natural generalization of this concept beyond linear frameworks. This generalized measure is highly relevant for modern machine learning applications, where the underlying functional relationships are complex and not explicitly specified.

We propose a fully nonparametric estimator for this generalized coefficient. Crucially, our approach enables the efficient estimation of predictive power directly from the data, without the prerequisite of first fitting a predictive model. Because it bypasses the model-fitting stage, this estimator serves as a scalable and objective tool for initial feature selection—a critical advantage given the substantial computational expense associated with training large-scale machine learning models.

Furthermore, we establish the asymptotic distribution of the estimator under the null hypothesis of zero predictive power, which enables the construction of an efficient test.

Nonparametric Foundations of Generative AI

11:00 - 11:30

Nonparametric undirected graphical model selection using diffusion models

M. Chae[1], H.K. Kwon[1]

[1]POSTECH, South Korea

Undirected graphical models provide a fundamental framework for representing conditional independence structures among high-dimensional random variables. Although undirected graphical model selection has become a central problem in high-dimensional statistics and machine learning, most existing methods are restricted to parametric settings such as Gaussian graphical models and Ising models. In this talk, we consider a nonparametric approach to undirected graphical model selection based on diffusion models. Recent work has shown that density estimators based on diffusion models can adapt to the unknown undirected graph structure of the data; however, diffusion models themselves do not provide an explicit estimator of the graph. To address this limitation, we propose a novel method for undirected graphical model selection that does not rely on parametric assumptions.

Provable statistical and computational efficiency of diffusion models

C. Cai[1]

[1]University of Michigan, United States

Score-based diffusion models have emerged as a foundational paradigm for modern generative modeling, achieving remarkable success across diverse applications from image synthesis to scientific computing. Despite their empirical prominence, fundamental questions about their theoretical foundations remain: How efficient can diffusion samplers be? What are the fundamental statistical limits of these samplers? In this talk, I will present recent theoretical advances that address both the computational and statistical frontiers of diffusion models. First, I will introduce a novel accelerated stochastic sampler that provably reduces iteration complexity under minimal assumptions, offering sampling speedups without sacrificing statistical optimality. Second, I will present the first comprehensive end-to-end analysis for deterministic ODE-based samplers, establishing (nearly-)minimax optimal statistical guarantees under mild assumptions on the target distribution. Together, these results provide a rigorous mathematical foundation that narrows the gap between the practical success and theoretical understanding of diffusion models.

Statistical foundations of representation learning in generative models

B. Aragam[1]

[1]University of Chicago, United States

One of the key paradigm shifts in statistical machine learning over the past decade has been the transition from handcrafted features to automated, data-driven representation learning. A crucial step in this pipeline is to identify latent and learn structured representations from data. In many applications, meaningful concepts are not directly observed, and must be learned from data, often using flexible, nonparametric models such as deep generative models. These settings present new statistical and computational challenges that will be focus of this talk. We will re-visit the statistical foundations of nonparametric latent variable models as a lens into the problem of identifying representations in deep generative models. We discuss our recent work on developing methods for identifying and learning causal representations from data with rigourous guarantees, and discuss how even basic statistical properties are surprisingly subtle. Along the way, we will explore the connections between deep generative models, nonparametric latent variable models, and causal graphical models.

Structural inference in high-dimensional models

11:00 - 11:30

Sequential Return to Baseline Testing

P. Serra[1], M. Regis[2]

[1]Vrije Universiteit Amsterdam, Netherlands
[2]TU Eindoven, Netherlands

We consider the problem of detecting a Return to Baseline (RtB) in high-frequency monitoring data preceding and following an intervention, where the aim is to identify the time at which the data-generating distribution realigns with its pre-intervention distribution. We propose a sequential, distribution-free testing procedure that does not rely on specifying a parametric null model and provides anytime-valid error control. The method relies on ideas from universal inference to define a discrepancy measure that is aggregated into a non-negative super-martingale, and is then empirically calibrated to form an E-process. The calibration is performed using the baseline data, and is thus subject specific. We establish finite-sample bounds for the calibration error (under a flexible non-parametric assumption), discuss the impact of tuning parameters and computational complexity, and demonstrate through simulations and a clinical case study that the procedure accurately detects RtB in quasi-periodic monitoring data.

Change-point detection In High-Dimensional Vector Autoregression

F. Enikeeva[1], O. Klopp[2], M. Rousselot[1]

[1]University of Poitiers, France
[2]ESSEC, CREST, France

I will talk about the problem of change-point testing in high-dimensional vector autoregression (VAR). We propose a test for a change in the transition matrix of the VAR process under low-rank assumptions on the transition matrix and show its minimax-rate optimality in some regimes. The proposed test is adaptive to an unknown change-point location without any assumption about the minimum spacing between the change-point and the boundaries of the observation interval.

Anytime-Valid Tests for Sparse Anomalies

R. Castro[1], M. Pérez-Ortiz[1], I. Stoepker[1]

[1]Eindhoven University of Technology, Netherlands

We consider the problem of testing sequentially for the presence of sparse anomalies among a large number of data streams. To this end, we design and analyze Anytime-Valid (AV) tests, which retain type-I error control at arbitrary stopping times. Existing results address exclusively the nonsequential case, which exhibits a subtle phase transition between two regimes where tests are either powerless or powerful. In our sequential setting, we argue, two challenges arise: (1) the standard analysis of AV tests cannot be executed in the relevant sample-size regime; and (2) standard constructions of parameter-adaptive AV tests are either analytically intractable or computationally unfeasible. This work addresses these challenges. Borrowing insights from the nonsequential literature, we propose a framework to analyze AV tests and their shortest possible sample sizes. Under this framework, we show that, in the Gaussian location setting, the oracle AV test has a delicate threshold behavior that is related to---but not implied by---the phase transition observed in optimal nonsequential tests. Our main result is a computationally efficient, parameter-adaptive AV test; we show that it achieves the same threshold behavior as the oracle AV test. Numerical simulations illustrate these theoretical findings (joint work with M. F. Pérez-Ortiz and I. V. Stoepker)

Improving variable selection properties with data integration and transfer learning.

P. Rognon-Vael[1]

[1]Bocconi University, Italy

We study variable selection (also called support recovery) in high-dimensional sparse linear regression when one has external information on which variables are likely to be associated with the response. Consistent recovery is only possible under somewhat restrictive conditions on sample size, dimension, signal strength, and sparsity. We investigate how these conditions can be relaxed by incorporating said external information. A key application that we consider is structural transfer learning, where variables selected in one or more source datasets are used to guide variable selection in a target dataset. We introduce a family of likelihood penalties that depend on the external information, motivated by connections to Bayesian variable selection. We show that these methods achieve variable selection consistency in regimes where any method ignoring external information fails, and that they achieve consistency at faster rates. We first quantify the potential gains under ideal, oracle-chosen, penalties.
We then propose computationally efficient empirical Bayes procedures that learn suitable penalties from the data. We prove that these procedures have improved variable selection properties compared to methods that do not use external information. We illustrate our approach using simulations and a genomics application, where results
from mouse experiments are used to inform variable selection for gene expression data in humans.

Nonparametric Copula-Based Methods for Dependence Modeling

11:00 - 11:30

Kendall’s tau-based inference for gradually changing dependence structures

F. Camirand Lemyre[1], J. Quessy[2]

[1]Université de Sherbrooke, Canada
[2]Université du Québec à Trois-Rivières, Canada

Suppose that a sequence of random pairs (X1,Y1),..., (Xn,Yn) is subject to a gradual change in the sense that for K1 ≤K2 ∈{1,...,n}, the joint distribution is F before K1, G after K2, and gradually moving from F to G between the two times of change K1 and K2. This setup elegantly generalizes the abrupt-change model that is usually assumed in the change-point analysis. Under this configuration, asymptotically unbiased estimates of Kendall’s tau up to and after the change are proposed, as well as tests and estimators of change points related to these measures. The asymptotic behaviour of the introduced estimators and test statistics is rigorously investigated, in particular by demonstrating a general result on weighted indexed U-statistics computed under a heterogeneous pattern. A simulation study is conducted to examine the sampling properties of the proposed methods under different scenarios of change in the dependence structure of bivariate series. An illustration is given on a time series of monthly atmospheric carbon dioxide concentrations and global temperature for the period 1959–2015.

A Unified Multivariate L_p Regression Model with Skewed Exponential Power Marginals and Copula Dependence Modelling

K. Oualkacha[1], Y. Touijer[1], F. Badaoui[2]

[1]Université du Québec à Montréal, Canada
[2]Institut National de Statistique et d'Economie Appliquée, Morocco

Quantile and expectile regression provide complementary views of conditional distributions, yet both face important limitations under skewness, heavy tails, and multivariate dependence. We develop a unified multivariate L_p regression framework with Skewed Exponential Power Distribution (SEPD) marginals linked via copulas. The SEPD shape parameter p enables a continuous transition between quantile-like ($p \approx 1$) and expectile-like ($p \approx 2$) targets, while a skewness parameter τ flexibly captures asymmetry.

To ensure stable model parameter estimation over $p \in [1,2]$, we introduce a Bernstein-smoothed asymmetric L_p loss that is convex and twice continuously differentiable. This construction supports efficient gradient-based optimization while preserving sensitivity to tail behavior. Regression, scale, and dependence parameters are estimated, with uncertainty quantified using sandwich-robust standard errors when available or bootstrap methods otherwise.

Monte Carlo experiments with trivariate responses show low bias, rapid mean squared error decay, and near-nominal coverage. The proposed approach substantially outperforms margin-wise p-expectile methods that ignore dependence. An application to a Dutch boys cohort study demonstrates smoothly varying covariate effects across L_p levels and strong joint fit.

Tests of independence and randomness for arbitrary data using copula-based covariances

B. Nasri[1]

[1]Université de Montréal, Canada

In this talk, we discuss tests of independence for data with arbitrary distributions in both non-serial and serial settings. These tests are constructed using copula-based covariances and their multivariate extensions via Möbius transforms. We examine their asymptotic properties, present results from numerical experiments, and illustrate the methodology with a data-driven example.

Conditional independence tests between arbitrary time series

B. Remillard[1, 2], B. Nasri[3], K. Ghoudi[1]

[1]United Arab Emirates University, United Arab Emirates
[2]HEC Montreal, Canada
[3]Universite de Montreal, Canada

Generalized innovations are defined and associated with generalized error models having arbitrary distributions, i.e., mixtures of continuous and discrete distributions. The main novelty of this article is to be able to test for conditional independence between several time series with arbitrary distributions. Families of empirical processes and their Moebius transformations are defined from lagged generalized errors, with asymptotically independent non-parametric Gaussian distributions. Several test statistics are then proposed based on Cramer von Mises-type statistics and dependence measures, as well as graphical methods to visualize the dependence. In addition, numerical experiments are performed to assess the power of the proposed tests. All developed methodologies are implemented in the CRAN package IndGenErrors.

Extremes, Statistical learning and Applications

11:00 - 11:30

Fixed-k Inference for Explosive Drift

Y. He[1], J. Li[2], Y. Li[3], Y. Zhu[2]

[1]Eastern Institute of Technology, Ningbo, China
[2]Singapore Management University, Singapore
[3]University of Manchester, United Kingdom

We propose a new framework for uniform inference on explosive drifts in high-frequency data. Standard large-bandwidth asymptotics often fail in this context because the spot test statistics computed over short windows are far from Gaussian. By treating the window size k as fixed, we show that the spot statistics are coupled with a sequence of dependent Student-t variables, and their maximum converges to a Frechet distribution rather than the conventional Gumbel limit. We establish a novel anti-clustering condition for dependent Student-t processes to justify this limit theory under overlapping estimation windows. A local power analysis reveals that explosive drifts induce a multiplicative power transformation of the limiting distribution, contrasting with the additive location shift characteristic of Gaussian theory. Empirically, we show that the proposed coupling-based test offers superior size control and reveals that statistically significant intraday price explosions are far rarer than suggested by Gaussian-based methods.

Polar depth and anomaly detection in heavy tailed data

S. Clémençon[1], C. Fernández[1], P. Mozharovskyi[1], A. Sabourin[2]

[1]LTCI, Telecom Paris, Institut Polytechnique de Paris, France
[2]Université Paris Cité, ENS Paris Saclay, Centre Borelli, France

Motivated by the analysis of the behaviour of extremes from multivariate heavy-tailed distributions, we introduce a novel notion of statistical depth, referred to as ``polar depth''. The polar depth function is naturally expressed in polar coordinates, as is the limiting distribution of a regularly varying r.v. beyond asymptotically large thresholds, once its marginals have been appropriately normalized. Not only does the polar depth function make it easy to order the extreme values taken by a heavy-tailed random variable X and finds natural applications in anomaly detection, but it is also possible to show that the polar depth of the largest observations, i.e., observations X whose norm exceeds a threshold t, converges to the polar depth of the limiting distribution as t goes to infinity. Although designed to quantify the depth of multivariate extremes, the polar depth is interesting in its own right. In particular, this notion is more relevant than present in the literature alternatives (including the half-space depth) for distributions whose support is included in a half-space. We demonstrate its properties and analyze statistical issues related to its estimation from both finite-sample and asymptotic points of view. We present numerical results to empirically demonstrate its relevance, particularly for the statistical analysis of extreme observations and more specifically for the identification of anomalies among them.

Likelihood-based inference for the block maxima method in time series

S. Padoan[1], D. Carl[2], S. Rizzelli[3]

[1]Università Bocconi, Italy
[2]Bocconi University, Italy
[3]University of Padova, Italy

We study likelihood-based inference within the Block Maxima (BM) framework for stationary time series. While Bayesian methods under the BM approach have been extensively investigated in the independent setting, a rigorous asymptotic theory for dependent data is currently lacking. To address this gap, we first develop a comprehensive likelihood theory for the misspecified Generalized Extreme Value (GEV) model under serial dependence. Building on this foundation, we establish the asymptotic properties of Bayesian inference for the GEV parameters, the extremal index, T-horizon return levels, and extreme quantiles (Value at Risk). Under general prior conditions, we prove posterior consistency, $\sqrt{k}$-contraction rates, Bernstein–von Mises results, and asymptotic coverage of credible intervals. For inference on the extremal index, we introduce an adjusted posterior distribution that corrects the poor coverage of a naive Bayesian approach. Simulation results demonstrate strong finite-sample performance of the proposed methodology.

Trends in tail dependence of heteroscedastic extremes

J. Einmahl[1], C. Zhou[2]

[1]Tilburg University, Netherlands
[2]Erasmus University Rotterdam, Netherlands

We consider multivariate extreme value statistics for independent but nonidentically distributed random vectors. In particular, the data may have varying tail copulas and also heteroscedastic marginal distributions. Assuming smoothly changing tail copulas, we propose a nonparametric estimator for the integrated tail copula, as well as one for the local tail copula. We establish the asymptotic behavior of both estimators. Notably, the heteroscedastic marginals do not affect the limiting processes. Finally we use the main result for the integrated tail copula to test for a constant tail copula across all observations.

Non- and semiparametric methods for biomedical applications

11:00 - 11:30

A Bayesian Nonparametric Approach for Semi-Competing Risks with Application to Cardiovascular Health

M. Daniels[1], K. Gelis Cadena[1], J. Siddique[2]

[1]University of Florida, United States
[2]Northwestern University, United States

We address causal estimation in semi-competing risks settings, where a non-terminal event may be precluded by one or more terminal events. We define a principal-stratification causal estimand for treatment effects on the non-terminal event, conditional on surviving past a specified landmark time. To estimate joint event-time distributions, we use both vine-copula constructions and Bayesian nonparametric enriched Dirichlet-process mixtures (EDPM), enabling inference under minimal parametric assumptions. We index our causal assumptions with sensitivity parameters. We illustrate the proposed method using data from a cardiovascular health study.

AI-enabled Model for Predicting Type 1 Diabetes Progression

S. Wu[1]

[1]University of South Florida, United States

The progression of type 1 diabetes (T1D) is characterized as a transition from autoimmune activity (Stage 1) to dysglycemia (Stage 2) and ultimately to the clinical onset of symptomatic diabetes (Stage 3). The most used prediction tools and risk scores in the T1D literature are based on regression models that assume linear and additive relationships between risk factors and T1D outcomes. However, as more studies acknowledged the interaction effects between risk factors, there has been a growing recognition of the advantage of machine learning methods that can flexibly capture these complexities and potentially improve predictive accuracy. We investigated multi-modal models to predict the progression from autoimmunity to T1D based on a unified latent representation for each patient by integrating demographic, autoantibody, metabolic, and genotyping data from the TrialNet Pathway to Prevention study. The AI prediction models may facilitate more personalized patient risk assessments and monitoring recommendations and improve the development of new interventions by enhancing the efficiency of clinical trials.

Semiparametric High-Dimensional Joint Model

S. Basu[1], S. Sahu[2], J. Sun[1]

[1]University of Illinois Chicago, United States
[2]M.D. Anderson Cancer Center, University of Texas, United States

We consider the setting of joint modeling of a time-to-event outcome and a large number of longitudinally measured processes that are posited to prognosticate the outcome. The literature on joint modeling is diverse; however, current approaches can typically jointly analyze one or a few longitudinal processes. In modern precision medicine, the ability to update a patient's risk profile in real-time is crucial. The motivating application for this work comes from a study on age-related macular degeneration (AMD), a disease that affects 12.6% of US adults aged ≥ 40 years (19.6 million) and more than 190 million people globally. We develop a novel regularized joint model that can handle high-dimensional longitudinal biomarkers. We also develop a computationally efficient algorithm and a two-stage selection procedure. In simulation studies and an application to AMD, we demonstrate that the proposed approach improves both biomarker discovery and performance compared to existing joint models.

The Curious Case of Informative Cluster Size: Testing for Informativeness in Three-Level Designs with An Application to Periodontal Disease

S. DATTA[1]

[1]University of Florida, United States

Multilevel data are frequently encountered in biomedical research, and several statistical methods have been developed to analyze such data. Informativeness of the number of units on certain levels often manifests itself in multilevel data analysis and failure to account for this phenomenon will lead to biased inference. Moreover, utilizing an incorrect marginalization approach will also lead to invalid conclusions. To identify the appropriate marginal distribution to be tested in multilevel designs, we propose a sequential testing procedure to test for informativeness of unit sizes in multilevel structures with three levels. At a given level of the design, a bootstrap method is developed to estimate the null distribution of no informativeness of unit size. Simulation studies confirm the efficacy of our sequential procedure in maintaining an overall Type I error rate. Additionally, we extend our testing procedure to a multilevel regression setting, enhancing its practical applicability. We demonstrate the utility of our proposed methods through the analysis of data from a study on periodontal disease and a study on stress levels of preschoolers.

Recent developments in online/sequential procedures

11:00 - 11:30

The Root Finding Problem Revisited: Beyond the Robbins-Monro procedure

Y. Yu[1], M. Banerjee[1], Y. Ritov[1]

[1]University of Michigan, United States

We introduce Sequential Probability Ratio Bisection (SPRB), a novel stochastic approximation algorithm that adapts to the local behavior of the (regression) function of interest around its root. We establish theoretical guarantees for SPRB's asymptotic performance, showing that it achieves the optimal convergence rate and minimal asymptotic variance even when the target function's derivative at the root is small (at most half the step size), a regime where the classical Robbins-Monro procedure typically suffers reduced convergence rates. Further, we show that if the regression function is discontinuous at the root, Robbins-Monro converges at a rate of whilst SPRB attains exponential convergence. If the regression function has vanishing first-order derivative, SPRB attains a faster rate of convergence compared to stochastic approximation. As part of our analysis, we derive a nonasymptotic bound on the expected sample size and establish a generalized Central Limit Theorem under random stopping times. Remarkably, SPRB automatically provides nonasymptotic time-uniform confidence sequences that do not explicitly require knowledge of the convergence rate. We demonstrate the practical effectiveness of SPRB through simulation results.

Precise Regret and Adaptive Inference in Multi-Armed Bandits

C. Zhang[1], Q. Han[1], K. Khamaru[1]

[1]Rutgers University, United States

Despite extensive research over the widely used UCB algorithms in multi-armed bandits, a precise understanding of their regret behavior remains elusive. This gap has not only hindered the evaluation of their actual algorithmic efficiency, but also limited further developments in statistical inference in sequential experiments. We bridge these fundamental aspects through a deterministic characterization of the number of arm pulls for an UCB index algorithm. The resulting precise regret formula not only accurately captures the actual behavior of the UCB algorithm for finite time horizons and individual problem instances, but also provides significant new insights into the regimes not covered by existing theoretical frameworks. The deterministic characterization of the number of arm pulls for the UCB algorithm also has major implications in adaptive statistical inference. We show that the UCB algorithm satisfies certain `stability' properties that lead to quantitative central limit theorems in two settings: for the empirical means of unknown rewards in the bandit setting, and for a class of Ridge estimators when the arm means exhibit a structured relationship through covariates. Our technical approach relies on an application of a new comparison principle between the UCB algorithm and its noiseless, continuous-time minimax counterpart. We expect this new principle to be broadly applicable for general UCB index algorithms.

Data-mixing in LLM pretraining

Y. Sun[1]

[1]University of Michigan, United States

Data mixing is one of the most consequential yet poorly understood engineering challenges in LLM training. The problem is deceptively hard: there is no single objective to optimize against, optimal mixtures shift with model and dataset scale, and pre-training metrics often fail to predict downstream capabilities that matter. In this talk, I present a (sequential) optimization formulation of the data mixing problem, and how we tackled some of the challenges when training the K2-series of LLMs.

Stochastic regret guarantees for online zeroth-andfirst-order bilevel optimization

G. Michailidis[1]

[1]University of Florida, United States

Online bilevel optimization (OBO) is a powerful framework for machine learning problems where both outer and inner objectives evolve over time, requiring dynamic updates. Current OBO approaches rely on deterministic\textit {window-smoothed} regret minimization, which may not accurately reflect system performance when functions change rapidly. In this work, we introduce a novel search direction and show that both first-and zeroth-order (ZO) stochastic OBO algorithms leveraging this direction achieve sublinear {stochastic bilevel regret without window smoothing}. Beyond these guarantees, our framework enhances efficiency by:(i) reducing oracle dependence in hypergradient estimation,(ii) updating inner and outer variables alongside the linear system solution, and (iii) employing ZO-based estimation of Hessians, Jacobians, and gradients. Experiments on online parametric loss tuning and black-box adversarial attacks validate our approach.

Contributed: Time Series, Spectra and Dynamic Dependence

11:00 - 11:20

Nonparametric two sample test of spectral densities

I. Nadin[1], T. Krivobokova[1], F. Enikeeva[2]

[1]University of Vienna, Austria
[2]University of Poitiers, France

A novel nonparametric test for the equality of the covariance matrices of two Gaussian stationary processes, possibly of different lengths, is proposed. The test translates to testing the equality of two spectral densities and is shown to be minimax rate-optimal. Test performance is validated in a simulation study, and the practical utility is demonstrated in the analysis of real electroencephalography data. The test is implemented in the R-package sdf.test.

Empirical energy distance for locally stationary processes

C. Beering[1,2], M. Armillotta[3], K. Fokianos[4]

[1]Helmut Schmidt University, Germany
[2]Otto von Guericke University, Germany
[3]University of Rome Tor Vergata, Italy
[4]University of Cyprus, Cyprus

Energy distance as introduced by Székely and Rizzo (2013) is used to examine equality of distributions. Hence, its wide field of application ranges from testing for symmetry over decomposition of distances to clustering. In addition to that, it is connected to distance covariance and distance correlation. More recently, Davis et al. (2023) proposed a methodology to cluster multivariate time series data based on energy distance. At it, they focussed on stationary time series fulfilling certain mixing conditions. We seize on that idea and transfer it to locally stationary processes. Hereby, we do not restrain ourselves to points in time. Instead, we incorporate a time-span component, which can be used to cover the whole time period of interest as well as only contiguous parts of it. After showing consistency of the empirical version of energy distance over time for time-varying linear processes, we investigate the limit distributions. Conclusory, our theoretical findings are illustrated by varied numerical results.

Singular Spectrum Analysis Revisited: Frequency Recovery, Spectral Consistency, and Window Length Selection

G. Martos[1], P. Poncela[2], D. Fresoli[2]

[1]Universidad Torcuato Di Tella, Argentina
[2]Universidad Autónoma de Madrid, Spain

Singular Spectrum Analysis (SSA) is a nonparametric method for time series analysis. Via the Singular Value Decomposition on the so--called trajectory matrix, or equivalently, by diagonalizing the empirical second moment matrix, it decomposes a time series into quasi--orthogonal components that aim to maximize variance. The resulting trendlines provide natural estimates of latent structures such as trend, cycles, and noise. However, in contrast with spectral methods, the components extracted are not intrinsically associated to particular frequencies. This paper introduces a related technique that reconciles frequency identification with variance--based decomposition. As a byproduct, our decomposition yields a consistent estimator of the spectral density. A key parameter in SSA is the window length, which critically influences the resulting decomposition and its interpretation. We provide practical guidance for selecting the window length based on asymptotic results and adapt well--established inferential tools to group components, thereby enabling the identification of statistically significant signals associated with specific frequencies. The performance of the proposed methodology is illustrated through simulation studies and empirical applications to temperature and high-frequency electricity consumption data, where meaningful latent structures are successfully identified.

Quantifying linear instantaneous Granger causality

G. Gkyzis[1], D. Kugiumtzis[1]

[1]Aristotle University of Thessaloniki, Greece

Granger causality constitutes a foundational framework for the identification of directional dependencies in time series. A wide range of extensions has been developed to characterize lag causal effects and, more recently, to incorporate both lagged and instantaneous interactions. Nonetheless, existing approaches remain limited in their capacity to consistently identify and disentangle instantaneous causal effects in multivariate systems. Such effects may arise from direct contemporaneous interactions, latent common drivers, or practical constraints whereby the temporal resolution of the observed data is coarser than that of the underlying causal dynamics. These considerations imply that instantaneous causality, particularly in the presence of latent variables, may manifest across arbitrary subsets of system variables and should therefore be interpreted as a system-level property [1]. To address this limitation, we introduce Extended Granger Causality for Instantaneous effects (EGCI), a novel linear metric for the detection and quantification of instantaneous causality in multivariate time series. We develop a comprehensive methodological framework encompassing reduced-form vector autoregressive estimation, structural vector autoregressive parameterization and identification, and the associated statistical inference procedures underpinning EGCI. The EGCI test statistic is shown to follow a Fisher–Snedecor distribution under the null hypothesis, enabling closed-form significance testing, while a non-parametric surrogate-based procedure is additionally developed to ensure valid inference under potential model misspecification. The proposed measure is evaluated, compared against an existing nonlinear metric for instantaneous causality, and benchmarked using both simulated and empirical datasets. The results indicate that EGCI enables reliable identification of linear instantaneous causal structure, thereby extending the scope of Granger-based causality analysis to multivariate systems exhibiting contemporaneous interactions.

[1] C. Koutlis, V. K. Kimiskidis, and D. Kugiumtzis, "Identification of hidden sources by estimating instantaneous causality in high-dimensional biomedical time series," International Journal of Neural Systems, vol. 29, no. 4, p. 1850051, 2019.

Detection of Higher-Order Interactions in Complex Dynamical Systems

A. Fotiadis[1], I. Vlachos[1], D. Kugiumtzis[1]

[1]Aristotle University of Thessaloniki, Greece

The Partial Mutual Information from Mixed Embedding (PMIME) is a non-parametric, information-based method for detecting direct Granger-causal effects in multivariate time series, searching for the most informative subset of driving lag variables for a given response. However, due to its reliance on scalar candidate terms (the lag variables), PMIME searches only for pair correlations and fails to identify higher-order interactions. Such interactions are the purely synergistic interactions, where predictive information about a target arises only through combinations of variables. Recent extensions of PMIME to search also for pair of lag variables addressed this limitation, yet they ignore the conditional structure induced by the progressively constructed embedding.

We propose an extension of PMIME for the detection of conditional synergies, integrating the evaluation of multivariate candidate terms directly into the embedding procedure. At each iteration, both scalar and vector candidate drivers are assessed for their mutual information to the response, given the current embedding. This enables the identification of higher-order interactions that emerge only under specific lag configurations and are undetectable through marginal or pairwise analysis.

The method is evaluated on simulated dynamical systems with controlled conditional synergy structures, across varying time series lengths and system dimensions.

Recent advances in Bayesian nonparametrics

14:30 - 15:00

Multiplex exchangeability for stochastic block models

V. Ghidini[1], B. Franzolini[2], F. Gaffi[3], D. Durante[4]

[1]Università della Svizzera Italiana, Italy
[2]University of Milano Bicocca, Italy
[3]University of Bergamo, Italy
[4]Bocconi University, Italy

Multiplex networks encode multiple types of relationships among a common set of nodes and arise in a wide range of applications, including neuroscience and criminology. A fundamental challenge in modeling such data is to identify a notion of exchangeability that appropriately reflects the multiplex structure, preserving node identity across layers while allowing for heterogeneity in connectivity patterns.
In this work, we introduce a notion of multiplex exchangeability for random network arrays, which formalizes the probabilistic symmetries inherent to multiplex data and ensures coherent inference across both node and layer dimensions. Building on this principle, we develop a class of Bayesian stochastic block models that are consistent with multiplex exchangeability and that jointly infer layer-specific and global community structures. The proposed construction relies on a novel prior on multivariate partitions based on conditionally partially exchangeable random partitions induced by hierarchical species sampling processes. This formulation enables principled borrowing of information across layers at the node level, automatic learning of the number of communities at both local and global scales, and analytical tractability.
We study the dependence structure induced by the model and characterize the symmetries propagated to the observable adjacency arrays. An efficient algorithm for posterior inference is developed. Simulation studies demonstrate accurate recovery of multiscale community structures under heterogeneous scenarios, and an application illustrates the ability of the method to disentangle persistent global communities from layer-specific connectivity patterns.

Bayesian nonparametric approach for ranking and selection problems

J. Griffin[1], M. Kalli[2]

[1]University College London, United Kingdom
[2]King's College London, United Kingdom

League tables, where performance is ranked based on some measures, are common in many fields from sports to education, healthcare to finance. These tables are used to inform decisions, and so a lot rests in accurately quantifying the uncertainty of the statistics/estimates used for ranking performance. We consider methods where the underlying performance measure is generated from some unknown distribution and we focus on estimating this distribution using a Bayesian non parametric approach.

Estimation of heterogeneous treatment effects via Pitman-Yor processes with spike-and-slab baseline

S. Bianchi[1], A. Lijoi[1], I. Pruenster[1]

[1]Bocconi University, Italy

The talk proposes a Pitman-Yor process prior, with a spike-and-slab baseline measure, for Bayesian nonparametric estimation of heterogeneous treatment effects in a setting with multiple treatments. The baseline measure has a spike at 0 that accounts for the control group, while the slab component is a Gaussian process (GP). We emphasize the design stage, focusing on matching to improve balance in the distribution of confounders across treatment groups and thereby support causal interpretation. To address the possible high dimensionality of the confounder space, we summarize covariates via generalized propensity scores (GPS) and incorporate matching directly into the GP covariance structure through a newly proposed stationary kernel. Finally, we derive theoretical guarantees on the optimal number of matches per unit, balancing computational tractability with accurate imputation of missing potential outcomes.

Exact finite mixture representations for species sampling processes

R. Mena[1]

[1]UNAM-IIMAS, Mexico

Discrete random probability measures are central to Bayesian inference, particularly as priors for mixture modeling and clustering. A broad and unifying class is that of proper species sampling processes (SSPs), encompassing many Bayesian nonparametric priors. We show that any proper SSP admits an exact conditional finite-mixture representation by augmenting the model with a latent truncation index and a simple reweighting of the atoms, which yields a conditional random finite-atom measure whose marginalized distribution matches the original SSP. This yields at least two consequences: (i) distributionally exact simulation for arbitrary SSPs, without user-chosen truncation levels; and (ii) posterior inference in SSP mixture models via standard finite-mixture machinery, leading to tractable MCMC algorithms without ad hoc truncations. We explore these consequences by deriving explicit total-variation bounds for the conditional approximation error when this truncation is fixed, and by studying practical performance in mixture modeling, with emphasis on Dirichlet and geometric SSPs.

Network and Tensor Time Series

14:30 - 15:00

Copula tensor count autoregressions

M. Armillotta[1], P. Gorgi[2], A. Lucas[2]

[1]University of Rome Tor Vergata, Italy
[2]Vrije Universiteit Amsterdam, Netherlands

This paper presents a novel copula-based autoregressive framework for multi-layer arrays of integer-valued time series with tensor structure. Our framework generalizes recent advances in tensor time series models for real-valued data to a context that accounts for the unique properties of integer-valued data, such as discreteness and non-negativity. The model incorporates feedback effects for the counts' temporal dynamics and allows for low-rank structures of the parameter matrices. An asymptotic theory is developed for a Two-Stage Maximum Likelihood Estimator (2SMLE) for the model's parameters. The estimator balances the challenges of dimensionality, interdependence of the different count series, and computational stability. Together, this substantially pushes the frontier for modeling multi-dimensional, structured tensor time series of counts. An application to tensor crime counts demonstrates the practical usefulness of the proposed methodology.

Tensor time series change-point detection in cryptocurrency network data

A. Anastasiou[1], I. Cribben[2]

[1]University of Cyprus, Cyprus
[2]University of Alberta, Canada

Financial fraud has been growing exponentially in recent years. The rise of cryptocurrencies as an investment asset has simultaneously seen a parallel growth in cryptocurrency scams. To detect possible cryptocurrency fraud, previous research focused on the detection of changes in the network of trades, however, scammers are now trading across multiple cryptocurrency platforms, making their detection more difficult. Hence, it is important to consider the identification of changes across several trading networks or a ‘network of networks’ over time. To this end, in this talk, we will propose a new change-point detection method in the network structure of tensor-variate data. This new method, firstly employs a tensor decomposition, and secondly detects multiple change-points in the second-order (cross-covariance or network) structure of the decomposed data. It allows for change-point detection in the presence of frequent changes of possibly small magnitudes and is computationally fast. We will present results on simulated data as well as to a cryptocurrency data set, which consists of network tensor-variate data from the Ethereum Blockchain.

A Bayesian Dynamic Latent Space Model for Weighted Networks

m. iacopini[1], R. Casarin[2], A. Peruzzi[2]

[1]Luiss University of Rome, Italy
[2]Ca' Foscari University of Venice, Italy

A new dynamic latent space eigenmodel (LSM) is proposed for weighted temporal networks.
The model accommodates integer-valued weights, excess of zeros, time-varying node positions (features), and time-varying network sparsity.
The latent positions evolve according to a vector autoregressive process that accounts for lagged and contemporaneous dependence across nodes and features, a characteristic neglected in the LSM literature.
A Bayesian approach is used to address two of the primary sources of inference intractability in dynamic LSMs: latent feature estimation and the choice of latent space dimension.
We employ an efficient auxiliary-mixture sampler that performs data augmentation and supports conditionally conjugate prior distributions.
A point-process representation of the network weights and the finite-dimensional distribution of the latent processes are used to derive a multi-move sampler in which each feature trajectory is drawn in a single block, without recursions.
This sampling strategy is new to the network literature and can significantly reduce computational time while improving chain mixing.
To avoid trans-dimensional samplers, a Laplace approximation of the partial marginal likelihood is used to design a partially collapsed Gibbs sampler.
Overall, our procedure is general, as it can be easily adapted to static and dynamic settings, as well as to other discrete or continuous weight distributions.

Structured High-Dimensional Inference

14:30 - 15:00

A Bayesian Proof of Talagrand's Majorizing Measure Theorem

I. Zadik[1]

[1]Yale University, United States

We will discuss a short Bayesian proof of Talagrand’s celebrated majorizing-measure theorem (MMT). Unlike previous approaches, our proof does not rely on existing Gaussian processes lower bounds techniques, nor on combinatorial, geometric, or coding-theoretic constructions. Instead, we derive the lower bound from two Bayesian area identities for Gaussian additive/sequence models.

A novel statistical approach to analyze image classification

J. Chen[1], S. Langer[2], J. Schmidt-Hieber[3]

[1]Xiamen University, China
[2]Ruhr University Bochum, Germany
[3]University of Twente, Netherlands

The recent statistical theory of neural networks focuses on nonparametric denoising problems that treat randomness as additive noise. Variability in image classification datasets does, however, not originate from additive noise but from variation of the shape and other characteristics of the same object across different images. To address this problem, we introduce a tractable model for supervised image classification. While from the function estimation point of view, every pixel in an image is a variable, and large images lead to high-dimensional function recovery tasks suffering from the curse of dimensionality, increasing the number of pixels in the proposed image deformation model enhances the image resolution and makes the object classification problem easier. We introduce and theoretically analyze three approaches. Two methods combine image alignment with a one-nearest neighbor classifier. Under a separation condition, it is shown that perfect classification is possible. The third method fits a convolutional neural network (CNN) to the data. We derive a rate for the misclassification error that depends on the sample size and the complexity of the deformation class. An empirical study corroborates the theoretical findings.

Estimation of discrete distributions in relative entropy, and the deviations of the missing mass

J. Mourtada[1]

[1]ENSAE, France

We consider the problem of estimating a distribution over a finite alphabet from an i.i.d. sample, with accuracy measured in relative entropy (Kullback-Leibler divergence). While optimal bounds on the expected risk are known, high-probability guarantees remain less well-understood. First, we characterize the performance of the classical Laplace (add-one) estimator, obtaining matching upper and lower bounds on its performance and establishing its optimality among confidence-independent estimators. We then characterize the minimax-optimal high-probability risk and show that it is achieved by a simple confidence-dependent smoothing technique. Notably, the optimal non-asymptotic risk incurs an additional logarithmic factor compared to the ideal asymptotic rate. Next, motivated by modern regimes in which the alphabet size exceeds the sample size, we discuss methods that adapt to the sparsity of the underlying distribution. We introduce an estimator using data-dependent smoothing, for which we establish a high-probability risk bound depending on two effective sparsity parameters. As part of our analysis, we also derive a sharp high-probability upper bound on the "missing mass", namely the total probability of symbols that do not appear in the sample.

New Frontiers in Inference for Complex Data

14:30 - 15:00

Inference in monotone regression with applications to Lipschitz regression

A. Kuchibhotla[1]

[1]Carnegie Mellon University, United States

In this talk, I will present a general asymptotic theory for least squares estimation of monotone regression function that allows for arbitrary local behavior. I will also describe an adaptive uniformly valid confidence interval construction. Finally, I will show how inference in monotone regression allows for inference in Lipschitz model. Presentations draws from work with Soham Mallick, Siddharth Sarkar, and Kenta Takatsu.

ON HIGH-DIMENSIONAL CHANGE-POINT DETECTION BASED ON PAIRWISE DISTANCES

B. Banerjee[1]

[1]National University of Singapore, Singapore

In change-point analysis, one aims at finding the locations of abrupt distributional changes (if any) in a sequence of multivariate observations. In this article, we propose some nonparametric methods based on averages of pairwise distances for this purpose. These distance-based methods can be conveniently used for high-dimensional data even when the dimension is much larger than the sample size (i.e., the length of the sequence). We carry out some theoretical investigations on the behaviour of these methods not only when the dimension of the data remains fixed and the sample size grows to infinity, but also in situations where the dimension diverges to infinity while the sample size may or may not grow with the dimension. Several high-dimensional datasets are analyzed to compare the empirical performance of these proposed methods against some state-of-the-art methods.

Causal partial identification via optimal transport

Z. Gao[1], Z. Gao[1]

[1]University of Southern California, United States

In causal inference, only one of multiple potential outcomes is observed for each unit, leaving many causal quantities only partially identified (PI). This missingness closely parallels the optimal transport (OT) problem, where marginal distributions are observed but the joint coupling between them is unknown. In this talk, we formalize this connection by casting the causal partial identification problem within the OT framework, enabling the PI sets to be analyzed using tools from the rapidly developing optimal transport literature. When treatment effects are heterogeneous, incorporating covariate information can further sharpen the PI sets. We then generalize the OT formulation to a conditional optimal transport (COT) framework and develop new statistical tools for COT to extend the analysis of PI sets. We conclude by discussing real-world applications and future research directions.

Separating Phase from Amplitude: Geodesic Two-Sample Permutation Tests for Multivariate Functional Data Under the Latent Deformation Model

C. Carroll[1]

[1]University of San Francisco, United States

Distinguishing phase variation (when features occur) from amplitude variation (how large those features are) remains a central challenge in two-sample inference for multivariate functional data. We develop geodesic two-sample tests for multivariate functional data under the Latent Deformation Model (LDM), which decomposes each functional component into a shared latent curve together with subject-specific warping functions, amplitude scalars, and population-level component transport functions. This representation separates timing and amplitude variation, enabling targeted inference on distinct sources of functional variability.

We propose three tests: 1. a component transport test comparing population-level component timing structure across groups via geodesic distances; 2. a subject warping test comparing distributions of subject-specific warping functions using a geodesic energy statistic on the square-root velocity manifold; and 3. an amplitude test comparing multivariate amplitude distributions using Euclidean energy distance. We establish asymptotic size control and consistency for each. Applied jointly as a diagnostic battery, they identify whether group differences arise from population timing structure, individual timing variability, or amplitude variation, or any combination of the three. We illustrate the methods with an application to testing phase and amplitude differences in boys' and girls' growth curves on multiple growth dimensions using data from the Zürich Longitudinal Growth Study.

Specification Testing

14:30 - 15:00

Empirical Likelihood Goodness-of-fit Test for Panel Data Models with Interactive Fixed Effects

J.M. Rodriguez-Poo[1], L.A. Arteaga-Molina[1]

[1]Universidad de Cantabria, Spain

This paper develops an empirical likelihood specification test for panel data models with interactive fixed effects. The test evaluates whether a parametric functional form provides an adequate fit in the presence of cross-sectional dependence induced by unobserved common factors. We establish the asymptotic normality of the test statistic under the null hypothesis and local alternatives, and propose a wild bootstrap procedure to improve finite-sample performance. The empirical likelihood framework offers two key advantages over existing approaches: automatic studentization without requiring variance estimation, and an asymptotic distribution free of nuisance parameters. We also develop a computationally efficient alternative test statistic that preserves more signal variation and exhibits substantially higher power in small to moderate samples. Monte Carlo simulations demonstrate good size control and power comparable to the benchmark test of Su et al. (2020), while avoiding computational costs of variance estimation. In an empirical application to the Environmental Kuznets Curve using panel data from 136 countries (1990--2020), we find no evidence against a linear specification for the relationship between CO_2 emissions and GDP per capita, contradicting the inverted U-shaped pattern predicted by the Environmental Kuznets Curve hypothesis.

Model Checks in a Kernel Ridge Regression Framework

Y. Li[1]

[1]Xi'an Jiaotong-Liverpool University, China

We propose new reproducing kernel-based tests for model checking in conditional moment restriction models. By regressing estimated residuals on kernel functions via kernel ridge regression (KRR), we obtain a coefficient function in a reproducing kernel Hilbert space (RKHS) that is zero if and only if the model is correctly specified. We introduce two classes of test statistics: (i) projection-based tests, using RKHS inner products to capture global deviations, and (ii) random location tests, evaluating the KRR estimator at randomly chosen covariate points to detect local departures. The tests are consistent against fixed alternatives and sensitive to local alternatives at the $n^{-1/2}$ rate. When nuisance parameters are estimated, Neyman orthogonality projections ensure valid inference without repeated estimation in bootstrap samples. The random location tests are interpretable and can visualize model misspecification. Simulations show strong power and size control, especially in higher dimensions, outperforming existing methods.

Kernel-Based Specification Testing with High-Dimensional Nuisance Parameters

J.C. Escanciano[1]

[1]Universidad Carlos III de Madrid, Spain

We develop kernel-based specification tests for semiparametric models defined by conditional moment restrictions in the presence of high-dimensional and machine-learned nuisance parameters. The proposed statistics are built from orthogonalized Reproducing Kernel Hilbert Space (RKHS) embeddings and combined with cross-fitting, yielding inference that is first-order invariant to regularization and slow nuisance convergence. We establish Gaussian approximations and multiplier bootstrap validity, with limiting laws matching the oracle case. Beyond size control, we introduce diagnostics that quantify the effective degree of overidentification in high-dimensional environments, providing practical measures of moment informativeness for specification testing.

A goodness-of-fit test for the latency in a mixture cure model with covariates

W. González-Manteiga[1], M.D. Martínez Miranda[2], I. Van Keilegom[3], M. Conde-Amboage[1]

[1]Universidade de Santiago de Compostela, Spain
[2]Universidad de Granada, Spain
[3]KU Leuven, Belgium

In classical survival analysis, it is assumed that all individuals will eventually experience the event of interest. However, in many situations a subset of subjects never experiences the event and is therefore considered “cured,” with infinite survival time. This phenomenon is addressed using cure models.

Throughout this talk, a general goodness-of-fit test is proposed 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. The asymptotic behaviour of the test statistic will be derived and to calibrate the test in practice it is suggested a bootstrap method. In addition, an extensive simulation study and a real data application will be presented to show the performance of the new proposal in practice.

Nonparametric Estimation for Complex Systems

14:30 - 15:00

Model selection methods for efficient nonparametric estimation in L_2

E. Pchelintsev[1], M. Leshchinskaya[1], S. Pergamenchtchikov[2]

[1]Tomsk State University, Russia
[2]Rouen University, France

The problem of nonparametric estimation of an unknown function in a continuous-time regression model is considered. The observed process is described by a stochastic differential equation with semimartingale noise of small intensity. The estimation quality is measured using robust quadratic risk. An adaptive setting of the problem is considered when the smoothness parameters are unknown. A model selection procedure based on oracle inequalities is proposed. Using the sharp oracle inequality and Pinsker's method for obtaining a lower bound for robust risks, the asymptotic efficiency of the proposed procedure is established. The Lévy and Ornstein-Uhlenbeck processes are considered as examples of semimartingale noises.

Association study in metric space with survival outcome

I. Rodionov[1]

[1]University of Essex, United Kingdom

From high-throughput sequencing to neuroimaging, the clinical data for medical research become increasingly complex, challenging the statistical analysis in a great manner. During the past decades, deep learning contributed vitally in utilizing image data and presented convincing performance in prediction tasks. However, its lack of interpretation is a major concern for the development of clinical diagnosis and intervention. In this work, we propose an analytical method for the association study between objects in the metric space and time-to-event survival outcome, where the predictor can be a vector of high-dimensional non-Euclidean elements with non-negligible interactions. We propose MRBcov and its sample estimator to measure the dependency between a non-Euclidean object and an observed survival outcome. The properties of this measure, as well as the consistency and asymptotic properties of its estimator, are provided. Simulation studies show the promising performance of our method in identifying the association between a complex predictor and a survival outcome. Real datasets are analysed for illustration.

Nonparametric Learning Non-Gaussian Quantum States

L. Markovich[1], X. Liu[1], J. Tura[1]

[1]Leiden university, Netherlands

Continuous-variable quantum systems are foundational to quantum computation, communication, and sensing. While traditional representations using wave functions or density matrices are often impractical, the tomographic picture of quantum mechanics provides an accessible alternative by associating quantum states with classical probability distribution functions called tomograms.
Despite its advantages, including compatibility with classical statistical methods, the tomographic method remains underutilised due to a lack of robust estimation techniques. This work addresses this gap by introducing a non-parametric kernel quantum state estimation (KQSE) framework for reconstructing quantum states and their trace characteristics from noisy data, without prior knowledge of the state. In contrast to existing methods, KQSE yields estimates of the density matrix in various bases, as well as trace quantities such as purity, higher moments, overlap, and trace distance, with a near-optimal convergence rate of $\tilde{O}\bigl(T^{-1}\bigr)$, where T is the total number of measurements. KQSE is robust for multimodal, non-Gaussian states, making it particularly well suited for characterizing states essential for quantum science.

Threshold selection for nonparametric estimation of extremal index by discrepancy method

N. Markovich[1], I. Rodionov[2]

[1]V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Russia
[2]University of Essex, United Kingdom

We propose the discrepancy method as a new threshold data-driven selection tool for the nonparametric estimation of the extremal index of stochastic processes. The discrepancy method was introduced to estimate the probability density function in Vapnik, Markovich, Stephanyuk 1992, where Kolmogorov-Smirnov (K-S) and the Cramer-von Mises-Smirnov (C-M-S) statistics were used as measures in the space of distribution functions. The K-S distance
between the empirical distribution of the upper k observations and the power-law distribution was used to find k in the Hill estimator of the extreme value index in Clauset et al. 2009, Wan et al. 2020 (the minimum distance method). We consider the estimation of the extremal index by the discrepancy method based on the C-M-S statistic that is calculated only by the k largest order statistics instead of the entire sample, Markovich and Rodionov 2023.
The order statistics correspond to the set of interexceedance times normalized by the tail function introduced in Ferro and Segers 2003. A normalization of the discrepancy statistic is proposed due to the usage of the k largest order statistics to provide the coincidence of its asymptotic distribution to the limit distribution of the C-M-S statistic as k tends to infinity. The consistency for known and estimated extremal index and the convergence rate of the extremal index estimators coupled with the discrepancy method regarding k are presented. The discrepancy method is oriented to threshold-based estimators of the extremal index like intervals and K-gaps ones. The method can be applied to other estimators to find only one parameter like the block size as a function of the threshold. The algorithm of the discrepancy method is provided. The exposition is accompanying by the simulation study which demonstrates the best accuracy of the K-gaps estimator together with the discrepancy method and of the application to real data.

Statistics on Non-Euclidean spaces

14:30 - 15:00

Support of Continuous Smeary Distributions on Spheres

S. PAL[1]

[1]Vrije Universiteit Brussel, Belgium

We investigate the support of smeary, directionally smeary, and finite sample smeary probability measures mu with densities on unit m-spheres.

First, we establish support conditions ensuring non-smeariness. In the rotationally symmetric case, we show that a distribution is non-smeary whenever its support lies in a geodesic ball centered at the Fréchet mean of radius R_m, where R_m=pi/2+O(1/m). In the general case, non-smeariness holds whenever the support is contained in a closed ball of radius pi/2.

Second, we prove sharpness of this threshold. For every positive epsilon, we show there exists m_0 depending on epsilon such that for all m at least m_0(epsilon), there exists a rotationally symmetric continuous smeary probability measure on unit m-sphere whose support lies in a ball of radius pi/2+epsilon around the Fréchet mean.

Third, in every dimension we construct directionally smeary continuous distributions supported in a ball of radius pi/2+\epsilon whose Fréchet function has Hessian of rank one.

Finally, we study finite sample smeariness. We show that any continuous non-smeary distribution supported in a geodesic ball of radius \pi/2 is necessarily Type I finite sample smeary, i.e. its variance modulation satisfies stays above 1 asymptotically. Lastly, in the rotationally symmetric case, we prove a curse-of-dimensionality phenomenon: the variance modulation increases with the dimension and can become arbitrarily large depending on the support. Time permitting, we may also discuss some relevant statistical tests.

Symmetry-preserving Geodesic Regression on Lie Groups for Longitudinal Medical Imaging

C. von Tycowicz[1], M. Hanik[1], J. Schade[1]

[1]Zuse Institute Berlin, Germany

Many medical imaging tasks require statistical modeling of continuous transformations, including longitudinal anatomical shape change and articulated skeletal motion. These transformations naturally live on Lie groups, where meaningful statistical analysis should respect group symmetries to remain invariant to arbitrary coordinate choices and reference frames.
In this talk, I will present a geodesic regression framework on Lie groups for longitudinal imaging data. Common approaches rely on Riemannian metrics, but many Lie groups do not admit a metric fully compatible with the group structure. This mismatch breaks symmetry and leads to unstable regression estimates. We therefore introduce a non-metric, bi-invariant estimator that is equivariant under both left and right group actions.
We evaluate the method on synthetic data and on an open-access clinical dataset of longitudinal knee joint configurations acquired for osteoarthritis research. The proposed approach yields stable trajectories and reproducible statistical conclusions, while state-of-the-art Riemannian methods exhibit sensitivity and instability. These results highlight the practical advantages of symmetry-preserving statistical modeling in longitudinal medical imaging studies.

An application of the holonomic gradient method for a model in 3-D shape analysis

A. Kume[1], T. Sei[2], A. Wood[3]

[1]University of Kent, United Kingdom
[2]University of Tokyo, Japan
[3]Australian National University, Australia

The Holonomic gradient method(HGM) has been shown to be a powerful tool in exactly evaluating the intractable normalising constants of probability distributions. While the principle is general enough to be applied to a wide class of distributions, its implementation is in many cases challenging. This is because a good understanding of the differential structure of the relevant likelihood function is essential. In this talk, we introduce the general methodology and apply it to a particular distribution, motivated by applications in 3D shape analysis, where the Pfaffians for constructing the resulting ordinary differential equations are derived. We show that the corresponding normalising constant has rank 8 and utilise this for showing that the HGM
seems to perform well.

Groupwise Image Registration and Template Generation Preserving Individual Features

M. Glock[1], T. Hotz[1]

[1]TU Ilmenau, Germany

Images of objects may be seen as non-Euclidean data as they typically lack an intrinsic coordinate system; so, to facilitate statistical analyses, they need to be put in a common coordinate system. Typically, this is achieved by registering the images at a template – which is a kind of mean image (after registration). From that mean, individual images vary in two essentially different ways – "horizontally" by variation in shape and "vertically" by variation in image intensity. These modes of variation are hard to differentiate between, and to complicate matters further, some images may exhibit features not present in most of the others, leading to implausible deformations when matched to a common template. We therefore propose the use of a spatially weighted distance measure that prioritises the alignment of common features while preserving individual ones. It is implemented within a diffeomorphic, stationary vector field framework and utilizes a multiscale strategy to handle initial misalignments. We demonstrate our approach on the 2D MNIST dataset, on 3D brain MRI data, and on a synthetic benchmark with known ground truth.

Recent Developments in Functional Methods

14:30 - 15:00

Nonparametric functional regression via signature transforms from rough path theory

S. Dabo[1, 2], C. Frévent[1]

[1]University of Lille, France
[2]Inria Datavers, France

In this talk, we study nonparametric regression and classification for function-valued data using a kernel estimator that incorporates the signature transform from rough path theory, enabling to encode sequential data via iterated integrals. Our method uses components extracted from signatures within a kernel regression setting. We derive finite-sample convergence guarantees and analyze finite-sample behavior on both synthetic and real datasets, covering tasks such as learning regression functions and classifying functional observations

Modeling Spatio-Temporal Smoothness using Nonlinear Differential Operators

A. Clemente[1], A. Palummo[1], E. Arnone[2], J. Aston[3], F. Panzica[4], L. Sangalli[1]

[1]Politecnico di Milano, Italy
[2]Università degli Studi di Torino, Italy
[3]University of Cambridge, United Kingdom
[4]Fondazione IRCCS Istituto Neurologico Carlo Besta, Italy

This work introduces a novel smoothing technique for functional data observed over space and time, incorporating a regularization term based on a time-dependent nonlinear partial differential equation (PDE). The proposed approach extends existing physics-informed statistical models, which have so far focused primarily on linear PDE-based penalties. By incorporating nonlinear PDEs into the regularization term, this method significantly broadens the applicability of smoothing models, while also presenting new theoretical and computational challenges.
Simulation studies highlight the potential of integrating additional physical knowledge into statistical modeling, with promising implications for real-world applications. The work presents an initial exploration in the context of neuroimaging data, where the nonlinear PDE regularization is used to model the progression of Alzheimer's disease, based on longitudinal Positron Emission Tomography imaging.

Staged Physics-Informed State Reconstruction for Reinforcement Learning under Partial Observability

V. Biancacci[1], P. Libin[1]

[1]Vrije Universiteit Brussel, Belgium

Reinforcement Learning (RL) methods are typically formulated within the framework of Markov Decision Processes (MDPs), which assume full observability of the environment's state. However, many real-world control problems are inherently partially observable.

In this work, we propose a staged physics-informed state reconstruction framework for reinforcement learning in pandemic mitigation problems. It addresses partial observability by decoupling trajectory-level system identification from RL agent training. In the first stage, noisy and partially observed trajectories of the pandemic are collected and reconstructed using physics-informed neural networks (PINNs), which infer hidden state variables and (time-varying) system parameters while enforcing known system dynamics. These reconstructed trajectories are then used to train an offline RL agent and recover the optimal policy.

Functional compositional regression models with application to cause-specific mortality in Italy

M. Stefanucci[1]

[1]University of Rome. - Tor Vergata, Italy

In this work we introduce a novel class or regression models for functional compositional data, a framework that simultaneously captures multivariate functional structures and compositional constraints often observed in biological and demographic studies. Our approach includes several model variants and estimation algorithms, leveraging a constrained B-splines basis to ensure interpretability and computational efficiency. We demonstrate the utility of these models through an application to cause-specific mortality rates across Italian regions, uncovering common temporal patterns and the influence of socio-economic factors. The results highlight the potential of functional compositional regression as a flexible and informative tool applicable across diverse fields.

Contributed: Functional Data, Density Regression and Dimension Reduction

14:30 - 14:50

Sufficient Dimension Reduction for Conditional Quantiles for Functional Data

E. Christou[1], E. Solea[2], S. Wang[1], J. Song[3]

[1]University of North Carolina at Charlotte, United States
[2]Queen Mary University of London, United Kingdom
[3]Korea University, South Korea

Functional data analysis is an important research area with the potential to transform numerous fields. However, existing work predominantly relies on the more traditional mean regression methods, with surprisingly limited research focusing on quantile regression. Furthermore, the infinite dimensional nature of the functional predictors necessitates the use of dimension reduction techniques. Therefore, in this work, we address this gap by developing dimension reduction techniques for the conditional quantiles of functional data. We derive the convergence rates of the proposed estimators and demonstrate their finite sample performance using simulation examples and a real dataset from fMRI studies.

Nonparametric quantile regression in the fully functional model

T. Greger[1], M. Birke[1]

[1]Universität Bayreuth, Germany

Compared to mean regression, quantile regression has several advantages. Concerning quantile regression, we look at a double functional model, i.e. a conditional model where both variables are functional. In functional settings the definition of a quantile itself is not straightforward. This problem arises because ordering is not possible for multivariate or even functional elements, which yields that the Lebesgue measure does not exist for infinite dimensions. In this talk, we discuss the projection of functional random variables onto one of their directions, when the direction is either fixed or random. Asymptotic properties are given for a quantile estimator, which is based on a smoothed kernel estimator of the distribution function. The performance of the estimator in finite samples is in addition checked in a simulation study.

A novel approach to covariate selection in the asynchronous functional concurrent model

L. Freijeiro González[1], M. Febrero Bande[2, 3], W. González Manteiga[2, 3]

[1]Universidade de Vigo, Spain
[2]Centro de Investigación y Tecnología Matemática de Galicia (CITMAGA), Spain
[3]Universidade de Santiago de Compostela, Spain

A functional concurrent model arises when both the response Y and the covariates X depend on a common argument
t and are related pointwise. In some settings, data can be observed asynchronously, with Y(t) and X(t) recorded at different time points for each individual. This asynchronous observation scheme complicates the analysis, as many standard methods are designed for aligned data. Existing approaches often rely on structural assumptions, such as linearity or additivity, which may be restrictive, particularly with multiple covariates.

In this work, we propose a novel covariate selection procedure for asynchronous functional concurrent models that avoids such constraints. The method reduces dimensionality by identifying and removing irrelevant covariates while remaining flexible with respect to the underlying model form. It is based on conditional distance correlation (Wang et al., 2015), which enables the measurement of conditional dependence in a fully nonparametric way.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional distance correlation. Journal of the American Statistical Association, 110(512):1726–1734.

Density-valued time series: Nonparametric density-on-density regression

H.L. Shang[1], F. Ferraty[2]

[1]Macquarie University, Australia
[2]University of Toulouse, France

This paper is concerned with forecasting probability density functions. Density functions are nonnegative and have a constrained integral; thus, they do not constitute a vector space. Implementing unconstrained functional time-series forecasting methods is problematic for such nonlinear and constrained data. A novel forecasting method is developed based on a nonparametric function-on-function regression, where both the response and the predictor are probability density functions. The asymptotic properties of our nonparametric regression estimator are established, as well as its finite-sample performance, through a series of Monte-Carlo simulation studies. Using Bovespa intraday 5-minute returns and age-specific period life tables from the United States, we assess and compare the finite-sample forecast accuracy of the proposed method with several existing methods.

Scalar-on-density regression with sparsely observed densities

J. Feeser[1], S. Greven[1]

[1]Humboldt-Universität zu Berlin, Germany

We present a novel method to estimate regression models for scalar outcomes with a density-valued regressor, when only sparsely sampled draws from the regressor density are observed, but not the density itself. To properly account for the constrained nature of density functions, our model assumes a functional-linear relationship using the inner product of the latent true regressor density with a functional (density-valued) coefficient in the sense of the so-called Bayes Hilbert space of densities.
The centered log-ratio (clr) transformations of latent regressor densities are modeled as realizations of a finite-dimensional Gaussian process, and the functional regression coefficient is represented using an appropriately constrained finite-dimensional spline basis. For estimation of the model, we use a Monte-Carlo Expectation-Maximization (MCEM) algorithm employing an importance sampling procedure for which we derive an efficient proposal density. To ensure computational feasibility, we use a truncated Karhunen-Loève decomposition of the clr-transformed regressor densities.
We apply our method to a motivating application from education research using data from the Student Teacher Achievement Ratio (STAR) study. Specifically, we test for the presence of peer- effects in education by estimating how students' test scores are influenced by the distribution of their class mates' previous test scores.

Modern nonparametric Bayesian statistics

8:30 - 9:00

A Bayesian Nonparametric Framework for Dynamic Item-Response Theory

M. De Iorio[1]

[1]National University of Singapore, Singapore

Item-response theory (IRT) is widely used for the statistical analysis of questionnaire data, allowing for the differentiation of respondent profiles and the characterisation of questionnaire items through interpretable parameters. However, conventional IRT models are typically cross-sectional and limited in their ability to capture complex longitudinal and hierarchical data structures.
We propose a Bayesian semiparametric extension of IRT that introduces temporal dependence across repeated questionnaire administrations, accommodates repeated measurements, and jointly models responses from related subject groups (e.g., mothers and children) to enable information sharing across hierarchies. The framework further incorporates covariate information, allows for the joint modelling of questionnaire data with other longitudinal markers, and supports clustering of subjects based on their latent response profiles.
Our approach is built on Bayesian nonparametric priors, specifically the Dirichlet Process and the Normalized Generalized Gamma Process, facilitating the identification of clinically meaningful subgroups within the population.
We demonstrate the utility of the proposed methodology through the analysis of longitudinal psychometric questionnaire data collected from mothers and their children, aiming to investigate how various factors influence growth trajectories, developmental outcomes, and mental health. This application, using data from the Singaporean GUSTO cohort study, highlights the potential of our modelling strategy to provide a nuanced understanding of child development by capturing complex dependencies in questionnaire data as well as the relationships between psychometric measures and other growth markers.

Constructing stationary time series of CRM's via Bayesian conjugacy

R. ARGIENTO[1], A. Colombi[2], J. Griffin[3]

[1]University of Bergamo, Italy
[2]Bocconi University, Italy
[3]University College London, United Kingdom

A flexible approach to constructing stationary time-dependent processes builds on the concept of conjugacy within a Bayesian framework. In this setting, the transition law of the process is defined as the predictive distribution of an underlying Bayesian model. When conjugacy holds, the transition kernel can be derived analytically, which makes the approach particularly attractive.
We pursue such analytical tractability in the context of completely random measures (CRMs), focusing on cases where the time-dependent variables themselves are CRMs. To exploit conjugacy, we consider the broad class of exponential-family CRMs, which yields a simple autoregressive representation and provides a foundation for extensions to more complex forms of temporal dependence.
The proposed process can be readily used to extend CRM-based Bayesian nonparametric models, such as feature allocation models, to time-dependent data. Applications span various modern domains, from computer science to biology. In particular, we develop a dependent latent feature model for dynamic image feature identification and a time-evolving Poisson factor analysis for topic modeling, both of which are demonstrated on synthetic and real datasets.

Conditional Dirichlet Processes and Functional Condition Models

J. Lee[1], K. Lee[2], J. Lee[3], S. Jo[4]

[1]Seoul National University, South Korea
[2]Chonnam National University, South Korea
[3]SK innovation, South Korea
[4]Inha University, South Korea

In this paper, we study the conditional Dirichlet process (cDP) when a functional of a random distribution is specified. Specifically, we apply the cDP to the functional condition model, a nonparametric model in which a finite-dimensional parameter of interest is defined as the solution to a functional equation of the distribution. We derive both the posterior distribution of the parameter of interest and the posterior distribution of the underlying distribution itself. We establish two limiting theorems for the posterior: one as the total mass of the Dirichlet process parameter tends to zero, and another as the sample size tends to infinity.
We consider two specific models—the quantile model and the regression model—and propose algorithms for posterior computation, accompanied by illustrative data analysis examples.
As a byproduct, we show that, in the quantile model, the Jeffreys substitute likelihood emerges as the limit of the marginal posterior under a cDP prior, thereby providing a theoretical justification for its use in this context.

Moment Condition Models with Conditional Dirichlet Process Priors

S. Jo[1], K. Lee[2], K. Lee[3], J. Lee[4]

[1]Inha University, South Korea
[2]Chonnam National University, South Korea
[3]Sungkyunkwan University7, South Korea
[4]Seoul National University9, South Korea

A moment condition model is a type of nonparametric model in which the parameter of interest is characterized through a set of moment equations. Such models allow one to focus on finite-dimensional parameters of interest without fully specifying the infinite-dimensional aspects of the data-generating process. The class of moment condition models is broad, encompassing standard moment models, linear regression models, nonlinear regression models, and instrumental variable models. In this paper, we propose a method for computing the posterior distribution of a moment condition model under the conditional Dirichlet process prior introduced by Lee et al. (2025). Previous approaches have relied on importance sampling based on the Bayesian bootstrap posterior, which can be computationally efficient but becomes challenging to implement in high-dimensional parameter spaces. We address this limitation by developing a posterior sampling scheme using the constrained Hamiltonian Monte Carlo method. The proposed approach is applicable to a wide range of moment condition models and scales effectively to high-dimensional settings. We demonstrate the method through applications to multiple datasets across various moment condition model specifications.

High-dimensional Time Series

8:30 - 9:00

Trend Filtering with Fractional Splines

T. Proietti[1]

[1]University of Rome Tor Vergata, Italy

The paper introduces a signal plus noise model, where the signal is a fractional spline process. The latter is generated by a linear combination of shocks, where the impulse response function follows a generalized harmonic sequence, governed by a memory parameter. The fractional spline process encompasses standard trend models, such as the random walk and the integrated random walk, while for fractional memory parameter it provides an alternative to standard fractional noise. We present statstical inference for the fixed effects and random effects specifications of the model and, in the latter case, its time series properties, also in comparison with fractional noise. We then illustrate its forecasting performance and its ability to capture persistent trends in economic data (inflation and realized volatility), and tree-ring based temperature reconstructions.

Sparse Tree-Based Aggregation for Time Series Regressions

M. Corillon[1], I. Wilms[1], S. Smeekes[1]

[1]Maastricht University, Netherlands

High-frequency and mixed-frequency time series often require numerous lags, increasing dimensionality, which complicates parameter estimation and interpretation. This paper introduces sparse tree-based aggregation for time series regressions (StarTime), a tree-guided estimator that jointly learns temporal aggregation and sparsity, delivering parsimonious and interpretable regressions, increasing the signal-to-noise ratio. This method arranges lags into a tree structure, where each level represents a higher degree of aggregation. Aggregation is induced by shrinking internal node parameters, allowing data to determine the appropriate temporal resolution at which groups of lags should enter the model. The framework extends to multiple predictors observed at different base frequencies. We provide a locally adaptive alternating direction method of multipliers algorithm to perform the estimation, followed by a post-estimation refit to reduce regularization bias. Furthermore, we present theoretical results for the prediction error of StarTime. Extensive simulations in autoregressive and mixed-frequency settings show systematic gains in estimation accuracy, recovery of aggregation structure, and sparsity relative to common benchmarks. Finally, we prove the practical advantages of our method through an empirical application to financial and macroeconomic data.

Threshold models for high-dimensional time series with network structure

C.Y. Yau[1]

[1]Chinese University of Hong Kong, Hong Kong

Threshold autoregressive (TAR) models form an important class of nonlinear time series models and have attracted great attentions in the literature. In order to extend threshold modeling to high-dimensional nonlinear time series, a threshold network autoregressive (TNAR) model is proposed in this paper to overcome the difficulty of over-parameterization by exploiting the available information of network relations. The proposed model can characterize the regime-switching feature in nonlinear complex network systems. Sufficient conditions for the strict stationarity and the ergodicity of the TNAR model are established. A computationally efficient method based on group LASSO is developed to estimate the multiple thresholds and the parameters. A grouped TNAR model is also proposed to further reduce the number of the parameters. The asymptotic behavior of the proposed method is explored and the estimation consistency of both number of groups and group membership structure is established.

Nonlinear Coherence for Vector Time Series: Defining Region-to-Region Functional Brain Connectivity

P. Redondo[1], R. Huser[1], H. Ombao[2], A. Y El Yaagoubi[2]

[1]KAUST, Saudi Arabia
[2]King Abdullah University of Science and Technology, Saudi Arabia

Alterations in functional brain connectivity characterize neurodegenerative disorders such as Alzheimer's disease (AD) and frontotemporal dementia (FTD). As a non-invasive and cost-effective technique, electroencephalography (EEG) is gaining increasing attention for its potential to identify reliable biomarkers for early detection and differential diagnosis of AD and FTD. Considering the behavioral similarities of signals from adjacent EEG channels, we propose a new spectral dependence measure, the nonlinear vector coherence (NVC), to capture beyond-linear interactions between oscillations of two multivariate time series observed from distinct brain regions. This addresses the limitations of conventional channel-to-channel approaches and defines a more natural region-to-region connectivity framework in the frequency domain. As a result, the NVC measure offers a new approach to investigate dependence between brain regions, which then enables to identify altered functional connectivity dynamics associated with AD and FTD. We further introduce a rank-based inference procedure that enables fast and distribution-free estimation of the proposed measure, as well as a fully nonparametric test for spectral independence. The empirical performance of our proposed inference methodology is demonstrated through extensive numerical experiments. An application to resting-state EEG data reveals that our novel NVC measure uncovers distinct and diagnostically meaningful connectivity patterns which effectively discriminate healthy individuals from those with AD and FTD.

Generative Models in Distribution Estimation

8:30 - 9:00

Adaptive nonparametric drift estimation for multivariate jump diffusions under sup-norm risk

N. Dexheimer[1]

[1]University of Twente, Netherlands

We investigate nonparametric drift estimation for multidimensional jump diffusions based on continuous observations. The results are derived under anisotropic smoothness assumptions and the estimators' performance is measured in terms of the sup-norm loss. We present two different Nadaraya--Watson type estimators, which are both shown to achieve the classical nonparametric rate of convergence under varying assumptions on the jump measure. Fully data-driven versions of both estimators are also introduced and shown to attain the same rate of convergence. The results rely on novel uniform moment bounds for empirical processes associated to the investigated jump diffusion, which are of independent interest.

Dimension-Independent Learning with Deep Neural Networks

S. Langer[1], T. Nagler[2]

[1]Ruhr University Bochum, Germany
[2]LMU Munich, Germany

Statistical learning theory for deep learning has primarily focused on analyzing neural networks as estimators in nonparametric regression and classification problems. A central question is under which assumptions neural networks can circumvent the well-known curse of dimensionality. Existing literature typically addresses this question by either imposing structural assumptions on the target function or by assuming that the covariates possess a low-dimensional structure.
In this talk, we present a unified framework that incorporates both types of assumptions within a single model. We show that neural networks can adapt to this framework and establish approximation and excess risk bounds that are independent of the ambient input dimension.

Distribution estimation via Flow Matching with Lipschitz guarantees

L. Kunkel[1]

[1]Ruhr University Bochum, Germany

Flow Matching, a promising approach in generative modeling, has recently gained significant attention. Despite its empirical success, the mathematical understanding of its statistical power remains limited, largely due to the strong dependence of existing theoretical bounds on the Lipschitz constant of the vector field driving the underlying ODE. In this talk, we investigate the assumptions under which this dependence can be controlled. We characterize classes of target distributions that admit vector fields with controlled Lipschitz constants, and exhibit examples where such control is impossible under arbitrary noise schedules. Building on these insights, we establish convergence rates in Wasserstein-1 distance between the learned and target distributions under stable noise schedules, improving upon previous results in high-dimensional settings.

Nonparametric estimation of conditional probability distributions using a generative approach based on conditional push-forward neural networks

L. Tedesco[1]

[1]University of Bergamo, Italy

We introduce Conditional Push-Forward Neural Networks (CPFN), a generative framework for estimating conditional distributions. Instead of directly modeling the conditional density, the method learns a stochastic transformation that, starting from a suitably chosen latent random vector, produces samples that approximately follow the same distribution as the observed data given a certain input. This approach enables efficient conditional sampling and straightforward estimation of conditional statistics via Monte Carlo methods. The model is trained using an objective function derived from the Kullback-Leibler divergence, without requiring invertibility or adversarial training. We also establish a near-asymptotic consistency result and demonstrate experimentally that CPFN can achieve performance comparable to, or even better than, state-of-the-art methods, including kernel estimators, tree-based algorithms, and popular deep learning techniques, while remaining lightweight and easy to train.

Recent advances in data beyond independence

8:30 - 9:00

GLS-Whitened Conformal Prediction for Spatially Correlated Data

A. Saha[1], A.B. Sen[2]

[1]University of California, Irvine, United States
[2]Indian Statistical Institute, Kolkata, India

Conformal prediction requires exchangeability, which spatial data violates due to correlated errors. When residuals are dependent, the conformal quantile is estimated from fewer effective observations than the nominal calibration size, and coverage deteriorates at locations far from training data. Existing spatial conformal methods address this by working within the dependence structure, using localization, kernel weighting, or mixing-based coverage corrections to achieve approximate validity. However, all such approaches effectively reduce the calibration information available for quantile estimation. We propose instead to remove the dependence before calibration. In GLS-based spatial models, a covariance estimate is already available from model fitting. We use its Cholesky factor to whiten calibration residuals, producing approximately independent conformity scores while retaining the full calibration set. The conformal quantile is then mapped back through the conditional variance at each test location, yielding spatially adaptive prediction intervals. We evaluate the proposed method across varied dependence regimes, spatial designs, and misspecification scenarios using simulated and real spatial data.

Variable Selection with Deep Neural Networks using Variational Inference

T. Maiti[1]

[1]Michigan State University, United States

Deep Learning has become an increasingly popular tool for a wide range of scientific domains. As a result, the so-called “black-box” is being validated more broadly than just showcasing its empirical success. To this end, there has been a growing interest in the problem of variable selection with deep neural networks. In this work, we propose a Bayesian solution to the problem of nonlinear variable selection by using a Sparse Group Lasso slab in conjunction with Dirac measure on the input weight matrix, followed by a simple Lasso prior on the remaining weights for optimal network estimation.

Data-Driven Estimation of Rare Event Probabilities for Spatial Extremes and Its Applications

M. Barman[1], L. Pereira[1], A. Deo[1], S. Deb[1]

[1]Indian Institute of Management Bangalore, India

In an era of climate volatility, unprecedented events are occurring with increasing frequency, shattering historical benchmarks. For decision-makers in infrastructure and supply chain management, the greatest challenge is the paucity of data: by definition, catastrophic events are rare and often absent from historical records. Traditional statistical models typically under-predict these extreme tail-risk events because they rely on empirical averages that fail to capture the heavy-tailed nature of extreme weather. When "once-in-a-century" events occur every decade, the limitations of Gaussian-based risk assessments become a liability.

This work introduces a robust non-parametric statistical framework designed to see beyond the available data by extrapolating into the unobserved extremes. The framework leverages Extreme Value Theory (EVT) and the principle of mathematical self-similarity to bridge the gap between observed data and unobserved catastrophic risks. This allows us to develop a methodology that learns the behavior of spatial processes deep into the tail. We demonstrate how the stability of intermediate order statistics can be utilized to estimate the "tail index," providing a reliable measure of how quickly probabilities decay for extreme events.Specifically, our contribution is organized into a cohesive three-fold structure: (1) modeling framework for spatial extremes, (2) statistical methodology for estimation of extremes, and (3) the application to estimation of probability of extreme rainfall, air quality index (AQI), and power-grid systems. The structural flexibility of the proposed framework ensures its utility across a broad spectrum of applications, particularly in sectors where stress-testing infrastructure against unobserved spatial extremes is critical for risk management.

On the Information-Theoretic Limits of Generative Language Models

S. Majumdar[1]

[1]Indian Institute of Management Bangalore, India

Large language models (LLMs) are evaluated as though perfect reliability is achievable on every task. We show this assumption is information-theoretically unjustified. Every generative task has a reliability ceiling, which is the fraction of output uncertainty resolvable from observable context. No model, regardless of scale or architecture, can exceed it. Autoregressive generation further degrades this ceiling at a rate governed by what we call the task's dependency kernel, which captures the inter-token correlation structure of the output. Tasks with local dependencies (e.g. code) contain generation errors locally, while tasks with global dependencies (e.g. poetry) suffer cascading amplification. From these two primitives, we derive a power-law relationship between model performance, parameter count, and training data volume. This relationship has three terms: an irreducible floor, a data-scaling term, and a capacity-scaling term. The irreducible floor equals the conditional entropy of the output given the input. The scaling exponents are determined by the eigenspectrum of the dependency kernel. Performance is limited by whichever resource, data or capacity, is more scarce. The familiar Chinchilla scaling law emerges along the compute-optimal frontier, where both are balanced. We validate our proposal empirically and connect the framework to evaluation practices, context injection strategies, constrained decoding, and model fine-tuning.

Recent Advances in Financial Econometrics

8:30 - 9:00

Sequentially Valid Forecast Comparisons

T. Dimitriadis[1]

[1]Goethe University Frankfurt, Germany

Time series forecasting and its evaluation are inherently sequential. However, their classical comparison via the Diebold–Mariano test is valid only for fixed evaluation time points and does not allow for continuous monitoring. We develop sequentially valid inference on forecast performance for general, potentially unbounded loss differentials in multi-step-ahead forecasting based on asymptotic confidence sequences. These confidence sequences provide asymptotic coverage guarantees uniformly over time, thereby enabling online monitoring and tracking of loss differentials while permitting data peeking. We document the favorable finite-sample properties of our procedure in simulations and demonstrate its practical relevance in an application to volatility forecasting.

Generative Predictive Distributions for Time Series

J. Llorens-Terrazas[1], M. Meitz[2]

[1]Universidad Carlos III de Madrid, Spain
[2]University of Helsinki, Finland

We propose a flexible framework for modeling the predictive distributions of nonlinear, possibly multivariate time series. Our approach expresses a general predictive distribution in an appropriate generative representation that is based on a folklore result from measure theoretic probability. This representation provides a direct simulation-based approximation to the predictive distribution, enabling straightforward computation of forecasts for the conditional mean and variance, fan charts, value at risk, expected shortfall, joint tail risks, and other quantities of interest. We estimate this generative representation using a version of generative adversarial networks and provide a formal statistical analysis of estimation under temporal dependence. Specifically, estimation is expressed as a particular minimax problem and we establish consistency of approximate minimax solutions in Hausdorff distance. The empirical relevance of the approach is illustrated using applications to equity returns, realized variance, and realized covariances. The proposed method is also computationally manageable, with estimation in our applications taking approximately one minute on a standard laptop.

Systems of Equations for Expectile

H. Chuang[1], O. Chuang[2], Z. Du[3], Z. Huang[4]

[1]National Taipei University, Taiwan
[2]Hubei University of Economics, China
[3]Fudan University, China
[4]University of Hong Kong, Hong Kong

xpectiles have recently received considerable attention in econometric theory and empirical applications, yet existing approaches predominantly focus on single-equation frameworks. This paper introduces the systems of equations for expectiles (SEE), including seemingly unrelated regressions for expectiles and panel expectiles as special cases, to capture the interdependencies among the expectiles of multiple units. We propose a system asymmetric least squares (SALS) estimator for SEE and develop a feasible generalized SALS (FGSALS) estimator to account for cross-equation error dependence. We further extend the two estimators to high-dimensional settings. We derive the asymptotic properties of the estimators in both settings. As an empirical application, we apply our methods to the 2020 list of global systemically important banks. Formal out-of-sample model evaluation tests document our models significantly outperform single equation models. We also find clear asymmetric effects of past positive and negative shocks of different units as well as their volatilities on the expectiles.

Debiased Tests for Sharpe Ratios of Machine Learning Portfolios

S. Barendse[1], M. Cheng[1]

[1]University of Amsterdam, Netherlands

We propose a debiased test for the Sharpe ratio of market-timing portfolios constructed from machine-learning return predictions, which corrects for estimation error in the predicted returns. Existing inference procedures typically ignore the estimation error, which can lead to under- or overstated performance, particularly in high-dimensional settings where flexible methods such as neural networks converge slowly. Our approach generalizes classical comparative Sharpe ratio tests by introducing Neyman-orthogonal moment conditions for portfolio return moments and a cross-fitting design that respect the time-series nature of the data. The resulting test remains valid for large panels of asset returns, allowing for the rate N = o(T^2), where N and T denote the amount of assets and sample size, respectively.
Monte Carlo simulations using neural-network predictions demonstrate the finite-sample performance of our method and document the size distortion of the standard test. Empirically, we apply the procedure to portfolio returns obtained from neural-network-based predictions of individual stock returns in a large panel and to aggregate equity premium prediction using a linear model.

Complex, partial or surrogate data in prediction

8:30 - 9:00

Partially identified linear quantile regression under dependent censoring

I. Willems[1], I. Van Keilegom[1], J. Beyhum[1]

[1]KU Leuven, Belgium

We study the linear quantile regression model for a censored event time, imposing only minimal assumptions on the underlying censoring mechanism. Notably, we allow for any form of dependence between event and censoring. In this context, it is well known that the linear quantile regression parameters are not point identified. Instead, we obtain bounds on these parameters as the solution to a linear programming problem with unknown but estimable feasible region. Recent advances in the field of econometrics are leveraged to prove asymptotically correct coverage of the resulting bounds. In a second stage, we show how the bounds can be improved if one can distinguish dependent from independent censoring. The finite sample performance of our method is tested through simulations, and a data application shows its practicability on real data.

Pseudo-observation process regression for survival outcomes

Y. Goldberg[1]

[1]Technion - Israel Institute of Technology, Israel

Pseudo-observation methods have become useful tools for regression with censored survival data, especially when the quantity of interest is not a standard hazard-based parameter. Most current applications focus on estimating outcomes at a fixed and limited number of time points, such as survival probability, cumulative incidence, or restricted mean survival. In many applications, however, the main goal is to describe how the entire outcome curve evolves over time for patients with different covariate profiles.

We propose a general framework that extends pseudo-observation regression from a small set of time points to the full time course of a survival outcome. The idea is to construct pseudo-observations over time and combine them with a process-regression approach, so that covariate effects can be studied continuously rather than only at selected times. This yields an estimated covariate-specific survival curve and includes existing pseudo-observation regression methods as a special case.

The proposed framework is flexible and can be used for several survival-type outcomes, including overall survival, cumulative incidence in competing risks, and related functionals. It also provides a bridge between familiar pseudo-observation methods and broader regression approaches for time-varying outcomes. We discuss the main methodological ideas, practical computation, and extensions to settings with covariate-dependent censoring. The approach is motivated by the need for interpretable regression methods that target clinically meaningful quantities directly, while making fuller use of follow-up information over time.

This work aims to broaden the scope of pseudo-observation methods from pointwise analysis to process-level estimation and prediction.

When Distribution-Free Falls Short for Prediction Intervals under Right-Censored Covariate

T. Garcia[1], K. Han[2], Y. Ma[2]

[1]University of North Carolina at Chapel Hill, United States
[2]Penn State University, United States

Prediction intervals provide a data-driven range for an unmeasured outcome at a prespecified probability level, and are essential for prognosis and patient monitoring in clinical studies. Unstable prediction intervals---ones that vary substantially in length or coverage rate from study to study---cannot support reliable clinical decisions about patient screening, resource allocation, or intervention timing. A common source of this instability is a right-censored time-to-event covariate: the event has not yet occurred by the end of the study, so the covariate value needed for prediction is unknown. No existing method constructs prediction intervals directly from a right-censored time-to-event covariate, and adapting distribution-free methods to this setting produces exactly this instability. We develop a semiparametric prediction method that incorporates an outcome model, a model for the time-to-event covariate, and a model for the censoring time into the estimation of the prediction interval length. The method achieves the smallest possible variance in interval length estimation---a formal improvement over distribution-free methods---and remains consistent even when the model for the time-to-event covariate or the model for the censoring time is misspecified. Simulation studies confirm substantially more stable interval lengths and coverage rates than distribution-free methods across censoring rates. In a Huntington disease study with 77\% censoring, our method achieves reliable coverage with stable interval lengths, while distribution-free methods produce either persistent undercoverage or intervals too wide to be informative.

Online Surrogate Evaluation with Streaming Data

L. Parast[1], A. Delaigle[2]

[1]The University of Texas at Austin, United States
[2]University of Melbourne, Australia

Surrogate markers, which are intermediate or easier-to-measure biological measurements, offer a practical alternative for assessing treatment effects more efficiently when they are used in place of a primary outcome that is difficult to measure. Many useful methods exist to evaluate whether a surrogate can be used in place of a primary outcome, but these existing methods tend to assume that the individual-level data are available and accessible in their entirety. However, in a clinical trial setting, data are often available over time in batches and additionally, increasingly strict restrictions around data storage and access warrant online estimation approaches where individual-level data from a batch are used once, and thereafter considered unavailable. While online estimation of streaming data has been extensively studied for decades, the existing work does not offer methods for surrogate evaluation. In this paper, we propose robust nonparametric methods for online surrogate evaluation with streaming data. Our proposed approach is based on kernel-based methods and recursive estimators that use only patient-level data from the current batch, along with summaries retained from prior batches. We pay particular attention to the problem of bandwidth selection which is uniquely important and complex in this setting. We examine both the theoretical properties of our proposed estimators, as well as the finite sample performance via a simulation study. Using data from the Diabetes Control and Complications Trial, we use our proposed methods to investigate hemglobin A1c as a surrogate marker for albumin excretion rate, which reflects damage to the kidneys and is used to diagnose kidney disease, a diabetes-related complication.

Contributions to Understandable Machine Learning

8:30 - 9:00

Statistics in the Shadow of AI: Crisis or Transformation?

J. Lederer[1]

[1]University of Hamburg, Germany

Artificial intelligence is reshaping research and education at an extraordinarily rapid pace. But unlike engineering and computer science, the field of statistics is largely a passenger instead of an active driver of this transformation. This situation has generated concern and uncertainty about the future of the field, including questions about whether the
discipline remains necessary at all. This talk outlines a future in which statistics evolves, rather than disappears, and remains relevant for both research and education.

Multi-target semi-supervised learning with application to small area estimation

K. Reluga[1], N. Salvati[2], M. van der Laan[3]

[1]Humboldt-Universität zu Berlin, Germany
[2]University of Pisa, Italy
[3]University of California, Berkeley, United States

In the classical single-target semi-supervised learning (SSL) setting, one has access to (i) a moderately sized labelled dataset containing both response values and associated features, and (ii) a much larger unlabelled dataset with only covariates observed. SSL naturally arises in settings where collecting features is easy, but obtaining labels is expensive or time-consuming, for example, in electronic health records or survey data, where full data is available for only a small subset of the population. We extend this framework to multi-target semi-supervised learning, where the goal is to estimate several parameters of interest across different subpopulations, but labelled data are sparse. Classical SSL methods can suffer from excessive variability in this setting. We propose novel estimation methods tailored to this problem and demonstrate how they improve stability and efficiency. Finally, we show how small area estimation emerges as a special case of this broader learning framework.

Bayesian Additive Regression Tree Copula Processes for Scalable Distributional Prediction

N. Klein[1]

[1]Karlsruhe Institute of Technology, Germany

We show how to construct the implied copula process of response values from a Bayesian additive regression tree (BART) model with prior on the leaf node variances. This copula process, defined on the covariate space, can be paired with any marginal distribution for the dependent variable to construct a flexible distributional BART model. Bayesian inference is performed via Markov chain Monte Carlo on an augmented posterior, where we show that key sampling steps can be realized as those of Chipman et al. (2010), preserving scalability and computational efficiency even though the copula process is high dimensional. The posterior predictive distribution from the copula process model is derived in closed form as the push-forward of the posterior predictive distribution of the underlying BART model with an optimal transport map. Under suitable conditions, we establish posterior consistency for the regression function and posterior means and prove convergence in distribution of the predictive process and conditional expectation. Simulation studies demonstrate improved accuracy of distributional predictions compared to the original BART model and leading benchmarks. Applications to five real datasets with 506 to 515,345 observations and 8 to 90 covariates further highlight the efficacy and scalability of our proposed BART copula process model.

Timely identification of critical indicators within fully localized prediction

s. sperlich[1], F. Loessl[1]

[1]University of Geneva, Switzerland

Suppose we are provided with a large number of observations over time and space.
However, only along a thin line within this data set the variable of interest, say a dummy, exhibits variation over time remaining zero elsewhere. By fully localized procedures we then try to first predict this variation, and then identify the main drivers or say (as we don't employ tools specific for causal analysis) the most important predictors. We also want to allow these to change locally though not necessarily over time.
Being provided with limited computational power and storage capacity but to nonetheless obtain timely results, we resort to importance sampling. To account for the location of the variation of interest we apply ideas similar to self attention but modified to become a positional attention and training. This is achieved by constructing a directional distance kernel. The identification of locally important predictors is based on kernel - localized linear lasso.
Our application considers the changes of the frontline course in the Russian-Ukrainian war. With the invasion of Ukraine by Russian ground forces in 2022, an unprecedented volume of near real time geospatial data became
publicly available. These exhibit both, a relatively high frequency and a quite high quality. Those data are provided by open access websites like liveUAmap.com, and is compiled from information sources close to action via a diverse set of
internet based channels like telegram and facebook. While such data has been extensively used for qualitative analysis of the conflict so far, we are not aware of any sophisticated quantitative statistical analysis. We created maps along the front separating the terrain in small contiguous hexagons (100 x 100 meter) to keep high spatial precision. The dummy variable of interest is an occupational change, and for its prediction we have about 50 covariates.

Recent advances on data with complex structure

8:30 - 9:00

'HARMLESS' sequential sampling for estimating level sets of response functions

M. Banerjee[1], Y. Yu[1], Y. Ritov[1]

[1]University of Michigan, United States

We propose a two-stage design for estimating a root or boundary defined by an unknown regression function observed with noise. In many applications, including dose-finding clinical trials, it is desirable to learn the zero level set while allocating observations mainly on one \emph{safe} side of the unknown boundary. Classical one-sided stochastic approximation procedures guarantee finite overshoot but achieve convergence rates governed by the imposed bias, which are slower than the parametric rate.

In Stage I, a one-sided stochastic approximation is used to obtain a conservative preliminary estimator while preserving finite overshoot. In Stage II, additional observations are collected locally on the safe side and a local regression estimator is constructed. We show that the resulting estimator is asymptotically normal with the parametric rate, despite the finite-overshoot constraint. We further show that a variant of Polyak–Juditsky averaging fails to recover the parametric rate under finite-overshoot designs. Extensions to multivariate boundary estimation are developed for both linear and smooth nonparametric level sets. To demonstrate the effectiveness of our procedure for risk-controlled decision making, we conduct extensive simulations and apply it to a bank loan dataset to estimate the maximal borrower feature level that keeps the default probability below a prescribed threshold.

This is joint work with Ya'acov Ritov and Yue Yu.

Inference of dependence among latent processes from aggregated measurements

H. Li[1], R. Requadt[1]

[1]University of Goettingen, Germany

In many applications, including ion channel analysis, only aggregated signals are observed rather than individual system components. This raises the question of how to recover dependence structures among latent processes from such limited data. We introduce a novel framework based on a continuous-time Markov chain on a binary state space, termed a sum-dependent Markov chain (SDMC), which encodes dependence through the dynamics of the aggregated sum. We establish identifiability, provide an interpretable structural characterization, and propose a cooperativity index to quantify dependence strength. Using discrete-time observations of the aggregated process, we develop a likelihood-based inference approach within a hidden Markov model framework, proving consistency and asymptotic normality of the maximum likelihood estimator, and introducing a test for dependence with asymptotic control of size and power. We extend the framework to general finite state spaces, where identifiability is governed by an informativity criterion linked to the solvability of certain Diophantine equations, and introduce a robust version to account for measurement noise.

Asymptotics in high dimensional regression under dependence

S. Lahiry[1]

[1]National University of Singapore, Singapore

Over the past decade, the proportional asymptotics regime—where the number of features and samples grow proportionally—has become a central framework in high-dimensional statistics. While this paradigm yields sharp predictions without relying on sparsity, extending existing techniques to dependent data remains a significant challenge.

In this talk, we study three models capturing different forms of dependence and obtain precise asymptotic characterizations. First, we establish a universality result for regression with dependent but uncorrelated sub-Gaussian designs, showing that under a block dependence structure the estimation risk matches that of the i.i.d. Gaussian case. This enables sharp risk predictions for nonlinear models such as additive and functional regression.

Second, we analyze regression with temporal dependence in the non-sparse regime from a Bayesian perspective. We derive asymptotic formulas for mutual information and minimum mean square error (MMSE), and show that the optimal error is achievable via an iterative algorithm.

Third, we study inference in linear structural equation models, to quantify how an exposure affects an outcome both directly and indirectly through multiple mediators. We construct a debiased estimator for mediation effects, establish its asymptotic normality, and fully characterize its limiting variance. While debiasing is a standard procedure in high dimensional inference, the mediating variables create complex dependence structure requires a far nuanced analysis. This yields a practical procedure for valid hypothesis testing of mediation effects.

Parsimonious Modeling of Multivariate Time Series with Heteroskedasticity

S.Y. Samadi[1], T. Valizadeh[2]

[1]SIUC, United States
[2]Texas State University, United States

Economic and financial time series often exhibit heteroskedasticity, posing challenges for modeling and inference, particularly in high-dimensional settings. As dimensionality increases, conventional multivariate approaches can become heavily parameterized, limiting both interpretability and computational feasibility. This talk introduces a parsimonious framework for multivariate time series that accommodates heteroskedasticity while maintaining a flexible representation of dependence. The approach leverages low-dimensional subspace structures to jointly capture key features of the mean and second-order behavior, effectively eliminating redundant information. This yields a scalable parameterization that improves estimation efficiency without imposing restrictive assumptions on the data-generating process. Theoretical properties of the resulting estimators are outlined, and the framework is illustrated through simulation studies and empirical applications, demonstrating improvements in estimation accuracy and efficiency.

Random networks, clustering and embeddings

8:30 - 9:00

Statistically and Computationally Optimal Estimation and Inference of Common Subspaces

J. Agterberg[1]

[1]University of Illinois Urbana-Champaign, United States

Given multiple data matrices, many problems in statistics and data science rely on estimating a common subspace that captures certain structure shared by all the data matrices.
In this paper we investigate the statistical and computational limits for the common subspace model in which one observes a collection of symmetric low-rank matrices perturbed by noise, where each low-rank matrix shares the same common subspace. Our main results identify several regimes of the signal-to-noise ratio (SNR) such that estimation and inference is statistically or computationally optimal, and we refer to these regimes as weak SNR, moderate SNR, strong estimation SNR, and strong inference SNR. First, we propose an estimator based on projected gradient descent initialized via spectral sum of squares and show that it achieves the optimal $\sin\Theta$ error rate under strong estimation SNR. These results are complemented by both statistical and computational lower bounds identifying the weak and moderate estimation SNR regimes. Next, we turn to statistical inference for the $\sin\Theta$ distance itself, and we show that our estimator has an asymptotically Gaussian distribution in the strong inference SNR regime. Based on this limiting result we propose confidence intervals and show that they are adaptively minimax optimal in the strong inference SNR regime, where adaptivity is measured in terms of the SNR. Finally, we then show that adaptive confidence intervals are information-theoretically impossible below the strong inference SNR regime. Consequently, our results unveil a novel phenomenon: despite the SNR being ``above'' the computational limit for estimation, adaptive statistical inference may still be information-theoretically impossible.

Network link prediction via isotonic regression

S. Sengupta[1]

[1]North Carolina State University, United States

We consider the problem of predicting unobserved links in one network (the target network) using information from a second, fully observed network (the source network) that shares some similarity in its underlying structure. For example, consider social media platforms such as Facebook and Twitter/X where pairs of users with a higher probability of connection in one network should also tend to have a higher probability of connection in the other, even if the two networks operate on different scales. We exploit this monotone relationship between edge probabilities through isotonic regression. Specifically, using a commonly observed part of the two networks, we first estimate the relationship between their latent edge probabilities, and then use this estimated relationship to predict unobserved edges in the second network. This framework gives rise to a nonstandard theoretical challenge since isotonic regression is carried out with estimated, rather than directly observed, predictor values, and these estimated predictors may have arbitrary dependence structures. We develop new theoretical tools to address this issue and establish theoretical guarantees for the proposed method. The resulting approach provides a flexible framework that integrates ideas from shape-constrained inference and statistical network analysis to address the link prediction problem.

Statistical hypothesis testing for differences between layers in dynamic multiplex networks

F. Sanna Passino[1]

[1]Imperial College London, United Kingdom

With the emergence of dynamic multiplex networks, corresponding to graphs where multiple types of edges evolve over time, a key inferential task is to determine whether the layers associated with different edge types differ in their connectivity. In this talk, we introduce a semiparametric hypothesis testing framework, under a latent space network model, for assessing whether the layers share a common latent representation. The method we propose enables global testing of differences between layers in multiplex graphs, providing a natural generalisation to the problem of pairwise testing for random graphs extensively covered in previous literature. While we introduce the method as a test for differences between layers, it can easily be adapted to test for differences between time points. We construct a test statistic based on a spectral embedding of an unfolded representation of the graph adjacency matrices and demonstrate its ability to detect differences across layers in the asymptotic regime where the number of nodes in each graph tends to infinity. The finite-sample properties of the test are empirically demonstrated by assessing its performance on both simulated data and a biological dataset describing the neural activity of larval Drosophila. This is joint work with Maximilian Baum and Axel Gandy (Department of Mathematics, Imperial College London).

Clustering and inference for very sparse diverse multiplex networks

M. Pensky[1]

[1]University of Central Florida, United States

The talk considers the DIverse MultiPLEx Generalized Random Dot Product Graph (DIMPLE-GRDPG) network model
where all layers of the network have the same collection of nodes and follow the Generalized Random Dot Product Graph (GRDPG) model. In addition, all layers can be partitioned into groups such that the layers in the same group are embedded in the same ambient subspace but otherwise all matrices of connection probabilities can be different. While this is already a very difficult model, in addition we assume that layers of the network are very sparse. We shall use tensor-based approaches to recovery of the groups of layers in such network and subsequent estimation of the ambient subspaces.

Contributed: Spatial, Spatio-Temporal and Dependent Processes

8:30 - 8:50

Spatio-temporal modeling for record-breaking events: An application to temperature daily series

A.C. Cebrian[1], A.E. Gelfand[2], J. Castillo[3], J. Asín[1]

[1]Universidad de Zaragoza, Spain
[2]Duke University, United States
[3]University of Zaragoza, Spain

The occurrence of record-breaking temperature events is one of the evidences of climate change. In this context, we present a Bayesian framework to investigate both the occurrence and the magnitude of record-breaking temperatures across years for any given day of the year, within a space-time setting. To this end, we develop two models: one for the occurrence of records and another for the record values, conditional on a record-breaking event having occurred. For the occurrence model, we propose a hierarchical conditional specification for the indicators, incorporating explicit trends, necessary autoregressive terms, additional covariates, and both spatial and strong daily random effects. For the record values, rather than modeling them directly, we model the increments relative to the previous record, conditional on the occurrence of a record-breaking event. This approach allows us to treat increments as conditionally independent and to avoid the order constraint inherent in record values. The analysis is based on more than sixty years (1960-2021) of daily maximum temperature data across peninsular Spain. The fitted models provide evidence of the impact of climate change on record-breaking temperatures, enabling us to quantify these effects while identifying spatial patterns and seasonal differences.

Modeling group heterogeneity in spatio-temporal data via physics-informed semiparametric regression

M.F. De Sanctis[1], E. Arnone[2], F. Ieva[1], L.M. Sangalli[1]

[1]Politecnico di Milano, Italy
[2]Università degli Studi di Torino, Italy

Modeling spatio-temporal phenomena driven by complex underlying processes requires flexible semiparametric tools capable of embedding prior physical knowledge in the statistical framework. Motivated by the analysis of nitrogen dioxide concentrations across Lombardy (Italy), we propose a space-time mixed-effects regression to model hourly pollutants' dynamics. The proposed model incorporates a nonparametric component regularized by a second-order elliptic partial differential equation. This formulation captures the non-stationary advective-diffusive wind processes driving air pollutant patterns. We further incorporate random effects in the model to address structured latent variability, successfully accounting for the measurement heterogeneity emerging from the use of distinct sensor technologies across the monitoring network. We develop a novel two-step procedure relying on a functional version of the Iterative Reweighted Least Squares algorithm to estimate both the nonparametric and parametric components. We establish the asymptotic properties of the semiparametric estimators, and we evaluate their performance through extensive simulation studies, demonstrating clear advantages over existing literature.

Record probabilities under dependence: operator-norm bounds and extremal control

C. Al Tannoury[1], A. Hoayek[1]

[1]Ecole Nationale Supérieure des Mines de Saint-Etienne, France

We study record probabilities for strictly stationary sequences of random variables with continuous underlying distributions and quantify their deviation from the classical i.i.d. benchmark. After uniformization, we first derive an exact conditional-law representation of the record probability. We then introduce an absolutely continuous smoothing transformation that converts the record event into a well-defined Lebesgue covariance integral. The resulting
formulation allows us to control record probabilities via covariance operator norms between present and past σ-fields. This in turn yields general bounds in terms of operator-based measures of dependence, which can in turn be combined with classical mixing tools, including Davydov-Rio covariance inequalities and interpolation between α- and ρ-mixing coefficients . To complement the dependence analysis, we derive local bounds for the running maximum under a finite-sample exponential tail condition. The resulting estimates are fully explicit and are illustrated on several stationary models, including Gaussian AR(1), Gaussian MA(1), and a uniformly ergodic refresh chain.

Nonparametric Covariance Estimation with Bias Correction for Spatial Lattice Processes

A. Fioravanti[1], C. Jentsch[1]

[1]TU Dortmund, Germany

The accurate estimation of covariances from spatial data is crucial in many applications. While parametric approaches have been widely used for modeling spatial covariances, the parametric assumption may not be correctly specified, which may lead to wrong conclusions. In contrast, nonparametric approaches are often neglected in practice because their finite-sample estimation accuracy suffers from the large number of covariance parameters to be estimated, and they are often prone to severe bias issues.
In this paper, we develop a novel framework for bias correction for sample covariance estimators of stationary lattice processes indexed in $\mathbb{Z}^2$. We derive the exact finite-sample biases of different versions of sample covariance estimators for spatial data. Our results show that the estimators' expectations can be expressed as linear combinations of the population covariances that depend only on the spatial lag of the sample covariance and on the sample size of the observed lattice data. Exploiting this characterization enables the construction of jointly bias-corrected covariance estimators that are nearly unbiased, depending on the strength of spatial dependence in the lattice process. Additionally, we derive exact formulas for the joint mean-squared error (MSE) and provide asymptotic normality results. In particular, we show that the bias-corrected estimators are asymptotically equivalent to their uncorrected versions.
Simulation studies for Gaussian lattice processes demonstrate that our proposed bias correction can achieve substantial bias reduction while often also reducing the MSE, particularly when the underlying process exhibits strong spatial persistence.

Structural analysis in matrix-autoregressive models

C. Wurtz[1], C. Jentsch[1]

[1]TU Dortmund University, Germany

We consider a structural matrix-autregressive (SMAR) model to conduct impulse response analysis for structural shocks to matrix-valued time series. The MAR model of order p offers a parsimonious and interpretable framework for these time series, thus addressing issues of high-dimensionality in corresponding vector-autoregressive (VAR) models. To interpret the dynamics, we resort to impulse response analysis as a popular tool from the SVAR context. Its conclusions rely on the valid identification of structural shocks that are mutually contemporaneously uncorrelated and interpretable. In contrast to the existing literature, the proposed SMAR model enables the identification of multiple structural shocks. To address the restrictive nature of the single-term MAR(p) model, we discuss the extension to a multi-term SMAR(p) model as a compromise between the single-term SMAR and the (unrestricted) SVAR model, trading off parsimony against flexibility. We discuss its identification, focusing in particular on issues that arise due to the typical Kronecker-product structure of the coefficient matrices in the MAR framework. Further, we discuss estimation and inference in the general multi-term SMAR(p) model. Specifically, we consider a bootstrap method to compute confidence bands for the impulse response curves and provide corresponding results. In this context, a key point concerns model misspecification and the use of MAR models to approximate more general (S)VAR data generating processes. Finally, we demonstrate the performance and practical use of our approach by Monte Carlo simulations and a real data application.

Keynote Talk

11:00 - 12:00

Testing hypotheses via orthogonalization

D. Witten[1]

[1]University of Washington, United States

Classical hypothesis testing frameworks break down in contemporary settings in which null hypotheses are increasingly abstract, the same data are used to both generate and test hypotheses, and minimal assumptions about the underlying data are made. In this work, we propose a new framework for conducting valid hypothesis tests in broad contexts. We propose to add and subtract external noise generated from a symmetric shift-family to our data, X, to partition it into two pieces, X1 and X2. We provide a generic strategy for orthogonalizing X2 against X1 under the null hypothesis H0, then show that testing whether the orthogonalization was successful provides a valid test of H0 under mild assumptions. Remarkably, this framework extends naturally to the post-selection inference setting with minimal modifications: we simply select a hypothesis on X1, then perform orthogonalization under the selected null. As our approach neither requires pre-specification of the selection mechanism, nor is restricted to a small class of data-generating distributions, it dramatically expands the settings for which valid post-selection inference can be conducted. We showcase the flexibility of our proposal in a number of case studies. This is joint work with Ameer Dharamshi (University of Washington).

At the frontier of modern nonparametric statistics

13:30 - 14:00

Extended Fiducial Inference for Models with Unknown Random Error Distributions

W. Wang[1], F. Liang[1]

[1]Purdue University, United States

Gaussian noise assumptions underpin much of classical statistical inference and machine learning, yet real-world data often deviate substantially from this ideal, rendering Gaussian-based inference unreliable. Existing approaches for unknown error distributions—such as robust statistics, flexible parametric families, and conformal prediction—typically fail to simultaneously provide valid parameter inference, automatic adaptation to complex noise structures, and principled uncertainty quantification. To address this gap, we propose Nonparametric Error-Extended Fiducial Inference (NE-EFI), a multi-stage framework that generalizes Extended Fiducial Inference (EFI) to settings with unknown error distributions. NE-EFI leverages EFI's error recovery property to impute latent errors, reconstructs the noise distribution via kernel density estimation, and iteratively refines inference. We establish theoretical convergence guarantees and demonstrate through simulations and real data that NE-EFI achieves accurate inference and valid uncertainty quantification, while consistently outperforming robust statistical methods and conformal approaches across a broad range of non-Gaussian error distributions. This is a joint work with Wen-Hung Wang.

A Bayesian approach to learning mixtures of nonparametric components

Y. Zhang[1], Y. Wei[2], A. Guha[3], L. Nguyen[1]

[1]University of Michigan, United States
[2]University of Texas, Dallas, United States
[3]AT&T Labs, United States

Mixture models are widely used in modeling heterogeneous data populations. A standard approach of mixture modeling assumes that the mixture component takes a parametric kernel form. In many applications, making parametric assumptions on the latent subpopulation distributions may be unrealistic, which motivates the need for nonparametric modeling of the mixture components themselves. In this paper, we study finite mixtures with nonparametric mixture components, using a Bayesian nonparametric modeling approach. In particular, it is assumed that the data population is generated according to a finite mixture of latent component distributions, where each component is endowed with a Bayesian nonparametric prior such as the Dirichlet process mixture. We present conditions under which the individual mixture component's distribution can be identified, and establish posterior contraction behavior for the data population's density, as well as densities of the latent mixture components. We develop an efficient MCMC algorithm for posterior inference and demonstrate via simulation studies and real-world data illustrations that it is possible to efficiently learn complex forms of probability distribution for the latent subpopulations. In theory, the posterior contraction rate of the component densities is nearly polynomial, which is a significant improvement over the logarithmic convergence rates of estimating mixing measures via deconvolution. This work is joint with Yilei Zhang, Yun Wei and Aritra Guha.

Nonparametric Functional Tolerance Bands for Anomaly Detection

D. Young[1]

[1]University of Kentucky, United States

Two primary types of anomalies (outliers) in functional data are magnitude anomalies, which are curves that at some point have a significant spike away from the rest of the curves, and shape anomalies, which are curves that display a different behavior than the other curves. Depth-based approaches have been employed to address the detection of these anomalies. In this talk, we propose the use of tolerance bands to address both types of anomalies. A pointwise approach (relative to the functional domain) based on order statistics of deviations is used for detection of magnitude anomalies. A simultaneous approach based on functional depth is used for detection of shape anomalies. A coverage study is performed along with analysis of a real dataset.

Nonparametric smoothed likelihood-based estimation of finite mixtures of symmetric distributions

M. Levine[1]

[1]Purdue University, United States

Our work is dedicated to the study of the semiparametric mixture of k unknown symmetric distributions that are equal up to à location parameter. Such a model is usually considered a semiparametric one since the unknown density functions do not belong to any predetermined parametric family of distributions. It has a fairly extensive history in non- and semiparametric statistics although its study has been somewhat hampered by the fact that sufficient identifiability conditions for it are only known in the case of k=2 and k=3. We propose a new method of estimation of these model's parameters based on optimizing a criterion that is, effectively, a negative nonparametric penalized likelihood of this model. The approach results in an iterative algorithm possessing a descent property with respect to the proposed criterion. Several algorithms of this kind have been proposed for estimation of non- and semiparametric mixtures so far; all of them, however, employ convolution operators at every step of iteration which makes them somewhat time consuming. Our approach avoids this pitfall and, consequently, its implementation is easier than that of other comparable algorithms. Interestingly enough, this algorithm is neither an EM nor an MM algorithm. We establish convergence of the algorithm and illustrate its performance using both synthetic data sets and the well-known benchmark datasets used in the literature earlier.

Complex time series analysis

13:30 - 14:00

The Unseen Species Problem Revisited

E. Eriksson[1]

[1]MPI Leipzig, Germany

Imagine an ecologist who wants to predict how many new species a return expedition would discover. Formally, given n i.i.d. samples from an unknown distribution over an unknown set, how many new outcomes will be observed if we collect m more samples? We will discuss the problem for short, intermediate and long return trips, as well as generalizations to set-valued samples. This generalized version is equivalent to forecasting the growth of the vertex set in an edge exchangeable (hyper-)graph. The highlight result is a new estimator which in the classical problem is, up to an explicit multiplicative constant, optimal in the class of estimators respecting a natural problem symmetry. We also enable the construction of prediction intervals for the classical Good-Toulmin estimator and give concentration inequalities for certain natural functionals of i.i.d. discrete-set-valued random variables.

A Multi-threshold Change Plane Autoregressive Model

W. Zhang[1]

[1]University of Macau, Macau

This paper proposes a novel multi-threshold change plane autoregressive model with an unknown number of regimes, where the transition between regimes is governed by multiple threshold variables. We introduce an estimation approach that combines adaptive group fused Lasso with linear programming to simultaneously estimate the number of regimes and model parameters. Asymptotic properties of the estimators are established. Simulation studies confirm the method's effectiveness in finite samples. Additionally, an application to U.S. real GNP data reveals insightful findings and demonstrates a marked gain in forecast accuracy over existing alternatives.

Tensor dynamic conditional correlation model with applications to portfolio selection

K. Zhu[1]

[1]University of Hong Kong, Hong Kong

Style investing creates asset classes (or the so-called "styles") with low correlations, aligning well with the principle of "Holy Grail of investing" in terms of portfolio selection. The returns of styles naturally form a tensor-valued time series, which requires new tools for studying the dynamics of the conditional correlation matrix to facilitate the aforementioned principle. Towards this goal, we introduce a new tensor dynamic conditional correlation (TDCC) model, which is based on two novel treatments: trace-normalization and dimension-normalization. These two normalizations adapt to the tensor nature of the data, and they are necessary except when the tensor data reduce to vector data. Moreover, we provide an easy-to-implement estimation procedure for the TDCC model, and examine its finite sample performance by simulations. Finally, we assess the usefulness of the TDCC model in international portfolio selection across ten global markets and in large portfolio selection for 1800 stocks from the Chinese stock market.

An Adversarial Approach for Goodness-of-Fit Tests

Q. YAO[1]

[1]London School of Economics, United Kingdom

Diagnostic checking for goodness-of-fit is one of the important and routine steps in building a statistical model. The most frequently used approach for checking the goodness-of-fit is the residual analysis in the context of regression analysis. However for many statistical models there exist no natural residuals, which includes the models for the underlying distributions behind data, or the models for some complex dynamic structures such as the dynamic network models with dependent edges. Furthermore, there are scenarios in which there exist several competing models but none of them are the clear favorite. One then faces a task to choose the best approximation among the wrong models. We propose an adversarial approach in this paper. For checking the goodness-of-fit, we generate a synthetic sample from the fitted model and construct a classifier to classify the original sample and the synthetic sample into two different classes. The hardness of the classification is then taken as a measure for the goodness-of-fit. For identifying the best model among several candidate models, the classifier will create a distance between the original sample and the synthetic sample generated from each of the candidate model, and the model with the shortest distance can be taken as the best approximation for the truth.

Kernel, Bandit, and Preference-Based Methods

13:30 - 14:00

The statistical price of few updates: efficiency and adaptivity in batched contextual bandits

C. Ma[1], R. Jiang[1]

[1]University of Chicago, United States

Sequential decision-making is central to modern statistics, with applications ranging from clinical trials to online recommendation systems. Classical theory typically assumes that policies can be updated after every observation. In many real-world experiments, however, updates are restricted to a small number of discrete time points, leading to batching constraints.

This raises a fundamental question: how much statistical efficiency is lost when updates are rare, and how many batches are needed to achieve optimal learning? In this talk, we address this question within the framework of contextual bandits with smooth reward functions. We begin with a success story: when the margin parameter is known, we show that only $\log\log T$ batches suffice to attain the minimax regret rates of the fully online setting. This result demonstrates that surprisingly limited adaptivity can yield optimal performance. We then turn to the more subtle and practically relevant case where the margin parameter is unknown. In contrast to the fully online regime, where adaptation is free, we show that batching incurs a provable statistical price, even under adaptive batching schedules. We conclude by describing recent results that sharply characterize this adaptation cost.

Deep kernel learning based Gaussian processes for Bayesian image regression analysis

Y. Zhong[1], W. Shen[2], J. Kang[1]

[1]University of Michigan, United States
[2]University of California, Irvine, United States

Regression models are widely applied in neuroimaging studies to learn associations between clinical and image variables. Gaussian process (GP) priors are common Bayesian nonparametric approaches to model unknown regression functions in high dimensional data. However, existing GP methods depend on pre-specified parametric covariance kernels, which often lack the flexibility to capture data complexity, have limited generalizability across study populations, and face computational challenges in large-scale datasets. We propose a scalable fully Bayesian deep kernel learning framework for GP priors with applications in various image regression models. Our method leverages the estimation power of deep neural networks (DNNs) to adaptively learn the kernel basis functions from the data to capture complex spatial correlations. We establish theoretical properties by deriving posterior concentration rates of regression and kernel function estimation. Simulation studies demonstrate improved accuracy in estimation and signal detection across different scenarios. We further validate the proposed method through two neuroimaging applications: fMRI activation detection for general cognitive ability and EEG-based brain-computer interfaces.

Kernel Methods for High-Dimensional Data Integration

R. Ma[1]

[1]Harvard University, United States

I will present two recent and closely related kernel methods for nonlinear joint embedding of high-dimensional datasets. The first method builds on ideas from entropic optimal transport, while the second is based on duo-landmark integral operators. Both are principled approaches for aligning and jointly embedding multiple datasets, supported by rigorous theoretical guarantees. We show that for a pair of noisy, high-dimensional datasets, these methods consistently recover the shared underlying manifold structure while mitigating dataset-specific nuisance structures. I will provide an intuitive geometric explanation of each methodology, along with the theoretical foundations that justify their performance. I will demonstrate their effectiveness in analyzing a single-cell multiomic dataset for human brain cells, which uncovers interesting cell-type-specific interactions between transcription and epigenomic regulation. This talk is based on recent work in collaboration with Xiucai Ding, Boris Landa, and Yuval Kluger.

Nonparametric methods for multivariate extremes

13:30 - 14:00

Flow-matching and transport for geometric extremes

I. Papastathopoulos[1], L. De Monte[1], X. Song[1]

[1]University of Edinburgh, United Kingdom

We study statistical models for geometric extremes based on transport to a center-outward reference distribution. Our approach combines optimal-transport-based and flow-matching ideas to learn distributional structure in the bulk while retaining geometric features of the tail. In particular, transport to a product-uniform reference provides a natural way to encode radial and angular extremal behaviour, and to examine how inverse rays and transport contours reflect tail geometry. We discuss both Brenier-type and entropic regularizations, with emphasis on the trade-off between smoothness, invertibility, and fidelity in the extremes. The resulting perspective points toward a broader connection between multivariate regular variation, geometric extreme-value analysis, and modern generative modeling.

Estimation of the number of principal components in high-dimensional multivariate extremes

V. Fasen-Hartmann[1], L. Butsch[1]

[1]Karlsruhe Institute of Technology, Germany

For multivariate regularly random vectors, the dependence structure of the extremes is modeled by the so-called angular measure. Estimating the angular measure is challenging in high dimensions due to its complexity. In this talk, we use Principal Component Analysis (PCA) to reduce the dimension and estimate the number of significant principal components of the covariance matrix of the angular measure, assuming a spiked covariance structure. Therefore, we develop an Akaike Information Criterion (AIC) and a Bayesian Information Criterion (BIC) to estimate the location of the spiked eigenvalue of the covariance matrix, reflecting the number of significant components. We then explore the consistency of these information criteria. When the dimension grows in proportion to the number of extremes, we apply random matrix theory techniques to derive sufficient conditions for the consistency of the AIC and the BIC. Finally, the performance of the AIC and BIC is compared in a simulation study.

Inference on marginal expected shortfall under multivariate regular variation

S. Rizzelli[1]

[1]University of Padova, Italy

Expected shortfall is a widely used risk measure and has been extensively studied in the literature, particularly in the context of univariate time series models and frameworks exhibiting either short- or long-range dependence. In multivariate settings, marginal expected shortfall (MES) plays a central role in the assessment of systemic risk, and understanding its extreme behaviour is crucial for evaluating the impact of severe downturns in global financial markets. Standard inference typically relies on bivariate extreme-value models (tail copulas) linking a variable of interest with a proxy for system-wide risk. We show that, when multivariate regular variation can be reasonably assumed, it enables a more comprehensive characterization of extremal dependence across many financial institutions and leads to sharper statistical inference on the MES. Within this framework, we derive an approximation formula for the extreme MES and propose a corresponding estimator, together with a bias-corrected version. Under a general beta-mixing framework—covering widely used time series models with heavy-tailed innovations—we establish contraction rates for the proposed estimators and prove their asymptotic normality, which allows the construction of confidence intervals. A simulation study shows that the proposed estimators substantially improve upon the performance of existing methods. This is joint work with Simone Padoan (Bocconi University) and Matteo Schiavone (University of Padova).

Separation-based causal discovery for extremes

J. Jiang[1]

[1]king abdullah university of science and technology, Saudi Arabia

Structural causal models (SCMs), with an underlying directed acyclic graph (DAG), provide a powerful analytical framework to describe the interaction mechanisms in large-scale complex systems. However, when the system exhibits extreme events, the governing mechanisms can change dramatically, and SCMs with a focus on rare events are needed. We propose a new class of SCMs, called XSCMs, which leverage transformed-linear algebra to model causal relationships among extreme values. Similar to traditional SCMs, we prove that XSCMs satisfy the causal Markov and causal faithfulness properties with respect to partial tail (un)correlatedness. This enables estimation of the underlying DAG for extremes using separation-based tests, and makes many state-of-the-art constraint-based causal discovery algorithms directly applicable. We further consider the problem of undirected graph estimation for relationships among tail-dependent (and potentially heavy-tailed) data. The effectiveness of our method, compared to alternative approaches, is validated through simulation studies on large-scale systems with up to 50 variables, and in a well-studied application to river discharge data from the Danube basin. Finally, we apply the framework to investigate complex market-wide relationships in China's derivatives market.

Topics in econometrics

13:30 - 14:00

Robust Estimation in Conditional Moment Models with Time-Varying Parameters

B. Antoine[1], F. Shahryarpoor[2]

[1]Simon Fraser University, Canada
[2]University of British Columbia, Canada

In a parametric conditional moment model with time-varying parameters, we develop a new integrated conditional moment (ICM) estimator which uses all information from conditional restrictions seamlessly. Our approach builds on the ICM principle originally proposed by Bierens (1982) and combines it with local smoothing to deliver estimates of the time-varying parameters. Under general regularity conditions - including local stationarity and restrictions on physical dependence - we show that our estimator is pointwise consistent and asymptotically normally distributed. Importantly, our approach is a one-step approach that is robust to parametrizing - and estimating - the relationship between endogenous variables and instruments. In addition, we derive uniform results that enable us to construct asymptotic simultaneous confidence tubes. Our simulation study document the reliability and power of our approach in a variety of cases - and, especially, when the underlying relationship between the endogenous variables and the instruments cannot be reliably estimated - even with flexible time-varying approaches.
Our estimation of the traditional Phillips curve that links inflation to unemployment with US data from 1960 to 2024 reveals important fluctuations over time, including the diminishing importance of unemployment, especially after 2006.

Machine Learning for Unobserved Heterogeneity in Panel Data

A. Babii[1], M. Carrasco[2], X. Chen[3]

[1]UNC-Chapel Hill, United States
[2]University of Montreal, Canada
[3]Yale University, United States

Unobserved heterogeneity is central to recovering the distribution of treatment effects, constructing value-added measures, and conducting counterfactual policy analysis. Yet its estimation in panel data models is often challenging, generating the incidental parameters problem. This paper develops a unified nonparametric approach to identifying and estimating the distribution of latent heterogeneity in a broad class of panel data models, including difference-in-differences with heterogeneous treatment effects, duration models, count outcomes, and binary choice. We show that in each case the target distribution solves an ill-posed inverse problem in which the relevant linear operator may be non-compact. We propose novel minimax-optimal regularized estimators built on projected conjugate gradient iterations with early stopping. The proposed estimators are easy to implement and perform well in simulations.

Testing for Bayesian–Nash Equilibrium in Binary Games with Incomplete Information

E. Lapenta[1], P. Lavergne[2]

[1]University of Exeter, United Kingdom
[2]Toulouse School of Economics, France

We propose a test to assess if the distribution of observed data can be rationalized by a unique Bayesian–Nash equilibrium of a binary game with incomplete information. This hypothesis is common in empirical models of incomplete information games. The game structure is nonparametrically specified. Our statistic relies on preliminary nonparametric estimators constructed through a multistep procedure. To control for nonparametric estimation bias, we construct the test statistic using a locally robust approach. Because the asymptotic null distribution of the statistic depends on unknown features of the data, we obtain critical values using a novel multinomial bootstrap and prove its validity. This procedure resamples observations while imposing that a unique Bayesian–Nash equilibrium is played. Monte Carlo experiments demonstrate good small-sample performance of the test. In an empirical application, we implement our procedure to study firms' entry decisions in local markets.

Pivotal and identification-robust nonparametric inference in linear IV models

P. Lavergne[1], B. Antoine[2]

[1]Toulouse School of Economics, France
[2]Simon Fraser University, Canada

We propose new inference procedures for a parametric structural model defined by conditional moments, that are robust to weak identification and heteroskedasticity of unknown form. Our first test is taylored for inference on parameters on endogenous explanatory variables. The test statistic builds on the ICM statistics of Birens (1982) and
Antoine and Lavergne (2023). However, our new statistic directly accounts for heteroskedasticity of unknown form, so it is asymptotically pivotal, and inference is greatly facilitated in practice. We also develop (i) an identification-robust subvector inference procedure that does not rely on the knowledge of identification strength for the remaining parameters, and (ii) a pure specification test. In both cases, the tests are conservative but powerful irrespective of the precise form of the link between instruments and endogenous variables. We show that our procedures are computationally friendly and competitive with existing procedures in simulations and an application.

Resampling method in non-standard frameworks

13:30 - 14:00

Jackknife for periodically correlated time series

A. Dudek[1, 2], P. Bertail[3]

[1]Aix-Marseille Universite, France
[2]AGH University of Krakow, Poland
[3]Paris Nanterre University, France

We develop a jackknife methodology for nonstationary time series, with a particular focus on periodically correlated processes. We show that, as in the stationary case, the jackknife variance provides a consistent estimator of the asymptotic variance. In addition, we consider two bootstrap approaches: the Generalized Seasonal Block Bootstrap and the Extension of the Moving Block Bootstrap, which are known to be consistent for various first- and second-order parameters of periodically correlated series. For the overall mean and seasonal means, we demonstrate that the bootstrap variances closely match the jackknife variance, allowing the derivation of optimal bootstrap block lengths. Finally, we extend the jackknife to general differentiable functionals in the nonstationary setting and illustrate the theoretical results through a small simulation study.

Validity of parametric bootstrap in time series models

Z. Maciszewska[1, 2], P. Bertail[1], A. Dudek[2, 3]

[1]Universite Paris Nanterre, MODAL'X, France
[2]AGH University of Krakow, Poland
[3]Aix-Marseille University, CNRS, I2M, France

We study the validity of bootstrap procedures in parametric models from the perspective of local asymptotic theory. Within this framework, asymptotic inference can be formulated in terms of local experiments that approximate the original statistical model. In locally asymptotically normal (LAN) models, bootstrap consistency is closely related to the local asymptotic equivariance of estimators, a property that extends naturally to many dependent time-series settings.
We generalize these results to models satisfying the Locally Asymptotically Mixed Normal (LAMN) property. This broader class is particularly relevant for nonstationary time-series models, many of which exhibit a LAMN structure, and therefore provides a theoretical framework for understanding bootstrap procedures in such contexts.
Finally, we aim to further extend these results to semiparametric models.

Deep Limit Model-free Prediction in Regression

K. Wu[1], D. Politis[2]

[1]Loyola University Chicago, United States
[2]University of California San Diego, United States

In this paper, we provide a novel Model-free approach based on Deep Neural Network (DNN) to accomplish point prediction and prediction interval under a general regression setting. Usually, people rely on parametric or non-parametric models to bridge dependent and independent variables (Y and X). However, this classical method relies heavily on the correct model specification. Even for the non-parametric approach, some additive form is often assumed. A newly proposed Model-free prediction principle sheds light on a prediction procedure without any model assumption. Previous work regarding this principle has shown better performance than other standard alternatives. Recently, DNN, one of the machine learning methods, has received increasing attention due to its great performance in practice. Guided by the Model-free prediction idea, we attempt to apply a fully connected forward DNN to map X and some appropriate reference random variable Z to Y. The targeted DNN is trained by minimizing a specially designed loss function so that the randomness of Y conditional on X is outsourced to Z through the trained DNN. Our method is more stable and accurate compared to other DNN-based counterparts, especially for optimal point predictions. With a specific prediction procedure, our prediction interval can capture the estimation variability so that it can render a better coverage rate for finite sample cases. The superior performance of our method is verified by simulation and empirical studies.

Quantum Bootstrap

W. Zhong[1], P. Ma[1], Y. Chen[2]

[1]University of Georgia, United States
[2]Harvard University, United States

The bootstrap is a foundational tool in statistical inference, but its classical implementation relies on Monte Carlo resampling, introducing approximation error and incurring high computational cost—especially for large datasets and complex models. We present the Quantum Bootstrap (QBOOT), a quantum algorithm that computes the ideal bootstrap estimate exactly by encoding all possible resamples in quantum superposition, evaluating the target statistic in parallel, and extracting the aggregate via quantum amplitude estimation. Under mild circuit-efficiency assumptions, QBOOT achieves a near-quadratic speedup over the classical bootstrap in approximating the ideal estimator, independent of the statistic or the underlying distribution. We provide a rigorous theoretical analysis of its statistical error properties—addressing a gap in the quantum algorithms literature—and validate our results through experiments on the IBM quantum simulator for the sample mean problem. Our findings demonstrate that QBOOT preserves the asymptotic properties of the ideal bootstrap while substantially improving computational efficiency and precision, establishing a scalable and principled framework for quantum statistical inference.

Privacy in Statistics : Recent advances

13:30 - 14:00

Differential privacy with dependent data

M. Avella Medina[1], V. Roth[2]

[1]Columbia University, United States
[2]ISTA, Austria

Dependent data underlies many statistical studies in the social and health sciences, which often involve sensitive or private information. Differential privacy (DP) and in particular userlevel DP provide a natural formalization of privacy requirements for processing dependent data where each individual provides multiple observations to the dataset. However, dependence introduced, e.g., through repeated measurements challenges the existing statistical theory under
DP-constraints. In i.i.d. settings, noisy Winsorized mean estimators have been shown to be minimax optimal for standard (item-level) and user-level DP estimation of a d-dimensional mean parameter Yet, their behavior on potentially dependent observations has not previously been studied. We fill this gap and show that Winsorized mean estimators can also be used under dependence for unbounded data, and can lead to asymptotic and finite sample guarantees that resemble their i.i.d. counterparts under a weak notion of dependence. For this, we formalize dependence via log-Sobolev inequalities on the joint distribution of observations. This enables us to adapt the stable histogram by Karwa and Vadhan (2018) to a non-i.i.d. setting, which we then use to estimate the private projection intervals of the Winsorized estimator. The resulting guarantees for our item-level mean estimator extend to user-level mean estimation and transfer to the local model via a randomized response histogram. Using the mean estimators as building blocks, we provide extensions to random effects models, longitudinal linear regression and nonparametric regression. Therefore, our work constitutes a first step towards a systematic study of DP for dependent data.

Optimality Theory for Adaptation under Differential Privacy

L. Vuursteen[1]

[1]Duke, United States

In classical high-dimensional or nonparametric statistics, it frequently occurs that estimators have to adapt to unknown properties of the underlying parameter class or distribution, such as smoothness or sparsity. Under differential privacy constraints, however, adapting to unknown hyperparameters is known to be significantly more challenging, as typical adaptation schemes such as Lepski's method or cross-validation require iterated re-use of the data, which is costly under the differential privacy framework.

In the talk, I will discuss a general optimality theory for adaptation under the federated differential privacy framework, which generalizes local and central differential privacy: data is distributed across many data holders, each imposing a differential privacy constraint. I will present matching upper and lower bounds that precisely quantify the cost of adaptation under federated differential privacy. Specifically, we delineate when adaptation is possible with little to no cost, and when adaptation incurs more significant penalties or is impossible altogether.

High-Dimensional Private Linear Regression with Optimal Rates

M. Mondelli[1], S. Bombari[1], I. Seroussi[2], J. Luo[3]

[1]Institute of Science and Technology Austria (ISTA), Austria
[2]Tel Aviv University, Israel
[3]Oxford University, United Kingdom

Differentially private (DP) linear regression has received significant attention in the recent theoretical literature, with several approaches proposed to improve error rates. Our work focuses on the popular high-dimensional regime where the number of training samples n and the input dimension d grow at a proportional rate d / n \to \gamma, and it considers a family of one-pass DP gradient descent (DP-GD) algorithms satisfying \rho^2/2-zero-concentrated DP. In this setting, we establish a deterministic equivalent for the DP-GD trajectory in terms of a system of ordinary differential equations. This in particular allows to analyze the effect of gradient clipping constants that are smaller than the typical norm of the per-sample gradients -- a setup that has been shown to improve performance in practice. For Gaussian and well-conditioned data, we show that DP-GD, upon properly choosing clipping constant and learning rate, achieves the non-asymptotic risk of O(\gamma + \gamma^2 / \rho^2), and we establish that this rate is minimax optimal. Then, we consider the ill-conditioned case, focusing on data covariance spectra following a power-law distribution, and we show that the risk displays a power-like scaling law in \gamma, highlighting the change in the exponent as a function of the privacy parameter \rho. Overall, our analysis demonstrates the benefits of practical algorithmic design choices, including aggressive gradient clipping and decaying learning rate schedules.

Nonparametric Spectral Density Estimation using Interactive Mechanisms under Local Differential Privacy

K. Klockmann[1], C. Butucea[2], T. Krivobokova[3]

[1]University of Kassel, Germany
[2]CREST ENSAE IP Paris, France
[3]University of Vienna, Austria

We study the problem of estimating the spectral density of a centered stationary Gaussian time series under local differential privacy constraints. Specifically, we propose new interactive privacy mechanisms for three tasks: recovering a single covariance coefficient, recovering the spectral density at a fixed frequency, and global recovery. Our approach achieves faster rates through a two-stage process: we first apply the Laplace mechanism to the truncated value, and then use the resulting privatized sample to learn about the dependence mechanism in the time series.
For spectral densities belonging to Hölder and Sobolev smoothness classes, we demonstrate that our algorithms improve upon the non-interactive mechanism of Kroll (2024) for small privacy parameter a, since the pointwise rates depend on na^2 instead of na^4. Moreover, we show that the rate 1/(na^4) is optimal for estimating a covariance coefficient with non-interactive mechanisms. However, the L2-rate of our interactive estimator is slower than the pointwise rate. We show how to use these procedures to provide a bona fide locally differentially private estimator of the entire covariance matrix. A simulation study validates our findings.

Advances in Survival Analysis

13:30 - 14:00

Goodness-of-fit tests under dependent censoring

A. Lago[1], J.C. Pardo-Fernández[1], I. Van Keilegom[2]

[1]Universidade de Vigo, Spain
[2]KU Leuven, Belgium

A large number of statistical methods for right-censored data rely on the key assumption of independence between the target and censoring variables. However, this assumption is often violated in real applications and, as reported in several studies, may lead to inconsistent statistical procedures.

In this talk, we propose new goodness-of-fit tests for right-censored data that account for dependence between the target and censoring variables through an Archimedean copula framework. The asymptotic distribution of the proposed test statistics is derived under the null hypothesis, and the consistency of the proposed methodology is also established. Since the direct application of these asymptotic results may be challenging in practice, a bootstrap resampling scheme is developed to approximate the null distribution of the tests. The finite-sample performance of the proposed procedures is evaluated through Monte Carlo simulations, and the methodology is illustrated with a real dataset.

Unsure about the Markov assumption? A comparison of transition probability estimators in multi-state models

C. Drenda[1], D. Dobler[2], M. Munko[3], A. Titman[4]

[1]TU Dortmund University, Germany
[2]RWTH Aachen University, Germany
[3]Otto von Guericke University Magdeburg, Germany
[4]Lancaster University, United Kingdom

Multi-state models extend classical survival models by allowing transitions between multiple states over time. Various estimators for modelling the transition probabilities in multi-state models have been proposed, e.g., the Aalen-Johansen estimator, the landmark Aalen-Johansen estimator, and a hybrid Aalen-Johansen estimator. While the Aalen-Johansen estimator is generally only consistent under the rather restrictive Markov assumption, the landmark Aalen-Johansen estimator can handle non-Markov multi-state models. However, the landmark Aalen-Johansen estimator leads to a strict data reduction and, thus, to an increased variance. The hybrid Aalen-Johansen estimator serves as a compromise by, firstly, checking with a log-rank-based test whether the Markov assumption is satisfied. Secondly, landmarking is only applied if the Markov assumption is rejected.
In this work, we propose a new hybrid Aalen-Johansen estimator that uses a Cox model instead of the log-rank-based test to check the Markov assumption in the first step. Furthermore, we compare the four estimators in an extensive simulation study across Markov, semi-Markov, and distinct non-Markov settings. In order to get deep insights into the performance of the estimators, we consider four different measures: bias, variance, root mean squared error, and coverage rate. Additionally, further influential factors on the estimators, such as the form and degree of non-Markov behaviour, the different transitions, and the starting time, are analysed. The main result of the simulation study is that the hybrid Aalen-Johansen estimators yield favourable results across various measures and settings.

Survival analysis under label shift

Y. Zong[1], Y. Ma[2], I. Van Keilegom[1]

[1]KU Leuven, Belgium
[2]Penn State University, United States

We study a setting with a source population P where both the covariates Z and the response T are observed, and a target population Q, where only the covariates Z are available. The two populations have different joint distributions but share the same conditional distribution of covariates Z given the response T, which is known as label shift. In the source population P, the response T is subject to random censoring.

Our goal is to estimate parameters of interest in the target population Q by using information from the label-shifted and label-censored source population P. We propose a class of semiparametric models for T given Z in Q and develop a sieve maximum likelihood estimation method. In this approach, the baseline hazard function in Q is approximated by Bernstein polynomials, and the marginal hazard function of T in P is estimated by nonparametric maximum likelihood. Our method allows conditional censoring given the covariates in P and provides flexibility that includes a wide range of classical survival models. The asymptotic properties of the proposed estimator are established, and its performance is assessed through simulation studies and a real data example.

Conformal prediction for time-to-event outcomes subject to truncation and censoring

R. Betensky[1], W. Wang[2], J. Qian[2]

[1]NYU School of Global Public Health, United States
[2]University of Massachusetts, United States

Accurate prediction of clinically meaningful times-to-event is essential for patient prognosis, optimal resource allocation, such as insurance coverage for costly Alzheimer's disease drugs and determination of eligibility for clinical trials. Conformal prediction offers a flexible framework for quantifying uncertainty with arbitrary prediction algorithms, yielding prediction intervals that achieve valid marginal coverage without requiring distributional assumptions. Although conformal prediction methods have been recently developed for censored data, corresponding methods for truncated data are still lacking. We propose a conformal prediction method for left-truncated and right-censored data, enabled by inverse probability of censoring and truncation weighting. We further extend the approach to accommodate additional forms of truncation, including right, double, and sequential truncation. Simulation studies and semi-synthetic examples demonstrate the effectiveness and robustness of our methods across diverse settings

Topics on high-dimensional and complex data

13:30 - 14:00

Cross-temporal forecast reconciliation using machine learning

I. Wilms[1], M. Ternes[1], J. Rombouts[2]

[1]Maastricht University, Netherlands
[2]ESSEC, France

Many forecasting tasks involve multiple, interrelated time series that must satisfy linear aggregation constraints, where the components collectively sum to the total. Ensuring such coherence across all aggregation levels is the goal of forecast reconciliation, which is essential for consistent and aligned decision-making. In cross-temporal frameworks, the focus of this talk, these aggregation constraints extend across both cross-sectional and temporal dimensions. Existing literature primarily relies on linear reconciliation methods, which adjust base forecasts through linear transformations within a least-squares framework to satisfy aggregation constraints. In this work, we move beyond this paradigm and introduce a non-linear forecast reconciliation approach for cross-temporal frameworks. Our method directly and automatically produces cross-temporal coherent forecasts by leveraging popular machine learning techniques. We empirically validate our framework on large-scale streaming datasets from a leading European on-demand delivery platform and a bicycle-sharing system in New York City.

Hypothesis Testing for Penalized Estimating Equations with Cross-Fitted Covariance Calibration

J. Zhou[1]

[1]University of Manchester, United Kingdom

We study hypothesis testing for penalized estimators in settings where the full marginal distribution of a multivariate response is difficult to specify, such as longitudinal data with correlated measurements or high-dimensional heteroscedastic regression. Assuming that the conditional mean model is correctly specified, we establish that the penalized estimating equations admit a $\sqrt{n}$-consistent solution, even when the working covariance structure is misspecified. Our inferential target is a low-dimensional subvector of parameters associated with the mean model. We show that the resulting test statistic converges to a χ^2 distribution, and that its asymptotic power depends on the nuisance covariance function. To mitigate this dependence, we propose estimating the covariance function via cross-fitting, which provides a calibrated and robust procedure for inference.

Selective Inference in DAGs

S. Guglielmini[1], G. Claeskens[1]

[1]KU Leuven, Belgium

Directed Acyclic Graphs, or DAGs, offer an interpretable framework for modelling directed dependency structures in multivariate data. Estimating a DAG requires specific sparsity constraints, producing a data-dependent selected model for which the inference guarantees of classical statistical theory do not hold. We consider DAG selection via regularization and pruning, with added Gaussian random noise, and propose a method for inference conditional on both selection steps. By applying a change of variable, we derive the conditional density of the estimated selected parameters. The additional randomization yields a valid post-selection likelihood, enabling inference in the selected DAG with controlled type I error rate. The proposed approach achieves valid inference while improving the trade-off between selection accuracy and confidence interval length when compared to existing methods.

Revisiting online sufficient dimension reduction

A. Artemiou[1]

[1]University of Limassol, Cyprus

Sufficient Dimension Reduction (SDR) methods are supervised dimension reduction methods which have become very popular in the last 35 years due to their computational power and simplicity. Recently there was an interest on real-time SDR methods. Real time SDR methods have been developed for algorithms which belong to two subclasses of the sDR methodology. For Sliced- Inverse Regression (SIR) for inverse-moment based class and for Principal Least Squares Support Vector Machines (SVM) for algorithms in the SVM-based class. In this talk we will talk about some recent developments on real-time algorithms for the SVM-based class

Contributed: Testing, Goodness-of-Fit and Distribution Theory

13:30 - 13:50

Goodness-of-fit tests for infinite-dimensional regression models: the kernel approach

D. Diz-Castro[1], M. Febrero-Bande[1], W. González-Manteiga[1]

[1]University of Santiago de Compostela, Spain

In this contribution, we address the problem of testing the goodness-of-fit (GoF) of regression models involving infinite-dimensional parameters and covariates. In particular, we explore a kernel-based approach and show that the theoretical developments introduced in recent proposals for the Euclidean case can be extended to more general settings in which the parameters take values in Banach spaces and the covariates take values in Polish spaces. To this end, we introduce the concepts of quasi Neyman orthogonalization of kernels and representatives of Fréchet derivatives. We propose an easily computable test statistic and show that, under suitable regularity conditions on the regression function, the model errors, and the convergence rate of an estimator of the unknown parameters, the limiting distribution of the test statistic exhibits the same behavior as in the finite-dimensional case. Moreover, the test can be consistently calibrated using a multiplier bootstrap scheme. Finally, we illustrate the finite-sample performance of the procedure through a succinct simulation study involving functional data.

Goodness-of-Fit testing for the hazard rate function

P. Mavridis[1], D. Bagkavos[1]

[1]Department of Mathematics, University of Ioannina, Greece

This paper introduces a novel goodness-of-fit test of a continuous parametric hazard rate function. Construction of the test is based on the integrated square error between the classical nonparametric kernel hazard rate estimate and the parametrically estimated hazard function under the null model. The theoretical contributions of the article include analytic quantification of the test statistic's asymptotic distribution under both the null and alternative hypotheses, including closed-form expressions for its asymptotic power under Pitman local alternatives. The power of the test for finite samples is also investigated numerically via distributional data simulations which reveal excellent performance under various sample sizes and amounts of censoring. Finally, the practical usefulness of the new test is demonstrated in the analysis of a real-world dataset.

Testing homogeneity across circular k-samples

A. Fernández de Marcos[1], E. García Portugués[1]

[1]Universidad Carlos III Madrid, Spain

We develop a unified framework for k-sample homogeneity testing on the circle based on Sobolev statistics. The proposed class of k-sample Sobolev tests generalizes the two-sample Sobolev tests based on uniform scores, and encompasses the existing multisample tests as particular cases. Within this class, we introduce the first Anderson-Darling-type homogeneity test for circular data, and we further propose two new tests designed to detect multimodal departures from homogeneity, constructed from Softmax and Poisson kernels. We derive the asymptotic null distribution of the class and prove its consistency against a broad class of fixed alternatives. The tests are distribution-free and therefore do not require resampling. A comprehensive simulation study demonstrates the superior power of the Anderson-Darling-type test compared with Cramér-von Mises-type competitors across several scenarios, and the effectiveness of multimodal tests under suitable alternatives. The practical relevance of the methodology is illustrated through a real-data application arising in zoology.

A test for the equality of a U-estimable parameter

M. Romero-Madroñal[1], M.R. Sillero-Denamiel[1], M.D. Jiménez-Gamero[1]

[1]Universidad de Sevilla, Spain

In this talk, we propose a general, unified framework for testing the equality of a broad class of estimable parameters, defined via U-statistics, across multiple independent populations. This framework encompasses various common statistical problems, such as comparing variances, correlation coefficients, and Gini indices. We consider two test statistics: a Wald-type and an ANOVA-type. While the asymptotic distribution of the former is derived under a fixed-dimension regime, the latter is analyzed in both fixed and high-dimensional settings, where the parameter dimension is permitted to grow with the sample size. These results yield testing procedures that enable asymptotically exact inference without parametric assumptions. We also discuss an alternative approach to approximating the null distribution based on a weighted bootstrap. The performance of the proposed procedures is illustrated through simulations, and their practical utility is shown via an application to a real dataset.

Testing exponentiality for fixed and diverging number of populations

J.S. Allison[1], M.D. Jiménez-Gamero[2], M.V. Alba-Fernández[3], L. Santana[4]

[1]University of South Africa, South Africa
[2]Universidad de Sevilla, Spain
[3]Universidad de Jaén, Spain
[4]North-West University, South Africa

We study the problem of simultaneously testing whether k separate populations of nonnegative values each have an exponential law, potentially with differing parameter values, by using k independent samples drawn from each population. Two asymptotic settings are considered: (i) k is assumed to be fixed, and the sample sizes are allowed to increase; and (ii) k is assumed to be large, while the sample sizes may be bounded or increase with k.

Contributed: Quantiles, Ranks and Robust Inference

13:30 - 13:50

Massive parallelization of projection-based depths

L. Leone[1], D. Bounie[2], P. Mozharovskyi[3]

[1]Institut Polytechnique de Paris, France
[2]CREST, Telecom Paris, France
[3]LTCI, Telecom Paris, France

Providing a statistically-meaningful center-outward ordering for a data set constitutes a powerful tool of analysis and inference over a range of fields, such as fraud detection being one of them. In the multivariate setting, due to complex character of contemporary data sets, assumptions on the data-generating process should be avoided as much as possible, while maintaining high level of data description is mandatory. While absence of assumptions limits statistical modelling methodology, absence of natural ordering in the Euclidean space impedes univariate tools. Being a non-parametric and robust technique that, in an agnostic way, generalizes distribution function and quantiles to higher dimensions, statistical data depth function comes as a remedy for multivariate anomaly detection.

Despite numerous recent advances in the field of data depth, the computational complexity and time consumption are often returned as critics. In this article, projection depth notions are studied as a fast and efficient anomaly detection method in the multivariate context. Due to the nature of the projection depth algorithm, the growth in big data technologies and hardware (such as multiple-kernel processors and graphics processing units) availability, computation can be optimized by parallelizing its non-sequential constituent without precision loss.

The article introduces a novel methodology for the massive parallelization of projection-based depths, addressing the computational challenges of data depth in high-dimensional spaces. We propose an algorithmic framework based on Refined Random Search (RRS) and demonstrate significant speedup (up to 7,000 times faster) on GPUs. Empirical results on synthetic data show improved precision and reduced runtime, making the method suitable for large-scale applications. The RRS algorithm (and other depth functions) are available in the Python-library data-depth with ready-to-use tools to implement and to build upon this work.

Nonparametric inference based on expected order statistics

T. Lando[1]

[1]University of Bergamo, Italy

As well known, the j-th order statistic of a sample of size m from a random variable X is defined as the j-th smallest value in the sample and is denoted by X_{j:m}. These quantities play an important role in nonparametric inference. Given a random sample of size n (possibly different from m) drawn from X, we investigate the estimation of the expected order statistics E[X_{j:m}] and, more importantly, the development of novel statistical methods based on these estimators. We introduce L-estimators for the expected order statistics and establish their almost sure (a.s.) consistency, even in scenarios where the mean of X is not finite. Letting F denote the cumulative distribution function (CDF) of X, and G a reference (known) distribution, these estimators serve as the foundation for tests assessing the convexity or concavity of the composition G^{-1}(F), as compared to equality in distribution. This approach includes well-known cases, such as distributions with increasing hazard rate, and extends to other relevant distribution families that have recently received growing interest. The proposed tests are shown to be consistent and unbiased. Additionally, we propose a novel nonparametric estimator of the distribution function F, constructed from the estimated expected order statistics.

Quasi-Monte Carlo confidence intervals using quantiles of randomized nets

Z. Pan[1]

[1]Zhejiang University, China

Recent advances in quasi-Monte Carlo integration have shown that for linearly scrambled digital net estimators, the convergence rate can be dramatically improved by taking the median rather than the mean of multiple independent replicates. In this work, we demonstrate that the quantiles of such estimators can be used to construct confidence intervals with asymptotically valid coverage for high-dimensional integrals. By analyzing the error distribution for a class of infinitely differentiable integrands, we prove that as the sample size increases, the integration error decomposes into an asymptotically symmetric component and a vanishing remainder. Consequently, the asymptotic error distribution is symmetric about zero, ensuring that a quantile-based interval constructed from independent replicates captures the true integral with probability converging to a nominal level determined by the binomial distribution.

A Unified Approach for Rank-Based Trend Inference with Missing Values

M. Geroldinger[1]

[1]Research Program Biomedical Data Science, Paracelsus Medical University, Salzburg, Austria

Testing for monotonic trends across groups is a common objective in clinical studies, particularly in small samples such as in research on rare diseases, where nonparametric methods are often preferred as parametric assumptions are frequently violated. In practice, however, such analyses are often complicated by missing observations resulting from study dropouts or incomplete measurements, which can compromise the validity of standard procedures, especially when the sample size is limited.
We propose a unified, nonparametric, rank-based approach for trend tests in univariate settings under the 'Missing Completely At Random' (MCAR) assumption, covering both independent and dependent samples. The methodology is based on linear rank statistics and incorporates inverse probability weighting to account for missing values. A key component is a Horvitz–Thompson-type variance estimator, which provides an unbiased and consistent variance estimate in the presence of missing data and remains valid under dependent data. The proposed estimator was also compared with other variance estimators to demonstrate its superior performance.
Simulation studies show that the proposed methodology ensures reliable control of type-I error and high statistical power across a wide range of scenarios, including limited sample sizes and a high proportion of missing data.
The proposed framework provides a flexible and robust tool for nonparametric trend testing of univariate independent and dependent data with missing values.

Shape Constrained Statistical Inference

16:00 - 16:30

The log-concave MLE in high dimensions -- adaptation and suboptimality for densities with polytopal support

A. Guntuboyina[1], G. Kur[2]

[1]University of California Berkeley, United States
[2]ETH Zurich, Switzerland

We study the adaptive behavior of the log-concave Maximum Likelihood Estimator (MLE) across all dimensions. While the log-concave MLE is known to be globally minimax optimal over the full class of log-concave densities, its performance on structured subclasses has been well understood only for dimensions up to three. We extend these results to all dimensions and uncover a qualitative shift in behavior beginning at dimension five, where the estimator exhibits fundamentally different adaptive properties compared to lower dimensions. This is joint work with Gil Kur from ETH Zurich.

Testing sufficient follow-up in cure models via monotone tail density constraints

T.P. Yuen[1], E. Musta[2], I. Van Keilegom[1]

[1]KU Leuven, Belgium
[2]University of Amsterdam, Netherlands

Cure models are used in survival analysis when a fraction of subjects will never experience the event of interest. Estimating cure rates reliably requires sufficiently long follow-up, but determining how long is "long enough" remains challenging. In practice, follow-up can be considered sufficient if the probability of an event occurring after the study ends is negligibly small. We develop a nonparametric test for this condition by exploiting monotonicity of the density function of the survival times in the tail region and then extend this approach to account for categorical covariates. A naive intersection-union approach, requiring rejection of insufficient follow-up for every covariate value, is overly conservative in practice and lacks power. To improve upon this, we propose a new test procedure that relies on the test decision for one properly chosen covariate value. We show that both methods yield tests of asymptotically level α and investigate their finite sample performance through simulations. The practical application of the methods is illustrated using a melanoma dataset.

Isotonic distributional regression with different notions of order and constraints

L. Duembgen[1]

[1]University of Bern, Switzerland

Distributional regression aims at estimating the conditional distributions of a response Y, given a covariate X. A nonparametric approach is to assume isotonicity with respect to some stochastic order. We discuss briefly the usual stochastic order and the likelihood ratio order, presenting a new algorithm for the latter setting. Then we illustrate the potential benefits of the additional assumption the the conditional laws are log-concave.

Bayesian Nonparametric Monotone Regression: Contraction Rates and Coverage of Credible Intervals

M. Chakraborty[1]

[1]University of Texas Medical Branch-Galveston, United States

For nonparametric univariate regression under monotonicity constraint on the regression function, we study global posterior contraction rate and the point-wise coverage of a Bayesian credible interval. For point estimation and credible sets, we use a finite random series of piece-wise constant functions with normal basis coefficients as a prior for the regression function. Compared to a prior distribution directly complying with the monotonicity constraint, a lot more tractability in the contraction rates and asymptotic coverage is achieved by considering a ``projection posterior" distribution of the regression function, where a sample from the resulting conjugate posterior distribution is projected on the space of monotone increasing functions to obtain a monotone function closest to the true function, inducing the projection posterior. We show that the projection posterior contracts at the minimax rate with respect to the L1-distance, and also with respect to the empirical Lp-distances up to a logarithmic factor. For the limiting coverage of a credible interval of the regression function evaluated at a fixed point, we observe an interesting phenomenon that the coverage may be higher than the nominal credibility level, the opposite of a phenomenon observed by Cox in the context of smooth nonparametric signal estimation. We then show that a re-calibration technique can give the right coverage, leading us to introduce the ``Bayes-Chernoff" distribution mapping the asymptotic coverage of this credible interval to its nominal credibility level. The reverse-Cox phenomenon following the Bayes-Chernoff distribution is shown to emerge in other shape constrained problems such as univariate monotone probability density estimation and regression quantile estimation for a univariate monotone regression problem, in addition to the multivariate extensions of these problems.

Uncertainty Quantification for Machine Learning

16:00 - 16:30

Diffusion based time series inference

L. Chen[1], W. Wang[2]

[1]Washington University in St Louis, United States
[2]University of Bristol, United Kingdom

We propose a diffusion-based framework for conditional density estimation in time series forecasting, designed to capture complex, nonlinear, and potentially multimodal predictive distributions. Unlike conventional forecasting methods that focus mainly on conditional means or impose restrictive parametric assumptions, our approach directly models the full conditional distribution of future observations given past information. To improve interpretability and performance in high-dimensional settings, we incorporate a lasso penalty into the conditional input layer of the network. This regularization encourages sparsity in the conditioning variables, effectively selecting the most relevant lagged observations and exogenous predictors while suppressing irrelevant or noisy inputs. This hybrid framework is particularly suitable for modern forecasting problems where the dependence structure may be complex and the number of candidate predictors can be large.

Online Monitoring of Time Series via Maximum Mean Discrepancy

B.C. Boniece[1], L. Horváth[2], L. Trapani[3]

[1]Drexel University, United States
[2]University of Utah, United States
[3]University of Pavia, Italy

Kernel two-sample methods based on maximum mean discrepancy (MMD) provide a flexible tool for detecting distributional differences in complex data. In this talk, we develop online monitoring procedures for time series based on MMD, aimed at detecting general distributional shifts without specifying the type of structural change in advance.

The proposed statistics are constructed from degenerate U-statistics that compare an historical baseline sample with incoming observations. We establish asymptotic theory for a general class of monitoring schemes under weakly dependent time series and develop feasible tests via spectral approximations of the kernel operator.

This perspective connects kernel two-sample testing with sequential change detection and highlights consideration of the choice of kernels in dependent data settings.

Conformalized Percentile Interval

W. Zhu[1], R. Zou[1], B. Nan[1]

[1]University of California Irvine, United States

Conformal prediction usually provides distribution-free predictive intervals with finite-sample marginal coverage, but achieving conditional validity remains challenging, particularly in complex regression settings with heteroskedasticity and skewed or heavy-tailed responses. We propose a method that combines conditional distribution estimation via neural networks with conformal-style calibration on transformed responses via the probability integral transformation (PIT) to construct a finite-sample--adjusted percentile interval. Calibrating in PIT space is effective because PIT values are asymptotically feature-independent when the CDF estimator is consistent, which mitigates feature-dependent miscoverage and improves conditional calibration. We further introduce an adjustment step that improves efficiency by shortening intervals while preserving validity, and we prove finite-sample marginal coverage and asymptotic conditional coverage under mild consistency conditions. Experiments on diverse synthetic and real-world benchmarks demonstrate stronger conditional calibration and substantially shorter intervals than existing methods.

Domain Adaptation & Semi-supervised Methods

16:00 - 16:30

Several Studies in Label Shift

Y. MA[1]

[1]PSU, United States

We provide an introduction to label shift problems.
In the context of discrete response, we study the importance weights confidence set problem by a paradigm shift from traditional inversion-based inference to a direct matrix constraint framework. We use this framework to characterize a joint confidence region and extract marginal intervals via linear programming, deriving provably tighter bounds for importance weights while maintaining exact finite-sample validity.
In the context of continuous response, we study the estimation and inference of a general target population characteristic by developing doubly and singly robust estimators as well as the efficient estimator. Many ongoing and future developments will be discussed too.

Semi-supervised Clustering Through Representation Learning of High-dimensional Count Data

M. Li[1], L. Wang[2], Z. Xia[3], M. Liu[4], T. Cai[2]

[1]Bentley University, United States
[2]Harvard University, United States
[3]University of Pittsburgh, United States
[4]Peking University, United States

High-dimensional count data arise in many modern applications, where observations are sparse, heterogeneous, and driven by complex latent structure. These characteristics make modeling and prediction challenging, especially when labels are limited. We propose SCORE, a semi-supervised representation learning framework for clustering and embedding construction from high-dimensional count features. SCORE is built on a Poisson-adapted latent factor mixture model that supports incorporating external pre-trained feature embeddings when available, enabling efficient characterization of feature patterns and extraction of meaningful latent cluster membership and low-dimensional representations. To scale inference and learning to large datasets, SCORE uses a hybrid algorithm that combines expectation maximization with Gaussian variational approximation, leveraging a small labeled subset to refine estimation on a large pool of unlabeled samples. We establish convergence guarantees for this hybrid procedure, quantify approximation errors from Gaussian variational approximation, and derive error rates under increasing embedding dimensions. Our theory and experiments show that incorporating unlabeled data improves accuracy and reduces sensitivity to label scarcity, yielding strong finite sample performance relative to existing approaches. We demonstrate SCORE on electronic health record count features for predicting disability status in multiple sclerosis, where it learns informative and predictive patient embeddings, illustrating its practical value.

Unsupervised domain adaptation beyond label shift

X. de Luna[1], M. Ghasempour[1], Y. Ma[2]

[1]Umeå University, Sweden
[2]Penn State University, United States

Consider the motivating example where we have K groups (source labelled datasets) born in consecutive years, and a target unlabelled group T born in year K+1. Labels Y represent the number of days of hospitalisation for an individual in any of the groups the year they turn 80 years old. Since cohort T is 79 years old, we do not have their labels. However, we want to estimate the expected number of hospitalisation days for this group at 80 year of age, using labels from all other groups and covariates X available for all datasets. This situation falls into the realm of unsupervised domain adaptation. We introduce and study a framework beyond label-shift to perform inference on the expected number of hospitalisation days and other general target estimands. We leave the density functions for Y given unspecified in all cohorts, while allowing for a conditional shift from one group to another, whereby the density function of X given Y for cohort K+1 is modelled as an exponential tilt of the density function of X given Y for cohort K. We provide conditions for identifying the target estimand, deduce the efficient influence function, and thereby propose semiparametric estimators, which are consistent, asymptotically normal, and locally efficient. Several nuisance functions need to be estimated, and we propose feasible and robust estimators allowing for the simultaneous misspecification of three of the nuisance functions. The implementation of the estimators is provided to complement the proven asymptotic theory and to illustrate the finite sample performance of the estimators and their inference. If time allows a hospitalization data from Swedish registers will be used to illustrate the results.

Flexible Deep Neural Networks for Partially Linear Survival Data

M. Gorfine[1], A. Ben Arie[1]

[1]Tel Aviv University, Israel

We propose a flexible deep neural network (DNN) framework for modeling survival data within a partially linear regression structure. The approach preserves interpretability through a parametric linear component for covariates of primary interest, while a nonparametric DNN component captures complex time–covariate interactions among nuisance variables. We refer to the method as FLEXI–Haz, a FLEXIble Hazard model with a partially linear structure. In contrast to existing DNN approaches for partially linear Cox models, FLEXI–Haz does not rely on the proportional hazards assumption. We establish theoretical guarantees: the neural network component attains minimax optimal convergence rates that depends on composite Holder classes, and the linear estimator is sqrt-n-consistent, asymptotically normal, and semiparametrically efficient. Extensive simulations and real-data analyses demonstrate that FLEXI–Haz provides accurate estimation of the linear effect, offering a principled and interpretable alternative to modern methods based on proportional hazards. Code for implementing FLEXI-Haz, as well as scripts for reproducing data analyses and simulations is available at GitHub site
https://github.com/AsafBanana/FLEXI-Haz.

Statistical inference for complex data structures

16:00 - 16:30

Inference under Long-Range Dependence in the Presence of Hermite Ranks

D. Nordman[1], S. Lahiri[2], H. Yu[3]

[1]Iowa State University, United States
[2]Washington University, United States
[3]University of Rhode Island, United States

Long-range dependent time series are often formulated by some unknown transformation of an underlying long-memory Gaussian process. The so-called Hermite rank of this transformation is a process parameter that critically impacts statistical inference because sampling distributions change depending on this rank. A compounding issue is that the rank is typically unknown and can further vary between statistics computed from the same time series. Furthermore, over the past 50 years, no approach has existed to generally approximate Hermite ranks from data. This talk describes a method for approximating both the Hermite rank as well as dependence parameter of the underlying Gaussian process, without knowledge of the underlying transformation that defines the observed long-memory time series. The estimation approach can also be coupled with a bootstrap method for approximating the sampling distribution of statistics in practice. The inference method is illustrated through numerical studies and examples.

Conformal Prediction for Dyadic Regression with Structured Missingness

R. Lunde[1], M. Yang[2], E. Levina[3], J. Zhu[3]

[1]Washington University in Saint Louis, United States
[2]Washington University in St. Louis, United States
[3]University of Michigan, United States

Dyadic regression, which involves modeling a relational matrix given covariate information, is an important task in statistical network analysis. We
consider uncertainty quantification for dyadic regression models using conformal prediction. We establish finite-sample validity of our procedures for various sampling mechanisms under a joint exchangeability assumption. We also show that, under certain conditions, it is possible to construct asymptotically valid prediction intervals for a missing entry under a
structured missingness assumption.

Statistical inference for subgraph counts using network sampling in a sparse Stochastic Block Model framework

A. Mandal[1], A. Chatterjee[1]

[1]Indian Statistical Institute, Delhi, India

We obtain limit laws for network sampling based estimates of subgraph counts and clustering coefficient of a large population network, and use them for predictive inference. A model based approach is used, where the population network is assumed to be generated from a sparse Stochastic Block Model (SBM). In order to quantify the effects of node sampling under resource constraints, a sparse Bernoulli node sampling scheme is introduced, where the node selection probability is allowed to decay to zero as the population size increases. Both induced and ego-centric network formation approaches are explored. Quantitative bounds on the speed of normal approximation for estimated subgraph counts are obtained in a joint model and design based asymptotic framework. We find that the accuracy of statistical inference depends intricately on the level of sparsity in the model, the sparsity level of the sampling scheme, and on features like the edge density and minimum vertex cover size of the target subgraph. The ego-centric approach requires a very delicate analysis and it can handle higher levels of model and sampling sparsity. We also find that, for every target subgraph, the level of model sparsity has no effect on the quality of inference if it stays below an initial threshold. Beyond which, the quality degrades rapidly. The sufficient conditions for obtaining a Gaussian limit law also turn out to be necessary. For strictly balanced target subgraphs, we obtain sharp transitions from Gaussian to various types of Poisson based limit laws, as the model and sampling sparsity levels increase. A complete description of all possible limit laws for estimated subgraph counts is found in the induced case, and a near-complete description is obtained in the ego-centric case. A simulation study provides strong support for the theoretical results at various levels of model sparsity and sampling sparsity.

Statistical Inference under snowball sampling

S. Lahiri[1]

[1]Washington University in St Louis, United States

We investigate asymptotic distributional properties of subgraph counts under snowball sampling and consider applications of the results to statstical inference on the population subgraph counts.

Advances in nonparametric estimation

16:00 - 16:30

ADAPTIVE ESTIMATION OF L2-NORM OF A PROBABILITY DENSITY AND RELATED TOPICS

O. Lepski[1]

[1]Aix Marseille Université, France

We deal with the problem of the adaptive estimation of the L2–
norm of a probability density on R^d, d ≥ 1, from independent observations.
The unknown density is assumed to be uniformly bounded
and to belong to the union of balls in the isotropic/anisotropic Nikolskii's
spaces. We prove that the optimally adaptive estimators
over the collection of considered functional classes do no exist. Also,
in the framework of an abstract density model we present several
generic lower bounds related to the adaptive estimation of an arbitrary
functional of a probability density. Next, we show that these bounds
are tight and present the adaptive estimator which is obtained by a
data-driven selection from a family of kernel-based estimators. The
proposed estimation procedure as well as the computation of its risk
are heavily based on new concentration inequalities for decoupled U-statistics.
It is also worth noting that all obtained results are derived from the unique oracle inequality.

Estimating a regression function under weak assumptions on the errors: Application to shape-constrained regression

Y. Baraud[1], G. Maillard[2]

[1]University of Luxembourg, Luxembourg
[2]ENSAI, France

In collaboration with Guillaume Maillard, we focus on estimating a regression function under weak assumptions on the error distribution. In particular, we do not assume that the errors are i.i.d. with a finite variance or possess an exponential moment; we only assume that they are independent and centered (hence integrable). In particular, they may be heteroscedastic whenever they are square-integrable.

Within this statistical framework, we present a generic estimation method that yields estimators whose performance automatically adapts to the integrability properties of the errors. Moreover, when the errors are square-integrable and heteroscedastic, we show that our procedure yields stable and (nearly) rate-optimal estimators, even in settings where classical least squares are not consistent.

We provide non-asymptotic risk bounds for our estimator and illustrate them in the specific setting of estimating a regression function under shape constraints. The constraints we consider include monotonicity, unimodality, convexity, among others. We show that the estimator is not only robust with respect to this a priori shape assumption but also exhibits remarkable adaptive properties when the target regression function possesses additional structure.

Unmatched regression: Asymptotic results under identifiability

F. Balabdaoui[1]

[1]ETHZ, Switzerland

Consider the regression problem where the response and the covariate are unmatched. Under this scenario we do not have access to pairs of observations from their joint distribution, but instead we have separate data sets of responses and covariates, possibly collected from different sources. We study this problem assuming that the regression function is linear and the noise distribution is known or can be estimated. We introduce an estimator of the regression vector based on deconvolution (the DLSE) and demonstrate its consistency and asymptotic normality under parametric identifiability. Under non-identifiability of the regression vector but identifiability of the distribution of the predictor, we construct an estimator of the latter based on the DLSE and show that it converges to the true distribution of the predictor at the parametric rate in the Wasserstein distance of order 1. We illustrate the theory with several simulation results.

Reliable error bounds for sparse recovery by convex optimization

A. Juditsky[1], A. Nemirovski[2]

[1]University Grenoble-Alpes, France
[2]Georgia Institute of Technology, United States

Our focus is on the problem of recovery of an uknown signal from a linear noisy observation. We assume that 1) that the signal in question is sparse, i.e., has a "small number" of nonvanishing entries, and 2) is known to be belong to a given convex set. Our objective is to build an estimate of the signal along with corresponding bounds of the estimation error. In the "standard" setting of this problem, the "sensing matrix" (regressor matrix) is either assumed to be "randomly generated" or is expected to satisfy an RIP, restricted eigenvalue, or compatibility conditions. Such assumptions lead to theoretically sound bounds of the estimation risk; however, they are hard to verify, making construction of reliable recovery error bounds essentially impossible.

In this work, we analyze the performance of polyhedral estimates---a particular class of nonlinear estimates as introduced in A. Juditsky and A. Nemirovski: On polyhedral estimation of signals via indirect observations. Electronic Journal of Statistics, 14(1):458––502, 2020.---and show how "presumably good" estimates of the sort may be
constructed by means of efficient convex optimization in the situation where a computationally tractable signal set.

Topics in Time Series and Functional Data Analysis

16:00 - 16:30

Learning population and individual structure in dynamic networks with degree heterogeneity

R. von Sachs[1], M. Li[1], E. Pircalabelu[1]

[1]UCLouvain, Belgium

Dynamic networks provide a powerful framework for characterizing time-varying functional connectivity in neuroimaging studies. In practice, such networks are typically collected from multiple subjects across time and exhibit both temporal dynamics and subject-specific heterogeneity. Brain functional connectivity networks also contain hub nodes, defined as highly connected regions that play critical roles in understanding brain functional connectivity. In this talk, we propose a mixed-effect dynamic stochastic block model with degree heterogeneity, which simultaneously disentangles the population connectivity structure from individual variability and recovers the trajectories of hub regions through time-varying degree parameters. We develop an efficient local approximate estimation procedure and evaluate its performance through extensive simulations and a case study of dynamic functional connectivity from the Human Connectome Project.

Some contributions to changepoint analysis

J. Freyermuth[1]

[1]Aix-Marseille Université, France

This talk consists of two parts. First, we introduce a novel framework for detecting regime changes in non-stationary data from multi-trial experiments. Our approach focuses on changes occurring across trials and is robust to a range of within-trial distortions, such as temporal misalignment and subtle frequency modulations. Second, we discuss the application of a well-established changepoint detection method to the spectral characteristics of multivariate EEG time series. The analysis is performed on resting-state EEG data collected from a large cohort of participants, together with an extensive battery of psychological and behavioral measures. This allows us to reveal new insights into brain dynamics.

Topics on Multivariate Integer-Valued Time Series

K. Fokianos[1]

[1]University of Cyprus, Cyprus

We study inference and modelling techniques for multivariate count time series. In particular, we focus on autoregressive, network, and space–time models. The development of both the underlying theory and its applications depends critically on the choice of an appropriate response distribution for multivariate count data, as well as on the framework of observation driven models. Our aim is to highlight recent methodological advances in this area and to outline several directions for future research.

Graph estimation based on multivariate functional data with different coarseness scales

E. Pircalabelu[1]

[1]UC Louvain, Belgium

We develop a high-dimensional graphical modelling approach for functional data where the number of functions exceeds the available sample size. This is accomplished by proposing a sparse estimator for a concentration matrix when identifying linear manifolds and by making use of regularized procedures that estimate sparse undirected graphical models. By doing so, one also gets insight into the conditional independence structure that governs the multivariate functional data, while at the same time estimating linear combinations of the components.

By working in a penalized setting our contribution enriches the functional data framework by estimating sparse undirected graphs that show how functional nodes connect to other functional nodes. As such, the proposed procedure extends the ideas of the manifold representation for functional data to high-dimensional settings where the number of functions is larger than the sample size. It allows multiple coarseness scales to be present in the data and proposes a simultaneous estimation of several related graphs, that are constrained to share similarities due to the design of the problem.

Simulated and real data examples show beneficial performance pointing to the procedure being a useful tool for modeling multiscale, multivariate functional data.

Statistical machine learning for complex data

16:00 - 16:30

Autotune: fast, accurate, and automatic tuning parameter selection for Lasso

S. Basu[1]

[1]Cornell Uniersity, United States

Least absolute shrinkage and selection operator (Lasso), a popular method for high-dimensional regression, is now used widely for estimating high-dimensional time series models such as the vector autoregression (VAR). Selecting its tuning parameter efficiently and accurately remains a challenge, despite the abundance of available methods for doing so. We propose autotune, a strategy for Lasso to automatically tune itself by alternately estimating regression coefficients and noise standard deviation. Simulation experiments on regression and VAR models shows that autotune is faster than alternatives, and more accurate in low signal-to-noise regimes. It also offers a new estimator of noise scale and new diagnostic plots to check model sparsity. Finally, we demonstrate the utility of autotune on a real-world financial data set and provide an R package based on C++ on GitHub.

Detecting changepoints for human activity events: A point process model

I.M. Hernandez[1]

[1]Lancaster University, United Kingdom

Monitoring changes in people's behaviour or activities is a vital tool for early health intervention, as these changes often signal underlying medical issues. However, natural circadian rhythms—such as activity spikes during meals and rest at night—create "noise" that can mask significant long-term trends. Our proposed approach focuses on filtering out these expected within-day fluctuations to detect meaningful shifts in behavioural patterns from one day to the next. To achieve this, we treat a full day of events as a single observation, modelled as the realisation of a point process (a similar approach to functional data analysis). By analysing the sequence of these point processes rather than individual event times, we can identify subtle changes in intensity that indicate potential health decline. This approach is particularly effective for home activity data, providing a framework that is both highly sensitive and easy to visualise any trend or change.

Estimating "Realized" Skewness using Convolutional Neural Network

H. Jiang[1], O. Okhrin[1, 2], M. Rockinger[3]

[1]Technische Universität Dresden, Germany
[2]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Germany
[3]University of Lausanne, Switzerland

We propose a new estimator of low-frequency skewness that exploits high-frequency data through a direct functional mapping consisting of layers of convolutional neural networks followed by layers of MLPs. We show that the relevant high-frequency features converge to a continuous limit and that the latent skewness admits a continuous functional representation. This allows us to establish the unbiasedness of our NN estimator using classical universal approximation results and Rademacher complexity arguments. Monte Carlo experiments under stochastic volatility models, with and without jumps, show that the estimator reduces finite-sample bias relative to existing realized-skewness estimators and remains accurate under model misspecification. Empirically, our estimator exhibits temporal stability and delivers superior cross-sectional pricing performance in skewness-sorted portfolios. Another application finds no evidence that ESG-oriented firms exhibit lower crash risk. Overall, the results demonstrate how learning-based functionals can improve the estimation of nonlinear distributional characteristics from high-frequency data.

Regularized score matching in statistics and AI

J. Lederer[1]

[1]University of Hamburg, Germany

Estimating the score is a fundamental task in statistics and AI. This talk presents two examples: multivariate extreme-value theory and diffusion models. We explain why the score is so important, how it can be estimated, and how regularization can improve both accuracy and computational efficiency. We also discuss limitations of classical score matching and potential solutions.

Advances in modeling random functions data

16:00 - 16:30

U-processes and their application in mean estimation

S. Minsker[1], S. Yao[2]

[1]University of Southern California, United States
[2]Hong Kong Baptist University, Hong Kong

This talk presents new results on robust multivariate location estimation built on U-statistical methodology. First, we derive a deviation a Bousquet-type inequality for U-processes with near-optimal constants. Second, using the derived technical results, we introduce a new location estimator based on a U-statistic version of Tukey's median. We show that the proposed estimator is asymptotically normal and achieves efficient asymptotic covariance, in contrast to the original Tukey's median.

Statistical Detection of Local Dynamical Instability in Stochastic Processes

R. Ichiya[1], R. Sagawa[1], Y. Liu[1]

[1]Waseda University, Japan

Chaos is characterized by sensitivity to initial conditions and is commonly quantified using Lyapunov exponents as indicators of nonlinear dynamics. Most existing studies focus on deterministic systems and analyze chaotic behavior from a microscopic perspective. This study proposes statistical methods for quantifying local dynamical instability arising from chaotic behavior in stochastic processes. We introduce the local block Lyapunov exponent and the diagonal Lyapunov dispersion ratio as statistical tools for identifying locally unstable behaviors in stochastic processes. The diagonal Lyapunov dispersion ratio serves as a macroscopic measure for investigating distributional distortions within each block of the stochastic process. We establish asymptotic properties of these statistical tools under a general framework. Numerical simulations demonstrate the effectiveness of the proposed approach under various parameter settings. Finally, we apply the method to financial market data, providing evidence of possible local dynamical instability in the observed time series.

Adaptive Inference for Functional Time Series and Local Regularity

H. Maissoro[1,2], V. Patilea[1], M. Vimond[1]

[1]CREST, ENSAI / Univ Rennes, France
[2]Capgemini Invent, France

We study statistical inference and prediction for stationary functional time series observed through noisy and possibly irregular sampling designs. Such data arise applications where curves are recorded at discrete, potentially random time points and are contaminated by measurement error.

First, we introduce adaptive nonparametric estimators for the mean and autocovariance functions under Lp − m−approximability assumptions. The proposed procedures accommodate both sparse and dense designs and automatically adapt to the local regularity of the underlying functional process through data-driven bandwidth selection. We also derive the asymptotic normality of the mean estimator, which allows honest inference for irregular mean functions. Building on these results, we investigate adaptive prediction methods for irregular functional time series within a best linear unbiased prediction framework. We present methodological developments together with partial theoretical results and discuss their implications for prediction accuracy. Simulations and a real data application illustrate the performance of the proposed procedures.

Network autoregression for binary responses

L. Wang[1]

[1]City University of Hong Kong, Hong Kong

Studying the propagation of binary responses on nodes in a large-scale social network is critical for understanding how individual behaviors and decisions are shaped by social structures and for predicting collective outcomes. We propose a network autoregressive model for binary-valued responses, in which the probability of response at each node is influenced by its neighbors' past decisions, its own past decision, and node-specific covariates, through a logistic link function. The model accounts for network noise and community structure by assuming the underlying network is generated from a block model, with autoregressive parameters that are community-specific. We establish conditions under which the long term behavior of the high-dimensional binary vector converges to a community-specific distribution and the associated convergence rate, illustrating when individuals in the same community or across the whole network reach a consensus regardless of their initial positions. Given an observed network and response vectors, we show asymptotic consistency and normality of the maximum likelihood estimators. We demonstrate the efficiency and validity of the inference procedure through simulated and real data. In particular, we show the model can be used to study the dynamics of strike occurrences in China and highlight the impact of online social networks in facilitating collective actions.

Statistical Learning in Network Models

16:00 - 16:30

Bayesian inference of planted matchings: Local posterior approximation and infinite-volume limit

Z. Fan[1], T.L.H. Wee[2], K.Y. Yang[1]

[1]Yale University, United States
[2]Georgia Institute of Technology, United States

We study Bayesian inference of an unknown matching π^* between two correlated random point sets $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ in $[0,1]^d$, under a critical scaling $\|X_i-Y_{\pi^*(i)}\|_2 \asymp n^{-1/d}$, in both an exact matching model where all points are observed and a partial matching model where a fraction of points may be missing. Restricting to the simplest setting of $d=1$, in this work, we address the questions of (1) whether the posterior distribution over matchings is approximable by a local algorithm, and (2) whether marginal statistics of this posterior have a well-defined limit as $n \to \infty$. We answer both questions affirmatively for partial matching, where a decay-of-correlations arises for large n. For exact matching, we show that the posterior is approximable locally only after a global sorting of the points, and that defining a large-n limit of marginal statistics requires a careful indexing of points in the Poisson point process limit of the data, based on a notion of flow. We leave as an open question the extensions of such results to dimensions $d \geq 2$.

Semi-Supervised Learning on Graphs with GNNs

O. Klopp[1]

[1]ESSEC, France

We study semi-supervised node prediction on graphs where responses arise from a graph-aware feature operator followed by a smooth regression map. Within a class combining skip-connected GCN propagation with a fully connected ReLU network, we (i) derive an oracle inequality for population risk under random label masks that separates approximation and estimation error and exposes dependence on the labeled fraction, covering numbers, and a receptive-field constant; (ii) show skip connections exactly represent multi-hop polynomial filters, mitigating over-smoothing; (iii) give covering-number bounds; and (iv) quantify robustness of our algorithm. These results link classical graph regularization and modern GNN design.

Signal Recovery from Random Dot-Product Graphs Under Local Differential Privacy

S. Vishwanath[1]

[1]UC San Diego, United States

Differential privacy (DP) has emerged as the standard for ensuring formal privacy, allowing for population-level inference, while limiting the risk of exposing the contribution of any single individual in a database. We consider the problem of recovering latent information from graphs under ε-edge local DP—where the presence of relationships/edges between two users/vertices remains confidential, even from the data curator.

For the class of generalized random dot-product graphs, we show that a standard local DP mechanism induces a specific geometric distortion in the latent information. Leveraging this insight, we develop a principled statistical framework that achieves consistent recovery of latent positions by adjusting the inference procedure for the privatized graph. We show that this procedure is minimax-optimal under local edge DP constraints, establishing the fundamental limits of private graph inference. Furthermore, we show that this framework extends to consistent recovery of geometric and topological information underlying the latent positions, as encoded in their persistence diagrams.

Our results extend previous work from the private community detection literature to a substantially richer class of models and inferential tasks.

Kernel and Graphical Methods for Comparing Conditional Distributions

B. Bhattacharya[1], A. Chatterjee[2], Z. Niu[1]

[1]University of Pennsylvania, United States
[2]University of Chicago, United States

In this talk we will discuss various nonparametric methods for comparing conditional distributions based on kernels and nearest-neighbor graphs. The methods can be readily applied to a broad range of problems, ranging from classical nonparametric statistics to modern machine learning. Specifically, we will discuss applications in testing model calibration, regression curve evaluation, and validation of emulators in simulation-based inference.

Nonparametric Methods for Complex and Nonstationary Data

16:00 - 16:30

Theoretical Comparison of Independent-Based Samplers

f. bertholom[1], r. douc[2], f. roueff[3]

[1]Telecom SudParis, France
[2]Telecom Sudparis, France
[3]Telecom Paris, France

We analyze the convergence properties of four independent-based Markov Chain Monte-Carlo samplers, namely the Importance Markov Chain and three algorithms based on multiple tries. We prove that the Multiple-Try Metropolis with Independent Sampling dominates Multiple-Try Metropolis with Independent Balancing and Independent Sampling Importance Resampling in the Peskun sense. Furthermore, we derive a general formula for obtaining exact geometric convergence rates of these algorithms, assuming that the importance weights are bounded. In the unbounded case, while none of the three algorithms based on multiple tries are uniformly ergodic, the Importance Markov Chain only requires an exponential moment condition on the weights to be uniformly ergodic, or a polynomial moment condition to guarantee polynomial ergodicity.

Composite Entropy Minimization: A New Approach to Clustering Extensions to Segmentation and Regression

D. Thierry[1]

[1]Université Paris Nanterre, France

Clustering is often addressed either through geometric methods such as k-means or through model-based approaches relying on finite mixture models. While mixture models provide a natural probabilistic framework, determining the appropriate number of components remains a difficult problem. Classical criteria such as BIC aim to recover the true number of mixture components but may diverge when the underlying distribution is not itself a finite mixture. Conversely, criteria such as ICL incorporate an entropy penalty to better reflect the clustering objective but lack clear asymptotic guarantees regarding the selected number of clusters.

In this work, we introduce a new framework based on composite entropy minimization. The proposed criterion evaluates mixture decompositions of a distribution by combining the entropy of the mixture weights with the cross-entropy between component distributions and a parametric model family. This leads to a principled notion of optimal mixture decomposition.

Within this framework, we define the Relative Entropic Order (REO), corresponding to the smallest number of components beyond which the optimal decompositions remain unchanged. We establish theoretical results showing that the empirical REO consistently estimates its population counterpart under mild assumptions. The proposed criterion is closely related to the Classification EM (CEM) algorithm, which can be interpreted as performing a descent of the composite entropy.

Beyond clustering, the framework naturally extends to several statistical learning problems. In particular, it can be adapted to segmentation tasks and to mixture regression models, where different subpopulations follow distinct regression relationships. These extensions illustrate the flexibility of the composite entropy perspective for identifying latent structure in heterogeneous data.

Experiments on synthetic and real datasets demonstrate that the proposed approach provides stable and interpretable decompositions while offering a unified view of clustering, segmentation, and regression within an entropy-based framework.

On Properties of Oscillatory Time Series with Random Amplitude and Phase

Ł. Lenart[1]

[1]Krakow University of Economics, Poland

This talk examines stationarity properties of oscillatory time series with random latent amplitude and phase components and an unknown frequency. We present general conditions under which such processes are second order stationary, stationary of any order, or strictly stationary. In particular, we show that stationarity of any order requires a specific symmetry property related to a spherical distribution, which implies a uniform distribution of phase shifts. We point out that commonly adopted assumptions, such as Gaussianity, can be overly restrictive and may substantially limit the flexibility of the amplitude dynamics. We further introduce a flexible class of oscillatory time series in which the amplitude process is stationary of any order, while the phase process follows a random walk. Within this framework, the resulting time series remains stationary of any order and exhibits a pseudo cyclical dependence structure with a potentially very slow decay of the autocovariance function. The framework is extended to allow for asymmetric cyclical behavior, and we show that the proposed time series is alpha mixing. The proposed approach fills an important gap in the modeling of cyclical data by allowing flexible amplitude and phase dynamics.

Innovations State-Space Model for Cyclical Data

Ł. Lenart[1], Ł. Kwiatkowski[2], B. Majewski[3], J. Wróblewska[2]

[1]Krakow University of Economics, Department of Mathematics, Poland
[2]Krakow University of Economics, Department of Econometrics and Operations Research, Poland
[3]AGH University of Krakow, Faculty of Applied Mathematics, Poland

We examined the statistical properties of a time series model that is an almost periodic function at time, with stochastic stationary amplitude modulation and stochastic time-warped components in the time. Our specification belongs to the family of non-linear innovations state space models. It was shown that if the time-warped component is stationary, then such a model has an almost periodic mean function with the same frequencies. However, if the time-warped component contains a random walk process, then such a model is zero-mean, and any cyclical phenomenon related to the modulated and time-warped almost periodic function is pushed to the second-order properties, with a clearly pseudo-cyclical autocovariance function that vanishes similarly to a damped almost periodic function with relevant frequencies. Moreover, it was proven that such a model is weakly stationary of any order.

Computer-intensive statistical inference for complex data

16:00 - 16:20

Local Machine Learning for Distributed Data Giants with an Application to Hedonic Modeling

M. Scholz[1,2], S. Sperlich[3]

[1]Graz University of Technology, Austria
[2]Joanneum Research, Austria
[3]Univeristy of Geneva, Switzerland

Recent advances in hedonic modeling indicate that nonparametric approaches outperform previously proposed parametric and semiparametric specifications. However, these findings are typically based on small and sometimes even aggregated datasets or are limited to individual cities. In practice, transaction price and rental data are large in scale. They are distributed across multiple stakeholders and are often not merged into a single database due to legal or institutional constraints. We propose a nonparametric framework that combines the strengths of local polynomial regression with concepts from modern predictive algorithms to enable efficient estimation and prediction in decentralized environments. Our locally adaptive method incorporates LASSO-type variable selection, effectively turning localization into an advantage. An application to German housing data demonstrates its practical relevance, particularly for policymakers.

Regularizing BELIEF with Sequency Lasso

W. Zhang[1]

[1]Academy of Mathematics and Systems Science,CAS, China

As the complexity of models and the volume of data increase, interpretable methods for modeling complicated dependence are greatly needed. A recent framework of Binary Expansion Linear Effect (BELIEF) provides a "divide and conquer" approach to decompose any complex form of dependency into small linear regressions over data bits. While BELIEF provides a flexible approximation to arbitrary relationships, it faces an important challenge of high dimensionality. To overcome this obstacle, we propose a novel definition of smoothness for binary interactions through an interesting connection to the sequency of Walsh functions. We investigate this connection and study related theory and algorithms. Based on this connection, we develop a smoothness-based regularization of BELIEF. Specifically, we propose the sequency Lasso, a generalized Lasso model that imposes larger penalties on less smooth terms to model smooth form of dependence. The proposed method yields a highly competitive yet interpretable nonparametric machine learning tool. Numerical studies demonstrate that the sequency Lasso has advantages in clear interpretability and effectiveness for nonlinear and high-dimensional data.

SSGEN: Simulation of Distribution Tails using Generative models

A. Deo[1], M. Gupta[1]

[1]Indian Institute of Management Bangalore, India

We introduce Self-Similar Generative Estimation (SS-GEN), a method for simulating multivariate tail events and estimating tail risk in both heavy- and light-tailed settings. SS-GEN exploits asymptotic structure to decompose the tail distribution into an explicit radial component and a nonparametric angular component, reducing tail learning to a compact-domain problem that can be handled by off-the-shelf deep generative models. The resulting sampler generates representative extreme scenarios and supports tail risk estimation far beyond the observed sample. Under mild tail regularity conditions, we show that the SS-GEN density is asymptotically exact in the tail, with vanishing uniform relative error for regularly varying distributions and vanishing uniform log-relative error for Weibull-type distributions. We also establish data-driven guarantees that bound the density-approximation error of SS-GEN in terms of the learning error of the deep generative model for the angular law, thereby linking tail-simulation accuracy directly to finite-sample generative performance on exceedance data. Unlike existing approaches that rely on specialised architectures or parametric tail specifications, SS-GEN leverages asymptotic tail structure to enable standard generative models to simulate representative extreme samples and estimate tail risk functionals in practically unobserved regions.

Assessing the quality of denoising diffusion models in Wasserstein distance: noisy score and optimal bounds

V. Arsenyan[1], E. Vardanyan[1], A. Dalalyan[1]

[1]ENSAE Paris, France

Generative modeling aims to produce new random examples from an unknown target distribution, given access to a finite collection of examples. Among the leading approaches, denoising diffusion probabilistic models (DDPMs) construct such examples by mapping a Brownian motion via a diffusion process driven by an estimated score function. In this work, we first provide empirical evidence that DDPMs are robust to constant-variance noise in the score evaluations. We then establish finite-sample guarantees in Wasserstein-2 distance that exhibit two key features: (i) they characterize and quantify the robustness of DDPMs to noisy score estimates, and (ii) they achieve faster convergence rates than previously known results. Furthermore, we observe that the obtained rates match those known in the Gaussian case, implying their optimality.

Student Poster Session (includes drinks and snacks)

18:00 - 19:00

Pivotal inference for linear predictions in stationary processes

H. Dette[1], S. Kühnert[1]

[1]Ruhr University Bochum, Germany

We develop pivotal inference for the final (FPE) and relative final prediction error (RFPE) of linear forecasts in stationary processes. Our approach is based on self-normalization and avoids the estimation of the asymptotic variances of the empirical autocovariances. We provide pivotal confidence intervals for the (R)FPE, develop estimates for the minimal order of a linear prediction that is required to obtain a prespecified forecasting accuracy and also propose (pivotal) statistical tests for the hypotheses that the (R)FPE exceeds a given threshold. Additionally, we provide pivotal uncertainty quantification for the commonly used coefficient of determination obtained from a linear prediction based on the past observations and develop new (pivotal) inference tools for the partial autocorrelation, which do not require the assumption of an autoregressive process.

Singular Spectrum Analysis Revisited: Frequency Recovery, Spectral Consistency, and Window Length Selection

G. Martos[1], P. Poncela[2], D. Fresoli[2]

[1]Universidad Torcuato Di Tella, Argentina
[2]Universidad Autónoma de Madrid, Spain

Singular Spectrum Analysis (SSA) is a nonparametric method for time series analysis. Via the Singular Value Decomposition on the so--called trajectory matrix, or equivalently, by diagonalizing the empirical second moment matrix, it decomposes a time series into quasi--orthogonal components that aim to maximize variance. The resulting trendlines provide natural estimates of latent structures such as trend, cycles, and noise. However, in contrast with spectral methods, the components extracted are not intrinsically associated to particular frequencies. This paper introduces a related technique that reconciles frequency identification with variance--based decomposition. As a byproduct, our decomposition yields a consistent estimator of the spectral density. A key parameter in SSA is the window length, which critically influences the resulting decomposition and its interpretation. We provide practical guidance for selecting the window length based on asymptotic results and adapt well--established inferential tools to group components, thereby enabling the identification of statistically significant signals associated with specific frequencies. The performance of the proposed methodology is illustrated through simulation studies and empirical applications to temperature and high-frequency electricity consumption data, where meaningful latent structures are successfully identified.

Distributed Sparse Linear Regression under Communication Constraints

R. Fonseca[1], B. Nadler[2]

[1]Federal University of Bahia, Brazil
[2]Weizmann Institute of Science, Israel

In multiple domains, statistical tasks are performed in distributed settings, with data split among several end machines that are connected to a fusion center. In various applications, the end machines have limited bandwidth and power, and thus a tight communication budget. In this work we focus on distributed learning of a sparse linear regression model, under severe communication constraints. We propose several two round distributed schemes, whose communication per machine is sublinear in the data dimension. In our schemes, individual machines compute debiased lasso estimators, but send to the fusion center only very few values. On the theoretical front, we analyze one of these schemes and prove that with high probability it achieves exact support recovery at low signal to noise ratios, where individual machines fail to recover the support. We show in simulations that our scheme works as well as, and in some cases better, than more communication intensive approaches.

Estimating the Probability of Default Cascades in Financial Networks

S. Ioannidis[1], D. Ioannides[2]

[1]Royal Holloway University of London, United Kingdom
[2]University of Macedonia, Greece

The recent theoretical and empirical literature on financial contagion has Investigated the relationships between the interbank exposure network and the financial stability of the banking system by Glasserman and Young (2016) for a nice recent surveys. The financial network has been recognized to be a source of financial crisis as shocks, which initially affect only few institutions, propagate through the entire banking system producing a contagion cascade. The present paper studies the issue of financial contagion assuming that the shocks are random variables (r.v.'s). Aim of the present paper is to study the issue of financial contagion assuming that the shocks are random variables (r.v.'s). As is pointed in Glasserman and Young(2015) the interconnections among the banks create potential channels for contagion and amplification of
the shocks. Contagion occurs when defaults by some banks tiger defaults by other banks through a domino effect, while amplification occurs where contagion stops but the losses among defaulting banks keep escalating because of their indebtedness to one another. They analyzed the probability of default cascades and consequent losses of value that are attributable to network connections by assuming that the shocks r.v.'s follow well known distributions. The contribution of this paper is to extend the results for the modeling of shocks in Gasserman and Young(2015) and the estimation of the probability of default cascades for a more general r.v.'s .

References

Glasserman , P., Young , H. (2016). Contagion in financial networks. Journal of Economic Literature , 54, 770-831.
Glasserman , P., Young , H. (2015). How likely is contagion in financial networks?. Journal of Banking and Finance 50, 383-399.

.

Asymptotic limit and distribution of the Information Imbalance coefficient

P. Chandra[1], D. Sulem[1], A. Mira[1]

[1]Università della Svizzera italiana, Switzerland

The Information Imbalance coefficient (II), introduced by Glielmo et al. (2022) gives a measure of information flow from one random variable to another. In fact, it can be viewed as a measure of association between two sets of random variables which also takes into account the direction of information flow. The coefficient has its foundation in classical information theory by its representation as the conditional expectation of copula variables. It operates non-parametrically by measuring association using ranks of nearest neighbours. There is however, a major gap in the statistical foundations of II. Its asymptotic limit was not previously known. We prove an intuitive theoretical limit of Information Imbalance (II). Beyond the deterministic limit, we establish asymptotic normality of the II estimator under mild regularity conditions: after appropriate centering and scaling, the estimator converges in distribution to a Gaussian. The proof leverages the Hájek representation theorem, following the strategy of Lin and Han (2025), adapted to the multivariate structure of II. This result provides the first rigorous characterisation of the fluctuations of II around its limit. We also provide an upper bound of its almost sure convergence rate. Finally, we present a framework for significance testing under the null hypothesis based on permutation distribution theory developed by Daniels (1944) for the Friedman-Rafsky coefficient (1983).

Testing for Isotropy of Function-Valued Random Fields

J. Baumhakel[1], S. Hörmann[1], M. Neumann[1]

[1]Graz University of Technology, Austria

We propose a new nonparametric approach for testing isotropy, i.e. invariance in distribution under rotations around the orgin, of function-valued random fields. The key idea is to extract local, directional-dependent characteristics of the field and analyze their distributions with appropriate adjustments to account for spatial dependence. While function-valued random fields are well studied in spatial statistics, nonparametric methods for assessing their isotropy have not yet been established, and our framework is designed to fill this gap.

We outline how such a test can be used for applications in materials science. In particular, we use it to assess cylindrical isotropy of local volume fractions in paper-based materials which have been extracted from 3D image data.

Extrinsic Total Variance and Coplanarity via Oriented Projective Shape Analysis

M. Alamoudi[1, 2], R. Paige[3], V. Patrangenaru[1]

[1]Florida State University, United States
[2]King Abdulaziz University, Saudi Arabia
[3]Missouri University of Science and Technology, United States

Projective shape analysis provides a geometric framework for studying landmarked digital images acquired by pinhole cameras. In classical projective shape (PS), landmark configurations are represented in a product of projective spaces and are invariant under the full projective group; however, the representation is sign-blind, so opposite directions in three-dimensional space correspond to the same projective point and the front-back orientation of a surface is not recorded. Oriented projective shape (OPS) restores this information by working on a product of spheres and restricting to orientation-preserving projective transformations. We introduce an extrinsic total-variance index for OPS within an extrinsic Frechet framework based on a directional embedding into Euclidean space. In the planar pentad case, the sample total extrinsic variance admits a closed-form expression in terms of the mean resultant length of the oriented projective coordinates on the sphere. As an illustration, using an oriented projective frame, we analyze the Sope Creek stone data set, a benchmark nearly planar example with 41 images and 5 landmarks. Using a delta-method large-sample approximation and a leave-two-out diagnostic, we assess coplanarity through the hypothesis of vanishing OPS total variance and obtain conclusions consistent with earlier coplanarity results under PS.

On Stein’s test of uniformity on the hypersphere

P. Axmann[1], B. Ebner[1], E. García Portugués[2]

[1]Karlsruhe Institute of Technology, Germany
[2]Carlos III University of Madrid, Spain

Testing for uniformity is a fundamental problem in directional statistics. We propose a new test of uniformity on the hypersphere based on a Stein characterization associated with the Laplace-Beltrami operator, together with a suitable class of test functions linked to the moment generating function. Exploiting the eigenfunctions of the operator, we obtain a harmonic decomposition in terms of Gegenbauer polynomials and show that the proposed procedure belongs to the class of Sobolev tests. We derive closed-form expressions for the distribution of the test statistic under the null hypothesis and under fixed alternatives. To enhance power against a range of alternatives, we introduce a tuning parameter into the characterization and study its impact on rejection probabilities. We also discuss data-driven strategies for selecting this parameter and compare the resulting procedure with related parametric tests and competing Sobolev-type methods in numerical experiments.

A Random Projection-based Kernel Density Estimator

A. Deb[1], M. Mukhopadhyay[1, 2], S. Dutta[1, 2]

[1]Indian Institute of Technology, Kanpur, India
[2]Indian Statistical Institute, Kolkata, India

Nonparametric density estimation is a fundamental problem in statistics, with the kernel density estimator (KDE) being a widely used method in view of its efficient performance in low-dimensional settings. However, the accuracy of KDE deteriorates and becomes highly sensitive to the choice of the smoothing parameter (say, h) as the data dimension increases. In this work, we develop a random projection-based kernel density estimator. The proposed KDE is computationally quite fast, making it attractive in high dimensions. It also incorporates an optimal and data-adaptive choice of h, which depends on both the test point as well as the training sample points. We establish pointwise consistency of the proposed estimator under a general assumption of continuity of the underlying density function. Numerical studies are conducted on non-parametric classification and regression to illustrate the good empirical performance of the proposed estimator.

Flexible and Nonparametric Methods in Modelling

9:00 - 9:30

Improved Estimation for Generalised Additive Models

I. Kosmidis[1], O. Kemp[1]

[1]University of Warwick, United Kingdom

Generalised additive models (GAMs) provide a flexible framework for modelling nonlinear relationships between predictors and a response variable through a combination of scalar covariate effects and smooth functions of covariates. In practice, estimation is performed using penalised likelihood methods, where smoothing penalties control the complexity of the fitted functions. However, when sample sizes are moderate or small, the resulting estimators can exhibit noticeable bias. Standard bias-reduction techniques developed for unpenalised likelihood problems ignore the role of the smoothing penalty and, therefore, depending on the estimated penalty strength, may fail in reducing bias. In this work, we develop a bias-reduction approach for GAMs that explicitly accounts for the smoothing penalty. We show that the asymptotic bias of the penalised estimator can be decomposed into a classical likelihood-based bias term and an additional bias term induced by the penalty. This insight allows us to derive adjusted estimating equations that remove the leading term in the bias expansion of the maximum penalised likelihood estimator. Simulation studies and a real-data example illustrate that the proposed method can deliver marked improvements in estimating both scalar effects and smooth functions in GAMs, particularly in small- to moderate-sample settings.

A Generalized Additive Partial-Mastery Cognitive Diagnostic Model

C. Cardenas-Hurtado[1], Y. Chen[1], I. Moustaki[1]

[1]London School of Economics and Political Science, United Kingdom

Cognitive diagnosis models (CDMs) are restricted latent class models widely used for measuring attributes of interest in diagnostic assessments in education, psychology, biomedical sciences, and related fields. Partial-mastery CDMs (PM-CDMs) are an important extension of CDMs. They model individuals' status for each attribute to be continuous for measuring the partial mastery level, which relaxes the restrictive discrete-attribute assumption of classical CDMs. As a result, PM-CDMs often yield better fits for real-world data and refined measurement of the substantive attributes of interest. However, these models inherit some strong parametric assumptions from the traditional CDMs about the item response functions and, thus, still suffer from a significant risk of model misspecification. This paper proposes a generalized additive PM-CDM (GaPM-CDM) that substantially relaxes the parametric assumptions of PM-CDMs. This proposal leverages model parsimony and interpretability by modeling each item response function as a mixture of nonparametric monotone functions of attributes. A method for the estimation of GaPM-CDM is developed, which combines the marginal maximum likelihood estimator with a sieve approximation of the nonparametric functions. The new model is applicable under both confirmatory and exploratory settings, depending on whether prior knowledge is available about the relationship between observed variables and attributes. The proposed method is applied to two measurement problems from educational testing and healthcare research, respectively, and further evaluated and compared with PM-CDMs through extensive simulation studies.

Coverage correlation: detecting singular dependencies between random variables

M. Azadkia[1], T. Wang[1], X. Yang[1]

[1]London School of Economics, United Kingdom

We introduce the coverage correlation coefficient, a novel nonparametric measure of statistical association designed to quantifies the extent to which two random variables have a joint distribution concentrated on a singular subset with respect to the product of the marginals. Our correlation statistic consistently estimates an f-divergence between the joint distribution and the product of the marginals, which is 0 if and only if the variables are independent and 1 if and only if the copula is singular. Using Monge--Kantorovich ranks, the coverage correlation naturally extends to measure association between random vectors. It is distribution-free, admits an analytically tractable asymptotic null distribution, and can be computed efficiently, making it well-suited for detecting complex, potentially nonlinear associations in large-scale pairwise testing.

Flexible and robust Epidemic Modelling with applications to Intervention assessment and optimal control

N. Demiris[1]

[1]AUEB, Greece

This presentation is concerned with a wide class of Stochastic Epidemic models appropriate for assessing the impact of interventions. We will show how one may model disease transmission using (i) continuous time models such as families of Geometric Brownian motions and other transformed multi-task Gaussian processes and (ii) multi-phasic stochastic models driven by point processes or Hidden semi-Markov models. The models will be applied to real data and used to (i) assess the effect of non-pharmaceutical interventions and (ii) to inform reinforcement learning-based strategies for optimal control. If time permits we will discuss assessing model mis-specification via generalised Bayes.

New Directions in Nonparametric Methods

9:00 - 9:30

SEMIPARAMETRIC COVARIANCE ESTIMATION FOR SPARSE FUNCTIONAL DATA

J. Racine[1], V. Patilea[2]

[1]McMaster, Canada
[2]ENSAI, France

Abstract. We study semiparametric covariance estimation for sparse functional data, where each trajectory may be observed at only a few irregular design points and is contaminated by measurement error. The goal is to estimate the covariance surface by combining nonparametric estimation of the mean and marginal variance functions with a low-dimensional parametric model for the correlation structure. Our proposed estimator uses cosine-series estimation for the mean and variance, a nearest-neighbor estimator of the measurement-error variance, semiparametric least-squares fitting over a compact set of stationary and nonstationary correlation families, and convex model averaging across fitted covariance candidates. We summarize the finite-sample evidence from four Monte Carlo designs and illustrate the approach with the CD4 count data distributed with the refund package. An R package implementing these methods is available for interested readers through the spcovar package.

Clustering and community detection via k-rays

J. Arroyo[1]

[1]Texas A&M University, United States

The k-means algorithm is a widely used clustering method that identifies clusters as point clouds associated with centroids based on their means. However, in many applications, data exhibit structures that do not conform to point clouds, requiring clustering techniques that account for manifold structure. A notable example is network data, where clusters correspond to communities, and spectral embeddings of adjacency matrices often display a ray-like structure. This paper introduces k-rays clustering, a method that groups observations along rays. We propose an objective function and solve it using a variant of the Lloyd algorithm. Theoretical analysis establishes the method's accuracy under flexible conditions. Experimental results demonstrate the advantages of the method compared to other clustering approaches.

Nonparametric Sparse Graphon Estimation with Vertex Covariates

M. Donath[1], A. Fuchs-Kreiß[1]

[1]University Hildesheim, Germany

Our goal in this talk is to model undirected networks with node covariates. Therefore, we extend the graphon model. Specifically, we model the latent positions ξ_i of the nodes in dependence of observed node covariates X_i. Based on a least-squares approach, we simultaneously estimate the graphon and the influence of the covariates. While the effect of the covariates follows a parametric model, the graphon is estimated in a non-parametric fashion by fitting a step graphon that is constant on subsquares of $[0,1]^2$. This estimator avoids computationally demanding discrete optimization steps because, in our model, the covariates provide probabilistic information about the location of the latent positions. Therefore, our estimator does not require to optimize over different allocations of vertices to communities. Our procedure is consistent for a large class of graphons because Hölder continuous graphons can be approximated by step graphons. We prove consistency of the estimator under certain regularity assumptions of the model by approximating the mean squared error. The proof utilize the ULLN for U-processes.

Bootstrapping network statistics using overlapping partitions

S. Chakrabarty[1]

[1]University of Michigan, United States

Bootstrapping network data efficiently is a challenging task. The existing methods tend to make strong assumptions on both the network structure and the statistics being bootstrapped, and are computationally costly. This paper introduces a general algorithm, OPBoot, for network bootstrap that partitions the network into multiple overlapping subnetworks and then aggregates results from bootstrapping these subnetworks to generate a bootstrap sample of the network statistic of interest. This approach tends to be much faster than competing methods as most of the computations are done on smaller subnetworks. We show that OPBoot is consistent in distribution for a large class of network statistics under minimal assumptions on the network structure, and demonstrate with extensive numerical examples that the bootstrap confidence intervals produced by OPBoot attain good coverage without substantially increasing interval lengths in a fraction of the time needed for running competing methods.

Non-parametric methods for complex Time Series

9:00 - 9:30

Frequency-domain Regularization for Dynamic factor models: Forecasting and Sparsity

M. Eichler[1], G. Motta[2]

[1]Maastricht University, Netherlands
[2]Columbia University, United States

The generalized dynamic factor model (GDFM) by Forni et al. (2001) rely on Brillinger's dynamic principal components and thus involve two-sided filters, which are unsuitable for forecasting. Forni et al. (2015, 2017) derive a semi-parametric estimator of common components and common shocks based on one-sided filters. However, this newer approach relies on the additional assumptions that the common components have a rational spectral density. Moreover, their estimation procedure involves several time-domain and frequency-domain steps.

To overcome these limitations, we have recently introduced a novel approach that estimates the common components and the common shocks directly in the frequency domain. Our model does not require the assumption of a rational spectral density, and our estimation method is computationally simpler and faster.

In this companion work on frequency-domain estimation of the GDFM, we extend our methodology in two important directions: forecasting and sparsity. We present a one-sided regularized estimator specifically designed for forecasting. Additionally, we introduce regularization techniques directly in the frequency domain to induce sparsity in the transfer-function (or frequency-domain loading) matrix. Our regularized frequency-domain estimator retains the computational advantages of our previous approach while delivering improved forecast performance and parsimonious representations.

Some hypothesis tests for covariance matrix of high-dimensional time series

Y. Liu[1], Y. Yoshida[1], M. Kita[1]

[1]Waseda University, Japan

We consider the testing problem for the sphericity hypothesis regarding the covariance matrix of high-dimensional time series. Under the regime of (n, p)-asymptotics, we derive the asymptotic null distributions of the U- and V-statistics, which play a central role when the data dimension is large. We propose a spherical bootstrap method for high-dimensional time series for the practical use of these statistics. The numerical simulations align well with our theoretical findings. Some real data applications are also provided.

Semi-parametric estimation of non-stationary autoregressive models

G. Motta[1, 2], Q. Wang[3]

[1]City University of New York, United States
[2]Columbia University, Data Science Institute, United States
[3]Texas A&M University, United States

We develop a novel semi-parametric approach for accurately estimating time-varying mean and variance in autoregressive (AR) models. By combining B-splines with (i) generalized least squares (GLS) estimation to account for serial correlation, and (ii) weighted least squares (WLS) for smooth parametrization, our approach addresses the challenges posed by time-varying dynamics in time series data. The covariance matrix in the GLS estimation of the spline coefficients is iteratively updated through the WLS estimation of the AR coefficients in a band-limited manner. Meanwhile, a new autoregressive model is proposed that incorporates time-varying variance with a finite bounded envelope function, and a novel method is introduced to estimate it through splines. Additionally, the order of the AR model is determined through a generalized Bayesian information criterion (GBICp) that incorporates prior information. The effectiveness of the methodology is demonstrated through extensive simulations and applications to the Federal Reserve Economic Data. We derive asymptotic theory for our approach and compare our rates of convergence with those achieved by existing methods.

Non-parametric modeling of time-varying quasi-periodic oscillations

D. Soto[1], G. Motta[2], F. Cuevas[3], M. Sobolewska[4]

[1]Universidad del Bío Bío, Chile
[2]Columbia University, United States
[3]Universidad Técnica Federico Santa María, Chile
[4]Smithsonian Astrophysical Observatory, United States

Quasi-periodic oscillations (QPOs) are commonly observed in the variability of astrophysical sources such as X-ray binaries and active galactic nuclei. These oscillatory components appear as narrow peaks in the power spectral density (PSD) of observed light curves, revealing processes near compact objects such as black holes or neutron stars.

Continuous-time autoregressive moving average (CARMA) models have been successfully used to represent the spectral structure of astronomical time series. However, stationary models assume a time-invariant spectral structure, while in many observations the frequency and amplitude of QPOs evolve over time. A popular approach astronomers use to estimate QPOs involves fitting a sum of Lorentzian functions to the periodogram. Albeit widely used, this two-step approach decouples estimation from modeling, limiting coherence and interpretability.

In this paper, we introduce an alternative approach to estimate QPOs using locally stationary CARMA (LS-CARMA) processes. We prove that the spectral density of an LS-CARMA process can be analytically decomposed into a finite number of component functions that depends on the nature (real or complex) of the roots of the autoregressive polynomial. We show that each component corresponds to a spectral peak, and we derive closed-form expressions for their frequencies, offering a feasible alternative to Lorentzian fitting.

The advantage of adopting our approach is twofold. First, by allowing the spectral density to vary smoothly over time, we can capture quasi-periodic components whose frequency changes gradually. Second, the closed-form nature of our novel decomposition permits direct estimation of the time-varying frequencies associated with the QPOs.

Leveraging AI for Problems in Public Health

9:00 - 9:30

Principal stratification causal effects estimation with continuous auxiliary variables

G. Tang[1]

[1]University of Pittsburg, United States

Principal stratification offers a popular framework in causal inference that involves an intermediate response variable. Estimating principal causal effects (PCEs) is challenging as the principal strata are not observed. Earlier works by Tan et al. (2022) addressed the identification issue under a monotonicity assumption by leveraging a linear trend in an ordinal auxiliary variable in a counterfactual model. However, the linearity assumption may be unrealistic in practice. To allow flexibility and applicability, we extend this method to accommodate a continuous auxiliary variable X. We first propose a novel weighted least square (WLS) kernel estimator of the principal strata distribution with continuous X under the monotonicity assumption. Subsequently we extend the estimation of PCEs with a continuous X. Our method can be readily extended to high dimensional covariates. We evaluate its performance through simulations and apply it to the motivating NSABP B-40 trial for early breast cancer.

Prediction-Powered Conditional Inference

X. Dai[1]

[1]UCLA, United States

We study prediction-powered conditional inference in the setting where labeled data are scarce, unlabeled covariates are abundant, and a black-box machine-learning predictor is available. The goal is to perform statistical inference on conditional functionals evaluated at a fixed test point, such as conditional means, without imposing a parametric model for the conditional relationship. Our approach combines localization with prediction-based variance reduction. First, we introduce a reproducing kernelbased localization method that learns a data-adaptive weight function from covariates
and reformulates the target conditional moment at the test point as a weighted unconditional moment. Second, we incorporate machine-learning predictions through a correction-based decomposition of this localized moment, yielding a prediction-powered estimator and confidence interval that reduce variance when the predictor is informative while preserving validity regardless of predictor accuracy. We establish nonasymptotic error bounds and minimax-optimal convergence rates for the resulting estimator, prove pointwise asymptotic normality with consistent variance estimation, and provide an explicit variance decomposition that characterizes how machine-learning predictions and unlabeled covariates improve statistical efficiency. Numerical experiments on simulated and real datasets demonstrate valid conditional coverage and substantially sharper confidence intervals than alternative methods.

High-Dimensional Markov-switching Ordinary Differential Processes

M. Kolar[1]

[1]USC & MBZUAI, United States

We investigate the parameter recovery of Markov-switching ordinary differential processes from discrete observations, where the differential equations are nonlinear additive models. This framework has been widely applied in biological systems, control systems, and other domains; however, limited research has been conducted on reconstructing the generating processes from observations. In contrast, many physical systems, such as human brains, cannot be directly experimented upon and rely on observations to infer the underlying systems. To address this gap, this manuscript presents a comprehensive study of the model, encompassing algorithm design, optimization guarantees, and quantification of statistical errors. Specifically, we develop a two-stage algorithm that first recovers the continuous sample path from discrete samples and then estimates the parameters of the processes. We provide novel theoretical insights into the statistical error and linear convergence guarantee when the processes are β-mixing. Our analysis is based on the truncation of the latent posterior processes and demonstrates that the truncated processes approximate the true processes under mixing conditions. We apply this model to investigate the differences in resting-state brain networks between the ADHD group and normal controls, revealing differences in the transition rate matrices of the two groups.

Reinforcement learning methods for adaptive respondent-driven sampling

E. Laber[1]

[1]Duke University, United States

Respondent-driven sampling (RDS) is widely used to study hidden or hard-to-reach populations by incentivizing study participants to recruit their social connections.
The success and efficiency of RDS can depend critically on the nature of the incentives, including their number, value, call to action, etc. Standard RDS uses an incentive structure that is set {\em a priori} and held fixed throughout the study. Thus, it does not make use of accumulating information on which incentives are effective and for whom.
We propose a reinforcement learning (RL) based adaptive RDS study design in which the incentives are tailored over time to maximize cumulative utility during the study. We show that these designs are more efficient, cost-effective, and can generate new insights into the social structure of hidden populations.
In addition, we develop methods for valid post-study inference which are non-trivial due to the adaptive sampling induced by RL as well as the complex dependencies among subjects due to latent (unobserved) social network structure. We provide asymptotic regret bounds and illustrate its finite sample behavior through a suite of simulation experiments.

Randomization and Causation

9:00 - 9:30

Learning density ratios in causal inference using Bregman-Riesz regression

O. Hines[1]

[1]Columbia University, United States

The ratio of two probability density functions is a fundamental quantity that appears in many areas of statistics: causal inference, reinforcement learning, covariate shift, outlier detection, independence testing, importance sampling, and diffusion modelling. Naively estimating the numerator and denominator densities separately can lead to unstable performance. For this reason, several methods have been developed for learning the density ratio directly based on (a) Bregman divergences or (b) recasting the density ratio as the odds in a probabilistic classification model that predicts whether an observation is sampled from the numerator or denominator distribution. Additionally, the density ratio can be viewed as a Riesz representing function, making it amenable to learning via (c) minimization of the so-called Riesz loss. In this talk we outline how these approaches are related under a unifying framework, which we call Bregman-Riesz regression. We further show how data augmentation techniques can be used to apply density ratio learning methods to causal problems in practice, where the numerator typically represents an unobserved intervention distribution.

Estimating the Wasserstein Barycenter of One-Dimensional Distributions with Sparse Sampling

F. Stijven[1], J. Peng[2], L. Wang[3], P. Gilbert[4]

[1]KU Leuven, Belgium
[2]University of Washington, United States
[3]University of Toronto, Canada
[4]Fred Hutchinson Cancer Center, United States

We investigate the statistical analysis of distributional data where each unit is represented by a distribution on the real line, observed only through finite, often small, samples of observations. A natural structural mean for such data is the Wasserstein barycenter, a Fréchet mean rooted in optimal transport theory. For univariate distributions, the barycenter possesses a convenient representation: its quantile function is the mean of the quantile functions of the underlying random distributions. The standard empirical barycenter estimates this mean by averaging the quantile functions of observed empirical distributions. However, this estimator is severely biased in sparse sampling settings -- common in randomized studies where the number of observations per unit is limited. We introduce the Marginal-Corrected Barycenter (MCB) estimator, which explicitly corrects for the bias induced by sparse sampling. The MCB is motivated by the observation that the Wasserstein barycenter depends on the distribution of distributions through an intermediate object: the marginal distribution of the random cumulative distribution functions (CDFs). For each $x \in \mathbb{R}$, we estimate these marginals by framing the problem as a series of binomial mixture models. This reduces complex functional estimation to a well-studied non-parametric mixing distribution problem. We establish the conditions under which the MCB is consistent and asymptotically normal. Furthermore, we demonstrate through simulation studies that the MCB substantially outperforms the empirical barycenter when sample sizes per distribution are small. This provides a robust framework for non-parametric inference when the target of interest is a structural mean in the presence of significant sampling randomization.

Nonparametric Assessment of Causal Interactions with Continuous-Valued Exposures Under Time-Varying Confounding

D. Benkeser[1], A. Codi[1], E. Rogawski McQuade[1]

[1]Emory University, Rollins School of Public Health, United States

Enteric infections are a major contributor to childhood illness and death, particularly in low-resource settings where children are exposed to multiple pathogens. Understanding whether multiple pathogens interact to produce worse clinical outcomes is critical for prioritizing prevention strategies. Evaluating such interactions is challenging as pathogen burdens are continuous-valued, high-dimensional, and can be confounded by antibiotic treatment decisions—a major determinant of outcomes for bacterial infections. To address these challenges, we propose a framework to characterize pathogen interactions by mimicking factorial challenge trials. Our estimands are defined by interventions that draw quantities for each pathogen from a specified distribution and compare regimes in which one or more pathogens are shifted to a high-level distribution while remaining pathogens are shifted to a low-level distribution. The estimands allow the use of doubly robust, machine-learning–based estimators that flexibly learn interaction effects. We apply the methods to enteric infection data from five large multi-site studies and discuss implications for intervention targeting and clinical management.

Randomization Inference with Sample Attrition

X. Li[1], P. Sheng[1], Z. Yu[2]

[1]University of Chicago, United States
[2]Princeton University, United States

Although appealing, randomization inference for treatment effects can suffer from severe size distortion due to sample attrition. We propose new, computationally efficient methods for randomization inference that remain valid under a range of potentially informative missingness mechanisms. We begin by constructing valid p-values for testing sharp null hypotheses, using the worst-case p-value from the Fisher randomization test over all possible imputations of missing outcomes. Leveraging distribution-free test statistics, this worst-case p-value admits a closed-form solution, connecting naturally to bounds in the partial identification literature. Our test statistics incorporate both potential outcomes and missingness indicators, allowing us to exploit structural assumptions, such as monotone missingness, and information about the distribution of missingness types to increase power. The methods are illustrated through simulations and an empirical application.

In memory of Professor Alain Berlinet

9:00 - 9:30

A stochastic version of Iterated Biased Reduction

P. Cornillon[1], N. Hengartner[2], E. Le Pennec[3], E. Matzner-Lober[4]

[1]IRMAR Université Rennes 2, France
[2]Los Alamos National Laboratory, United States
[3]CMAP Ecole Polytechnique, France
[4]CEPE ENSAE, France

Iterated Bias Reduction (IBR) is an iterative algorithm for regression estimation. Like boosting, it starts with a weak learner (a biased estimator) and corrects it at each iteration by using information contained in the residuals. Unlike classical boosting, the initial learner can be a nonparametric estimator with certain properties.

However, IBR also has limitations. It can be computationally expensive due to repeated iterations, and defining what constitutes “acceptable” bias is often subjective and context-dependent. Additionally, IBR requires computing the smoothing matrix S, which can be problematic when the size of the dataset is large. In this presentation, we propose a stochastic version of IBR that can be used regardless of the size of the dataset.

Dynamic CAPM with long memory factors

C. Francq[1], J. Royer[2], J. Zakoian[1]

[1]CREST-ENSAE, France
[2]CREST, Institut Polytechnique de Paris, France

Factor models are widely used in finance, and recent advances allow slope coefficients (known as 'betas') to vary over time.
However, estimation theory for these dynamic conditional betas typically relies on short-memory volatility models, which can be restrictive in empirical applications. Moreover, exogenous variables such as realized moments have proven useful in recent volatility modeling studies.
In this paper, we introduce a multivariate framework that allows for time-varying betas where covolatilities can exhibit higher persistence than standard exponential decay. We incorporate covariates in the dynamics of both conditional variances and betas. We establish stationarity conditions for the proposed model, prove the consistency and asymptotic normality of the quasi-maximum likelihood (QML) estimator, and propose goodness-of-fit tests. Monte Carlo experiments assess the finite-sample performance of the estimation procedure. Finally, we discuss the choice of relevant exogenous variables and illustrate the model's effectiveness through real data applications.

The Pivotal Information Criterion

S. Sardy[1], M. van Cutsem[1], S. van de Geer[2]

[1]Université de Genève, Switzerland
[2]ETHZ, Switzerland

We derive an information criterion that has three advantages over BIC: selection of λ is (asymptotically) pivotal, the information criterion is a continuous function of the parameters, and selection of λ allows to retrieve the correct input covariates with high probability. The Pivotal Information Criterion can be seen as the extension of square-root LASSO to the location scale and exponential families and to survival analysis. PIC can be employed for a linear model or an artificial neural network.

Recent Advances in Statistical Ranking

9:00 - 9:30

Model-free Rank Aggregation: a Maximum Score Approach

H. Zhang[1], Y. CHEN[2]

[1]Southern University of Science and Technology, China
[2]London School of Economics and Political Science, United Kingdom

This paper addresses the rank aggregation problem using multi-way comparison data derived from raters' scores. Unlike traditional methods that rely on parametric models (e.g., Bradley-Terry or Plackett-Luce models), we consider a model-free setting that assumes only the existence of a latent global ranking and a weak monotonicity constraint for each rater. This flexible setting accounts for heterogeneity in raters and also covers the weak stochastic transitivity setting for pairwise comparison data as a special case. To learn this global ranking, we introduce a distribution-free maximum score estimator that automatically accounts for rater heterogeneity. This estimator is effective across both dense and sparse regimes, accommodating scenarios where the number of ratings per rater is either a fixed constant or diverges with the total number of items. We establish the theoretical foundations of the proposed estimator by proving its consistency, showing that the proportion of discordant pairs (Kendall's tau) converges to zero in probability as the number of raters diverges. Furthermore, we derive tight upper and lower bounds for a performance metric based on Kendall's tau. We show that under specific asymptotic regimes, these bounds match, establishing the minimax optimality of our method. The practical utility of the estimator is demonstrated through extensive simulations and two real-data applications, one on ranking sport players and the other on aggregating preferences of survey respondents.

Deep Ranking with Heterogeneous Effect

Y. Luo[1], S. Fang[1], R. HAN[1], Y. Xu[2]

[1]The Hong Kong Polytechnic University, Hong Kong
[2]University of Kentucky, United States

Classical latent-score ranking models often fail to distinguish objects' intrinsic scores from contextual effects, which are typically nonlinear and can dominate the observed outcomes. To address this, we introduce a semiparametric ranking framework in which the log-score of each object is modeled as the sum of a utility parameter and a nonparametric covariate effect. We establish model identifiability under mild regularity and connectivity conditions. For estimation, we approximate the covariate effect using a neural network and estimate the parameters via the maximum likelihood estimator (MLE). We prove that the MLE exists with high probability under random design assumptions and derive non-asymptotic error bounds that achieve minimax optimality. Numerical experiments on both synthetic data and an ATP tennis dataset are conducted to support our findings.

Spectral Ranking Inferences Based on General Multiway Comparisons

W. Wang[1], J. Fan[2], Z. Lou[3], M. Yu[4]

[1]University of Hong Kong, Hong Kong
[2]Princeton University, United States
[3]UC San Diego, United States
[4]Washington University in St. Louis, United States

This paper studies the performance of the spectral method in the estimation and uncertainty quantification of the unobserved preference scores of compared entities in a general and more realistic setup. Specifically, the comparison graph consists of hyper-edges of possible heterogeneous sizes, and the number of comparisons can be as low as one for a given hyper-edge. Such a setting is pervasive in real applications, circumventing the need to specify the graph randomness and the restrictive homogeneous sampling assumption imposed in the commonly used Bradley-Terry-Luce (BTL) or Plackett-Luce (PL) models. Furthermore, in scenarios where the BTL or PL models are appropriate, we unravel the relationship between the spectral estimator and the maximum likelihood estimator (MLE). We discover that a two-step spectral method, where we apply the optimal weighting estimated from the equal weighting vanilla spectral method, can achieve the same asymptotic efficiency as the MLE. Given the asymptotic distributions of the estimated preference scores, we also introduce a comprehensive framework to carry out both one-sample and two-sample ranking inferences, applicable to both fixed and random graph settings. It is noteworthy that this is the first time effective two-sample rank testing methods have been proposed. Finally, we substantiate our findings via comprehensive numerical simulations and subsequently apply our developed methodologies to perform statistical inferences for statistical journals and movie rankings.

Preference-based Centrality and Ranking in General Metric Spaces

L. Lyu[1], D. Zhou[2]

[1]University of Science and Technology of China, China
[2]National University of Singapore, Singapore

Ranking or assessing centrality in multivariate and non-Euclidean data is difficult because there is no canonical order and many depth notions become computationally fragile in high-dimensional or structured settings. We introduce a preference-based notion of centrality defined through population proximity comparisons with respect to a random reference draw, yielding a metric-intrinsic statistical functional that is well-defined on general metric spaces. Because the induced pairwise preferences may be non-transitive, we map them to a coherent one-dimensional score via a Bradley--Terry--Luce cross-entropy projection, viewed as a calibrated aggregation device rather than a correctly specified model. We develop two finite-sample estimators a convex M-estimator and a fast spectral estimator based on a comparison operator, and establish identifiability and consistency under mild conditions. Simulations and real-data examples, including high-dimensional and functional observations, illustrate that the proposed scores provide stable, interpretable rankings aligned with the underlying preference centrality.

Novel False Discovery Rate Methodology

9:00 - 9:30

The local false discovery rate without a prior: from frequentist foundations to nonparametric estimation and control

W. Fithian[1], J. Soloff[2], D. Xiang[3]

[1]UC Berkeley, United States
[2]University of Michigan, United States
[3]University of Chicago, United States

The false discovery rate (FDR) has become the dominant error criterion in large-scale multiple testing, but it measures only the average quality of a set of rejections, leaving open whether each individual discovery is worth pursuing. The local false discovery rate (lfdr) — the probability that a specific hypothesis is truly null given its test statistic — addresses this directly, but has traditionally required Bayesian assumptions difficult to justify in many scientific settings. In this talk I present a unified framework, developed across two papers, that grounds the lfdr in frequentist foundations using nonparametric tools. I first define a frequentist lfdr as the relative frequency of null hypotheses at each point in the sample space, requiring no prior. This frequentist lfdr preserves the key properties of its Bayesian counterpart: it gives a conditional probability that a hypothesis whose test statistic takes a given value is null, it produces calibrated forecasts of hypothesis truth status, and thresholding it yields the optimal separable rejection rule under weighted classification loss. I then present the Support Line (SL) procedure which, under independence assumptions, controls the probability that the least promising rejection is a false discovery in finite samples without knowledge of the prior. The SL procedure is equivalent to thresholding a plug-in lfdr estimate based on Grenander's nonparametric monotone density estimator, requiring only the shape constraint that smaller p-values represent stronger evidence against the null. Its asymptotic behavior is governed by cube-root convergence, arising from the shape-constrained estimation at the rejection boundary, and it achieves optimal empirical Bayes regret at rate $m^{-2/3}$. Together, these results show that the lfdr's advantages over the FDR can be realized without Bayesian assumptions or parametric modeling of the alternative (only nonparametric shape constraints), enabling fine-grained, individually interpretable multiple testing in complex scientific applications.

Conformal novelty detection with false discovery rate control at the boundary

Z. Gao[1], D. Xiang[2], E. Roquain[3]

[1]USC Marshall Business School., United States
[2]University of Chicago, United States
[3]Sorbonne Université, France

Conformal novelty detection is a classical machine learning task for which uncertainty quantification is essential for providing reliable results. Recent work has shown that the BH procedure applied to conformal p-values controls the false discovery rate (FDR). Unfortunately, the BH procedure can lead to over-optimistic assessments near the rejection threshold, with an increase of false discoveries at the margin as pointed out by Soloff et al. (2024). This issue is solved therein by the support line (SL) correction, which is proven to control the boundary false discovery rate (bFDR) in the independent, non-conformal setting. The present work extends the SL method to the conformal setting: first, we show that the SL procedure can violate the bFDR control in this specific setting. Second, we propose several alternatives that provably control the bFDR in the conformal setting. Finally, numerical experiments with both synthetic and real data support our theoretical findings and show the relevance of the new proposed procedures.

Selecting Informative Conformal Prediction Sets with an Optimized FCR-Controlled Approach

r. heller[1], I. Solomon[1], s. rosset[1], E. Roquain[2]

[1]tel-aviv university, Israel
[2]Sorbonne Université, France

Conformal methods provide prediction sets for outcomes with confidence guarantees. We study their use in a selective inference setting, where inference is performed only when the prediction set is informative. The analyst may consider as informative, for example, cases with prediction sets that are sufficiently small, excluding null values, or satisfy other appropriate monotone constraints. A general framework for constructing such informative conformal prediction sets while controlling the false coverage rate (FCR) on the selected sample was suggested in Gazin et al. (2025). In this work we focus on oracle-guided procedures. We derive the optimal decision policy under a suitable power objective in the oracle setting where the probability of belonging to each prediction set can be computed. In practice, of course, only estimated probabilities are available. So we introduce a calibration procedure that adjusts the oracle policy to maintain finite sample FCR control. We show that this approach can achieve substantially higher power than available alternatives. We demonstrate the effectiveness of our new methods for classification outcomes on both real and simulated data.

Rank confidence intervals via FDR control

Y. Benjamini[1], Y. Benjamini[2]

[1]Hebrew University of Jerusalem, Israel
[2]Tel Aviv University, Israel

The interpretation of feature importance and the benchmarking of model performance are two modern challenges that rely heavily on the relative ordering of items (ranking) rather than their raw values. However, because these rankings can be highly unstable, recent frameworks have proposed quantifying this uncertainty through simultaneous confidence intervals for ranks based on pairwise comparisons. While statistically sound, this simultaneous approach becomes overly conservative as the number of items increases—a common characteristic of modern datasets.
In this paper, we propose a more powerful alternative that adapts the logic of the False Discovery Rate (FDR). By allowing a controlled proportion of errors in the coverage of rank confidence intervals relative to their severity, we provide a more flexible and scalable inference framework. We further develop methods to select a set of top items while maintaining rigorous error control.

Recent Modeling and Computational Advances in Bayesian Nonparametrics

9:00 - 9:30

Feature allocation models for multiple populations with imperfect detection

F. Stolf[1]

[1]Duke University, United States

Feature allocation models are a widely used class of Bayesian nonparametric methods for modeling binary latent feature matrices with a potentially unbounded number of features. They are particularly well-suited for species occurrence data, as they naturally accommodate an ever-growing number of species. However, their application to biodiversity studies has been limited by unrealistic assumptions of full exchangeability and perfect detection. Here, we introduce a novel and tractable class of feature allocation models for partially exchangeable data with imperfect detection. Our approach incorporates key ideas from ecological occupancy models to explicitly account for detectability, while allowing for heterogeneity across groups of samples corresponding to distinct regions. We provide a comprehensive theoretical analysis, deriving closed-form expressions for quantities of interest, including in-sample and predictive distributions for the total number of species and for the number of species shared between groups. This analytical tractability enables scalable and efficient computation. We further introduce a novel measure for quantifying diversity across regions under a coherent probabilistic framework, yielding interpretable and practically relevant tools for large-scale biodiversity monitoring studies.
Simulation studies and applications to global fungal biodiversity data demonstrate the effectiveness of the proposed modeling class.

Graphical Pitman-Yor Process for Clustering in Bayesian Networks

I. Golovko[1], A. Lijoi[1], I. Prünster[1]

[1]Bocconi University, Italy

Hierarchical Bayesian nonparametric models have been highly successful for analysing grouped data, yet most available methods treat groups as exchangeable. We address settings where groups are related but not symmetric by representing their dependence with a directed acyclic graph (DAG). Building on the graphical Dirichlet process, we introduce the graphical Pitman–Yor (GPY) process, which adds a discount parameter that induces heavy-tailed cluster-size behaviour and delivers two practical advantages: more realistic power-law clustering in network settings and adaptive information pooling along the DAG. We characterise the predictive distributions and derive tractable conditional distributions that yield an exact marginal Gibbs sampler on DAGs, enabling efficient inference without truncation. Simulation studies show consistent gains over the Dirichlet special case and standard baselines, particularly when cluster sizes follow power-law patterns and when the DAG captures meaning- ful relations among groups. The approach applies broadly to network-structured problems.

Bayesian calculus and predictive characterizations of extended feature allocation models

L. Ghilotti[1], M. Beraha[2], F. Camerlenghi[2]

[1]Duke University, United States
[2]University of Milano-Bicocca, Italy

We introduce and study a unified Bayesian framework for extended feature allocations which flexibly captures interactions -- such as repulsion or attraction -- among features and their associated weights. We provide a complete Bayesian analysis of the proposed model and specialize our general theory to noteworthy classes of priors. This includes novel priors based on (i) determinantal point processes, which yield promising results in a spatial statistics application, and (ii) shot noise Cox processes, illustrated with a genetics example.
Within the general class of extended feature allocations, we further characterize those priors that yield predictive probabilities of discovering new features depending either solely on the sample size or on both the sample size and the distinct number of observed features.
These predictive characterizations, known as "sufficientness" postulates, have been extensively studied in the literature on species sampling models starting from the seminal contribution of the English philosopher W.E. Johnson for the Dirichlet distribution.
Within the feature allocation setting, existing predictive characterizations are limited to very specific examples; in contrast, our results are general, providing practical guidance for prior selection.

Conformalized Bayesian Inference: Uncertainty Quantification for Nonstandard Parameter Spaces

N. Bariletto[1]

[1]University of Texas at Austin, United States

Bayesian posterior distributions naturally represent parameter uncertainty informed by data. When the parameter space is complex — as in nonparametric settings where it is infinite-dimensional or combinatorially large — standard summaries such as posterior means, credible intervals, or assessments of multimodality are often unavailable, hindering interpretable posterior uncertainty quantification. We introduce Conformalized Bayesian Inference (CBI), a broadly applicable and computationally efficient framework for posterior inference on nonstandard parameter spaces. Requiring only Monte Carlo samples from the posterior and a notion of discrepancy between parameters, CBI yields three key outputs: a point estimate, a credible region with assumption-free posterior coverage guarantees, and a principled analysis of posterior multimodality. The method constructs a pseudo-density score for each parameter value, producing a MAP-like point estimate and a credible region derived from conformal prediction principles. The central conceptual contribution is a reinterpretation of posterior inference as a prediction problem on the parameter space, enabling the importation of conformal coverage guarantees into Bayesian uncertainty quantification. A density-based clustering step further identifies representative posterior modes. We establish theoretical and methodological properties of CBI and demonstrate its practicality, scalability, and versatility through applications to Bayesian clustering and variable selection problems.

Contributed: Survival, Censoring and Event-Time Modeling

9:00 - 9:20

Nonparametric Estimation of the Cross Ratio Function under Right Censoring

Ö. Sercik[1], A. Verhasselt[1], S. Abrams[1, 2]

[1]Hasselt University, Belgium
[2]University of Antwerp, Belgium

The cross-ratio function (CRF) is a commonly used local dependence measure describing the strength of association between two time-to-event variables, for such as infection times for two pathogens in the same individual, or failure times for system components in reliability theory. The CRF can be written in terms of (first and second order derivatives of) the joint survival function of these random variables. Although, parametric, semi-parametric and non-parametric estimators for the CRF have been proposed in the literature forbivariate right-censored time-to-event data, theseestimators are either based on very strong parametric assumptions regarding the underlying association structure, or these are of little practical use due to their rough behaviour yielding unsatisfying finite sample performance. In this work, we propose a novel non-parametric estimator for the CRF under univariate right censoring, based on Bernstein polynomials. We will discuss its performance using a simulation study and show theoretical properties of the estimator. Moreover, application of the proposed estimator in a real-life data application is demonstrated. In general, the Bernstein-based estimator shows good finite sample performance though it depends on the selection of the Bernstein order which is explored in detail.

Penalized Variable Selection with Broken Adaptive Ridge Regression for Semi- competing Risks Data

F. Mahmoudi[1], X. Lu[2]

[1]Mount Royal Univeristy, Canada
[2]University of Calgary, Canada

Semi-competing risks data arise when both non-terminal and terminal events are considered in an illness-death model. Such data with multiple events of interest are frequently encountered in medical research and clinical trials. Unlike some recent works on penalized variable selection that deal with the competing risks separately without incorporating possible correlation between them, we perform variable selection in the illness-death model using shared frailty. We propose a broken adaptive ridge (BAR) penalty to encourage sparsity and perform variable selection in an event-specific manner so that the potential risk factors can be selected and their effects can be estimated simultaneously, corresponding to each event in the study. The oracle property of the proposed BAR procedure is established, and its performance is evaluated and compared with other commonly used methods by simulation studies. The proposed method is then applied to the real-life data arising from a colon cancer study.

Kernel estimators of the reliability/survival function and related indicators for semi-Markov processes

V.S. Barbu[1], C. Ayhar[2], F. Mokhtari[3], S. Rahmani[4]

[1]University of Rouen - Normandy, France
[2]University Center of El Bayadh Nour El Bachir, Algeria, Algeria
[3]University of Saida Algeria, Algeria
[4]LSMSA, University of Saida–Doctor Moulay Taher, Algeria

Our presentation introduces nonparametric kernel estimators of the reliability/survival function, availability and failure rate of a semi-Markov process. After constructing the nonparametric kernel estimators, we establish asymptotic properties of these estimators, when the sample size becomes large. The qualities of the estimators are illustrated by a numerical example.
This is a joint work with:
Chafiâa AYHAR (University Center of El Bayadh Nour El Bachir, Algeria; ayharchafiaa@yahoo.com),
Fatiha MOKHTARI (LSMSA, University of Saida–Doctor Moulay Taher, Algeria; fatiha.mokhtari@univ-saida.dz),
Saâdia RAHMANI (LSMSA, University of Saida–Doctor Moulay Taher, Algeria; saadia.rahmani@univ-saida.dz).

Order restricted estimation of the parameter functions in an additive hasard model

D. Anevski[1], E.M. Merai[2]

[1]Lund University, Sweden
[2]University of Constantine 1, Algeria

We derive limit distributions for order restricted estimators of the individual functions in a Aalen additive hasard model, in an a survival analysis data setting. The ultimate goal for the work is to provide a graphical model approach in a survival analysis setting, thereby enabling modelling and interpretation of direct and indirect effects of covariates on the (distribution of the) time to event.

Contributed: High-Dimensional Learning, Neural Methods and Modern Computation

9:00 - 9:20

From robust neural networks toward robust nonlinear quantile estimation

J. Kalina[1, 2]

[1]The Czech Academy of Sciences, Institute of Information Theory and Automation, Czech Republic
[2]The Czech Academy of Sciences, Institute of Computer Science, Czech Republic

Regression quantiles provide a flexible framework for modeling the conditional distribution of a response variable by estimating different parts of its distribution, thereby offering valuable insights into the relationship between predictors and outcomes. However, existing nonlinear regression quantile methods may be sensitive to the presence of severe
outliers in the data. This paper starts with investigating robust versions of neural networks. The study includes a proposal of a sequential outlier detection procedure based on sequential example selection for robust neural networks. Further, robust quantile estimators for nonlinear regression are introduced. The proposed quantiles are inspired by least
weighted squares regression. To enhance robustness to outliers, they assign implicit weights to individual samples and are specifically tailored for multilayer perceptrons, radial basis function networks, and regularized networks. Numerical experiments demonstrate that the robust quantiles improve generalization and outlier resistance. Simulations confirm
that the proposed method outperforms traditional non-robust quantiles.

A comparative study of Reservoir Computing and LSTM neural network for trajectory reconstruction in dynamical systems

E. Papadopoulou[1], D. Kugiumtzis[1]

[1]School of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece

The task of modelling and forecasting dynamical systems has been a key challenging subject of research across a wide range of scientific fields. Machine learning models are found to be capable of predicting effectively future states of the underlying dynamical system based on its past states. However, little attention has be given on assessing whether the forecasts preserve the dynamical properties of the underlying system, and the machine learning model has learned the complex underlying dynamics from the observed timeseries data [1]. We attempt to do this in this study. Two different recurrent neural networks models are utilized to reproduce the trajectories of deterministic dynamical systems: a Long Short-Term Memory neural network (LSTM) and an Echo State Network (ESN) [2]. The first model is specifically designed to capture the long-term dependencies in time series and to overcome the vanishing gradient problem encountered in standard recurrent neural networks. The second model belongs to the class of Reservoir Computing networks and is known for its high computational efficiency and low training complexity because of the fixed weight matrices related to the recurrent part of the model. The two neural networks are trained on multivariate time series from a complex dynamical system, and then they make long term predictions without any external forcing, i.e., they generate future states for long times, simulating the dynamical system trajectory. Their forecast performance is evaluated and compared not only through prediction error metrics but also through statistical dependence measures, which can quantify similarity between reference (ground-truth) and generated (model-produced) trajectories.

References

[1] Z. Fang, G. Mengaldo. Dynamical errors in machine learning forecasts. Chaos, Solitons & Fractals, 201 (2025) 117376.

[2] L. A. Hurley, S. E. Shaheen. Reservoir computing with large valid prediction time for the Lorenz system. arXiv, (2025) 2508.06730.

Minimum Norm Interpolation via The Local Theory of Banach Spaces: The Role of 2-Uniform Convexity

G. Kur[1]

[1]ETH Zurich, Switzerland

The minimum-norm interpolator (MNI) framework has recently attracted considerable attention as a tool for understanding generalization in overparameterized models, such as neural networks. In this work, we study the MNI under a 2-uniform convexity assumption, which is weaker than requiring the norm to be induced by an inner product, and it typically does not admit a closed-form solution. At a high level, we show that this condition yields an upper bound on the MNI bias in both linear and nonlinear models. We further show that this bound is sharp for overparameterized linear regression when the unit ball of the norm is in isotropic or, John's position, and the covariates are isotropic, symmetric, i.i.d. sub-Gaussian, such as vectors with i.i.d. Bernoulli entries. Finally, under the same assumption on the covariates, we prove sharp generalization bounds for the ℓ_p-MNI when $p \in \bigl(1 + C/\log d, 2\bigr]$. To the best of our knowledge, this is the \emph{first} work to establish sharp bounds for non-Gaussian covariates in linear models when the norm is not induced by an inner product. This work is deeply inspired by classical works on K-convexity, and more modern work on the geometry of 2-uniform and isotropic convex bodies.

Keynote Talk

11:30 - 12:30

Distributed-oracle estimation for high-dimensional quantile regression

X. He[1]

[1]Washington University in St. Louis, United States

Quantile regression (QR) is a valuable tool for analyzing heterogeneous covariate effects across the entire outcome distribution including lower and upper tails. However, implementing QR in high-dimensional settings where data are decentralized presents computational and communication hurdles. We propose a communication-efficient estimator for high-dimensional QR designed for data distributed across multiple machines. To use folded-concave penalties, we develop an iterative multi-step (IM) algorithm utilizing a surrogate smoothed quantile loss. This approach effectively balances statistical efficiency with communication constraints. To provide a theoretical foundation for our method, we introduce the concept of a distributed-oracle estimation and demonstrate that the IM estimator converges to this oracle with high probability. Furthermore, we extend our framework to enable distributed inference for specific low-dimensional components of interest. This talk is based on joint work with SongshanYang, Yifan Gu, and Hangfang Yang.

Hypothesis testing and related topics

8:30 - 9:00

Testing independence of errors and covariates in functional linear models

M. Birke[1], N. Neumeyer[2]

[1]University of Bayreuth, Germany
[2]University of Hamburg, Germany

In functional linear models the assumption of independence of errors and covariates is often essential for data analysis procedures. Neglecting certain dependencies will result in inconsistent parameter estimates and, as a consequence, misleading conclusions in applications. In the case of functional covariates and scalar responses we suggest two different hypothesis tests based on empirical residual characteristic functionals. In a first one the squared distance between the characteristic functionals in general and under independence is integrated with respect to some measure on the Borel-sigma filed on a Hilbert space. The second approach uses a truncated basis representation of the Hilbert space valued observations. Indepence can then be expressed via the characteristic functions of residuals and projections onto the basis functions which results in a sum of double integrals of the squared difference of the characteristic functions over the real numbers. For both approaches we show consistency of the resulting tests as well as the asymptotic distributions under the null hypothesis. As well known, bootstrap procedures work better in such settings. Therefore we base our tests on a residual bootstrap. In a simulation study we investigate the finite sample performance of both tests and discuss the limitations and advantages.

Detecting Practically Significant Dependencies in Metric Spaces via Distance Correlations

M. Kroll[1], H. Dette[1]

[1]Ruhr University Bochum, Germany

We take a different look at the problem of testing the independence of two metric-space-valued random variables using the distance correlation. Instead of testing if the distance correlation vanishes exactly, we are interested in the hypothesis that it does not exceed a certain threshold. Our testing problem is motivated by the observation that in many cases it is more reasonable to test for a practically significant dependency since it is rare that a hypothesis of perfect independence is exactly satisfied. This point of view also reflects statistical practice, where one often classifies the strength of the association in categories such as `small', `medium' and `large' and the precise definitions depend on the specific application. To address these problems we develop a pivotal test for the hypothesis that the distance correlation between two random variables does not exceed a pre-specified threshold Δ. The new test is applicable to processes taking values in separable metric spaces of strong negative type, covering Euclidean as well as functional data. We do not assume independent observations, and instead prove our results for absolutely regular sample generating processes, which includes many time series such as ARMA and GARCH models. Our approach is based on a new functional limit theorem for the sequential distance correlation process, and can also be used to construct confidence intervals for the distance correlation without the need for resampling.

Change-Point Detection in Functional Regression Models

A. Batsidis[1], B. Milošević[2], M. Veljović[3]

[1]University of Ioannina, Greece
[2]Faculty of Mathematics, University of Belgrade, Serbia
[3]University of Belgrade, Faculty of Mathematics, Serbia

Change-point detection in the error distribution of functional linear regression models is investigated. While existing approaches rely on residual-based empirical distribution functions, a new class of test statistics based on the empirical characteristic function of the estimated residuals is proposed.
The asymptotic properties of the proposed test statistic are established, and its finite-sample performance is evaluated through simulations under several alternatives, including mixture and skewed error distributions. The results indicate that the proposed method performs competitively in detecting structural changes in the error distribution.

Testing for changes in the error distribution in functional regression models

L. Selk[1], N. Neumeyer[1]

[1]University of Hamburg, Germany

In this talk regression models with scalar responses and functional covariates are considered. The aim is to detect change points in the error distribution, based on sequential residual empirical distribution functions. A nonparametric Nadaraya-Watson type method is proposed to estimate the regression function, and the resulting change-point tests in this setting and in a functional linear model are compared. In both cases, the suggested change point test is asymptotically distribution-free and consistent for one-change point alternatives.

Temporal Modeling and Causal Inference

8:30 - 9:00

Understanding tree water usage and stress via sap flux density time series

M. Gong[1], R. Killick[1], A. Hirons[2]

[1]Lancaster University, United Kingdom
[2]Myerscough College, United Kingdom

Understanding tree water usage patterns and their responses to extreme weather conditions are crucial to the irrigation management of agriculture, forestry and horticulture, especially under climate change. Modern sensors provide an opportunity to carry out real-time monitoring of the behaviour of trees. High-frequency time series data obtained by sensors enable the investigation of the complex relationships between tree sap flux density and weather patterns, and the construction of prediction models for forecasting tree water usage.

Here we developed an ensemble prediction approach based on non-parametric regression models capturing the non-linear relationship between tree sap flux density and its environmental drivers. In particular, the ensemble elements are taken to be the additive models fitted to individual trees monitored in the field study. Considering the large variation in the temporal patterns of the sap flux density among individual trees, even if they belong to the same species, we proposed to combine the predictions from the individual tree models via appropriate weighting to form an ensemble prediction that is more representative of the species. The method demonstrated good performance on the data collected from nine species over three growing seasons at Hillier Nurseries in England, with the only exception being two species during the UK heatwaves of summer 2022.

To explore the impact of extreme weather events, such as heatwaves, on modelling and prediction, we took inspiration from plant science literature and extended the basic sap flux density model to a threshold regression model using soil moisture content as the threshold variable. This allows us to investigate the impact of soil water availability on the change of water extraction behaviour of trees. We plan to further bring in the dendrometer (which measures the growth of trees) time series to investigate the stress response of trees to heatwaves.

Conformal time series forecasting

S. Ben Taieb[1]

[1]MBZUAI, United Arab Emirates

Conformal prediction provides a flexible, model-agnostic framework for uncertainty quantification with finite-sample, distribution-free guarantees. However, these guarantees rely on exchangeability, an assumption often violated in time series due to temporal dependence and non-stationarity. As a result, standard conformal methods can produce unreliable prediction regions in time series applications. This presentation introduces the key challenges of applying conformal prediction to time series forecasting and reviews recent methodological advances designed to address non-exchangeability. We outline conditions under which approximate validity can still be achieved and survey a range of approaches, including reweighting schemes, adaptive residual updates, and online calibration strategies that dynamically recalibrate prediction regions to maintain target coverage over time. Finally, we present empirical comparisons highlighting the trade-offs between coverage accuracy, efficiency, and computational cost. The talk concludes with open questions and practical considerations for deploying conformal methods in forecasting.

Assumption-lean Aalen Regression

N.R. Hansen[1], A. Mangulad Christgau[1]

[1]University of Copenhagen, Denmark

We propose the Aalen covariance measure as a novel quantification of the conditional association between a time-to-event and an exposure given a set of covariates. It can be interpreted as a weighted hazard difference and viewed as an assumption-lean generalization of the cumulative exposure coefficient in the Aalen additive hazards model. We develop an estimation method based on the principles of double machine learning of the Aalen covariance measure. We prove that our method is doubly rate robust meaning that with modest rate conditions on learning the nuisance functions, the resulting estimator converges at a root-n-rate. The estimator is made available via an R package, and a simulation study shows that it is as efficient as the Aalen regression estimator in the additive hazards model while being robust to model misspecifications.

Can language models boost the power of randomized experiments without statistical bias?

W. Wei[1]

[1]University of Southern California, United States

Randomized experiments or randomized controlled trials (RCTs) are gold standards for causal inference, yet cost and sample-size constraints limit power. We introduce CALM (Causal Analysis leveraging Language Models), a statistical framework that integrates large language models (LLMs) generated insights of RCTs with established causal estimators to increase precision while preserving statistical validity. In particular, CALM treats LLM-generated outputs as auxiliary prognostic information and corrects their potential bias via a heterogeneous calibration step that residualizes and optimally reweights predictions. We prove that CALM remains consistent even when LLM predictions are biased and achieves efficiency gains over augmented inverse probability weighting estimators for various causal effects. In particular, CALM develops a few-shot variant that aggregates predictions across randomly sampled demonstration sets. The resulting U-statistic-like predictor restores i.i.d. structure and also mitigates prompt-selection variability. Empirically, in simulations calibrated to a mobile-app depression RCT, CALM delivers lower variance relative to other benchmarking methods, is effective in zero- and few-shot settings, and remains stable across prompt designs. By principled use of LLMs to harness unstructured data and external knowledge learned during pretraining, CALM provides a practical path to more precise causal analyses in RCTs.

Recent advances in survival analysis

8:30 - 9:00

Semiparametric Regression Analysis of Interval-Censored Multi-State Data with An Absorbing State

D. Zeng[1]

[1]University of Michigan, United States

In studies of chronic diseases, the health status of a subject can often be characterized by a finite number of transient disease states and an absorbing state, such as death. The times of transitions among the transient states are ascertained through periodic examinations and thus interval-censored. The time of reaching the absorbing state is known or right-censored, with the transient state at the previous instant being unobserved. We provide a general framework for analyzing such multi-state data. We formulate the effects of potentially time-dependent covariates on
the multi-state disease process through semiparametric proportional intensity models with random effects. We combine nonparametric maximum likelihood estimation with sieve estimation and develop a stable expectation-maximization algorithm. We establish the asymptotic properties of the proposed estimators and assess the performance of the proposed methods through extensive simulation studies. Finally, we provide an illustration with a cardiac allograft vasculopathy study.

Goodness-of-fit tests for censored and truncated data: maximum mean discrepancy over regular functionals

J.C. Escanciano[1], J. de Uña-Álvarez[2]

[1]Universidad Carlos III de Madrid, Spain
[2]Universidade de Vigo, Spain

We develop a systematic, omnibus approach to goodness-of-fit testing for parametric distributional models when the variable of interest is only partially observed due to censoring and/or truncation. In many such designs, tests based on the nonparametric maximum likelihood estimator are hindered by nonexistence, computational instability, or convergence rates too slow to support reliable calibration under composite nulls. We avoid these difficulties by constructing a regular (pathwise differentiable) Neyman-orthogonal score process indexed by test functions, and aggregating it over a reproducing kernel Hilbert space ball. This yields a maximum-mean-discrepancy-type supremum statistic with a convenient quadraticform representation. Critical values are obtained via a multiplier bootstrap that keeps nuisance estimates fixed. We establish asymptotic validity under the null and local alternatives and provide concrete constructions for left-truncated rightcensored data, current status data, and random double truncation; in particular, to the best of our knowledge, we give the first omnibus goodness-of-fit test for a parametric family under random double truncation in the composite-hypothesis case. Simulations and an empirical illustration demonstrate size control and power in practically relevant incomplete-data designs.

Flexible Archimedean copula models for dependent censoring

M. D'Haen[1], A. Verhasselt[2], I. Van Keilegom[1]

[1]KU Leuven, Belgium
[2]Hasselt University, Belgium

When studying survival data, one is often presented with the phenomenon of right censoring. As the common assumption of independent censoring is questionable in many applications, copula models have emerged as a popular alternative that can take dependence into account. It is well known, however, that the joint distribution of survival and censoring times cannot be nonparametrically identified without additional assumptions. The traditional copula-graphic estimator adopts nonparametric margins at the cost of a copula structure that needs to be fully known. In more recent literature, it is shown that parametric margins can be combined with a one-parameter copula family whose association parameter is identifiable from the data. While follow-up articles have focused on relaxing assumptions on the margins, in the current project we study flexibility in the dependence structure by proposing a new multi-parameter Archimedean copula family. In particular, we construct Archimedean generator functions starting from Bernstein polynomials, in such a way that both the strength as well as the shape of the resulting copula are flexible. Convergence properties of quantities relevant for maximum likelihood estimation are established; full model identifiability turns out to be a delicate issue.

Cox Regression on the Plane

Y. Travis-Lumer[1], M. Mandel[2], I.D. Fabian[3,4], R. Betensky[5], M. Gorfine[6]

[1]Technion, Israel
[2]The Hebrew University of Jerusalem, Israel
[3]Sheba Medical Center, Israel
[4]London School of Hygiene & Tropical Medicine, United Kingdom
[5]New York University, United States
[6]Tel Aviv University, Israel

The Cox proportional hazards model is the most widely used regression model in univariate survival analysis. Extensions of the Cox model to bivariate survival data, however, remain scarce. We propose two novel extensions based on a Lehmann-type representation of the survival function. The first, the simple Lehmann model, is a direct extension that retains a straightforward structure. The second, the generalized Lehmann model, allows greater flexibility by incorporating three distinct regression parameters and includes the simple Lehmann model as a special case.
For both models, we derive the corresponding regression formulations for the three bivariate hazard functions and discuss their interpretation and model validity. To estimate the regression parameters, we adopt a bivariate pseudo-observations approach. For the generalized Lehmann model, we extend this approach to accommodate a trivariate structure: trivariate pseudo-observations and a trivariate link function. We then propose a two-step estimation procedure, where the marginal regression parameters are estimated in the first step, and the remaining parameters are estimated in the second step. Finally, we establish the consistency and asymptotic normality of the resulting estimators. The proposed approach is validated through simulation studies and an application to data from the Global Retinoblastoma Outcome Study.

Time series analysis, modelling and applications

8:30 - 9:00

Augmented Dynamic Regression

Y. Li[1], L. Giraitis[2], G. Kapetanios[1], E. Hill[2]

[1]King's College London, United Kingdom
[2]Queen Mary University of London, United Kingdom

The recent work on regression modeling that permits general heterogeneity is extended to allow for lagged dependent variables. The purpose is to explore to what extent the generality of the setting, the simplicity of assumptions, and the ease of computation of standard errors can be preserved. Theoretical properties of regression estimation and inference is accompanied by Monte Carlo experiments and an empirical application.

Robust CDF-Filtering of a LocationParameter

A. Harvey[1], L. Catania[2], A. Luati[3, 4]

[1]University of Cambridge, United Kingdom
[2]University of Aarhus, United Kingdom
[3]Imperial College London, United Kingdom
[4]University of Bologna, Italy

This paper introduces a novel framework for designing robust
filters associated with signal plus noise models having symmetric
observation density.
The filters are obtained by a recursion where the innovation term is a transform of the cumulative distribution function of the residuals. The latter downweights extreme values by construction and
allows the filters to be analytically tractable. The updating scheme
naturally arises as the solution of an optimization problem, where the
objective function is a continuous version of the quantile check function,
formerly employed as a proper scoring function for quantiles and used to
construct robust minimum contrast estimators.
Stationarity, ergodicity and invertibility are derived under minimal
assumptions and preserved under different parametric specifications.
Estimation is carried out by the method of maximum likelihood and the
asymptotic theory is developed under misspecification. As an illustration, the new filters are
applied to brain scan data and compared, across Gaussian,
Student-t, Cauchy and Logistic density specifications, with alternative methods. Additional results include a novel class of score-driven models
and a subgaussian density suitable for robust filtering and modelling, arising as the infinite sum of independent non identically distributed uniform random variables.

Combining Prediction Intervals Using an Interval Score An Application to Electricity Price Forecasting

A. Giovannelli[1], T. Proietti[2], A. Cerasa[3], F. Nan[4]

[1]University of L'Aquila, Italy
[2]University of Rome, Italy
[3]European Commission Joint Research Centre, Italy
[4]EU Agency for the Cooperation of Energy Regulators (ACER), Italy

The aim of this study is to introduce a new procedure for combining prediction intervals from a set of candidates obtained by different models or conformal procedures. The goal is to obtain intervals that are more efficient according to a criterion based on an interval scoring rule. While conformal prediction guarantees marginal coverage, it does not ensure interval efficiency. In particular, when multiple predictors are available, the problem of combining the associated prediction intervals is non-trivial since both coverage and interval width should be considered jointly. To address this issue, we adopt a criterion that jointly accounts for sharpness and miscoverage by using a measure of interval width with an explicit penalty for noncoverage. The procedure is designed to preserve nominal coverage, while the combination weights are chosen by interval score optimisation. A simulation study is used to assess the finite-sample properties of the proposed procedure, confirming its ability to produce more efficient intervals while maintaining nominal coverage. It also confirms that a natural benchmark such as simple equal-weight aggregation of prediction interval does not, in general, lead to an efficient combined interval. An empirical application to electricity price forecasting in the Italian market further indicates that neither weights obtained from simple equal-weight interval combination nor rules based solely on interval width or miscoverage lead to superior results relative to our approach.

Change-Point Testing for Spectral Densities under Cyclical Long Memory

P. Sibbertsen[1]

[1]Univrsity of Hannover, Germany

This paper considers testing for changes in the spectral density matrix of a multivariate cyclical long-memory process. The spectral density is locally estimated by a lag-window spectral estimator before spectral average statistics are considered. For change-point testing a CUSUM-type test statistic based on these statistics is proposed and its limiting distribution is derived where possible. Its finite sample properties are investigated in a Monte Carlo study proposing a frequency-based sieve bootstrap approach. The paper is motivated by the debate if the Milankovitch cycles consisting of orbital earth variables such as eccentricity, obliquity and precession have a new or since long persisting effect on earth climate history. These Milankovitch cycles can be identified as poles in the spectral density of earth climate time series and thus the question arises if the location of these poles remain constant with time. Applying the change-point test to data of earth climate variables for the past 67 million years shows that the influence of the Milankovitch cycles is only recently a driving factor beginning with the icing of the Artic about 13 million years ago.

At the intersection between geometry and statistics

8:30 - 9:00

Data foliations in machine learning

G. NOVELLI[1], R. Fioresi[2], E. Latini[2], E. Tron[3]

[1]UNIMORE, Italy
[2]UNIBO, Italy
[3]ENAC, France

The aim of this work is to endow the data space of a simple neural network with a meaningful geometric structure. The approach relies on the Stefan–Sussmann theorem concerning singular foliations. This result will be used to obtain a decomposition of the data space into disjoint weakly embedded submanifolds called leaves.

The core of the construction is the Data Information Matrix, which is used to define two singular, non-smooth distributions on the data space: one corresponding to the level sets of the neural network, and the other given by its euclidean orthogonal complement. Although these structures are not smooth everywhere, their singularities are well behaved: the set of non-smooth points is a union of hyperplanes, closed and of measure zero. Therefore, the distributions are smooth almost everywhere.

The work also shows that the lack of smoothness can be resolved and presents some methods to achieve this. The simplest approach consists in slightly modifying the foliation by forcing the distribution to vanish on the singular hyperplanes. This in practice, forces the leaves on the hyperplanes to be points.

Some new results on kernel density estimation on a known Riemannian manifold.

A.F. Yao[1, 2], F. Nicol[3], M. Abdillahi Isman[1, 4], D.G. Kouadio[5, 6], V. Monsan[5]

[1]Université Clermont Auvergne/ Laboratoire de Mathématiques Blaise Pascal, France

[2]Centre de Mathématiques Appliquées, Ecole polytechnique, Paris, France

[3]Université de Toulouse, Laboratoire ENAC, France

[4]Université de Djibouti, France

[5]Université Félix Houphouet Boigny, Abidjan/Laboratoire de Mathématiques Informatique et Mécanique, Côte d'Ivoire

[6]Université des Lagunes, Abidjan, Côte d'Ivoire

In this work, we focus on the problem of estimating the density f of a variable X with values in a finite-dimensional Riemannian manifold, M. Specifically, we investigate Kernel Density Estimation (KDE) of f based on independent, identically distributed observations of X. We assume that M is compact and known.
In fact, one can find a KDE in M where the kernel is locally a transported density from a tangent space to M, as studied by Pelletier (2005) and Abdilahi et al. (2025). Here, we study some alternatives to Pelletier's KDE. While these KDE methods are easier to compute in practice and are often used in machine learning, their theoretical behaviour has not yet been studied. This work aims to contribute to addressing this issue.

Alongside the theoretical considerations, we will illustrate the practical behaviour through numerical applications on simulations.

References

Pelletier, B. (2005). Kernel density estimation on riemannian manifolds. Statistics & probability letters, 73(3):297–304.
Abdillahi Isman, M., Nefzi, W., Mbaye, P., Khardani, S. and Yao, A.-F. (2025), ‘Kernel density estimation for a stochastic process with values in a riemannian manifold’, Journal of Nonparametric Statistics 37(2), 344–363.

This work is partly supported by the Data Science Institute program in Côte d’Ivoire of Ecole polytechnique Paris, l’X (France).

Intrinsic regression on Riemannian manifolds : statistical modelling and properties.

F. Nicol[1], J. Aubray[1], A.F. Yao[2, 3]

[1]Université de Toulouse, Laboratoire ENAC, France
[2]Université Clermont Auvergne/ Laboratoire de Mathématiques Blaise Pascal, France
[3]Centre de Mathématiques Appliquées, Ecole polytechnique, Paris, France

This work addresses the issue of determining the position and orientation of an aircraft from noisy data in the context of air traffic management. Such data naturally lie on a special Euclidean Lie group that naturally models the joint evolution of position and orientation. Geodesic regression generalises linear regression to Riemannian manifolds by replacing straight lines with geodesics, defined through an affine connection (see Fletcher (2011) and Aubray and Nicol (2024)). Locally, the exponential map parametrises geodesics and expresses the geodesic distance. Estimation relies on minimizing the mean squared error defined by this distance, usually via gradient descent methods, as no global closed-form solution exists for the estimators.The main contribution of this work lies in the statistical analysis of the model, explicitly highlighting its fundamental statistical properties. The mean squared error of the estimation problem is reformulated as an integral of conditional Riemannian covariances along the regression solution curve by using the notion of Riemannian covariance developed in Abuqrais and Pigoli (2026). While many authors have addressed the regression problem in previous works, to the best of our knowledge, there has been no research into the statistical properties of these estimators as in those known in the Euclidean framework, except for peculiar cases such as symmetric spaces. We propose a nonparametric empirical estimator of the conditional Riemannian covariance to express the empirical mean squared criterion, thereby making the model’s statistical properties explicit. The geometric assumptions necessary for the existence and uniqueness of the relevant objects are stated, and numerical studies on simulated data are presented.

Fletcher T. (2011) Geodesic Regression on Riemannian Manifolds.
M Abuqrais, D Pigoli (2026). A Riemannian covariance for manifold-valued data. Journal of Mathematical Imaging and Vision 68 (2), 7
Aubray J and Nicol F. (2024) Polynomial regression on Lie groups and applications to SE(3). Entropy.

Convex generalized Fréchet means in a metric tree

G. Romon[1], V. Brunel[2]

[1]University of Luxembourg, Luxembourg
[2]ENSAE/CREST, France

Many datasets record where events occur along a network: accident locations on a road system, service requests on a utility network, or observations along a branched biological structure. In such settings the natural notion of distance is shortest-path length along the network. In this talk we study parameters of central tendency for a population on a tree-shaped network, modelled as a metric tree.

In this non-Euclidean setting, we develop location parameters called generalized Fréchet means, which are obtained by replacing the squared loss in the usual objective function with a generic convex increasing loss function. We extend to a tree the notion of stickiness defined by Hotz et al. (2013): a generalized Fréchet mean is either sticky, one-sided partly sticky or two-sided partly sticky.

Estimation of the population parameter is performed by minimizing an empirical objective function. The estimator exhibits radically different behaviours according to whether the unknown parameter is sticky or partly sticky. We present a sticky law of large numbers and central limit theorems. We also demonstrate how to determine from the data whether the location parameter is sticky or partly sticky. Finally, we present an extension of the classical nonparametric sign test to the tree, where the null hypothesis is equality of the Fréchet median with a specific point in the tree.

Modeling complex functional data

8:30 - 9:00

Estimation of Functional Principal Components from Sparse Functional Data

U. Mbaka[1], J. Cao[2], M. Carey[3]

[1]University College Dublin, Ireland
[2]Simon Fraser University, Canada
[3]UCD, Ireland

Sparse functional data are characterized by infrequent, irregular measurements, often observed with error, and require methods that accommodate these features. We propose a new method that combines basis expansion with maximum likelihood estimation to extract functional principal components from such data. Orthogonality of the estimated eigenfunctions is maintained during optimization using modified Gram–Schmidt orthonormalization. An information criterion is developed to select both the number of basis functions and the rank of the covariance structure. Principal component scores are estimated by conditional expectation, enabling accurate reconstruction of the underlying trajectories across the full domain despite sparse sampling. Simulation studies show that the proposed method performs favorably relative to existing approaches. Its practical utility is illustrated using CD4 cell count data from the Multicenter AIDS Cohort Study and somatic cell count data from Irish research dairy cattle.

Principal component analysis in Bayes spaces for sparsely sampled density functions

L. Steyer[1], S. Greven[1]

[1]Chair of Statistics, Humboldt-Universität zu Berlin, Germany

This paper presents a novel approach to functional principal component analysis (FPCA) in Bayes spaces in the setting where densities are the object of analysis, but only few individual samples from each density are observed. We use the observed data directly to account for all sources of uncertainty, instead of relying on prior estimation of the underlying densities in a two-step approach, which can be inaccurate if small or heterogeneous numbers of samples per density are available. To account for the constrained nature of densities, we base our approach on Bayes spaces, which extend the Aitchison geometry for compositional data to density functions. For modeling, we exploit the isometric isomorphism between the Bayes space and the L2 subspace L2_0 with integration-to-zero constraint through the centered log-ratio transformation. As only discrete draws from each density are observed, we treat the underlying functional densities as latent variables within a maximum likelihood framework and employ a Monte Carlo Expectation Maximization (MCEM) algorithm for model estimation. Resulting estimates are useful for exploratory analyses of density data, for dimension reduction in subsequent analyses, as well as for improved preprocessing of sparsely sampled density data compared to existing methods. The proposed method is applied to analyze the distribution of maximum daily temperatures in Berlin during the summer months for the last 70 years, as well as the distribution of rental prices in the districts of Munich.

Nonparametric Regression with Non-Stationary Penalization: An Application to Mobility Data

I. Di Battista[1], E. Arnone[2], L.M. Sangalli[1], P. Secchi[1]

[1]Politecnico di Milano, Italy
[2]University of Turin, Italy

This work presents a novel methodology for modeling anisotropy and non-stationarity in spatio-temporal phenomena. Many real-world processes across diverse fields, such as telecommunications and urban development, are inherently non-stationary and anisotropic, exhibiting varying statistical properties across both spatial and temporal domains. Traditional spatio-temporal models often assume stationarity, limiting their ability to capture such variability.
The proposed methodology introduces a nonparametric spatio-temporal regression model that incorporates a partial differential equation regularization to effectively model anisotropy and non-stationarity. This regularization term is expressed through a second-order linear differential operator, providing the flexibility to integrate problem-specific knowledge.
The utility of this model is demonstrated through its application to the Telecom Italia dataset, which consists of mobile phone data collected over a fine spatio-temporal grid. The analysis focuses on the metropolitan area of Milan, where spatial dynamics are strongly influenced by proximity to highways and roads, which serve as preferential signal directions, and temporal dynamics reflect daily and weekly human activity cycles. The resulting signal patterns are highly complex, displaying localized spatio-temporal behaviour driven by geography and urban mobility. This highlights the importance of accounting for non-stationarity when analyzing, for instance, mobility patterns.
Information on non-stationary anisotropy is mathematically encoded using diffusion tensors, positive definite matrices that provide a powerful representation of directional dependencies. However, working with tensors poses several mathematical challenges, such as preserving their positive definite property during operations like averaging or interpolating, as exemplified by modeling road intersections in the Telecom dataset. To address these challenges, we employ the Log-Euclidean metric to develop a mathematical framework that enables tensor operations while allowing the derivation of eigenvalue and eigenvector properties of tensors obtained through averaging.

Inferential Methods for Density Data in the Bayes Space: A Nonparametric Permutation Approach

A. Pini[1], S. Vantini[2], A. Menafoglio[2]

[1]Università Cattolica del Sacro Cuore, Italy
[2]Politecnico di Milano, Italy

This work presents inferential methods for null hypothesis significance testing of density data within the Bayes space framework. The Bayes space is a geometric framework that treats probability density functions as compositional objects, equipping the space of densities with a Hilbert space structure via the centered log-ratio transformation. This connection to Hilbert spaces naturally bridges Bayes space methodology with functional data analysis (FDA), enabling the adaptation of FDA tools to density-valued data while respecting their compositional geometry. Building on this foundation, we introduce nonparametric permutation-based procedures for one-sample and two-sample hypothesis tests, with extensions to analysis of variance and regression settings. The proposed methods guarantee finite-sample exactness without requiring distributional assumptions on the data, making them broadly applicable across scientific domains. The framework is illustrated through an analysis of fertility data, demonstrating its practical utility. These results contribute to the growing toolkit for the statistical analysis of distributional and compositional data, at the intersection of functional data analysis and compositional methods.

Resampling for Complex Data Problems

8:30 - 9:00

New explanations and inference for least angle regression

K. Gregory[1], D.J. Nordman[2]

[1]University of South Carolina, United States
[2]Iowa State University, United States

We present new results for least angle regression (LAR), introduced in Efron et al. (2004), which make possible clearer explanations of its strategy for admitting new variables in the construction of stepwise predictors and which allow an analysis of the behavior of the algorithm. Rather than treating LAR as a tool only for variable selection, we consider whether the path taken by LAR on data approximates a population-level LAR path and discover conditions under which the data-level path is consistent for its population-level counterpart, where the latter becomes our target for inference. We derive distributional results for LAR output, which determine variable entrances, leading to a stopping rule for LAR that has been missing in the literature. Notably, we do not encounter any condition like the irrepresentable condition appearing with Lasso; rather, we require only that on each step of LAR the variable newly contributing to the predictor must offer a contribution greater than that of all other not-yet-contributing variables by some margin. We formally establish a bootstrap for uncertainty quantification with LAR sample output and provide examples of practical LAR inference on real data sets along with simulation studies supporting our results.

A Modified t-test for Treatment Means in Unreplicated Classroom Comparisons

U. Genschel[1], D. Nordman[1]

[1]Iowa State University, United States

Discipline-based education research (DBER), with a focus on evidence-based teaching, has grown immensely over the last few decades. A common interest in DBER studies is identifying superior pedagogical approaches through rigorous, scientific methodology.

Researchers may have few classrooms available when comparing classroom-level treatments or conditions, so that one classroom per treatment is not uncommon in many DBER studies.

Because data and analysis options are then limited, an approach often seen in the DBER literature is to compare treatment means with a two-sample t-test applied to student-level responses from each classroom. This strategy, however, carries particular risks for statistical inference, where p-values can be misleading to an extent that is often under-appreciated and also much worse than possibly overstating practical significance.

We demonstrate that, even in the absence of any treatment difference, a mathematical guarantee exists that the p-value from a standard two-sample t-test applied to student-level responses in this setting can be made arbitrarily close to zero with probability 1, simply as an artifact of sufficient student enrollment.

Existing options to remedy the t-test, as we review, are typically intractable. As a more reasonable assessment of the evidence, we propose a modified two-sample t-test for comparing treatment means, which includes a smoothing step to account for classroom-level experimental error rather than ignoring it or possible correlations among student responses. Our numerical studies show that the modified t-test performs better than the standard t-test in controlling false rejection rates. The method is also illustrated with applications to several real data sets from educational studies.

Blockwise Empirical Likelihood for Spatial Regression Models

J.U. Soh[1], S. Bandyopadhyay[2], D. Nordman[3], C.Y. Lim[1]

[1]Seoul National University, South Korea
[2]Colorado School of Mines, United States
[3]Iowa State University, United States

In this presentation, we introduce theoretical properties of empirical likelihood methods for spatial regression models. Specifically, we discuss asymptotic results for regression parameters in both lattice and irregularly located settings. In the irregular-data case, the analysis is conducted under pure increasing domain and mixed increasing domain asymptotic frameworks. We further investigate a penalized extension and establish its oracle property under suitable regularity conditions. Simulation studies and a real data application illustrate the finite-sample performance and practical relevance of the theoretical findings.

A simple spectral goodness-of-fit test for general time series models

H. Yu[1], X. Han[2], C.Y. Yau[2]

[1]University of Rhode Island, United States
[2]The Chinese University of Hong Kong, Hong Kong

We revisit a test statistic originally proposed by Milhøj (1981) for assessing autoregressive and moving average (ARMA) models with independent innovations, and modify it to obtain a goodness-of-fit test for general time series models. Under the null hypothesis, the modified statistic asymptotically removes the contribution of fourth-order cumulant structures, which are typically the main obstacle to extending linear time series inference tools to general nonlinear processes. This removal obviates the need to estimate high-order terms, resulting in a tuning-parameter-free test that is computationally simple to implement. Importantly, the elimination of fourth-order effects is induced by the test's form and does not depend on the specific underlying model. The proposed test is therefore widely applicable, and its asymptotic null distribution is established under ?-weak dependence conditions that encompass both linear and nonlinear time series. Numerical studies demonstrate that the test not only achieves accurate type I error control but also attains power comparable to existing bootstrap-based and data-driven portmanteau tests.

Estimators and tests for nonparametric models

8:30 - 9:00

Goodness-of-Fit Testing for Point Processes in Large Populations

S.U. Can[1], R. Laeven[1], E. Khmaladze[2]

[1]University of Amsterdam, Netherlands
[2]Victoria University of Wellington, New Zealand

Suppose we are given an observed path from a population point process (e.g. deaths in a population) and we would like to test the goodness-of-fit of a particular parametric model for the conditional intensity of the event occurrences. In this paper, we propose a novel approach to conducting such goodness-of-fit tests. The idea is to consider the compensated point process, where the compensator is estimated parametrically, and to transform this process into a Poisson process compensated by its own estimated compensator. Then it is sufficient to know the asymptotic behavior of the latter process to test the goodness-of-fit of a wide class of parametric intensity models. We demonstrate the applicability of our approach through Monte Carlo simulations and data analyses.

Analysis of gradual changes in nonparametric regression based on a new optimization method in the non-unique case

N. Neumeyer[1], M. Hušková[2], L. Selk[1]

[1]University of Hamburg, Germany
[2]Charles University, Prague, Czech Republic

Consider a nonparametric regression model with one-dimensional covariate values and a continuous regression function. Assume that the regression function from the left of the covariate support starts equal to zero and then changes at some unknown point. Our aim is to estimate this gradual change point, which can be defined as the maximal value minimizing an objective function. Although one has uniform convergence of an empirical version of the objective function, using the maximal value minimizing the empirical version typically does not give a consistent estimator. We define a new general optimization method in this non-unique case, and apply it for the gradual change point estimation to obtain various consistent estimators. We discuss rates of convergence, asymptotic distribution and estimating the regression function based on the gradual change structure.

Efficient Density Estimation in an AR(1) Model

A. Schick[1]

[1]Binghamton University, United States

A class of plug-in estimators of the stationary density of an autoregressive model with autoregression parameter in the interval (0,1) is studied.
Two types of estimators of the innovation density are used,
a standard kernel estimator and a weighted kernel estimator with weights
chosen to mimic the condition that the innovation density has mean zero.
Bahadur expansions are obtained for this class of estimators in the space
of integrable functions. These stochastic expansions establish root-n
consistency in the L1 norm. It is shown that the density estimators
based on the weighted kernel estimators are asymptotically efficient if
an asymptotically efficient estimator of the autoregression parameter is used.

Adaptive exact recovery in sparse functional models

N. Stepanova[1], M. Turcicova[2]

[1]Carleton University, Canada
[2]Institute of Computer Science, Czech Republic

An unknown regression function f defined on the unit d-dimensional cube is observed in a Gaussian white noise model of small intensity. We assume that the function f is regular and that it is a sum of k-variate functions, where k varies from 1 to s for some integer s between 1 and d. These functions are unknown to us and only a few of them are nonzero, meaning that the function f is sparse. In this talk, we address the problem of identifying the nonzero components of f in the case when d tends to infinity as the noise intensity tends to zero and s is either fixed or tends to infinity but stays small relative to d. This may be viewed as a variable selection problem. We derive the conditions under which exact variable selection in the model at hand is possible and provide a selection procedure that achieves this type of selection. The procedure is adaptive to the degree of sparsity of the function f. We also derive the conditions that make exact variable selection in this model impossible. In view of these conditions, the proposed selector is asymptotically optimal.

Transfer Learning in Statistics

8:30 - 9:00

Transfer learning for piecewise-constant mean estimation

F. Wang[1], Y. Yu[2]

[1]University of Melbourne, Australia
[2]University of Warwick, United Kingdom

We study transfer learning for estimating piecewise-constant signals when source data, which may be relevant but disparate, are available in addition to target data. We first investigate transfer learning estimators that respectively employ
l_1 and l_0 penalties for unisource data scenarios and then generalize these estimators to accommodate multisources. To further reduce estimation errors, especially when some sources significantly differ from the target, we introduce an informative source selection algorithm. We then examine these estimators with multisource selection and establish their minimax optimality. Unlike the common narrative in the transfer learning literature that the performance is enhanced through large source sample sizes, our approaches leverage higher observational frequencies and accommodate diverse frequencies across multiple sources. Our extensive numerical experiments show that the proposed transfer learning estimators significantly improve estimation performance compared to estimators that only use the target data.

Minimax optimal transfer learning for high-dimensional additive regression

S.H. Moon[1], B.U. Park[1]

[1]Seoul National University, South Korea

This paper concerns transfer learning in additive regression where one observes a random sample from a target population together with auxiliary samples from different but potentially related regression models. We study
the minimax optimality of the local linear smooth backfitting estimator under a transfer learning framework in high-dimensional regimes. For this, we first derive minimax optimal error bounds under weak conditions for the baseline
procedure of estimating additive regression functions. We then introduce a novel transfer learning procedure and demonstrate that the transfer-learned estimator is minimax optimal under the transfer learning framework.We provide
numerical evidences that support the theoretical results in simulation studies and a real data analysis.

Domain Adaptation Targeting Heterogeneous and Imbalanced Subgroups

D. Zhou[1]

[1]National University of Singapore, Singapore

Domain adaptation enables generalizable and efficient data-driven research. However, existing work has largely focused on domain adaptation for some intrinsically homogeneous target cohort, overlooking inherent heterogeneity within the target, which can exacerbate biases and unfairness in the presence of subgroups with imbalanced sample sizes. We develop a novel domain adaptation framework that addresses more complicated target data that consists of heterogeneous and data-sparse subgroups and lacks gold-standard label observations. Our method simultaneously handles high-dimensionality, covariate shift, and outcome model heterogeneity by combining a model-assisted debiasing step used for covariate shift correction with an adaptive knowledge-guided sparsification procedure used to mitigate the issue of sample disparity. We also introduce a new model selection strategy to avoid negative knowledge transfer in the absence of labels in target data. Our method is theoretically justified to be robust to nuisance model misspecification and adaptive to heterogeneity between the subgroups. Numerical experiments and two real-world applications---genetic risk modeling of type II diabetes and prediction of mutation-induced protein stability changes---demonstrate our method's practical advantages.

Adaptive Transfer Learning for High-Dimensional Cox Models: A Two-Step Procedure with Source Screening Optimal Rates, and Valid Inference

E.R. Lee[1]

[1]Sungkyunkwan University, South Korea

We propose Trans-Cox, a transfer learning procedure for high-dimensional Cox proportional hazards models in which a small target cohort is augmented by multiple heterogeneous source cohorts. Unlike i.i.d. loss settings, the Cox partial likelihood couples individuals through random risk sets whose composition varies across cohorts, complicating both source screening and inference. The proposed method comprises three components: (1) a pooled lasso–Cox fit via stratified partial likelihood; (2) a target-only offset correction for the pooled residual bias; and (3) risk-set–aware source screening using within-fold validation gaps, which avoids combinatorial search and mitigates negative transfer. Under sub-Gaussian covariates and martingale score concentration, we establish nonasymptotic oracle inequalities that separate pooled stochastic error from residual transfer bias. We prove detection consistency for the screening procedure and develop cross-fitted debiased inference that yields valid confidence intervals after data-driven source selection. When no source is beneficial, the method automatically recovers the target-only rate. Applications to SEER breast cancer data demonstrate improved predictive performance relative to the target-only approach, and an analysis of FLChain survival data confirms that the method avoids the severe negative transfer incurred by naive pooling.

Contributed: Geometry, Manifolds and Object Data

8:30 - 8:50

A minimax manifold estimator for hypersurfaces

H. González-Vázquez[1], B. Pateiro-López[1], A. Rodríguez-Casal[1]

[1]Universidade de Santiago de Compostela - CITMAga, Spain

The manifold hypothesis posits that high-dimensional real data usually exhibit a lower-dimensional structure; in particular, they are supported on or close to a lower-dimensional submanifold embedded in the ambient Euclidean space. The goal of manifold estimation is to estimate the unknown underlying manifold based on the sample of points. This work proposes a new manifold estimator in a tubular noise model. Our estimator builds upon the Euclidean Distance Transform (EDT) estimator, originally proposed by Genovese et al. (2012), and generalizes it by allowing for the use of any pilot estimator of the support, not just the Devroye-Wise estimator employed in the original EDT estimator. Moreover, we propose to estimate the support using the r-convex hull of the sample. This estimator is related to a shape restriction, called r-convexity, that generalizes convexity and provides greater flexibility. Our manifold estimator achieves the minimax rate of convergence in Hausdorff distance (up to logarithmic factor) to estimate hypersurfaces, that is, manifolds whose dimension is one unit less than that of the ambient space. Finally, we present a real data application regarding the estimation of filaments in the two-dimensional plane.

Adaptive thresholding for heteroskedastic variance estimation on the sphere

C. Durastanti[1], R. Shevchenko[2]

[1]Sapienza Università di Roma, Italy
[2]J. A. Dieudonné Laboratory - Côte d'Azur University, and Centrale Méditerranée, Nice., France

This talk discusses the nonparametric estimation of a heteroskedastic variance function on the sphere within a regression framework, assuming that the variance belongs to a Besov regularity class. Problems of this type naturally arise in the analysis of spherical data, where the variability of the observations may change across the domain and needs to be estimated without imposing restrictive parametric assumptions.
To address this problem, we propose a needlet-based estimator that combines multiresolution analysis with a hard-thresholding procedure. Needlets form a class of spherical wavelets with remarkable spatial and spectral localization properties, which make them particularly suitable for statistical inference on the sphere. By exploiting these localization features, the proposed method is able to capture local irregularities of the variance function while maintaining good global approximation properties.
The estimator is constructed through a needlet expansion of the target function and a suitable thresholding rule applied to the empirical coefficients. This approach allows the procedure to automatically adapt to the unknown smoothness of the variance function. We show that the resulting estimator achieves minimax-optimal convergence rates over Besov spaces under standard loss functions.

Metric Skewness for Object Data

V. Zamanifarizhandi[1], J. Kujala[1], O. Rainio[1], J. Virta[1]

[1]University of Turku, Finland

As datasets become increasingly complex, more advanced methods for extracting insight from data are required. Recently, analyzing data in metric spaces has gained attention and several descriptive statistics have been developed. However, important distributional characteristics, such as skewness, which provides valuable information about its structure, remain largely unexplored for object data in metric spaces. In this work a novel method is introduced to compute the metric skewness. Moreover, its use in testing for the presence of skewness exhibits promising outcome both in level and power, surpassing its multivariate skewness counterparts in Euclidean space. Finally, the proposed metric skewness is applied as an inferential statistic to positron emission tomography (PET) data, illustrating its practical applicability.

{Multivariate Quantile-Based Permutation Tests with Application to Functional Data

D. Hlubinka[1], Š. Hudecová[1], Z. Hlávka[1]

[1]Univerzita Karlova, Czech Republic

Permutation tests enable testing statistical hypotheses in situations when the distribution of the test statistic is complicated or not available. In some situations, the test statistic under investigation is multivariate, with the multiple testing problem being an important example. The corresponding multivariate permutation tests are then typically based on a suitable one-dimensional transformation of the vector of partial permutation p-values via so called combining functions. This contribution describes a new approach that uses the discrete optimal measure transportation concept. The final single p-value is computed from the empirical center-outward distribution function of the permuted multivariate test statistics. This method avoids computation of the partial p-values and it is easy to be implemented. In addition, it allows to compute and interpret contributions of the components of the multivariate test statistic to the overall non-conformity score and to the rejection of the null hypothesis. Apart from this method, the measure transportation is applied also to the vector of partial p-values as an alternative to the classical combining functions.

Keynote Talk

11:00 - 12:00

When Censoring Is Not Innocent: Challenges and Solutions for Dependent Censoring in Survival Analysis

I. Van Keilegom[1]

[1]KU Leuven, Belgium

Survival analysis relies on a deceptively simple assumption: censoring is independent of the event of interest. In practice, however, this assumption is often violated. Patients may drop out because their health deteriorates, treatments may be discontinued in response to emerging risk, and follow-up may depend on factors that are themselves related to survival. Such dependent censoring can introduce substantial bias, undermine the validity of standard estimators, and lead to misleading scientific conclusions.

In this talk, I will explore why dependent censoring is one of the most challenging—and frequently overlooked—threats to reliable time-to-event analysis. Using examples from clinical and observational studies, I will illustrate how dependent censoring arises, how it distorts inference, and why conventional survival methods can fail.

I will then discuss methodological approaches for addressing dependent censoring, with particular emphasis on copula-based models that explicitly characterize the dependence between survival and censoring times. These models provide a flexible framework for sensitivity analysis and for quantifying the impact of departures from the independent censoring assumption. In addition, I will present recent developments on the partial identification of survival quantities under dependent censoring, as well as new results on identifiability and verifiability in enriched data settings. Dependent censoring will serve as a motivating example of a broader question at the heart of statistical inference: what can—and cannot—be learned from incomplete data?

Partitioning Methods

13:30 - 14:00

Accuracy Limits of Causal Trees

M. Cattaneo[1], J. Klusowski[1], R. Yu[1]

[1]Princeton University, United States

Recursive decision trees are widely used to estimate heterogeneous causal treatment effects in experimental and observational studies. These methods are typically implemented using CART-type recursive partitioning and are often viewed as adaptive procedures capable of discovering treatment effect heterogeneity in high-dimensional settings. We study causal tree estimators based on adaptive recursive partitioning and establish lower bounds on their estimation accuracy. We show that causal trees constructed via standard CART-type splitting rules cannot achieve polynomial-in-n convergence rates in the uniform norm (where n denotes the sample size) under basic conditions, even when the true treatment effect is constant across the covariate space.

Inferring Treatment Effects in Large Panels by Uncovering Latent Similarities

B. Deaner[1], A. Zeleneev[1], C. Hsiang[1]

[1]UCL, United Kingdom

The presence of unobserved confounders is one of the main challenges in identifying treatment effects. In this paper, we propose a new approach to causal inference using panel data with large N and T. Our approach imputes the untreated potential outcomes for treated units using the outcomes for untreated individuals with similar values of the latent confounders. In order to find units with similar latent characteristics, we utilize long pre-treatment histories of the outcomes. Our analysis is based on a nonparametric, nonlinear, and nonseparable factor model for untreated potential outcomes and treatments. The model satisfies minimal smoothness requirements. We impute both missing counterfactual outcomes and propensity scores using kernel smoothing based on the constructed measure of latent similarity between units, and demonstrate that our estimates can achieve the optimal nonparametric rate of convergence up to log terms. Using these estimates, we construct a doubly robust estimator of the period-specifc average treatment effect on the treated (ATT), and provide conditions, under which this estimator is $\sqrt{N}$-consistent, and asymptotically normal and unbiased. Our simulation study demonstrates that our method provides accurate inference for a wide range of data generating processes.

Unknown Group Structures in Econometric Models

D.(. Kang[1], V. Marmer[2], J. Catalano[3], P. Schrimpf[2]

[1]Xiamen University, China
[2]University of British Columbia, Canada
[3]Charles River Associates, Canada

We consider a regression model in which coefficients vary across unknown groups determined by an unknown partition of the covariate space. We propose a new estimation procedure that builds on the regression tree algorithm and consistently recovers both the group partition and group-specific coefficients. Our innovations include a joining step to mitigate over-fitting by consolidating spurious splits, and a local version of the splitting algorithm to address highly correlated partitioning variables. We show that the procedure can achieve the oracle property.

The Blessings of Overparameterization: Applications in Solving Economic Models

M. Ebrahimi Kahou[1], J. Fernández-Villaverde[2]

[1]Bowdoin College, United States
[2]University of Pennsylvania, United States

We investigate the effects of overparameterization when using neural networks to solve dynamic programming problems in economics. As the number of parameters in a neural network grows, test loss on Bellman and Euler residuals decreases, and the resulting approximate policy and value functions converge toward their benchmark counterparts. A central finding is that overparameterization dramatically improves algorithmic stability: with small networks, solutions across different random initializations can disagree substantially, but as network width increases, this disagreement collapses and solutions concentrate tightly around the truth. We demonstrate these properties across three canonical economic models with known benchmark solutions: the McCall job-search model, the Linear-Quadratic Regulator, and the Real Business Cycle model. Our results suggest that practitioners solving economic models with neural networks should not fear large networks, as overparameterization is a blessing rather than a curse.

Learning in complex data settings

13:30 - 14:00

Adaptive transfer learning from multiple heterogeneous sources

S. Maity[1]

[1]University of Waterloo, Canada

Exploiting information from related source domains to enhance performance in a target domain is a central objective of transfer learning — an objective that has broad practical impact and has attracted considerable research attention in recent years. However, handling multiple heterogeneous source domains remains a complex challenge. In transfer learning, the relevance of a source domain is usually quantified by its bias relative to the target domain. Since these biases are rarely known in practice, the existing literature heavily relies on assuming specific, restrictive structures for admissible biases to develop adaptive methodologies. By contrast, our research explores the theoretical limits of adaptation under more general conditions. Within the context of parameter estimation, we demonstrate a striking fundamental limit: adaptation is impossible under the most general class of bias structures if there are at least two source domains. To resolve this, we propose two new structural paradigms: order-based and separation-based biases, under which adaptation becomes feasible, and we construct the corresponding estimators to achieve it. Ultimately, our separation-based framework offers a strictly more general approach than commonly adopted bias structures, significantly broadening the theoretical foundations of adaptive transfer learning from multiple sources.

On regression with estimated covariates and conditional effects given the propensity score

M. Bonvini[1], J. Wu[1], E. Kennedy[2], J. Brand[3], Y. Xie[4]

[1]Rutgers University, United States
[2]Carnegie Mellon University, United States
[3]University of California, Los Angeles, United States
[4]Princeton University, United States

Motivated by the study of heterogeneous returns to education in Brand and Xie (2010), which considers how the effect of completing college on earnings varies with the (unknown) probability of completing college, we analyze the problem of estimating a nonparametric regression function when certain covariates are estimated in a first step. Plug-in estimators that treat the estimated covariates as known generally suffer from first-stage estimation error. To mitigate this issue, we analyze two debiasing approaches within a framework that is agnostic to the choice of the first-stage estimation method and consider both local- and sieves-based methods for the second-stage regression. The methods considered are: (i) influence function-based estimators of pathwise differentiable parameters that approximate the target estimand, (ii) a variant of plug-in estimators that directly aim to correct the bias. For each method, we upper bound the estimation risk and characterize conditions under which oracle rates can be approached, highlighting the possible gains in terms of convergence rates relative to the plug-in. Simulation results corroborate our theoretical findings. We apply our methodology to data from the National Longitudinal Survey of Youth 1997 and find evidence that completing college yields the largest benefits for individuals least likely to do so, consistent with earlier findings in the literature (Brand and Xie, 2010; Brand, 2019).

Beyond euclidean domains: nonparametric inference for point processes on linear networks

W. González Manteiga[1], M.I. Borrajo-García[1]

[1]University of Santiago de Compostela, Spain

Point processes defined on linear networks arise in many scientific contexts where events occur along one-dimensional structures embedded in planar regions, such as traffic accidents on road systems, crimes on urban street networks or neuronal activity along dendritic trees. The geometry of these domains poses specific methodological challenges because classical spatial techniques developed for Euclidean spaces are not directly applicable.
This talk focuses on nonparametric inference for point processes constrained to linear networks, with particular emphasis on the estimation and testing of the first-order intensity function and related characteristics. Nonparametric methods offer a flexible framework that avoids restrictive structural assumptions and lets the data reveal their spatial structure along the network.
For applications, we analyse real-world datasets in which events are naturally defined on network structures, using accident data as a primary motivating example to illustrate the methodology. We also discuss how these techniques can be extended to other types of linear structures, such as river systems, highlighting the versatility of the framework across different applied settings.
Overall, this work advances nonparametric inference for point processes beyond the classical planar setting by providing a coherent framework for studying event patterns on linear networks. The approach combines theoretical insight with practical applicability, laying the groundwork for further methodological developments and applications in diverse domains

A Framework for Multivariate Goodness-of-Fit Testing Based on Matrix Distances

M. Markatou[1]

[1]University at Buffalo, United States

Measures of discrepancy between two probability distributions are used in the scientific literature to develop goodness-of-fit methods. In this talk, we will discuss a unified framework for the study of two-sample and k-sample goodness-of-fit testing that is based on the concept of matrix distance, which we define first. We then use its elements to construct test statistics for testing equality (in distribution) of k-samples. We show that the two-sample test statistic is a special case of the k-sample test statistic. Furthermore, we derive the asymptotic distributions of the tests under the null hypothesis and illustrate the connection of the MMD statistic with the proposed tests. We then present computational considerations and illustrate the implementation of these methods via the QuadratiK software, Simulation results exemplify the performance of the new methods and compare it with that of the state-of-the-art procedures.

Biometrika session

13:30 - 14:00

Boosting the power of kernel two-sample tests

A. Chatterjee[1], B. Bhattacharya[2]

[1]University of Chicago, United States
[2]University of Pennsylvania, United States

Maximum Mean Discrepancy (MMD) is a widely used kernel method for two-sample testing in complex data. This talk presents a principled approach to improve its power by aggregating MMD statistics across multiple kernels using a Mahalanobis distance. The resulting test is universally consistent and achieves strong finite-sample performance by adapting to a broad range of alternatives. We will outline the asymptotic theory of the method, including the null distribution, behavior under fixed and local alternatives, and non-trivial Pitman efficiency. I will also discuss consistency when the number of kernels grows with sample size and show how a multiplier bootstrap enables efficient computation. Empirical results on synthetic and real datasets demonstrate clear gains over single-kernel tests. The analysis relies on a general framework for the joint distribution of MMD estimators via multiple stochastic integrals, with broader implications for adaptive and linear-time multi-kernel testing.

Bayesian clustering of high-dimensional data via latent repulsive mixtures

A. Guglielmi[1], L. Ghilotti[2], M. Beraha[3]

[1]Politecnico di Milano, Italy
[2]Duke University, United States
[3]University of Milan-Bicocca, Italy

Model-based clustering of moderate- or large-dimensional data is notoriously difficult. We propose a model for simultaneous dimensionality reduction and clustering by assuming a mixture model for a set of latent scores, which are then linked to the observations via a Gaussian latent factor model. This approach was recently investigated by Chandra et al. (2023). The authors used a factor-analytic representation and assumed a mixture model for the latent factors. However, performance can deteriorate in the presence of model misspecification. Assuming a repulsive point process prior for the component-specific means of the mixture for the latent scores is shown to yield a more robust model that outperforms the standard mixture model for the latent factors in several simulated scenarios. The repulsive point process must be anisotropic to favour well-separated clusters of data, and its density should be tractable for efficient posterior inference. We address these issues by proposing a general construction for anisotropic determinantal point processes. We illustrate our model in simulations, as well as a plant species co-occurrence dataset.

Noise-Induced Randomization in Regression Discontinuity Designs

N. Ignatiadis[1]

[1]University of Chicago, United States

Regression discontinuity designs assess causal effects in settings where treatment is determined by whether an observed running variable crosses a pre-specified threshold. Here we propose a new approach to identification, estimation, and inference in regression discontinuity designs that uses knowledge about exogenous noise (e.g., measurement error) in the running variable. In our strategy, we weight treated and control units to balance a latent variable of which the running variable is a noisy measure. Our approach is driven by effective randomization provided by the noise in the running variable, and complements standard formal analyses that appeal to continuity arguments while ignoring the stochastic nature of the assignment mechanism.

Dynamic Factor Analysis of High-dimensional Recurrent Events

K. Zhou[1]

[1]Columbia University, United States

Recurrent event time data arise in many studies, including biomedicine, public health, marketing, and social media analysis. High-dimensional recurrent event data involving many event types and observations have become prevalent with advances in information technology. This paper proposes a semiparametric dynamic factor model for the dimension reduction of high-dimensional recurrent event data. The proposed model imposes a low-dimensional structure on the mean intensity functions of the event types while allowing for dependencies. A nearly rate-optimal smoothing-based estimator is proposed. An information criterion that consistently selects the number of factors is also developed. Simulation studies demonstrate the effectiveness of these inference tools. The proposed method is applied to grocery shopping data, for which an interpretable factor structure is obtained.

ISNPS Student Paper Competition

13:30 - 14:00

Semiparametric KSD test: unifying score and distance-based approaches for goodness-of-fit testing

Z. Huang[1], Z. Niu[1]

[1]University of Pennsylvania, United States

Goodness-of-fit (GoF) tests are fundamental tools for assessing model adequacy. Score-based GoF tests are particularly appealing because they require fitting the model only once under the null. However, extending these tests to powerful nonparametric settings is challenging, mainly due to the lack of suitable scores. Through a class of exponentially tilted models, we show that the resulting score-based GoF tests are equivalent to the tests based on integral probability metrics (IPMs) indexed by a function class. When the class is rich, the test is universally consistent. This simple yet insightful perspective enables reinterpretation of classical distance-based testing procedures—including those based on Kolmogorov–Smirnov distance, Wasserstein-1 distance, and maximum mean discrepancy—as arising from score-based constructions. Building on this insight, we propose a new nonparametric score-based GoF test through a special class of IPM induced by kernelized Stein function class, called semiparametric kernelized Stein discrepancy (SKSD) test. Compared with other nonparametric score-based tests, the SKSD test is computationally efficient and accommodates general nuisance-parameter estimators, supported by a generic parametric bootstrap procedure. The SKSD test is universally consistent and attains Pitman efficiency. Moreover, SKSD test provides simple GoF tests for models with intractable likelihoods but tractable score functions with the help of Stein's identity. We apply the SKSD testing framework to two widely used models of this type to demonstrate the power of our method. Our method also achieves power comparable to task-specific normality tests, such as the Anderson-Darling and Lilliefors tests, despite being designed for general goodness-of-fit problems.

Graphical Pitman-Yor Process for Clustering in Bayesian Networks

I. Golovko[1], A. Lijoi[1], I. Prünster[1]

[1]Bocconi University, Italy

Hierarchical Bayesian nonparametric models have been highly successful for analysing grouped data, yet most available methods treat groups as exchangeable. We address settings where groups are related but not symmetric by representing their dependence with a directed acyclic graph (DAG). Building on the graphical Dirichlet process, we introduce the graphical Pitman–Yor (GPY) process, which adds a discount parameter that induces heavy-tailed cluster-size behaviour and delivers two practical advantages: more realistic power-law clustering in network settings and adaptive information pooling along the DAG. We characterise the predictive distributions and derive tractable conditional distributions that yield an exact marginal Gibbs sampler on DAGs, enabling efficient inference without truncation. Simulation studies show consistent gains over the Dirichlet special case and standard baselines, particularly when cluster sizes follow power-law patterns and when the DAG captures meaning- ful relations among groups. The approach applies broadly to network-structured problems.

Nonparametric inference for ratios of densities via uniformly valid and powerful permutation tests

A. Bordino[1], T. Berrett[1]

[1]University of Warwick, United Kingdom

We propose the density ratio permutation test, a hypothesis test that assesses whether the ratio between two densities is proportional to a known function based on independent samples from each distribution. The test uses an efficient Markov Chain Monte Carlo scheme to draw weighted permutations of the pooled data, yielding exchangeable samples and finite sample validity. For power, if the statistic is an integral probability metric, our procedure is consistent under mild assumptions on the defining function class; specializing to a reproducing kernel Hilbert space, we introduce the shifted maximum mean discrepancy and prove minimax optimality of our test when a normalized difference between the densities lies in a Sobolev ball. We extend to the case of an unknown density ratio by estimating it on an independent training sample and derive type~I error bounds in terms of the estimation error as well as power results. This allows adapting our method to conditional two sample testing, making it a versatile tool for assessing covariate-shift and related assumptions, which frequently arise in transfer learning and causal inference. Finally, we validate our theoretical findings through experiments on both simulated and real-world datasets.

New nonparametric inference & modeling on network data

13:30 - 14:00

Dynamic Factor Model for Poisson Time Series in Matrix and Tensor Form

R. Chen[1]

[1]Rutgers University, United States

In many applications, we often encounter integer-valued time series observations in the form of tensors (multi-dimensional array). Motivated by the applications in geo-political event prediction, crime data analysis and transportation and trading networks modeling, in this paper we study the modeling, interpretation and prediction for Poisson tensor time series through dynamic factor models. The approach contains an observation layer specified for the Poisson distribution, and a latent layer to account for the dynamic and concurrent dependence. The dynamics is introduced through the factor structure with tensor Tucker decomposition. We propose an autocovariance-based approach to estimate the loading matrices and vectors, which takes advantage of the dependence to reduce the noise level. The estimation of the factors is through an efficient computational method for Poisson Lon-Normal regression model. An autoregressive modeling of the factors is considered to enable predictions. Theoretical investigations and numerical experiments are presented.

Statistical Inference for Large Potts Models

F. Yang[1], Y. Lin[2], W. Zhou[3], Z. Ren[1]

[1]University of Pittsburgh, United States
[2]Chongqing Normal University, China
[3]New York University, United States

The Potts model is a fundamental graphical model for multistate categorical data, yet statistical inference for edge interactions, represented by matrix-valued parameters, in high-dimensional settings remains largely unexplored. This article develops a likelihood-based inference framework for high-dimensional Potts models. Building on penalized node-wise multinomial regression, we propose a block-wise debiasing procedure that constructs one-step corrected estimators via an orthogonalized score matrix derived from the likelihood. The resulting estimators are shown to be asymptotically normal, enabling valid chi-square tests with asymptotic control of type I error and nontrivial power against local alternatives. To facilitate large-scale inference, we further develop an FDR-controlled multiple testing procedure and a simultaneous testing method for collections of interaction blocks. The multiple testing procedure is shown to asymptotically control the FDR, while the simultaneous test is calibrated via a bootstrap approximation for maxima of dependent quadratic forms, supported by Gaussian approximation theory. Our analysis introduces new technical tools, including a quadratic-to-quadratic comparison result and anti-concentration bounds for maxima of chi-square-type statistics. The effectiveness of the proposed methods is demonstrated through simulation studies.

Statistical Inference for Latent Space Models of Network Data with Edge Covariates

J. Zhu[1]

[1]University of Michigan, United States

Latent space models (LSMs) provide a powerful framework for analyzing network data by embedding nodes in a latent space. Incorporating covariate information via edge covariates offers an important generalization that strengthens both the interpretability and practical utility of the model. However, we show that coefficient estimates for edge covariate effects obtained through maximum likelihood estimation exhibit asymptotic bias due to high-order geometric effects and errors in latent variable estimation. To address this issue, we propose a plug-in bias-correction estimator that enables asymptotically valid and unbiased statistical inference for the effects of edge covariates. We establish theoretical guaranties, including consistency and asymptotic normality, under various network structures. Extensive simulations and real-world data examples demonstrate that our method effectively reduces estimation bias and improves the accuracy of inference. Our findings contribute to the statistical methodology of LSMs by providing a principled framework for unbiased parameter estimation in network models with edge covariates.

Frontiers in Nonparametric Learning: From Classical Regression to Complex High-Dimensional Models

13:30 - 14:00

Analysis of Variance of Tensor Product Reproducing Kernel Hilbert Spaces on Metric Spaces

Y. Wang[1]

[1]University of California - Santa Barbara, United States

Many methods have been developed to analyze complex data, such as non-Euclidean shape, network, and manifold data. However, there is a lack of methods for studying interactions among complex data. In this paper, we first propose a novel kernel function for a metric space and construct its associated reproducing kernel Hilbert space. The new nonstationary kernel function provides a flexible and powerful tool for learning complex structures in non-Euclidean data. We then construct an analysis of variance (ANOVA) decomposition of the nonparametric regression function defined on metric space, which provides a hierarchical structure for investigating the main effects and interactions. We develop estimation and computational methods for a semi-parametric model with a multivariate function on a product of metric spaces modeled by the ANOVA decomposition. We establish the convergence rates of parameter and nonparametric function estimates. The application of the proposed methods to the Alzheimer's Disease Neuroimaging Initiative hippocampus shape data confirms some existing and suggests some new interactions among hippocampal regions. Simulations indicate that the proposed methods work well.

Reluctant Interaction Inference after Additive Modeling

G. Yu[1]

[1]University of California Santa Barbara, United States

Additive models enjoy the flexibility of nonlinear models while still being readily understandable to humans. By contrast, other nonlinear models, which involve interactions between features, are not only harder to fit but also substantially more complicated to explain. Guided by the principle of parsimony, a data analyst therefore may naturally be reluctant to move beyond an additive model unless it is truly warranted. To put this principle of interaction reluctance into practice, we formulate the problem as a hypothesis test with a fitted sparse additive model (SPAM) serving as the null. Because our hypotheses on interaction effects are formed after fitting a SPAM to the data, we adopt a selective inference approach to construct p-values that properly account for this data adaptivity. Our approach makes use of external randomization to obtain the distribution of test statistics conditional on the SPAM fit, allowing us to derive valid p-values, corrected for the over-optimism introduced by the data-adaptive process prior to the test. Through experiments on simulated and real data, we illustrate that—even with small amounts of external randomization—this rigorous modeling approach enjoys considerable advantages over naive methods and data splitting.

Deep Generative Models: Complexity, Dimensionality, and Approximation

K. Wang[1], H. Niu[1], D. Li[1]

[1]UNC CH, United States

Generative networks have shown remarkable success in learning complex data distributions, particularly in generating high-dimensional data from lower-dimensional inputs. While this capability is well-documented empirically, its theoretical underpinning remains unclear. One common theoretical explanation appeals to the widely accepted manifold hypothesis, which suggests that many real-world datasets, such as images and signals, often possess intrinsic low-dimensional geometric structures. Under this manifold hypothesis, it is widely believed that to approximate a distribution on a d-dimensional Riemannian manifold, the latent dimension needs to be at least d or $d+1$. In this work, we show that this requirement on the latent dimension is not necessary by demonstrating that generative networks can approximate distributions on d-dimensional Riemannian manifolds from inputs of any arbitrary dimension, even lower than d, taking inspiration from the concept of space-filling curves. This approach, in turn, leads to a super-exponential complexity bound of the deep neural networks through expanded neurons. Our findings thus challenge the conventional belief on the relationship between input dimensionality and the ability of generative networks to model data distributions. This novel insight not only corroborates the practical effectiveness of generative networks in handling complex data structures, but also underscores a critical trade-off between approximation error, dimensionality, and model complexity.

Optimal-k Sequence for Difference-based Methods in Nonparametric Regression

T. Tong[1]

[1]Hong Kong Baptist University, Hong Kong

Difference-based methods have been attracting increasing attention in nonparametric regression, in particular for estimating the residual variance. To implement the estimation, one needs to choose an appropriate difference sequence, mainly between the optimal difference sequence and the ordinary difference sequence. This difference sequence selection is a fundamental problem in nonparametric regression, and it remains unresolved until recently. In this paper, we propose to further advance the difference sequence selection from another unique perspective, which creates a new family of difference sequence called the optimal-k sequence. Our proposed difference sequence not only provides a better bias-variance trade-off, but also includes the optimal and the ordinary difference sequences as two important special cases. Through theoretical and numerical studies, we demonstrate that the optimal-k sequence has been pushing the boundaries of our knowledge in difference-based methods in nonparametric regression, and more importantly, it always performs the best in practical situations.

Topics in Econometrics I

13:30 - 14:00

Bayesian Double Machine Learning for Causal Inference

L. Liu[1], F. DiTraglia[2]

[1]University of Pittsburgh, United States
[2]University of Oxford, United Kingdom

This paper proposes a simple, novel, and fully-Bayesian approach for causal inference in partially linear models with high-dimensional control variables. Off-the-shelf machine learning methods can introduce biases in the causal parameter known as regularization-induced confounding. To address this, we propose a Bayesian Double Machine Learning (BDML) method, which modifies a standard Bayesian multivariate regression model and recovers the causal effect of interest from the reduced-form covariance matrix. Our BDML is related to the burgeoning frequentist literature on DML while addressing its limitations in finite-sample inference. Moreover, the BDML is based on a fully generative probability model in the DML context, adhering to the likelihood principle. We show that in high dimensional setups the naive estimator implicitly assumes no selection on observables--unlike our BDML. The BDML exhibits lower asymptotic bias and achieves asymptotic normality and semiparametric efficiency as established by a Bernstein-von Mises theorem, thereby ensuring robustness to misspecification. In simulations, our BDML achieves lower RMSE, better frequentist coverage, and shorter confidence interval width than alternatives from the literature, both Bayesian and frequentist.

Testing the Solvability of Systems of Linear Inequalities

E. Mbakop[1], L. Goff[2]

[1]OSU, United States
[2]U Calgary, Canada

This paper studies the problem of testing whether a system of linear equality and inequality constraints admits a solution when the coefficients may have to be estimated. We show that a wide range of inferential questions in partially identified models can be formulated as hypotheses of this form. Our approach exploits an alternative characterization of the hypothesis based on whether the value of a certain linear program is equal to zero. Building on this characterization, we develop bootstrap-based testing procedures and establish their uniform validity over large classes of data-generating processes. Simulation results demonstrate good finite-sample performance, even for moderate sample sizes. We illustrate the usefulness of the approach in two empirical applications.

Shape-aware deep learning for models of production

A. Prokhorov[1], Z. Wei[2], H. Sang[3], Y. Ma[3]

[1]U Sydney, Australia
[2]TAMU, United States
[3]tamu, United States

The stochastic frontier model (SFM) is widely employed in the analysis of productivity and efficiency, yet strict parametric forms, such as the Cobb-Douglas and Translog functions, are often assumed for modeling production, leading to potential misspecification issues. While semi- and nonparametric SFMs offer greater flexibility, they face challenges in imposing monotonicity and concavity to maintain their desirable economic interpretation. We develop a framework which enforces the shape restrictions within deep neural networks (DNNs). The stochastic frontier model we develop (DNN-SFM) leverages the flexibility and predictive power of DNNs while preserving key properties of a production function, such as free disposability and diminishing marginal product. Additionally, we demonstrate how to use Shapley values to measure and interpret global and local effects of individual inputs on the production frontier in cases when model parameters to not admit a simple interpretation. The performance of the proposed method is assessed using simulations while a real-world application to rice production in the Philippines illustrates empirical relevance of the proposed method.

Identification and Estimation of Translation Invariant Panel Data Models

S. Khan[1], E. Tamer[2]

[1]Boston College, United States
[2]Harvard University, United States

This paper considers identification and estimation of a class of models with additive fixed effects. Our main result is that for models which satisfy a translation invariance property, parameters of interest can be identified for longitudinal data, which character- izes most panel data sets in empirical microeconometric settings. Our leading example is the quantile regression fixed effect model introduced in Koenker (2004). We show that under stated conditions our proposed procedure, which involves optimization of a convex objective function, can consistently estimate the regression coefficients in a model with only a finite number of time periods, for any quantile. This is in contrast to Koenker (2004) and subsequent quantile regression panel data papers which generally assume the number of time periods to grow with the sample size, inferring an inciden- tal parameters problem. To allow for quantile specific unobserved heterogeneity in our setting, we impose a factor structure and consistently estimate the quantile specific factor load. Simulation studies indicate adequate finite sample properties using our procedure. Extensions to other panel data models, notably with discrete outcomes are also proposed, where novel set identification results are obtained.

Modern Methods and Applications in Robust Nonparametric Statistics

13:30 - 14:00

Empirical Likelihood and Density Ratio Models: A Framework for Modern Forestry Data Analysis

J. Chen[1]

[1]the University of British Columbia, Canada

The strength and quality of lumber evolve over time under the combined influence of climate change, forest disturbances, and industrial practices. To ensure that wood products continue to meet safety standards, the forest products industry conducts ongoing monitoring based on data collected from multiple mills, grades, and years.
These data sets are typically modest in size, clustered in structure, and only partially comparable across time, posing unique challenges for reliable statistical inference.

This talk presents a unified semiparametric framework for analyzing such data using the \emph{density ratio model} (DRM) in combination with \emph{empirical likelihood} (EL). The DRM links multiple related populations through exponential tilts, allowing efficient information pooling while retaining nonparametric flexibility. Within this framework, we develop inference procedures based on \emph{dual empirical likelihood}, \emph{composite empirical likelihood} for clustered samples, and \emph{permutation methods} for complex rotating sampling plans. These approaches yield valid and efficient tests and confidence regions for means, percentiles, and distributional changes, even under model misspecification or dependence within clusters.

Illustrated by examples from long-term lumber-strength monitoring programs, the methods demonstrate how modern semiparametric tools can enhance industrial data analysis. The presentation emphasizes both the theoretical development and its practical impact through collaboration between UBC statisticians and the forest products industry.

Independent Component Analysis by Robust Distance Correlation

P. Rousseeuw[1]

[1]KU Leuven, Belgium

This is joint work with Sarah Leyder, Jakob Raymaekers, Tom Van Deuren, and Tim Verdonck. Independent component analysis (ICA) is a powerful tool that attempts to decompose a multivariate signal or distribution into fully independent sources, not just uncorrelated ones like PCA does. ICA is harder to do, but it has many important applications. Unfortunately, most approaches to ICA are not robust against outliers. Here we propose a robust ICA method called PICARD, which estimates the components by minimizing a robust measure of dependence between multivariate random variables. The dependence measure used is the distance correlation (dCor). In order to make it more robust we first apply a new transformation called the bowl transform, which is bounded, continuous, injective, and maps far outliers to points close to the origin. This preserves the crucial property that a zero dCor implies independence. PICARD estimates the independent sources sequentially, by looking for the component that has the smallest dCor with the remainder. We prove that PICARD is strongly consistent. Its robustness is investigated by a simulation study, in which it generally outperforms its competitors. The method is illustrated on three applications, including the well-known cocktail party problem.

Counting cycles with AI

J. Jin[1]

[1]Carnegie Mellon University, United States

Despite recent progress, AI still struggles on advanced mathematics. We consider a difficult open problem: How to derive a Computationally Efficient Equivalent Form (CEEF) for the cycle count statistic? The CEEF problem does not have known general solutions, and requires delicate combinatorics and tedious calculations. Such a task is hard to accomplish by humans but is an ideal example where AI can be very helpful. We solve the problem by combining a novel approach we propose and the powerful coding skills of AI. Our results use delicate graph theory and contain new formulas for general cases that have not been discovered before. We find that, while AI is unable to solve the problem all by itself, it is able to solve it if we provide it with a clear strategy, a step-by-step guidance and carefully written prompts. For simplicity, we focus our study on DeepSeek-R1 but we also investigate other AI approaches.

Slacked Empirical Likelihoods for Post-Criterion Inference

Y. She[1]

[1]Westlake University, China

Statistical inference under nonsmooth penalties presents significant challenges for traditional empirical likelihood methods developed in low dimensions. We introduce SEL from the penalized criterion itself and use slack variables to convert its structural conditions into a tractable dual formulation over a family of divergence functions. This makes penalty-induced structural bias explicit, yielding a transparent distinction between noncentral and central limiting behavior. In particular, SEL introduces a data-driven dual centering scheme that cancels the bias terms and recovers a central chi-square limit. The resulting theory covers classical and high-dimensional regimes, attains sample-complexity scalings standard in sparse inference.

Recent Advances in Statistical Analysis for Graphs and Networks

13:30 - 14:00

Representation Learning with Blockwise Missingness and Signal Heterogeneity

W. Tang[1]

[1]Carnegie Mellon University, United States

Unified representation learning for multi-source data integration faces two important challenges: blockwise missingness and blockwise signal heterogeneity. The former arises from sources observing different, yet potentially overlapping, feature sets, while the latter involves varying signal strengths across subject groups and feature sets. While existing methods perform well with fully observed data or uniform signal strength, their performance degenerates when these two challenges coincide, which is common in practice. To address this, we propose Anchor Projected Principal Component Analysis (APPCA), a general framework for representation learning with structured blockwise missingness that is robust to signal heterogeneity. APPCA first recovers robust group-specific column spaces using all observed feature sets, and then aligns them by projecting shared "anchor" features onto these subspaces before performing PCA. This projection step induces a significant denoising effect. We establish estimation error bounds for embedding reconstruction through a fine-grained perturbation analysis. In particular, using a novel spectral slicing technique, our bound eliminates the standard dependency on the signal strength of subject embeddings, relying instead solely on the signal strength of integrated feature sets. We validate the proposed method through extensive simulation studies and an application to multimodal single-cell sequencing data. This is the joint work with Ziqi Liu and Ye Tian.

U-aggregation: Unsupervised Aggregation of Multiple Learning Algorithms

R. Duan[1]

[1]Harvard University, United States

Across various domains, the growing advocacy for open science and open-source machine learning has made an increasing number of models publicly available. These models allow practitioners to integrate them into their own contexts, reducing the need for extensive data labeling, training, and calibration. However, selecting the best model for a specific target population remains challenging due to issues like limited transferability, data heterogeneity, and the difficulty of obtaining true labels or outcomes in real-world settings. In this paper, we propose an unsupervised model aggregation method, U-aggregation, designed to integrate multiple pre-trained models for enhanced and robust performance in new populations. Unlike existing supervised model aggregation or super learner approaches, U-aggregation assumes no observed labels or outcomes in the target population. Our method addresses limitations in existing unsupervised model aggregation techniques by accommodating more realistic settings, including heteroskedasticity at both the model and individual levels, and the presence of adversarial models. Drawing on insights from random matrix theory, U-aggregation incorporates a variance stabilization step and an iterative sparse signal recovery process. These steps improve the estimation of individuals' true underlying risks in the target population and evaluate the relative performance of candidate models. We provide a theoretical investigation and systematic numerical experiments to elucidate the properties of U-aggregation. We demonstrate its potential real-world application by using U-aggregation to enhance genetic risk prediction of complex traits, leveraging publicly available models from the PGS Catalog.

Wedge sampling: Efficient tensor completion with nearly-linear sample complexity

Y. Zhu[1]

[1]University of Southern California, United States

We introduce \emph{Wedge Sampling}, a new non-adaptive sampling scheme for low-rank tensor completion. We study recovery of an order-k low-rank tensor of dimension $n\times\cdots\times n$ from a subset of its entries. Unlike the standard uniform entry model (i.e., i.i.d.\ samples from $[n]^k$), wedge sampling allocates observations to structured length-two patterns (wedges) in an associated bipartite sampling graph. By directly promoting these length-two connections, the sampling design strengthens the spectral signal that underlies efficient initialization, in regimes where uniform sampling is too sparse to generate enough informative correlations. Our main result shows that this change in sampling paradigm enables polynomial-time algorithms to achieve both weak and exact recovery with nearly linear sample complexity in n. The approach is also plug-and-play: wedge-sampling–based spectral initialization can be combined with existing refinement procedures (e.g., spectral or gradient-based methods) using only an additional $\tilde O(n)$ uniformly sampled entries, substantially improving over the $\tilde O(n^{k/2})$ sample complexity typically required under uniform entry sampling for efficient methods. Overall, our results suggest that the statistical-to-computational gap highlighted in \cite{barak.moitra_2016_noisy} is, to a large extent, a consequence of the uniform entry sampling model for tensor completion, and alternative non-adaptive measurement designs that guarantee a strong initialization can overcome this barrier.

Random geometric graphs with smooth kernels: sharp detection threshold and a spectral conjecture

C. Mao[1], Y. Wu[2], J. Xu[3]

[1]Georgia Institute of Technology, United States
[2]Yale University, United States
[3]Duke University, United States

A random geometric graph with kernel K is generated by sampling hidden points x_1 through x_n independently and uniformly on the d-dimensional sphere, and then connecting each pair of vertices with probability K(). We study the sharp detection threshold: the largest dimension for which this model can still be distinguished from an Erdős–Rényi graph with the same edge density. We show that for smooth kernels the critical scaling is d = n^(3/4) in the dense regime, which is much smaller than the d = n^3 threshold known for hard random geometric graphs with step-function kernels. We also extend the result to kernels whose signal-to-noise ratio varies with n, and propose a unifying conjecture that the critical dimension is characterized by a condition involving the standardized kernel operator.

Our main technical contribution is a new analysis of the posterior distribution of the latent points given the observed graph, especially the overlap between two independent posterior samples. As a by-product, we show that d = n^(1/2) is the critical dimension for nontrivial estimation of the latent inner products.

Computer-intensive statistical inference for complex data

13:30 - 14:00

Improved rate of convergence of generalized subsampling estimator for nonstationary time series

P. Bertail[1], A.E. Dudek[2,3], L. Lenart[4]

[1]MODALX - University Paris-Nanterre, France
[2]AGH university in Krakow, Poland
[3]Université Aix-Marseille, France
[4]Krakow University of Economics, France

It is well-established that for independent data, aggregating subsampling estimators—a technique known as "subagging"—can enhance convergence rates. This paper extends this principle to general nonstationary time series. We introduce a generalized subsampling estimator that aggregates the mean, median, and trimmed mean of individual subsample estimates, and establish its mean-square consistency under standard assumptions, including finite moments and α-mixing dependence. Beyond consistency, we prove that, in the case of the mean, the proposed estimator satisfies a Bernstein-type concentration inequality. Critically, we show that under conditions of negligible or zero bias, the aggregated estimator achieves an improved rate of convergence, mirroring the favorable properties observed in the independent case. This demonstrates the validity and power of the subagging principle even under dependence. We apply our results to estimate the Fourier coefficients of the time-varying autocovariance function for periodically correlated (cyclostationary) time series—a problem of significant practical interest. By aggregating subsampled Fourier coefficients, our method yields more reliable inference for the underlying periodic structure. Simulation studies corroborate our theoretical findings, empirically validating the improved convergence rates and concentration properties across various nonstationary scenarios.

Subgraph counts for hypergraphons with node covariates: asymptotic and bootstrap inference

E. Barthel[1], C. Jentsch[1]

[1]TU Dortmund University, Germany

As a natural extension of the graphon random graph model, we consider the hypergraphon model, which is an inhomogeneous model for exchangeable uniform hypergraphs. These are also called m-graphs and allow for edges that connect multiple (exactly m many) nodes. For $m=2$, the classical graphon model is recovered. In this setup, we equip the nodes with covariates and study weighted subhypergraph counts, where the weights are determined by the node covariates. Specifically, each found copy of a fixed subhypergraph is weighted by a function of its node covariates. While the node covariates and the (node-wise) latent positions of the hypergraphon are assumed to be jointly i.i.d., the dependence of the covariates and the latent positions may be arbitrary. For instance, our setup allows to count only those copies of a subhypergraph, whose node covariates satisfy certain properties. In this general setup, using the theory of generalized U-statistics, we give a full characterization of the joint limiting distributions of covariate-weighted subhypergraph counts. While the resulting limiting distribution is often normal, in general, it is a sum of random variables in different orders of the Wiener chaos. In particular, this extends known results on the asymptotic distribution of ordinary subgraph counts in the classical graphon random graph model (with m=2). Furthermore, assuming estimable hypergragphons, we show first-order consistency of a class of network bootstrap approaches for vectors of such counts in the case of a limiting normal distributions. In the special case of no node covariates, we extend these consistency results also to the non-normal limiting distribution case. To establish the latter, we propose a novel joint limiting theorem for generalized U-statistics on a triangular array type sample with sample-size dependent kernels, which may be of independent interest. The performance of these bootstrap methods is explored in simulations.

Analysis of quadratic forms of high-dimensional non-stationary time series, with application to ANOVA and independent testing

Y. Zhang[1]

[1]The Chinese University of Hong Kong, Shenzhen, China

Quadratic forms of time series are common in statistical applications, making analysis of these structures important for understanding statistics of time series data. This talk introduces theoretical results, including a concentration inequality, a Gaussian approximation theorem, and a consistent variance estimator, for a quadratic form of high-dimensional non-stationary time series. Building on these results, we introduce an ANOVA procedure, as well as a hypothesis testing method for independence of two time series. We further develop distributional results of the test statistics, and propose dependent wild bootstrap algorithms to facilitate hypothesis testing through Monte-Carlo simulations. Numerical studies and real-life data applications demonstrate the good performance of the proposed test statistics.

Nonparametric Density Estimation in High-dimensions via Tensor Train

D. Wang[1]

[1]UCSD, United States

We propose a tensor-based linear algebraic framework for density estimation and sampling. The method consists of two simple steps: first, smoothing the empirical density tensor with cluster-basis kernels to reduce estimator variance; second, compressing the smoothed empirical density tensor into tensor-train format. Numerical results show that the proposed method achieves accurate density estimation and sampling in dimensions up to 100.

Contributed: Structured Data, Ranking and Classification

13:30 - 13:50

Robustifying the conditional independence testing via marginal permutation

Z. Niu[1], J. Ai[1]

[1]University of Pennsylvania, United States

Permutation tests are classically appealing because they are finite-sample valid under a permutation-invariant null, but in conditional independence testing the true null is weaker and generally not permutation invariant. In this talk, I will show that permutation can nevertheless remain valid well beyond its classical scope for a broad class of modern conditional independence tests, from permuted score tests to semiparametric procedures such as GCM and PCM. The central insight is that these methods can be written as studentized residual-product statistics: under a strong exchangeable null, permutation delivers exact finite-sample calibration, while under the weaker conditional independence null it still yields asymptotically valid inference. For orthogonalized tests such as GCM and PCM, this robustness is especially attractive because their existing double-robustness properties are inherited, and permutation adds an additional, essentially free layer of robustness at the calibration stage without sacrificing first-order local power. Overall, the talk presents permutation calibration as a simple but powerful principle that unifies parametric and semiparametric conditional independence testing.

Debiased Ill-Posed Functional Operator Estimation

E.D. Roth[1]

[1]Universidad Carlos III de Madrid, Spain

We study debiased estimation and tuning-parameter selection for ill-posed functional operator problems characterized by compact inverse equations. In the functional IV model, we derive a first-order expansion of the projected error in Bochner \(L^2\) geometry, which yields an influence-function correction for the empirical criterion. The resulting debiased loss is locally robust to first-stage nuisance estimation and exhibits a second-order bias structure, with robustness to misspecification of one nuisance component when the other is consistently estimated. Monte Carlo evidence from a fully functional endogenous design shows that the debiased procedure improves operator recovery, reduces structured bias, and selects more stable regularization levels than the naive plug-in rule. These results suggest that orthogonalized projected-error criteria can be a useful tool for regularization choice in functional inverse problems.

On the Edgeworth expansion of the maxima and the blessings of dimensionality

D. Peer[1], M. Jirak[1]

[1]University of Vienna, Austria

Let $X_1,\ldots, X_n \in \mathbb{R}^d$ be a sequence of i.i.d. random vectors, where d may be potentially much larger than n. A fundamental problem in high-dimensional statistics concerns normal approximations and convergence properties of the maximum statistic $$M_n=\max_{1\leq k\leq d} \frac{1}{\sqrt{n}}\sum_{i=1}^n X_{i,k},$$ whose study was initiated in seminal works by Chernozhukov, Chetverikov and Kato. A next step in understanding the asymptotic properties of M_n and accompanying quantile approximations is the development of Edgeworth-type expansions and corresponding bootstrap methods. A very recent result in this direction was established by Koike, developing an Edgeworth expansion for $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ based on Stein kernels, subject to some regularity conditions. In our project, we view the problem through the lens of Poisson-approximations to directly construct an Edgeworth expansion for M_n. Our main assumptions are a Cram\'{e}r-type condition for all pairs of components of X_i and a notion of weak dependence across the dimension. Utilizing this expansion, we obtain second order approximations for $\mathbb{P}(M_n\leq x)$ and the quantiles of M_n. Under suitable uniformity assumptions on the moments across components, we improve these convergence rates to third and higher orders. Furthermore, we extend our results to studentized case, that is to the statistic $\max_{1\leq k\leq d} T_{n,k}$, where $T_{n,k}$ are the component-wise Student-t statistics.

Partition-based discriminant analysis

B. Nipoti[1]

[1]University of Milano Bicocca, Italy

We propose a Bayesian nonparametric framework that enhances classical discriminant analysis in settings characterized by limited sample sizes and high estimation uncertainty. The proposed method provides a flexible approach that encompasses both linear and quadratic discriminant analysis as special cases. This is accomplished through a scale-only nonparametric mixture model defined on the space of positive definite matrices. Within this framework, observations are modeled as Gaussian with class-specific mean vectors and covariance matrices that may be unique or shared across classes. The key innovation lies in allowing information sharing across classes, which improves the estimation of covariance matrices and stabilizes classification boundaries in small-sample regimes. A conjugate nonparametric prior ensures remarkable ease of implementation and tractability. This allows closed-form characterization of several posterior quantities of interest, including the induced partition of classes and the number of covariance clusters. The resulting methodology is straightforward to implement and avoids the need for numerical approximations. The tractability of the proposed model further enables theoretical investigation of its asymptotic behavior, providing insight into its large-sample properties. Through experiments on both simulated and real datasets, we demonstrate the adaptability and effectiveness of the proposed methodology.

Recent developments in missing data and causal inference

16:00 - 16:30

Estimation of conditional cumulative incidence functions under generalized semiparametric regression models with missing covariates, with application to analysis of biomarker correlates in vaccine trials

Y. Sun[1], F. Heng[2], U. Lee[3], P. Gilbert[4]

[1]University of North Carolina at Charlotte, United States
[2]University of North Florida, United States
[3]CBER, Food and Drug Administration, United States
[4]Fred Hutchinson Cancer Center, United States

This article studies generalized semiparametric regression models for conditional cumulative incidence functions with competing risks data when covariates are missing by sampling design or happenstance. A doubly robust augmented inverse probability weighted complete-case (AIPW) approach to estimation and inference is investigated. This approach modifies IPW complete-case estimating equations by exploiting the key features in the relationship between the missing covariates and the phase-one data to improve efficiency. An iterative numerical procedure is derived to solve the nonlinear estimating equations. The asymptotic properties of the proposed estimators are established. A simulation study examining the finite-sample performances of the proposed estimators shows that the AIPW estimators are more efficient than the IPW estimators. The developed method is applied to the RV144 HIV-1 vaccine efficacy trial to investigate vaccine-induced IgG binding antibodies to HIV-1 as correlates of acquisition of HIV-1 infection while taking account of whether the HIV-1 sequences are near or far from the HIV-1 sequences represented in the vaccine construct.

Encoding and inference on separable effects for sustained treatments

M. Stensrud[1], I. Gonzalez Perez[2]

[1]Ecole Polytechnique Fédérale de Lausanne, Switzerland
[2]EPFL, Switzerland

In many applications, especially medicine, treatments are given repeatedly over time. In such settings, what often matters most is the effect of following a treatment strategy throughout follow-up, rather than the effect of being assigned or starting treatment. This is also relevant in mediation and mechanistic studies, where the conceptual goal is to understand how treatment effects operate through different pathways, when the treatment is actually taken versus not taken. Here we consider the separable effect of sustained use of a time-varying treatment. Despite the potential usefulness of this estimand, the theory of separable effects has yet to be extended to settings with sustained treatment strategies.

In this talk, I discuss the use of an unconventional encoding of of time-varying treatment strategies. This allows to obtain concise formulations of identifying assumptions with better practical properties; for example, they admit frugal graphical representations and formulations of identifying functionals. I introduce conditions under which sustained separable effects can be identified, and discuss how to assess them and when they might be violated. The identification formulas are used to propose a collection of estimands, one of which achieves double-robustness properties. As an application, I will an analysis of the Systolic Blood Pressure Intervention Trial (SPRINT), where we estimated a sustained separable effect of modified blood pressure treatments on the risk of acute kidney injury.

Private Rate-Double-Robust Inference

M. Kormos[1], A. van der Vaart[2]

[1]Ghent University, Belgium
[2]Delft University of Technology, Netherlands

We reconcile privacy protection and rate-double-robust inference. The privacy of individuals is protected by a local privacy mechanism: injecting noise into their sensitive data, revealing only the noisy data for inference. Hence, privacy protection hinders inference. In contrast, the inference of a target parameter is rate-double-robust when the large-sample bias of an estimator of the parameter is characterised by a trade-off between the estimation errors of two other, nuisance, parameters. Hence, rate-double-robustness facilitates inference. Our starting point of reconciliation is a novel class of rate-double-robust target parameters allowing for nonlinear dependencies on low-dimensional regressions. Among others, this includes causal parameters. To infer these targets privately, we show how suitable privacy mechanisms transfer the semiparametric properties of the sensitive-data model to the private setting. Rate-double-robustness is transferred, enabling locally-private, unbiased and semiparametrically efficient inference of our target parameters. Finally, we transform general nonparametric nuisance estimators into private ones, which inherit convergence properties of their nonprivate counterparts. For parametric nuisance models, we develop a private method-of-moments estimator and its large-sample inference theory.

Cross-Balancing for Data-Informed Design and Efficient Analysis of Observational Studies

Y. Jin[1], J. Zubizarreta[2]

[1]University of Pennsylvania, United States
[2]Harvard University, United States

Causal inference starts with a simple idea: compare groups that differ by treatment, not much else. Traditionally, comparable groups are constructed using only observed covariates; however, it remains a long-standing challenge to incorporate available outcome data into the study design while preserving valid inference. In this paper, we study the general problem of covariate adjustment, effect estimation, and statistical inference when balancing features are constructed or selected with the aid of outcome information from the data. We propose cross-balancing, a method that uses sample splitting to separate the error in feature construction from the error in weight estimation. Our framework addresses two cases: one where the features are learned functions and one where they are selected from a potentially high-dimensional dictionary. In both cases, we establish mild and general conditions under which cross-balancing yields consistent, asymptotically normal, and efficient estimators. In the learned-function case, cross-balancing achieves finite-sample bias reduction relative to plug-in estimators and is multiply robust when the learned features converge at slow rates. In the variable-selection case, cross-balancing requires only a product condition on the approximation quality of the selected variables. We illustrate cross-balancing in extensive simulations and an observational study, demonstrating that careful use of outcome information can substantially improve both estimation and inference while maintaining interpretability.

Goodness-of-fit and specification tests

16:00 - 16:30

A portmanteau test for multivariate non-stationary functional time series with an increasing number of lags

L. Bai[1], H. Dette[1], W. Wu[2]

[1]Ruhr University Bochum, Germany
[2]Tsinghua University, China

Multivariate locally stationary functional time series provide a flexible framework for modeling complex data structures exhibiting both temporal and spatial dependencies while allowing for time-varying data generating mechanism. In this paper, we introduce a portmanteau-type test for assessing white noise assumptions tailored for multivariate lo-
cally stationary functional time series without dimension reduction. A simple bootstrap procedure is proposed to implement the test because the limiting distribution can be non-standard or even does not exist. Our approach is based on a Gaussian approximation result for a degenerate ?-statistic of second-order functional time series involving an increasing
number of lags, which is of independent interest. Through theoretical analysis and simulation studies, we demonstrate the efficacy and adaptability of the proposed method in detecting departures from white noise assumptions in multivariate locally stationary functional time series.

Almost goodness-of-fit tests

A. Baíllo[1], J. Cárcamo[2]

[1]Universidad Autónoma de Madrid, Spain
[2]Universidad del País Vasco, Spain

The almost goodness-of-fit test is a procedure to decide if a (parametric) model provides a good representation of the probability distribution generating the sample. We consider the approximate model determined by an M-estimator of the parameters as the best representative of the unknown distribution within the parametric class. The objective is the
approximate validation of a distribution or an entire parametric family up to a pre-specified error margin. The methodology also allows quantifying the percentage improvement of the proposed model compared to a non-informative (constant) one. The test statistic is the Lp distance between the empirical distribution function and the corresponding one of the
estimated (parametric) model. We present an easy-to-implement and flexible bootstrap scheme to carry out the test. The performance of the proposal is illustrated via the analysis of real data sets.

New tests of complete randomness of 2D-binary image data based on geometric functionals

B. Ebner[1]

[1]Karlsruhe Institute of Technology (KIT), Germany

We present new methods for testing the randomness of binary image data. The proposed tests are based on geometric functionals, also known as Minkowski functionals, such as area, perimeter, and the Euler–Poincaré characteristic. We develop methods for both the simple hypothesis, in which the coloring probability is known, and the composite hypothesis, in which it is unknown. Using Stein's method together with dependency graphs, we derive the limiting null distributions of the test statistics and establish convergence in Kolmogorov distance. A Monte Carlo simulation study shows that the tests are able to detect alternatives. We then apply them to data associated with irrational numbers and mathematical constants, including pi, e, the Euler–Mascheroni constant, and Catalan's constant.

Change-Point Detection via Characteristic Functions and Optimal Measure Transport

Š. Hudecová[1]

[1]Charles University, Czech Republic

Detecting structural changes in a sequence of observations is a central problem in statistical inference, particularly when the observations are multivariate. In many applications, it is desirable to operate in a fully nonparametric setting, without imposing restrictive assumptions on the underlying distribution or on the nature of the change.

In this contribution, we consider the problem of detecting an abrupt change in the distribution of multivariate observations. We propose a distribution-free test based on multivariate ranks constructed via optimal measure transport, whose test statistic is defined through empirical characteristic functions. We describe a method for constructing an exact test and also provide the asymptotic distribution of the test statistic. The performance of the proposed procedure is illustrated through data examples.

Topics in Causal Inference

16:00 - 16:30

Longitudinal Regression Discontinuity Designs under Local Sequential Randomization

L. Forastiere[1], A. Mattei[2], F. Mealli[3]

[1]Department of Biostatistics, Yale School of Public Health., United States
[2]Department of Statistics, Computer Science, Applications; University of Florence, Italy
[3]Department of Economics; European University Institute (EUI), Italy

We study longitudinal regression discontinuity (RD) designs in which treatment is dynamically assigned over time through a sequence of cutoff rules. We focus on a two-period setting and allow the forcing variable in the first period to affect the forcing variable in the second period. Within the potential outcomes framework, we formally characterize longitudinal RD designs as local sequentially latent regular designs. Under this framework, we define causal estimands that capture the effects of treatment sequences on well-defined, though generally unknown, sub-populations for which local overlap conditions and local Stable Unit Treatment Value Assumptions (SUTVA) hold. For each subpopulation, we specify the treatment assignment mechanism by invoking local longitudinal unconfoundedness assumptions. Specifically, we assume local conditional independence between the potential outcomes for the primary endpoint and the first-period forcing variable, conditional on observed covariates. In addition, we consider alternative local conditional independence governing the relationship between the potential outcomes of the second-period forcing variable and those of the primary endpoint. A central challenge of the proposed approach is the identification of the subpopulations for which valid causal inference is possible. To address this issue, we extend the Bayesian model-based finite mixture approach proposed by Forastiere et al. (AOAS, 2025) to the longitudinal setting. We probabilistically clusters observations into subpopulations in which the required assumptions either hold or fail, based on their observable implications. We then derive posterior distributions of the target causal effects by marginalizing over uncertainty in subpopulation membership. We apply the proposed methodology to the evaluation of Italian university student-aid policies on academic outcomes.

Randomization Inference of Treatment Effects with the Estimated Propensity Score

P. Toulis[1], S. Huang[1], J. Du[2], A. Shaikh[3]

[1]University of Chicago, Booth School of Business, United States
[2]University of Chicago, Department of Statistics, United States
[3]University of Chicago, Department of Economics, United States

In this paper, we study the properties of design-based (or randomization) inference for treatment effects when analyzing observational data under unconfoundedness. In such settings, the design is defined by the propensity score, i.e., the distribution of treatment status given the covariates, which is typically unknown and must be estimated from the data. For the sharp null hypothesis of no treatment effect, we derive non-asymptotic bounds on the size distortion of such tests that depend only on the error in estimating the propensity score. In contrast to prior related work, our results do not require that the propensity score be estimated using independent auxiliary data, as is commonly assumed in methods based on sample splitting. For the weak null hypothesis of no average treatment effect, we show that such tests are asymptotically valid when based on common estimators of the average treatment effect, including inverse-propensity-weighted and doubly robust estimators. We further compare our tests with conventional (non-randomization) tests for the weak null hypothesis and, since these tests are first-order equivalent, develop higher-order comparisons using Edgeworth expansions. Our analysis reveals that, from this perspective, neither approach uniformly dominates the other; however, as one might expect, the randomization test enjoys higher-order accuracy “close” to the sharp null hypothesis—for example, when treatment effects are sufficiently small or rare.

Rerandomized Saturation Designs

A. Frosini[1], T. Arduini[1], P. Ding[2], L. Forastiere[3]

[1]Tor Vergata University of Rome, Italy
[2]University of California, Berkeley, United States
[3]Yale University, United States

Interference arises when a unit's outcome depends on other units' treatments. Randomized saturation designs, which randomly assign cluster to saturations and individual treatments within clusters according to the assigned saturation, are used to estimate direct, indirect, total and overall effects under partial interference. To improve covariate balance, we propose rerandomized saturation designs that rerandomize saturations and individual assignments until prespecified balance criteria are met. Under randomization inference with stratified interference and without distributional or modeling assumptions on covariates or outcomes, we derive the efficiency gain as well as the joint asymptotic sampling distribution of standard weighting estimators. Rerandomization yields a symmetric, unimodal limit that is more concentrated around the true effects and supports conservative inference. Motivated by settings where the two stages occur at different times, we further analyze a stepwise protocol that rerandomizes saturations to balance cluster covariates and, conditional on acceptance, rerandomizes individual treatments to balance individual covariates. We compare the joint and the stepwise strategies and provide a decision-theoretic framework for choosing stage-specific thresholds under a fixed overall acceptance probability.

Orthogonal Representation Learning for Estimating Causal Quantities

V. Melnychuk[1], D. Frauen[1], J. Schweisthal[1], S. Feuerriegel[1]

[1]LMU Munich & Munich Center for Machine Learning, Germany

End-to-end representation learning has become a powerful tool for estimating causal quantities from high-dimensional observational data, but its efficiency remained unclear. Here, we face a central tension: End-to-end representation learning methods often work well in practice but lack asymptotic optimality in the form of the quasi-oracle efficiency. In contrast, two-stage Neyman-orthogonal learners provide such a theoretical optimality property but do not explicitly benefit from the strengths of representation learning. In this work, we step back and ask two research questions: (1) When do representations strengthen existing Neyman-orthogonal learners? and (2) Can a balancing constraint – commonly proposed technique in the representation learning literature – provide improvements to Neyman-orthogonality? We address these two questions through our theoretical and empirical analysis, where we introduce a unifying framework that connects representation learning with Neyman-orthogonal learners (namely, OR-learners). In particular, we show that, under the low-dimensional manifold hypothesis, the OR-learners can strictly improve the estimation error of the standard Neyman-orthogonal learners. At the same time, we find that the balancing constraint requires an additional inductive bias and cannot generally compensate for the lack of Neyman-orthogonality of the end-to-end approaches. Building on these insights, we offer guidelines for how users can effectively combine representation learning with the classical Neyman-orthogonal learners to achieve both practical performance and theoretical guarantees.

Complex data and depth

16:00 - 16:30

Blessing of dimensionality in cross-validated bandwidth selection on the sphere

J.E. Chacón[1], E. García-Portugués[2], A. Meilán-Vila[2]

[1]Universidad de Extremadura, Spain
[2]Universidad Carlos III de Madrid, Spain

We study the asymptotic behavior of least-squares cross-validation (LSCV) bandwidth selection in kernel density estimation on the d-dimensional hypersphere. First, we show the existence of the optimal bandwidth minimizing the mean integrated squared error (MISE) under mild non-uniformity conditions. Then, we obtain the exact rate of convergence of the LSCV bandwidth selector with respect to the MISE-optimal bandwidth. Surprisingly, this rate approaches the root-n parametric rate as d grows. This "blessing of dimensionality" in bandwidth selection offers theoretical support for utilizing the conceptually simpler LSCV selector over plug-in techniques for larger dimensions. Numerical experiments corroborate the speed of this convergence in an array of scenarios and dimensions, precisely illustrating the tipping dimension where cross-validation outperforms plug-in approaches.

A probabilistic characterization of k-means uniqueness

J. Cárcamo[1], A. Cuevas[2], L. Rodríguez[3]

[1]University of the Basque Country, Spain
[2]Universidad Autónoma de Madrid, Spain
[3]Georg-August-Universität Göttingen, Germany

We give necessary and sufficient conditions for the uniqueness of the k-means set of a probability distribution. This uniqueness property is closely related to the choice of the parameter k, since some values may produce multiple optimal clusterings, complicating interpretation and affecting algorithmic stability. Within a risk minimization framework and under standard Donsker-type conditions, we derive the asymptotic distribution of the empirical within-cluster sum of squares (WCSS). This, in turn, provides a criterion for k-means uniqueness in terms of the limiting behavior of the empirical WCSS. Building on this result, we construct a practical bootstrap test for assessing k-means uniqueness. The finite-sample performance of the procedure is investigated in a simulation study.

This is a joint work with Antonio Cuevas (Universidad Autónoma de Madrid) and Luis-Alberto Rodríguez (Institut für Mathematische Stochastik, Georg-August-Universität Göttingen).

Sequential Monte Carlo Depth Computation

A. Nieto-Reyes[1], C. Kirch[2], F. Gnettner[3]

[1]University of Cantabria, Spain
[2]Heinrich-Heine-Universität Düsseldorf, Germany
[3]South Dakota State University, United States

Statistical depth functions rank observations from the centre outwards in spaces where no natural ordering exists, but their computation can be very expensive. We propose a new sequential Monte Carlo method, seMCD, to approximate depth functions and related quantities with rigorous statistical guarantees. The method returns an interval, among a set of intervals pre-specified by the user, that contains the target value with high user-specified probability and often needs far fewer simulations than standard Monte Carlo. It applies to a variety of depth functions, multivariate and functional, and performs well in tasks such as outlier detection, classification and depth-region computation.

Uniform consistency of Tukey depth for fuzzy sets

L. González-De La Fuente[1], A. Nieto-Reyes[1], P. Terán[2]

[1]Universidad de Cantabria, Spain
[2]Universidad de Oviedo, Spain

The computation of depth functions for fuzzy sets is fundamental for their use in depth-based statistical procedures. In this work, we examine the Tukey depth for fuzzy data and explore the conditions ensuring that the depth function with respect to the empirical measure converges uniformly to its counterpart under the true distribution. This property is referred to as the uniform consistency of the depth function. As a result, we prove that the sample 1-median of a fuzzy random number converges to the population 1-median under various metrics defined on the fuzzy space.

Recent Advances of Time Series Analysis

16:00 - 16:30

Sharp oracle inequalities for covariate selection via the AIC

G. Köstenberger[1], J.M. Jirak[1]

[1]University of Vienna, Austria

Prediction is a core task in time series analysis, and the choice of the forecasting model is of fundamental importance.
To address this problem (among other things), Akaike introduced the AIC and FPE, and demonstrated their significant usefulness for prediction in two landmark papers. In subsequent seminal works, Shibata developed a notion of asymptotic efficiency and showed that both AIC and FPE are optimal, setting the stage for decades-long developments and research in this area and beyond.

Most of the literature on the usage of AIC for prediction focuses on the case of nested models. However, there is no fundamental information theoretic reason for this restriction, as the AIC is essentially an optimal estimator of the prediction error, which is based on minimizing the empirical Kullback-Leibler divergence between any two models. This point of view suggests that the AIC (and its variants) should be able to select an (optimal) forecasting model from any set of candidate models. In this work, we establish sharp, finite-sample oracle inequalities for the AIC in the non-nested case, subject only to a very general notion of weak dependence. This establishes a unversality property of the AIC, in the sense that it can not just be used to compare between nested models, but rather between arbitrary models. Our framework contains many prominent dynamical systems such as random walks on the regular group, functionals of iterated random systems, functionals of (augmented) Garch models of any order, functionals of (Banach space valued) linear processes, possibly infinite memory Markov chains, dynamical systems arising from SDEs, and many more. The proofs require a new notion of uniform integrability in the Wiener algebra, a new uniform version of the Wiener-Levy theorem and a uniform Baxter-type inequality, which may be of independent interest.

An operator-level ARCH Model

A. Aue[1], S. Kühnert[2], G. Rice[3], J. Wanderdoes[3]

[1]University of California-Davis, United States
[2]Ruhr University Bochum, Germany
[3]University of Waterloo, Canada

AutoRegressive Conditional Heteroscedasticity (ARCH) models are standard for modeling time series exhibiting volatility, with a rich literature in univariate and multivariate settings. In recent years, these models have been extended to function spaces. However, functional ARCH and generalized ARCH (GARCH) processes established in the literature have thus far been restricted to model "pointwise" variances. We propose a new ARCH framework for data residing in general separable Hilbert spaces that accounts for the full evolution of the conditional covariance operator. We define a general operator-level ARCH model. For a simplified Constant Conditional Correlation version of the model, we establish conditions under which such models admit strictly and weakly stationary solutions, finite moments, and weak serial dependence. Additionally, we derive consistent Yule--Walker-type estimators of the infinite-dimensional model parameters. The practical relevance of the model is illustrated through simulations and a data application to high-frequency cumulative intraday returns.

Mercer Expansions in Sobolev Spaces and Applications to Stochastic Processes

D. Rademacher[1]

[1]TU Graz, Austria

Mercer's celebrated theorem is refined and extended by introducing a novel class of higher order kernel operators,
that includes the common integral operator only as a special case. These operators genuinely take into account information encoded in the (weak) derivatives of a kernel and their natural domains are Sobolev spaces of order k over some bounded d-dim. domainn, where k depends on the order of (weak) differentiability.
The spectral decomposition of such higher order kernel operators leads to Mercer-type expansions, which are optimal in term of the Sobolev norm and, if k>d, also converge uniformly without requiring the kernel to be positive definite.
Nuclearity of higher order kernel operators is confirmed for positive definite kernels and a major refinement of Mercer's theorem is obtained that implies trace formulas and a simple rate for the uniform convergence (including derivatives) in terms of the eigenvalues. A further immediate consequence are novel spectral representations of RKHS's.
Finally, applied to the covariance kernel of a (weakly) differentiable stochastic process, these refinements also yield novel Karhunen-Loève-type expansions allowing for simultaneous approximations of the process and its (weak) derivatives in a mean square optimal sense.

Principal Components Analysis for Irregular Data

A. Stoecker[1], V. Panaretos[1], K. Waghmare[2]

[1]EPFL, Switzerland
[2]ETH Zurich, Switzerland

Functional principal component analysis (FPCA) is a fundamental tool for exploring variation in samples of random curves or surfaces. We propose a new approach to FPCA for functional data observed irregularly and sparsely over their domains, based on smoothing directly at the level of the eigenfunctions. Our formulation leads to an efficient optimization-based procedure whose computational and storage costs are comparable to those of standard multivariate PCA for regularly observed data. The method is flexible with respect to domain geometry and model class, and accommodates structural constraints and penalties. More broadly, the underlying philosophy extends naturally to time series of irregularly observed functional data, where it enables coherent modeling and inference without resorting to ill-posed inverse problem formulations.

Modern Network Analysis I

16:00 - 16:30

Modelling multiplex networks

S. Olhede[1], C. Dufour[1], A. Skeja[2]

[1]EPFL, Switzerland
[2]Uppsala University, Sweden

This talk will discuss how to model and make inferences of multiplex networks. One point-of-view starts from a special class of models based on multivariate graph limits (see Skeja and Olhede (2024)). These can be represented as a limiting object known as a decorated graphon (see Dufour and Olhede (2024)), where a more complex object is represented by a generalization of a graph limit. The choice of object will depend on the particular application at hand, and we will discuss the particular case of coma patients (Verdeyme et al (2025)). We will discuss choices of parameterization, and choices of summary statistics for particular models, highlighting both the benefits and challenges of multiplex observations.

Euclidean mirrors and changepoints in network time series

A. Athreya[1], T. Chen[1], Z. Lubberts[2], Y. Park[1], C. Priebe[1]

[1]Johns Hopkins University, United States
[2]University of Virginia, United States

We describe a model for a network time series whose evolution is governed by an underlying stochastic process, known as the latent position process, in which network evolution can be represented in Euclidean space by a curve, called the Euclidean mirror. We define a class of changepoints for network time series and construct a family of latent position process networks with such changepoints. We show how a spectral estimate of the associated Euclidean mirror can localize these changepoints, and we provide simulated and real data examples of such localization.

Network Goodness-of-Fit for the block-model family

J. Wang[1], J. Jin[2]

[1]University of Virginia, United States
[2]Carnegie Mellon University, United States

The block model family is widely used in network modeling and includes four popular models: SBM, DCBM, MMSBM and DCMM. However, the question of which block model best fits real networks has received limited attention in the literature. In this talk, I will introduce a novel approach using cycle count statistics to address the Goodness-of-Fit for these block models. By leveraging the cycle count statistics and a network fitting scheme, we construct four GoF metrics with parameter-free limiting distributions of N(0,1) under the assumed models. We apply these GoF-metrics to some frequently-used real networks for comparison. The numerical results suggest that DCMM is particularly promising for modeling undirected networks.

How Networks Change Shape

I. Gallagher[1], M. Hajimir[1], P. Menendez[1]

[1]The University of Melbourne, Australia

Real-world networks rarely stand still. Social networks grow, biological systems evolve, and infrastructure adapts – understanding how and why networks change is essential. We model the observed dynamic network structure using node representations evolving through an unknown latent space. Recovering this latent structure is key to understanding network dynamics.

Dynamic embeddings represent the evolving behaviour of each node in a network as points moving through low-dimensional space. Nodes with similar behaviour have similar embeddings enabling the detection of anomalous nodes, communities merging or splitting, and entire networks changing structure over time.

In this talk, we explore how the topological features we observe in dynamic embeddings directly mirror the topological features of the underlying latent space. Persistent homology reveals the topological structure of a network by tracking connected components, cycles, and higher-dimensional holes as the network evolves. Using this approach, we prove that tracking these topological features is a principled method for understanding how networks change shape.

Theory & Computation for Gaussian processes

16:00 - 16:30

Deep Gaussian Processes and Vecchia Approximation

T. Randrianarisoa[1]

[1]University of Toronto, Canada

Deep Gaussian Processes (DGPs) have emerged as powerful tools for modeling complex, compositional structures in modern data, bridging the gap between Bayesian nonparametrics and deep learning. However, their widespread adoption is often hindered by severe computational bottlenecks; as a composition of GPs, their computational complexity inherently exceeds the already prohibitive O(n^3) cost of standard GP regression.
In the first part, we explore the theoretical guarantees of DGPs in nonparametric regression. Building upon the Deep Horseshoe Gaussian Process (Deep-HGP) framework with squared-exponential and Matérn kernels, we establish posterior contraction rates that achieve adaptive, near-optimal recovery under a squared loss. These rates simultaneously adapt to the unknown smoothness and compositional structure of the true regression function. Finally, we also show that they can achieve spatial adaptation, effectively capturing signals with spatially varying smoothness.
In the second part of this talk, we address this scalability challenge by introducing the Deep Vecchia Gaussian Process. We demonstrate that this approach yields scalable, valid Bayesian inference, while retaining minimax optimality over a broad class of composite functions.
Together, these contributions establish DGPs as both practically scalable and theoretically robust priors for complex inference.

Approximate posteriors in Gaussian process regression based on spectral concentration phenomena of kernel matrices.

B. Stankewitz[1]

[1]University of Potsdam, Germany

Due to their flexibility and theoretical tractability, Gaussian process regression models have become a central topic in modern statistics and machine learning. While the true posterior in these models is given explicitly, numerical evaluations depend on the inversion of the augmented kernel matrix $K + \sigma^2 I$, which requires up to $O(n^3)$ operations.
For large sample sizes n, which are typically given in modern applications, this is computationally infeasible and necessitates the use of an approximate version of the posterior. Although such methods are widely used in practice, they have limited theoretical underpinning. In this context, we analyze a class of recently proposed approximation algorithms from the field of probabilistic numerics. They can be interpreted in terms of Lanczos approximate eigenvectors of the kernel matrix or a conjugate gradient approximation of the posterior mean, which are particularly advantageous in truly large scale applications, as they are only based on matrix vector multiplications amenable to the GPU acceleration of modern software frameworks. We combine bounds from numerical analysis with state of the art concentration results for spectra of kernel matrices to obtain minimax contraction rates.

Bayesian nonparametrics for stochastic reaction diffusion equations

R. Altmeyer[1], S. Gaudlitz[2]

[1]Imperial College London, United Kingdom
[2]Humboldt-Universität zu Berlin, Germany

Stochastic partial differential equations (SPDEs) are a major subject of current research in probability and analysis, with rich methodologies for studying existence and regularity. At the same time, SPDEs are increasingly used as statistical models for spatially and temporally structured data, where inference requires learning unknown parameters or functions from observations. In this talk, we consider Bayesian inference for the reaction function in a stochastic reaction-diffusion equation, based on a single solution trajectory observed continuously in space over a fixed time interval. We place a Gaussian process prior on the reaction function and derive posterior contraction rates in a novel asymptotic regime in which the spatial domain grows while the observation horizon remains fixed. In this setting, the SPDE solution becomes spatially ergodic and converges to a stationary process, which allows for proving concentration inequalities for spatial averages of the solution. The proofs combine tools from Malliavin calculus - most notably the Clark–Ocone formula - with sharp bounds on the marginal densities of the SPDE.

Dependent p-values, exchangeability, and causal coupling

16:00 - 16:30

Aggregating dependent signals: validity and power of Heavy-Tailed Combination Tests

J. Wang[1]

[1]The University of Chicago, United States

A common problem in statistics is how to combine results from many tests when the tests are not independent. Traditional approaches, like Bonferroni, are valid but often overly conservative and lose power. Recently, “heavy-tailed” methods, such as the Cauchy combination test and the harmonic mean p-value, have become increasingly popular as more efficient ways to handle unknown dependence.
In this talk, I will present a unified perspective on the validity and power of these methods under different dependence structures. I will explain when they behave like Bonferroni—offering little advantage—and when they can deliver clear power gains. In particular, I will show that when p-values are pairwise asymptotically independent (such as pairwise normally distributed), heavy-tailed combination tests are equivalent to the Bonferroni test as significance thresholds shrink. However, under stronger forms of tail dependence modeled via multivariate regularly varying copulas, they remain asymptotically valid and achieve substantial improvements. A key insight is that the tail index of the transformation distribution determines the tradeoff between validity and power, with $\gamma=1$ maximizing power while preserving validity, whereas Bonferroni emerges as the degenerate case when $\gamma \to 0$. These results offer both theoretical insights and practical guidance, and I will illustrate a few examples in genetic analysis.

Combining exchangeable p-values

M. Gasparin[1]

[1]Università della Svizzera Italiana, Italy

The problem of combining p-values is an old and fundamental one, and the classic assumption of independence is often violated or unverifiable in many applications. There are many well-known rules that can combine a set of arbitrarily dependent p-values (for the same hypothesis) into a single p-value. We show that essentially all these existing rules can be strictly improved when the p-values are exchangeable, or when external randomization is allowed (or both). For example, we derive randomized and/or exchangeable improvements of well known rules like “twice the median” and “twice the average”, as well as geometric and harmonic means. Our work also improves rules for combining arbitrarily dependent p-values, since the latter becomes exchangeable if they are presented to the analyst in a random order. The main technical advance is to show that all existing combination rules can be obtained by calibrating the p-values to e-values (using an α-dependent calibrator), averaging those e-values, converting to a level-α test using Markov’s inequality, and finally obtaining p-values by combining this family of tests; the improvements are delivered via recent randomized and exchangeable variants of Markov’s inequality (joint work with R. Wang and A. Ramdas).

User-friendly causal coupling: Nonparametric instrumental variable models and beyond

R. Guo[1]

[1]University of Michigan, United States

Relaxing strong, often untestable assumptions in causal inference often leads to partial identification, where effects of interest become identified as intervals rather than single values. Recent computational and theoretical advances have made partial identification more practically relevant for credible, assumption-lean causal inference. However, obtaining sharp bounds of an effect from the observed distribution remains a technical challenge: existing tools are often not generally applicable or can be too cumbersome to use. In this talk, I will introduce a user-friendly framework for partial identification based on the coupling between the counterfactual distribution and the observed distribution. Drawing on polytope geometry and a seminal result by Volker Strassen, I will show that this framework leads to a mathematically minimal characterization of the set of counterfactual distributions and, consequently, the sharp bounds. I will illustrate its application through nonparametric instrumental variable models. Time permitting, I will also discuss how to incorporate sampling variability via convex programming.

On the universal calibration of heavy-tailed combination tests

P. Chakraborty[1], R. Guo[1], S. Stoev[1], K. Shedden[1]

[1]University of Michigan, Ann Arbor, United States

Recently, several heavy-tailed combinations tests, such as the harmonic mean test and the Cauchy combination test, have been proposed to test a global null hypothesis while controlling the Type-I error rate. They transform p-values into heavy-tailed random variables before combining them into a single test statistic. The resulting tests, which are calibrated under some form of independence assumption among the p-values, have shown to be rather robust to dependence asymptotically as the test level gets small. Yet, it has remained an open problem to understand this phenomenon generically and characterize how such tests behave under dependence.

In this work, we use the framework of multivariate regular variation (MRV) to study this problem. We show that for a class of combination tests that are homogeneous, the asymptotic level of the test can be expressed using the angular measure under MRV, which characterizes the dependence of the transformed variables in their upper tails, or equivalently, the dependence of the p-values near zero. We use this result to study several tests. In particular, the Pareto linear combination test, also known as the harmonic mean test, is shown to be universally calibrated regardless of the tail dependence; further, this test is shown to be the only one that achieves universal calibration among all homogeneous heavy-tailed combination tests. In contrast, the Cauchy combination test is shown to be universally honest but often conservative; the Tippett combination test, while being honest, is calibrated if and only if the underlying p-values are independent near zero.

Recent developments in differential privacy

16:00 - 16:30

Query-Efficient Locally Private Hypothesis Selection via the Scheffé Graph

M. Regehr[1], G. Kamath[1], A. F. Pour[1], D. P. Woodruff[2]

[1]University of Waterloo, Canada
[2]Carnegie Mellon University, Canada

We propose an algorithm with improved query-complexity for the problem of hypothesis selection under local differential privacy constraints. Given a set of k probability distributions Q, we describe an algorithm that satisfies local differential privacy, performs $\tilde{O}(k^{3/2})$ non-adaptive queries to individuals who each have samples from a probability distribution p, and outputs a probability distribution from the set Q which is nearly the closest to p. Previous algorithms required either $\Omega(k^2)$ queries or many rounds of interactive queries.

Technically, we introduce a new object we dub the Scheffé graph, which captures structure of the differences between distributions in Q, and may be of more broad interest for hypothesis selection tasks.

Goodness-of-fit Testing under Local Differential Privacy: From One (User), Many (Samples)

C. Canonne[1], A. Gentle[1], V. Singhal[2]

[1]The University of Sydney, Australia
[2]University of Copenhagen, Denmark

We initiate the study of distribution testing under user-level local differential privacy, where each of n users contributes m samples from the unknown underlying distribution. This setting, albeit very natural, is significantly more challenging that the usual locally private setting, as for the same parameter ε the privacy guarantee must now apply to a full batch of m data points. While some recent work consider distribution learning in this user-level setting, nothing was known for even the most fundamental testing task, uniformity testing (and its generalization, identity testing).

We address this gap, by providing (nearly) sample-optimal user-level LDP algorithms for uniformity and identity testing (i.e., one-sample goodness-of-fit). Motivated by practical considerations, our main focus is on the private-coin, symmetric setting, which does not require users to share a common random seed nor to have been assigned a globally unique identifier.

Permutation testing under local differential privacy

T. Berrett[1], A. Kent[1], Y. Yu[1]

[1]University of Warwick, United Kingdom

In this talk I will discuss recent work on two-sample testing under a local differential privacy constraint where a permutation procedure is used to calibrate the tests. While permutation testing is a classical resampling technique, popular due to its ease of implementation and uniform Type I error control, its use under local privacy constraints is complicated by the fact that access to the data is limited. In this work we design appropriate privacy mechanisms, both interactive and non-interactive, that allow for permutation tests. Our analysis shows that these lead to minimax optimal separation rates in both discrete and continuous settings, with interactive procedures being significantly more powerful. This is recent joint work with Alexander Kent and Yi Yu (https://arxiv.org/abs/2505.24811).

Advances in Object Data Analysis I

16:00 - 16:30

The Probability of the Cut Locus of a Fréchet Mean

s. huckemann[1]

[1]Universtity of Göttingen, Germany

We show that the cut locus of a Fréchet mean of a random variable on a connected and complete Riemanian manifold has zero probability, a result known previously in special cases (Le and Barden, 2014) and conjectured in general. The proof is based on first order and second order considerations, where the latter are based on a recent result by Générau (2020) on "Laplacians in the barrier sense". This generalizes to Fréchet p-means for p > 2. The former allow also to rule out stickiness on Riemannian manifolds, and for generalization to 1 <= p < 2, with a conjecture. We close with discussing and conjecturing extensions to noncomplete manifolds and more general metric spaces.

This is joint work with Alexander Lytchak.

Underlying papers are: Générau, F. (2020). Laplacian of the distance function on the cut locus on a Riemannian manifold.
Nonlinearity 33(8), 3928.
Le, H. and D. Barden (2014). On the measure of the cut locus of a Fréchet mean. Bulletin of the London Mathematical Society 46(4), 698–708.
Lytchak, A. and S. F. Huckemann (2025). Zero mass at the cut locus of a Fréchet mean on a Riemannian manifold. arXiv preprint arXiv:2508.00747

Extrinsic dynamical data analysis on manifolds

M. Pricop-Jeckstadt[1, 2], V. Patrangenaru[3], R. Paige[4]

[1]National University of Science and Technology POLITEHNICA Bucharest, Romania

[2]Center for Research and Training in Innovative Techniques of Applied Mathematics in Engineering “Traian Lalescu”, Romania

[3]Florida State University, United States

[4]Missouri University of Science and Technology, United States

In this talk, we introduce and develop an extrinsic dynamical stochastic analysis on manifolds. A novel Karhunene-Loeve transform for time-dependent random objects is presented starting from concepts of extrinsic covariance and extrinsic PCA available in [1]. The computational advantages of this approach is illustrated with the help of an application to a projective shape-from-motion problem (see [2]).

References

1. Wong K.C., Patrangenaru V., Paige R.L.,Pricop-Jeckstadt M., "Extrinsic Principal Component Analysis", arXiv}(2024).
2. Pricop-Jeckstadt M., Garthe A., Patrangenaru V., Paige R.L., "Projective Shape Analysis for Spatial Orientation in Virtual Environments", 2025 JSM Proceedings, https://doi.org/10.5281/zenodo.17234775 (2025).

Learning and Shape Analysis of Geometric Image Spaces of 3D Objects

B. Beaudett[1], S. Liang[2], A. Srivastava[3]

[1]Durham University, United Kingdom
[2]Florida State University, United States
[3]Johns Hopkins University, United States

Despite high-dimensionality of images, the sets of images of 3D objects have long been hypothesized to form low-dimensional spaces, the so-called manifold hypothesis. What is the nature of such spaces? How do they differ across objects and object classes? Answering these questions can provide key insights in explaining and advancing success of machine learning algorithms in computer vision. We investigate dual tasks -- learning and analyzing shapes of image spaces -- by revisiting a classical problem of manifold learning but from a novel geometrical perspective. We use geometry-preserving transformations to map the pose image spaces, sets of images formed by rotating 3D objects, to low-dimensional latent spaces. The pose image spaces of different objects in latent spaces are found to be nonlinear spaces which include manifolds as special cases. We then compare shapes of these sets for different objects using Kendall's shape analysis, modulo rigid motions and global scaling, and cluster objects according to these shape metrics. Interestingly, image spaces for objects from the same classes are frequently clustered together.

Contributed: Estimation, Semiparametrics and Prediction

16:00 - 16:20

Prediction of Non-Negative Variables under Misspecification and Nonstationarity

C. Francq[1,2], G. Sucarrat[3]

[1]CREST, France
[2]Universite de Lille, France
[3]BI Norwegian Business School, Norway

In empirical practice, it is commonly assumed that the maintained model is correctly specified. While this simplifies theory and interpretation, it is unlikely to hold in reality. We consider a broad class of predictive specifications for non-negative variables (e.g. volatility, spreads, volume and unemployment) that are optimal under Quasi Likelihood (QLIKE) loss. The predictive specifications need not be correctly specified, and the non-negative variables can be nonstationary. Examples of predictive specifications within our class include GARCH specifications and nonstationary seasonality predictors, and combinations thereof. Consistent and asymptotically normal estimation of the "pseudo-true" parameter is established. The finite sample properties of the estimator is studied by simulation, and an empirical application illustrates our results.

Semi-structured Additive Density Regression

M. Jung[1,2], D. Rügamer[2], S. Greven[1]

[1]Humboldt-Universität zu Berlin, Germany
[2]LMU Munich, Germany

We propose a novel semi-parametric regression framework for modeling the conditional density of a scalar response, which allows the entire density to depend on both tabular and non-tabular features (such as images or text). This is achieved by embedding structured additive density-on-scalar regression models into neural networks. These models are then extended to semi-structured additive density regression models by adding a suitable unstructured predictor that captures non-tabular feature dependencies through a deep neural subnetwork. By viewing probability densities as elements of a Bayes Hilbert space, this approach inherently preserves non-negativity and integration to unity under addition and scalar multiplication. Importantly, the additive model structure enables a meaningful ceteris paribus interpretation of estimated effects, in which density ratios take a role similar to well-known odds ratios. In simulation studies, our approach is benchmarked against an existing implementation of additive density regression for tabular features and exhibits comparable performance in most scenarios. Additionally, we demonstrate the practical utility of this framework through an application to Airbnb listing price data, where the picture of an Airbnb home, together with its type and review score, is used to model conditional price densities. The developed models provide an innovative approach to flexible yet interpretable conditional density modeling in complex, partially non-tabular data settings.

Generalised Exponential Kernels for Nonparametric Density Estimation

L. Craig[1], W. Barreto-Souza[1]

[1]University College Dublin, Ireland

This paper introduces a novel kernel density estimator (KDE) based on the generalised exponential (GE) distribution, designed specifically for positive continuous data. The proposed GE KDE offers a mathematically tractable form that avoids the use of special functions, for instance, distinguishing it from the widely used gamma KDE, which relies on the gamma function. Despite its simpler form, the GE KDE maintains similar flexibility and shape characteristics, aligning with distributions such as the gamma, which are known for their effectiveness in modelling positive data. We derive the asymptotic bias and variance of the proposed kernel density estimator, and formally demonstrate the order of magnitude of the remaining terms in these expressions. We also propose a second GE KDE, for which we are able to
show that it achieves the optimal mean integrated squared error, something that is difficult to establish for the former. Through numerical experiments involving simulated and real data sets, we show that GE KDEs can be an important alternative and competitive to existing KDEs.

Nonparametric estimation of evolutionary distances from incomplete genomic data

M. Garba[1]

[1]Northumbria University, United Kingdom

Ancestral sequence reconstruction — inferring the sequences of ancient proteins from the sequences of their present-day descendants — is a central problem in evolutionary biology with applications in protein engineering, vaccine design, and the study of disease evolution. Existing methods are limited in two important ways. First, they treat each position in a protein sequence as evolving independently of all others, ignoring the statistical dependencies between positions that arise from coevolution. Second, they condition on a single fixed evolutionary tree assumed to be fully known, ignoring both missing taxa — sequences absent from some species — and gene tree uncertainty — the fact that individual gene trees can differ in topology from the underlying species tree.

These limitations become increasingly consequential at genomic scale, where datasets contain thousands of gene families each with incomplete taxon coverage and potentially discordant evolutionary histories. No existing probabilistic framework adequately addresses scalability, missing taxa, and gene tree uncertainty within a single model.

We introduce a nonparametric estimator of the full pairwise evolutionary distance matrix from incomplete genomic data. For each pair of species, evolutionary distances are observed across all gene trees where both species are present and averaged to form a consensus distance matrix, with the variance across gene trees capturing gene tree uncertainty. Missing entries are estimated using the distances of co-occurring species. We establish the consistency of this estimator and characterise its bias and convergence rate as the number of gene trees increases theoretically.

The consensus distance matrix will be incorporated into a tree-structured Ornstein-Uhlenbeck Gaussian process variational autoencoder as part of a larger research programme aimed at scalable ancestral sequence reconstruction under sparse variational inference. We will present a proof of concept illustrating the proposed estimator on a small simulated dataset, with a discussion of planned extensions to large-scale genomic analyses.

Nonparametric Statistics on Complex Data

9:00 - 9:30

Minimax optimal rates of convergence in monotone shuffled and unlinked regression models under vanishing noise

C. DUROT[1], D. MUKHERJEE[2]

[1]Universite Paris Nanterre, France
[2]Boston University, France

Shuffled regression and unlinked regression represent intriguing challenges that have garnered considerable attention in many fields, including ecological regression, multi-target tracking problems, image denoising, and others. However, a notable gap exists in the existing literature, particularly in vanishing noise, i.e., how the rate of estimation of the underlying signal scales with the error variance. This paper aims to bridge this gap by delving into the monotone function estimation problem under vanishing noise variance, i.e., we allow the error variance to go to 0 as the number of observations increases. Our investigation reveals that, asymptotically, the shuffled regression problem is comparatively simpler than the unlinked regression; if the error variance is smaller than a threshold, then the minimax risk of the shuffled regression is smaller than that of the unlinked regression. On the other hand, the minimax estimation error is of the same order in the two problems if the noise level is larger than that threshold. Our analysis is quite general in that we do not assume smoothness assumptions on the link function (which only needs to be left-continuous, but may even have jump discontinuities). Because these problems are related to deconvolution, we also provide bounds for deconvolution in a similar context. Through this exploration, we contribute to understanding the intricate relationships between these statistical problems and shed light on their behaviors under vanishing noise.

A general theory for extremal regression in heavy-tailed models

Y. Abbas[1], A. Daouia[1], G. Stupfler[2]

[1]Toulouse School of Economics, France
[2]Angers University, France

Studying rare events at the heavy tails of conditional Pareto-type distributions, in the presence of high-dimensional covariates, is a burgeoning science with many applications in actuarial, financial and environmental risk management. The most prominent risk measures to quantify these events utilize conditional quantiles, Expected Shortfall, and expectiles at extreme levels. The few attempts to tackle this extreme value problem involve location-scale regression models with heavy-tailed noise. In this work, we employ a more flexible and complex model which better balances model generality with estimation efficiency. We develop a general theory that relies on residual-based estimators of the three regression risk measures at both intermediate and extreme levels, and fully explore their asymptotic behavior in generic settings. We also provide simple sufficient criteria for verifying our main high-level assumption, which facilitates the construction of weighted Gaussian approximations for the tail quantile residual process, ultimately ensuring the asymptotic normality of all produced extreme value estimators. We then apply this generic extremal regression framework to linear, nonlinear and nonparametric estimation scenarios. Simulations show the undeniable potential of our methodology for various distribution types, outperforming the best available competing estimation approaches. An application to real financial data further solidifies their dominance.

Automatic Debiased Machine Learning of Structural Parameters with General Conditional Moments

F. Argañaraz[1]

[1]Sciences Po Paris, France

This paper proposes a method to automatically construct or estimate Neyman-orthogonal moments in general models defined by a finite number of conditional moment restrictions (CMRs), with possibly different conditioning variables and endogenous regressors. CMRs are allowed to depend on non-parametric components, which might be flexibly modeled using Machine Learning tools, and non-linearly on finite-dimensional parameters. The key step in this construction is the estimation of Orthogonal Instrumental Variables (OR-IVs)—"residualized" functions of the conditioning variables, which are then combined to obtain a debiased moment. We argue that computing OR-IVs necessarily requires solving potentially complicated functional equations, which depend on unknown terms. However, by imposing an approximate sparsity condition, our method finds the solutions to those equations using a Lasso-type program and can then be implemented straightforwardly. Based on this, we introduce a GMM estimator of finite-dimensional parameters (structural parameters) in a two-step framework. We derive theoretical guarantees for our construction of OR-IVs and show √n-consistency and asymptotic normality for the estimator of the structural parameters. Our Monte Carlo experiments and an empirical application on estimating firm-level production functions highlight the importance of relying on inference methods like the one proposed.

Estimation of distribution functions, their jumps and interval probabilities under measurement error

C. Martins-Filho[1], K. Mynbaev[2], C. Brown[3]

[1]University of Colorado - Boulder, United States
[2]Kazakh-British Technical University, Kazakhstan
[3]University of Manchester, United Kingdom

In the classical error-in-measurement model, where a random variable Y is observed with error Z that has known distribution, we propose three new estimators for: a) the distribution function F_Y of Y at its points of continuity; b) interval probabilities F_Y(y)-F_Y(x) for x

Topological and geometric data analysis

9:00 - 9:30

A statistical framework for analyzing shape in a time series of random geometric objects

A. van Delft[1], A. Blumberg[1]

[1]Columbia University, United States

We introduce a new framework to analyze shape descriptors that capture the geometric features of an ensemble of point clouds. At the core of our approach is the point of view that the data arises as sampled recordings from a metric space-valued stochastic process, possibly of nonstationary nature, thereby integrating geometric data analysis into the realm of functional time series analysis. Our framework allows for natural incorporation of spatial-temporal dynamics, heterogeneous sampling, and the study of convergence rates. Further, we derive complete invariants for classes of metric space-valued stochastic processes in the spirit of Gromov, and relate these invariants to so-called ball volume processes. Under mild dependence conditions, a weak invariance principle in D([0,1]\times [0,R]) is established for sequential empirical versions of the latter, assuming the probabilistic structure possibly changes over time. Finally, we use this result to introduce novel test statistics for topological change, which are distribution-free in the limit under the hypothesis of stationarity. We explore these test statistics on time series of single-cell mRNA expression data, using shape descriptors coming from topological data analysis.

Asymptotic Normality of Topological Statistics in Dynamic Random Geometric Networks

C. Hirsch[1], N. Lundbye[1]

[1]Aarhus University, Denmark

We study dynamic random geometric networks generated by a marked Poisson process in $\mathbb R^d$, where vertices arrive over time and edges are formed through spatial interactions. This produces an evolving random intersection graph whose connectivity and higher-order structure change dynamically. We analyze topological characteristics of this network, including connected components and higher-order cycles encoded by persistent Betti numbers of the associated nerve complex. Under subcritical percolation assumptions, we prove a functional central limit theorem for suitably recentered and rescaled persistent Betti numbers in expanding observation windows. The fluctuations converge to a centered Gaussian field in a two-parameter Skorokhod space.

These results provide a fluctuation theory for time-evolving geometric networks and establish asymptotic normality for a broad class of topological network statistics. If time permits, we illustrate the theory with simulations and with data from black silicon materials. This talk is based on ongoing joint work with Nikolaj N. Lundbye.

Subsampling Euler Characteristic Curves for Imaging Data

B. Roycraft[1], A. Thomas[2]

[1]University of Florida, United States
[2]University of Iowa, United States

Euler characteristic curves (ECCs) provide functional summaries of imaging data by tracking topological changes in excursion sets across filtration thresholds. This presentation develops a subsampling-based framework for inference with ECCs under spatial dependence. Treating ECCs as function-valued statistics, we employ block-based subsampling to approximate their sampling distributions, enabling the construction of confidence bands and nonparametric tests for spatial homogeneity and stationarity. The framework integrates topological data analysis with resampling methods for dependent data and provides a flexible basis for inference in various imaging settings, with potential applicability to problems involving spatially structured data.

Universality of persistent homology over triangulable spaces

U. Lim[1], O. Bobrowski[1], P. Skraba[1]

[1]Queen Mary University of London, United Kingdom

Persistent homology is a notion of homology group for a finite metric space, and its statistics is of significant interest to data science and machine learning. In this work, we prove that persistent homology computed from Poisson point processes living on a compact C2-triangulable space satisfies a universal statistical law. Universality refers to the surprising phenomenon that the statistics of multiplicative persistence does not depend on the probability density generating the point process. Our result extends the result by Bobrowski and Skraba that was proven for the Euclidean space to the more general class of triangulable spaces. To achieve our aim, we use Freudenthal-Kuhn subdivisions and control interference effects across adjacent simplices. This is a joint work with Omer Bobrowski and Primoz Skraba.

Modern Network Analysis II

9:00 - 9:30

Randomization tests for model specification in causal inference under network interference

S. Tiwari[1], P. Basu[1]

[1]Indian School of Business Hyderabad, India

Analysis of experimental data becomes challenging when the underlying population is connected by a network. Exposure mapping is a common tool in the literature for defining and estimating spillover effects. These mappings reduce the dimensionality of the estimand, thereby facilitating identifiability. It is assumed that this mapping is correctly specified, leaving the choice of the exposure mapping to the analyst. This makes estimators of the spillover effect, such as the Horvitz-Thompson estimator, vulnerable to bias from model misspecification. Although these estimators have been shown to be robust to certain forms of controlled misspecification, there has been little methodological progress in empirically investigating appropriate exposure mappings. In this paper, we propose a novel design-based model specification framework for causal inference. Building on this, we develop a randomization testing procedure to assess the correct specification of an exposure mapping model under network interference. We provide theoretical guarantees for the asymptotic validity of the proposed testing procedure. We establish the favorable power properties of our method via an extensive simulation study and illustrate our method on a field experiment investigating the effect of anti-conflict norms among adolescents.

Learning the kernel in latent space network models

A. Modell[1]

[1]Imperial College London, United Kingdom

Latent space approaches to network modelling usually aim to learn a set of node-specific latent variables under the assumption that edge formation probabilities were determined by some fixed, known kernel function of them.
In this talk, I'll discuss the opposite problem of learning the kernel function given access to some known node-specific latent variables. I'll introduce a new supervised learning algorithm which learns the manifold geometry of high-dimensional node embeddings to directly model the kernel feature map, and I'll demonstrate its practical utility in an exploratory analysis workflow.

Bridging Theory and Practice: Statistical Inference of Latent Space Models for Networks

Y. He[1]

[1]University of Wisconsin-Madison, United States

Latent space models have been widely adopted in modeling network data. Developing statistical inference for estimated model parameters enables quantifying associated uncertainty and is pivotal for downstream tasks. Despite recent progress on statistical inference of maximum likelihood estimation, there exist crucial gaps between asymptotic theory and practical use. What is the relationship between theoretical maximum likelihood estimators and solutions from algorithms in practice? Can theoretical guarantees be obtained without unnecessary restrictions in existing algorithms? To address these foundational questions, we develop a unified analytical framework bridging theory and practice of conducting statistical inference under the latent space models. First, for the maximum likelihood estimator, we relax technical eigen-gap constraints in its existing asymptotic theory to broaden the applicability. Second, we overcome the dependence on unknown ground truth in prior algorithmic analysis by developing novel adaptive criteria and theoretical tools. For the widely used algorithm based on the projected gradient descent and the singular value thresholding, we explicitly connect the output to the maximum likelihood estimator without relying on unknown information, laying a solid foundation for practically useful and statistically principled statistical inference in network analysis.

Community detection via curvature gaps

Z. Lubberts[1], C. Li[1], M. Weber[2], Y. Tian[3]

[1]University of Virginia, United States
[2]Harvard University, United States
[3]Center for Systems Biology Dresden, Germany

We consider clustering in stochastic blockmodel graphs from the perspective of Ollivier's Ricci Curvature, an extension of Ricci Curvature on manifolds to this discrete setting. The gap between the distributions of edge curvatures for within-cluster edges and between-cluster edges allows us to identify these two groups of edges by their curvature, guaranteeing effective clustering. This curvature gap is studied under multiple signal strength regimes, identifying the limiting distributions and exploring the limits of curvature-based clustering. These distributional limits for edge curvatures are the first of their kind in the literature, and show that curvature-based clustering can be an effective competitor to traditional clustering methods, even in low signal strength settings.

Advances in Object Data Analysis II

9:00 - 9:30

Extrinsic Principal Component Analysis on a Projective shape space for 3D scenes derived from digital images

A. Algahtani[1]

[1]King Saud University, Saudi Arabia

In recent years, imaging data has emerged as one of the most dominant forms of data, driven largely by the widespread accessibility of smartphones. This article leverages statistical shape analysis to interpret such data. It introduces nonparametric methods to perform extrinsic principal component analysis on a projective shape space, denoted as P(R^(m+1)), which represents k-ads in general position within R^(m+1). Practical applications include 3D bioshape analysis derived from digital camera images. In this article, we consider 3D Projective Shapes and introduce the corresponding shape space as well as some statistics on this object space. We then compute extrinsic principal components for a concrete example. The general reference for this section is Mardia and Patrangenaru (2005) , Patrangenaru and Ellingson (2015) and Ka Wong et. al. .
keywords: Extrinsic mean; Extrinsic Principal Components ; Random Object;
Projective Space; Veronese Whitney embedding of a Projective projective space
1

Signed probability distributions: examples of numerical simulation

I. Podlubny[1]

[1]Technical University of Kosice, Slovakia

The notion of negative probability is almost one hundred years old, and so far some results have been obtained only in the direction of theoretical development of the notion of extended probability. However, there is still the strong need of computational methods and tools for applications, and this presentation is aimed on filling this gap. Several examples, including Feynman's problem, of numerical simulation of signed probability distributions are provided along with the results of simulation using the developed MATLAB toolbox. The presented methods and results might help using signed probabilities in the various fields of probability, statistics, decision making, finances, insurance, large language models and artificial intelligence, and other fields where the use of signed probability distributions can extend the current level of mathematical modeling.

Extrinsic Data Analysis on BHV4

V. Patrangenaru[1], T. Chen[2]

[1]FSU, United States
[2]Florida State University, United States

RNA and DNA evolutionary tree spaces with a given number of current species (see Billera-Holmes-Vogtmann [1]) are stratified spaces (see Patrangenaru and Osborne[3]). We investigate the extrinsic statistical analysis on the space of Billera-Holmes-Vogtmann tree space with four leaves (BHV4) based on its recently proposed novel representation (see Ordway et al [2])– as Spiky Excavated Dodecahedron (SPED). Due to the symmetry of the SPED, the Veronese – Whitney embedding can produce a natural extrinsic metric for a statistical analysis on BHV4. We demonstrate our new method on a yeast genome dataset.
[1] Billera, L., Holmes, S. & Vogtmann, K. (2001). Geometry of the Space of Phylogenetic Trees. Advances In Applied Mathematics. 27, 733-767 (2001)
[2] Garett Ordway, Tingan Chen, Vic Patrangenaru(2026). Continuing Investigations in the BHV Tree Space. DOI: 10.13140/RG.2.2.21497.94569 (2026)
[3] Vic Patrangenaru and Daniel Osborne (2026). Nonparametric Statistics on Stratified Spaces and Their Applications in Object Data Analysis, Chapman & Hall. ISBN9781138043138.

Nonparametric Methods for Dependent Data

9:00 - 9:30

Frequency Domain Resampling for Spatial Data

S. Bera[1], D. Nordman[2], S. Bandyopadhyay[1]

[1]Colorado School of Mines, United States
[2]Iowa State University, United States

In frequency domain analysis for spatial data, spectral averages based on the periodogram often play an important role in understanding spatial covariance structure, but also have complicated sampling distributions due to complex variances from aggregated periodograms. In order to nonparametrically approximate these sampling distributions for purposes of inference, resampling can be useful, but previous developments in spatial bootstrap have faced challenges in the scope of their validity, specifically due to issues in capturing the complex variances of spatial spectral averages. As a consequence, existing frequency domain bootstraps for spatial data are highly restricted in application to only special processes (e.g. Gaussian) or certain spatial statistics. To address this limitation and to approximate a wide range of spatial spectral averages, we propose a practical hybrid-resampling approach that combines two different resampling techniques in the forms of spatial subsampling and spatial bootstrap. Subsampling helps to capture the variance of spectral averages while bootstrap captures the distributional shape. The hybrid resampling procedure can then accurately quantify uncertainty in spectral inference under mild spatial assumptions. Moreover, compared to the more studied time series setting, this work fills a gap in the theory of subsampling/bootstrap for spatial data regarding spectral average statistics.

Blockwise Empirical Likelihood and Efficiency for Markov Chains

U.U. Müller[1, 2]

[1]Texas A&M University, United States
[2]OvGU Magdeburg, Germany

Suppose we observe an ergodic Markov chain on an arbitrary state space. The usual nonparametric estimator of a linear functional of the stationary distribution is the empirical estimator. If the stationary distribution obeys finitely many known linear constraints, we can improve the empirical estimator by empirical likelihood weights. Since the observations are dependent, an optimal choice of weights is determined by weighting averages over disjoint blocks of observations with slowly increasing length. We show that the improved empirical estimator is efficient. We also introduce two additively corrected empirical estimators that are asymptotically equivalent to the weighted empirical estimator, hence also efficient.

This talk is based on joint work with Anton Schick (Binghamton University) and Wolfgang Wefelmeyer (University of Cologne)

Likelihood-Based Nonparametric Causal Discovery under Latent Confounding

Y. Liang[1]

[1]Technical University of Munich, Germany

Causal discovery with latent confounding amounts to learning an acyclic directed mixed graph (ADMG) over observed variables and unmeasured confounders. Existing approaches often rely on discrete combinatorial search, which becomes computationally prohibitive for large-scale problems. Recent methods alleviate this challenge by introducing a differentiable acyclicity and bow-freeness constraint. However, these methods either assume linear structural equations or model nonlinear causal relationships involving both observed variables and confounders. The latter approximates intractable posteriors via variational inference, resulting in complex objectives and making performance more challenging. In this work, we propose a nonlinear causal model with correlated errors that encode latent confounding. We establish structural identifiability of the proposed model under bow-free graphs and parameter identifiability under ancestral graphs. This model yields a simple maximum-likelihood objective. We further develop a differentiable optimization scheme incorporating constraints for ADMG discovery. Experiments on synthetic and real-world datasets demonstrate that our method achieves competitive performance compared to relevant baselines.

Learning graphical models for multivariate nonstationary time series

S. Subbarao[1], N. Bolanos[2], J. Krampe[3]

[1]Texas A&M, United States
[2]Texas A&M University, United States
[3]Gesundheitsforen Leipzig, Germany

NonStGM is a general nonparametric graphical modeling framework for studying dynamic associations among the components of a nonstationary multivariate time series. The proposed framework captures conditional correlations/noncorrelations in the form of an undirected graph. In addition, to describe the more nuanced nonstationary relationships among the components of the time series, we incorporate within the graph architecture the notion
of conditional nonstationarity/stationarity.
This allows one to distinguish between direct and indirect nonstationary relationships among system components, and can be used to search for small subnetworks that
serve as the ``source'' of nonstationarity in a system.

In this talk, we describe a method for estimating the graph from data based on the Discrete Fourier Transform of the time series. We study the sampling properties of the estimator and describe a data adaptive method for selecting the tuning parameters. We illustrate the method with simulations and real data.

Advances in Directional Statistics

9:00 - 9:30

A pivotal Goodness-of-Fit test for the Spherical Cauchy family

D. Bolón[1], D. Paindaveine[1]

[1]Université Libre de Bruxelles, Belgium

The Spherical Cauchy family is a parametric family of distributions that is an extension of the multivariate t-distribution to directional data. One of its main properties is that it is a transformation model under the group of Möbius transformations in the sphere: given any two distributions belonging to the Spherical Cauchy family, there exists a Möbius transformation that maps one distribution onto the other. We use this property to introduce a general approach for testing whether the distribution of a given sample belongs to this family. We introduce a class of test statistics that are pivotal, i.e. their distribution under the null hypothesis does not depend on the value of the parameters. We then derive the limiting distribution of the proposed test statistics and show their consistency against general alternatives. Finally, we illustrate the behavior in practice of this procedure with a simulation study.

Semiparametric circular regression with flexible conditional densities

J. Ameijeiras-Alonso[1], I. Gijbels[2]

[1]Universidade de Santiago de Compostela, Spain
[2]KU Leuven, Belgium

This talk presents a semiparametric regression model for circular response variables depending on either linear or circular covariates. The conditional distribution is specified through a flexible parametric family of circular densities that allows for asymmetry and varying levels of peakedness around the modal direction. Both the modal direction and the concentration parameter are allowed to vary with the covariate and are estimated nonparametrically using local polynomial fitting with kernel weights.

We derive the asymptotic distribution of the estimators of the conditional modal direction and concentration. Based on these theoretical results, an expression for the optimal smoothing parameter is obtained and a practical data-driven bandwidth selector is proposed. The methodology is illustrated through an ecological application analyzing how the orientation of migratory birds varies with flight altitude and wind direction.

Circular single-index regression

M. Alonso-Pena[1], G. Claeskens[2], I. Gijbels[2]

[1]Universidade de Santiago de Compostela, Spain
[2]KU Leuven, Belgium

In this work, we propose a semiparametric single-index model for circular responses that provides flexible estimation while reducing the dimensionality of the problem. An estimation algorithm for the proposed model is developed, and the asymptotic properties of the estimators for both the parametric and nonparametric components are investigated. A data-driven bandwidth selection procedure is also introduced, and the finite-sample performance of the estimators is assessed through simulation studies. The methodology is illustrated through an application to the analysis of wave directions influenced by several maritime covariates at a strategically important and hazardous coastal location, where the proposed model is used to predict wave direction from the observed covariates.

Quasi-likelihood estimation for semiparametric circular regression models

A. Meilán-Vila[1], A. Gottard[2], A. Panzera[2]

[1]Universidad Carlos III de Madrid, Spain
[2]Università degli Studi di Firenze, Italy

Motivated by the need for flexible and interpretable models to handle circular data, this work introduces a semiparametric regression model for a circular response that can include both linear and circular covariates in its parametric and nonparametric components. The nonparametric component allows for modeling complex effects while avoiding restrictive parametric assumptions.
Rather than imposing a particular parametric distribution on the error term, we adopt a circular quasi-likelihood formulation, which is useful when the underlying distribution is unknown. Model estimation relies on a backfitting algorithm that iteratively updates the parametric and nonparametric components using circular partial residuals.
We establish the asymptotic properties of the resulting parametric and nonparametric estimators and assess their finite-sample performance through simulation studies. An application to the migratory patterns of willow warblers illustrates the advantages of the proposed approach for assessing genetic effects on circular responses and provides new insights into how specific genomic elements influence migratory behaviour.

Optimal Transport and Statistical Inference

9:00 - 9:30

The Optimal Transport Barycenter problem as a statistical toolbox

E. Tabak[1]

[1]New York University, United States

This talk will discuss a framework for data analysis based on the [Monge] optimal transport barycenter problem (OTBP), which removes from a set of variables X any variability that a set of covariates Z can explain. Solving this problem enables a number of tasks, including conditional density estimation and simulation, time-series analysis, data assimilation, factor discovery, consolidation of data bases and transfer learning. We will discuss this framework and its applications and describe an effective methodology for solving the data-driven OTBP through an adversarial characterization of independence.

Sparse regularized optimal transport without curse of dimensionality

A. González Sanz[1]

[1]Columbia University, United States

It is well known that optimal transport suffers from the curse of dimensionality: when the marginals are approximated by i.i.d. samples, the convergence of the empirical problem to the population one slows exponentially with dimension. Entropic regularization overcomes this barrier, achieving parametric sample complexity. This success rests on two pillars: smoothness of the dual potentials and strong concavity of the dual problem. But entropic OT has a price: it produces couplings with full support, leading to mass overspreading, and becomes numerically unstable when regularization is small. Quadratic regularization offers a compelling alternative. By penalizing the L^{2} norm instead of entropy, it yields sparse couplings and remains stable across regularization parameters. Yet its potentials lack smoothness, and its dual problem is not strongly concave. For these reasons, quadratic OT has been widely assumed to inherit the curse of dimensionality.

In this talk, I will show that this assumption is false. Quadratic OT achieves parametric sample complexity. I will present central limit theorems for its dual potentials, optimal couplings, and optimal cost. The proof introduces a new geometric perspective. Instead of relying on smoothness, we exploit the regularity of the support of the optimal coupling. We prove that its sections are Lipschitz and use VC theory to control statistical complexity. Along the way, we obtain gradient estimates of independent interest, including $C^{1,1}$ regularity of the population potentials.

Dimension reduction with Optimal Transport Barycenters for Transfer Learning - The Gaussian case.

M. Sued[1]

[1]Universidad de San Andrés - CONICET, Argentina

Dimensionality reduction is a fundamental tool in modern statistics. In this talk, we propose a method to extract d invariant features W = f(X) for predicting a response variable Y while avoiding confounding effects from variables Z. This plays a crucial role in transfer learning, since the conditional distribution of Z | Y may differ between source and target domains.
The approach is based on penalizing statistical dependence between W and Z conditional on Y . For computational tractability, we instead enforce indepen dence between W and a transformed variable T(Z,Y) arising from the Monge optimal transport barycenter problem for Z | Y . In the Gaussian case, these two formulations are equivalent. The resulting linear feature extractor admits a closed-form solution given by the first d eigenvectors of an explicit matrix.

Bridging sufficient dimension reduction and conditional optimal transport

K. Zeng[1], E. Bura[1]

[1]TU Wien, Austria

Most sufficient dimension reduction (SDR) methods rely on coverage conditions and primarily explore the first two conditional moments. However, the fundamental conditional independence assumption in SDR applies to the entire conditional distribution. We investigate the broader implications of the SDR assumption by exploring conditional optimal transport (COT) under a general nonlinear SDR framework.

Using the conditional Wasserstein space, we show that under mild regularity conditions, the non-identity component of the optimal triangular transport map is measurable with respect to the central sigma-algebra generated by the sufficient reduction. Consequently, this transport map can be factorised as a composition involving a reduced representation of the covariates. We further extend this analysis to the dynamic formulation of COT, proving that the optimal triangular velocity field inherently possesses this same structural property.

To translate these theoretical guarantees into a practical algorithm, we solve the dynamic COT problem under the nonlinear SDR assumption using conditional flow matching. By parameterising the velocity field to isolate the reduced representation, we evaluate the distance correlation between the true and learned sufficient predictors. Numerical experiments on synthetic datasets validate our theoretical results, demonstrating that our proposed method successfully captures the underlying central sigma-algebra.

Contributed: Applied Nonparametric Methods in Risk, Dependence and Decision

9:00 - 9:20

Robust estimation and inference for categorical data

M. Welz[1]

[1]University of Zurich, Switzerland

Empirical research in the social, health, and economic sciences is often based on categorical variables, such as questionnaire responses, self-reported health, or counting processes. Yet, just like in continuous variables, contamination might be present in such data, for instance (but not limited to) careless responses in questionnaires. Such contamination may cause severe biases in the commonly employed maximum likelihood (ML) estimation. However, robustifying estimation against contamination is challenging because categorical variables, by their very nature, cannot take arbitrarily large values, and may not even admit a numerical interpretation in the first place. Consequently, the extensive literature on outlier-robust M-estimation may not be applicable.

As a remedy, we propose a general framework for contamination-robust estimation of models for categorical data, called C-estimation ("C" for categorical). C-estimators achieve robustness by implicitly downweighting observations discrepant to a postulated model. The ensuing weights are useful practical tools for detecting possibly contaminated observations. Moreover, we show that C-estimators are consistent, asymptotically Gaussian, and fully efficient. The latter property starkly contrasts M-estimation, which is characterized by a fundamental tradeoff between robustness and efficiency. C-estimators avoid this tradeoff by exploiting the categorical nature of the data. Furthermore, C-estimators do not incur any additional computational cost and are therefore also attractive from a practical perspective.

In this talk, we present the theoretical properties of C-estimators as well as simulation studies to corroborate the derived theory. We also show how models for high-dimensional categorical data can be robustly estimated. As an empirical application, we consider a dataset of questionnaire responses from empirical psychology that are being modeled by a factor model. We show that estimation with C-estimators yields a substantially improved model fit compared to ML. It turns out that the observations downweighted by the robust estimator are those with inherently contradictory response patterns, indicating inattentive or careless responding.

Conditional copula estimation under two consecutive censored gap times

E. Strzalkowska-Kominiak[1]

[1]Universidad Carlos III de Madrid, Spain

We study nonparametric estimation of the conditional copula associated with two consecutive censored times in the presence of a continuous covariate. This framework allows us to model the bivariate conditional distribution of the two gap times while letting the dependence structure vary with the covariate. Building on recent work on conditional cumulative distribution estimation under censoring [1], we propose kernel-based estimators of the conditional copula and of related conditional dependence measures, and establish their asymptotic normality under suitable regularity assumptions. Their finite-sample behaviour is investigated by simulation. The methodology is illustrated with bladder cancer data. In this application, we also show how a single-index model [2] can be used to incorporate higher-dimensional covariate information.

Refereces:

[1] Strzalkowska-Kominiak, E., Molanes-López, E. M. and Letón, E. (2024). Non-parametric estimation of the covariate-dependent bivariate distribution for censored gap times. SORT 48, 183–208.

[2] Strzalkowska-Kominiak, E. and Cao, R. (2013). Maximum likelihood estimation for conditional distribution single-index models under censoring. Journal of Multivariate Analysis 114, 74–98.

Policy Learning with Compliance Guarantee

V. Marmer[1], T. Chan[1], K. Song[1]

[1]University of British Columbia, Canada

We study optimal policy learning where a policymaker (PM) uses data from a source population to design treatment assignments for a target population under a budget constraint. Because of the budget constraint, the PM needs to consider both treatment effects and individuals' incentives for treatment participation to minimize wasted resources. The main challenge is that treatment participation incentives may differ between the two populations. We develop a maximin approach that maximizes the minimum of the PM's expected objective across all possible incentive configurations. We show that this optimal policy learning problem can be reformulated using stochastic dominance constraints, where the optimal assignment prioritizes individuals most likely to comply with the treatment.

Nonparametric Tail-Risk Connectedness in Eurozone Energy-Related industries

J. Ascorbebeitia[1], J. Barberá Vilaplana[2], S. Orbe Mandaluniz[1]

[1]University of the Basque Country UPV/EHU, Spain
[2]KMPG-Spain, Spain

This paper investigates the evolution of connectedness and systemic tail-risk spillovers among Eurozone energy-related stocks over the period 2015-2025. We select the most capitalized firms in the Energy, Utilities, and Basic Materials sectors and construct a network that maps the interactions among energy producers, distributors, and consumers. Methodologically, we employ the semiparametric TENET framework, which relies on nonparametric tail dependence estimation to capture nonlinear and asymmetric propagation mechanisms that standard parametric models fail to detect.
The analysis is conducted at three complementary levels: system-wide, by industry, and at the firm level. The results reveal three pronounced peaks of connectedness, each associated with major disruptions to European energy markets: the US-China trade conflict, the COVID-19 outbreak, and the Russia-Ukraine war. Across the sample, the upstream Energy sector consistently emerges as a net transmitter of extreme spillovers, whereas Utilities and Basic Materials act as net receivers of tail risk. Finally, the strong within-industry linkages highlight the predominance of intra-industry contagion relative to cross-industry effects.

Random Coefficient extensions for integer autoregressive models: a non-parametric approach

D. Karlis[1]

[1]Athens University of Economics and Business, Greece

In this talk, we use the class of integer autoregressive (INAR) time series models, widely used to model time series of counts. They extend the idea of classical time series for continuous data in an admissible way to keep the discrete nature of the data. Random Coefficient INteger AutoRegressive (RCINAR) models constitute an important extension of the simple INAR model when the thinning parameters are considered to be random. In the present paper, we propose the case that the thinning parameter follows a discrete distribution with positive probability at a finite number of points. We provide an EM algorithm to estimate the model while we link the model to the non-parametric Maximum Likelihood estimate of the mixing distribution. The ideas are then extended to the bivariate case where the thinning parameters are now a matrix of random variables. The finite mixture representation helps a lot to account for the extra variability but also extra correlation to the model.

Contributed: Goodness-of-Fit, Smoothing and Methodological Foundations

9:00 - 9:20

The Weak-Feature-Impact Phase Transition of the NPMLE in Monotone Binary Regression

D. Kieffer[1], A. Rohde[1]

[1]Albert-Ludwigs-Universität Freiburg, Germany

The nonparametric maximum likelihood estimator (NPMLE) in monotone binary regression models is considered here when the impact of the features on the labels is weak, where weakness is colloquially understood as "close to flatness" of the feature-label relationship x -> P(Y=1 | X=x). Statistical literature provides limiting distributions of the NPMLE for the two extremal cases: If the feature-label relation is strictly monotone and sufficiently smooth, then it converges at a nonparametric rate pointwise and in L^1 with scaled Chernoff-type and Gaussian limiting distribution, respectively, and it converges at the parametric n^1/2-rate if the underlying relation is flat. To explore the distributional transition of the NPMLE from the nonparametric to the parametric regime, we introduce a novel mathematical scenario. New restricted minimax lower bounds and matching pointwise and L^1-rates of convergence of the NPMLE in the weak-feature-impact scenario together with corresponding limiting distributions are derived. They are shown to exhibit an elbow and a phase transition, solely characterized by the level of feature impact.

Pinelis' Inequality for Weighted Sums of Random Vectors: Statistical Motivations, Proof Arguments, and Conjectured Extensions

E. Gautherat[1]

[1]Université de Reims, laboratoire CRIEG-REGARDS, France

Concentration inequalities for sums of random vectors play a central role in nonparametric statistics, particularly in the analysis of resampling procedures, bootstrap consistency, and randomization tests. Among these, a remarkable result due to Pinelis controls the tail of the L^2-norm of a weighted sum of arbitrary random vectors — where the weights are Rademacher random variables — by the corresponding tail for the same vectors weighted by standard Gaussian variables, up to an explicit multiplicative constant.

This comparison inequality is powerful and can be obtained without any moment conditions, no independence beyond the weights, no geometric assumptions on the space of the vectors . This generality makes it particularly attractive in a nonparametric statistical settings, when the underlying distribution of the random vectors is unknown.

In this talk, we revisit Pinelis' result from a statistical perspective. We first motivate why such a Rademacher-to-Gaussian comparison is natural and useful in nonparametric inference. We then revisit the core arguments underlying Pinelis' proof, clarifying the key analytical mechanisms at play: in particular, the role of hypercontractivity, symmetrization, and comparison principles for stochastic processes. This re-examination reveals that some conditions assumed in the original result may not be strictly necessary, and suggests that the inequality could potentially hold in a broader framework.

To support this conjecture, we present simulation evidence illustrating the inequality beyond its current proven regime, suggesting that the explicit constant remains valid under weaker assumptions. These numerical experiments motivate future theoretical work aimed at establishing the conjecture rigorously.

Goodness-of-fit on metric spaces using distance profiles

D. Serrano[1], E. García-Portugués[1], I. Van Keilegom[2]

[1]Universidad Carlos III de Madrid, Spain
[2]KU Leuven, Belgium

We propose a general goodness-of-fit framework for distributions on separable metric spaces. Under suitable identifiability conditions, probability distributions are characterized by distance profiles, which motivates their use in goodness-of-fit testing, for simple and composite null hypotheses. For composite null hypotheses, parameter estimation is incorporated via a Bahadur-type expansion, and the asymptotic distribution of the empirical process for distance profiles is obtained under the null. We define test statistics based on this empirical process and derive their asymptotic null distributions. We further study the behavior of the proposed tests under fixed and local alternatives, establishing consistency results. The methodology is illustrated with simulation studies and real-data experiments.

Robust smoothing splines in the presence of discontinuities

H. Jankowski[1], K. Ramsay[1], J. Diaz-Rodriguez[1]

[1]York University, Canada

We study the estimation of a function which is smooth (twice differentiable) except at a finite number of locations, at which discontinuities exist. Our approach is robust, in that we use the Huber loss function instead of the more standard least squares loss. We show both theoretical and computational properties of the resulting estimator.

Neyman-orthogonal goodness-of-fit tests for generalized partially linear models

R. Costa-Miranda[1], C. Heumann[2], W. González-Manteiga[3], R. Gaio[1]

[1]University of Porto, Portugal
[2]Ludwig Maximilian University of Munich, Germany
[3]University of Santiago de Compostela, Spain

Many well-established goodness-of-fit hypothesis tests rely on data-dependent asymptotic distributions of the test statistics and are locally sensitive to model estimation errors. When contrasting composite null hypotheses, this sensitivity typically requires bootstrap procedures involving repeated estimations of the model for the bootstrapped data – a process which can be computationally intensive when dealing with models that are more complex or depend on a large number of covariates. For testing conditional moment restrictions with finite-dimensional nuisance parameters, in parametric models, these limitations have recently been overcome by integrating squared Neyman orthogonal function-parametric gaussian processes. This improvement introduces lack of sensitivity to changes in the model parameters around their true values, thus allowing fast-bootstrap procedures where the model does not need to be re-estimated.
In this talk, we extend the test to regression models with infinite-dimensional nuisance parameters, focusing on generalized partially linear models. We propose a hypothesis test based on local polynomial finite-dimensional approximations of the null conditional moment restrictions. The test was implemented using a fast-bootstrap procedure where Gaussian processes with zero mean and distance covariance kernels were considered.
We compare the finite-sample performance (n = 50, 100, 500) of the proposed test against preceding kernel-based counterparts, in Gaussian, logistic and Poisson partially linear models. Simulation results were based on 1000 replications; for each, the bootstrap process was repeated 999 times. The empirical rejection probabilities were close to the nominal level (alpha = 0.01, 0.05, 0.10) in all considered scenarios. While simulation results indicated a reduction in power on the three analysed deviations from the null, this novel approach offered computational efficiency and robustness to estimation, making it a promising procedure for checking goodness-of-fit in generalized partially linear models.

Keynote Talk

11:30 - 12:30

Predictive Inference in Nonparametric Regression: Model-free Bootstrap, Conformal Prediction, and Pertinent Prediction Intervals

D.N. Politis[1]

[1]University of California at San Diego, United States

Predictive inference in a general regression setting is attracting progressively more attention in the big-data era. Given a nonparametric model driven by i.i.d. errors, model-based bootstrap procedures can be devised to yield prediction intervals for a future response associated with a regressor value of interest. Ideally, the bootstrap procedure will be designed to mimic/incorporate the variability of all estimated components; in this case, the resulting prediction intervals are called `pertinent'. If a model equation is not available, there are three possible avenues: Model-free bootstrap, conformal prediction, and quantile estimation. The three approaches will be contrasted via theoretical analysis as well as numerical experiments with a focus on conditional coverage; in particular, three notions of conditionality will be described that are nested in terms of increasing strength. Under mild conditions, we can show that the Model-free bootstrap yields prediction intervals with guaranteed better conditional coverage compared to quantile estimation using any one of the three notions of conditionality. We also extend the concept of pertinence of prediction intervals to the nonparametric regression setting, and give concrete examples where its importance emerges under finite sample scenarios. Time-permitting, an application of Model-free and conformal prediction to Markov sequences will be given.

[Joint work with Yiren Wang, Dehao Dai, and Kejin Wu.]