Zaheer Ahmed

ahmed@ifado.de

Leibniz Research Center for Wroking Enviornment and Human Factors (IfADo) of TU Dotmund Germany

Talk: *Uncovering Latent Structures in Large-Scale Cognitive Conflict Data Using a Novel Probabilistic Model-based Clustering Approach*

Abstract: Uncovering latent structures and interactions in cognitive data is essential for advancing our understanding of complex cognitive processes in humans. Despite the abundance of data from diverse studies, effective approaches for disentangling these latent patterns are yet to be developed. We employed a recently established probabilistic model-based clustering approach to uncover latent structures in an extensive dataset, the Dortmund Vital Study (DVS, ClinicalTrials.gov Identifier: NCT05155397), a prospective study on healthy cognitive aging, which includes data recorded from about 600 participants in various cognitive tasks. Beyond uncovering latent structures, this method is suitable for identifying and eliminating outlier participants, indicating that outcome variables might differ at the level of various participant groups identified by the analysis. The main advantage of the proposed analytical tool is that it can simultaneously identify latent groups of participants and tasks, offering a data-driven, more comprehensive understanding of latent structures. By capturing these latent structures and participant groups, the probabilistic model-based clustering method allows for a more sophisticated analysis of how participants' cognitive profiles relate to the tasks or conditions they encounter.

Keywords: Probabilistic model-based clustering, Latent class analysis, Outlier detection, Cognitive tasks, Cognitive profiles

Abdulrahman Aldawsari

abd.aldawsari@psau.edu.sa

Prince Sattam bin Abdulaziz University

Talk: *A Predictive Bootstrap Technique for Double-Censored Data*

Abstract: Bootstrap methods are widely used because they are simple to implement and have strong theoretical and practical performance. We introduce a novel predictive bootstrap technique for double-censored data, designed specifically for predictive inference within a fully parametric setting. The proposed method is evaluated using simulation scenarios commonly employed to assess the predictive performance of bootstrap procedures. The proposed approach is compared with Efron's classical bootstrap in terms of prediction interval coverage probabilities. In addition, a global measure of coverage accuracy is examined, and a chi-squared goodness-of-fit test is applied to provide a broader evaluation of performance. The results show that the proposed procedure achieves strong predictive performance, which is consistent with its explicitly predictive construction.

Keywords: Bootstrap methods, Double-censored data, Predictive inference, Coverage accuracy, Prediction intervals

Abdulmajeed Alharbi

abdulmajeed1@KSU.EDU.SA

Department of Statistics and Operations Research, College of Science, King Saud University, Riyadh, Saudi Arabia

Talk: *A Predictive Approach for Classifying Future Observations into Ordered Groups*

Abstract: In diagnostic testing and decision-making problems, classification often relies on selecting suitable decision thresholds. The choice of these thresholds plays an important role in how observations are assigned to different groups and can substantially influence classification performance, particularly when the groups have a natural ordering.

This work develops a predictive approach for selecting thresholds that focuses on classification performance for future observations while relying on minimal distributional assumptions. The proposed method explicitly incorporates multiple future observations and can be applied in a range of classification settings, including those involving more than two groups.

Illustrative examples and extensive simulation studies compare the proposed method with commonly used threshold selection approaches. The results show that the proposed method consistently achieves a higher number of correctly classified future observations across groups, particularly in more challenging scenarios.

Keywords: Predictive inference, Classification, Ordered groups, Future observations, Thresholds selection

Anas Alharshan

af.alharshan@psau.edu.sa

Prince Sattam bin Abdulaziz University

Talk: *Probability Distributions for Component Failures in Multi-Type Systems*

Abstract: This research develops a formal probability distribution for the number of failed components in systems comprising multiple distinct types. Although predicting failure counts is essential for spare part logistics and reliability centering, the existing literature is largely confined to single-component types or restricted architectures. Furthermore, current models often neglect the "survival history'' of the system: the fact that the distribution of failed components at the moment of system collapse is contingent on the system having been operational up to a specific time t.

By leveraging the survival signature framework, we derive a generalized expression for the joint probability of failed components across multiple types at the exact moment of system failure. This model explicitly accounts for the system's functional status at time t and provides the distribution for failures occurring within specific future time intervals. In this work, we demonstrate the derivation of these probabilistic results and provide illustrative examples of their application in complex, multi-type system configurations.

Keywords: Reliability, coherent system, survival signature, spare parts, failed components

Nicolai Amann

nicolai.amann@univie.ac.at

University of Vienna

Talk: *Conditional validity and a fast approximation formula of full conformal prediction sets*

Abstract: Prediction sets based on full conformal prediction have seen an increasing interest in statistical learning due to their universal marginal coverage guarantees and versatility across a wide range of applications. However, practitioners have refrained from using it in real-world implementations for two reasons: Firstly, it comes at very high computational costs, exceeding even that of cross-validation. Secondly, an applicant is typically not interested in a marginal coverage guarantee which averages over all possible (but not available) training data sets, but rather in a guarantee conditional on the specific training data.

We tackle these problems by, firstly, showing that full conformal prediction provides reliable prediction sets conditional on the training data if the conformity score is stochastically bounded and satisfies a stability condition. Secondly, we propose an approximation for the full conformal prediction set that has asymptotically the same training conditional coverage as full conformal prediction under the stability assumption derived before, and can be computed more easily. Furthermore, we show that under the stability assumption, another variant, n-fold cross-conformal prediction, also has the same asymptotic training conditional coverage guarantees as full conformal prediction. If the conformity score is defined as the out-of-sample prediction error, our approximation of the full conformal set coincides with the symmetrized Jackknife.

Keywords: Uncertainty quantification, Conformal prediction, High-dimensional statistics, Conditional predictive inference, Algorithmic stability

Lujia Bai

lujia.bai@rub.de

Ruhr University Bochum

Talk: *Validating spatial-temporal separability for stationary processes*

Abstract: A crucial assumption to reduce computational complexity in spatial-temporal data analysis is separability, which factors the covariance structure into a purely spatial and a purely temporal component. In this paper, we develop statistical inference tools for validating this assumption for a second-order stationary process under both domain-expanding-infill asymptotics

and domain-expanding asymptotics. In contrast to previous work on this subject, the methodology does neither require the assumption of normal distributed data, nor uses spectral methods. Our approach is based on nonparametric estimates of measures for the deviation between the covariance matrix and separable approximations, which vanish if and only

if the assumption of separability is satisfied. We derive the asymptotic distributions of appropriate estimators for these measures with non-standard limiting distributions and use these results to develop inference tools for validating the assumption of separability. More specifically, we derive confidence intervals for the deviation measures, tests for the hypothesis of exact separability, and for the hypothesis that the deviation from separability is smaller than a pre-specified threshold.

Keywords: Spatial-temporal data, U -statistics, Central limit theorem for dependent data, Best rank-one approximation, Non-parametrics

Julius Baumhakel

baumhakel@tugraz.at

Graz University of Technology

Talk: *Testing for Isotropy of Function-Valued Random Fields*

Abstract: We propose a new nonparametric approach for testing isotropy, i.e. invariance in distribution under rotations around the origin, 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.

Keywords: Spatial Statistics, Nonparametric Statistics, Functional Data, Isotropy Testing, Directional Statistics

Fabian Bleile

f.bleile@tum.de

TUM, MCML

Talk: *Efficient Learning of Stationary Diffusions with Stein-type Discrepancies*

Abstract: Learning a stationary diffusion amounts to estimating the parameters of a stochastic differential equation whose stationary distribution matches a target distribution. We build on the recently introduced kernel deviation from stationarity (KDS), which enforces stationarity by evaluating expectations of the diffusion's generator in a reproducing kernel Hilbert space. Leveraging the connection between KDS and Stein discrepancies, we introduce the Stein-type KDS (SKDS) as an alternative formulation. We prove that a vanishing SKDS guarantees alignment of the learned diffusion's stationary distribution with the target. Furthermore, under broad parametrizations, SKDS is convex with an empirical version that is ϵ-quasiconvex with high probability. Empirically, learning with SKDS attains comparable accuracy to KDS while substantially reducing computational cost and yields improvements over the majority of competitive baselines.

Keywords: causal models, SDEs, kernels, stationary diffusions, causality

Rajdeep Brahma

rbrahma2@illinois.edu

University of Illinois Urbana-Champaign

Talk: *Two-Sample Hypothesis Testing for Subspace Equality in Network Data*

Abstract: In many settings one is often interested in determining whether two networks share some joint structural connectivity patterns such as communities. However, while communities may be shared across networks, edge probabilities may differ significantly. Therefore, in this paper we consider testing a general null hypothesis that two networks have the same underlying subspace, which in particular includes the setting that communities are the same for either stochastic blockmodels or mixed-membership stochastic blockmodels (even if edge probabilities are different). We propose a test statistic based on the Frobenius norm of the difference of the leading subspace projection matrices, and we prove that our test statistic, after appropriate centering and scaling, converges in distribution to a Gaussian random variable as long as the average expected degree grows at least logarithmically in the number of vertices. We then provide estimators for the asymptotic mean and variance and show consistency under a stronger signal condition, and we give the local power of our test when the networks are sufficiently dense. Our theoretical results are based on a limit theorem for the projection difference of empirical and true eigenvectors which can also be viewed as the one-sample version of our test statistic, and this result may be of independent interest. We demonstrate our results through numerical simulations and an application to US Flight data.

Keywords: Low-rank matrices, Subspace perturbation, Network data, Signal-plus-noise Model, Random Matrices

Johannes Brutsche

johannes.brutsche@stochastik.uni-freiburg.de

University of Freiburg

Talk: *Self-organized criticality under arbitrary sampling frequencies*

Abstract: We derive the rate of convergence and limiting distribution of the profile MLE for the level of self-organized criticality of an oscillating Brownian motion (OBM) under equidistant sampling schemes, including low-frequency asymptotics for which the observations are realizations of a null recurrent Markov chain.

The derived nonstandard Cox-type limit, caused by the discontinuity of the profile likelihood in the parameter of interest, turns out to be independent of the sampling frequency while the rate of convergence interpolates between that for the low- and the high-frequency observation scheme. The limit depends on the true level of self-organized criticality as well as the unknown volatility levels and this dependence is shown to be continuous w.r.t. the topology of weak convergence, enabling statistical inference.

Our proof relies crucially on the observation that the OBM, if started in the true parameter and shifted by the latter, possesses a desirable distributional self-similarity. This relates the long-term behavior of the Markov chain to the infill asymptotics in this special case. On basis of the strong Markov property, the deficency of starting in the unknown true parameter is finally overcome by a coupling argument.

Keywords: Oscillating Brownian motion, Null recurrent Markov chain, Profile maximum likelihood estimation, Cox-type limiting distribution, Self-similarity

Ferdinand Buchner

f.buchner@tum.de

Technische Universität München

Talk: *Identifying risky flight landings using multivariate vine copula based regression models*

Abstract: Runway overruns remain a significant safety concern in commercial aviation, particularly during landing, which consistently accounts for a notable share of serious incidents. Identifying landings with elevated overrun risk is challenging because landing performance is driven by several interacting phases and operational factors. Existing statistical models often consider only a single rollout component or classify flights into overrun versus non overrun events, limiting their ability to capture dependencies within the landing process.

This work proposes a trivariate distributional vine copula regression framework that jointly models three key components of landing performance: the touchdown distance, the distance until the aircraft reaches controllable speed, and the average deceleration. By modeling their joint distribution and incorporating meteorological, aircraft specific, and operational covariates, the approach provides a more comprehensive representation of landing dynamics and enables probabilistic assessments of overrun risk.

The method is evaluated using QAR data from 711 landings of identical cargo aircraft on a single runway under dry conditions. The dataset exhibits nonlinear relationships and non Gaussian dependence structures, motivating the use of flexible multivariate modeling techniques such as vine copula based regression. The proposed framework offers a more quantitative and operationally relevant basis for identifying flights at increased overrun risk.

Keywords: vine copula, vine copula regression models, multivariate regression models, flight data, runway overrun

Yo Joong Choe

yojoong.choe@insead.edu

INSEAD

Talk: *Betting on Bets: Anytime-Valid Tests for Stochastic Dominance*

Abstract: How can we monitor, in real time, whether one uncertain prospect has any upside over another? To answer this question, we develop a novel family of sequential, anytime-valid tests for stochastic dominance (SD), a widely used framework for comparing entire distribution functions. We first derive powerful, nonparametric e-processes that quantify evidence against the null hypothesis that one prospect is dominated by another. For first-order SD, these e-processes are shown to be growth-rate optimal and to achieve power one asymptotically. The approach further generalizes to sequential testing for SD beyond the first order, including any higher-order SD. Finally, we sketch the complementary and challenging problem of testing the non-SD null hypothesis, which asks whether a prospect has a definite upside, and describe the conditions under which we can derive a nontrivial anytime-valid test. We validate our approaches in simulations as well as real-world player comparison tasks using sports data.

This talk is based on a joint work with Sebastian Arnold (CWI Amsterdam), Marco Scarsini (Luiss University), and Ilia Tsetlin (INSEAD).

Keywords: stochastic dominance, sequential anytime-valid inference (SAVI), e-process, testing-by-betting, decision theory

jisoo choi

ddsy999@gmail.com

Seoul national university

Talk: *Dual-Homotopy Framework for Constrained EM Algorithm*

Abstract: We propose a new constrained EM algorithm that is applicable to general constrained estimation problems. This proposed method is based on a novel framework, the `dual homotopy framework,' which combines deterministic annealing EM with a barrier-based optimization, enabling stable estimation under parameter constraints. Building on this framework, we further introduce an adaptive constrained EM algorithm that preserves likelihood monotonicity, regardless of the underlying distributional form or the specific structure of the constraints.

Through simulation studies using two models and a real-data analysis under parameter constraints, we demonstrate that the proposed algorithm yields more stable and accurate estimates than existing methods, including the standard EM algorithm.

Keywords: Constrained EM algorithm, Barrier optimization, Likelihood monotonicity, Deterministic annealing, Dual homotopy

BRENTON CLARKE

b.clarke@murdoch.edu.au

Murdoch University

Talk: *Reconciling Least Squares and Robustness Approaches in the Two-Way Layout: An Example*

Abstract: The competition between least squares and robust methodologies documented in many books including Hampel, F.R.,et al.(1986) and Clarke,B.R.(2018), Wiley publishers, has led many either to embrace least squares analysis or instead embrace robust methods, and bury their heads in the sand when it comes to the alternative methodology. In 1983 the author gave a justification of the influence curve, later called the influence function of Hampel, that was revealed through conditions sufficient for Fréchet differerentiability of M-estimators, a critical requirement for robustness. An example of a Fréchet differentiable and robust estimator uses the Tukey bi-square estimator of location, coupled with the MADN estimator of scale, both later being replaced by the MM-estimator. The practical example of the two-way layout of Fisher(1949) illustrates least squares methodology in Analysis of Variance specifically for the randomized complete block design. In this talk we compare least squares and MM-estimation using the Tukey bi-square. Why this is enlightening is revealed in Clarke and Godolphin (1992, pp. 2520-2522) Commun. in Statist, 21(9),2501-2526, this data being subject to possible swapping of observations as suggested by Daniel, C. (1976, p.159) Wiley, New York, and also, large residuals for the variety Trebi, not necessarily outliers, but leading one to doubt least squares analysis, without further remodeling. We show both least squares and robust inferences essentially agree.

Keywords: Influence Function, Fréchet differerentiability, M-estimators, Least squares, Outliers

Inez De Batselier

inez.debatselier@student.uhasselt.be

Hasselt University

Talk: *A flexible parametric regression approach to handle left-censored outcome and covariate information*.

Abstract: Left-censored data, where observations are only known to fall below a threshold, are common in environmental and biomedical sciences. While methods for censored responses are well-established, handling censored covariates remains challenging, often relying on biased “naive” substitutions or restrictive parametric assumptions.

We propose a flexible parametric regression framework handling left censoring in both response and a covariate. To relax rigid assumptions of existing methods (e.g. Tran et al., 2021), we model the covariate distribution using a power piecewise exponential distribution with fixed cut-off points. Offering semi-parametric flexibility while maintaining parametric tractability. Under mild regularity conditions, we establish model identifiability.

We introduce one-stage and two-stage maximum likelihood estimators for the regression and covariate parameters. The one-stage approach estimates all parameters simultaneously, maximising efficiency. The two-stage approach first estimates covariate distribution parameters, then plugs these into the full likelihood to estimate regression parameters, reducing dimensionality and computational burden. Consistency and asymptotic normality are established for both procedures.

Finite-sample performance and robustness are assessed through extensive simulations. We apply this framework to an immunological study of transplacental antibody transfer, quantifying the relationship between maternal and infant antibody levels.

Keywords: Left censoring, Censored covariates, Power piecewise exponential distribution, Maximum likelihood estimation, Asymptotic theory

Maike de Jongh

m.c.dejongh@utwente.nl

University of Twente

Talk: *Mitigating large-scale avalanches in the Abelian sandpile model through strategic interventions*

Abstract: The Abelian sandpile model is a classic example of a system in which critical behavior emerges independently of external parameters, a phenomenon known as self-organized criticality. In this model, sand grains are added randomly to vertices of a graph. Whenever a vertex collects too many sand grains, it topples and redistributes its mass to its neighbors, a mechanism that may trigger large cascading events. Such avalanches are frequently used to model catastrophic phenomena in real-world systems, such as earthquakes, financial market crashes, and forest fires. Motivated by these applications, we study intervention strategies aimed at mitigating avalanche sizes in the Abelian sandpile model. We present theoretical results on how removing sand grains at designated locations affects the expected avalanche size. In addition, we evaluate several heuristic strategies and investigate their long-term impact on the evolution of the system. Our findings demonstrate the potential of strategic interventions in stabilizing self-organized critical systems.

Keywords: Abelian sandpile model, Control, Self-organized criticality, Probability theory, Markov decision theory

Nicolas Dietrich

nicolaspascal.dietrich@plus.ac.at

University of Salzburg

Talk: *Nonparametric Estimation of Archimax Copulas and Consequences for Modeling Dependence*

Abstract: Modeling the dependence structure of extreme events has been a central topic in extreme value theory. Archimax copulas provide a flexible framework capturing both, asymptotic dependence between extreme events, and intermediate-level dependence. They are characterized by an Archimedean generator ψ and a Pickands dependence function A. In this work, we first investigate identifiability in the Archimax class. Building on these results, we propose a novel nonparametric estimator for Archimax copulas C, based on the Kendall distribution function F_K of C and the CFG estimator A_n^{CFG} of A. We establish consistency of the proposed estimator under mild regularity conditions and examine its performance. Finally, we demonstrate how it can be used to estimate dependence when the underlying data-generating process is assumed to follow an Archimax copula.

Keywords: Copula, Extreme value theory, Archimax copula, Nonparametric estimation, Dependence modeling

Frieder Droullier

frieder.droullier@gmail.com

Department of Mathematics and Statistics, University of Konstanz

Talk: *On Non-Markovian INAR(1) Processes*

Abstract: We study INAR(1) processes with random coefficients. INAR(1) models were originally introduced by McKenzie (1985) and Al-Osh and Alzaid (1987). They are based on the binomial thinning operator and are Markovian. Integer-valued processes have many applications, such as high-frequency trading and epidemiology. We introduce random coefficient INAR(1) processes with long memory and examine their characteristics. Estimation and prediction is considered. Asymptotic rates of convergence are derived.

The methods are illustrated by applications to trading volumes of Microsoft and Netflix.

This is joint work with Jan Beran.

Keywords: Integer-valued time series, INAR Process, Long Memory, Random coefficient process, Gaussian subordination

Chenguang Duan

duan@igpm.rwth-aachen.de

RWTH Aachen University

Talk: *Inference-Time Alignment for Diffusion Models via Variationally Stable Doob's Matching*

Abstract: Inference-time alignment for diffusion models aims to adapt a pre-trained reference diffusion model toward a target distribution without retraining the reference score network, thereby preserving the generative capacity of the reference model while enforcing desired properties at the inference time. A central mechanism for achieving such alignment is guidance, which modifies the sampling dynamics through an additional drift term. In this work, we introduce variationally stable Doob's matching, a novel framework for provable guidance estimation grounded in Doob's h-transform. Our approach formulates guidance as the gradient of logarithm of an underlying Doob's h-function and employs gradient-regularized regression to simultaneously estimate both the h-function and its gradient, resulting in a consistent estimator of the guidance. Theoretically, we establish non-asymptotic convergence rates for the estimated guidance. Moreover, we analyze the resulting controllable diffusion processes and prove non-asymptotic convergence guarantees for the generated distributions in the 2-Wasserstein distance. Finally, we show that variationally stable guidance estimators are adaptive to unknown low dimensionality, effectively mitigating the curse of dimensionality under low-dimensional subspace assumptions.

Keywords: generative learning, Bayesian inference, convergence analysis, posterior sampling, diffusion models

Chenguang Duan

duan@igpm.rwth-aachen.de

RWTH Aachen University

Talk: *Provable Diffusion Posterior Sampling for Bayesian Inversion*

Abstract: This paper proposes a novel diffusion-based posterior sampling method within a plug-and-play (PnP) framework. Our approach constructs a probability transport from an easy-to-sample terminal distribution to the target posterior, using a warm-start strategy to initialize the particles. To approximate the posterior score, we develop a Monte Carlo estimator in which particles are generated using Langevin dynamics, avoiding the heuristic approximations commonly used in prior work. The score governing the Langevin dynamics is learned from data, enabling the model to capture rich structural features of the underlying prior distribution. On the theoretical side, we provide non-asymptotic error bounds, showing that the method converges even for complex, multi-modal target posterior distributions. Finally, we present numerical experiments demonstrating the effectiveness of the proposed method across a range of inverse problems.

Keywords: Generative model, Bayesian inverse problems, Machine learning, convergence rate analysis, posterior sampling

Dominic Edelmann

dominic.edelmann@dkfz.de

German Cancer Research Center

Talk: *Omnibus Goodness-of-Fit Testing for Distributions on Stiefel Manifolds*

Abstract: In this talk, a comprehensive framework for goodness-of-fit testing for distributions on Stiefel manifolds is developed. The approach is based on integrals of the squared differences between empirical and theoretical characteristic functions, yielding test statistics that are consistent against all fixed alternatives. For the Fisher-Bingham family of distributions, explicit computable forms of the test statistic are derived. Simplified expressions for important special cases, including the matrix Fisher, matrix Bingham, and uniform distributions are provided. In the case of testing uniformity on hyperspheres, we obtain the complete asymptotic distribution of the test statistic, enabling computationally efficient asymptotic testing. For general Fisher-Bingham distributions, we establish theoretically justified Monte Carlo testing procedures for both simple and composite hypotheses. Simulation studies demonstrate accurate Type I error control and strong power across a wide range of alternatives. The practical relevance of the proposed methodology is illustrated using an application to data on the orbits of comets.

Keywords: goodness-of-fit, hypersphere, Stiefel manifold, Fisher-Bingham, characteristic function

Xiao Fang

xfang@cuhk.edu.hk

The Chinese University of Hong Kong

Talk: *Central limit theorem for high temperature Ising models via martingale embedding*

Abstract: We use martingale embeddings to prove a central limit theorem (CLT) for projections of high-dimensional random vectors satisfying a Poincaré inequality. We obtain a non-asymptotic error bound for the CLT in 2-Wasserstein distance involving two-point and three-point covariances. We present two illustrative applications to Ising models: one with finite-range interactions and the other in the ferromagnetic case under the Dobrushin condition.

This talk is based on joint work with Yi-Kun Zhao. The manuscript is available at: https://arxiv.org/abs/2511.06196

Keywords: Central limit theorem, correlation decay, Dobrushin’s condition, Ising model, martingale embedding

Maxim Fedotov

maxim.fedotov@upf.edu

Universitat Pompeu Fabra

Talk: *Projected Information Criteria*

Abstract: We introduce Projected Information Criteria, a new class of variable selection criteria for regression and classification problems. The proposed approach offers a computationally efficient alternative to classical criteria such as AIC, BIC, and EBIC by allowing to reduce problem dimensionality and avoiding costly repeated optimization for each candidate subset. The projected criteria are based on a global estimate computed once using all available covariates, which is then being reused during inference. We show that projected criteria exhibit strong theoretical guarantees: in a broad class of models, including generalized linear models, they achieve selection consistency under mild conditions on the model and the global estimator. We also establish non-asymptotic error probability bounds for additive models with Gaussian noise, providing finite-sample guarantees and supporting their use beyond the fixed-dimensional asymptotic regime. In addition, we show that the resulting selection problem can be formulated as a binary quadratic optimization problem, enabling practical computation with modern solvers. We argue that the computational gain compared to classical criteria is particularly pronounced when the global estimate is sparse. In simulation studies, the projected criteria demonstrate highly competitive performance relative to popular alternatives, including the regular and adaptive Lasso.

Keywords: Variable selection, Information criteria, High-dimensional inference, Selection consistency, Binary quadratic optimization

Davide Ferrari

davferrari@unibz.it

Free University of Bolzano

Talk: *A Single-Observation Uniformity Test via Multiscale Lacunary Harmonics*

Abstract: A test of uniformity on $[0,1]$ is developed for the setting of a single observation recorded with sufficient precision. Although consistency against general alternatives is not attainable with only one draw in the classical large-sample sense, a multiscale harmonic digit expansion provides a framework for structured inference. By aggregating trigonometric components across digit scales at suitably separated frequencies, a quadratic test statistic is constructed whose null distribution converges to a chi-square law via a lacunary central limit theorem. Under contiguous alternatives, the statistic admits a noncentral chi-square limit, with noncentrality determined by a Fourier projection of the alternative density. The resulting procedure detects multiscale harmonic structure that remains invisible to classical digit-frequency methods.

Keywords: Uniformity testing, Single-observation inference, Lacunary trigonometric series, Central limit theorem;, Multiscale harmonic analysis

Si Cheng Fong

ryan.fong@link.cuhk.edu.hk

The Chinese University of Hong Kong

Talk: *Optimal Subsampling and EM Algorithms for Non-Markovian Semiparametric Regression with Interval-Censored Multi-State Data*

Abstract: Interval-censored multi-state data frequently arise in longitudinal studies of chronic diseases and credit rating transitions, where exact transition times are unobserved and only known to lie within specific intervals. Inference in such settings is complicated by non-Markovian dynamics and high computational demands in large-scale datasets.

To address these challenges, we propose a semiparametric proportional intensity model with clearly established identifiability conditions that ensure unique parameter estimation. Flexible B-spline basis functions are used to capture non-Markovian effects, while a stable expectation-maximization (EM) algorithm is developed for nonparametric maximum likelihood estimation under general interval censoring. We further introduce a general subsampling EM framework and advance it into an optimal subsampling EM algorithm by deriving sampling probabilities that minimize the asymptotic variance of the parameter estimates. This approach substantially enhances computational efficiency while preserving semiparametric efficiency.

Theoretical results establish the consistency, asymptotic normality, and semiparametric efficiency of the proposed estimators. Extensive simulation studies confirm the method's robustness, scalability, and strong performance in analyzing complex interval-censored multi-state processes. Application study in credit risk modelling are extensively demonstrated.

Keywords: EM algorithm, Optimal subsampling, Interval-censored, Multi-state transition, Semiparametric proportional intensity model

Benedikt Fröhlich

benedikt.froehlich@informatik.uni-regensburg.de

University of Regensburg

Talk: *Reducing correlation bias of MDI importance for tree-based estimators through local sample weighting*

Abstract: Feature importance (FI) statistics provide valuable insight into the degree to which individual features contribute to the prediction of a machine learning model. However, standard metrics such as Mean Decrease in Impurity (MDI) for tree-based estimators are notoriously biased when features are not independent, attributing inflated FI to noise features that are highly correlated with signal variables.

To address this issue, we propose "local sample weighting", which can be integrated locally in the split selection of tree-based models. At each node, it reweights the sample for each candidate feature with an estimated density ratio, reducing the correlation of the candidate with the other variables, such that the split can be chosen purely based on the marginal effect of the candidate.

Our approach comes with a natural tuning parameter, the minimum effective sample size of the weighted population. It introduces an interpretation-prediction-tradeoff by moderating between the degree of decorrelation and the effective sample size used for estimation.

We validate our approach with an extensive simulation study and support our simulation findings with theory on consistency of our model as well as asymptotic behavior of the MDI. We also investigate the relationship between the minimal effective sample size and population MDI as a key component of the interpretation-prediction-tradeoff.

Part of this work is joint work with Alison Durst and Merle Behr, see arXiv: 2508.06337.

Keywords: Machine Learning, Nonparametric, Causal Inference, Random Forest, Interpretability

Sebastian Fuchs

sebastian.fuchs@plus.ac.at

University of Salzburg

Talk: *Continuity of Chatterjee's rank correlation and related dependence measures*

Abstract: While measures of concordance such as Spearman's rho and Kendall's tau are continuous with respect to weak convergence, Chatterjee's rank correlation, recently introduced by Azadkia and Chatterjee (2021), does not share this property, causing drawbacks in statistical inference as pointed out by Bücher and Dette (2025). As we show, Chatterjee's rank correlation is instead weakly continuous with respect to conditionally independent copies, namely the Markov products. To establish weak continuity of Markov products, we provide several sufficient conditions, including copula-based criteria and conditions relying on the concept of conditional weak convergence from Sweeting (1989). As a consequence, we obtain continuity results for Chatterjee's rank correlation and related dependence measures and verify their continuity in standard models such as multivariate elliptical and l1-norm symmetric distributions.

Keywords: conditional weak convergence, elliptical distribution, Markov product, noise resistance, stochastically increasing

Agustín García Nogales

agnogales@gmail.com

Universidad de Extremadura

Talk: *Some Examples on Optimal Bayesian Estimation\\ of the Sampling Distribution*

Abstract: This paper addresses, from a Bayesian perspective, the problem of estimating the unknown probability distribution governing a random experiment and, in the dominated case, estimating its density function. It is shown that the posterior predictive distribution is the optimal solution (Bayes estimator) to the first of these problems for the squared total variation loss function. The posterior predictive density is the Bayes estimator of the density for the L1-squared loss function. The results are presented within a very general framework encompassing continuous and discrete, uni- and multivariate, parametric and non-parametric cases, without making any explicit assumptions about the prior distribution. Some examples of applications of these results are shown, two of them of a nonparametric nature.

Keywords: Bayes estimator of a probability distribution., Bayes estimator of the density., Posterior predictive distribution or density., Bayesian nonparametrics., Markov kernel.

Celia Garcia Pareja

celiagp@kth.se

KTH Royal Institute of Technology

Talk: *Universal diffusion limit for Markovian models of evolution in structured populations with migration*

Abstract: The evolution of microbial subpopulations that migrate within spatial structures has gained interest in recent years. Questions of relevance include, for instance, the ability of a migrant mutant to take over the population (fixate). Estimating fixation probabilities is, however, usually hindered by the lack of analytical formulas and by computational complexity of simulation-based strategies when considering large populations. In this work, we study several population genetics models where the population is divided into D subpopulations (called demes) consisting of two types of individuals, mutants and wild-types, that evolve through discrete Markovian updates. We prove that under certain assumptions all the considered models converge to the same diffusion approximation, which we call \textit{universal}. This diffusion approximation is amenable to simulation strategies that underly methods of statistical inference while significantly reducing computational costs. In all models, each Markovian update follows two phases: First, a local growth phase in each subpopulation, where the growth of each type of individual depends on its fitness, and then a sampling phase that implements migration between subpopulations. Our proof relies on existing diffusion approximation results for degenerate diffusions but requires further technicalities due to fact that sample sizes in each deem are not necessarily fixed but change randomly with each update.

This is joint work with Alia Abbara

Keywords: Diffusion approximation, Evolution in structured populations, Martingale problem, Stochastic processes, Population genetics

Shubhangi Ghosh

sg4156@columbia.edu

Department of Statistics, Columbia University

Talk: *Signal-to-noise ratio aware minimax analysis of sparse linear regression*

Abstract: We consider parameter estimation under sparse linear regression– an extensively studied problem in high-dimensional statistics and compressed sensing. While the minimax framework has been one of the most fundamental approaches for studying statistical optimality in this problem, we identify two important issues that the existing minimax analyses face: (i) The signal-to-noise ratio appears to have no effect on the minimax optimality, while it shows a major impact in numerical simulations. (ii) Estimators such as best subset selection and Lasso are shown to be minimax optimal, yet they exhibit significantly different performances in simulations. In this paper, we tackle the two issues by employing a minimax framework that accounts for variations in the signal-to-noise ratio (SNR), termed the SNR-aware minimax framework. We adopt a delicate higher-order asymptotic analysis technique to obtain the SNR-aware minimax risk. Our theoretical findings determine three distinct SNR regimes: low-SNR, medium-SNR, and high-SNR, wherein minimax optimal estimators exhibit markedly different behaviors. The new theory not only offers much better elaborations for empirical results, but also brings new insights to the estimation of sparse signals in noisy data.

Keywords: high-dimensional linear regression, sparsity, minimax optimality, asymptotic analysis, signal-to-noise ratio

Deep Ghoshal

deepg2@illinois.edu

University of Illinois Urbana-Champaign

Talk: *Inference for Change Point in Marginal Distribution of High Dimensional Time Series*

Abstract: This work addresses the challenge of nonparametric inference for change points in the context of marginal distribution of high-dimensional, temporally dependent Euclidean or functional data. The literature on change-point testing for marginal distribution of high-dimensional data is still limited but growing. However, in our knowledge, there is currently no theoretically valid method for this problem that handles temporal dependence; applying methods designed for independent data typically results in severe size distortion. We propose a kernel-based test that embeds the data into the Cartesian product of reproducing kernel Hilbert spaces, constructing a CUSUM-statistic via sample splitting, trimming, projection, and self-normalization techniques. For stationary and absolutely regular processes, we establish that under weak cross-sectional dependence and standard moment assumptions, the test statistic possesses a pivotal limiting null distribution as both sample size and dimension diverge. We further analyze the asymptotic power of our test in the presence of a single change point. Our theoretical framework leverages Berbee's coupling result and introduces a novel martingale approximation argument to establish a functional central limit theorem in the high-dimensional setting. Finally, we validate the finite-sample size and efficiency of our method through simulation studies.

Keywords: Change-point detection, Marginal distribution, High-dimensional time series, Sample-splitting, Self-normalization

Hanke Guo

hanke.guo@tum.de

Technical University of Munich / Munich Center for Machine Learning (MCML)

Talk: *Rank-restricted Wald tests for goodness-of-fit testing*

Abstract: We present a modified version of a Wald test that allows one to test multiple constraints in the case where the estimated restrictions have an asymptotic covariance matrix that is singular. In such settings, it is tempting to replace the matrix inverse in the construction of the test statistic by the Moore-Penrose pseudoinverse. However, this approach does not provide a valid test for problems in which the rank of the covariance matrix estimate may exceed the asymptotic rank. To address this issue, we propose the use of a rank-restricted pseudoinverse, defined for a specified fixed rank. For this construction, it is not necessary to know the exact rank of the asymptotic covariance matrix, which can be difficult to determine when testing complex sets of constraints. We then clarify under which conditions the rank-restricted Wald statistics admit an asymptotic chi-square distribution.

As a motivating example, we consider goodness-of-fit testing for Independent Component Analysis (ICA) models. In ICA, one observes a collection of linear combinations of independent and non-Gaussian source variables. This setting, which also encompasses linear non-Gaussian structural equation models (LiNGAM), leads to algebraic relations among second and third moments, which we test via the proposed rank-restricted Wald tests.

Keywords: Wald test, Pseudoinverse, Goodness-of-fit test, Independent component analysis, Singular restrictions

Wenqing He

whe@stats.uwo.ca

University of Western Ontario

Talk: *Boosting methods for interval censored data with regression and classification*

Abstract: Boosting has gained significant interest across both machine learning and statistical communities. Traditional boosting algorithms, designed for fully observed random samples, often struggle with real-world problems, particularly with interval censored data. This type of data is common in survival analysis where exact event times are unobserved but fall within known intervals. Effective handling of such data is crucial in fields like medical research, reliability engineering, and social sciences. In this work, we introduce novel nonparametric boosting methods for regression and classification tasks with interval censored data. Our approaches leverage censoring unbiased transformations to adjust loss functions and impute transformed responses while maintaining model accuracy. Implemented via functional gradient descent, these methods ensure scalability and adaptability. We rigorously establish their theoretical properties. Our proposed methods not only offer a robust framework for enhancing predictive accuracy in domains where interval censored data are common but also complement existing work, expanding the applicability of existing boosting techniques. Empirical studies demonstrate robust performance across various finite-sample scenarios, highlighting the practical utility of our approaches.

Keywords: Boosting, Interval-censored, Regression, Classification, Robust

Masayuki Hemmi

henmi@ism.ac.jp

The Institute of Statistical Mathematics

Talk: *Infinite-dimensional information geometry for semiparametric statistics*

Abstract: In information geometry, a parametric model (a parametric family of probability distributions) is treated as a finite-dimensional differentiable manifold, where the Riemannian metric called Fisher metric and the pair of two dual affine connections called the exponential and mixture connections play essential roles for geometrical consideration of statistical inference. For example, the maximum likelihood estimation in an exponential family is seen as the orthogonal projection of the geodesic defined by the mixture connection and its statistical properties can be understood in terms of geometry. However, semi-parametric statistical methods cannot be understood geometrically in this framework because the statistical models behind these methods are essentially infinite-dimensional. Since Pistone and Sempi (1995) introduced an infinite-dimensional manifold structure of probability

distributions using the notion of Orlicz space, the geometrical theory of infinite-dimensional statistical models has been developed from a mathematical point of view. However, it has not sufficiently provided a geometrical understanding of semi-parametric statistical methods yet. The aim of this work is to reconsider the construction of infinite-dimensional manifold structure of probability distributions and to develop geometrical tools for geometrical understanding of statistical inference so that it may provide an extension of the semiparametric estimation theory.

Keywords: Information geometry, Semiparametric model, Semiparametric estimation theory, Infinite-dimensional manifold, Tangent space

Jieliang Hong

hongjl@sustech.edu.cn

Southern University of Science and Technology

Talk: *Rescaled SIR epidemic processes converge to super-Brownian motion in four or more dimensions*

Abstract: In dimensions four or more, by choosing a suitable scaling parameter, we show that the rescaled spatial SIR epidemic process converges to a super-Brownian motion with drift, thus complementing the previous results by Lalley (Prob. Th. Rel. Fields, 144 (2009), 429--469) and Lalley-Zheng (Prob. Th. Rel. Fields, 148 (2010), 527--566) on the convergence of SIR epidemics in dimensions three or less. The scaling parameters we choose also agree with the corresponding asymptotics for the critical probability p_c of the range-R bond percolation on Z^d as R goes to infinity.

Keywords: spatial epidemic, critical scaling, super-Brownian motion, branching random walk, percolation

Peter Hyland

p.hyland24@imperial.ac.uk

Imperial College London

Talk: *SGD Convergence for GLMs via Local Curvature: A Concentration of Measure Approach*

Abstract: We analyse the convergence rates of stochastic gradient descent (SGD) to a maximum likelihood estimator for the parameters of a generalised linear model. Many classical convergence results for SGD require strong convexity of the loss function, which is not satisfied for GLMs, even with the canonical link. Despite this, empirical experience suggests that SGD converges at a rate expected for a strongly convex loss close to the optimum. We provide convergence guarantees for this behaviour in wide range of settings, including logistic regression, with bounded or sub-Gaussian covariates. By exploiting concentration of measure results and exit time calculations, we show that SGD adapts to the local curvature close to the optimum with high probability.

Keywords: Stochastic Gradient Descent, Generalised Linear Models, Local Strong Convexity, Concentration of Measure, Exit Time Analysis

matteo iacopini

miacopini@luiss.it

Luiss University of Rome

Talk: *A Bayesian Dynamic Latent Space Model for Weighted Networks*

Abstract: A new dynamic latent space eigenmodel (LSM) is proposed for time-series of weighted networks.

The model accounts for integer-valued weights with excess zeros, time-varying node positions (features), and time-varying sparsity probabilities. The latent positions evolve according to a vector autoregressive process that accounts for possible lagged and contemporaneous dependence across nodes and features, which is 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. Regarding the first task, 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 the mixing of the chain. To avoid trans-dimensional samplers, a Laplace approximation to a partial marginal likelihood is used. Overall, our partially collapsed Gibbs sampler is general, as it can be easily extended to static and dynamic settings, as well as to discrete or continuous network weights.

Keywords: Latent space, Count time series, Bayesian statistics, Dyamic network, Data augmentation

Nicolas Alexander Ihlo

Nicolas.Ihlo@informatik.uni-regensburg.de

Faculty of Informatics and Data Science, University of Regensburg

Talk: *Provable Recovery of Locally Important Signed Features and Interactions from Random Forest*

Abstract: Feature and Interaction Importance (FII) methods are essential in supervised learning for assessing the relevance of input variables and their interactions in complex prediction models. In many domains, such as personalized medicine, global scores summarizing overall feature importance are insufficient; instead, local interpretations for individual predictions are essential. Random Forests (RFs) are widely used in these settings, and existing interpretability methods typically exploit tree structures and split statistics to provide model-specific insights. However, theoretical understanding of local FII methods for RF remains limited, making it unclear how to interpret high importance scores for individual predictions.

We propose a novel, local, model-specific FII method that identifies frequent co-occurrences of features along decision paths, combining global patterns with those observed on paths specific to a given test point.

We prove that our method consistently recovers the true local signal features and their interactions under a Locally Spike Sparse (LSS) model and also identifies whether large or small feature values drive a prediction. We illustrate the usefulness of our method and theoretical results through simulation studies and a real-world data example.

This is joint work with Kata Vuk and Merle Behr, see preprint https://arxiv.org/abs/2512.11081.

Keywords: Tree ensembles, Interpretability, Consistency, Local scores, Model-specific scores

Jef Jonkers

jef.jonkers@ugent.be

Ghent University

Talk: *Functional Calibration for Survival Prediction under Right Censoring*

Abstract: Many survival models produce full predictive distributions, yet calibration is often assessed only through fixed-horizon summaries. We develop a general framework for functional calibration under right censoring, focusing on whether decision-relevant summaries of a predicted survival distribution, such as horizon-specific risks, quantiles, expectiles, or moments, are reliable conditional on their issued values. Using inverse-probability-of-censoring weighting, we show how these calibration targets can be identified from censored data. We then introduce censoring-aware functional reliability diagrams based on a weighted H-pool-adjacent-violators algorithm, yielding monotone empirical recalibration maps for identifiable forecast functionals. The same construction induces score decompositions into miscalibration, discrimination, and uncertainty, extending Murphy-type decompositions to the right-censored setting. We also study large-sample behavior in discrete- and continuous-support regimes, providing a basis for uncertainty bands around the reliability diagrams. The framework unifies several existing survival calibration tools and provides a flexible way to diagnose where and how predictive survival distributions are miscalibrated under censoring.

Keywords: Survival analysis, Probabilistic forecasting, Calibration, Reliability diagrams, Isotonic regression

Pushkar Kale

pushkar.kale@u.nus.edu

National University of Singapore

Talk: *Moment constrained cutting feedback for modular Bayesian models*

Abstract: Statistical models are often constructed from multiple linked submodels (modules), each informed by different data sources and domain expertise. Misspecification in any module can distort posterior inference and propagate across the entire model. Cutting feedback methods address this by modifying the joint posterior so that unreliable modules cannot distort inference in trusted components. The standard formulation finds the Kullback-Leibler closest distribution to the full posterior whose marginal for the shared parameters matches the posterior from the trusted module alone. However, sampling from this posterior is challenging because it requires evaluating an intractable marginal density at each step. Consequently, naïve nested MCMC is typically used, leading to high computational cost. An alternative formulation, called the moment-constrained cut (MCC) posterior, replaces the full distributional constraint with a finite set of marginal mean and variance constraints on the shared parameters. The resulting distribution admits an exponential tilting representation, with the tilting coefficients estimated via stochastic gradient descent coupled with MCMC updates, whose convergence properties are established theoretically. This approach eliminates the need to run a separate Markov chain for each sample. Three benchmark examples demonstrate that the MCC posterior produces results similar to the conventional cut posterior while being much easier to compute.

Keywords: Constrained sampling, Cutting feedback, Exponential tilting, Modular inference, Moment constraints

Seungwoo Kang

kangsw@skku.edu

Sungkyunkwan University

Talk: *Multiview Functional Clustering Using Latent Representations of Phase and Amplitude Components*

Abstract: Functional data often exhibit two distinct sources of variability: amplitude variation in function values and phase variation arising from domain misalignment. While many preprocessing methods have been proposed to separate these two types of variation, their role in functional clustering has not been fully understood from a probabilistic perspective. Moreover, modeling the phase component within the usual functional data geometry fails to capture its intrinsic structural characteristics. In this work, we propose multiview functional k-means and k-medians clustering methods that explicitly incorporate both amplitude and phase variation into the clustering procedure. Within this framework, clustering is performed on latent representations consisting of both variations, allowing different views of the data to be explored by varying the relative contributions of phase and amplitude variation. With particular attention to the geometric structure of the phase component, the latent representations are defined in a separable Hilbert space, which enables theoretical guarantees under a principled probabilistic model. Simulation studies and real data analyses demonstrate that the proposed methods achieve improved clustering performance and provide multiview insights into the structure of functional data.

Keywords: Functional clustering, Curve registration, Aitchison distance, Random functions, Functional data analysis

Lucas Kania

lucaskania@cmu.edu

Carnegie Mellon University

Talk: *Optimal Inference with Black-box Predictions*

Abstract: Powerful black-box predictive models have motivated many proposals for combining observed data with predictions to perform valid statistical inference, including prediction-powered inference, dimension-agnostic inference, and universal inference. Despite this progress, the field lacks a unifying principle that explains how hypothesis tests should integrate data and predictions in a way that is both valid and efficient. In this talk, I address this gap by characterizing the information-theoretic limits of inference with black-box predictions in the high-dimensional Gaussian sequence model. Building on this characterization, I develop practical hypothesis tests that adapt optimally to the unknown accuracy of the predictions. Finally, I discuss how these results provide insight into the strengths and limitations of existing methodologies, and how the proposed methodology extends to nonparametric settings.

This is joint work with Abhinav Chakraborty, Edward Kennedy, Larry Wasserman, and Sivaraman Balakrishnan.

Keywords: adaptive inference, hypothesis testing, local minimax optimality, algorithms with predictions, structure-agnostic optimality

Dario Kieffer

dario.kieffer@stochastik.uni-freiburg.de

Albert-Ludwigs-Universität Freiburg

Talk: *The weak-feature-impact phase transition of the NPMLE in monotone binary regression*

Abstract: 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.

Keywords: Binary regression model, nonparametric maximum likelihood estimation, phase transition, weak feature impact, limiting distribution

Nick Koning

n.w.koning@ese.eur.nl

Erasmus University Rotterdam

Talk: *Equivalence testing with data-dependent and post-hoc equivalence margins*

Abstract: Equivalence testing compares the hypothesis that an effect is large against the alternative that it is negligible. Here, 'large' is classically expressed as being larger than some 'equivalence margin'. A longstanding problem is that this margin must be specified but can rarely be objectively justified in practice. We lay the foundation for an alternative paradigm, arguing to instead report a data-dependent margin that bounds the true effect with high probability. Our key argument is that this data-dependent margin is more useful than a test outcome at a fixed margin, as measured by the guarantees it offers to decision makers. We generalize this to a curve of margins, uniformly valid under the post-hoc selection of the margin. These ideas rely on e-values, which we derive for models that are strictly totally positive of order 3, nesting the classical z-test and t-test settings.

Keywords: Equivalence testing, E-values, Post-hoc testing, Hypothesis testing, Evidence-based decisions

Tomasz Kozubowski

tkozubow@unr.edu

University of Nevada, Reno

Talk: *Flexible skew models via coordinatewise Gaussian mixtures*

Abstract: The normal mean - variance mixture model is a versatile tool in probability and statistics, with applications in hierarchical modeling, Bayesian inference, stochastic processes, and machine learning. In this talk, we propose an extension that allows for coordinatewise mean - variance mixing in a multivariate Gaussian framework. A particular case based on gamma mixing leads to a new class of multivariate marginally generalized asymmetric Laplace (MGAL) distributions, extending the classical multivariate generalized asymmetric Laplace (GAL) model. The proposed class offers enhanced flexibility for modeling multivariate data with heterogeneous skewness and tail behavior across dimensions. We present key distributional properties of the MGAL family and discuss parameter estimation via the Expectation - Maximization (EM) algorithm. An application to real data illustrates the practical usefulness of the proposed approach. This is joint work with Amos Natido.

Keywords: Skew multivariate distribution, Financial data, Mixtures of Gaussian distributions, Skew Laplace distribution, Variance-gamma distribution

Yurou Liang

yurou.liang@tum.de

Technical University of Munich, Chair of Mathematical Statistics

Talk: *Likelihood-Based Nonparametric Causal Discovery under Latent Confounding*

Abstract: 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.

Keywords: Causal Discovery, Latent Confounding, Structural Identifiability, Parameter identifiability, Acyclic Directed Mixed Graphs

Jingbo Liu

jingbol@illinois.edu

University of Illinois

Talk: *Simple and sharp generalization bounds via lifting*

Abstract: We develop an information-theoretic framework for bounding the supremum of stochastic processes, offering a simpler and sharper alternative to classical chaining and slicing arguments for generalization bounds. The key idea is a lifting argument that produces information-theoretic analogues of empirical process bounds, such as Dudley's entropy integral. Lifting introduces permutation symmetry, yielding sharp bounds when the classical Dudley integral is loose. This gives a simple proof of the majorizing measure theorem via the sharpness of Dudley's entropy integral for stationary processes, a result known well before the proof of the majorizing measure theorem. Furthermore, the information-theoretic formulation provides soft versions of classical localized complexity bounds in generalization theory, but is simpler and does not require the slicing argument. We apply this approach to empirical risk minimization over Sobolev ellipsoids, obtaining sharp convergence rates in settings where previous methods are suboptimal. (arXiv:2508.18682)

Keywords: empirical risk minimization, empirical processes, rate-distortion function, covering numbers, Dudley's integral

Mengqi Lou

mlou30@gatech.edu

Georgia Institute of Technology

Talk: *Computationally efficient reductions between some statistical models*

Abstract: Average-case reductions establish rigorous connections between different statistical models, allowing us to show that if one problem is computationally hard, then another must be as well. Reductions from the planted clique problem have revealed statistical-to-computational gaps in many statistical problems with combinatorial structure. However, several important models remain beyond the reach of existing reduction techniques—for example, no reduction-based hardness results are currently known for sparse phase retrieval.

In this talk, we introduce a computationally efficient procedure that approximately transforms a single observation from certain source models with continuous-valued sample and parameter spaces into a single observation from a broad class of target models. I will present several such reductions and highlight their applications in computational lower bounds, including universality results and hardness in sparse generalized linear models. I will also discuss a potential application in transforming one differentially private mechanism into another.

Keywords: Computationally efficient reductions, Tensor PCA, Mixtures of sparse linear regressions, Sparse generalized linear model, Planted submatrix model

QILONG LU

lu_qilong@stu.pku.edu.cn

Peking University

Talk: *Versatile Differentially Private Learning for General Loss Functions*

Abstract: We provide a versatile privacy-preserving release mechanism along with a unified approach for subsequent parameter estimation and statistical inference. We propose a privacy mechanism based on zero-inflated symmetric multivariate Laplace (ZIL) noise, which requires no prior specification of subsequent analysis tasks, allows for general loss functions under minimal conditions, imposes no limit on the number of analyses, and is adaptable to increasing data volume in online scenarios. We derive the trade-off function for the proposed ZIL mechanism, which characterizes its privacy protection level. Furthermore, to formalize the local differential privacy (LDP) property of the ZIL mechanism, we extend the classical epsilon-LDP to a more general f-LDP framework. To address scenarios where only individual attribute values require protection, we propose attribute-level differential privacy (ADP) and its local version. Within the M-estimation framework, we introduce a novel doubly random (DR) corrected loss for the ZIL mechanism, which yields consistent and asymptotically normal M-estimates under differential privacy constraints. The proposed approach is computationally efficient and does not require numerical integration or differentiation for noisy data. It applies to a broad class of loss functions, including non-smooth ones. In addition, our method can handle unbounded data, while ensuring that discrete data remain discrete and bounded data stay within their original range.

Keywords: Differential privacy, M-estimation, Symmetric multivariate Laplace distribution, Zero-inflated symmetric multivariate Laplace distribution, Doubly random corrected loss

Sarah Lumpp

sarah.lumpp@tum.de

Technical University of Munich

Talk: *Identifiability in continuous Lyapunov models with heterogeneous noise*

Abstract: Stationary distributions of multivariate diffusion processes provide a new approach to probabilistic modelling of causal systems in statistics and machine learning. By assuming each observation to arise as a one-time cross-sectional snapshot of a temporal process in equilibrium, they allow one to naturally model dependence structures that may include feedback loops.

Specifically, the graphical continuous Lyapunov model consists of Gaussian stationary distributions of multivariate Ornstein-Uhlenbeck processes. Their covariance matrices are parametrized as solutions to the continuous Lyapunov equation, with sparsity assumptions on the processes' drift matrices represented by a directed graph.

Allowing for coordinate-wise different scaling of the driving noise process, we show generic parameter identifiability for certain classes of directed graphs, including tree-structured graphs. Additionally, we prove sufficient conditions for generic non-identifiability and characterize the generically identifiable structures in small acyclic graphs up to 6 nodes.

Keywords: Graphical models, Lyapunov models, Parameter identifiability, Stationary processes, Ornstein-Uhlenbeck process

Anna Lyubarskaja

annalyu@stanford.edu

Stanford University

Talk: *Chatterjee’s coefficient is optimal for detecting highly oscillating signals*

Abstract: We investigate the power of rank-based independence tests in detecting deterministic dependence for a class of k-oscillating functions. We establish that tests based on integrated CDF distances, such as Hoeffding’s coefficient , Blum--Kiefer--Rosenblatt’s coefficient , and Bergsma--Dassios--Yanagimoto’s coefficient, require a sample size of $n=k^2$ for detection. In contrast, Chatterjee’s coefficient succeeds with only n of order $k^{4/5}$ data points. Finally, by bounding the total variation distance between rank distributions, we prove that this rate is optimal: no rank-based test can detect k-oscillation when n is of order less than $k^{4/5}$

Keywords: Rank-based Independence Tests, K-oscillating functions, Chatterjee's Coefficient, Detection rates, Total variation distance

Lea Maislinger

lea.maislinger@plus.ac.at

University of Salzburg

Talk: *Multivariate Archimedean copulas may have fractal support*

Abstract: We study the family of all d-dimensional Archimedean copulas with d greater than or equal to three, a well-known class with broad applicability. Induced by a continuous univariate function, the so-called Archimedean generator, Archimedean copulas might be expected to distribute their mass in a rather regular manner on the unit cube. Motivated by a recently shown result stating that bivariate Archimedean copulas may have fractal support (in the sense of having non-integer Hausdorff dimension) we investigate the general multivariate case.

In particular, we show that for every d and arbitrary s between d-1 and d, there exists some d-dimensional Archimedean copula whose support has Hausdorff dimension s.

Finally, considering absolute continuity of the corresponding marginal copula our results are used to point out that ignoring at least one coordinate can obscure this irregularity, thereby creating a misleading impression of regularity.

Keywords: Archimedean copula, Fractal, Markov Kernel, Hausdorff Dimension, Stochastic measure

Leopold Mareis

leopold.mareis@tum.de

Technical University of Munich

Talk: *Cost-Aware Optimized Front-Door Experimental Design*

Abstract: Causal effect estimation often succeeds cost-constrained sequential data collection. This work considers multivariate linear front-door models with arbitrary unobserved confounding on treatment and response. We optimize the experimental design by balancing the statistical efficiency and measurement costs through partial data. The full-data efficient influence function for the causal effect is derived, together with the geometry of all observed-data influence functions. This characterization yields a closed-form optimal sampling policy and an estimator to minimize the asymptotic variance of regular asymptotically linear (RAL) estimators within a class of augmented full-data influence functions. The resulting design also covers back-door estimation. In simulations and applications to biological, medical, and industrial datasets, the optimized designs achieve substantial efficiency gains (5.3% to 31.9%) over naive full-sampling strategies.

Keywords: Causal inference, Experimental design, Front-door estimation, Linear SEM, Semiparametric estimation

Lorenz Matz

lorenz.matz@wu.ac.at

Vienna University of Economics and Business

Talk: *Analyzing the relationship between universal and classical likelihood ratio inference: Almost sure-relations and asymptotic rates*

Abstract: Universal Inference is a general method for constructing confidence sets and tests in statistical models, proposed by Wasserman et al. (2020). It is based on a combination of the likelihood ratio approach and sample splitting. Unlike procedures built on asymptotic results or most resampling methods, the resulting confidence sets/tests have finite-sample guarantees and require virtually no regularity conditions on the model or the set of null distributions. However, for many inference problems of interest, universal confidence sets and tests tend to be conservative in comparison to established methods with (finite sample or asymptotic) coverage/type I error guarantees. Currently, the relation between the universal method and classical approaches to inference in terms of power and confidence set diameters is still poorly understood. To shed some light on this, we investigate the relationship between the split LR and the classical LR confidence set in general settings. For one- and two-dimensional models, we establish that the classical LR set is almost surely contained in the split LR set when the maximum likelihood estimator is used for the latter. Furthermore, we find conditions under which the diameters of both confidence sets shrink at the same rate by proving some general results about these rates for 'M-type' sets.

Keywords: Universal Inference, Likelihood Ratio, Confidence Set, Sample Splitting, E-Values

Maria Chiara Menicucci

mariachiara.menicucci@carloalberto.org

University of Torino and Collegio Carlo Alberto

Talk: *Bayesian Nonparametric Privacy-Preserving Synthetic Data Generation*

Abstract: Synthetic data generation is a well-established method for releasing data while protecting individual privacy. From a statistical perspective, a release mechanism should satisfy differential privacy (DP) guarantees, and preserve statistical utility of the data. We propose two new release mechanisms based on Bayesian nonparametric models: modelling the private data as a sample from a random probability distribution, endowed with a nonparametric prior, we generate synthetic data sampling from the posterior predictive distribution. Specifically, when the private data have ties, we consider the Pitman-Yor process (PYP) prior and show that, in general, only instance-level (ε, δ)-DP is achievable. However, if the discount parameter σ of the PYP is zero, which corresponds to the Dirichlet process, then global (ε, δ)-DP can be obtained depending on the size of the released sample. When the private data have no ties, we consider the Pólya tree prior and show that global (ε, δ)-DP is achievable, with constraints on the size of the released sample whose stringency depends on the rate of growth of the tree coefficients. In both cases, we quantify the statistical utility of the mechanism by deriving rates for the expected 1-Wasserstein distance between the empirical measure of the released data and the data generating distribution. For the Dirichlet process, such a result is based on a novel concentration result for the posterior of a Dirichlet process, which is of independent interest.

Keywords: Differential Privacy, Dirichlet Process, Pólya Tree, Synthetic Data, Wasserstein Metric

Volodymyr Minin

vminin@uci.edu

University of California, Irvine

Talk: *Probabilistic forecasting via continuously weighted, proper score-optimal ensembles*

Abstract: When forecasting with parametric models, one usually employs the classical statistical approach of selecting for parameter values that best explain observed data, generally either by maximizing the likelihood function or approximating the Bayesian posterior distribution derived from it. A class of novel procedures has arisen to shift away from this paradigm, selecting instead for parameter values that induce a predictive distribution best matching the empirical distribution. Recent work, such as Focused Bayesian Prediction and Predictive Variational Inference, modifies the Bayesian update to achieve this idea in a generalized Bayesian setting. Building on these developments but not limiting ourselves to the Bayesian framework, we introduce a new paradigm for probabilistic forecasting with an arbitrary family of parametric time series models. Each member of the family generates its own probabilistic forecast, and the forecasts are combined into a continuously-weighted ensemble. We find optimal weights by minimizing a proper score (e.g., continuous ranked probability score) of the ensemble forecast over the training period. Our preliminary research compares the predictive performance of this new framework with the Bayesian posterior predictive distribution—the state of the art in many applications, such as infectious disease forecasting. Specifically, we apply our technique to the forecasting of healthcare utilization during peaks of respiratory virus activity.

Keywords: probabilistic forecasting, proper scoring rules, time series, Bayesian inference, infectious diseases

Gilles Mordant

mordantgilles@gmail.com

Yale University

Talk: *Statistics for regularized optimal transport with decreasing regularization*

Abstract: In this talk, we will discuss the statistical fluctuations of entropy-regularized optimal transport between empirical measures when the regularization parameter is a decreasing function of the sample size and the latter goes to infinity. We shall discuss how the dual potentials relates to the density of the data and how entropy-regularized optimal transport is related to diffusions.

Keywords: Optimal transport, Diffusion, Score estimation, Asymptotic statistics, Pseudodiffential operators

Ziang Niu

ziangniu@wharton.upenn.edu

University of Pennsylvania

Talk: *The blessing of refitting in doubly-robust conditional independence testing procedures*

Abstract: Conditional randomization tests (CRTs) is a powerful framework for conditional independence (CI) testing. Two practical drawbacks are often cited: (i) they require exact knowledge of the predictor's conditional distribution given covariates, and (ii) they involve refitting a model for each resampled statistic. In practice, the conditional distribution is rarely known and must be estimated from data. We show that when this distribution is learned in-sample, the refitting in CRT is not merely a computational burden; it can improve Type-I error control in finite samples. This improvement can be understood through the lens of data sharpening—a classical bias-correction technique in nonparametric regression. Building on this insight, we first generalize the data sharpening to accommodate broad learners and then propose new CI testing procedures based on the generalized sharpening approach. These procedures require at most two additional refits and enjoy provably more robust Type-I error control compared to the vanilla unsharpened version. Our theory accommodates a range of learners, including kernel ridge regression, lasso, and local linear regression. In particular, we extend our approach to CI testing based on doubly robust functional and show that data sharpening reduces second-order bias, leading to improved Type-I error control.

Keywords: conditional independence testing, conditional randomization test, data sharpening, generalized covariance measure, second-order bias

Jing Ouyang

jingoy@hku.hk

University of Hong Kong

Talk: *Inference on Generalized Latent Variable Models with High-Dimensional Responses and Covariates*

Abstract: Regression models with both high-dimensional responses and covariates have attracted growing attention. Standard multivariate regression models become inadequate when the response variables depend not only on observed covariates but also on latent variables that capture key unobserved characteristics. To draw statistical inferences on covariate effects while accounting for latent variables, we consider a high-dimensional generalized latent variable model that accommodates mixed-type responses and allows for flexible dependence between covariates and latent variables, which is more suitable for many real-world applications than traditional methods that either rely on a linear regression form or restricted assumptions on the dependence between covariates and latent variables. We develop an alternating algorithm that iteratively updates the regression parameters and the latent variables, transforming an intractable nonconvex problem into a sequence of tractable convex subproblems. Theoretically, we provide algorithmic guarantees by deriving error bounds for the resulting estimator, proving statistical consistency, and characterizing its convergence rate. Further, building on this estimator, we construct debiased estimates for the covariate effects and establish their asymptotic normality.

Keywords: Nonlinear Factor Models, Latent Confounders, Regularized Estimation, Nonconvex Optimization, Alternating Algorithm

Davy Paindaveine

Davy.Paindaveine@ulb.be

Université libre de Bruxelles

Talk: *Stopping on the last success with unknown odds: Impossibility barriers and quantitative oracle bounds*

Abstract: Optimal stopping problems lie at the interface of probability, statistics, and sequential decision theory. They model situations in which decisions must be made online, using only the information revealed so far and without access to future outcomes. We focus on the classical last-success problem: one observes a sequence of Bernoulli trials and seeks to stop exactly on the final success. When the success probabilities are known, the problem admits an elegant and complete solution via Bruss' sum-the-odds theorem (Bruss, 2000, Annals of Probability), which yields a simple threshold rule and an explicit formula for the optimal win probability. In most applications, however, these probabilities are unknown, so the oracle rule cannot be implemented. We therefore study the last-success problem under a minimal information structure, where the decision maker observes only the sequential Bernoulli outcomes. Our contribution provides a quantitative theory of oracle-freeness: we analyze the performance of the natural plug-in odds rule and, more importantly, characterize what is possible---and what is fundamentally impossible---when the success probabilities are unknown.

Keywords: Optimal stopping, Last-success problem, Plug-in rule, Oracle inequalities, Concentration bounds

Seeun Park

jo9373@snu.ac.kr

Seoul National University

Talk: *Tail-oriented Factor Models Using Expectiles*

Abstract: We propose a new factor model that captures tail-sensitive dependence structures in high-dimensional panel data, called the expectile factor model (EFM). While conventional factor models primarily focus on central co-movements, EFM targets conditional expectiles and extracts latent factors that characterize common dynamics in the tails. Importantly, the use of expectiles allows the model to exploit their smoothness and tail sensitivity through an asymmetric least squares criterion, which offers two key advantages. First, it yields numerically stable estimation even at very low or high expectile levels. Second, it enhances the ability to capture extreme events. We develop an iterative estimation algorithm and investigate the consistency and convergence rates of the proposed estimators. We further propose two consistent procedures for determining the number of factors, based on an information criterion and a rank-based approach. Simulation studies demonstrate that EFM remains numerically stable in extreme regions and better captures tail-driven latent components compared to existing quantile-based approaches. Empirical analyses of U.S. stock return data show that expectile factors provide substantial explanatory power for downside market risk, especially during periods of market stress. An additional application to macroeconomic forecasting demonstrates that tail-sensitive expectile factors capture information overlooked by mean-based factors and improve predictive performance.

Keywords: Asymmetric least squares, Extreme events, High-dimensional panel data, Tail dependence, Tail-driven factors

Paloma Pérez Fernández

palomaperezfernandez@gmail.com

Universidad de Extremadura

Talk: *On a Property of Conditional Expectation for Markov Kernels*

Abstract: A well-known property of conditional expectation states that, given three random variables such that the third is independent of the other two, then the conditional expectation of the first given the last two coincides with the conditional expectation of the first given the second. A Markov kernel can be considered as an extension of the concept of a random variable (and of a sub-sigma-field), and the authors have studied in previous works extensions to Markov kernels of certain probabilistic or statistical concepts (including independence and conditional expectation) and their properties. The main result of this communication considers the extension of the aforementioned result. A version in terms of densities is also presented.

Keywords: Markov kernel., Conditional expectation for Markov kernels., Independence of Markov kernel., Density. clinical diagnosis., Clinical diagnosis.

Constantinos Petropoulos

costas@math.upatras.gr

Department of Mathematics, University of Patras, Greece

Talk: *Improved estimators for functions of scale parameters in mixture models*

Abstract: Estimation of the scale parameter of the scale mixture of a location–scale family under the scale-invariant loss function is considered. The technique of Strawderman (Ann Stat 2(1):190–198, 1974) is used to obtain a class of estimators improving upon the best affine equivariant estimator of the scale parameter under certain conditions. Further, integral expressions of risk difference approach of Kubokawa (Ann Stat 22(1):290–299, 1994) is used to derive similar improvements for the reciprocal of the scale parameter. Using the improved estimator of the scale parameter and the improved estimator of the reciprocal of the scale parameter, classes of improved estimators for the ratio of scale parameters of two populations have been derived. In particular, Stein type and Brewster–Zidek type estimators are provided for the ratio of scale parameters of two mixture models. These results are applied to the scale mixture of exponential distributions, this includes the multivariate Lomax and the modified Lomax distributions.

Keywords: Affine equivariant estimator, Kubokawa's IERD approach, Stein type estimator, Brewster–Zidek type estimator, Strawderman-type estimator

Serge Provost

provost@stats.uwo.ca

The University of Western Ontario

Talk: *An Alternative Copula Construction for Samples Exhibiting Duplicate Component Values*

Abstract: After a brief introductory overview, the presentaton will introduce a construction that maps pseudo-observations obtained from a sample onto the cells of a grid partitioning the unit square, producing a random vector having uniformly distributed margins. This formalism is then refined into a coherent framework for representing stochastic dependence in settings where coordinate values may coincide. It will be shown that Sklar's theorem continues to hold within this extended context, thereby situating the approach within a familiar theoretical foundation. A smooth version of the resulting distribution is subsequently achieved through Bernstein polynomial approximation. The presentation will feature illustrative examples supported by graphical displays, and a connection to a notable work of art will be highlighted.

Keywords: Copulas, Pseudo-observations, Sklar's theorem, Bernstein polynomial approximation, Tied observations

Johannes Resin

resin@econ.uni-frankfurt.de

Goethe University Frankfurt

Talk: *Debiasing Score Decompositions for Reliable Inference*

Abstract: Proper scoring rules are widely used to quantify the predictive performance of probabilistic predictions. Score decompositions uncover contributions of miscalibration (reliability) and discrimination (resolution) on the overall predictive performance.

Recent work has proposed the use of isotonic regression to estimate score decompositions nonparametrically, allowing for consistent estimation under a natural monotonicity assumption. We demonstrate that these estimators are susceptible to bias and show empirically that cross-validation can be used to effectively debias the decomposition in simulations. We then discuss how to leverage large sample theory for cross-validated generalization errors to enable inference for the debiased score components. Finally, we illustrate the debiased decomposition in an application to common image classifiers from the machine learning literature.

Keywords: calibration, cross-validation, isotonic regression, probabilistic classifiers, proper scoring rules

Angelika Rohde

angelika.rohde@stochastik.uni-freiburg.de

University of Freiburg

Talk: *Asymptotic equivalence for nonparametric additive regression*

Abstract: We prove asymptotic equivalence of nonparametric additive regression and an appropriate Gaussian white noise experiment in which a multidimensional shifted Wiener process is observed, whose dimension equals the number of additive components. The shift depends on the additive components of the regression function and solely the one- and two-dimensional marginal distributions of the covariates via an explicitly specified bounded but non-compact linear operator Γ. The number of additive components d is allowed to increase moderately with respect to the sam-

ple size. In the special case of pairwise independent components of the covariates, the white noise model decomposes into d independent univariate processes. Moreover, we study approximation in some semiparametric setting where Γ splits into a multiplication operator and an asymptotically negligible Hilbert-Schmidt operator.

Keywords: Additive model, asymptotic sufficiency, high-dimensional regression, Le Cam distance, white noise models

Christopher Saunders

christopher.saunders@sdstate.edu

South Dakota State University

Talk: *Statistical Analysis of Forensic Black Box Studies*

Abstract: "Following the recommendations of the President's Council of Advisors on Science and Technology, researchers in forensic science have focused on the foundational validity of forensic source identification disciplines. A commonly performed study focuses on treating the forensic examiner as a “black box

Keywords: providing the examiner with a pair of items to compare and recording the result of the forensic examiner’s analysis on an ordinal point scale. The results of the study are then summarized in a manner that measures either the error rates or accuracy of the forensic examiners when applying the standard techniques to a given population of sources.

A typical black box study design has the following components: a sample of forensic examiners, a sample of sources, and a sample of objects from each source. These different components of variability make it difficult to approximate the standard error of the examiner accuracy. This difficulty has led to numerous methods being proposed for constructing confidence intervals that all lead to different results with the degree of the difference being dependent on the number of examiners and sources used in the study. In this presentation, we will review the experimental design of these studies, the various proposed methods for analyzing the experimental results, and use a functional data based Monte Carlo experiment to compare the different statistical approaches to constructing confidence intervals in these studies., Experimental Design, Forensic Source Identification, Black Box Studies, Categorical Data Analysis

Kazuma Sawaya

sawaya@g.ecc.u-tokyo.ac.jp

University of Tokyo

Talk: *Provable false discovery rate control for deep feature selection*

Abstract: We develop a flexible feature selection framework based on deep neural networks that approximately controls the false discovery rate (FDR), a measure of Type-I error. The method applies to architectures whose first layer is fully connected. From the second layer onward, it accommodates multilayer perceptrons (MLPs) of arbitrary width and depth, convolutional and recurrent networks, attention mechanisms, residual connections, and dropout. The procedure also accommodates stochastic gradient descent with data-independent initializations and learning rates. To the best of our knowledge, this is the first work to provide a theoretical guarantee of FDR control for feature selection within such a general deep learning setting.

Our analysis is built upon a multi-index data-generating model and an asymptotic regime in which the feature dimension n diverges faster than the latent dimension q, while the sample size, the number of training iterations, the network depth, and hidden layer widths are left unrestricted. Under this setting, we show that each coordinate of the gradient-based feature-importance vector admits a marginal normal approximation, thereby supporting the validity of asymptotic FDR control. As a theoretical limitation, we assume B-right orthogonal invariance of the design matrix, and we discuss broader generalizations. We also present numerical experiments that underscore the theoretical findings.

Keywords: variable selection, high-dimensional statistics, false discovery rate control, deep neural network, normal approximation

Kai Schärer

kai.schaerer@plus.ac.at

University of Salzburg

Talk: *Non-Parametric Distributional Regression via Sklar's Theorem and Empirical Copula Approximations*

Abstract: Nonparametric regression is a powerful tool for estimating unknown regression functions, yet most existing methods treat each functional, such as conditional means, quantiles, or expectiles, separately. We answer the natural question of whether there is a unified framework for regression on a broad class of functionals, using nonparametric estimation of the conditional distribution functions. We show that this can be done by exploiting Sklar's theorem, which separates the joint distribution into a copula and its marginal components. Under suitably chosen resolution and mild regularity assumptions, we establish uniform convergence to the conditional distributions of the underlying copula using the empirical checkerboard and Bernstein approximations. By plugging in the univariate empirical marginal distribution functions, we establish uniform convergence to the joint conditional distribution function. By simple aggregation of the estimator, we obtain consistent non-parametric estimators for a broad class of regression functionals, including those mentioned earlier and many others. The method provides a clear computational advantage over traditional kernel-based regressors. Simulation studies and a real-data application illustrate the finite-sample performance and flexibility of the proposed approach.

Keywords: Non-Parametric Regression, Copulas, Distributional Regression, Consistency, Checkerboard Approximations

Dominic T. Schickentanz

info@dominic-schickentanz.com

Technion and Paderborn University

Talk: *Critical Decoration of Branching Brownian Motion via Time-Inhomogeneous Additive Penalization Methods*

Abstract: Our aim is to provide a construction of the decoration point measure appearing in the extremal process of the branching Brownian motion, by constructing the law of this process seen from its topmost particle. This construction involves the law of a Brownian motion penalized by a solution of the F-KPP equation. We therefore give a general description of the law of a Brownian motion penalized by a time-inhomogeneous potential, which is of independent interest. This is joint work with Bastien Mallein (Toulouse).

Keywords: Stochastic processes, Branching Brownian motion, Decoration point measure, Brownian motion, Time-inhomogeneous penalizations

Ulrike Schneider

ulrike.schneider@tuwien.ac.at

TU Wien

Talk: *A Unified Framework for Pattern Recovery in Penalized Estimation*

Abstract: We consider the framework of least-squares penalized estimation where the penalty term is given by a polyhedral gauge, which encompasses methods such as LASSO and generalized LASSO, SLOPE, OSCAR, PACS and others. Each of these estimators can uncover a different structure or "pattern" of the unknown parameter vector. We define a novel and general notion of patterns based on subdifferentials and formalize an approach to measure pattern complexity. For pattern recovery, we provide a minimal condition for a pattern to be detected with positive probability, the so-called accessibility condition. We introduce the stronger noiseless recovery condition which is shown to play exactly the same role as the well-known irrepresentability condition for the LASSO in that the probability of pattern recovery is bounded by 1/2 if the condition fails, thereby unifying and extending irrepresentability to a general framework. We also look at the "gap" between accessibility and the noiseless recovery condition and illustrate that it is more pronounced for simple patterns. Finally, we prove that the noiseless recovery condition can be relaxed when turning to so-called thresholded penalized estimation: in this setting, the accessibility condition is already sufficient for sure pattern recovery provided that the signal of the pattern is large enough. We demonstrate how our findings can be interpreted through a geometric lens and illustrate our results for various estimation procedures.

Keywords: penalized estimation, regularization, (generalized) LASSO, pattern recovery, irrepresentability condition

Christof Schötz

christof.schoetz@tum.de

Technical University of Munich

Talk: *Power Fréchet Means for Robust Statistics in Metric Spaces*

Abstract: The 2-Fréchet mean generalizes the expectation to metric spaces by minimizing the expected squared distance to a random object, while the 1-Fréchet mean generalizes the median. Interpolating between these two notions, the alpha-Fréchet means (or power Fréchet means), for alpha between one and two, inherit robustness properties that likewise interpolate between those of the expectation and the median. We establish finite-sample bounds in expectation for their empirical counterparts in Hadamard spaces (complete metric spaces of non-positive curvature) under the sole assumption of a finite moment of order alpha. Notably, the moment of order (2 alpha minus 2) emerges as the dominant term in the error bound, making power Fréchet means particularly suitable for heavy-tailed distributions. These results hold in infinite-dimensional settings and are new even in Euclidean spaces.

Keywords: Metric Space, Fréchet Mean, Robust Statistics, Heavy Tails, Finite-Sample Bounds

Richard Schwank

richard.schwank@tum.de

Technical University of Munich

Talk: *Non-parametric recovery of causal diffusion mechanisms from equilibrium observations*

Abstract: When studying a causal system, it can be beneficial to consider its time evolution, particularly if there are feedback loops that are expected to unfold at small timescales. However, in many practical scenarios (e.g., for single-cell gene expression data), observations can only be obtained at a single point in time. We present methodology to recover the system's time-infinitesimal transition mechanism from such cross-sectional data. Precisely, we assume the system follows a time-homogeneous diffusion process that has reached an equilibrium distribution at observation time. Further, we assume the causal mechanism is fully described by the diffusion drift, is acyclic, and its causal graph is known. In this setting, we show that the full causal mechanism, i.e., the drift function, can be non-parametrically identified under mild conditions, and we provide a practical algorithm to solve this challenging inverse problem.

Keywords: Causality, SDE, PDE inverse problem, Ergodic system, Kernel method

Armin Schwartzman

armins@ucsd.edu

University of California, San Diego

Talk: *Consistency of fraction of variance estimation from summary statistics in high-dimensional linear models*

Abstract: In Genome-Wide Association Studies (GWAS), heritability is defined as the fraction of variance of an outcome explained by a large number of genetic predictors in a high-dimensional polygenic linear model. This work studies the asymptotic properties of the most common estimator of heritability from summary statistics called linkage disequilibrium score (LDSC) regression, together with a simpler and closely related estimator called GWAS heritability (GWASH). We show that, with some variations, two conditions which we call weak dependence (WD) and bounded-kurtosis effects (BKE) are sufficient for consistency of both the basic LDSC with fixed intercept and GWASH estimators, for both Gaussian and non-Gaussian predictors. For Gaussian predictors it is shown that these conditions are also necessary for consistency of GWASH (with truncation) and simulations suggest that necessity holds too when the predictors are non-Gaussian. We also show that, with properly truncated weights, weighting does not change the consistency results, but standardization of the predictors and outcome, as done in practice, introduces bias in both LDSC and GWASH if the two essential conditions are violated. Finally, we show that, when population stratification is present, all the estimators considered are biased, and the bias is not remedied by using the LDSC regression estimator with free intercept, as originally suggested by the authors of that estimator.

Keywords: Genome-Wide Association Studies, Heritability, Random-effects model, High-dimensional dependence, Kurtosis

Ömer Sercik

omer.sercik@uhasselt.be

Hasselt University

Talk: *Goodness-of-Fit Tests for the Survival Copula and Cross-Ratio Function based on Nonparametric Bernstein-based Estimators*

Abstract: In multivariate statistics, and particularly in bivariate time-to-event data analysis, inference regarding the underlying dependence structure is of crucial importance. In recent years, copulas have been widely used for modeling dependence between random variables. Alongside the application of the copula framework, several global and local association measures have been developed.In this work, we study the dependence structure between two non-negative random variables focusing on the cross ratio function (CRF) as a measure of local association. Astime-to-event data are often subject to (right) censoring, leading to partial information about these event times for some subjects, association measures such as the CRF should be estimated explicitly accommoding censoring. Most existing approaches for copula modeling and CRF estimation are parametric in nature, leading to the need of reliable methods to assess parametric assumptions. In this work, we propose Bernstein-based estimators for the survival copula and the CRF under both uncensored and univariate right-censored data. Specifically, we 1) introduce the proposed estimators for the survival copula and CRF, 2) develop goodness-of-fit tests for survival copula and CRF estimation, and 3) evaluate the performance of the proposed tests through an in-depth simulation study. Finally, we illustrate the performance of the proposed tests based on real-life data-applications.

Keywords: Bernstein-based estimation, Survival copula, Cross ratio function, Goodness-of-fit, Bivariate survival data

Ha-Young Shin

hayoung.shin@gmail.com

Department of Statistics and Actuarial Science, Soongsil University

Talk: *Treatment effects on Hadamard spaces*

Abstract: We propose a new framework for causal inference with Hadamard space-valued response variables, in which treatment effects are defined by a non-negative real number representing magnitude and a point at infinity representing direction. In particular, the average treatment effect (ATE) is defined as a magnitude-direction pair satisfying the following: when the the control group is transported a distance equal to that magnitude in that direction, its Frechet mean coincides with that of the treatment group. We define an ATE estimator, explore properties including strong consistency, and detail algorithms for computation of the ATE, using Broyden's method, and confidence regions. Simulations in which we investigate consistency and confidence region coverage are included. We perform causal inference on real diffusion tensor imaging (DTI) data and compare with Euclidean alternatives. Interpretability, and the testability of the null hypothesis of zero ATE, are major advantages over existing attempts to generalize the ATE to non-linear spaces; this is also illustrated with the DTI data.

Keywords: Geometric statistics, Causal inference, Treatment effects, Hadamard spaces, Diffusion tensor imaging

Bernhard Spangl

bernhard.spangl@boku.ac.at

BOKU University, Vienna

Talk: *An IDEAL algorithm for stream-based selective sampling in regression scenarios*

Abstract: In statistical learning and machine learning, the efficiency of data acquisition plays a crucial role in model performance, especially when labeled data is expensive or time-consuming to obtain. This challenge has given rise to the field of active learning, a paradigm where the learning algorithm actively selects the most informative data points for labeling to optimize the training process. While many active learning strategies consider query synthesis or pool-based sampling we focus on the problem of online active learning

in regression scenarios. We propose a new algorithm to iteratively select new data points from a data stream adapting ideas of the

Inverse Distance based Exploration for Active Learning (IDEAL) algorithm using thresholding of a properly chosen acquisition function. The IDEAL algorithm was originally developed for the pool-based sampling scenario and balances between model-based

learning and learning based on the pure exploration of the feature-vector space to promote diversity. We compare the proposed active selection strategy to other state-of-the-art methods and to the passive selection strategy that selects the next input points from the unlabelled data stream at random with a given probability and serves as a baseline. The performance of our method is evaluated in numerical tests on illustrative synthetic problems and real-world datasets. This talk is based on a joint ongoing work with Xiaojian Xu.

Keywords: online active learning, stream-based selective sampling, regression, inverse distance weighting, supervised learning

Dominik Stöger

dominik.stoeger@ku.de

KU Eichstätt-Ingolstadt

Talk: *Non-convex matrix sensing: Breaking the quadratic rank barrier in the sample complexity*

Abstract: For the problem of reconstructing a low-rank matrix from a few linear measurements, two classes of algorithms have been widely studied in the literature: convex approaches based on nuclear norm minimization, and non-convex approaches that use factorized gradient descent. Under certain statistical model assumptions, it is known that nuclear norm minimization recovers the ground truth as soon as the number of samples scales linearly with the number of degrees of freedom of the ground-truth. In contrast, while non-convex approaches are computationally less expensive, existing recovery guarantees assume that the number of samples scales at least quadratically with the rank r of the ground-truth matrix. In this talk, we close this gap by showing that the non-convex approaches can be as efficient as nuclear norm minimization in terms of sample complexity. Namely, we consider the problem of reconstructing a positive semidefinite matrix from a few Gaussian measurements. We show that factorized gradient descent with spectral initialization converges to the ground truth at a linear rate as soon as the number of samples scales linearly in the degrees of freedom of the ground truth. This improves the previous rank-dependence in the sample complexity of non-convex matrix factorization from quadratic to linear. Furthermore, we extend our theory to the noisy setting where we achieve minimax optimal error rates.

Keywords: Low-rank matrix estimation, matrix sensing, Non-convex optimization, Leave-one-out sequence, factorized gradient descent

Alex Trindade

alex.trindade@ttu.edu

Texas Tech UniTexas Tech University

Talk: *Optimized Inference for High-Dimensional Vector Autoregressions*

Abstract: Two computationally feasible estimators for high-dimensional sparse vector autoregressive (VAR) models are proposed, in which its dimension (d) or its lag-order p, or both increase with the sample size n. Both methods rely on the p-values derived from the least-squares estimator (LSE) as the basis for sparsification. The thresholding method (TLSE) sets to zero the coefficients with p-values exceeding a cutoff which depends on n, and the constrained model is refitted. The data-adaptive method (BLSE) first ranks the p-values and then fits increasingly larger models, selecting the model with the smallest information criterion from this sequence. Asymptotic normality of LSE is established, and under additional regularity conditions it is shown that both TLSE and BLSE have the oracle property, whereby the true sparsity structure can be recovered asymptotically. Weaker conditions result in the sure-screening property, guaranteeing that the true nonzero coefficients are included in the selected active set. Simulations provide a comparison with existing methods, and suggest that TLSE and BLSE are generally better at sparsity pattern recovery while maintaining computational feasiblity. The methodology is applied to high-dimensional neuroscience and econometrics real datasets, illustrating two types of Granger-causality recently proposed in the literature.

Keywords: least squares, threshold estimation, information criteria, oracle property, Granger-causality

Wolfgang Trutschnig

wolfgang.trutschnig@plus.ac.at

University of Salzburg

Talk: *d-stochastic measures with uniform (d-1)-marginals and fractal support*

Abstract: We study the family P_d of d-stochastic measures on [0,1]^d, whose (d-1)-dimensional marginals coincide with the uniform distribution on [0,1]^{d-1}, and show that this family contains elements with fractal support (in the sense that the support is self-similar with non-integer Hausdorff dimension). In fact, working with Iterated Function Systems with Probabilities (IFSP) it is possible to prove the following surprising theorem: for every dimension d>=3 the set of Hausdorff dimensions of supports of elements in P_d is dense in [d-1,d]. Complementing this result with some additional observations on the interplay between complete dependence and uniform (d-1)-marginals allows to conclude that the family P_d is much richer than one might conjecture.

Keywords: d-stochastic measure, copula, fractal, Hausdorff dimension, dependence

Guenther Walther

gwalther@stanford.edu

Stanford University

Talk: *Fast and Optimal changepoint detection and localization with Bonferroni triplets*

Abstract: The talk concerns the problem of detecting and localizing changepoints in a sequence of independent observations.

We propose to evaluate a local test statistic on a triplet of time points, for each such triplet in a particular collection.

This collection is sparse enough so that the results of the local tests can simply be combined with a weighted

Bonferroni correction. This results in a simple and fast method, {\sl Lean Bonferroni Changepoint detection} (LBD),

that provides finite sample guarantees for the existance of changepoints as well as simultaneous

confidence intervals for their locations. LBD is free of tuning parameters, and we

show that LBD allows optimal inference for the detection of changepoints. Joint work with Jayoon Jang.

Keywords: changepoint detection, Bonferroni adjustment, confidence interval, finite sample guarantees, optimal inference

Jian Wang

jianwang@mdanderson.org

The University of Texas MD Anderson Cancer Center

Talk: *A Bayesian Monitoring Approach Based on the Win Ratio for Randomized Phase II Trials with Multiple Time-to-Event Outcomes*

Abstract: Futility monitoring in Phase II randomized clinical trials with multiple time-to-event outcomes is typically based on a single survival endpoint or a composite time-to-first-event outcome. However, these approaches are inadequate, because a single endpoint may not adequately capture treatment effects, while composite endpoints treat all events as equally important, ignoring differences in clinical priority. We develop a Bayesian monitoring design for two-arm randomized Phase II trials using the win ratio framework to account for the clinical importance of multiple time-to-event endpoints. Operating characteristics of the design are evaluated via simulation and compared with conventional monitoring approaches. The proposed design stops early when deterioration occurs in higher-priority outcomes (e.g., death) while avoiding premature stopping driven by lower-priority outcomes (e.g., recurrence). This proposed approach provides a more tailored approach to decision-making in clinical trials with multiple time-to-event endpoints.

Keywords: Bayesian futility monitoring, multiple time-to-event endpoints, two-arm Phase II clinical trial, win ratio, Gibbs sampling

Jixin Wang

jixin.wang23@imperial.ac.uk

Imperial College London

Talk: *Drift Estimation and Donsker Theorems for SDEs with Non-smooth Drift*

Abstract: Ergodic diffusions with rough drift arise naturally in stochastic dynamics and nonparametric inference, but obtaining variance asymptotics and concentration inequality in this setting remains difficult. We study uniformly elliptic SDEs in $\mathbb{R}^d$ with Hölder diffusion coefficient and non-smooth, possibly discontinuous drift. Under a weak dissipativity condition guaranteeing an invariant measure and polynomial β-mixing, we establish a linear asymptotic variance formula for additive functionals via a Clark–Ocone representation, combining Aronson-type gradient estimates with long-time mixing control. We further prove a concentration inequality for kernel estimators of the invariant density indexed by translated kernel classes, as well as a Donsker theorem for the corresponding empirical process. Together, these results extend empirical-process methods to a rough drift diffusion setting and provide the probabilistic foundation for uniform analysis of Nadaraya–Watson type drift estimators under only local drift regularity.

Keywords: Diffusion processes, Non-smooth drift, Concentration inequalities, Donsker theorems, Nonparametric drift estimation

Liqun Wang

liqun.wang@umanitoba.ca

University of Manitoba

Talk: *Adaptive Lq-regularized Estimation of High-dimensional Sparse Covariance Matrix*

Abstract: The estimation of high-dimensional covariance matrix plays an important role in many application fields such as economics, biology, social and health sciences. A mainstream structural assumption for enhancing estimator accuracy is that the covariance matrix is sparse or approximately sparse. This paper proposes an adaptive Lq-regularized estimator with minimum eigenvalue constraint for high-dimensional sparse covariance matrix. This method eliminates the need for the conventional two-stage framework of sequential correlation and covariance matrix estimation. Under appropriate regularity conditions, we analyze its asymptotic and finite sample properties. The proposed iterative reweighted minimization method and its inexact variant can be employed to find a desired estimate. Simulation studies confirm that the proposed estimator performs better than some other state-of-the-art methods. The method is applied to a cancer RNA sequence data analysis. This is a joint work with X Wang, H Zhao, Z Zhou, and L Kong.

Keywords: Sparse covariance matrix, Regularized estimation, Finite sample properties, Iteratively reweighted optimizaiton, Inexact optimization method

Weijing Wang

wjwang@stat.nycu.edu.tw

National Yang Ming Chiao Tung University

Talk: *Two-Sample Comparison in the Presence of Long-Term Survivors*

Abstract: Clinical trials of modern cancer therapies frequently violate the proportional hazards assumption because of delayed treatment effects, crossing survival curves, and the presence of long-term survivors. In these settings, hazard ratio–based inference can be difficult to interpret.

We introduce the tau process, an intuitive and robust measure of treatment effect under nonproportional hazards that can also identify when the hazards change their ordering over time. We further extend this framework to accommodate long-term survivors by adopting a mixture (cure) modeling perspective and proposing estimands that compare both cure fractions and treatment effects within the susceptible (non-cured) subpopulation. In particular, we develop the susceptible tau process to quantify treatment effects among patients who remain at risk. A case study of the CheckMate 067 melanoma trial illustrates how these tools provide clinically meaningful insights into long-term treatment benefit.

This talk is based on joint work with Professor Yi-Cheng Tai (National Chengchi University) and Professor Martin T. Wells (Cornell University).

Keywords: Nonproportional hazards, Cure model, Estimands, Oncology trials, Tau Process

Yihua Wang

wangyi@yuntech.edu.tw

Department of Industrial Engineering and Management, National Yunlin University of Science and Technology

Talk: *Parametric and Nonparametric Approaches for Phase II Monitoring of Nonlinear Profiles*

Abstract: In many manufacturing and continuous process applications, product or process quality can be characterized by a functional relationship between a response variable and one or more explanatory variables. Such data are commonly referred to as profile data, and profile monitoring aims to detect changes in this functional relationship.

In practice, the underlying profile is sometimes nonlinear, making parameter estimation and monitoring more challenging in Phase II applications. To address this issue, we propose two Phase II monitoring approaches for nonlinear profiles. The first approach approximates the nonlinear function using a Taylor series expansion, transforming it into a linear form, and monitors the resulting parameters using a multivariate exponentially weighted moving average (MEWMA) control chart. The second approach is based on a nonparametric area-based statistic, which is incorporated into exponentially weighted moving average (EWMA) and homogeneously weighted moving average (HWMA) control charts.

Both approaches are designed to detect changes in profile parameters, and error variance. The performance of the proposed methods is evaluated through Monte Carlo simulations using the average run length (ARL) as the performance measure. Comparisons with the existing EWMA-based (NEWMA) monitoring method show that the proposed nonparametric schemes are more effective in detecting moderate to large shifts in profile parameters as well as increases in the error variance.

Keywords: Nonlinear profile monitoring, Phase II, EWMA, HWMA, Nonparametric control charts

Jan-Lukas Wermuth

wermuth@econ.uni-frankfurt.de

Goethe University Frankfurt

Talk: *Asymptotic Inference for Rank Correlations*

Abstract: Kendall's tau and Spearman's rho are widely used tools for measuring dependence. Surprisingly, when it comes to asymptotic inference for these rank correlations, some fundamental results and methods have not yet been developed, in particular for discrete random variables and in the time series case, and concerning variance estimation in general. Consequently, asymptotic confidence intervals are not available. We provide a comprehensive treatment of asymptotic inference for classical rank correlations, including Kendall's tau, Spearman's rho, Goodman-Kruskal's gamma, Kendall's tau-b, and grade correlation. We derive asymptotic distributions for both iid and time series data, resorting to asymptotic results for U-statistics, and introduce consistent variance estimators. This enables the construction of confidence intervals and tests, generalizes classical results for continuous random variables and leads to corrected versions of widely used tests of independence. We analyze the finite-sample performance of our variance estimators, confidence intervals, and tests in simulations and illustrate their use in case studies.

Keywords: Kendall’s tau, Spearman’s rho, Goodman-Kruskal’s gamma, Confidence Intervals, Statistical Tests

Thomas Wilkinson

tswilkinson@protonmail.com

University of Exeter

Talk: *Testing the calibration of fixed-event probability forecast sequences*

Abstract: The martingale property for a sequence of real-valued random variables can also be stated for a

sequence of random measures on a common measurable space. The martingale property for random

measures has been studied in Bayesian statistics, where the random measures in the sequence are

predictive probability distributions for a sequence of observations. We show that for a sequence of

random probability measures representing a sequence of forecasts made for a common observation,

the martingale property is equivalent to a basic requirement that the forecasts are calibrated for the

observation. We present a simple statistical test of the martingale property for a finite sequence of

random probability measures on $\mathbb{R}$, and interpret it as a test of forecast calibration. As part of the

test, we construct 'synthetic PIT values', based on the popular Probability Integral Transform

method for testing the calibration of a rolling-event sequence of forecasts. We also show how to

extend our method to test the martingale property of any finite sequence of random probability

measures on a common measurable space which is a Borel space.

Keywords: Random measures, Forecast calibration, Probability forecasts, Martingale, Measure theory

Rainer Winkelmann

rainer.winkelmann@econ.uzh.ch

University of Zurich

Talk: *Underreporting in count data models: some new results and an application*

Abstract: Winkelmann and Zimmermann (1993), although only published as a working paper, has

initiated a small but persistent literature on underreporting in count data

models. Their Poisson-logistic (Pogit) model was initially developed to estimate

the determinants of the number of job offers during a given time interval, when

the researcher only observes the number of accepted offers. Later Pogit

applications can be found in epidemiology, engineering, public health,

and criminology, among others. I review some of the recent

developments in the literature. I also consider an extension of the original

model where the reporting process can be estimated separately from the event

generating process due to the availability of a validation dataset.

Keywords: binomial thinning, infeasible Poisson, validation data, two-step Pogit, labor mobility

Fabrice Wunderlich

fabrice.wunderlich@uni.lu

University of Luxembourg

Talk: *Weak convergence of Volterra-type stochastic integrals in Skorokhod space*

Abstract: We study weak convergence of Volterra-type stochastic integrals whose integrands depend on both the integration variable and the terminal time. Such stochastic convolutions arise naturally in non-Markovian limit problems.

We obtain weak convergence results in both the J1 and M1 Skorokhod topologies under natural conditions on the semimartingale integrators, the integrands, and their interaction. The framework applies in particular to stochastic convolutions with singular kernels and is motivated by scaling limits arising in nearly unstable Hawkes models and rough volatility.

This is joint work with Giulia Livieri and Andreas Søjmark.

Keywords: Weak convergence, Stochastic processes, Stochastic integrals, Skorokhod topology, Volterra processes

Hongquan Xu

hqxu@stat.ucla.edu

UCLA

Talk: *A Stratified L2-Discrepancy with Application to Space-Filling Designs*

Abstract: Space-filling designs are widely used in computer experiments. We propose a stratified L2-discrepancy to evaluate the uniformity of a design when the design domain is stratified into various subregions. Weights are used to adjust preferences for the uniformity over subregions in each stratification. The stratified L2-discrepancy is easy to compute, satisfies a Koksma-Hlawka type inequality, and overcomes the curse of dimensionality that exists for other discrepancies. It is applicable to a broad class of designs, and covers several minimum aberration-type criteria as special cases. Strong orthogonal arrays of maximum strength are shown to have low stratified L2-discrepancies, and thus are suitable for computer experiments. In addition, we develop a lower bound for the stratified L2-discrepancy and provide a construction method for designs that achieve the lower bound. We further introduce a general version of the stratified L2-discrepancy for evaluating designs with flexible stratification properties.

Keywords: Computer experiment, curse of dimensionality, generalized minimum aberration, space-filling hierarchy principle, strong orthogonal array

Jiaming Xu

jx77@duke.edu

Duke University

Talk: *Random geometric graphs with smooth kernels: sharp detection threshold and a spectral conjecture*

Abstract: We show that the critical dimension for distinguishing a random geometric graph (RGG) with a smooth kernel from its Erdos-Renyi counterpart is given by $d* = n^{3/4}$, much lower than $d* = n^3$ for the hard RGG [Bubeck-Ding-Eldan-Racz '16]. To unify these results, we formulate a conjecture that the critical dimension is spectrally determined by $d = n^{3/4} b_1^{3/2}$, where b_1/d is the second eigenvalue of the kernel operator.

Keywords: Random graph, Hypothesis testing, Spectral method, Spherical harmonics, High-dimensional geometry

Yisha Yao

yy3381@columbia.edu

Columbia University

Talk: *Wasserstein Spatial Depth*

Abstract: Modeling observations as random distributions embedded within Wasserstein spaces is becoming increasingly popular across scientific fields, as it captures the variability and geometric structure of the data more effectively. However, the distinct geometry and unique properties of Wasserstein spaces pose challenges to the application of conventional statistical tools designed for Euclidean spaces. The space of distributions on $\mathbb{R}^d$ with $d>1$ is not linear, and ` ` mimic'' the geometry of a Riemannian manifold. Developing new methodologies for analysis within Wasserstein spaces has become essential. In this paper, we extend the concept of statistical depth to distribution-valued data, introducing the notion of Wasserstein spatial depth (WSD). This new measure provides a way to rank and order distributions, enabling the development of order-based clustering techniques and inferential tools. We show that the WSD preserves critical properties of conventional statistical depths, notably, ranging within $[0,1]$, transformation and geodesic invariance, vanishing at infinity, reaching a maximum at the geometric median, and continuity. Its robustness is established via the breakdown points of the empirical depth regions and its influence function. Additionally, the population WSD has a straightforward plug-in estimator based on sampled empirical distributions. We establish the estimator's consistency and asymptotic normality, and develop a two-sample test based on WSD.

Keywords: Distributional data analysis, High dimensional data, Statistical depths, Wasserstein distance, Robust inference

MI HONG YIM

mh.yim@kiom.re.kr

KOREA INSTITUTE OF ORIENTAL MEDICINE

Talk: *Preventive Use of Korean Medicine Among Children and Adolescents in Korea: A Statistical Analysis of National Real-World Data*

Abstract: Objectives: This study identified determinants of Korean Medicine (KM) use for preventive purposes among children and adolescents. Using Andersen's behavioral model, we investigated how socioeconomic characteristics influence KM utilization within a nationally representative dataset.

Methods: We analyzed the 2021 Korea Health Panel Survey (n=732). To account for the multi-stage stratified cluster sampling design, all statistical analyses were performed using the survey package in R. We employed multivariable logistic regression incorporating compensatory weights and Taylor series linearization for accurate variance estimation. Generalized Variance Inflation Factors (GVIF) were used to ensure model stability by addressing multicollinearity among predictors.

Results: Multivariable analyses showed KM utilization was significantly lower in the Busan/Gyeongsang region (aOR 0.28, 95% CI 0.10–0.83) versus the capital area and in larger households (≥5members; 0.30, 0.09–0.98). Conversely, those in the highest income quartile exhibited substantially higher odds (8.80, 1.80–43.03). KM visits primarily involved herbal preparations (88.25%), while conventional visits were dominated by immunizations (97.61%).

Conclusion: Findings suggest KM plays a unique role in pediatric wellness through herbal strategies. By employing rigorous complex-sample inference, this study provides robust evidence for Health Technology Assessments and integrated preventive care policies.

Keywords: Korean medicine, Pediatric preventive care, Complex sample survey design, Andersen's behavioral model, Real-world evidence

Adriano Zambom

adriano.zambom@csun.edu

California State University Northridge

Talk: *Consistent Variable Selection for GARCH-X Models*

Abstract: In this talk, we present a consistent variable selection procedure for GARCH-X models designed to identify which exogenous variables truly drive volatility dynamics. While GARCH-X models allow many potential covariates to enter the volatility equation, determining which ones genuinely matter remains a challenging problem.

The proposed approach combines Wald-type test statistics with a multiple testing framework based on the Benjamini–Yekutieli False Discovery Rate procedure, allowing us to systematically control false discoveries while retaining relevant predictors. We outline the intuition behind the method and discuss the theory of consistency — meaning that as the sample size grows, the procedure correctly recovers the true set of volatility drivers.

Simulation studies demonstrate that the method performs well across different distributions and dependence structures. We conclude with an empirical application to S&P 500 volatility, incorporating macroeconomic and commodity indicators, illustrating how the procedure can uncover meaningful volatility determinants in practice.

Keywords: Variable selection, GARCH-X models, False discovery rate, multiple hypothesis tests, Volatility modeling

Yizhe Zhu

yizhezhu@usc.edu

University of Southern California

Talk: *Minimax optimal differentially private synthetic data for smooth queries*

Abstract: Differentially private synthetic data enables the sharing and analysis of sensitive datasets while providing rigorous privacy guarantees for individual contributors. A central challenge is to achieve strong utility guarantees for meaningful downstream analysis. Many existing methods ensure uniform accuracy over broad query classes, such as all Lipschitz functions, but this level of generality often leads to suboptimal rates for statistics of practical interest. Since many common data analysis queries exhibit smoothness beyond what worst-case Lipschitz bounds capture, we ask whether exploiting this additional structure can yield improved utility. We study the problem of generating (ε,δ)-differentially private synthetic data from a dataset of size n supported on the hypercube $[-1,1]^d$, with utility guarantees uniformly for all smooth queries having bounded derivatives up to order k. We propose a polynomial-time algorithm that achieves a minimax error rate of $n^{-\min \{1, \frac{k}{d}\}}$, up to a $\log(n)$ factor. This characterization uncovers a phase transition at $k=d$. Our results generalize the Chebyshev moment matching framework in the literature and strictly improve the error rates for k-smooth queries. Moreover, we establish the first minimax lower bound for the utility of (ε,δ)-differentially private synthetic data with respect to k-smooth queries.

Keywords: differential privacy, synthetic data, minimax lower bound, moment matching, smooth queries