PLENARY TALK
Named Lectures

Kolmogorov Lecture
JULY 23 (FRI), 10:00 - 11:00 KST

From algorithm to theorem
Persi Diaconis*
Stanford University

Abstract: When you meet a new object (parking functions, perfectoid spaces,...) a natural question to ask is 'what does a typical such object 'look like'? This points to the need to have methods of generating 'typical objects'— those are the algorithms. In many, many cases, the availability of explicit algorithms can be turned into limit theorems (often with sharp error terms) allowing proofs for the patterns observed. This is a familiar activity; there are many ways of generating random permutations of n things, points on a line, points in space, the Feller coupling, Fisher - Yates and others. Each of these makes some kind of limit theorems 'easy'. I will give examples drawn from many areas, but focus on using the little known algorithm of Stam to generate random set partitions and my joint work with Chern, Kane and Rhodes. Proofs often need conditioned limit theory (dePoissonization) in fairly sophisticated forms for algorithms that entail randomizing a parameter to make components independent. There is also the dual task of going from theorems to algorithms (and then,...)

Bernoulli Lecture
JULY 19 (MON), 20:00 - 21:00 KST

Some models of spatially distributed populations: the effect of crowding
Alison Etheridge*
University of Oxford

Abstract: We consider some models of spatially distributed populations in which we take account of the effects of local crowding on both the number of offspring produced by individuals, and the chance that those offspring survive to maturity. In particular, we would like to understand the way in which ancestry of individuals that survive in such populations is affected by different responses to crowding. A special case would be a birth-death process with an additional logistic term controlling local population growth, but the novelty here is that we can also see an influence on the way in which mature individuals disperse. As time permits, we will touch on work with lots of people including Tom Kurtz (Madison), Peter Ralph (Oregon), Ian Letter, Aaron Smith, and Terence Tsui (all Oxford).

Levy Lecture
JULY 20 (TUE), 19:00 - 20:00, KST

A variational method for Euclidean quantum fields
Massimiliano Gubinelli*
University of Bonn

Abstract: I will talk about recent progresses in understanding the probabilistic structure of certain (bosonic) Euclidean Quantum Field theories in terms of a variational representation of their Laplace transform. This approach gives an alternative construction of the $\Phi^4_3$ measure in finite volume and a tool to investigate some of its properties.
It also generates some new interesting mathematical objects like a new kind of equations which allows to describe the infinite volume measures.

**Laplace Lecture**
JULY 21 (WED), 09:00 - 10:00 KST

**Transfer learning: optimality and adaptive algorithms**

Tony Cai*
University of Pennsylvania

**Abstract:** Human learners have the natural ability to use knowledge gained in one setting for learning in a different but related setting. This ability to transfer knowledge from one task to another is essential for effective learning. In this talk, we consider statistical transfer learning in various settings with a focus on nonparametric classification based on observations from different distributions under the posterior drift model, which is a general framework and arises in many practical problems. We first establish the minimax rate of convergence and construct a rate-optimal weighted K-NN classifier. The results characterize precisely the contribution of the observations from the source distribution to the classification task under the target distribution. A data-driven adaptive classifier is then proposed and is shown to simultaneously attain within a logarithmic factor of the optimal rate over a large collection of parameter spaces.

**Tukey Lecture**
JULY 22 (THU), 20:00 - 21:00 KST

**Max-margin classification and other interpolation methods**

Sara van de Geer*
ETH Zurich

**Abstract:** John Tukey writes that detective work is an essential part of statistical analysis (Tukey [1969]). In this talk we discuss methods that do the opposite of detective work: data interpolation. This was often considered forbidden, but then again, statistical paradigms are not to be sanctified. We consider basis pursuit and one-bit compressed sensing. We re-establish the $\ell_2$-rates of convergence for noisy basis pursuit of Wojtaszczyk [2010]. For one-bit compressed sensing we study the algorithm of Plan and Vershynin [2013] and re-derive $\ell_2$-rates as well. The techniques used also allow deriving novel results for the max-margin classifier - related to the ada-boost algorithm – as given in Liang [2020].

[This is joint work with Geoffrey Chinot, Felix Kuchelmeister and Matthias Löffler.]

**References**
Random walks and fractal graphs
Martin Barlow*
University of British Columbia

Abstract: This series of talks will study random walks on graphs with irregular, random or fractal structure. The motivation goes back to a 1976 article by the physicist Pierre de Gennes on percolation. Calling the simple random walk on a percolation cluster 'the ant in the labyrinth' he asked about its properties. It was conjectured in 1976, and has been proved in a number of cases since, that critical models in statistical physics have fractal structure. I will review de Gennes' questions. Since random fractals are hard, a first step was to look at deterministic exact fractals, and the graphs that can be associated naturally with them. The simplest of these is the Sierpinski gasket graph (SGG), and I will start with this example. Early work in this area used direct probabilistic methods, which were often very specific to the particular graph. The search for a more robust theory leads one to look for more flexible tools, of which the first is given by the connection, between random walks and electrical networks.

Low dimensional random fractals
Martin Barlow*
University of British Columbia

Abstract: The behaviour of the random walk can often be described by two indices, called by physicists the 'fractal' and 'walk' dimensions, and denoted by $d_f$ and $d_w$. This lecture will look at the tools which enable us to calculate these, and obtain the associated transition probability or heat kernel bounds. Three kinds of estimate are needed: (1) control of the size of balls (2) control of the resistance across annuli, and (3) a smoothness result (a Harnack inequality). In the 'low dimensional case' the Harnack inequality is not needed, and (2) can be replaced by easier bounds on the resistance between points. Many random fractals of interest are low dimensional: examples include critical branching processes, the incipient infinite cluster (IIC) for percolation in high dimensions, and the uniform spanning tree. Critical percolation in $d=2$ remains a challenge however.

Higher dimensional spaces
Martin Barlow*
University of British Columbia

Abstract: I will begin by discussing supercritical percolation clusters, where the random walk has Gaussian type bounds, and also satisfies an invariance principle. I will then discuss the challenge posed by fractal graphs with high dimension. I will conclude by reviewing models in the plane, such as the uniform infinite planar triangulation, a topic with connections to the Gaussian free field and Liouville quantum gravity.
**Estimating the mean of a random vector**
Gabor Lugosi*
ICREA & Pompeu Fabra University

**Abstract:** One of the most basic problems in statistics is the estimation of the mean of a random vector, based on independent observations. This problem has received renewed attention in the last few years, both from statistical and computational points of view. In this talk we review some recent results on the statistical performance of mean estimators that allow heavy tails and adversarial contamination in the data. In particular, we are interested in estimators that have a near-optimal error in all directions in which the variance of the one dimensional marginal of the random vector is not too small.

[The material of this talk is based on a series of joint papers with Shahar Mendelson.]

**Parking on Cayley trees and Frozen Erdös-Rényi**
Nicolas Curien*
Paris-Saclay University

**Abstract:** Consider a uniform Cayley tree $T_n$ with $n$ vertices and let $m$ cars arrive sequentially, independently, and uniformly on its vertices. Each car tries to park on its arrival node, and if the spot is already occupied, it drives towards the root of the tree and park as soon as possible. Using combinatorial enumeration, Lackner & Panholzer established a phase transition for this process when $m$ is approximately $n/2$. We couple this model with a variation of the classical Erdös–Rényi random graph process. This enables us to completely describe the phase transition for the size of the components of parked cars using a modification of the standard multiplicative coalescent which we named the frozen multiplicative coalescent. The geometry of critical parked clusters in the parking process is also studied. Those trees are very different from usual random trees and should converge towards the growth-fragmentation trees canonically associated to $3/2$-stable process that already appeared in the study of random planar maps.

[Based on joint work with Alice Contat.]

**Balloons in space(s)**
Omer Angel*
University of British Columbia

**Abstract:** We study a process of growing balls that annihilating on contact in several spaces, and explore the connection to stable matchings, random graphs, and conjectures regarding the Weil-Petersson measure.

[With Gourab Ray and Yinon Spinka.]
**Random determinants and the elastic manifold**
Gerard Ben Arous*
New York University

**Abstract:** The elastic manifold is a paradigmatic representative of the class of disordered elastic systems. These are surfaces with rugged shapes resulting from a competition between random spatial impurities (preferring disordered configurations), on the one hand, and elastic self-interactions (preferring ordered configurations), on the other. The elastic manifold model is interesting because it displays a depinning phase transition and has a long history as a testing ground for new approaches in statistical physics of disordered media, for example for fixed dimension by Fisher (1986) using functional renormalization group methods, and in the high-dimensional limit by Mézard and Parisi (1992) using the replica method. We study the energy landscape of this model, and compute the (annealed) topological complexity both of total critical points and of local minima, in the Mezard-Parisi high dimensional limit. Our main result confirms the recent formulas by Fyodorov and Le Doussal (2020). It gives the phase diagram and identifies the boundary between simple and glassy phases. Our approach relies on new exponential asymptotics of random determinants, for non-invariant random matrices.

[This is joint work with Paul Bourgade and Benjamin McKenna (Courant Institute, NYU).]

---

**High-dimensional interpolators: From linear regression to neural tangent models**
Andrea Montanari*
Stanford University

**Abstract:** Modern machine learning methods --most noticeably multi-layer neural networks - require to fit highly non-linear models comprising tens of thousands to millions of parameters. However, little attention is paid to the regularization mechanism to control model’s complexity and the resulting models are often so complex as to achieve vanishing training error. Despite this, these models generalize well to unseen data : they have small test error. I will discuss several examples of this phenomenon, leading to two-layers neural networks in the so-called lazy regime. For these examples precise asymptotics could be determined mathematically, using tools from random matrix theory, and a unifying picture is emerging. A common feature is the fact that a complex unregularized nonlinear model becomes essentially equivalent to a simpler model, which is however regularized in a non-trivial way.

[Based on joint papers with: Michael Celentano, Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, Feng Ruan, Youngtak Sohn, Jun Yan, Yiqiao Zhong]
**Simplicity and complexity of belief-propagation**

Elchanan Mossel*

MIT

**Abstract:** Belief Propagation is a very simple and popular algorithm for the inference of posteriors for probability models on trees based on iteratively applying Bayes’ rule. It is widely used in coding theory, in machine learning, in evolutionary inference, among other areas. We will survey the distributional properties and statistical efficiency of Belief Propagation in some of the simplest models and applications of these to phylogenetic reconstruction and to detection of block models. Finally, we will discuss the computational complexity of this seemingly simple algorithm.

**Gambler’s ruin problems**

Laurent Saloff-Coste*

Cornell University

**Abstract:** The classical gambler’s ruin problem asks for the probability that player A wins all the money in a fair game between two players, A and B. For this lecture, our starting point is a fair game of this sort involving three players, A, B, and C, holding a total on N tokens. That’s already quite interesting. More generally, I will discuss techniques that allow us to understand the behavior of certain finite Markov chains before the time the chain is absorbed at a given boundary.

[This is based on joint work with Persi Diaconis and Kelsey Houston-Edwards.]

**Selective inference for trees**

Daniela Witten*

University of Washington

**Abstract:** As datasets grow in size, the focus of data collection has increasingly shifted away from testing pre-specified hypotheses, and towards hypothesis generation. Researchers are often interested in performing an exploratory data analysis to generate hypotheses, and then testing those hypotheses on the same data. Unfortunately, this type of ‘double dipping’ can lead to highly-inflated Type 1 errors. In this talk, I will consider double-dipping on trees. First, I will focus on trees generated via hierarchical clustering, and will consider testing the null hypothesis of equality of cluster means. I will propose a test for a difference in means between estimated clusters that accounts for the cluster estimation process, using a selective inference framework. Second, I'll consider trees generated using the CART procedure, and will again use selective inference to conduct inference on the means of the terminal nodes. Applications include single-cell RNA-sequencing data and the Box Lunch Study.

[This is collaborative work with Lucy Gao (U. Waterloo), Anna Neufeld (U. Washington), and Jacob Bien (USC).]
Public Lecture

Structure and randomness in data
Young-Han Kim*
UCSD and Gauss Labs Inc.

Abstract: In many engineering applications ranging from communications and networking to compression and storage to artificial intelligence and machine learning, the main goal is to reveal, exploit, or even design structure in apparently random data. This talk illustrates the art and science of such information processing techniques through a variety of examples, with a special focus on data storage systems from memory chips to cloud storage platforms.

Keywords: Information theory, Noise, Manufacturing, Computer vision, Distributed computing, Probability laws
INVITED SESSION
Asymptotics of determinants of discrete Laplacians
Konstantin Izyurov*
University of Helsinki

Abstract: The zeta-regularized determinants of Laplace-Beltrami operators play an important role in analysis and mathematical physics. We show that for Euclidean surfaces with conical singularities that are glued of finitely many equal equilateral triangles or squares, these determinants appear in the asymptotic expansions of the determinants of discrete Laplacians, as the mesh size of a lattice discretization of the surface tends to zero. This establishes a particular case of a conjecture by Cardy and Peschel on the behavior of partition functions of critical lattice models, and their relation to partition functions of underlying Conformal field theories. Joint work with Mikhail Khristoforov.

Keywords: Statistical mechanics, Random Walks, Laplacian

On Loewner evolutions with jumps
Eveliina Peltola*
Rheinische Friedrich-Wilhelms-Universität Bonn

Abstract: I discuss the behavior of Loewner evolutions driven by a Levy process. Schramm’s celebrated version (Schramm-Loewner evolution), driven by standard Brownian motion, has been a great success for describing critical interfaces in statistical physics. Loewner evolutions with other random drivers have been proposed, for instance, as candidates for finding extremal multifractal spectra, and some tree-like growth processes in statistical physics. Questions on how the Loewner trace behaves, e.g., whether it is generated by a (discontinuous) curve, whether it is locally connected, tree-like, or forest-like, have been partially answered in the symmetric alpha-stable case. We shall consider the case of general Levy drivers.

[Joint work with Anne Schreuder (Cambridge).]

Extremal distance and conformal radius of a CLE4 loop
Titus Lupu1*, Juhan Aru2 and Avelio Sepúlveda3
1Laboratoire de Probabilités, Statistique et Modélisation, CNRS / Sorbonne Université, France
2Mathematics, EPFL, Switzerland
3Institut Camille Jordan, Université Lyon 1 Claude Bernard, France

Abstract: Consider CLE4 in the unit disk and let be the loop of the CLE4 surrounding the origin. Schramm, Sheffield and Wilson determined the law of the conformal radius seen from the origin of the domain surrounded by this loop. We complement their result by determining the law of the extremal distance between the loop and the boundary of the unit disk. More surprisingly, we also compute the joint law of these conformal radius and extremal distance. This law involves first and last hitting times of a one-dimensional Brownian motion. Similar techniques also allow us to determine joint laws of some extremal distances in a critical Brownian loop-soup cluster.

[This is a joint work with Juhan Aru (EPFL) and Avelio Sepúlveda (Université Lyon 1 Claude Bernard).]

Keywords: Schramm-Loewner evolution, Conformal loop ensemble, Gaussian free field, Local set
Exceptional geodesic pairs in the directed landscape
Erik Bates*
University of Wisconsin-Madison

Abstract: Within the Kardar-Parisi-Zhang universality class, the space-time Airy sheet is conjectured to be the canonical scaling limit for last passage percolation models. In recent work of Dauvergne, Ortmann, and Virág, this object was constructed and shown to be the limit after parabolic correction of one such model: Brownian last passage percolation. This limit object, called the directed landscape, admits geodesic paths between any two space-time points \((x, s)\) and \((y, t)\) with \(s < t\). Here we examine fractal properties of the set of these paths. Our main results concern exceptional endpoints admitting disjoint geodesics. First, we fix two distinct starting locations \(x_1\) and \(x_2\), and consider geodesics traveling \((x_1, 0) \to (y, 1)\) and \((x_2, 0) \to (y, 1)\). We prove that the set of \(y \in \mathbb{R}\) which these geodesics coalesce only at time 1 has Hausdorff dimension one-half. Second, we consider endpoints \((x, 0)\) and \((y, 1)\) between which there exist two geodesics intersecting only at times 0 and 1. We prove that the set of such \((x, y) \in \mathbb{R}^2\) also has Hausdorff dimension one-half. The proofs require several inputs of independent interest, including (i) connections to the so-called difference weight profile studied by Basu, Ganguly, and Hammond; and (ii) a tail estimate on the number of disjoint geodesics starting and ending in small intervals. The latter result extends the analogous estimate proved for the prelimiting model by Hammond.

[This talk is based on joint work with Shirshendu Ganguly and Alan Hammond.]

Keywords: Brownian last passage percolation, Directed landscape, Airy sheet, Polymers

Disorder relevance and the continuum random field Ising model
Adam Bowditch*
University College Dublin

Abstract: Since its introduction by Lenz in 1920, the Ising model has been one of the most studied statistical mechanics models. It has been particularly central in the theory of critical phenomena since Peierls famously proved that it undergoes a phase transition in dimension at least 2. We discuss the long considered question of whether this picture is changed by the addition of disorder acting as a small random external field and whether the model admits a disordered continuum limit.

Keywords: Disordered, Continuum, Critical, Ising

A CLT for KPZ on torus
Yu Gu\(^*\) and Tomasz Komorowski\(^2\)
1Math, CMU, USA
2Math, Polish Academy of Sciences, Poland

Abstract: I will present a joint work with Tomasz Komorowski on proving a central limit theorem for the KPZ equation on torus.

Keywords: KPZ equation, Directed polymer, Invariant measure
SDEs driven by multiplicative stable-like Levy processes
Zhen-Qing Chen*
University of Washington

Abstract: In this talk, I will present results on weak as well as strong well-posedness results for solutions to time-inhomogeneous SDEs driven by stable-like Levy processes with Holder continuous coefficients. The Levy measure of the Levy process can be anisotropic and singular with respect to the Lebesgue measure on $R^d$ and its support can be a proper subset of $R^d$. Based on joint work with Xicheng Zhang and Guohuan Zhao.

Keywords: SDE, Levy process, Strong solution, Weak solution

Periodic homogenization of non-symmetric Lévy-type processes
Takashi Kumagai*, Xin Chen2, Zhenqing Chen3 and Jian Wang4
1RIMS, Kyoto University, Japan
2Department of Mathematics, Shanghai Jiao Tong University, China
3Department of Mathematics, University of Washington, USA
4College of Mathematics and Informatics, Fujian Normal University, China

Abstract: We study homogenization problem for strong Markov processes on $R^d$ having infinitesimal generators in periodic media, where can be singular with respect to the Lebesgue measure. Under a proper scaling, we show the scaled processes converge weakly to Lévy processes on $R^d$. In particular, we completely characterize the limiting processes when $b(x)$ is bounded continuous, $k(x,z)$ is non-negative bounded continuous and 1-periodic both for $x$ and $z$, and in spherical coordinate $\rho_0$ being any finite measure on the unit sphere. Different scaling limits appear depending on the values of $\alpha$.

Keywords: Homogenization, Lévy-type process, Martingale problem, Corrector

Optimal Hardy identities and inequalities for the fractional Laplacian on $L^p$
Krzysztof Bogdan*
Wrocław University of Science and Technology

Abstract: We will present a route from symmetric Markovian semigroups to Hardy inequalities, to nonexplosion and contractivity results for Feynman-Kac semigroups on $L^p$. We will focus on the fractional Laplacian on $R^d$, in which case the constants, estimates of the Feynman-Kac semigroups and tresholds for contractivity and explosion are sharp. [Namely we will discuss selected results from joint work with Bartłomiej Dyda, Tomasz Grzywny, Tomasz Jakubowski, Panki Kim, Julia Lenczewska, Katarzyna Pietruska-Pałuba and Dominika Piłarczyk (see arXiv).]

Keywords: Fractional Laplacian, Markov semigroup, $L^p$ space, Hardy inequality
Mapping genetic ancestors
Graham Coop*
University of California at Davis

Abstract: Spatial patterns in genetic diversity are shaped by the movements of individuals dispersing from their parents and populations expanding and contracting. It has long been appreciated that these patterns of movement leave shape the underlying genealogies along the genome leading to geographic patterns of isolation by distance in contemporary population genetic data. The enormous amount of information contained in genealogies along recombining sequences has, up till now, not been amenable to this approach. However, it is now possible to infer a sequence of gene genealogies along a recombining sequence. Here we capitalize on this important advance and develop methods to use thousands of trees to estimate time-varying per-generation dispersal rates and to locate the genetic ancestors of a sample back through time. We take a likelihood approach using a simple approximate spatial model (Branching Brownian Motion) as our prior distribution of genealogies. After testing our method with simulations we apply it to the 1001 Genomes dataset of over one thousand Arabidopsis thaliana genomes sampled across a wide geographic extent. We detect a very high dispersal rate in the recent past, especially longitudinally, and use inferred ancestor locations to visualize many examples of recent long-distance dispersal and recent admixture events. We also use inferred ancestor locations to identify the origin and ancestry of the North American expansion, to depict alternative geographic ancestries stemming from multiple glacial refugia. Our method highlights the huge amount of largely untapped information about past dispersal events and population movements contained in genome-wide genealogies.

Cellular point processes: quantifying cell signaling
Barbara Engelhardt*
Princeton University

Abstract: This talk does not have an abstract.

Fitting stochastic epidemic models to gene genealogies using linear noise approximation
Vladimir Minin*
University of California, Irvine

Abstract: Phylodynamics is a set of population genetics tools that aim at reconstructing demographic history of a population based on molecular sequences of individuals sampled from the population of interest. One important task in phylodynamics is to estimate changes in (effective) population size. When applied to infectious disease sequences such estimation of population size trajectories can provide information about changes in the number of infections. To model changes in the number of infected individuals, current phylodynamic methods use non-parametric approaches (e.g., Bayesian curve-fitting based on change-point models or Gaussian process priors), parametric approaches (e.g., based on differential equations), and stochastic modeling in conjunction with likelihood-free Bayesian methods. The first class of methods yields results that are hard to interpret epidemiologically. The second class of methods provides estimates of important epidemiological parameters, such as infection and removal/recovery rates, but ignores variation in the dynamics of infectious disease spread. The third class of methods
is the most advantageous statistically, but relies on computationally intensive particle filtering techniques that limits its applications. We propose a Bayesian model that combines phylodynamic inference and stochastic epidemic models, and achieves computational tractability by using a linear noise approximation (LNA) - a technique that allows us to approximate probability densities of stochastic epidemic model trajectories. LNA opens the door for using modern Markov chain Monte Carlo tools to approximate the joint posterior distribution of the disease transmission parameters and of high dimensional vectors describing unobserved changes in the stochastic epidemic model compartment sizes (e.g., numbers of infectious and susceptible individuals). We illustrate our new method by applying it to Ebola genealogies estimated using viral genetic data from the 2014 epidemic in Sierra Leone and Liberia.

**Keywords:** Infectious disease dynamics, Bayesian inference, Markov chain Monte Carlo, Coalescent

---

**Global rates of convergence in mixture density estimation**

Arlene Kyoung Hee Kim*

Korea University

**Abstract:** In this talk, we consider estimating a monotone decreasing density $f_0$ represented by a scale mixture of uniform densities. We first derive a general bound on the Hellinger accuracy of the MLE over convex classes. Using this bound with an entropy calculation, we provide a different proof for the convergence of the MLE for $d=1$. Then we consider a possible multidimensional extension. We can prove, for $d \geq 2$, that the rate is as conjectured by Pavlides and Wellner under the assumption that the density is bounded from above and below and supported on a compact region. We are exploring strategies for weakening the assumptions.

**Keywords:** Scale mixture of uniforms, Hellinger accuracy, Maximum likelihood estimator, Bracketing entropy

---

**Convex regression in multidimensions**

Adityanand Guntuboyina*

University of California Berkeley

**Abstract:** I will present results on the rates of convergence of the least squares estimator for multidimensional convex regression with polytopal domains. Our results imply that the least squares estimator is minimax suboptimal when the dimension exceeds 5.

[This is joint work with Gil Kur, Frank Fuchang Gao and Bodhisattva Sen.]

---

**Multiple isotonic regression: limit distribution theory and confidence intervals**

Qiyang Han*, Hang Deng and Cun-Hui Zhang

Rutgers University, USA

**Abstract:** In the first part of the talk, we study limit distributions for the tuning-free max-min block estimators in multiple isotonic regression under both fixed lattice design and random design settings. We show that at a fixed interior point in the design space, the estimation error of the max-min block estimator converges in distribution to a non-Gaussian limit at certain rate depending on the number of vanishing derivatives and certain effective dimension and sample size that drive the asymptotic theory. The limiting distribution can be viewed as a generalization of the well-known Chernoff distribution in univariate problems. The convergence rate is optimal in a local asymptotic
In the second part of the talk, we demonstrate how to use this limiting distribution to construct tuning-free pointwise nonparametric confidence intervals in this model, despite the existence of an infinite-dimensional nuisance parameter in the limit distribution that involves multiple unknown partial derivatives of the true regression function. We show that this difficult nuisance parameter can be effectively eliminated by taking advantage of information beyond point estimates in the block max-min and min-max estimators through random weighting. Notably, the construction of the confidence intervals, even new in the univariate setting, requires no more efforts than performing an isotonic regression for once using the block max-min and min-max estimators, and can be easily adapted to other common monotone models.

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

**Keywords:** Network, Mcmc, Inference, Community detection

**Adversarial classification, optimal transport, and geometric flows**

Nicolas Garcia Trillos*
University of Wisconsin-Madison, USA

**Abstract:** The purpose of this talk is to provide an explicit link between the three topics that form the talk’s title, and to introduce a new perspective (more dynamic and geometric) to understand robust classification problems. For concreteness, we will discuss a version of adversarial classification where an adversary is empowered to corrupt data inputs up to some distance $\varepsilon$. We will first describe necessary conditions associated with the optimal classifier subject to such an adversary. Then, using the necessary conditions we derive a geometric evolution equation which can be used to track the change in classification boundaries as $\varepsilon$ varies. This evolution equation may be described as an uncoupled system of differential equations in one dimension, or as a mean curvature type equation in higher dimension. In one dimension we rigorously prove that one can use the initial value problem starting from $\varepsilon = 0$, which is simply the Bayes classifier, to solve for the global minimizer of the adversarial problem. Global optimality is certified using a duality principle between the original adversarial problem and an optimal transport problem. Several open questions and directions for further research will be discussed.
Capturing network effect via fused lasso penalty with application on shared-bike data
Yunjin Choi*
University of Seoul

Abstract: Given a dataset with network structures, one of the common research interests is to model nodal features accounting for network effects. In this study, we investigate shared-bike data in Seoul, under a spatial network framework focusing on the rental counts of each station. Our proposed method models rental counts via a generalized linear model with regularizations. The regularization is made via fused lasso penalty which is devised to capture network effect. In this model, parameters are posed in a station-specific manner. The fused lasso penalty terms are applied on the parameters associated with locationally nearby stations. This approach facilitates parameters corresponding to neighboring stations to have the same value and account for underlying network effect in a data-adaptive way. The proposed method shows promising results.

Distribution-free robust linear regression
Nikita Zhivotovskiy*
ETH Zurich

Abstract: We study random design linear regression with no assumptions on the distribution of the covariates and with a heavy-tailed response variable. When learning without assumptions on the covariates, we establish boundedness of the conditional second moment of the response variable as a necessary and sufficient condition for achieving deviation-optimal excess risk rate of convergence. In particular, combining the ideas of truncated least squares, median-of-means procedures and aggregation theory, we construct a non-linear estimator achieving excess risk of order d/n with the optimal sub-exponential tail. While the existing approaches to learning linear classes under heavy-tailed distributions focus on proper estimators, we highlight that the improperness of our estimator is necessary for attaining non-trivial guarantees in the distribution-free setting considered in this work. Finally, as a byproduct of our analysis, we prove an optimal version of the classical bound for the truncated least squares estimator due to Györfi, Kohler, Krzyzak, and Walk.

Keywords: Least squares, Robust estimation, Improper learning

Algorithmic high-dimensional robust statistics
Ilias Diaconicolas*
University of Wisconsin-Madison

Abstract: Fitting a model to a collection of observations is one of the quintessential questions in statistics. The standard assumption is that the data was generated by a model of a given type (e.g., a mixture model). This simplifying assumption is at best only approximately valid, as real datasets are typically exposed to some source of contamination. Hence, any estimator designed for a particular model must also be robust in the presence of corrupted data. This is the prototypical goal in robust statistics, a field that took shape in the 1960s with the pioneering works of Tukey and Huber. Until recently, even for the basic problem of robustly estimating the mean of a high-dimensional dataset, all known robust estimators were hard to compute. Moreover, the quality of the common heuristics degrades badly as the dimension increases. In this talk, we will survey the recent progress in algorithmic high-dimensional robust statistics. We will describe the first computationally efficient algorithms for robust mean and
covariance estimation and the main insights behind them. We will also present practical applications of these estimators to exploratory data analysis and adversarial machine learning. Finally, we will discuss new directions and opportunities for future work.

**Keywords:** Robust Statistics, High Dimensions, Unsupervised Learning, Efficient Algorithms, Mean and Covariance Estimation

**Robust estimation of a mean vector with respect to any norm: a minimax MOM and a Stahel-Donoho Median of means estimators**

Guillaume Lecué*

Laboratoire de statistiques at CREST

**Abstract:** We consider the problem of robust mean estimation w.r.t. any pseudo-norm of the form

\[ x \in \mathbb{R}^d \rightarrow \|x\|_S = \sup_{v \in S} \|x\|_S \]

where \( S \) is any symmetric subset of \( \mathbb{R}^d \). We show that the deviation-optimal minimax subgaussian rate for confidence \( 1 - \delta \) is

\[
\max \left( \frac{\ell^* \left( \Sigma^{1/2} S \right)}{\sqrt{N}}, \sup_{v \in S} \|\Sigma^{1/2} v\|_2 \sqrt{\frac{\log(1/\delta)}{N}} \right)
\]

where \( \ell^*(\Sigma^{1/2} S) \) is the Gaussian mean width of \( \Sigma^{1/2} S \) and \( \Sigma \) is the covariance matrix of the data (in the benchmark i.i.d. Gaussian case). We also show that this rate can be achieved by a solution to a convex optimization problem in the adversarial and \( \ell_2 \) heavy-tailed setup by considering minimum of some Fenchel-Legendre transforms constructed using the Median-of-means principle. In the special case of the norm \( \|\Sigma^{-1/2} \cdot \|_2 \) we construct a Median-of-means version of the Stahel-Donoho median which does not require to know or estimate \( \Sigma \) and show it achieves the optimal rate mentioned above. Our analysis also covers cases where the mean does not even exist but a location parameter does; in those cases we still recover the same subgaussian rates and the same price for adversarial contamination even though there is not even a first moment.

[Joint works with Jules Depersin.]

**IS08 Functional Data Analysis**

**Partially specified covariance operators and intrinsically functional graphical models**

Victor Panaretos*

EPFL

**Abstract:** Motivated by the problem of covariance recovery from functional fragments, we consider the problem of completing a partially specified covariance kernel on the unit square. By representing the underlying stochastic process as an undirected graphical model with uncountable vertices and edges, we show that a canonical completion always exists and can be explicitly described, under weak assumptions. For partial covariances specified on nearly banded domains containing the diagonal, we present necessary and sufficient conditions for unique completion, and characterise all completions under non-uniqueness. Finally, we show how the estimation of the canonical completion reduces to a system of ill-posed linear inverse problems in the space of Hilbert-Schmidt operators, and derive rates of convergence under standard source conditions.

[Based on joint work with K. Waghmare (EPFL).]

**Keywords:** Covariance completion, Functional fragments
Domain selection for functional linear models: a dynamic RKHS approach
Jane-Ling Wang* and Shu-Chin Lin
Department of Statistics, UC Davis, USA

Abstract: In conventional scalar-on-function linear regression model, the entire trajectory of the predictor process on the whole domain is used to model the response variable. However, the response may only be associated with the covariate process $X$ on a subdomain. We consider the problem of estimating the domain of association when assuming that the regression coefficient function is nonzero on a subinterval. We propose a solution based on the reproducing kernel Hilbert space (RKHS) approach to estimate both the domain and the regression function. A simulation study illustrates the effectiveness of the proposed approach. Asymptotic theory is developed for both estimators.

Keywords: Functional data, Reproducing kernel Hilbert space, Representer's theorem, Sobolev space

Simultaneous Inference for function-valued parameters: A fast and fair approach
Dominik Liebl*
University of Bonn

Abstract: Quantifying uncertainty using confidence regions is a central goal of statistical inference. Despite this, methodologies for confidence bands in Functional Data Analysis are underdeveloped compared to estimation and hypothesis testing. This work represents a major leap forward in this area by presenting a new methodology for constructing simultaneous confidence bands for functional parameter estimates. These bands possess a number of striking qualities: (1) they have a nearly closed-form expression, (2) they give nearly exact coverage, (3) they have a finite sample correction, (4) they do not require an estimate of the full covariance of the parameter estimate, and (5) they can be constructed adaptively according to a desired criteria. One option for choosing bands we find especially interesting is the concept of fair bands which allows us to do fair (or equitable) inference over subintervals and could be especially useful in longitudinal studies over long time scales. Our bands are constructed by integrating and extending tools from Random Field Theory, an area that has yet to overlap with Functional Data Analysis.

Estimation of quantum state and quantum channel
Masahito Hayashi*
Southern University of Science and Technology

Abstract: We review the results of quantum state estimation by using Cramer-Rao type bound. The quantum channel estimation has two scaling, standard quantum limit and Heisenberg scaling. In the standard quantum limit, we present that the extension of the above method works well. For the Heisenberg scaling, we present that the method of Fourier transform method works.

Keywords: Carmer-Rao type bound, Heisenberg scaling, Standard quantum limit, Fourier transform method
Information geometry and local asymptotic normality for quantum Markov processes
Madalin Guta*
University of Nottingham

Abstract: This talk deals with the problem of identifying and estimating dynamical parameters of quantum Markov processes, in the input-output formalism. I will discuss several aspects of this problem: The first aspect concerns the structure of the space of identifiable parameters for ergodic dynamics, assuming full access to the output state for arbitrarily long times. I will show that the equivalence classes of undistinguishable parameters are orbits of a Lie group acting on the space of dynamical parameters. The second aspect concerns the information geometric structure on this space. I will show that the space of identifiable parameters and carries a Riemannian metric based on the quantum Fisher information of the output. The metric can be computed explicitly in terms of the Markov covariance of certain fluctuation operators. The third aspect concerns the asymptotic statistical structure of the output state. I will show that the output state satisfy local asymptotic normality, i.e. they can be approximated by a Gaussian model constructed from the Markov covariance data.

Keywords: Quantum Markov process, Local asymptotic normality, Information geometry

Optimal adaptive strategies for sequential quantum hypothesis testing
Marco Tomamichel*
National University of Singapore

Abstract: We consider sequential hypothesis testing between two quantum states using adaptive and non-adaptive strategies. In this setting, samples of an unknown state are requested sequentially and a decision to either continue or to accept one of the two hypotheses is made after each test. Under the constraint that the number of samples is bounded, either in expectation or with high probability, we exhibit adaptive strategies that minimize both types of misidentification errors. Namely, we show that these errors decrease exponentially (in the stopping time) with decay rates given by the measured relative entropies between the two states. Moreover, if we allow joint measurements on multiple samples, the rates are increased to the respective quantum relative entropies. We also fully characterize the achievable error exponents for non-adaptive strategies and provide numerical evidence showing that adaptive measurements are necessary to achieve our bounds under some additional assumptions.

Keywords: Hypothesis testing, Quantum states, Adaptive tests

Two-sample tests for relevant differences in the eigenfunctions of covariance operators
Alexander Aue*, Holger Dette2 and Gregory Rice3
1Department of Statistics, UC Davis, USA
2Department of Mathematics, Ruhr-University Bochum, Germany
3Department of Statistics and Actuarial Science, University of Waterloo, Canada

Abstract: This talk deals with two-sample tests for functional time series data, which have become widely available in conjunction with the advent of modern complex observation systems. Here, particular interest is in evaluating whether two sets of functional time series observations share the shape of their primary modes of variation as encoded by the eigenfunctions of the respective covariance operators. To this end, a novel testing approach is introduced that connects with, and extends, existing literature in two main ways. First, tests are set up in the relevant
testing framework, where interest is not in testing an exact null hypothesis but rather in detecting deviations deemed sufficiently relevant, with relevance determined by the practitioner and perhaps guided by domain experts. Second, the proposed test statistics rely on a self-normalization principle that helps to avoid the notoriously difficult task of estimating the long-run covariance structure of the underlying functional time series. The main theoretical result of this paper is the derivation of the large-sample behavior of the proposed test statistics. Empirical evidence, indicating that the proposed procedures work well in finite samples and compare favorably with competing methods, is provided through a simulation study, and an application to annual temperature data.

**Keywords:** Functional data, Structural change, Time series

---

**Multiple change point detection under serial dependence**

Haeran Cho\(^1\) and Piotr Fryzlewicz\(^2\)

\(^1\)School of Mathematics, University of Bristol, UK

\(^2\)Department of Statistics, LSE, UK

**Abstract:** We propose a methodology for detecting multiple change points in the mean of an otherwise stationary, autocorrelated, linear time series. It combines solution path generation based on the wild energy maximisation principle, and an information criterion-based model selection strategy termed gappy Schwarz criterion. The former is well-suited to separating shifts in the mean from fluctuations due to serial correlations, while the latter simultaneously estimates the dependence structure and the number of change points without performing the difficult task of estimating the level of the noise as quantified e.g. by the long-run variance. We provide modular investigation into their theoretical properties and show that the combined methodology, named WEM.gSC, achieves consistency in estimating both the total number and the locations of the change points. The good performance of WEM.gSC is demonstrated via extensive simulation studies, and we further illustrate its usefulness by applying the methodology to London air quality data.

**Keywords:** Data segmentation, Autoregressive time series

---

**An asymptotic test for constancy of the variance in a time series**

Herold Dehling\(^1\), Sara Schmidt\(^1\), Max Wornowizki\(^2\) and Roland Fried\(^2\)

\(^1\)Department of Mathematics, Ruhr-University Bochum, Germany

\(^2\)Department of Statistics, TU Dortmund University, Germany

**Abstract:** We present a novel approach to test for heteroscedasticity of a non-stationary time series that is based on Gini’s mean difference of logarithmic local sample variances. In order to analyse the large sample behaviour of our test statistic, we establish new limit theorems for U-statistics of dependent triangular arrays. We derive the asymptotic distribution of the test statistic under the null hypothesis of a constant variance and show that the test is consistent against a large class of alternatives, including multiple structural breaks in the variance. Our test is applicable even in the case of non-stationary processes, assuming a locally varying mean function. The performance of the test and its comparatively low computation time are illustrated in an extensive simulation study.

**Keywords:** Change-point analysis, Test for heteroscedasticity, U-statistic of triangular arrays
Statistical learning with spatially dependent high-dimensional data
Taps Maiti*
Michigan State University

Abstract: The rapid development of information technology is making it possible to collect massive amounts of multidimensional, multimodal data with high dimensionality in diverse fields of science and engineering. New statistical and machine learning methods have been developing continuously to solve challenging problems arising out of these complex systems. This talk will discuss a specific type of statistical learning, namely feature selection and classification, when the features are multidimensional. More specifically, they are spatio-temporal by nature. Various machine learning techniques are suitable for this problem, although their underlying statistical theories are not well established. We start with linear discriminant analysis under spatial dependence, establish its statistical properties and then connecting to other machine learning tools for flexible data analysis in the context of brain imaging data.

Keywords: Bayes Risk, Covariance, LDA, High-dimension

Large-scale spatial data science with ExaGeoStat
Marc Genton*
King Abdullah University of Science and Technology (KAUST)

Abstract: Spatial data science aims at analyzing the spatial distributions, patterns, and relationships of data over a predefined geographical region. For decades, the size of most spatial datasets was modest enough to be handled by exact inference. Nowadays, with the explosive increase of data volumes, High-Performance Computing (HPC) can serve as a tool to handle massive datasets for many spatial applications. Big data processing becomes feasible with the availability of parallel processing hardware systems such as shared and distributed memory, multiprocessors and GPU accelerators. In spatial statistics, parallel and distributed computing can alleviate the computational and memory restrictions in large-scale Gaussian process inference and prediction. In this talk, we will describe cutting-edge HPC techniques and their applications in solving large-scale spatial problems with the new software ExaGeoStat.

Keywords: HPC, Large Datasets, Spatial Statistics, Supercomputer

Multivariate spatio-temporal Hawkes process models of terrorism
Mikyoung Jun†* and Scott Cook†
†Mathematics, University of Houston, USA
†Political Science, Texas A&M University, USA

Abstract: We develop a flexible bivariate spatio-temporal Hawkes process model to analyze patterns of terrorism. Previous applications of point process methods to political violence data have mainly utilized temporal Hawkes process models, neglecting spatial variation in these attack patterns. This limits what can be learned from these models, as any effective counter-terrorism strategy requires knowledge on both when and where attacks are likely to occur. Even the existing work that does exist on spatio-temporal Hawkes processes impose restrictions on the triggering function that are not well-suited for terrorism data. Therefore, we generalize the structure of the spatio-temporal triggering function considerably, allowing for nonseparability, nonstationarity, and cross-triggering (i.e., across the groups). To demonstrate the utility of our model, we analyze two samples of real-world terrorism data:
Afghanistan (2002-2013) and Nigeria (2010-2017). Jointly, these two studies demonstrate that our model dramatically outperforms standard Hawkes process models, besting widely-used alternatives in overall model fit and revealing spatio-temporal patterns that are, by construction, masked in these models (e.g., increasing dispersion in cross-triggering over time).

[This is joint work with Scott Cook.]

**Keywords:** Terrorism, Hawkes processes, Multivariate point process, Spatio-temporal point patterns

---

**Wasserstein regression**

Hans-Georg Müller¹*, Yaqing Chen¹ and Zhenhua Lin²

¹Department of Statistics, University of California, Davis, USA
²Department of Statistics, National University of Singapore, Singapore

**Abstract:** The analysis of samples of random objects that do not lie in a vector space has found increasing attention in statistics in recent years. An important class of such object data are univariate probability measures defined on the real line. Adopting the Wasserstein metric, we develop a class of regression models for data that include random distributions as predictors and distributions or scalars as responses. To study these regression models, we utilize the geometry of tangent bundles of the metric space of random measures with the Wasserstein metric and derive asymptotic rates of convergence for estimators of the regression coefficient function and for predicted distributions. We also study an extension to autoregressive models for distribution-valued time series. The proposed methods are illustrated with data that include distributional components in various regression settings.

**Keywords:** Distribution regression, Distribution time series, Wasserstein geometry, Tangent bundles

---

**Finite sample smeariness for Fréchet means**

Stephan Huckemann¹*, Benjamin Eltzner² and Shayan Hundrieser³

¹Felix-Bernstein-Institute for Mathematical Statistics in the Biosciences, Georg-August-Universitaet Goettingen, Germany
²Research Group Computational Biomolecular Dynamics, Max Planck Institute for Biophysical Chemistry, Germany
³Institute for Mathematical Stochastics, Georg-August-Universitaet Goettingen, Germany

**Abstract:** It is well known for the Euclidean setting that a variety of statistical asymptotic tests, e.g. T-tests or MANOVA, are robust under nonnormality. It is much less known, that this cannot be taken for granted, for similar tests based on manifolds data, in particular for data on compact spaces. The reason lies in a recently discovered phenomenon: Smeariness lowers the classical square-root-of-n-rate for Fréchet means. While true smeariness is only present for a nullset of most parametric families, it surfaces in a finite sample regime for a large class of distributions: For instance, all nontrivial distributions on spheres are affected and all distributions on circles whose support extends beyond a half circle, like, e.g. all Fisher-von-Mises distributions. We give finite sample smeariness a precise definition and illustrate some effects in theory and practice. In particular, the presence of finite sample smeariness renders tests based on quantiles of asymptotic distributions ineffective up to considerably high sample sizes. Suitably designed bootstrap tests remain valid, however.

**Keywords:** Asymptotic tests, Manifold data, Convergence rate
Score matching for microbiome compositional data
Janice Scealy*
Australian National University

**Abstract:** Compositional data and multivariate count data with known totals are challenging to analyze due to the non-negativity and sum constraint on the sample space. It is often the case with microbiome compositional data that many of the components are highly right-skewed, with large numbers of zeros. A major limitation of currently available estimators for compositional models is that they either cannot handle many zeros in the data or are not computationally feasible in moderate to high dimensions. We derive a new set of novel score matching estimators applicable to distributions on a Riemannian manifold with boundary, of which the standard simplex is a special case. The score matching method is applied to estimate the parameters in a new flexible model for compositional data and we show that the estimators are scalable and available in closed form. We apply the new model and estimators to real microbiome compositional data and show that the model provides a good fit to the data.

Recent results for critical lattice models in high dimensions
Mark Holmes*
University of Melbourne

**Abstract:** We'll discuss recent results concerning the limiting behaviour of critical lattice models (e.g. lattice trees and the voter model) in high dimensions. In particular (i) results with Ed Perkins on the scaling limit of the range (the set of vertices ever visited/occupied by the model), and (ii) results with Cabezas, Fribergh, and Perkins on weak convergence of the historical processes and of random walks on lattice trees.

**Keywords:** Range, Historical process, Super-Brownian motion

Near-critical avalanches in 2D frozen percolation and forest fires
Pierre Nolin*
City University of Hong Kong

**Abstract:** We discuss two closely related processes in two dimensions: frozen percolation, where connected components of occupied vertices freeze (they stop growing) as soon as they reach a given size, and forest fire processes, where connected components are hit by lightning (and thus become entirely vacant) at a very small rate. When the density of occupied sites approaches the critical threshold for Bernoulli percolation, both processes display a striking phenomenon: the appearance of what we call "near-critical avalanches". We study these avalanches, all the way up to the natural characteristic scale of each model, which constitutes an important step toward understanding the self-organized critical behavior of such processes. In the case of forest fires, it is crucial to analyze the effect of fires on the connectivity of the forest. For this purpose, a key tool is a percolation model where regions ("impurities") are removed from the lattice, in an independent fashion. The macroscopic behavior of this process is quite subtle, since the impurities are not only microscopic, but also allowed to be mesoscopic.

[This talk is based on joint works with Rob van den Berg (CWI and VU, Amsterdam) and with Wai-Kit Lam (University of Minnesota).]

**Keywords:** Near-critical percolation, Frozen percolation, Forest fires, Self-organized criticality
Quenched and annealed Ising models on random graphs
Cristian Giardinà*
Modena & Reggio Emilia University

Abstract: The ferromagnetic Ising model on a lattice is a paradigmatic model of statistical physics used to study phase transitions in lattice systems. In this talk, I shall consider the setting where the regular spatial structure of a lattice is replaced by a random graph, which is often used to model complex networks. I shall treat both the case where the graph is essentially frozen (quenched setting) and the case where instead it is rapidly changing (annealed setting). I shall prove that quenched and annealed may have different critical temperatures, provided the graph degrees are allowed to fluctuate. I shall also discuss how universal results (law of large numbers, central limit theorems, critical exponents) are affected by the disorder in the spatial structure.

[The picture that I will present emerges from several joint works, involving V.H. Can, S. Dommers, C. Giberti, R. van der Hofstad and M.L. Prioriello.]

Keywords: Ising model, Random graphs, Quenched, Annealed

Density estimation and conditional simulation using triangular transport
Youssef Marzouk†, Ricardo Baptista‡ and Olivier Zahm§
†Massachusetts Institute of Technology, USA
‡INRIA/Université Grenoble Alpes, France

Abstract: Triangular transformations of measures, such as the Knothe–Rosenblatt rearrangement, underlie many new computational approaches for density estimation and conditional simulation. This talk discusses two aspects of such constructions. First, is the problem of estimating a triangular transformation given a sample from a distribution of interest—and hence, transport-driven density estimation. We present a general functional framework for representing monotone triangular maps between distributions, and analyze properties of maximum likelihood estimation in this framework. We demonstrate that the associated optimization problem is smooth and, under appropriate conditions, has no spurious local minima. This result provides a foundation for a greedy semi-parametric estimation procedure. Second, we discuss a conditional simulation method that employs a specific composition of maps, derived from the Knothe-Rosenblatt rearrangement, to push forward a joint distribution to any desired conditional. We show that this composed-map approach reduces variability in conditional density estimates and reduces the bias associated with any approximate map representation. Moreover, this approach motivates alternative estimation objectives that focus on the removal of dependence. For context, and as a pointer to an interesting application domain, we elucidate links between conditional simulation with composed maps and the ensemble Kalman filter.

Keywords: Transport maps, Density estimation, Conditional simulation, Likelihood-free inference

Estimation of Wasserstein distances in the spiked transport model
Jonathan Niles-Weed¶ and Philippe Rigollet∥
¶Courant Institute of Mathematical Sciences/Center for Data Science, New York University, USA
∥Department of Mathematics, MIT, USA

Abstract: We propose a new statistical model, the spiked transport model, which formalizes the assumption that two
probability distributions differ only on a low-dimensional subspace. We study the minimax rate of estimation for the Wasserstein distance under this model and show that this low-dimensional structure can be exploited to avoid the curse of dimensionality. As a byproduct of our minimax analysis, we establish a lower bound showing that, in the absence of such structure, the plug-in estimator is nearly rate-optimal for estimating the Wasserstein distance in high dimension. We also give evidence for a statistical-computational gap and conjecture that any computationally efficient estimator is bound to suffer from the curse of dimensionality.

**Keywords:** Wasserstein distance, High-dimensional statistics

---

**Statistical estimation of barycenters in metric spaces and the space of probability measures**

Quentin Paris*

Higher School of Economics

**Abstract:** The talk presents rates of convergence for empirical barycenters over a large class of geodesic spaces with curvature bounds in the sense of Alexandrov. We show that parametric rates of convergence are achievable under natural conditions that characterise the bi-extendibility of geodesics emanating from a barycenter. We show that our results apply to infinite-dimensional spaces such as the 2-Wasserstein space, where bi-extendibility of geodesics translates into regularity of Kantorovich potentials.

---

**The right complexity measure in locally private estimation: It is not the Fisher information**

John Duchi*

Stanford University

**Abstract:** We identify fundamental tradeoffs between statistical utility and privacy under local models of privacy in which data is kept private even from the statistician, providing instance-specific bounds for private estimation and learning problems by developing the local minimax risk. In contrast to approaches based on worst-case (minimax) error, which are conservative, this allows us to evaluate the difficulty of individual problem instances and delineate the possibilities for adaptation in private estimation and inference. Our main results show that the local modulus of continuity of the estimand with respect to the variation distance—as opposed to the Hellinger distance central to classical statistics—characterizes rates of convergence under locally private estimation for many notions of privacy, including differential privacy and its relaxations. As a consequence of these results, we identify an alternative to the Fisher information for private estimation, giving a more nuanced understanding of the challenges of adaptivity and optimality.

---

**Sequentially interactive versus non-interactive local differential privacy: estimating the quadratic functional**

Lukas Steinberger*

University of Vienna

**Abstract:** We develop minimax rate optimal locally differentially private procedures for estimating the integrated square of the data generating density. A sequentially interactive two-step procedure is found to outperform the best possible non-interactive method even in terms of convergence rate. This is in stark contrast to many other private
estimation problems (e.g., those where the estimand is a linear functional of the data generating distribution) where it is known that sequential interaction between data owners can not lead to faster rates of estimation than that of an optimal non-interactive method.

**Keywords:** Differential privacy, Non-parametric estimation, Quadratic functional

---

**Gaussian differential privacy**

Weijie Su*  
University of Pennsylvania

**Abstract:** Privacy-preserving data analysis has been put on a firm mathematical foundation since the introduction of differential privacy (DP) in 2006. This privacy definition, however, has some well-known weaknesses: notably, it does not tightly handle composition. In this talk, we propose a relaxation of DP that we term "f-DP", which has a number of appealing properties and avoids some of the difficulties associated with prior relaxations. First, f-DP preserves the hypothesis testing interpretation of differential privacy, which makes its guarantees easily interpretable. It allows for lossless reasoning about composition and post-processing, and notably, a direct way to analyze privacy amplification by subsampling. We define a canonical single-parameter family of definitions within our class that is termed "Gaussian Differential Privacy", based on hypothesis testing of two shifted normal distributions. We prove that this family is focal to f-DP by introducing a central limit theorem, which shows that the privacy guarantees of any hypothesis-testing based definition of privacy (including differential privacy) converge to Gaussian differential privacy in the limit under composition. This central limit theorem also gives a tractable analysis tool. We demonstrate the use of the tools we develop by giving an improved analysis of the privacy guarantees of noisy stochastic gradient descent.  
[This is joint work with Jinshuo Dong and Aaron Roth.]

**Keywords:** Differential privacy

---

**Inference for nonlinear inverse problems**

Vladimir Spokoinyi*  
WIAS and HU Berlin

**Abstract:** Bayesian methods are actively used for parameter identification and uncertainty quantification when solving nonlinear inverse problems with random noise. However, there are only few theoretical results justifying the Bayesian approach. Recent papers, see e.g. (Nickl (2017); Lu (2017) bernsteinvon) and references therein, illustrate the main difficulties and challenges in studying the properties of the posterior distribution in the nonparametric setup. This paper offers a new approach for study the frequentist properties of the nonparametric Bayes procedures. The idea of the approach is to relax the nonlinear structural equation by introducing an auxiliary functional parameter and replacing the structural equation with a penalty and by imposing a prior on the auxiliary parameter. For the such extended model, we state sharp bounds on posterior concentration and on the accuracy of the penalized MLE and on Gaussian approximation of the posterior, and a number of further results. All the bounds are given in terms of effective dimension, and we show that the proposed calming device does not significantly affect this value.

**Keywords:** Posterior, Concentration, Gaussian approximation, Bernstein–von Mises Theorem
**Change point analysis for high-dimensional data**
Xiaohui Chen* and Mengjia Yu
Statistics, University of Illinois at Urbana-Champaign, USA

Abstract: Cumulative sum (CUSUM) statistics are widely used in the change point inference and identification. For the problem of testing for existence of a change point in an independent sample generated from the mean-shift model, we introduce a Gaussian multiplier bootstrap to calibrate critical values of the CUSUM test statistics in high dimensions. The proposed bootstrap CUSUM test is fully data-dependent and it has strong theoretical guarantees under arbitrary dependence structures and mild moment conditions. Specifically, we show that with a boundary removal parameter the bootstrap CUSUM test enjoys the uniform validity in size under the null and it achieves the minimax separation rate under the sparse alternatives when the dimension p can be larger than the sample size n. Once a change point is detected, we estimate the change point location by maximizing the $\ell^\infty$-norm of the generalized CUSUM statistics at two different weighting scales corresponding to covariance stationary and non-stationary CUSUM statistics. For both estimators, we derive their rates of convergence and show that dimension impacts the rates only through logarithmic factors, which implies that consistency of the CUSUM estimators is possible when p is much larger than n. In the presence of multiple change points, we propose a principled bootstrap-assisted binary segmentation (BABS) algorithm to dynamically adjust the change point detection rule and recursively estimate their locations. We derive its rate of convergence under suitable signal separation and strength conditions. Time permitting, we may also discuss some robust extension of the change point detection problem for high-dimensional location parameters.

Keywords: Change point analysis, High-dimensional data, Bootstrap, CUSUM

**Bootstrap test for multi-scale lead-lag relationships in high-frequency data**
Yuta Koike* and Takaki Hayashi

1Graduate School of Mathematical Sciences, University of Tokyo, Japan
2Graduate School of Business Administration, Keio University, Japan

Abstract: Motivated by recent empirical findings in high-frequency financial econometrics, we consider a pair of Brownian motions having possibly different lead-lag relationships at multiple time scales. Given their discrete observation data, we aim to test at which time scales these processes have non-zero cross correlations. For this purpose, we introduce maximum type test statistics based on scale-by-scale cross covariance estimators and develop a Gaussian approximation theory for these statistics. Since their null distributions are analytically intractable, we propose a wild bootstrap procedure to approximate them. Theoretical verification of these approximations are established through recent Gaussian approximation results for high-dimensional vectors of degenerate quadratic forms.

Keywords: Bootstrap, Gaussian process, High-dimensional CLT, High-frequency data

**IS17 Approximate Bayesian Computation**

**Approximate inference for ordinal linear regression**
Jean-Luc Dortet-Bernadet*
Université de Strasbourg

Abstract: Ordinal regression remains one of the most useful methods for analysing data arising from ordered
responses, such as those typically found in opinion surveys. We consider a flexible linear ordinal regression model which, unlike the majority of existing ordinal regression models, allows to differentiate covariate effects between different levels in the response. For scalable inference, we develop a VB approach based on truncated Gaussian distributions. A real application on data arising from student satisfaction surveys is given.

[Joint work with Yanan Fan.]

**Generalized Bayesian likelihood-free inference using scoring rules estimators**
Ritabrata Dutta\(^1\) and Lorenzo Pacchiardi\(^2\)
\(^1\)Department of Statistics, University of Warwick, UK
\(^2\)Department of Statistics, University of Oxford, UK

Abstract: We propose a framework for Bayesian Likelihood-Free Inference (LFI) based on Generalized Bayesian Inference using scoring rules (SR). SR are used to evaluate probabilistic models given an observation; a proper SR is minimised in expectation when the model corresponds to the true data generating process for the observation. Use of a strictly proper SR, for which the above minimum is unique, ensures posterior consistency of our method. Further, we prove outlier robustness of our posterior for a specific SR. As the likelihood function is intractable for LFI, we employ consistent estimators of SR using model simulations in a pseudo-marginal Monte Carlo Markov chain setup; we show the target of such a chain converges to the exact SR posterior with increasing number of simulations. Furthermore, we note popular LFI techniques such as Bayesian Synthetic Likelihood (BSL) can be seen as special cases of our framework using only proper (but not strictly so) SR. We empirically validate our consistency and outlier robustness results and show how related approaches do not enjoy these properties. Practically, we use the Energy and Kernel Scores, but our general framework sets the stage for extensions with other scoring rules.

Keywords: Generalized Bayesian inference, Likelihood-free inference, Proper scoring rules, Pseudo-marginal MCMC

**Dynamics and phase transitions in deep neural networks**
Yasaman Bahri*
Google Research

Abstract: The study of deep neural networks whose hidden layer widths are large has been fruitful in building theoretical foundations for deep learning. I will begin by surveying the result of our past work along these lines. For instance, infinitely-wide deep neural networks can be exactly described by Gaussian processes with particular compositional kernels, both in their prior and predictive posterior distributions. Furthermore, such infinite-width deep networks can be exactly described as linear models under gradient descent up to a maximum learning rate. At larger learning rates with squared loss, empirical evidence suggests a phase transition to a different, nonlinear regime with universal features across architectures and datasets. I will describe our theoretical understanding of this phase transition through the study of a class of simple dynamical systems distilled from neural network evolution in function space.

Keywords: Deep neural networks, Gaussian processes, Dynamical systems
Theoretical understanding of adding noises to deep generative models
Yongdai Kim¹*, Minwoo Chae², Dongha Kim³ and Lizhen Lin⁴
¹Department of Statistics, Seoul National University, South Korea
²Department of Industrial and Management Engineering, Pohang University of Science and Technology, South Korea
³Department of Statistics, Sungshin Women’s University
⁴Department of Computational and Applied Mathematics and Statistics, University of Notre Dame

Abstract: Deep generative models have received much attention recently since they can generate realistic synthetic images. Recently, some researches have reported that adding noises to data is helpful to learn deep generative models. In this talk, we provide theoretical justifications of this method. We derive the convergence rate of the maximum likelihood estimator of a deep generative model and show that the convergence rate can be improved by adding noises in particular when the noise level of data is small.

Adversarial examples in random deep networks
Peter Bartlett¹*, Sébastien Bubeck² and Yeshwanth Cherapanamjeri³
¹Statistics and EECS, UC Berkeley, USA
²Microsoft Research
³UC Berkeley

Abstract: Because the phenomenon of adversarial examples in deep networks poses a serious barrier to the reliable and robust application of this methodology, there has been considerable interest in why it arises. We consider ReLU networks of constant depth with independent gaussian parameters, and show that small perturbations of input vectors lead to large changes of outputs. Building on insights of Daniely and Schacham (2020) and of Bubeck et al (2021), we show that adversarial examples arise in these networks because the functions that they compute are very close to linear. The main result is for networks with constant depth, but we also show that some constraint on depth is necessary for a result of this kind: there are suitably deep networks that, with constant probability, compute a function that is close to constant.

Is your distribution in shape?
Ronitt Rubinfeld*
Massachusetts Institute of Technology

Abstract: Algorithms for understanding data generated from distributions over large discrete domains are of fundamental importance. We consider the sample complexity of “property testing algorithms” that seek to distinguish whether or not an underlying distribution satisfies basic shape properties. Examples of such properties include convexity, log-concavity, heavy-tailed, and approximability by $k$-histogram functions. In this talk, we will focus on the property of “monotonicity”, as tools developed for distinguishing the monotonicity property have proven to be useful for the all of the above properties as well as several others. We say a distribution $p$ is monotone if for any two comparable elements $x < y$ in the domain, we have that $p(x) < p(y)$. For example, for the classic $n$-dimensional hypercube domain, in which domain elements are described via $n$ different features, monotonicity implies that for every element, an increase in the value of one of the features can only increase its probability. We recount the development over the past nearly two decades of monotonicity testing algorithms for distributions over
various discrete domains, which make no a priori assumptions on the underlying distribution. We study the sample complexity for testing whether a distribution is monotone as a function of the size of the domain, which can vary dramatically depending on the structure of the underlying domain. Not surprisingly, the sample complexity over high dimensional domains can be much greater than over low dimensional domains of the same size. Nevertheless, for many important domain structures, including high dimensional domains, the sample complexity is sublinear in the size of the domain. In contrast, when no a priori assumptions are made about the distribution, learning the distribution requires sample complexity that is linear in the size of the domain. The techniques used draw tools from a wide spectrum of areas, including statistics, optimization, combinatorics, and computational complexity theory.

**Keywords:** Property testing algorithms, Monotonicity, Boolean hypercube

---

**Beyond independent rounding: Strongly Rayleigh distributions and traveling salesperson problem**
Shayan Oveis Gharan*
University of Washington

**Abstract:** Strongly Rayleigh (SR) distributions are a family of probably distributions that generalize product distributions and satisfy strongest forms of negative dependence. Over the last last decade, these distributions have found numerous application in algorithm design. I this talk I will survey several fundamental properties of SR distributions and some of their applications in going beyond the limits of the Independent Rounding method for designing approximation algorithms for combinatorial optimization problems.

**Keywords:** Strongly Rayleigh distributions, Negative association, Maximum entropy distributions, Traveling salesperson problem

---

**A survey of dependent randomized rounding**
Aravind Srinivasan*
University of Maryland, College Park

**Abstract:** In dependent randomized rounding, we take a point \( x \) inside a given high-dimensional object such as a polytope and “round” it probabilistically to a suitable point \( y \), such as one with integer coordinates. The “dependence” arises from the fact that the dimensions of \( x \) are rounded in a carefully-correlated manner, so that some properties hold with probability one or with high probability, while others hold in expectation. We will discuss this methodology and some applications in combinatorial optimization and concentration of measure.

**Keywords:** Randomized algorithms, Concentration of measure, Combinatorial optimization, Approximation algorithms

---

**Random linear functions of regularly varying vectors**
Bikramjit Das*
Singapore University of Technology and Design

**Abstract:** In various applications ranging from finance and insurance to network and environmental sciences, we encounter complex risk objects created using a combination of underlying risks which are heavy-tailed (or under
certain assumptions regularly varying). A well-known result from Breiman says that the tail distribution of a product of a regularly varying random variable with another random variable remains regularly varying with the same index. We show that an extension of this result to a multivariate setting helps in quantifying a variety of extreme risks for linear combination of heavy-tailed underlying objects. In particular, we give a characterization of regular variation on sub-cones of the d-dimensional non-negative orthant, under random linear transformations. This allows us to compute probabilities of a variety of extreme events, which classical multivariate regularly varying models would report to be asymptotically negligible. Our findings are illustrated with applications to risk assessment in financial systems and reinsurance markets under a bipartite network structure. We also indicate applications of the result in computing multivariate risk measures, dimensionality reduction, and further extensions to stochastic processes.

[This talk is based on joint work with Claudia Kluppelberg and Vicky Fasen-Hartmann.]

Keywords: Bipartite graphs, Heavy-tails, Multivariate regular variation, Networks

**Power laws and weak convergence of the Kingman coalescent**
Henrik Hult*
KTH Royal Institute of Technology

**Abstract:** The Kingman coalescent is an important and well studied process in population genetics modelling the ancestry of a sample of individuals. In this talk weak convergence results are presented that characterise asymptotic properties of the Kingman coalescent under parent dependent mutations, as the sample size grows to infinity. It is shown that the sampling probability satisfies a power-law and derive the asymptotic behaviour of transition probabilities of the block counting jump chain. For the normalised jump chain and number of mutations between types a limiting process is derived consisting of a deterministic component, describing the limit of the block counting jump chain, and independent Poisson processes with state-dependent intensities, exploding at the origin, describing the limit of the number of mutations. Finally, the results are extended to characterise the asymptotic performance of popular importance sampling algorithms, such as the Griffiths-Tavare algorithm and the Stephens-Donnelly algorithm.

[This is joint work with Martina Favero.]

**Keywords:** Power laws, Population genetics, Kingman coalescent, Importance sampling

**Limit theorems for topological invariants of extreme sample cloud**
Takashi Owada*
Purdue University

**Abstract:** The main objective of this work is to study the topological crackle from the viewpoints of Topological Data Analysis (TDA) and Extreme Value Theory. TDA is a growing research area that broadly refers to the analysis of high-dimensional datasets, the main goal of which is to extract robust topological information from datasets. Topological crackle typically appears in the statistical manifold learning problem, referring to the layered structure of homological cycles generated by “noisy” samples, where the underlying distribution has a heavy tail. We establish various limit theorems (e.g., central limit theorems, strong laws of large numbers) for topological objects, including Betti numbers — a basic quantifier of homological cycles, and the Euler characteristics.

**Keywords:** Extreme value theory, Topological data analysis, Heavy tail, Betti number
Portfolio liquidation games with self-exciting order flow

Guanxing Fu¹, Xiaonyu Xia² and Ulrich Horst³*
¹Applied Mathematics, Hong Kong Poly U, Hong Kong
²Mathematics, Wenzhou University, China
³Mathematics, Humboldt University Berlin, Germany

Abstract: We analyze novel portfolio liquidation games with self-exciting order flow. Both the \(N\)-player game and the mean-field game are considered. We assume that players' trading activities have an impact on the dynamics of future market order arrivals thereby generating an additional transient price impact. Given the strategies of her competitors each player solves a mean-field control problem. We characterize open-loop Nash equilibria in both games in terms of a novel mean-field FBSDE system with unknown terminal condition. Under a weak interaction condition we prove that the FBSDE systems have unique solutions. Using a novel sufficient maximum principle that does not require convexity of the cost function we finally prove that the solution of the FBSDE systems do indeed provide existence and uniqueness of open-loop Nash equilibria.

Keywords: Mean field games, Portfolio liquidation

A mean-field game approach to equilibrium pricing in renewable energy certificate markets

Sebastian Jaimungal¹*, Dena Firoozi² and Arvind Shrivats³
¹Statisticsl Sciences, University of Toronto, Canada
²Decision Sciences, University of Montreal, Canada
³ORFE, Princeton University, Canada

Abstract: Solar Renewable Energy Certificate (SREC) markets are a market-based system that incentivizes solar energy generation. A regulatory body imposes a lower bound on the amount of energy each regulated firm must generate via solar means, providing them with a tradeable certificate for each MWh generated. Firms seek to navigate the market optimally by modulating their SREC generation and trading rates. As such, the SREC market can be viewed as a stochastic game, where agents interact through the SREC price. We study this stochastic game by solving the mean-field game (MFG) limit with sub-populations of heterogeneous agents. Market participants optimize costs accounting for trading frictions, cost of generation, non-linear non-compliance costs, and generation uncertainty. Moreover, we endogenize SREC price through market clearing. We characterize firms' optimal controls as the solution of McKean-Vlasov (MV) FBSDEs and determine the equilibrium SREC price. We establish the existence and uniqueness of a solution to this MV-FBSDE, and prove that the MFG strategies form an \(\epsilon\)-Nash equilibrium for the finite player game. Finally, we develop a numerical scheme for solving the MV-FBSDEs and conduct a simulation study.

Keywords: Mean Field Games, Fiancial Mathematics, Clean Energy

Entropic optimal transport

Marcel Nutz*
Columbia University

Abstract: Applied optimal transport is flourishing after computational advances have enabled its use in real-world problems with large data sets. Entropic regularization is a key method to approximate optimal transport in high
dimensions while retaining feasible computational complexity. In this talk we discuss the convergence of entropic optimal transport to the unregularized counterpart as the regularization parameter vanishes, as well as the stability of entropic optimal transport with respect to its marginals.

[Based on joint works with Espen Bernton (Columbia), Promit Ghosal (MIT), Johannes Wiesel (Columbia).]

**Keywords:** Optimal transport, Entropic regularization

---

**IS22 Random Trees**

**JULY 22 (THU), 21:30 - 22:00 KST**

---

**1d Brownian loop soup, Fleming-Viot processes and Bass-Burdzy flow**

Elie Aidekon*

Fudan University

**Abstract:** We describe the connection between these three objects which appear in the problem of conditioning the so-called perturbed reflecting Brownian motion on its occupation field.

[Joint work with Yueyun Hu and Zhan Shi.]

**Keywords:** Loop soup, Fleming-Viot processes, Continuous-state branching processes, Local time

---

**A new state space of algebraic measure trees for stochastic processes**

Wolfgang Löhr*, Anita Winter¹ and Leonid Mytnik³

¹Mathematics, University of Duisburg-Essen, Germany
²Industrial Engineering and Management, Technion, Israel

**Abstract:** In the talk, I present a new topological space of "continuum" trees, which extends the set of finite graph-theoretic trees to uncountable structures, which can be seen as limits of finite trees. Unlike previous approaches, we do not use the graph-metric but formalize the tree-structure by a tertiary operation on the tree, namely the branch-point map. The resulting space of algebraic measure trees has coarser equivalence classes than the older space of metric measure trees, but the topology preserves more of the tree-structure in limits, so that it is incomparable to, and not coarser than, the standard topologies on metric measure trees. With the example of the Aldous chain on cladograms, I also illustrate that our new space can be very useful as state-space for stochastic processes in order to obtain path-space diffusion limits of tree-valued Markov chains.

**Keywords:** Algebraic continuum tree, Tree-valued diffusion, Aldous chain on cladograms, Random tree

---

**Scaling limits of critical rank-1 inhomogeneous random graphs**

Nicolas Broutin¹, Thomas Duquesne¹ and Minmin Wang³*

¹Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, France
²Department of Mathematics, University of Sussex, UK

**Abstract:** Continuum inhomogeneous random graphs arise in the scaling limits of critical rank-1 inhomogeneous random graphs. They are extensions of the continuum random graph introduced by Addario-Berry, Broutin & Goldschmidt which has appeared as the scaling limit of the Erdos–Renyi graph in the critical window. In this talk, we present a construction of these graphs from the Levy processes without replacement of Aldous & Limic. In particular, this construction reveals a close connection between the clusters of the graphs and Levy trees, which consists in an isometric embedding of the spanning trees of these clusters into Levy trees. As a consequence of this construction,
we provide near optimal conditions for the convergence of critical rank-1 inhomogeneous random graphs. We also deduce the Hausdorff dimension and the packing dimension of the limit graphs.

Keywords: Inhomogeneous random graph, Scaling limit, Branching process, Levy tree

IS23 Stochastic Partial Differential Equations
JULY 22 (THU), 22:30 - 23:00 KST

Phase analysis for a family of stochastic reaction-diffusion equations
Carl Mueller*
University of Rochester

Abstract: We consider a reaction-diffusion equation, that is, a one-dimensional heat equation with a drift function $V(u)$ and two-parameter white noise with coefficient $\sigma(u)$, subject to a “nice” initial value and periodic boundary, where $x$ lies in $T=[-1,1]$. The reaction term $V(u)$ belongs to a large family of functions that includes Fisher-KPP nonlinearities $V(x)=x(1-x)$ as well as Allen-Cahn potentials $V(x)=x(1-x)(1+x)$, the multiplicative nonlinearity $\lambda \sigma$ is non random and Lipschitz continuous, and $\lambda>0$ is a non-random number that measures the strength of the effect of the noise $W$. Our principal finding is that: (i) When $\lambda$ is sufficiently large, the above equation has a unique invariant measure; and (ii) When $\lambda$ is sufficiently small, the collection of all invariant measures is a non-trivial line segment, in particular infinite. This proves an earlier prediction of Zimmerman. Our methods also give information about the structure of these invariant measures.

Keywords: Stochastic partial differential equations, Phase transition

Regularization by noise for SPDEs and SDEs: a stochastic sewing approach
Oleg Butkovsky*
Weierstrass Institute

Abstract: We consider stochastic reaction-diffusion equation
$$\frac{\partial}{\partial t} u_t(x) = \frac{1}{2} \frac{\partial^2}{\partial x^2} u_t(x) + b(u_t(x)) + W_t(x),$$
where the drift $b$ is a Schwartz distribution and $W$ is the space-time white noise. We show that this equation has a unique strong solution if $b$ belongs to a Besov space of certain negative regularity. This extends the classical results of Gyöngy and Pardoux (1993) to distributional drifts. We also show that similar ideas allow to obtain well-posedness of skew fractional Brownian motion
$$dX_t = \delta_0(X_t)dt + dW_t^H,$$
where $\delta_0$ is the Dirac delta function and $W^H$ is the fractional Brownian motion with the Hurst index $H \in (0,1/3)$. This generalizes to the fractional context the skew Brownian motion studied by Harrison and Shepp (1981). To establish these results, we exploit the regularization effect of the white noise through a new strategy based on the stochastic sewing lemma.

[This is a joint work with L. Mytnik, S. Athreya and K. Lê.]

Keywords: SPDE, Sewing lemma, Regularization by noise
Stochastic quantization, Large N, and mean field limit
Hao Shen1, Scott Smith1, Rongchan Zhu2 and Xiangchan Zhu3
1Mathematics, University of Wisconsin-Madison, USA
2Mathematics, Beijing Institute of Technology, China
3Mathematics, Chinese Academy of Sciences, China

Abstract: “Large N problems” in quantum field theory refer to the study of models with N field components for N large. We study these problems using SPDE methods via stochastic quantization. In the SPDE setting this is formulated as mean field problems. In particular we will consider the vector φ4 model (i.e. linear sigma model), whose stochastic quantization is a system of N coupled dynamical φ4 SPDEs. We discuss a series of results. First, in 2D, we prove mean field limit for these dynamics as N goes to infinity. We also show that the quantum field theory converges to massive Gaussian free field in this limit, in both 2D and 3D. Moreover we prove exact formulae for some correlations of O(N)-invariant observables in the large N limit.

Keywords: Stochastic quantization, Large N, Mean field limit

IS24 Random Planar Geometries
JULY 19 (MON), 22:30 - 23:00 KST

Markovian infinite triangulations
Thomas Budzinski
ENS Lyon

Abstract: We say that a random infinite planar triangulation T is Markovian if for any small triangulation t with boundaries, the probability to observe t around the root of T only depends on the boundaries and the total size of t. Such a property can be expected from the local limits of many natural models of random maps. An important example is the UIPT of Angel and Schramm, which is the local limit of large uniform triangulations of the sphere. We will classify completely infinite Markovian planar triangulations, without any assumption on the number of ends. In particular, we will see that there is (almost) no model of multi-ended Markovian triangulation. As an application, we will prove, without to rely on enumerative combinatorics, that the convergence of uniform triangulations to the UIPT is robust under certain perturbations. One example of such a perturbation is to consider random maps with prescribed face degrees where almost all faces are triangles.

Keywords: Random planar maps, Triangulations, UIPT

Rotational invariance in planar FK-percolation
Ioan Manolescu
Université de Fribourg

Abstract: We prove the asymptotic rotational invariance of the critical FK-percolation model on the square lattice with any cluster-weight between 1 and 4. These models are expected to exhibit conformally invariant scaling limits that depend on the cluster weight, thus covering a continuum of universality classes. The rotation invariance of the scaling limit is a strong indication of the wider conformal invariance, and may indeed serve as a stepping stone to the latter. Our result is obtained via a universality theorem for FK-percolation on certain isoradial lattices. This in turn is proved via the star-triangle (or Yang-Baxter) transformation, which may be used to gradually change the square lattice into any of these isoradial lattices, while preserving certain features of the model. It was previously proved that throughout this transformation, the large scale geometry of the model is distorted by at most a limited amount.
the present work we argue that the distortion becomes insignificant as the scale increases. This hinges on the interplay between the inhomogeneity of isoradial models and their embeddings, which compensate each other at large scales. As a byproduct, we obtain the asymptotic rotational invariance also for models related to FK-percolation, such as the Potts and six-vertex ones. Moreover, the approach described here is fairly generic and may be adapted to other systems which possess a Yang-Baxter transformation.

[Based on joint work with Hugo Duminil-Copin, Karol Kajetan Kozłowski, Dmitry Krachun and Mendes Oulamara.]

**Keywords:** FK percolation, Scaling limit, Conformal invariance

---

**Brownian half-plane excursions, CLE₄ and critical Liouville quantum gravity**

Ellen Powell¹*, Nina Holden², Xin Sun³ and Juhan Aru⁴

¹Department of Mathematics, Durham University, UK
²Mathematics, ETH Zurich, Switzerland
³Mathematics, University of Pennsylvania, USA
⁴Mathematics, EPFL, Switzerland

**Abstract:** I will discuss a coupling between a Brownian excursion in the upper half plane and an exploration of nested CLE₄ loops in the unit disk. In this coupling, the CLE₄ is drawn on top of an independent "critical Liouville quantum gravity surface" known as a quantum disk. It turns out that there is a correspondence between loops in the CLE and (sub) half-planar excursions above heights in the Brownian excursion, where the "width" of the sub-excursion corresponds to the "quantum boundary length" of the loop and the height encodes a certain "quantum distance" from the boundary.

[This is based on a forthcoming joint work with Juhan Aru, Nina Holden and Xin Sun, and describes the analogue of Duplantier-Miller-Sheffield's "mating-of-trees correspondence" in the critical regime.]

**Keywords:** Liouville Quantum Gravity, Schramm-Loewner Evolutions

---

**An unexpected phase transition for percolation on scale-free networks**

Souvik Dhara*

Massachusetts Institute of Technology

**Abstract:** The talk concerns the critical behavior for percolation on finite, inhomogeneous random networks, where the weights of the vertices follow a power-law distribution with exponent $\tau \in (2,3)$. Such networks, often referred to as scale-free networks, exhibit critical behavior when the percolation probability tends to zero, as the network-size becomes large. We identify the critical window for percolation phase transition. Rather surprisingly, the critical window turns out to be of finite length, which is in sharp contrast with the previously studied critical behaviors for $\tau \in (3,4)$ and $\tau > 4$ regimes. The rescaled vector of maximum component sizes are shown to converge in distribution to an infinite vector of non-degenerate random variables that can be described in terms of components of a one-dimensional inhomogeneous percolation model studied in a seminal work by Durrett and Kesten (1990).

[Based on joint work with Shankar Bhamidi, Remco van der Hofstad.]

**Keywords:** Percolation, Critical behavior, Scale-free networks
Recent results for the graph alignment problem
Marc Lelarge*
INRIA

Abstract: Random graph alignment refers to recovering the underlying vertex correspondence between two random graphs with correlated edges. This can be viewed as an average-case and noisy version of the well-known NP-hard graph isomorphism problem. For the correlated Erdős-Rényi model, we give an impossibility result for partial recovery in the sparse regime. We also propose a machine learning approach to solve the problem and design a new graph neural network architecture showing great performances.

Local law and Tracy-Widom limit for sparse stochastic block models
Jong Yun Hwang¹, Ji Oon Lee²* and Wooseok Yang²
¹Stochastic Analysis and Application Research Center, KAIST, Korea
²Department of Mathematical Sciences, KAIST, Korea

Abstract: We consider the spectral properties of sparse stochastic block models, where N vertices are partitioned into K balanced communities. Under an assumption that the intra-community probability and inter-community probability are of similar order, we prove a local semicircle law up to the spectral edges, with an explicit formula on the deterministic shift of the spectral edge. We also prove that the fluctuation of the extremal eigenvalues is given by the GOE Tracy-Widom law after rescaling and centering the entries of sparse stochastic block models. Applying the result to sparse stochastic block models, we rigorously prove that there is a large gap between the outliers and the spectral edge without centering.

Keywords: Random matrix, Stochastic block model, Local law, Tracy-Widom distribution

Sig-Wasserstein Generative models to generate realistic synthetic time series
Hao Ni*
University College London

Abstract: Wasserstein generative adversarial networks (WGANs) have been very successful in generating samples, from seemingly high dimensional probability measures. However, these methods struggle to capture the temporal dependence of joint probability distributions induced by time-series data. Moreover, training WGANs is computational expensive due to the min-max formulation of the loss function. To overcome these challenges, we integrate Wasserstein GANs with mathematically principled and efficient path feature extraction called the signature of a path. The signature of a path is a graded sequence of statistics that provides a universal and principled description for a stream of data, and its expected value characterises the law of the time-series model. In particular, we develop a new metric, (conditional) Sig-W₁, that captures the (conditional) joint law of time series models, and use it as a discriminator. The signature feature space enables the explicit representation of the proposed discriminators which alleviates the need for expensive training. We validate our method on both synthetic and empirical dataset and our method achieved the superior performance than the other state-of-the-art benchmark methods.[This is the joint work with Lukasz Szpruch (Uni of Edinburgh), Magnus Wiese (Uni of Kaiserslautern), Shujian Liao (UCL), Baoren Xiao(UCL).]

Keywords: Rough path theory, Generative models, Time series
State space for the 3D stochastic quantisation equation of Yang-Mills
Ilya Chevyrev*
University of Edinburgh

Abstract: In this talk I will present a proposed state space of distributions for the stochastic Yang-Mills equations (SYM) in 3D. I will show how the notion of gauge equivalence extends to this space and how one can construct a Markov process on the space of gauge orbits associated with the SYM. This partly extends a recent construction done in the less singular 2D setting.
[Based on a joint work in progress with Ajay Chandra, Martin Hairer, and Hao Shen.]
Keywords: Stochastic PDEs, Quantum field theory, Gauge theory

A priori bounds for quasi-linear parabolic equations in the full sub-critical regime
Felix Otto¹, Jonas Sauer², Scott Smith³* and Hendrik Weber⁴
¹Max Planck Institute for Mathematics in the Sciences, Germany
²Department of Mathematics, TU Delft, Netherlands
³Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, China
⁴Department of Mathematics, University of Bath, United Kingdom

Abstract: We will discuss quasi-linear parabolic equations driven by an additive forcing, in the full sub-critical regime. Our arguments are inspired by Hairer's regularity structures, however we work with a more parsimonious model indexed by multi-indices rather than trees. Assuming bounds on this model, which is modified in agreement with the concept of algebraic renormalization, we prove local a priori estimates on solutions to the quasi-linear equations modified by the corresponding counter terms.

The scaling limit of the characteristic polynomial of a random matrix at the spectral edge
Elliot Paquette¹* and Gaultier Lambert²
¹Mathematics, McGill University, Canada
²Mathematics, University of Zurich, Switzerland

Abstract: The Gaussian beta-ensemble (GbetaE) is a 1-parameter generalization of the Gaussian orthogonal/unitary/symplectic ensembles which retains some integrable structure. Using this ensemble, Ramirez, Rider and Virag — building on a heuristic of Edelman and Sutton — constructed a limiting point process, the Airy-beta point process, which is the weak limit of the point process of eigenvalues or a random matrix in a neighborhood of the spectral edge. Jointly with Gaultier Lambert, we give a construction of a new limiting object, the stochastic Airy function (SAi); we show this is the limit of the characteristic polynomial of GbetaE in a neighborhood of the spectral edge. It is the bounded solution of the stochastic Airy equation, which is the usual Airy equation perturbed by a multiplicative white noise. We also give some basic properties of SAi.
Keywords: Random matrices, Characteristic polynomial, Beta ensemble
**Strong asymptotics of planar orthogonal polynomials: Gaussian weight perturbed by finite number of point charges**

Seung Yeop Lee*
University of South Florida

Abstract: We consider the orthogonal polynomial $p_n(z)$ with respect to the planar measure supported on the whole complex plane

$$e^{e^{-N|z|^2}} \prod_{j=1}^{\nu} |z - a_j|^{2c_j} \, dA(z)$$

where $dA$ is the Lebesgue measure of the plane, $N$ is a positive constant, $\{c_1, \ldots, c_\nu\}$ are nonzero real numbers greater than $-1$ and $\{a_1, \ldots, a_\nu\} \subset \mathbb{D}\setminus\{0\}$ are distinct points inside the unit disk. In the scaling limit when $n/N \to 1$ and $n \to \infty$, we obtain the strong asymptotics of the polynomial $p_n(z)$. We show that the support of the roots converges to what we call the "multiple Szegő curve," a certain connected curve having $\nu + 1$ components in its complement. We apply the nonlinear steepest descent method on the matrix Riemann-Hilbert problem of size $(\nu + 1) \times (\nu + 1)$.

Keywords: Averaged characteristic polynomial, Orthogonal polynomial, Two dimensional Coulomb gas, Riemann-Hilbert problem

**Secular coefficients and the holomorphic multiplicative chaos**

Joseph Najnudel1*, Elliot Paquette2 and Nick Simm3
1School of Mathematics, University of Bristol, UK
2Mathematics and Statistics department, McGill University, Canada
3School of Mathematical and Physical Sciences, University of Sussex, UK

Abstract: We study the coefficients of the characteristic polynomial (also called secular coefficients) of random unitary matrices drawn from the Circular Beta Ensemble (i.e. the joint probability density of the eigenvalues is proportional to the product of the power beta of the mutual distances between the points). We study the behavior of the secular coefficients when the degree of the coefficient and the dimension of the matrix tend to infinity. The order of magnitude of this coefficient depends on the value of the parameter beta, in particular, for beta = 2, we show that the middle coefficient of the characteristic polynomial of the Circular Unitary Ensemble converges to zero in probability when the dimension goes to infinity, which solves an open problem of Diaconis and Gamburd. We also find a limiting distribution for some renormalized coefficients in the case where beta > 4. In order to prove our results, we introduce a holomorphic version of the Gaussian Multiplicative Chaos, and we also make a connection with random permutations following the Ewens measure.

Keywords: Circular beta ensemble, Characteristic polynomial, Secular coefficients, Gaussian multiplicative chaos

**A goodness-of-fit test for exponential random graphs**

Gesine Reinert*
University of Oxford

Abstract: For assessing the goodness of fit of a model, often independent replicas are assumed. When the data are given in the form of a network, usually there is only one network available. If the data are hypothesised to come from
an exponential random graph model, the likelihood cannot be calculated explicitly. Using Stein’s method we introduce a kernelized goodness of fit test and illustrate its performance. [This talk is based on joint work with Nathan Ross and with Wenkai Xu.]

Keywords: Exponential random graph model, Kernelized goodness-of-fit-test, Stein’s method

Networks in the presence of informative community structure
Alexander Volfovsky*
Duke University

Abstract: The study of network data in the social and health sciences frequently concentrates on associating covariate information to edge formation and assessing the relationship between network information and individual outcomes. In much of this data, it is likely that latent or observed community structure plays an important role. In this talk we describe how to incorporate this community information into a class of latent space models by allowing the effects of covariates on edge formation to differ between communities (e.g. age might play a different role in friendship formation in communities across a city). This information is lost by ignoring explicit community membership and we show that ignoring such structure can lead to over- or underestimation of covariate importance to edge formation. We further demonstrate that when designing experiments on networks, if outcomes of interest are community driven (e.g. differential response to a treatment based on community behavior), incorporating this structure directly into the randomization procedure leads to an improvement in the ability to estimate causal effects.

Keywords: Networks, Causal inference, Randomization, Latent space models

Motif estimation via subgraph sampling: The fourth-moment phenomenon
Bhaswar Bhattacharya1*, Sayan Das2 and Sumit Mukherjee3
1Statistics, University of Pennsylvania, USA
2Mathematics, Columbia University, USA
3Statistics, Columbia University, USA

Abstract: Network sampling has emerged as an indispensable tool for understanding features of large-scale complex networks where it is practically impossible to search/query over all the nodes. Examples include social networks, biological networks, internet and communication networks, and socio-economic networks, among others. In this talk we will discuss a unified framework for statistical inference for counting motifs, such as edges, triangles, and wedges, in the widely used subgraph sampling model. In particular, we will provide precise conditions for the consistency and the asymptotic normality of the natural Horvitz–Thompson (HT) estimator, which can be used for constructing confidence intervals and hypothesis testing for the motif counts. As a consequence, an interesting fourth-moment phenomena for the asymptotic normality of the HT estimator and connections to fundamental results in random graph theory will emerge.

Keywords: Network motifs, Asymptotic inference, Stein’s method, Random graphs
**Minimax rates for derivative-free stochastic optimization with higher order smooth objectives**
Alexandre Tsybakov*
CREST, ENSAE, IP Paris

**Abstract:** We study the problem of finding the minimizer of a function by a sequential exploration of its values, under measurement noise. In the stochastic optimization literature, this problem is known under the name of derivative-free or zero-order optimization. We consider an approximation of the gradient descent algorithm where the gradient is estimated by procedures involving function evaluations in randomized points and a smoothing kernel. We first assume that the objective function is $\beta$-Hölder and $\alpha$-strongly convex with some $\alpha > 0$, and $\beta$ greater or equal to 2. Under general adversarial noise, we obtain non-asymptotic upper bounds, both for the optimization error and the cumulative regret of the algorithm, as functions of the quadruplet $(T, \alpha, \beta, d)$, where $T$ is the number of queries and $d$ is the dimension of the problem. Furthermore, we establish minimax lower bounds for any sequential search method implying that the suggested algorithm is nearly optimal in a minimax sense in terms of sample complexity and the problem parameters. Based on similar ideas, we solve several other problems. In particular, we propose an algorithm achieving almost sharp oracle behavior for the problem of estimating the minimum value of the function. Next, we extend the results to zero-order distributed optimization, where the aim is to minimize the average of local objectives associated to different nodes in a graph with an exchange of information permitted only between neighboring nodes. Finally, we apply similar ideas for the problem of non-convex optimization.
[The talk is based on a joint work with Arya Akhavan and Massimiliano Pontil.]

**High-dimensional, multiscale online changepoint detection**
Richard Samworth1*, Yudong Chen1 and Tengyao Wang2
1Statistical Laboratory, University of Cambridge, UK
2Department of Statistical Science, University College London, UK

**Abstract:** We introduce a new method for high-dimensional, online changepoint detection in settings where a $p$-variate Gaussian data stream may undergo a change in mean. The procedure works by performing likelihood ratio tests against simple alternatives of different scales in each coordinate, and then aggregating test statistics across scales and coordinates. The algorithm is online in the sense that both its storage requirements and worst-case computational complexity per new observation are independent of the number of previous observations. We prove that the patience, or average run length under the null, of our procedure is at least at the desired nominal level, and provide guarantees on its response delay under the alternative that depend on the sparsity of the vector of mean change. Numerical results confirm the practical effectiveness of our proposal, which is implemented in the R package `ocd`.

**Keywords:** Changepoint detection, High-dimensional data stream, Online inference, Multiscale method
Optimal transport and inference for stationary processes
Andrew Nobel*
The University of North Carolina at Chapel Hill

Abstract: Optimal transport ideas have found widespread use in a variety of practical and theoretical statistical problems. In most of these problems the objects under study are fixed and do not possess dynamic structure. However, there are many problems in which the objects of interest are themselves stationary processes, and in these cases it is natural to consider couplings that preserve stationarity. In the talk I will describe several ways in which stationary couplings arise in inference problems for families of processes, how in the Markov setting consideration of transition couplings can lead to fast algorithms for finding optimal couplings, and how optimal couplings may be estimated from data. I will illustrate the potential utility of the Markov case by considering the problem of graph alignment.

Keywords: Optimal transport, Inference for stationary processes, Graph alignment

Optimal network testing by the signed-polygon statistics
Tracy Ke*
Statistics, Harvard University, USA

Abstract: Given a symmetric social network, we are interested in testing whether it has only one community or multiple communities. The desired tests should (a) accommodate severe degree heterogeneity, (b) accommodate mixed-memberships, (c) have a tractable null distribution, and (d) adapt automatically to different levels of sparsity and achieve the optimal phase transition. How to find such a test is a challenging problem. Many existing tests do not allow for heterogeneity or mixed-memberships and cannot (a)-(d). We propose the Signed Polygon as a class of new tests. Fixing $m \geq 3$, for each $m$-gon in the network, define a score using the centered adjacency matrix. The sum of such scores is then the $m$-th order Signed Polygon statistic. The Signed Quadrilateral (SgnQ) is a special example of the Signed Polygon with $m=4$. We show that the SgnQ test satisfies (a)-(d), and especially, they work well for both very sparse and less sparse networks. We derive the asymptotic null distribution and the power of the SgnQ test. For the matching lower bound, we use a phase transition framework, which is more informative than the standard minimax argument. The SgnQ test is applied to a coauthorship network constructed from research papers in 36 statistics journals in a 41-year time span. We demonstrate using the SgnQ test to (a) measure the coauthorship diversity and (b) build a multi-layer community tree.

[This is a collaborated work with Jiashun Jin and Shengming Luo.]

Statistical inference for linear mediation models with high-dimensional mediators
Runze Li*, Xu Guo, Jingyuan Liu and Mudong Zeng
1Statistics, Penn State University, USA
2Statistics, Beijing Normal University, China
3Statistics, Xiamen University, China
4Statistics, Penn State University, USA

Abstract: Mediation analysis draws increasing attention in many scientific areas such as genomics, epidemiology and finance. In this paper, we propose new statistical inference procedures for high dimensional mediation models, in which both the outcome model and the mediator model are linear with high dimensional mediators. Traditional
Procedures for mediation analysis cannot be used to make statistical inference for high dimensional linear mediation models due to high-dimensionality of the mediators. We propose an estimation procedure for the indirect effects of the models via a partial penalized least squares method, and further establish its theoretical properties. We further develop a partial penalized Wald test on the indirect effects, and prove that the proposed test has a $\chi^2$ limiting null distribution. We also propose an $F$-type test for direct effects and show that the proposed test asymptotically follows a $\chi^2$-distribution under null hypothesis and a noncentral $\chi^2$-distribution under local alternatives. Monte Carlo simulations are conducted to examine the finite sample performance of the proposed tests and compare their performance with existing ones. A real data example is used to illustrate the proposed methodology.

**Keywords:** Mediation Analysis, Penalized Least Squares, Sparsity, Wald test

---

**Perturbation bounds for tensors and their applications in high dimensional data analysis**

Ming Yuan*
Columbia University

**Abstract:** We develop deterministic perturbation bounds for singular values and vectors of orthogonally decomposable tensors, in a spirit similar to classical results for matrices. Our bounds exhibit intriguing differences between matrices and higher-order tensors. Most notably, they indicate that for higher-order tensors perturbation affects each singular value/vector in isolation. In particular, its effect on a singular vector does not depend on the multiplicity of its corresponding singular value or its distance from other singular values. Our results can be readily applied and provide a unified treatment to many different problems involving higher-order orthogonally decomposable tensors. In particular, we illustrate the implications of our bounds in several high dimensional data analysis problems.

---

**Algorithmic optimal transport in Euclidean spaces**

Salman Beigi*
Institute for Research in Fundamental Sciences (IPM)

**Abstract:** Transportation cost inequalities in product spaces put an upper bound on the distance that a random point in the space should traverse in order to reach a point in a given target subset of the space. The main question in this talk is whether given the random starting point, the target point can be found algorithmically. This is a hard problem in general and whose answer depends on the underlying product space and its metric. In this talk after motivating this problem via applications in learning theory, answers to this question are given for Euclidean spaces. A main tool in the design and analysis of our algorithm is the tensorization property of transportation cost inequalities. [This talk is based on a joint work with Omid Etesami and Amin Gohari.]

**Keywords:** Optimal transport, Transportation cost inequality, Tensorization

---

**Entropy bounds for discrete log-concave distributions**

Sergey Bobkov*
University of Minnesota

**Abstract:** We will be discussing two-sided bounds for concentration functions and Renyi entropies in the class of log-concave distributions.
discrete log-concave probability distributions. They are used to derive certain variants of the entropy power inequalities.

[The talk is based on a joint work with Arnaud Marsiglietti and James Melbourne.]

**Keywords:** Renyi entropy, Concentration functions

---

**Entropy and convex geometry**

Tomasz Tkocz*

Carnegie Mellon University

**Abstract:** I shall survey several problems emerging from the interplay between convex geometry and information theory, pertaining mainly to reverse entropy power inequalities.

[Based mainly on joint works with Ball, Madiman, Melbourne, Nayar.]

**Keywords:** Entropy power inequalities, Convex geometry

---

**IS32 Statistical Learning**

**JULY 20 (TUE), 22:30 - 23:00 KST**

---

**Equivariant variance estimation for multiple change-point model**

Ning Hao*

University of Arizona, USA

**Abstract:** The variance of noise plays an important role in many change-point detection procedures and the associated inferences. Most commonly used variance estimators require strong assumptions on the true mean structure or normality of the error distribution, which may not hold in applications. In this talk, we introduce a framework of equivariant variance estimation for multiple change-point models. In particular, we characterize the set of all equivariant unbiased quadratic variance estimators for a family of change-point model classes, and develop a minimax theory for such estimators.

---

**A forward approach for sufficient dimension reduction in binary classification**

Jongkyeong Kang and Seung Jun Shin*

Korea University

**Abstract:** Since the seminal sliced inverse regression (SIR) proposed, the inverse-type methods have been canonical in sufficient dimension reduction (SDR). However, they often suffer in binary classification since the binary response yields two slices at most. In this article, we develop a forward approach for SDR in binary classification based on weighted large-margin classifiers. We first show that the gradient of a large-margin classifier is unbiased for SDR as long as the corresponding loss function is Fisher consistent. This leads us to propose what we call weighted outer-product of gradients (wOPG) method. The WOPG can recover the central subspace exhaustively without linearity or constant variance conditions routinely required for the inverse-type methods. We analyze the asymptotic behavior of the proposed estimator, and demonstrate its promising finite-sample performance for both simulated and real data examples.

**Keywords:** Dimension reduction, Gradient learning, Large-margin classifier, Outer product gradient
Variable selection for global Fréchet regression
Danielle Tucker\(^1\), Yichao Wu and Hans-Georg Müller\(^2\)
\(^1\)University of Illinois at Chicago, USA
\(^2\)University of California at Davis, USA

Abstract: Global Fréchet regression is an extension of linear regression to cover more general types of responses, such as distributions, networks and manifolds, which are becoming more prevalent. In such models, predictors are Euclidean while responses are metric space valued. Predictor selection is of major relevance for regression modeling in the presence of multiple predictors but has not yet been addressed for Fréchet regression. Due to the metric space valued nature of the responses, Fréchet regression models do not feature model parameters, and this lack of parameters makes it a major challenge to extend existing variable selection methods for linear regression to global Fréchet regression. In this talk, we share our recent work which addresses this challenge and proposes a novel variable selection method with good practical performance. We provide theoretical support and demonstrate that the proposed variable selection method achieves selection consistency. We also explore the finite sample performance of the proposed method with numerical examples and data illustrations.

Reversing nonequilibrium systems
Leonid Petrov*
University of Virginia

Abstract: A typical stochastic particle model for nonequilibrium thermodynamics starts from a densely packed initial configuration, and evolves by emanating particles into the "rarefaction fan". Imagine having air and vacuum in two halves of a room, and removing the separating barrier. I will explain how for very special (integrable) stochastic particle systems one can explicitly "undo" the rarefaction, and construct another Markov chain which "puts the air back into its half of the room". I will also discuss the corresponding stationary processes preserving each time-t nonequilibrium measure.

Keywords: TASEP, Markov duality, Integrable probability

Mapping KPZ models to free fermions at positive temperature
Takashi Imamura*, Matteo Mucciconi and Tomohiro Sasamoto
Chiba University

Abstract: We find a direct connection between solvable models in the Kardar-Parisi-Zhang (KPZ) universality class and free fermionic models at positive temperature. In studies of integrable probability during the last decade, Fredholm determinant formulas have been obtained for the one dimensional KPZ equation and its integrable discretized models. However it has been a long standing problem to understand the origin of such determinantal structures. Although the final formulas are very simple, they are usually found through complicated calculations. The situation was quite different around 2000. Using the Robinson-Shensted-Knuth (RSK) algorithm Johansson showed that the current of the totally asymmetric simple exclusion process (TASEP) corresponds to a marginal of a free fermionic system at zero temperature described by the Schur measure. In this talk, we will show that there exist more general connections between the discretized models of the KPZ equations and the free fermions. On the KPZ side the models we consider are solvable one parameter deformation of the TASEP. On the free fermionic side, such
deformations in fact correspond to bringing the system to positive temperature. The connections we find is enabled by a new fundamental identity between marginals of the q-Whittaker measures and the periodic Schur measures. It is obtained by comparing the Fredholm determinant formulas for the q-Whittaker measures by Imamura-Sasamoto (2019) and for the Periodic Schur measures by Borodin (2007). We will also report briefly further insights into this topic. One is the bijective combinatorics approach to the identity. There are deep mathematical structures behind it. The other is about applications of this approach to a few KPZ models. Also reported are deformations to this identity connecting the KPZ models in half spaces with Pfaffian point process. Details of these two topics will be explained by Matteo Mucciconi and Tomohiro Sasamoto respectively.

Keywords: KPZ class, Free fermions, Integrable probability

Relaxation time limit of TASEP on a ring
Jinho Baik*
University of Michigan

Abstract: TASEP, totally asymmetric simple exclusion process, is a standard example of an interacting particle system that belongs to the KPZ (Kadar-Parisi-Zhang) universality class. The 2-dimensional random field defined by the height fluctuations of the TASEP on an infinite line converges to an universal random field, the KPZ fixed point. In this talk we discuss what happens if the space is changed to a ring and the ring size grows large together with the time in a certain critical way so that all particles are critically correlated.

(This talk is based on a joint work with Zhipeng Liu.)

Random walk on a barely supercritical branching random walk
Remco van der Hofstad¹, Tim Hulshof¹ and Jan Nagel²*
¹Department of Mathematics and Computer Science, TU Eindhoven, Netherlands
²Department of Mathematics, TU Dortmund, Germany

Abstract: The motivating question behind this project is how a random walk behaves on a barely supercritical percolation cluster, that is, an infinite percolation cluster when the percolation probability is close to the critical value. As a more tractable model, we approximate the percolation cluster by the embedding of a Galton-Watson tree into the lattice. When the random walk runs on the tree, the embedded process is a random walk on a branching random walk. Now we can consider a barely supercritical branching process conditioned on survival, with survival probability approaching zero. In this setting the tree structure allows a fine analysis of the random walk and we can prove a scaling limit for the embedded process under a nonstandard scaling, when the tree becomes more critical the longer the random walk runs on it. This scaling limit allows us to interpolate between supercritical and critical behavior.

Keywords: Percolation, Branching random walk, Scaling limit

Universality of cutoff for graphs with an added random matching
Jonathan Hermon¹, Allan Sly² and Perla Sousi³*
¹Mathematics, University of British Columbia, Canada
²Mathematics, Princeton University, USA
³Mathematics, Cambridge University, UK
Abstract: We establish universality of cutoff for simple random walk on a class of random graphs defined as follows. Given a finite graph $G=(V,E)$ with $|V|$ even we define a random graph $G'=(V,E \cup E')$ obtained by picking $E'$ to be the (unordered) pairs of a random perfect matching of $V$. We show that for a sequence of such graphs of diverging sizes and of uniformly bounded degree, if the minimal size of a connected component of (the original graphs) is at least 3, then the obtained graphs are w.h.p. expanders and the random walk on them exhibits cutoff at a time with an entropic interpretation. This provides a simple generic operation of adding some randomness to a given graph, which results in cutoff. The emerging "cutoff at an entropic time" paradigm will be emphasized.

Keywords: Cutoff, Entropic time, Random graph, Perfect matching

Invariance principle for a random walk among a Poisson field of moving traps
Rongfeng Sun*
National University of Singapore

Abstract: For a random walk among an i.i.d. Bernoulli field of immobile traps on $\mathbb{Z}^d$, it is a classic result that conditioned on survival up to time $n$, the random walk is confined in a ball of radius $Rn^{1/(d+2)}$, and the rescaled path converges to a Brownian motion conditioned to stay inside a ball of radius $R$. When the traps are mobile and perform independent random walks, the only path result so far is that in dimension 1, the random walk (conditioned on survival) is sub-diffusive. We show that in dimension 5 and higher, instead of being confined on a sub-diffusive scale as in the case of immobile traps, the conditioned walk satisfies an invariance principle on the diffusive scale. Our proof is based on the theory of Thermodynamic Formalism.

[Joint work with S. Athreya and A. Drewitz.]

Keywords: Random walk, Mobile traps

From signature based models in finance to affine and polynomial processes and back
Christa Cuchiero*
University of Vienna

Abstract: Modern universal classes of dynamic processes, based on neural networks or signature methods, have recently entered the field of stochastic modeling, in particular in Mathematical Finance. This has opened the door to more data-driven and thus more robust model selection mechanisms, while first principles like no arbitrage still apply. We focus here on signature based models, i.e. (possibly Levy driven) stochastic processes whose characteristics are linear functions of an underlying process' signature and present methods how to learn these characteristics from data. From a more theoretical point of view, we show how these new models can be embedded in the framework of affine and polynomial processes, which have been -- due to their tractability -- the dominating process class prior to the new era of highly overparametrized dynamic models. Indeed, we prove that generic classes of models can be viewed as infinite dimensional affine processes, which in this setup coincide with polynomial processes. A key ingredient to establish this result is again the signature process. This then allows to get power series expansions for expected values of analytic functions of the process' marginals.

[The talk is based on joint works with Guido Gazzani, Francesca Primavera, Sara-Svaluto-Ferro and Josef Teichmann.]

Keywords: Signature methods, Affine and polynomial processes, Data driven models, Deep calibration
Optimal dividends with capital injections at a level-dependent cost
Ronnie Loeffen*
University of Manchester

Abstract: Assume the capital or surplus of an insurance company evolves randomly over time as in the Cramér-Lundberg model but where in addition the company has the possibility to pay out dividends to shareholders and to inject capital at a cost from shareholders. We impose that when the resulting surplus becomes negative the company has to decide whether to inject capital to get to a positive surplus level in order for the company to survive or to let ruin occur. The objective is to find the combined dividends and capital injections strategy that maximises the expected paid out dividends minus cost of injected capital, discounted at a constant rate, until ruin. Such optimal dividends and capital injections problems have been studied before but in the case where the cost of capital (injections) is constant whereas we consider the setting where the cost of capital is level-dependent in the sense that it is higher when the surplus is below 0 than when it is above 0. We investigate optimality of a 3-parameter strategy with parameters \(-r < 0 < c < b\) where dividends are paid out to keep the surplus below \(b\), capital injections are made in order to keep the surplus above \(c\) unless capital drops below the level \(-r\) in which case the company decides to let ruin occur.

[This is joint work with Zbigniew Palmowski.]

Keywords: Optimal control, Ruin theory

Exponential Lévy-type change-point models in mathematical finance
Lioudmila Vostrikova*
University of Angers

Abstract: The parameters of financial models are generally highly dependent on time (information in the press, price of raw materials, stock price hits some psychological level). This time-dependency of the parameters can often be described using a piece-wise constant functions. Of course, the time of change (change-point) for the parameters is not explicitly known, but it is often possible to make reasonable assumptions about its nature and use statistical tests for its detection. The models with change-point were studied since a long time: in a posteriori setting by Page (1955,1957), Vostrikova (1981,1983), Brodsky, Darkhovsky (1993), Brodsky (2017), Garreau (2017), Truong and al. and in its sequential version by Shiryaev (1963), Galchuk, Rozovskij (1971), Basseville, Nikiforov (1993), Gapeev, Pechkir (2004), Bayaraktar, Dayanik, Karatzas (2005), Brodsky (2017). In Mathematical Finance change-point models was considered in the papers of Di Masi, Kabanov (1998), Chen, Gupta (1995), Chen (2005), Dias, Embrechts (2004), Hawkins, Zamba (2005), Vostrikova (2012), Shiryaev, Zhitlukhin (2016), Zhitlukhin, Ziemba (2016).

In this talk we consider exponential Lévy-type models with change point, namely

\[ X_t = L_t I_{\{\tau \leq t\}} + (L_t + \bar{L}_t - \bar{L}_{\tau}) I_{\{\tau > t\}} \]

where \(L\) and \(\bar{L}\) are independent Lévy processes, \(\tau\) is independent from these processes random variable and \(I\) is indicator function. We consider three type of information, related to \(\tau\), by introducing the initially enlarged filtration, progressively enlarged filtration and also the natural filtration of \(X\). We give the conditions for existence and also the expressions of the minimal martingale measures corresponding to HARA utilities. Then we give the expressions of the corresponding utility maximizing strategies. For the initially enlarged filtration and progressively enlarged filtration, these strategies depend on the position of the change-point and the parameters of the Lévy processes involved in \(X\).

Keywords: Exponential Lévy model, Change-point, Initially enlarged filtration, Progressively enlarged filtration
Dynamic pricing and learning under the Bass model
Shipra Agrawal*, Steven Yin and Assaf Zeevi
Columbia University

Abstract: We consider a novel formulation of the dynamic pricing and demand learning problem, where the evolution of demand in response to posted prices is governed by a stochastic variant of the popular Bass model with parameters \((\alpha, \beta)\) that are linked to the so-called "innovation" and "imitation" effects. Unlike the more commonly used i.i.d. demand models, in this model the price posted not only affects the demand and the revenue in the current round but also the evolution of demand, and hence the fraction of market potential that can be captured, in future rounds. Finding a revenue-maximizing dynamic pricing policy in this model is non-trivial even when model parameters are known, and requires solving for the optimal non-stationary policy of a continuous-time, continuous-state MDP. In this paper, we consider the problem of dynamic pricing is used in conjunction with learning the model parameters, with the objective of optimizing the cumulative revenues over a given selling horizon. Our main contribution is an algorithm with a regret guarantee of \(O(m^{2/3})\), where \(m\) is mnemonic for the (known) market size. Moreover, we show that no algorithm can incur smaller order of loss by deriving a matching lower bound. We observe that in this problem the market size \(m\), and not the time horizon \(T\), is the fundamental driver of the complexity; our lower bound in fact indicates that for any fixed \(\alpha, \beta\), most non-trivial instances of the problem have constant \(T\) and large \(m\). This insight sets the problem setting considered here uniquely apart from the MAB type formulations typically considered in the learning to price literature.

Keyword: Dynamic Pricing, Multi-armed bandits, Bass model

TensorPlan: A new, flexible, scalable and provably efficient local planner for huge MDPs
Csaba Szepesvari1*, Gellert Weisz2, Philip Amortila3, Barnabas Janzer4, Yasin Abbasi-Yadkori2 and Nan Jiang3
1DeepMind, Canada
2DeepMind, UK
3Department of Computer Science, University of Illinois at Urbana-Champaign, USA
4Department of Mathematics, Cambridge University, UK

Abstract: In this talk I will consider provably efficient planning in huge MDPs when the planner is helped with a hint about the form of the optimal value function. In particular, a thoughtful oracle provides the planner with basis functions the linear combination of which give the optimal value function either exactly, or with small errors. The problem is to design a local planner, which, similarly to model-predictive control, is called to find a good action after every state transition, while it is given access to a simulator. We propose a new planner which when used continuously is guaranteed to induce a near-optimal policy. When the number of action is kept as a constant, the planner is shown to require only polynomially many simulator queries as a function of the horizon and the number of basis functions. The planner does not use dynamic programming as we know it, but is based on optimism and the “tensorization” of the Bellman optimality equation.

On the importance of (linear) structure in contextual multi-armed bandit
Alessandro Lazaric*
Facebook AI Research
Abstract: In this talk I will discuss how structural assumptions on the reward function impacts the regret performance of bandit algorithms. Notably, I will focus on linear contextual bandits and first review recent results showing how the structure of the arm set and reward function can be leveraged to achieve improved regret guarantees. Then, I will describe a novel incremental algorithm able to achieve asymptotic optimality, while ensuring finite-time worst-case optimality in the context-free case. Finally, I will discuss how stronger assumptions on context distribution and linear representation may be leveraged to achieve constant regret. This eventually leads to a representation-selection algorithm matching the regret of the best linear representation in a given set, up to a logarithmic factor in the number of representations.

Most relevant references
M. Papini, A. Tirinzoni, M. Restelli, A. Lazaric, M. Pirotta, "Leveraging Good Representations in Linear Contextual Bandits", 2021.

Hydrodynamic large deviations of strongly asymmetric interacting particle systems
Quastel Jeremy¹ and Li-Cheng Tsai²*
¹Mathematics, University of Toronto, Canada
²Mathematics, Rutgers University, USA

Abstract: We consider the large deviations from the hydrodynamic limit of the Totally Asymmetric Simple Exclusion Process (TASEP), which is related to the entropy production in the inviscid Burgers equation. Here we prove the full large deviation principle. Our method relies on the explicit formula of Matetski, Quastel, and Remenik (2016) for the transition probabilities of the TASEP.

Conformal loop ensembles on Liouville quantum gravity with marked points
Nina Holden*
ETH Zurich

Abstract: Liouville quantum gravity (LQG) surfaces are a natural family of random fractal surfaces, while the conformal loop ensemble (CLE) is a random collection of conformally invariant non-crossing loops in the plane. In a joint work with Matthis Lehmkuehler we study CLE-decorated LQG surfaces whose law has been reweighted according to CLE loop nesting statistics around a fixed number of marked points. In particular, we describe explicitly the law of a Markovian exploration of the surface, we compute some observables of interest via explicit computations for Levy processes, and we investigate admissibility and recursive computation of the partition function of the surface.
Keywords: Liouville quantum gravity, Conformal loop ensemble
Integrability of Schramm-Loewner evolution and Liouville quantum gravity

Xin Sun\textsuperscript{1}\textsuperscript{*} and Ang Morris\textsuperscript{2}

\textsuperscript{1}Math, University of Pennsylvania, USA
\textsuperscript{2}Math, MIT, USA

Abstract: It appears that there are rich integrable structures in Schramm-Loewner evolution and Liouville quantum gravity. Namely, many important observables admit exact expressions. In this talk, I will review two major resources of such integrability: conformal field theory and random planar maps decorated with statistical physics models. I will then present a recent work with Morris Ang that proves an integrable result for conformal loop ensembles, which is analogous to the DOZZ formula in Liouville conformal field theory. Our result is an example of a series of results that are proved by blending these two sources of integrability.

Efficient manifold approximation with spherelets

Didong Li\textsuperscript{*}
Princeton University / University of California

Abstract: Data lying in a high dimensional ambient space are commonly thought to have a much lower intrinsic dimension. In particular, the data may be concentrated near a lower-dimensional subspace or manifold. There is an immense literature focused on approximating the unknown subspace, and in exploiting such approximations in clustering, data compression, and building of predictive models. Most of the literature relies on approximating subspaces using a locally linear, and potentially multiscale, dictionary. In this talk, a simple and general alternative is introduced, which instead uses pieces of spheres, or spherelets, to locally approximate the unknown subspace. Theory is developed showing that spherelets can produce lower covering numbers and MSEs for many manifolds. We develop spherical principal components analysis (SPCA). Results relative to state-of-the-art competitors show gains in ability to accurately approximate the subspace with fewer components. In addition, unlike most competitors, our approach can be used for data denoising and can efficiently embed new data without retraining. The methods are illustrated with standard toy manifold learning examples, and applications to multiple real data sets.

Toward instance-optimal reinforcement learning

Ashwin Pananjady\textsuperscript{*}
Georgia Institute of Technology

Abstract: The paradigm of reinforcement learning has now made inroads in a wide range of applied problem domains. This empirical research has revealed the limitations of our theoretical understanding: popular RL algorithms exhibit a variety of behavior across domains and problem instances, and existing theoretical bounds, which are generally based on worst-case assumptions, can often produce pessimistic predictions. An important goal is thus to develop instance-specific analyses that help to reveal what aspects of a given problem make it "easy" or "hard", and allow distinctions to be drawn between ostensibly similar algorithms. Taking an approach grounded in nonparametric statistics, we initiate a study of this question for the policy evaluation problem. We show via information-theoretic lower bounds that many popular variants of stochastic approximation or "temporal difference learning" algorithms do not exhibit the optimal instance-specific performance in the finite-sample regime. On the other hand, making careful modifications to these algorithms does result in automatic adaptation to the intrinsic difficulty of the problem. When
there is function approximation involved, our bounds also characterize the instance-optimal tradeoff between approximation and estimation errors in solving projected fixed-point equations, a general class of problems that includes policy evaluation as a special case. These oracle inequalities, which are non-standard and involve a non-unit pre-factor multiplying the approximation error, may be of independent statistical interest.

**Keywords:** Non-asymptotic local minimax, Reinforcement learning, Stochastic approximation, Non-standard oracle inequalities

---

**Bayesian pyramids: Identifying interpretable discrete latent structures from discrete data**

Yuqi Gu¹* and David Dunson²

¹Department of Statistics, Columbia University, USA
²Department of Statistical Science, Duke University, USA

**Abstract:** High dimensional categorical data are routinely collected in biomedical and social sciences. It is of great importance to build interpretable models that perform dimension reduction and uncover meaningful latent structures from such discrete data. Identifiability is a fundamental requirement for valid modeling and inference in such scenarios, yet is challenging to address when there are complex latent structures. In this article, we propose a class of interpretable discrete latent structure models for discrete data and develop a general identifiability theory. Our theory is applicable to various types of latent structures, ranging from a single latent variable to deep layers of latent variables organized in a sparse graph (termed a Bayesian pyramid). The proposed identifiability conditions can ensure Bayesian posterior consistency under suitable priors. As an illustration, we consider the two-latent-layer model and propose a Bayesian shrinkage estimation approach. Simulation results for this model corroborate identifiability and estimability of the model parameters. Applications of the methodology to DNA nucleotide sequence data uncover discrete latent features that are both interpretable and highly predictive of sequence types. The proposed framework provides a recipe for interpretable unsupervised learning of discrete data, and can be a useful alternative to popular machine learning methods.

**Keywords:** Identifiability, Interpretable machine learning, Latent structure, Multivariate categorical data

---

**Convergence of randomized urn models with irreducible and reducible replacement**

Li-Xin Zhang*

Zhejiang University

**Abstract:** Generalized Friedman urn is a popular model in probability theory. Since Athreya and Ney (1972) showed the almost sure convergence of urn proportions in a randomized urn model with irreducible replacement matrix under the $L \log L$ moment assumption, this assumption has been regarded as the weakest moment assumption, but its necessity has never been shown. In this talk, we will consider the strong and weak convergence of generalized Friedman urns. It is proved that, when the random replacement matrix is irreducible in probability, the sufficient and necessary moment assumption for the almost sure convergence of the urn proportions is that the expectation of the replacement matrix is finite, which is less stringent than the $L \log L$ moment assumption, and that when the replacement is reducible, the $L \log L$ moment assumption is the weakest sufficient condition. The rate of convergence and the strong and weak convergence of non-homogeneous generalized Friedman urns are also derived.

**Keywords:** Generalized Friedman urn, Almost sure convergence, Branching process
Condensation phenomenon and metastability in interacting particle systems
Insuk Seo\textsuperscript{1,\*}, Claudio Landim\textsuperscript{1}, Diego Marcondes\textsuperscript{1} and Seonwoo Kim\textsuperscript{2}
\textsuperscript{1}Institute of Mathematics and Statistics, Universidade de Sao Paulo, Brazil
\textsuperscript{2}Department of Mathematical Sciences, Seoul National University, Korea

Abstract: In this talk, we discuss recent development in the study of the condensation phenomena appeared in interacting particle systems such as the super-critical or critical zero-range process and the inclusion process. In particular, we will focus on the metastable behavior, i.e., the evolution of the condensate after it is formed.

Keywords: interacting particle systems, Zero-range process, Inclusion process, Metastability

Time Correlation exponents in planar last passage percolation
Riddhipratim Basu\textsuperscript*\textsuperscript{1}
International Centre for Theoretical Sciences-TIFR

Abstract: Planar last passage percolation (LPP) models are canonical examples of stochastic growth in the Kardar-Parisi-Zhang universality class, where one considers oriented paths (moving forward in the "time" direction) between points in a random environment accruing the integral of the noise along itself as its weight. The maximal weight of a path joining two points is called the last passage time between the points. Although these models are expected to exhibit universal features under mild conditions on the underlying i.i.d. noise, rigorous progress has mostly been limited to a handful of exactly solvable models. One of the questions in this class of models that has drawn a lot of recent attention is that of the two time correlation, i.e., two understand the correlation decay of the last passage times to a sequence of points varying along the time direction (i.e., the diagonal direction) starting from different initial data. I shall describe some results obtaining the exponents governing the short and long range correlations in the context of the exactly solvable model of planar exponential last passage percolation starting from flat and step initial data.
[Based on joint works with Shirshendu Ganguly and Lingfu Zhang.]

Keywords: Last passage percolation, KPZ universality class

IS40 KSS Invited Session: Nonparametric and Semi-parametric Approaches in Survival Analysis
JULY 21 (WED), 21:30 - 22:00 KST

Smoothed quantile regression for censored residual lifetime
Kyuhyun Kim\textsuperscript{1} and Sangwook Kang\textsuperscript{2}\textsuperscript*\textsuperscript{2}
\textsuperscript{1}Statistics and Data Science, Yonsei University, Korea
\textsuperscript{2}Applied Statistics / Statistics and Data Science, Yonsei University, Korea

Abstract: We consider a regression modeling of the quantiles of residual lifetime at a specific time given a set of covariates. For estimation of regression parameters, we propose an induced smoothed version of the existing non-smooth estimating equations approaches. The proposed estimating equations are smooth in regression parameters, so solutions can be readily obtained via standard numerical algorithms. Moreover, smoothness in the proposed estimating equations enables one to obtain a closed form expression of the robust sandwich-type covariance estimator of regression estimators. To handle data under right censoring, inverse probabilities of censoring are incorporated as weights. Consistency and asymptotic normality of the proposed estimator are established. Extensive simulation studies are conducted to verify performances of the proposed estimator under various finite samples settings. We apply the proposed method to dental study data evaluating the longevity of dental restorations.
Keywords: Induced smoothing, Resampling, Sandwich estimator, Survival analysis

Superefficient estimation of future conditional hazards based on marker information
Enno Mammen¹, Alex Isakson², Jens Perch Nielsen² and Cécile Proust-Lima³
¹Institute of Applied Mathematics, Heidelberg University, Germany
²Bayes Business School, City, University of London, UK
³Bordeaux Population Health Research Center, Inserm, France

Abstract: We introduce a new concept for forecasting future events based on marker information. The model is based on a nonparametric approach with counting processes featuring so-called high quality markers. Despite the model having nonparametric parts we show that we attain a parametric rate of uniform consistency and uniform asymptotic normality. In usual nonparametric scenarios reaching such a fast convergence rate is not possible, so one can say that our approach is superefficient. We then use these theoretical results to construct simultaneous confidence bands directly for the hazard rate.

On a semiparametric estimation method for AFT mixture cure models
Ingrid Van Keilegom*
KU Leuven

Abstract: When studying survival data in the presence of right censoring, it often happens that a certain proportion of the individuals under study do not experience the event of interest and are considered as cured. The mixture cure model is one of the common models that take this feature into account. It depends on a model for the conditional probability of being cured (called the incidence) and a model for the conditional survival function of the uncured individuals (called the latency). This work considers a logistic model for the incidence and a semiparametric accelerated failure time model for the latency part. The estimation of this model is obtained via the maximization of the semiparametric likelihood, in which the unknown error density is replaced by a kernel estimator based on the Kaplan-Meier estimator of the error distribution. Asymptotic theory for consistency and asymptotic normality of the parameter estimators is provided. Moreover, the proposed estimation method is compared with several competitors. Finally, the new method is applied to data coming from a cancer clinical trial.

From infinite random matrices over finite fields to square ice
Leonid Petrov*
University of Virginia

Abstract: Asymptotic representation theory of the symmetric group is a rich and beautiful subject with deep connections with probability, mathematical physics, and algebraic combinatorics. I will discuss a one-parameter deformation of this theory related to infinite random matrices over a finite field, which has an interesting connection to the six vertex (square ice) model and traffic systems on a 1-dimensional lattice.
Keyword: Six vertex model, Hall-Littlewood polynomials, Robinson-Schensted-Knuth correspondence
A general frequency domain method for assessing spatial covariance structures
Soutir Bandyopadhyay\textsuperscript{1,*}, Matthew Van Hala, Daniel Nordman\textsuperscript{2} and Soumendra N. Lahiri\textsuperscript{3}
\textsuperscript{1}Department of Applied Mathematics and Statistics, Colorado School of Mines, USA
\textsuperscript{2}Department of Statistics, Iowa State University, USA
\textsuperscript{3}Department of Mathematics, Washington University, USA

Abstract: When examining dependence in spatial data, it can be helpful to formally assess spatial covariance structures that may not be parametrically specified or fully model-based. That is, one may wish to test for general features regarding spatial covariance without presupposing any particular, or potentially restrictive, assumptions about the joint data distribution. Current methods for testing spatial covariance are often intended for specialized inference scenarios, usually with spatial lattice data. We propose instead a general method for estimation and testing of spatial covariance structure, which is valid for a variety of inference problems (including nonparametric hypotheses) and applies to a large class of spatial sampling designs with irregular data locations. In this setting, spatial statistics have limiting distributions with complex standard errors depending on the intensity of spatial sampling, the distribution of sampling locations, and the process dependence. The proposed method has the advantage of providing valid inference in the frequency domain without estimation of such standard errors, which are often intractable, and without particular distributional assumptions about the data (e.g., Gaussianity). To illustrate, we develop the method for formally testing isotropy and separability in spatial covariance and consider confidence regions for spatial parameters in variogram model fitting. A broad result is also presented to justify the method for application to other potential problems and general scenarios with testing spatial covariance. The approach uses spatial test statistics, based on an extended version of empirical likelihood, having simple chi-square limits for calibrating tests. We demonstrate the proposed method through several numerical studies.

Keywords: Confidence sets, Spatial periodogram, Spatial testing, Spectral moment conditions
ORGANIZED CONTRIBUTED SESSION
Outliers for Coulomb gases
David Garcia-Zelada*
Institut de Mathématiques de Marseille, Aix-Marseille University, France

Abstract: We will be interested in a two-dimensional model of n positively charged particles at equilibrium. Our systems will be attracted to a background with negative charge distribution, and it will be seen that, as n goes to infinity and in the regions where there is no background charge, an interesting phenomenon occurs.


Keywords: Determinantal point process, Bergman kernel, Convergence in distribution

Edge behaviors of 2D Coulomb gases with boundary confinements
Seong-Mi Seo*
Mathematics, Korea Institute for Advanced Study, Korea

Abstract: In this talk, we will consider the local statistics of a planar Coulomb gas system which is determinantal. In a suitable external field, the Coulomb particles tend to accumulate on a set called a droplet, which is the support of the equilibrium measure associated with the external field. The most well-known boundary condition for the gas is the "free boundary", where the particles are admitted to be outside of the droplet. On the other hand, if a boundary confinement is imposed to force the particles to be completely confined to a set, the edge behavior may change. In the presence of a hard-wall constraint to change the equilibrium, the density of the equilibrium measure acquires a singular component at the hard wall and the Coulomb gas system properly rescaled at the hard wall converges to a determinantal point process which appears in the context of truncated unitary matrices. I will present the edge behaviors of the Coulomb gas under the different boundary confinements and explain two approaches using the asymptotics of orthogonal polynomials and the rescaled version of Ward's equation from the field theory.

Keywords: Coulomb gases, Boundary confinements, Scaling limits, Ward's equation

Large deviations in the quantum quasi-1D jellium
Christian Hirsch1*, Sabine Jansen2 and Paul Jung3
1Bernoulli Institute, University of Groningen, Netherlands
2Mathematisches Institut, LMU Munich, Germany
3Department of Mathematical Sciences, KAIST, Korea

Abstract: Wigner's jellium is a model for a gas of electrons. The model consists of N unit negatively charged particles lying in a sea of neutralizing homogeneous positive charge spread out according to Lebesgue measure, and interactions are governed by the Coulomb potential. In this work, we consider the quantum jellium on quasi-one-dimensional spaces with Maxwell-Boltzmann statistics. Using the Feynman-Kac representation, we replace particle locations with Brownian bridges. We then adapt the approach of Leblé and Serfaty (2017) to prove a process-level large deviation principle for the empirical fields of the Brownian bridges.

Keywords: Coulomb systems, Jellium, LDP
**Lemniscate ensembles with spectral singularities**

Sung-Soo Byun\(^1\), Seung-Yeop Lee\(^2\) and Meng Yang\(^3\)

\(^1\)Mathematical Sciences, Seoul National University, Korea

\(^2\)Mathematics and Statistics, University of South Florida, USA

\(^3\)Mathematical Sciences, University of Copenhagen, Denmark

**Abstract:** In this talk, I will discuss a family of determinantal Coulomb gases, which tend to occupy lemniscate type droplets in the large system. For these lemniscate ensembles under the insertion of a point charge, I will present the scaling limits at the singular boundary point, which are expressed in terms of the solution to the Painlevé IV Riemann-Hilbert problem. I will also explain the main ingredients of the proof, which include a version of the Christoffel-Darboux identity and the strong asymptotic behaviour of the associated orthogonal polynomials.

**Keywords:** Determinantal Coulomb gas, Lemniscate, Spectral singularity, Orthogonal polynomial

---

**OCS02 Nonlocal Operators Related to Probability**

**JULY 19 (MON), 11:30 - 12:00 KST**

---

**A Sobolev space theory for time-fractional stochastic PDE driven by Levy processes**

Kyeong-Hun Kim and Daehan Park*

Korea Advanced Institute of Science and Technology (KAIST)

**Abstract:** There are many kinds of studies related to time-fractional equations (both deterministic and stochastic). In this talk, the speaker will give a Sobolev space theory for time-fractional stochastic PDE driven by Levy processes. The existence of the kernel to represent solutions gives the way how we control the derivatives of solutions. Precisely, we find suitable decay of Fourier transform of the kernel and use the Littlewood-Paley theory. From this, we find a suitable condition to give regularity to solutions.

**Keywords:** Time-fractional equations, Levy processes, Littlewood-Paley theory, Regularity theory

---

**A regularity theory for stochastic modified Burgers’ equation driven by multiplicative space-time white noise**

Beom-Seok Han*

Pohang University of Science and Technology

**Abstract:** Consider a stochastic modified Burgers’ equation driven by multiplicative white noise:

\[ u_t = au_{xx} + bu_x + cu + \tilde{b}u^\lambda u_x + \sigma(u)\tilde{W}, \quad (t, x) \in (0, \infty) \times bR, \]

with nonnegative random initial data \( u(0, \cdot) = u_0. \) Here \( \lambda > 0 \) and \( \tilde{W} \) is a space-time white noise. The coefficients \( a, b \) and \( c \) depend on \( (\omega, t, x) \), and \( \tilde{b} \) depend on \( (\omega, t) \). The coefficient \( \sigma(u) \) depends on \( (\omega, t, x, u) \).

Depending on types of \( \sigma(u) \) and noise \( \tilde{W} \), we consider two different cases. If \( \sigma(u) \) is a bounded Lipschitz function in \( u, \lambda \) can be less than or equal to \( 1; \lambda \in (0, 1]. \) Otherwise, if \( \sigma(u) = u^{1+\lambda_0} \), then we consider the case \( \lambda, \lambda_0 \in (0, 1). \)

In each cases, we discuss the uniqueness, existence, \( L_p \)-regularity, and maximal Hölder regularity of a strong solution to the above equation. The maximal Hölder regularity of the solution depends on the nonlinear terms; \( \tilde{b}u^{\lambda} u_x \) and \( \sigma(u) \). For example, if \( \lambda = 1 \) and \( \sigma(u) \) is a Lipschitz function in \( u \), we have (a.s.)

\[ u \in C^{1/4 - \varepsilon, 1/2 - \varepsilon}_{\text{loc}}([0, T] \times R). \]

Otherwise, if \( \sigma(u) = u^{1+\lambda_0} \), then (a.s.)
General Law of iterated logarithm for Markov processes
Jaehun Lee¹*, Soobin Cho² and Panki Kim²
¹Mathematics, KIAS, Korea
²Mathematics, SNU, Korea

Abstract: In this talk, we discuss general criteria and forms of laws of iterated logarithm (LIL) for continuous-time Markov processes. We consider minimal assumptions for LILs to hold at zero (at infinity, respectively) in general metric measure spaces. We establish LILs under local assumptions near zero (near infinity, respectively) on uniform bounds of the first exit time from balls and uniform bounds on the tails of the jumping measure. We provide a general formulation of liminf and limsup LILs, which covers a large class of subordinated diffusions, jump processes with mixed polynomial local growths, jump processes with singular jumping kernels and random conductance models with long range jumps. We also introduce our recent results on the laws of iterated logarithm for occupation times of ball and local time for continuous-time Markov processes.
[This talk is based on the joint work with Soobin Cho and Panki Kim.]

Keywords: Markov process, Laws of iterated logarithm, Dirichlet form

Heat kernel estimates for subordinate Markov processes and their applications
Soobin Cho*
Department of Mathematical Sciences, Seoul National University, Korea

Abstract: In this talk, we discuss sharp two-sided estimates for transition densities of a large class of subordinate Markov processes. As applications, we show that parabolic Harnack inequality and Hölder regularity hold for parabolic functions of such processes, and derive sharp two-sided Green function estimates.

Keywords: Heat kernel, Subordinate Markov process

A maximal $L_p$-regularity theory to initial value problems with time measurable nonlocal operators generated by additive processes
Jae-Hwan Choi* and Ildoo Kim
Mathematics, Korea University, Korea

Abstract: Let $Z = (Z_t)_{t \geq 0}$ be an additive process with a bounded triplet $(0, 0, \Lambda_t)_{t \geq 0}$. Then the infinitesimal generator of $Z$ is given by time dependent nonlocal operators as follows:

$$A_Z(t)u(t, x) = \lim_{h \downarrow 0} \frac{E[u(t, x + Z_{t+h} - Z_t) - u(t, x)]}{h} = \int_{\mathbb{R}^d} (u(t, x + y) - u(t, x) - y \cdot \nabla u(t, x) 1_{|y| \leq 1}) \Lambda_t(dy).$$

Suppose that Lévy measures $\Lambda_t$ have a lower bound and satisfy a weak-scaling property. We emphasize that there is no regularity condition on Lévy measures $\Lambda_t$ and they do not have to be symmetric. In this paper, we establish the $L_p$-solvability to initial value problem (IVP)

$$\frac{\partial u}{\partial t} (t, x) = A_Z(t)u(t, x), \quad u(0, \cdot) = u_0, \quad (t, x) \in (0, T) \times \mathbb{R}^d,$$
where \( u_0 \) is contained in a scaled Besov space \( B^{p,q}_{\gamma,q} (\mathbb{R}^d) \) with a scaling function \( s \), exponent \( p \in [1, \infty) \), \( q \in [1, \infty) \), and order \( \gamma \in (0, \infty) \). We show that IVP is uniquely solvable and the solution \( u \) obtains full-regularity gain from the diffusion generated by a stochastic process \( Z \). In other words, there exists a unique solution \( u \) to IVP in \( L_q((0,T); H^{\mu,\gamma}_p(\mathbb{R}^d)) \) where \( H^{\mu,\gamma}_p(\mathbb{R}^d) \) is a generalized Besel potential space.

Moreover, the solution \( u \) satisfies

\[
\|u\| \leq N(1 + T^2) \|u_0\| B^{p,q}_{\gamma,q} (\mathbb{R}^d),
\]

where \( N \) is independent of \( u, u_0, \) and \( T \).

[This talk is based on joint work with Ildoo Kim.]

**Keywords:** Nonlocal operator, Additive process, Littlewood-Payley theory

---

### OCS03 Gaussian Processes

**JULY 21 (WED), 21:30 - 22:00 KST**

**Gaussian determinantal processes: A new model for directionality in data**

Subhroshekhar Ghosh\(^1\)* and Philippe Rigollet\(^2\)

\(^1\)National University of Singapore

\(^2\)Mathematics, MIT, USA

**Abstract:** Determinantal point processes (DPPs) have recently become popular tools for modeling the phenomenon of negative dependence, or repulsion, in data. However, our understanding of an analogue of a classical parametric statistical theory is rather limited for this class of models. In this work, we investigate a parametric family of Gaussian DPPs with a clearly interpretable effect of parametric modulation on the observed points. We show that parameter modulation impacts the observed points by introducing directionality in their repulsion structure, and the principal directions correspond to the directions of maximal (i.e., the most long- ranged) dependency. This model readily yields a viable alternative to principal component analysis (PCA) as a dimension reduction tool that favors directions along which the data are most spread out. This methodological contribution is complemented by a statistical analysis of a spiked model similar to that employed for covariance matrices as a framework to study PCA. These theoretical investigations unveil intriguing questions for further examination in random matrix theory, stochastic geometry, and related topics.

**Keywords:** Determinantal point processes, Spiked model, Dimension reduction

### Persistence exponents of Gaussian stationary functions

Ohad Noy Feldheim*

Einstein Institute of Mathematics, Hebrew University of Jerusalem, Israel

**Abstract:** Let \( f: \mathbb{R} \rightarrow \mathbb{R} \) be a Gaussian stationary process, that is, a random function, invariant to real shifts, whose marginals have multi-normal distribution. Persistence is the event that the process remains positive over the interval \([0,T]\). The asymptotics of this quantity as \( T \) tends to infinity has been long studied since the early 50’s with motivation stemming from Probability theory, Physics and Electric Engineering. In recent years, it has been discovered that persistence is best characterized in spectral terms. This view was used to describe the decay rate of persistence probability (up to a constant in the exponent). In this work we take this study one step further, showing mild conditions for the existence of persistence exponents, that is, \( C \) such that the probability of persistence on \([0,T]\) is \( e^{-CT^{1+o(1)}} \). This we obtain by establishing an array of continuity properties of the persistence probability and relating the problem to small ball exponents. In particular, we show that the persistence exponent is independent from the
singular component of the spectral measure away from the origin.

(Joint work with N. Feldheim and S. Mukherjee.)

**Keywords:** Persistence, Spectral measure, Hole probability, Stationary

---

**Connectivity of the excursion sets of Gaussian fields with long-range correlations**

Stephen Muirhead*

School of Mathematics and Statistics, University of Melbourne, Australia

**Abstract:** In recent years the global connectivity of the excursion sets of smooth Gaussian fields with rapidly decaying correlations has been fairly well understood (at least in the case of positively-correlated fields), and the general picture that emerges is that the connectivity undergoes a phase transition which is analogous to that of Bernoulli percolation. On the other hand, if the fields have long-range correlations then they are believed to lie outside the Bernoulli percolation universality class, with different scaling limits and critical exponents. The behaviour of the connectivity is not well-understood in this regime, and in this talk I will present some recent results and conjectures that shed some light on the behaviour.

**Keywords:** Gaussian fields

---

**Overcrowding estimates for the nodal volume of stationary Gaussian processes on** \( \mathbb{R}^d \)

Lakshmi Priya*

Department of Mathematics, Indian Institute of Science (IISc), Bangalore, India

**Abstract:** In earlier studies, under varying assumptions on the spectral measures of SGPs, the following statistics were obtained for the nodal volume in \([0, T]^d\): expectation, variance asymptotics, CLT, exponential concentration (only for \(d=1\)), and finiteness of moments. We study the unlikely event of overcrowding of the nodal set in \([0, T]^d\); this is the event that the volume of the nodal set in \([0, T]^d\) is much larger than its expected value. Under some mild assumptions on the spectral measure, we obtain estimates for the overcrowding event’s probability. We first get overcrowding estimates for the zero count of SGPs on \(\mathbb{R}\). In higher dimensions, we consider Crofton’s formula which gives the volume of the nodal set in terms of the number of intersections of the nodal set with all lines in \(\mathbb{R}^d\). We discretise this formula to get a more workable version of it; we use this and the ideas used to obtain the overcrowding estimates in one dimension to get the overcrowding estimates in higher dimensions.

**Keywords:** Stationary Gaussian processes, Nodal set, Overcrowding estimates

---

**Metastability in the reversible inclusion process**

Sander Dommers*

Department of Physics and Mathematics, University of Hull, UK

**Abstract:** In the reversible inclusion process with a fixed number of particles on a finite graph each particle at a site \(x\) jumps to site \(y\) at rate \((d + \eta_y)r(x, y)\), where \(d\) is a diffusion parameter, \(\eta_y\) is the number of particles on site \(y\) and \(r(x, y)\) is the jump rate from \(x\) to \(y\) of an underlying reversible random walk. When the diffusion \(d\) tends to 0 as the number of particles tends to infinity, the particles cluster together to form a condensate. It turns out that these condensates only form on the sites where the underlying random walk spends the most time. Once such a condensate
is formed the particles stick together and the condensate performs a random walk itself on much longer timescales, which can be seen as metastable (or tunnelling) behaviour. We study the rates at which the condensate jumps and characterize the behavior on the shortest time scale at which jumps occur. This generalizes work by Grosskinsky, Redig and Vafayi who study the symmetric case. Our analysis is based on the martingale approach by Beltrán and Landim. [This is joint work with Alessandra Bianchi and Cristian Giardinà.]

Keywords: Inclusion process, Metastability, Interacting particle systems, Condensation

---

**Metastability in the reversible inclusion process II: Toward the characterization of the relaxation time**

Alessandra Bianchi*
Mathematics Department, Università di Padova, Italy

**Abstract:** The inclusion process (IP) is a stochastic lattice gas where particles perform random walks subjected to mutual attraction, thus providing the natural bosonic counterpart of the well-studied exclusion process. Due to the attractive interaction between particles, the IP can exhibit a condensation transition, where a positive fraction of all particles concentrates on a single site. In this talk, following the setting and results presented by S. Dommers, we consider the reversible IP on a finite set $S$, in the limit of total number of particles going to infinity, and address the problem of characterizing the relaxation time of the dynamics. This will be connected with the study of the formation of a condensate starting from two half-condensates (nucleation phase) and will lead to the identification of possibly many transition time scales, which are generally longer than that characterizing the dynamics of the condensate on neighboring sites (discussed in Dommers’s talk). We approach the problem starting from potential theoretic techniques and following some recent related ideas developed in a few papers we will refer to. [Joint work with S. Dommers and C. Giardinà.]

Keywords: Inclusion process, Condensation, Metastability, Relaxation time

---

**Condensation and metastability of general inclusion processes**

Seonwoo Kim* and Insuk Seo
Seoul National University

**Abstract:** In this talk, we present various results regarding the phenomenon of condensation and metastability of inclusion processes on a wide class of underlying graphs. Condensation denotes the phenomenon when a majority of the particles assemble on a single site. It occurs in various interacting particle systems, including the current one, due to the attractive behavior of the particles. In a bigger time scale, because of the small randomness of the system, the formed condensate breaks up and forms another one on a different site. This is a typical example of the metastable behavior, which is quite ubiquitous in numerous stochastic systems. The metastable behavior of the current model is known to exhibit the scheme of multiple time scales; thus, we first explain the reason of presence of such a scheme. Moreover, the fundamental behavior of metastability dramatically differs between the reversible and non-reversible ones. Therefore, the main results are divided into two parts: reversible case and non-reversible case. In the reversible case, the results are known fairly in details. In the context of multiple time scales scenario, the metastable behavior is fully characterized up to the second time scale. We present the known results along with the conjectures for the unsolved regime. In the non-reversible case, a much more complicated scenario emerges even in the first time scale. The main difficulty is that we do not have an explicit formula for the invariant measure. We explain how to overcome this drawback and characterize the condensation and metastability in the first time scale which generalizes the previous results in the reversible case.

Keywords: Metastability, Inclusion process, Non-reversibility, Multiple time scales
**Condensation of SIP particles and sticky Brownian motion**
Gioia Carinci*
Department of Physics, Informatics and Mathematics, Università di Modena e Reggio Emilia, Italy

**Abstract:** The symmetric inclusion process (SIP) is a particle system with attractive interaction. We study its behavior in the condensation regime attained for large values of the attraction intensity. Using Mosco convergence of Dirichlet forms, we prove convergence to sticky Brownian motion for the distance of two SIP particles. We use this result to obtain, via duality, an explicit scaling for the variance of the density field in this regime, for the SIP initially started from a homogeneous product measure. This provides relevant new information on the coarsening dynamics of condensing particle systems on the infinite lattice.

[Joint works with M. Ayala, C Giardina and F. Redig.]

**Keywords:** Interacting Particle Systems, Scaling Limits

---

**Condensed phase structure in inclusion processes**
Watthanan Jatuviriyapornchai1*, Paul Chleboun2 and Stefan Grosskinsky3
1Department of Mathematics, Faculty of Science, Mahidol University, Thailand
2Department of Statistics, University of Warwick, UK
3Delft Institute of Applied Mathematics, TU Delft, Netherlands

**Abstract:** We establish a complete picture of condensation in the inclusion process in the thermodynamic limit. The condensed phase structure is derived by a size-biased sampling of occupation numbers. Our results cover the entire scaling regimes of the diffusion parameter, especially an interesting hierarchical structure characterized by the Poisson-Dirichlet distribution. Whereas our results are rigorous, Monte-Carlo simulation and recursive numerics for partition functions are presented to illustrate the main points.

**Keywords:** Condensation, Inclusion process, Poisson-Dirichlet distribution, Size-biased sampling

---

**Spectral large deviations for sparse random matrices**
Kyeongsik Nam*
Mathematics, University of California, Los Angeles, USA

**Abstract:** The large deviation problem for the spectrum of random matrices has attracted immense interest. It was first studied for GUE and GOE, which are exactly solvable, and subsequently studied for Wigner matrices with general distributions. Once the sparsity is induced (i.e. each entry is multiplied by the independent Bernoulli distribution Ber(p)), eigenvalues can exhibit a drastically different behavior. Constant average degree of sparsity, p~1/n, is an interesting case since universality breaks down at this regime of sparsity. In this talk, I will consider Gaussian ensembles with sparsity p=1/n and talk about the typical behavior and large deviation estimates for the largest eigenvalue.

[Joint work with Shirshendu Ganguly.]

**Keywords:** Random matrices, Large deviations, Eigenvalues
Robust hypergraph clustering via convex relaxation of truncated MLE
Jeonghwan Lee1, Daesung Kim2 and Hye Won Chung2*
1Department of Mathematical Sciences, KAIST, Korea
2School of Electrical Engineering, KAIST, Korea

Abstract: We study hypergraph clustering in the weighted d-uniform hypergraph stochastic block model (d-WHSBM), where each edge consisting of d nodes from the same community has higher expected weight than the edges consisting of nodes from different communities. We propose a new hypergraph clustering algorithm, called CRTMLE, and provide its performance guarantee under the d-WHSBM for general parameter regimes. We show that the proposed method achieves the order-wise optimal or the best existing results for approximately balanced community sizes. Moreover, our results settle the first recovery guarantees for growing number of clusters of unbalanced sizes. Involving theoretical analysis and empirical results, we demonstrate the robustness of our algorithm against the unbalancedness of community sizes or the presence of outlier nodes.

Keywords: Hypergraph, Clustering, MLE, SDP

Convergence rate to the Tracy-Widom laws for the largest eigenvalue of Wigner matrices
Kevin Schnelli* and Yuanyuan Xu
Department of Mathematics, KTH Royal Institute of Technology, Sweden

Abstract: In this talk we discuss quantitative versions of the edge universality for Wigner matrices and related random matrix models. The fluctuations of the largest rescaled eigenvalues for the Gaussian invariant ensembles are described by the Tracy-Widom laws. The universality of these laws has in recent years been established for many non-invariant random matrix models. We will present new results on the convergence rate to the universal laws for the largest eigenvalues of Wigner matrices and high-dimensional sample covariance matrices.

Keywords: Tracy-Widom laws, Edge universality, Random matrix

Attributed graph alignment
Lele Wang*
Electrical and Computer Engineering, University of British Columbia, Canada

Abstract: Motivated by various data science applications including de-anonymizing user identities in social networks, we consider the graph alignment problem, where the goal is to identify the vertex/user correspondence between two correlated graphs. Existing work mostly recovers the correspondence by exploiting the user-user connections. However, in many real-world applications, additional information about the users, such as user profiles, might be publicly available. In this paper, we introduce the attributed graph alignment problem, where additional user information, referred to as attributes, is incorporated to assist graph alignment. We establish sufficient and necessary conditions for recovering vertex correspondence exactly, where the conditions match for a wide range of practical regimes. Our results recover existing tight information-theoretic limits for models where only the user-user connections are available, and further span the full spectrum between these models and models where only attribute information is available.

[This is joint work with Ning Zhang and Weina Wang.]

Keywords: Graph matching, Graph alignment, Network de-anonymization
Minkowski content for the scaling limit of loop-erased random walk in three dimensions
Xinyi Li*
BICMR, Peking University, China

Abstract: In this talk, we will discuss loop-erased random walk in three dimensions and its scaling limit, and briefly explain how to prove the existence of the Minkowski content of the latter and why it gives the scaling limit of the former in natural parametrization.
[This is joint work with Daisuke Shiraishi (Kyoto).]

Keywords: Loop-erased random walk, Minkowski content

Let $K$ be a Kato class measure on $\mathbb{R}^d$ for some $d \geq 1$. We are interested in extending the definition of $K$ to the setting of symmetric Markov processes on $\mathbb{R}^d$. In this talk, we will discuss the relationship between $K$ and the Dirichlet form of the underlying Markov process.

Keywords: $L^p$-class measures, $L^p$-Dynkin class measures, Dirichlet form, Symmetric Markov process

Green function estimates and boundary Harnack principles for non-local operators whose kernels degenerate at the boundary
Panki Kim*
Department of Mathematical Sciences, Seoul National University, Korea

Abstract: In this talk, we discuss the potential theory of Markov processes with jump kernels decaying at the boundary of the half space. The boundary part of kernel is comparable to the product of three terms with parameters appearing as exponents in these terms. The constant $c$ in the killing term can be written as a function of a parameter p which is strictly increasing in $p$. We establish sharp two-sided estimates on the Green functions of these processes for all admissible values of $p$ and parameters in the boundary part of kernel. Depending on the regions where parameters and $p$ belong, the estimates on the Green functions are different. In fact, the estimates have three different forms depending on the regions the parameters belong to. As applications, we completely determine the region of the parameters where the boundary Harnack principle holds or not.
Heat kernel upper bounds for symmetric Markov semigroups
Jian Wang*
College of Mathematics and Informatics, Fujian Normal University, China

Abstract: It is well known that Nash-type inequalities for symmetric Dirichlet forms are equivalent to on-diagonal heat kernel upper bounds for associated symmetric Markov semigroups. In this talk, we show that both imply (and hence are equivalent to) off-diagonal heat kernel upper bounds under some mild assumptions. Our approach is based on a new generalized Davies's method. Our results extend that by Carlen-Kusuoka-Stroock for Nash-type inequalities with power order considerably and also extend that by Grigor'yan for second order differential operators on a complete non-compact manifold.
[The talk is based on a joint work with Z.-Q. Chen (Seattle), P. Kim (Seoul) and T. Kumagai (Kyoto).]

Inverse local time of one-dimensional diffusions and its comparison theorem
Lidan Wang*
School of Statistics and Data Science, Nankai University, China

Abstract: It is well known that for a reflecting Bessel process, the inverse local time at 0 is an \( \alpha \)-stable subordinator, then the corresponding subordinate Brownian motion is a \( 2\alpha \)-stable process. Based on discussions of some transforms and regenerative theory of general diffusions, we get a comparison result between inverse local times of Bessel processes and perturbed Bessel processes. An immediate application will be the stability of Green function estimates for trace processes.
[This talk is based on a joint work with Prof. Zhen-Qing Chen.]
Keywords: Diffusions, Inverse local time, Non-local operators, Comparison theorem

Archimedes' principle for ideal gas
Krzysztof Burdzy\(^1\)* and Jacek Malecki\(^2\)
\(^1\)Mathematics, University of Washington, USA
\(^2\)Mathematics, Technical University of Wroclaw, Poland

Abstract: I will present Archimedes' principle for a macroscopic ball in ideal gas consisting of point particles with non-zero mass. The main result is an asymptotic theorem, as the number of point particles goes to infinity and their total mass remains constant. Asymptotically, the gas has an exponential density as a function of height. The asymptotic inverse temperature of the gas is identified as the parameter of the exponential distribution.
Keywords: Archimedes principle
Rough path theory and the Stochastic Loewner Equation

Vlad Margarint*
Mathematics, NYU Shanghai, China

Abstract: In this talk, I will give an overview of some work carried at the intersection of Rough Path Theory and Schramm-Loewner Evolutions (SLE) Theory. Specifically, I will cover a study of the Loewner Differential Equation using Rough Path techniques (and beyond). The Loewner Differential Equation describes the evolution of a family of conformal maps. We rephrase this in terms of (Singular) Rough Differential Equations. In this context, it is natural to study questions on the stability, and approximations of solutions of this equation. First, I will present a result on continuity of the dynamics and related objects in a natural parameter that appears in the problem. The first approach will be based on Rough Path Theory, and a second approach will be based on a constructive method of independent interest: the square-root interpolation of the Brownian driver of the Loewner Differential Equation. In the second part, if time permits, I will present a result on the asymptotic radius of convergence of the Stochastic Taylor approximation of the Loewner Differential Equation and numerical simulations of the SLE trace using a novel numerical method: Ninomiya-Victoir (or Strang) splitting.

[The first part is based on joint work with Dmitry Belyaev, Terry Lyons, and the second part on a collaboration with James Foster.]

Keywords: Rough Path Theory, SLE

Rough path with jumps and its application in homogenization

Huilin Zhang*
School of Mathematics, Fudan University, China

Abstract: In this talk I will present recent progress on rough path theory with jumps, consisting of Ito/forward theory and the Stratonovich/Marcus theory. It allows us to handle stochastic differential equations driven by jump noise. As an application, we show how rough paths can be applied in the homogenization of a fast-slow system proposed by Melbourne and Stuart.

[This talk is based on works with Chevyrev, Friz, Korepanov and Melbourne.]

Keywords: Rough paths, Jumps, Homogenization

Probabilistic rough paths

William Salkeld* and Francois Delarue
Mathematics, Universite Cote d’Azur, UK

In this talk, I will explain some of the foundation results for a new regularity structure developed to study interactive systems of equations and their mean-field limits. At the heart of this solution theory is a Taylor expansion using the so called Lions measure derivative. This quantifies infinitesimal perturbations of probability measures induced by infinitesimal variations in a linear space of random variable.

[This talk is based on preprints and ongoing work with my supervisor Francois Delarue at Universite Cote d’Azur.]

Keywords: Regularity structure, Lions derivative, rough path, McKean-Vlasov equations
Transport and continuity equations with (very) rough noise
Carlo Bellingeri\textsuperscript{1}, Ana Djurdjevac\textsuperscript{2}, Peter K. Friz\textsuperscript{3} and Nikolas Tapia\textsuperscript{4}\textsuperscript{*}
\textsuperscript{1}Institute for Mathematics, TU Berlin, Germany
\textsuperscript{2}FB Mathematik und Informatik, FU Berlin, Germany
\textsuperscript{3}Institute for Mathematics, TU Berlin, Germany
\textsuperscript{4}FG6, Weierstrass Institute/TU Berlin, Germany

Abstract: We study the solution theory of linear transport equations driven with rough multiplicative noise. We show existence and uniqueness for rough flows driven by an arbitrary geometric rough path, and obtain a rough version of the classical method of characteristics, under a boundedness condition for the vector fields. We also obtain an adjoint RDE for the derivatives of the induced flow. Dually, we show existence and uniqueness for the associated continuity equation.

Keywords: Rough stochastic partial differential equations, Transport equation

Rough walks in random environment
Tal Orenshtein\textsuperscript{*}
Mathematics, WIAS Berlin, Germany

Abstract: Random walk in random environment (RWRE) is a model to describe propagation of heat or diffusion of matter through a highly irregular medium. The latter is expressed locally in the model in terms of a random environment according to which the process evolves randomly in time. In a few fundamental classes the phenomenon of homogenization of the media takes place. One way this can be expressed is in the fact that on large scales, the RWRE fluctuates as a Brownian motion with a deterministic covariance matrix given in terms the (law of the) environment. Rough path theory enables the construction of solutions to SDEs so that the solution map is continuous with respect to the noise. One important application guarantees that if the approximation converges to the noise in the rough path topology, the SDEs driven by the noise approximations converge, in an appropriate sense, to a well-defined SDE which is different than the original one, so that the correction term is explicit in terms of the noise approximation. In this talk we shall present our current program, in which one lifts RWRE in various classes to the rough path space and shows a convergence to an enhanced Brownian motion in the rough path topology. Interestingly, the limiting second level of the lifted RWRE may have a linear correction, called area anomaly, which we identify. Except for the immediate application to approximations of SDEs, and potentially to SPDEs, this adds some new information on the RWRE limiting path. Time permitted, we shall elaborate on the tools to tackle these problems. [Based on joint works with Olga Lopusanschi, with Jean-Dominique Deuschel and Nicolas Perkowski and with Johaness Bäumler, Noam Berger and Martin Slowik.]

Keywords: Rough paths, Area anomaly, Random walks in random environment, Homogenization

OCS09 Random Matrices and Infinite Particle Systems
JULY 21 (WED), 11:30 - 12:00 KST

Dynamical universality for random matrices
Hirofumi Osada\textsuperscript{*}
Department of Mathematics, Kyushu University, Japan

Abstract: We establish an invariance principle corresponding to the universality of random matrices. More precisely, we prove dynamical universality of random matrices in the sense that, if random point fields $\mu_N$ of $N$-particle
systems describing eigenvalues of random matrices or log-gases with general self-interaction potentials $V^N$ converge to some random point field $\mu$, then the associated natural $\mu^N$-reversible diffusion processes represented by solutions of stochastic differential equations (SDE) converge to some $\mu$-reversible diffusion processes given by a solution of the infinite-dimensional stochastic differential equations (ISDE). Our results are general theorems and can be applied to various random point fields related to random matrices such as sine, Airy, Bessel, and Ginibre random point fields. The representations of finite-dimensional SDEs describing the $N$-particle systems are very complicated in general. The limit ISDE has simple and universal representations, nevertheless, according to a class of random matrices such as bulk, soft-edge, and hard-edge scaling. We thus prove ISDE such that the infinite-dimensional Dyson model and Airy, Bessel, and Ginibre interacting Brownian motions are universal dynamical objects. The key ingredients are (1) Local uniform convergence of correlation functions to that of the limit point process. (2) The uniqueness of a weak solution of the limit ISDE, which deduces the uniqueness of Dirichlet forms. Concerning (2), we use the result in [1] and [2].


**Keywords:** Random matrices, Dynamical universality, Infinite-particle systems, Infinite-dimensional stochastic differential equations

---

**Signal processing via the stochastic geometry of spectrogram level sets**

Subhroshekhar Ghosh*

Department of Mathematics, National University of Singapore, Singapore

**Abstract:** Spectrograms are fundamental tools in the detection, estimation and analysis of signals in the time-frequency analysis paradigm. The spectrogram of a signal (usually corrupted with noise) is the squared magnitude of its short time Fourier transform (STFT), which in turn is a generalised version of the classical Fourier transform, augmented with a window in the time domain. Signal analysis via spectrograms has traditionally explored their peaks, i.e. their maxima, complemented by a recent interest in their zeros or minima. In particular, recent investigations have demonstrated connections between Gabor spectrograms of Gaussian white noise and Gaussian analytic functions (abbrv. GAFs) in different geometries. However, the zero sets (or the maxima or minima) of GAFs have a complicated stochastic structure, which makes a direct theoretical analysis of usual spectrogram based techniques via GAFs a difficult proposition. These techniques, in turn, largely rely on statistical observables from the analysis of spatial data, whose distributional properties for spectrogram extrema are mostly understood only at an empirical level. In this work, we investigate spectrogram analysis via the stochastic, geometric and analytical properties of their level sets. We obtain theorems demonstrating the efficacy of a spectrogram level sets based approach to the detection and estimation of signals, framed in a concrete inferential set-up. Exploiting these ideas as theoretical underpinnings, we propose a level sets based algorithm for signal analysis that is intrinsic to given spectrogram data. We substantiate the effectiveness of the algorithm by extensive empirical studies. Our results also have theoretical implications for spectrogram zero based approaches to signal analysis.

[Based on joint work with Meixia Lin and Dongfang Sun.]

---

**Logarithmic derivatives and local densities of point processes arising from random matrices**

Shota Osada*

Kyushu University

**Abstract:** We talk about a distribution (a generalized function) theory for point processes. We show that a logarithmic
derivative in the distributional sense can indicate the local density of the point process. This theory is especially effective for point processes appearing in random matrix theory. In particular, using this result, we solve infinite-dimensional stochastic differential equations associated with the point process given by de Branges spaces, so-called integrable kernels, and random matrices such as Airy, sine, and Bessel point processes. [2] Conventionally, the point process that describes an infinite particle system is described by the Dobrushin-Lanford-Ruelle (DLR) equation. The point process of an infinite particle system appearing in a random matrix has a logarithmic potential as an interaction potential. Because the logarithmic potential is not integrable at infinity, the DLR equation cannot describe the point process as it is. Logarithmic derivative for point process is a concept introduced in [1] to settle this problem. There must be a logarithmic derivative and local density of the point process to solve the infinite-dimensional stochastic differential equation. [3] With our result, the existence of a logarithmic derivative with suitable integrability is sufficient for the construction of the stochastic dynamics as a solution of infinite-dimensional stochastic differential equations.


**Keywords:** Determinantal point processes, Random matrices, Infinite-dimensional stochastic differential equations

---

**Stochastic differential equations for infinite particle systems of jump type with long range interactions**  
Hideki Tanemura*

Department of Mathematics, Keio university, Japan

**Abstract:** Infinite dimensional stochastic differential equations (ISDEs) describing systems with an infinite number of particles are considered. Each particle undergoes Levy process, and interaction between particles is given by long range interaction potential, which is not only of Ruelle's class but also logarithmic. We discuss the existence and uniqueness of strong solutions of the ISDEs.

[This talk is based on a collaboration with Shota Esaki (Fukuoka University).]

**Keywords:** Random Matrix, Levy process, Stochastic differential equation

---

**Loewner dynamics for the multiple SLE(0) process**  
Tom Alberts*1, Nikolai Makarov2 and Nam-Gyu Kang3

1Mathematics, The University of Utah, USA  
2Mathematics, California Institute of Technology, USA  
3Mathematics, KIAS, Korea

**Abstract:** Recently Peltola and Wang introduced the multiple SLE(0) process as the deterministic limit of the random multiple SLE(κ) curves as kappa goes to zero. They prove this result by means of a "small κ" large deviations principle, but the limiting curves also turn out to have important geometric characterizations that are independent of their relation to SLE(κ). In particular, they show that the SLE(0) curves can be generated by a deterministic Loewner
evolution driven by multiple points, and the vector field describing the evolution of these points must satisfy a particular system of algebraic equations. We show how to generate solutions to these algebraic equations in two ways: first in terms of the poles and critical points of an associated real rational function, and second via the well-known Calogero-Moser integrable system with particular initial velocities. Although our results are purely deterministic they are again motivated by taking limits of probabilistic constructions, which I will explain.

**Keywords:** Schramm-Loewner Evolution, Real rational functions, Calogero-Moser system, Large deviations

---

**Conformal field theory for annulus SLE**

Sung-Soo Byun and Nam-Gyu Kang

1Mathematical Science, Seoul National University, Korea
2School of Mathematics, Korea Institute for Advanced Study, Korea

**Abstract:** In this talk, I will present a constructive conformal field theory generated by central/background charge modifications of Gaussian free fields in a doubly conected domain and outline its connection to annulus SLE theory. Furthermore, I will explain some applications, which include Coulomb gas solutions to the null-vector equation for annulus SLE partition functions and hitting probabilities of level lines of Gaussian free fields.

**Keywords:** Gaussian free field, Conformal field theory, Annulus SLE, SLE martingale-observables

---

**Convergence of martingale observables in the massive FK-Ising model**

S. C. Park

School of Mathematics, Korea Institute of Advanced Study, Korea

**Abstract:** We show the convergence of fermionic martingale observables (MO) of the FK-Ising model in massive scaling limit. We generalise, along with a recent work by Chelkak-Izyurov-Mahfouf (2021), the discrete complex analytic machinery developed in Chelkak-Smirnov (2012) for the critical isoradial setup into the massive setting. No assumptions on domain regularity or the direction of massive perturbation are imposed. We then discuss implications on the interface curves and Russo-Seymour-Welsh (RSW) type crossing estimates, as well as ongoing work on the spin model with Chelkak and Wan.

**Keywords:** Probability, Statistical Mechanics, Complex Analysis

---

**Boundary Minkowski content of multi-force-point SLEκ(ρ) curves**

Dapeng Zhan

Mathematics, Michigan State University, USA

**Abstract:** We consider a transient chordal \( \text{SLE}_κ(ρ_1, ..., ρ_m) \) curve in \( \mathbb{H} \) started from 0 with force points \( 0 ≥ υ_1 ≥ ⋯ ≥ υ_m \), which intersects but is not boundary-filling on the interval \( (−∞, υ_m) \). The main result of this work is that, almost surely there is an atomless locally finite random Borel measure \( µ_η \) on \( η \cap (−∞, υ_m) \) such that for any \( υ < υ_m \), the \( d \)-dimensional Minkowski content of \( η \cap [υ, υ_m] \) exists and equals \( µ_η[υ, υ_m] \), where \( d = \frac{(|p|+4)(κ+4−2|p|)}{2κ} \) is the Hausdorff dimension of \( η \cap [υ, υ_m] \), and \( |p| = \sum_{j=1}^m ρ_j \). Such measure, called Minkowski content measure, satisfies conformal covariance properties. We prove that, in the case of chordal SLEκ with no force point, this measure agrees with the covariant measure derived in [Alberts-Sheffield, 2011] up to a constant depending on κ, and thus proved a conjecture raised there. To construct the Minkowski content measure, we follow the standard approach to study the one-point and two-point boundary Green’s functions of \( η \) at points on \( (−∞, υ_m) \), which are
the rescaled probability that $\eta$ passes through small discs centered at one or two points on $(-\infty, \nu_m)$. We derive the existence of these Green's functions as well as the convergence rate. We also obtain the exact formula of the one-point Green's function up to an unknown multiplicative constant.

**Keywords:** SLE, Minkowski content, Covariant measure, Green's function

---

### Holes in first-passage percolation

**Wai-Kit Lam**

Mathematics, University of Minnesota, USA

**Abstract:** In first-passage percolation (FPP), one places i.i.d. nonnegative weights $(\tau_e)$ on the nearest-neighbor edges on $\mathbb{Z}^d$, and studies the induced pseudo-metric. One of the main goals in FPP is to understand the geometry of the metric ball $B(t)$ centered at the origin with radius $t$. When the weights $(\tau_e)$ have a heavy-tailed distribution, it is known that $B(t)$ consists of a lot of small holes. It is natural to ask whether large holes typically exist; if yes, how large can they be? In an ongoing project with M. Damron, J. Gold and X. Shen, we show that for any distribution with $P(\tau_e = 0) < p_c$, a.s. for all large $t$, the size of the largest hole is of order at least $\log t$, and the number of holes is of order at least $t^{d-1}$. If we assume the limiting shape of $B(t)$ is uniformly curved (which is unproven), we can also show that in two dimensions, a.s. the size of the largest hole is at most of order $(\log t)^C$.

**Keywords:** First-passage percolation, Limit shape

---

### Coalescence estimates for the corner growth model with exponential weights

**Timo Seppäläinen and Xiao Shen**

Department of mathematics, UW-Madison, USA

**Abstract:** We establish estimates for the coalescence time of semi-infinite directed geodesics in the planar corner growth model with i.i.d. exponential weights. There are four estimates: upper and lower bounds for both fast and slow coalescence on the correct scale with exponent $3/2$. The lower bound for fast coalescence is new and has the optimal exponential order of magnitude. For the other three, we provide proofs that do not rely on integrable probability or on the connection with the totally asymmetric simple exclusion process, in order to provide a template for the extension to other models. We utilize a geodesic duality introduced by Pimentel and properties of the increment-stationary last-passage percolation process.

**Keywords:** Random growth model, Last-passage-percolation

---

### Scaling limits of sandpiles

**Ahmed Bou-Rabee**

Statistics, University of Chicago, USA

**Abstract:** The Abelian sandpile is a diffusion process on the integer lattice which produces striking, kaleidoscopic patterns. I will discuss recent progress towards understanding these patterns and their stability under randomness.

**Keywords:** Scaling limit, Abelian sandpile
DeepKriging: Spatially dependent deep neural networks for spatial prediction

Ying Sun*
KAUST

Abstract: In spatial statistics, a common objective is to predict the values of a spatial process at unobserved locations by exploiting spatial dependence. In geostatistics, Kriging provides the best linear unbiased predictor using covariance functions and is often associated with Gaussian processes. However, when considering non-linear prediction for non-Gaussian and categorical data, the Kriging prediction is not necessarily optimal, and the associated variance is often overly optimistic. We propose to use deep neural networks (DNNs) for spatial prediction. Although DNNs are widely used for general classification and prediction, they have not been studied thoroughly for data with spatial dependence. In this work, we propose a novel neural network structure for spatial prediction by adding an embedding layer of spatial coordinates with basis functions. We show in theory that the proposed DeepKriging method has multiple advantages over Kriging and classical DNNs only with spatial coordinates as features. We also provide density prediction for uncertainty quantification without any distributional assumption and apply the method to PM2.5 concentrations across the continental United States.

A model-free subsampling method based on minimum energy criterion

Wenlin Dai*
Institute of Statistics and Big Data, Renmin University of China, China

Abstract: The extraordinary amounts of data generated in science today pose heavy demands on computational resources and time, which hinders the implementation of various statistical methods. An efficient and popular strategy of downsizing data volumes and hence alleviating these challenges is subsampling. However, the existing methods either rely on specific assumptions for the underlying models or acquire only partial information from the available data. We propose a novel approach, termed adaptive subsampling, that is based on the minimum energy criterion (ASMEC). The proposed method requires no explicit model assumptions and ‘smartly’ incorporates information on covariates and responses. ASMEC subsamples possess two desirable properties: space-filling and spatial adaptiveness to the full data. We investigate the theoretical properties of the ASMEC estimator under the smoothing spline regression model and show that it converges at an identical rate to two recently proposed basis selection methods. The effectiveness and robustness of the ASMEC approach are also supported by a variety of simulated examples and two real-life examples.

Keywords: Basis selection, Massive data, Smoothing spline, Space-filling, Spatial adaptiveness

Global wind modeling with transformed Gaussian processes

Jaehong Jeong*
Department of Mathematics, Hanyang University, Korea

Abstract: Uncertainty quantification of wind energy potential from climate models can be limited because it requires considerable computational resources and is time-consuming. We propose a stochastic generator that aims at reproducing the data-generating mechanism of climate ensembles for global annual, monthly, and daily wind data. Inferences based on a multi-step conditional likelihood approach are achieved by balancing memory storage and
distributed computation for a large data set. In the end, we discuss a general framework for modeling non-Gaussian multivariate stochastic processes by transforming underlying multivariate Gaussian processes.

**Keywords:** Climate models, Non-Gaussian processes, Spatial Statistics

---

**Threshold estimation for continuous three-phase polynomial regression models with constant mean in the middle regime**

ChihHao Chang*
Institute of Statistics, National University of Kaohsiung, China

**Abstract:** This talk considers a continuous three-phase polynomial regression model with two threshold points for dependent data with heteroscedasticity. We assume the model is polynomial of order zero in the middle regime, and is polynomial of higher orders elsewhere. We denote this model by M2, which includes models with one or no threshold points, denoted by M1 and M0, respectively, as special cases. We provide an ordered iterative least squares (OiLS) method when estimating M2 and establish the consistency of the OiLS estimators under mild conditions. We also apply a model-selection procedure for selecting Mk; k=0,1,2. When the underlying model exists, we establish the selection consistency under the aforementioned conditions. Finally, we conduct simulation experiments to demonstrate the finite-sample performance of our asymptotic results.

**Keywords:** Threshold estimation, Polynomial regression, Model selection, Consistency

---

**Change points detection for high dimensional time series**

Likai Chen¹*, Wei Biao Wu² and Weining Wang³
¹mathematics and statistics, Washington University in Saint Louis, USA
²statistics, University of Chicago, USA
³Economics and Related Studies, University of York, UK

**Abstract:** We consider multiple change-points detection of high-dimensional time series. Asymptotic theory for testing the existence of breaks, the consistency and the asymptotic distribution of the breakpoint statistics and estimated break sizes will be provided. The theory backs up a simple two-step procedure for detecting and estimating multiple change-points. The proposed two-step procedure involves the maximum of a MOSUM (moving sum) type statistics in the first step and a CUSUM (cumulative sum) refinement step on an aggregated time series in the second step. Thus, for a fixed time-point, we can capture both the biggest break across different coordinates and aggregating simultaneous breaks over multiple coordinates. Extending the existing high-dimensional Gaussian approximation theorem to dependent data with jumps, the theory allows us to characterize the size and power of our multiple change-point test asymptotically. Moreover, we can make inferences on the breakpoints estimates when the break sizes are small. Our theoretical setup incorporates both weak temporal and strong or weak cross-sectional dependence and is suitable for heavy-tailed innovations. A robust long-run covariance matrix estimation is proposed, which can be of independent interest. An application on detecting structural changes of the U.S. unemployment rate is considered to illustrate the usefulness of our method.

**Keywords:** Multiple change points detection, Temporal and cross-sectional dependence

---

**OCS13 Recent Advances in Complex Time Series Analysis**

JULY 20 (TUE), 22:30 - 23:00 KST
Asymptotics of large autocovariance matrices
Monika Bhattacharjee*
Department of Mathematics, Indian Institute of Technology Bombay, India

Abstract: We consider the high-dimensional moving average process and explore the asymptotics for eigenvalues of its sample autocovariance matrices. Under quite weak conditions, we prove, in a unified way, that the limiting spectral distribution (LSD) of any symmetric polynomial in the sample autocovariance matrices, after suitable centering and scaling, exists and is non-degenerate. We use methods from free probability in conjunction with the method of moments to establish our results. In addition, we are able to provide a general description for the limits in terms of some freely independent variables. We also establish asymptotic normality results for the traces of these matrices. We suggest statistical uses of these results in problems such as order determination of high-dimensional MA and AR processes and testing of hypotheses for coefficient matrices of such processes.

Keywords: Autocovariances, Limiting spectral distribution, High-dimension, Time series

Factor models for matrix-valued high-dimensional time series
Xialu Liu*
Management information systems, San Diego State University, USA

Abstract: In finance, economics and many other fields, observations in a matrix form are often observed over time. For example, many economic indicators are obtained in different countries over time. Various financial characteristics of many companies are reported over time. Although it is natural to turn a matrix observation into a long vector then use standard vector time series models or factor analysis, it is often the case that the columns and rows of a matrix represent different sets of information that are closely interrelated in a very structural way. We propose a novel factor model that maintains and utilizes the matrix structure to achieve greater dimensional reduction as well as finding clearer and more interpretable factor structures. Estimation procedure and its theoretical properties are investigated and demonstrated with simulated and real examples.

Keywords: Factor model, Matrix valued-time series, High-dimensional data

Multi-level changepoint inference for periodic data sequences
Anastasia Ushakova1*, Simon A Taylor2 and Rebecca Killick1
1Mathematics and Statistics, Lancaster University, UK
2School of Mathematics, University of Edinburgh, UK

Abstract: Existing changepoint approaches consider changepoints to occur linearly in time; one changepoint happens after another and they are not linked. However, data processes may have regularly occurring changepoints, e.g. a yearly increase in sales of ice-cream on the first hot weekend. Using linear changepoint approaches here will miss more global features such as a decrease in sales of ice-cream in favour of sorbet. Being able to tease these global changepoint features from the more local (periodic) ones is beneficial for inference. We propose a periodic changepoint model to model this behaviour using a mixture of a periodic and linear time perspective. Built around a Reversible Jump Markov Chain Monte Carlo sampler, the Bayesian framework is used to study the local (periodic) changepoint behaviour. To identify the optimal global changepoint positions we integrate the local changepoint model into the pruned exact linear time (PELT) search algorithm. We demonstrate that the method detects both local and global changepoints with high accuracy on simulated and motivating applications that share periodic behaviour. Due to the multi-level nature of the analysis, visualisation of the results can be challenging. We additionally provide a unique multi-level perspective for changepoint visualisations in data sequences.
Keywords: PELT, Segmentation, Periodicity, Level shift

Bayesian spatial binary regression for label fusion in structural neuroimaging
Andrew Brown*
Clemson University

Abstract: Alzheimer’s disease is a neurodegenerative condition that accelerates cognitive decline relative to normal aging. It is of critical scientific importance to gain a better understanding of early disease mechanisms in the brain to facilitate effective, targeted therapies. The volume of the hippocampus can be used as an aid to diagnosis and disease monitoring. Measuring this volume via neuroimaging is difficult since each hippocampus must either be manually identified or automatically delineated, a task referred to as segmentation. Automatic hippocampal segmentation often involves mapping a previously manually segmented image to a new brain image and propagating the labels to obtain an estimate of where each hippocampus is located in the new image. A more recent approach to this problem is to propagate labels from multiple manually segmented atlases and combine the results using a process known as label fusion. To date, most label fusion algorithms either employ voting procedures or impose prior structure and subsequently find the maximum a posteriori estimator through optimization. We propose using a fully Bayesian spatial regression model for label fusion that facilitates direct incorporation of covariate information while making accessible the entire posterior distribution. Our results suggest that incorporating tissue classification (gray matter, white matter, etc.) into the label fusion procedure can greatly improve segmentation when relatively homogeneous, healthy brains are used as atlases for diseased brains. The fully Bayesian approach also allows quantification of the associated uncertainty, information which we show can be leveraged to detect significant differences between healthy and diseased populations that would otherwise be missed.

Convex clustering analysis for histogram-valued data
Cheolwoo Park*
KAIST

Abstract: In recent years, there has been increased interest in symbolic data analysis, including for exploratory analysis, supervised and unsupervised learning, time series analysis, etc. Traditional statistical approaches that are designed to analyze single-valued data are not suitable because they cannot incorporate the additional information on data structure available in symbolic data, and thus new techniques have been proposed for symbolic data to bridge this gap. In this article, we develop a regularized convex clustering approach for grouping histogram-valued data. The convex clustering is a relaxation of hierarchical clustering methods, where prototypes are grouped by having exactly the same value in each group via penalization of parameters. We apply two different distance metrics to measure (dis)similarity between histograms. Various numerical examples confirm that the proposed method shows better performance than other competitors.

Keywords: Clustering, Histogram-valued data, Wasserstein-Kantorovich metric
A geometric mean for multivariate functional data
Juhyun Park*
ENSIIE

Abstract: The analysis of curves has been routinely dealt with using tools from functional data analysis. However, its extension to multi-dimensional curves poses a new challenge due to its inherent geometric features that are difficult to capture with the classical approaches that rely on linear approximations. We propose an alternative notion of mean that reflects shape variation of the curves. Based on a geometric representation of the curves through the Frenet-Serret ordinary differential equations, we introduce a new definition of mean curvature and mean shape through the mean ordinary differential equation. We formulate the estimation problem in a penalized regression and develop an efficient algorithm. We demonstrate our approach with both simulated data and a real data example.

Keywords: Multivariate functional data, Curvature estimation, Shape analysis

A confidence region for the elastic shape mean of planar curves
Justin Strait*
Statistics, University of Georgia, USA

Abstract: Visualization is an integral component of statistical shape analysis, where the goal is to perform inference on shapes of objects. When interested in identifying shape variation, one typically performs principal component analysis (PCA) to decompose total variation into orthogonal directions of variation. In many cases, shapes observe multiple sources of variation; using PCA to visualize requires decomposition into several plots displaying each mode of variation, without the ability to understand how these components work together. In this talk, I will discuss a constructive confidence region associated with the elastic shape mean, with a significant emphasis on producing a succinct visual summary of this region. The use of elastic shape representations allows for optimal matching of shape features, yielding more appropriate estimation of shape variation compared to other approaches within the shape analysis literature. The proposed region is demonstrated on simulated data, as well as common shapes from the MPEG-7 dataset (popular in computer vision applications).

Keywords: Shape analysis, Principal component analysis, Geometry, Visualization

Estimation of particulate levels using deep Dehazing network and temporal prior
SungHwan Kim*
Applied Statistics, Konkuk University, Korea

Abstract: Particulate matters (PM) have become one of the important pollutants that deteriorate public health. Since PM is ubiquitous in the atmosphere, it is closely related to life quality in many different ways. Thus, a system to accurately monitor PM in diverse environments is imperative. Previous studies using digital images have relied on individual atmospheric images, not benefiting from both spatial and temporal effects of image sequences. This weakness led to undermining predictive power. To address this drawback, we propose a predictive model using the deep dehazing cascaded CNN and temporal priors. The temporal prior accommodates instantaneous visual moves and estimates PM concentration from residuals between the original and dehazed images. The present method also provides, as by-product, high-quality dehazed image sequences superior to the nontemporal methods.
improvements are supported by various experiments under a range of simulation scenarios and assessments using standard metrics.

**Keywords:** Particulates, Deep learning

**Graph-regularized contextual bandits with scalable Thompson sampling and semi-parametric reward models**

Young-Geun Choi¹*, Seunghoon Paik², Gi-Soo Kim³ and Myunghae Paik⁴

¹Department of Statistics, Sookmyung Women’s University, Korea  
²Department of Mathematical Sciences, Seoul National University, Korea  
³Department of Industrial Engineering, UNIST, Korea  
⁴Department of Statistics, Seoul National University, Korea

**Abstract:** Graph-based bandit algorithms for multiple users have received attention as the relationship among users captured by a social network or graph can improve personalized content recommendation. The graph-based Thompson sampling algorithm by Vaswani (2017) is one of the state-of-the-art method where the relationship between users is represented by a simple undirected graph, however, a large graph poses computational challenges. We propose a novel Thompson sampling algorithm for multiple users with graph. We show that the proposed algorithm improves the regret bound by \( \sqrt{n} \) from the algorithm by Vaswani (2017), where \( n \) is the number of users. Furthermore, we propose a method for a semi-parametric bandit problem with multiple users and graph, which is the first algorithm proposed in the same setting. We show that the upper bound of the cumulative regret has the same order as in the setting without the semi-parametric term. In establishing the proposed algorithms, novel local estimators play a crucial role in improving the bound with reduced computational cost.

**INN: A stable method identifying clean-annotated samples via consistency effect in deep neural networks**

Dongha Kim¹*, Yongchan Choi², Kunwoong Kim² and Yongdai Kim²

¹Department of Statistics, Sungshin Women’s University, Korea  
²Department of Statistics, Seoul National University, Korea

**Abstract:** In classification problems with deep neural networks, researchers have been in trouble with collecting massive clean-annotated data and a lot of efforts have been put in to handle data with noisy labels. Many recent solutions for noisy label problems share the key idea so-called the memorization effect. While the memorization effect is a powerful tool, the performances are sensitive to the choice of a training epoch necessary in utilizing the memorization effect. In this paper, we introduce a new method called INN (Integration with the Nearest Neighborhoods) for refining noisy labels, which is more stable as well as more powerful. Our method is based on a new finding, called consistency effect, that the discrepancies of predictions at neighbor regions of clean labeled data and noisy labeled data are consistently observed regardless of training epochs. By applying the INN to the DivideMix algorithm, we propose a new learning framework called INN-DivideMix which improves the INN. By conducting various experiments including both performance test and ablation study, we demonstrate the superiority and stability of our proposed two methods.

**Keywords:** Deep learning, Noisy label problem, Consistency effect
An efficient parallel block coordinate descent algorithm for large-scale precision matrix estimation using graphics processing units
Donghyeon Yu*
Statistics, Inha University, Korea

Abstract: Large-scale sparse precision matrix estimation has attracted wide interest from the statistics community. The convex partial correlation selection method (CONCORD) developed by Khare et al. (2015) has recently been credited with some theoretical properties for estimating sparse precision matrices. The CONCORD obtains its solution by a coordinate descent algorithm (CONCORD-CD) based on the convexity of the objective function. However, since a coordinate-wise update in CONCORD-CD is inherently serial, a scale-up is nontrivial. In this paper, we propose a novel parallelization of CONCORD-CD, namely, CONCORD-PCD. CONCORD-PCD partitions the off-diagonal elements into several groups and updates each group simultaneously without harming the computational convergence of CONCORD-CD. We guarantee this by employing the notion of edge coloring in graph theory. Specifically, we establish a nontrivial correspondence between scheduling the updates of the off-diagonal elements in CONCORD-CD and coloring the edges of a complete graph. It turns out that CONCORD-PCD simultaneously updates off-diagonal elements in which the associated edges are colorable with the same color. As a result, the number of steps required for updating off-diagonal elements reduces from p(p-1)/2 to p-1 (for even p) or p (for odd p), where p denotes the number of variables. We prove that the number of such steps is irreducible. In addition, CONCORD-PCD is tailored to single-instruction multiple-data (SIMD) parallelism. A numerical study shows that the SIMD-parallelized PCD algorithm implemented in graphics processing units (GPUs) boosts the CONCORD-CD algorithm multiple times.

Keywords: CONCORD, Parallel coordinate descent, Edge coloring, Graphics processing units

Functional linear regression model with randomly censored data: predicting conversion time to Alzheimer’s disease
Seong J. Yang*, Hyejin Shin†, Sang Han Lee‡ and Seokho Lee§
1Department of Statistics, Jeonbuk National University, Korea
2Samsung Research, Korea
3Department of Child and Adolescent Psychiatry, New York University School of Medicine, The Nathan S. Kline Institute for Psychiatric Research, USA
4Department of Statistics, Hankuk University of Foreign Studies, Korea

Abstract: Predicting the onset time of Alzheimer’s disease is of great importance in preventive medicine. Structural changes in brain regions have been actively investigated in the association study of Alzheimer’s disease diagnosis and prognosis. In this study, we propose a functional linear regression model to predict the conversion time to Alzheimer’s disease among mild cognitive impairment patients. Vertical thickness change in corpus callosum is measured from magnetic resonance imaging scan and put into the model as a functional covariate. A synthetic response approach is taken to deal with the censored data. The simulation studies demonstrate that the proposed model successfully predicts the unobserved true survival time but indicate that high censoring rate may cause poor prediction in time. Through ADNI data application, we find that the atrophy in the rear area of corpus callosum is a possible neuroimaging marker on Alzheimer’s disease prognosis.

Keywords: Alzheimer’s disease, Censored data, Functional regression, Reproducing kernel Hilbert space
Deconvolution estimation on hyperspheres
Jeong Min Jeon* and Ingrid Van Keilegom
ORSTAT, KU Leuven, Belgium

Abstract: This paper considers nonparametric estimation with contaminated data observed on the unit hypersphere $S^d$. For such data, we consider deconvolution density estimation and regression analysis. Our methodology and theory are based on harmonic analysis on $S^d$ which is largely unknown in statistics. We establish novel deconvolution density and regression estimators, and study their asymptotic properties including the rates of convergence and asymptotic distributions. We also provide asymptotic confidence intervals. We present practical details on implementation as well as the results of numerical studies.

Keywords: Hypersphere, Deconvolution, Measurement error, Nonparametric estimation

Confidence band for persistent homology of KDEs
Jisu Kim†, Jaehyeok Shin‡, Alessandro Rinaldo‡, Larry Wasserman‡ and Frédéric Chaza†
†DataShape, Inria, France
‡Statistics & Data Science, Carnegie Mellon University, USA

Abstract: The persistent homology of the upper level sets of a probability density function quantifies the salient topological features of data. Such a target quantity can be well estimated using the persistent homology of the upper level sets of a KDE (kernel density estimator). In this talk, I will present how the confidence band can be computed for determining the significance of the topological features in the persistent homology of KDEs, based on the bootstrap procedure. First, I will present how the confidence band can be computed for the persistent homology of KDEs computed on a grid. In practice, however, computing the persistent homology on a grid is infeasible when the dimension of the ambient space is high or topological features are in different scales. Hence, I will consider the persistent homology of KDEs on Vietoris-Rips complexes over the sample point. I will describe how to construct a valid confidence band for the persistent homology of KDEs on Vietoris-Rips complexes based on the bootstrap procedure.

Keywords: Persistent homology, Vietoris-Rips complex, Confidence band, Kernel density estimator

Analysis of chemical-gene bipartite network via a user-based collaborative filtering method incorporating chemical structure information
Namgil Lee*
Department of Information Statistics, Kangwon National University, Korea

Abstract: Drug repositioning refers to finding new applications and different uses of known drugs. In this study, we introduce a network analysis approach for drug repositioning. In particular, we introduce a user-based collaborative filtering method for analyzing bipartite networks between chemicals and genes. Moreover, under the assumption that structural similarity between chemicals is deeply related to functional similarity, an improved measure of similarity between chemicals is proposed. Numerical experiments are conducted to evaluate the statistical significance of the proposed method for the CTD database.

Keywords: Drug, Network, Collaborative filtering, Similarity
Interpretable, predictive spatio-temporal models via enhanced pairwise directions estimation
Sheng Li Tzeng*
Department of Applied Mathematics, National Sun Yat-sen University, Rep of China (Taiwan)

Abstract: Spatio-temporal phenomena are often complicated, but kriging methods are widely used in modeling such data, where only a very simple mean structure is assumed. We instead develop a novel approach based on supervised dimension reduction for such data in order to capture nonlinear mean structures without requiring a prespecified parametric model. In addition to prediction as a common interest, our approach focuses more on the exploration of geometric information in the data. The method of Pairwise Directions Estimation (PDE) is incorporated in our approach to implement the data-driven function searching of spatial structures and temporal patterns, useful in exploring data trends. We further enhance PDE, referring to it as PDE+, by using resolution adaptive fixed rank kriging to estimate the random effects not explained in the mean structures. Our proposal can not only produce more accurate and explainable prediction, but also increase the computation efficiency for model building. Illustrative applications to two real datasets are also presented. The results demonstrate that the proposed PDE+ method is very useful for exploring and interpreting the patterns of trend for spatio-temporal data.

Keywords: Dimension reduction, Kriging, Semi-parametric models, Visualization

Model selection with a nested spatial correlation structure
Chun-Shu Chen*
Graduate Institute of Statistics, National Central University, Rep of China (Taiwan)

Abstract: In spatial regression analysis, a suitable specification of the mean regression model is crucial for unbiased analysis. Suitably account for the underlying spatial correlation structure of the response variables is also an important issue. Here, we focus on selection of an appropriate mean model in spatial regression analysis under a general anisotropic nested spatial correlation structure. We propose a distribution-free model selection criterion which is an estimate of the weighted mean squared error based on assumptions only for the first two moments of the response data. The simulations under the settings of covariate selection reveal that the proposed criterion performs well for covariate selection in the mean model regardless of the underlying spatial correlation structure is nested/non-nested, isotropic/anisotropic. Also, the proposed criterion accommodates both continuous and count response data. Finally, a real data example regarding the fine particulate matter concentration is also analyzed for illustration.

Keywords: Anisotropy, Generalized method of moments, Information criterion, Variable selection

Consistent order selection for ARFIMA models
Kun Chen*
Southwestern University of Finance and Economics

Abstract: Estimating the orders of the autoregressive fractionally integrated moving average (ARFIMA) model has been a long-standing challenge in time series analysis. This paper tackles the challenge by establishing the consistency of the Bayesian information criterion (BIC) in the ARFIMA model with independent errors. Since we allow the model's memory parameter to be any unknown real number, our consistency result can apply simultaneously to
short-memory, long-memory, and non-stationary time series. We further extend BIC’s consistency to the ARFIMA model with conditional heteroskedastic errors, thereby broadening the criterion’s range of applications. Finally, the finite-sample implications of our theoretical results is illustrated using numerical examples.

**Whittle likelihood for irregularly spaced spatial data**
Soutir Bandyopadhyay*
Department of Applied Mathematics and Statistics, Colorado School of Mines, USA

**Abstract:** Under some regularity conditions, including that the process is Gaussian, the sampling region is rectangular, and that the parameter space \( \Theta \) is compact, Matsuda and Yajima (2009) showed that the Whittle estimator \( \hat{\theta}_n \) minimizing their version of Whittle likelihood is consistent (for \( d \leq 3 \)) and one can construct large sample confidence regions for covariance parameters \( \theta \) using the asymptotic normality of the Whittle estimator \( \hat{\theta}_n \). However, this requires one to estimate the asymptotic covariance matrix, which involves integrals of the spatial sampling density. Moreover, nonparametric estimation of the quantities in the asymptotic covariance matrix requires specification of a smoothing parameter and is subject to the curse of dimensionality. In comparison, we propose a spatial frequency domain empirical likelihood method (cf. Bandyopadhyay et al. (2015), Van Hala et al. (2017)) based approach which can be employed to produce asymptotically valid confidence regions and tests on \( \theta \), without requiring explicit estimation of such quantities.

**Keywords:** Whittle likelihood, Empirical Likelihood, Discrete Fourier transform, Spatial data

---

**OCS18 Advanced Learning Methods for Complex Data Analysis**
JULY 21 (WED), 11:30 - 12:00 KST

**Peel learning for pathway-related outcome prediction**
Rui Feng†, Yuantong Li‡, Mengying Yang§, Fei Wang¶ and Edward Cantu∥
†University of Pennsylvania, USA
‡Statistics, Purdue University, USA
§Biostatistics, Duke University, USA
¶Healthcare Policy and Research, Cornell University, USA
∥Surgery, University of Pennsylvania, USA

**Abstract:** Traditional regression models are limited in outcome prediction due to their parametric nature. Current deep learning methods allow for various effects and interactions and have shown improved performance, but they typically need to be trained on a large amount of data to obtain reliable results. Gene expression studies often have small sample sizes but high dimensional correlated predictors so that traditional deep learning methods are not readily applicable. In this talk, I present peel learning, a novel neural network that incorporates the prior relationship among genes. In each layer of learning, overall structure is peeled into multiple local substructures. Within the substructure, dependency among variables is reduced through linear projections. The overall structure is gradually simplified over layers and weight parameters are optimized through a revised backpropagation. We applied PL to a small lung transplantation study to predict recipients' postsurgery primary graft dysfunction using donors' gene expressions within several immunology pathways, where PL showed improved prediction accuracy compared to conventional penalized regression, classification trees, feed-forward neural network, and a neural network assuming prior network structure. Through simulation studies, we also demonstrated the advantage of adding specific structure among predictor variables in neural network, over no or uniform group structure, which is more favorable in smaller
Principal boundary for data on manifolds
Zhigang Yao*
Statistics and Applied Probability, National University of Singapore, Singapore

Abstract: We will discuss the problem of finding principal components to the multivariate datasets, that lie on an embedded nonlinear Riemannian manifold within the higher-dimensional space. Our aim is to extend the geometric interpretation of PCA, while being able to capture the non-geodesic form of variation in the data. We introduce the concept of a principal sub-manifold, a manifold passing through the center of the data, and at any point of the manifold, it moves in the direction of the highest curvature in the space spanned by the eigenvectors of the local tangent space PCA. We show the principal sub-manifold yields the usual principal components in Euclidean space. We illustrate how to find, use and interpret the principal sub-manifold, with which a classification boundary can be defined for data sets on manifolds.

Probabilistic semi-supervised learning via sparse graph structure learning
Li Wang*
University of Texas at Arlington

Abstract: We present a probabilistic semi-supervised learning (SSL) framework based on sparse graph structure learning. Different from existing SSL methods with either a predefined weighted graph heuristically constructed from the input data or a learned graph based on the locally linear embedding assumption, the proposed SSL model is capable of learning a sparse weighted graph from the unlabeled high-dimensional data and a small amount of labeled data, as well as dealing with the noise of the input data. Our representation of the weighted graph is indirectly derived from a unified model of density estimation and pairwise distance preservation in terms of various distance measurements, where latent embeddings are assumed to be random variables following an unknown density function to be learned and pairwise distances are then calculated as the expectations over the density for the model robustness to the data noise. Moreover, the labeled data based on the same distance representations is leveraged to guide the estimated density for better class separation and sparse graph structure learning. A simple inference approach for the embeddings of unlabeled data based on point estimation and kernel representation is presented. Extensive experiments on various data sets show the promising results in the setting of SSL compared with many existing methods, and significant improvements on small amounts of labeled data.

Bayesian modeling for paired data in genome-wide association studies with application to breast cancer
Min Chen* and Yashi Bu
Mathematical Sciences, University of Texas at Dallas, USA

Abstract: Genome-wide association studies (GWAS) has emerged as a useful tool to identify common genetic variants that are linked to complex diseases. Conventional GWAS are based on the case-control design where the individuals in cases and controls are independent. In cancer research, matched pair designs, which compare tumor tissues with
normal ones from the same subjects, are becoming increasingly popular. Such designs succeed in identifying somatic mutations in tumors while controlling both genetic and environmental factors. Somatic variation is one of the most important cancer risk factors that contribute to continuous monitoring and early detection of various cancers. However, most GWAS analysis methods, developed for unrelated samples in case-control studies, cannot be employed in the matched pair designs. A novel framework is proposed in this manuscript to accommodate for the particularity of matched-data in association studies of somatic mutation effects. In addition, we develop a Bayesian model to combine multiple markers to further improve the power of mapping genome regions to cancer risks.

**Keywords:** GWAS, Matched-pair design

---

**OCS19 Recent Advances in Complex Data Analysis**

**July 22 (Thu), 11:30 - 12:00 KST**

**Kernel density estimation and deconvolution under radial symmetry**

Kwan-Young Bak* and Ja-Yong Koo
Department of Statistics, Korea University, Korea

**Abstract:** This study illustrates a dimensionality reduction effect of radial symmetry in nonparametric density estimation. To deal with the class of radially symmetric functions, we adopt a generalized translation operation that preserves the symmetry structure. Radial kernel density estimators based on directly or indirectly observed random samples are proposed. For the latter case, we analyze deconvolution problems with four distinct scenarios depending on the symmetry assumptions on the signal and noise. Minimax upper and lower bounds are established for each scheme to investigate the role of the radial symmetry in determining optimal rates of convergence. The results confirm that the radial symmetry reduces the dimension of the estimation problems so that the optimal rate of convergence coincides with the univariate convergence rate except at the origin where a singularity occurs. The results also imply that the proposed estimators are rate optimal in the minimax sense for the Sobolev class of densities.

**Keywords:** Hankel transform, Kernel estimator, Minimax risk, Radial symmetry

**Penalized poly gram regression for bivariate smoothing**

Jae-Hwan Jhong*
Information of Statistics, Chungbuk National University, Korea

**Abstract:** We consider the problem of estimating a bivariate function over the plane using triangulation and penalization techniques. In order to provide a spatially adaptive method, total variation penalty for the bivariate spline function is considered to remove unnecessary common edges for the initial triangulation. A coordinate descent algorithm which we introduce can efficiently solve the convex optimization problem to handle the total variation penalty. The proposed estimator which is called Penalized Polygram Regression (PPR) has a form of piecewise linear and continuous within the adjacent polygons, not limited on triangles, and its corresponding basis functions can be obtained by the coordinate descent algorithm for eliminating common edges proceeds. Numerical studies using both simulated and real data examples are provided to illustrate the performance of the proposed method.

**Keywords:** Adaptive polygrams, Coordinate descent algorithm, Total variation, Triangulation
Penalized logistic regression using functional connectivity as covariates with an application to mild cognitive impairment

Jae-Hwan Jung¹, Seong-Jin Ji¹, Hongtu Zhu², Joseph Ibrahim², Yong Fan³ and Eunjee Lee¹*

¹Dept of Information and Statistics, Chungnam National University, Korea
²Department of Biostatistics, University of North Carolina at Chapel Hill, USA
³Department of Radiology, University of Pennsylvania, USA

Abstract: There is an emerging interest in brain functional connectivity (FC) based on functional Magnetic Resonance Imaging in Alzheimer's disease (AD) studies. The complex and high-dimensional structure of FC makes it challenging to explore the association between altered connectivity and AD susceptibility. We develop a pipeline to refine FC as proper covariates in a penalized logistic regression model and classify normal and AD susceptible groups. Three different quantification methods are proposed for FC refinement. One of the methods is dimension reduction based on common component analysis (CCA), which is employed to address the limitations of the other methods. We applied the proposed pipeline to the Alzheimer’s Disease Neuroimaging Initiative (ADNI) data and deduced pathogenic FC biomarkers associated with AD susceptibility. The refined FC biomarkers were related to brain regions for cognition, stimuli processing, and sensorimotor skills. We also demonstrated that a model using CCA performed better than others in terms of classification performance and goodness-of-fit.

Keywords: Penalized logistic regression, Common component analysis, fMRI, Alzheimer's disease

Resmax: detecting voice spoofing attacks with residual network and max filter map

Il-Youp Kwak*
Department of Applied Statistics, Chung-Ang University, Korea

Abstract: The 2019 automatic speaker verification spoofing and countermeasures challenge (ASVspoof) competition aims to facilitate the design of highly accurate voice spoofing attack detection systems. However, they do not emphasize model complexity and latency requirements such constraints are strict and integral in real-world deployment. Hence, most of the top performing solutions from the competition use an ensemble approach, and combine multiple complex deep learning models to maximize detection accuracy. This kind of approach would sit uneasily with real-world deployment constraints. To design a light weight system, we combine the notions of skip connection (from ResNet) and max filter map (from Light CNN), and evaluate its accuracy using the ASVspoof 2019 dataset by optimizing a well known signal processing feature called constant Q transform (CQT), our single model achieved a spoofing attack detection equal error rate (EER) of 0.16%, outperforming the top ensemble system from the competition that achieved an EER of 0.39%.

Keywords: Voice spoof detection, maMchine learning

Weighted validation of heteroscedastic regression models for better selection

Yoonsuh Jung* and Hayoung Kim
Department of Statistics, Korea University, Korea

Abstract: Statistical modeling can be divided into two processes: model fitting and model selection for the given task. For model fitting, it is vital to select the appropriate type of model to use. This step is taken first. For model selection, the model is fine-tuned via variable and parameter selection. Improving model selection in the presence of heteroscedasticity is the main goal of this talk. Model selection is usually conducted by measuring the prediction error. When there is heteroscedasticity in the data, observations with high variation tend to produce larger prediction errors. In turn, model selection is strongly effected by observations with large variation. To reduce the effect of
heteroscedastic data, we propose weighted selection during the model selection process. The proposed method reduces the impact of large prediction errors via weighted prediction and leads to better model and parameter selection. The benefits of the proposed method is demonstrated in simulations and with two real data sets. 

**Keywords:** Cross-validation, Heteroscedasticity, Model Assessment, Model Selection

---

**OCS20 Theories and Applications for Complex Data Analysis**
JULY 21 (WED), 21:30 - 22:00 KST

---

**Partly interval-censored rank regression**

Sangbum Choi* and Taehwa Choi  
Statistics, Korea University, Korea

**Abstract:** This paper studies estimation of the semiparametric accelerated failure time model for double and partly interval-censored data. Gehan-type weighted estimating function is constructed by contrasting comparable rank cases under interval-censoring. An extension to the general class of log-rank estimating functions can also be investigated, along with an efficient variance estimation procedure. Asymptotic behaviors of the proposed estimator are established under mild conditions by using empirical processes theory. Simulation studies demonstrate our method works very well with practical size of samples. Two data examples are given to illustrate the practical usefulness of our method.  

**Keywords:** Interval censoring, Linear model, Rank regression, Survival analysis

---

**Two-sample testing of high-dimensional linear regression coefficients via complementary sketching**

Tengyao Wang*  
University College London

**Abstract:** We introduce a new method for two-sample testing of high-dimensional linear regression coefficients without assuming that those coefficients are individually estimable. The procedure works by first projecting the matrices of covariates and response vectors along directions that are complementary in sign in a subset of the coordinates, a process which we call ‘complementary sketching’. The resulting projected covariates and responses are aggregated to form two test statistics, which are shown to have essentially optimal asymptotic power under a Gaussian design when the difference between the two regression coefficients is sparse and dense respectively. Simulations confirm that our methods perform well in a broad class of settings.  

**Keywords:** Two sample testing, High dimensional

---

**Optimal rates for independence testing via U-statistic permutation tests**

Thomas Berrett*  
Department of Statistics, University of Warwick, UK

**Abstract:** Independence testing is one of the most well-studied problems in statistics, and the use of procedures such as the chi-squared test is ubiquitous in the sciences. While tests have traditionally been calibrated through asymptotic theory, permutation tests are experiencing a growth in popularity due to their simplicity and exact Type I error control. In this talk I will present new, finite-sample results on the power of a new class of permutation tests, which show that their power is optimal in many interesting settings, including those with discrete, continuous, and functional data. A
simulation study shows that our test for discrete data can significantly outperform the chi-squared for natural data-generating distributions. Defining a natural measure of dependence \( D(f) \) to be the squared \( L^2 \)-distance between a joint density \( f \) and the product of its marginals, we first show that there is generally no valid test of independence that is uniformly consistent against alternatives of the form \( \{ f: D(f) \geq \rho^2 \} \). Motivated by this observation, we restrict attention to alternatives that satisfy additional Sobolev-type smoothness constraints, and consider as a test statistic a U-statistic estimator of \( D(f) \). Using novel techniques for studying the behaviour of U-statistics calculated on permuted data sets, we prove that our tests can be minimax optimal. Finally, based on new normal approximations in the Wasserstein distance for such permuted statistics, we also provide an approximation to the power function of our permutation test in a canonical example, which offers several additional insights.

[This is joint work with Ioannis Kontoyiannis and Richard Samworth.]

Keywords: Independence testing, Permutation test, U-statistics, Minimax separation rates

---

**Empirical Bayes PCA in high dimensions**

Zhou Fan*

Yale University, Department of Statistics and Data Science, USA

**Abstract:** When the dimension of data is comparable to or larger than the number of data samples, Principal Components Analysis (PCA) may exhibit problematic high-dimensional noise. In this work, we propose an Empirical Bayes PCA method that reduces this noise by estimating a joint prior distribution for the principal components. EB-PCA is based on the classical Kiefer-Wolfowitz nonparametric MLE for empirical Bayes estimation, distributional results derived from random matrix theory for the sample PCs, and iterative refinement using an Approximate Message Passing (AMP) algorithm. In theoretical “spiked” models, EB-PCA achieves Bayes-optimal estimation accuracy in the same settings as an oracle Bayes AMP procedure that knows the true priors. Empirically, EB-PCA significantly improves over PCA when there is strong prior structure, both in simulation and on quantitative benchmarks constructed from the 1000 Genomes Project and the International HapMap Project. An illustration is presented for analysis of gene expression data obtained by single-cell RNA-seq.

Keywords: Empirical Bayesian inference, Random matrix theory, Approximate message passing, Principal components analysis

---

**Identifiability of additive noise models using conditional variances**

Gunwoong Park*

Department of Statistics, University of Seoul, Korea

**Abstract:** The identifiability assumption of structural equation models (SEMs) is considered in which each variable is determined by an arbitrary function of its parents plus an independent error. It has been shown that linear Gaussian structural equation models are fully identifiable if all error variances are the same or known. Hence, we prove the identifiability of SEMs with both homogeneous and heterogeneous unknown error variances. Our new identifiability assumption exploits not only error variances, but edge weights; hence, it is strictly milder than prior work on the identifiability result. We further provide a statistically consistent and computationally feasible learning algorithm. We verify through simulations that the proposed algorithm is statistically consistent and computationally feasible in the high-dimensional settings, and performs well compared to state-of-the-art US, GDS, LISTEN, PC, and GES algorithms. We also demonstrate through real human cell signalling and mathematics exam data that our algorithm is well-suited to estimating DAG models for multivariate data in comparison to other methods used for continuous data.
Keywords: Graphical models, Causal inference, Bayesian network, Identifiability

Multivariate functional group sparse regression: functional predictor selection
Jun Song*
Mathematics and Statistics, UNC Charlotte, USA

Abstract: In this talk, I will present a method for functional predictor selection and the estimation of smooth functional coefficients simultaneously in a scalar-on-function regression problem under a high-dimensional multivariate functional data setting. In particular, we develop two methods for functional group-sparse regression under a generic Hilbert space of infinite dimension. Then we show the convergence of algorithms and the consistency of the estimation and selection under infinite-dimensional Hilbert spaces. Simulation and fMRI data application will be presented at the end to show the effectiveness of the methods in both the selection and estimation of functional coefficients.
Keywords: Functional predictor selection, Multivariate functional data, Oracle property, Group lasso

Causal foundations for fair and responsible machine learning
Joshua Loftus*
Statistics, London School of Economics, UK

Abstract: Recently the social impacts of new data and information technologies started receiving more attention from scholars and the public. Many are concerned that algorithmic decision systems using machine learning or "artificial intelligence" may affect people negatively, especially in ways that reinforce harmful patterns in historic data related to attributes like race, gender, and other ethically or legally important attributes. This talk will briefly survey recent work in the field and then focus on causal modeling as a pathway beyond impossibility results and toward consensus.
Keywords: Machine learning, Algorithmic fairness, Causal inference

Network change point detection
Yi Yu*
Statistics, University of Warwick, UK

Abstract: This talk will be on three different projects, parametric network change point detection, nonparametric network change point detection and online network change point detection. I will provide theoretically-justified and computationally-efficient change point estimators in these three different scenarios, and discuss the theoretical difficulties in all settings.

A Spline-based modeling approach for time-indexed multilevel data
Eun Ryung Lee*
Statistics of Department, Sungkyunkwan University, Korea
Abstract: This paper introduces a spline-based multilevel approach for analyzing a time-indexed data collected from an automated platform. The proposed method is computationally efficient and easy to implement, and useful for analyzing data that have hierarchical structure with varying complexity at different levels. An estimation procedure is developed combining Expectation-Maximization algorithm and nonparametric regression approach. The theoretical properties of the proposed methods are derived. The proposed estimator is shown to belong to a well-known linear smoother class. Thus, further statistical inference can be easily adopted from the existing literature on linear smoothers. An R package for the proposed methodology is provided in an open repository. The effectiveness of the approach is illustrated using music concert event data collected in the United States for several years.

Keywords: Expectation-maximization, Hurdle model, Nonparametric regression, Multi-level data

Impulse response analysis for sparse high-dimensional time series
Carsten Trenkler*, Jonas Krampe1 and Efstathios Paparoditis2
1Economics, University of Mannheim, Germany
2Department of Mathematics and Statistics, University of Cyprus, Cyprus

Abstract: We consider structural impulse response analysis for sparse high-dimensional vector autoregressive (VAR) systems. First, we present a consistent estimation approach in the high-dimensional setting. Second, we suggest a valid inference procedure. Inference is more involved since standard procedures, like the delta-method, do not lead to valid inference in our set-up. Therefore, by using the local projection equations, we first construct a de-sparsified version of regularized estimators of the moving average parameters that are associated with the VAR system. In order to obtain estimators of the structural impulse responses we combine these de-sparsified estimators with a non-regularized estimator of the contemporaneous impact matrix in such a way that the high-dimension is taken into account. We show that the estimators of the structural impulse responses have a Gaussian limiting distribution. Moreover, we also present a valid bootstrap procedure. Applications of the inference procedure are confidence intervals of the impulse responses as well as tests for forecast error variance decompositions that are often used to construct connectedness measures. Our procedure is illustrated by means of simulations.

Keywords: Bootstrap, Vector Autoregression, Sparse Models, Inference

Revealing cluster structures based on mixed sampling frequencies: with an application to the state-level labor market
Yeonwoo Rho1*, Yun Liu2 and Hie Joo Ahn3
1Mathematical Sciences, Michigan Technological University, USA
2Quicken Loans, Quicken Loans, USA
3Federal Reserve Board, Federal Reserve Board, USA

Abstract: This paper proposes a new linearized mixed data sampling (MIDAS) model and develops a framework to infer clusters in a panel regression with mixed frequency data. The linearized MIDAS estimation method is more flexible and substantially simpler to implement than competing approaches. We show that the proposed clustering algorithm successfully recovers true membership in the cross-section, both in theory and in simulations, without requiring prior knowledge of the number of clusters. This methodology is applied to a mixed-frequency Okun’s law model for state-level data in the U.S. and uncovers four meaningful clusters based on the dynamic features of state-level labor markets.

Keywords: Mixed data sampling regression mode, Panel data, Clustering, Forecasting
Semiparametric efficient estimators in heteroscedastic error models
Mijeong Kim*
Dept. of Statistics, Ewha Womans University, Korea

Abstract: In the mean regression context, this study considers several frequently encountered heteroscedastic error models where the regression mean and variance functions are specified up to certain parameters. An important point we note through a series of analyses is that different assumptions on standardized regression errors yield quite different efficiency bounds for the corresponding estimators. Consequently, all aspects of the assumptions need to be specifically taken into account in constructing their corresponding efficient estimators. This study clarifies the relation between the regression error assumptions and their, respectively, efficiency bounds under the general regression framework with heteroscedastic errors. Our simulation results support our findings; we carry out a real data analysis using the proposed methods where the Cobb–Douglas cost model is the regression mean.

Keywords: Heteroscedasticity, Semiparametric method, Standardized regression error, Variance function

Multivariate responses quantile regression for regional quantiles with applications to CCLE data
Seyoung Park*
Department of Statistics, Sungkyunkwan University, Korea

Abstract: Cancer Cell Line Encyclopedia(CCLE) is a large-scale project that have generated resources with cancer cell lines characterized by high-dimensional molecular profiles along with pharmacological profiles. In CCLE, identifying gene-drug interaction is important to elucidate mechanisms of drug actions. Considering interrelations between the pharmacological responses, multivariate responses regression can be applied to identify meaningful gene-drug interaction. Quantile regression, as an alternative to classical linear regression, may better reveal the relation-ship between molecular profiles and pharmacological responses because quantile regression permits investigation of heterogeneity across quantiles. In this study, we propose a new multivariate responses quantile regression framework considering an interval of quantile levels. We aim to select relevant variables to any τ-th conditional quantiles for multiple responses, where Δ is an interval of quantile levels of interest. We propose a penalized composite quantile regression framework with double group Lasso penalty to estimate the quantile coefficient function. In theory, we show the oracle property of the proposed estimator, in combination with a novel information criterion with theoretical guarantees. Numerical examples and applications to CCLE data demonstrate the effectiveness of the proposed method.

Keywords: Multivariate regression, Regional quantile, Variable selection, Adaptive Lasso

On sufficient graphical models
Kyongwon Kim*
Department of Statistics, Ewha Womans University, Korea

Abstract: We introduce a Sufficient Graphical Model by applying the recently developed nonlinear sufficient dimension reduction techniques to the evaluation of conditional independence. The graphical model is nonparametric in nature, as it does not make distributional assumptions such as the Gaussian or copula Gaussian assumptions. However, unlike a fully nonparametric graphical model, which relies on the high-dimensional kernel to
characterize conditional independence, our graphical model is based on conditional independence given a set of sufficient predictors with a substantially reduced dimension. In this way, we avoid the curse of dimensionality that comes with a high-dimensional kernel. We develop the population-level properties, convergence rate, and variable selection consistency of our estimate. By simulation comparisons and an analysis of the DREAM 4 Challenge data set, we demonstrate that our method outperforms the existing methods when the Gaussian or copula Gaussian assumptions are violated, and its performance remains excellent in the high-dimensional setting.

**Keywords:** Conjoined Conditional Covariance Operator, Reproducing Kernel Hilbert Space, Generalized Sliced Inverse Regression, Nonlinear Sufficient Dimension Reduction

**Principal component analysis in the wavelet domain**

Yaeji Lim*
Applied Statistics, Chung Ang University, Korea

**Abstract:** In this paper, we propose a new method of principal component analysis in the wavelet domain that is useful for dimension reduction of multiple non-stationary time series and for identifying important features. The proposed method is constructed by a novel combination of eigen analysis and the local wavelet spectrum defined in the locally stationary wavelet process. So, it can be expected that the proposed method reflects a more generalized non-stationary time series beyond some limited types of signals that existing methods have performed. We investigate the theoretical results of estimated principal components and their loadings. Results from numerical examples, including analysis of real seismic data and financial data, show the promising empirical properties of the proposed approach.

**Bayesian inference of evolutionary models from genomic data**

Yujin Chung*
Applied statistics, Kyonggi University, Korea

**Abstract:** The evolutionary history of a group of organisms explains the process of their genetic variation over time. Due to recent sequencing and computing advances, statistical inference has become an essential discipline in the study of evolutionary history from genomic data. However, typical analyses are either limited to a small amount of data or fail to estimate complex and diverse evolutionary models. In this talk, I will present a Bayesian method for estimating population/species-level history, including population sizes, splitting time of two populations, and migration rates. The method resolves statistical limitations and overcomes major roadblocks to analyze genome-scale data. Using importance sampling and a Markov chain representation of genealogy, the method scales to genomic data without mixing difficulty in a Markov chain Monte Carlo simulation. The method also provides for the calculation of the joint posterior density for all model parameters, thus resolving the problem of high false positive rates that arises for the likelihood ratio tests for migration rates using other existing Bayesian approaches. I will demonstrate the method with simulated data and real DNA sequences.

**Keywords:** MCMC, Bayesian, Evolution, IM model
On the verifiable identification condition in NMAR missing data analysis
Kosuke Morikawa* and Kenji Beppu
Graduate School of Engineering Science, Osaka University, Japan

Abstract: Efficient estimation for regression parameters in longitudinal data analysis requires the correct covariance structure. When the data involves some missing values, specification of a missing-data mechanism and a regression model is required as well as the covariance structure. There have been proposed a variety of double robust estimators with working models for the missing-data mechanisms and the regression models. However, the estimators have double robustness if either of the two types of working models are correct at all the time points, which is unrealistic when the observed time points are large. Also, no methods exist to obtain a consistent estimator of the working models even if they are correctly specified, so that it is impossible to obtain a real double robust estimator so far. In this talk, we give a consistent estimator of the working models, and propose a real double robust estimator. In addition, we show that our proposed estimator has the double robustness if either of the two types of working models are correctly specified at each time point, which is a much weaker condition. The estimator is also extended to a multiple robust one.

Keywords: Missing at random, Dropout, Generalized estimating equations

Bayesian hierarchical spatial model for small-area estimation with non-ignorable nonresponses and its application to the NHANES dental caries data
Ick Hoon Jin*
Statistics and Data Science, Yonsei University, Korea

Abstract: The National Health and Nutrition Examination Survey (NHANES) is a major program of the National Center for Health Statistics, designed to assess the health and nutritional status of adults and children in the United States. The analysis of NHANES dental caries data faces several challenges, including (1) the data were collected using a complex, multistage, stratified, unequal-probability sampling design; (2) the sample size of some primary sampling units (PSU), e.g., counties, is very small; (3) the measures of dental caries have complicated structure and correlation, and (4) there is a substantial percentage of nonresponses, which are expected not to be missing at random or non-ignorable. We propose a Bayesian hierarchical spatial model to address these analysis challenges. We develop a two-level Potts model that closely resembles the caries evolution process, and captures complicated spatial correlations between teeth and surfaces of the teeth. By adding Bayesian hierarchies to the Potts model, we account for the multistage survey sampling design, while also enabling information borrowing across PSUs for small-area estimation. We incorporate sampling weights by including them as a covariate in the model and adopt flexible B-splines to achieve robust inference. We account for non-ignorable missing outcomes and covariates using the selection model. We use data augmentation coupled with the noisy Monte Carlo algorithm to overcome the numerical difficulty caused by doubly-intractable normalizing constants and sample posteriors. Our analysis results show strong spatial associations between teeth and tooth surfaces, including that dental hygienic factors, such as fluorosis and sealant, reduce dental disease risks.

Keywords: Small Area Estimation
Raking-based relabeling classification method for highly imbalanced data
Seunghwan Park1* and Jongho Im2
1Department of Information Statistics, Kangwon National University, Korea
2Department of Statistics and data science, Yonsei University, Korea

Abstract: We consider binary classification on the imbalanced data. A dataset is called imbalanced if the proportion of classes are heavily skewed. Classification on the imbalanced data is often challengeable, especially for high dimensional data, in the sense that the unequal classes deteriorate the performance of classifiers. Undersampling the majority class and/or oversampling the minority class are popular methods to construct balanced samples and then it helps to improve the classification performance. However, many existing sampling methods cannot be easily extended to high dimensional data and mixed variables, because they often require to approximate the distribution of attributes and this becomes another critical issue rather than classification. In this paper, we propose new sampling strategy, called counter-matching sampling, such that the attribute values of the major class are imputed for the values of the minor class in construction of balanced samples. The proposed algorithms produce the same or similar performance with the existing methods but is more flexible to data shape and size of attributes. Our sampling algorithm is very attractive in practice considering that our sampling algorithm does not require any density estimation for synthetic data generation in oversampling and is not bothered from mixed variables. Also, the proposed sampling strategy can be easily combined with many existing classifiers.

Keywords: Imbalanced data, Calibration, Oversampling, Undersampling

Imputation approach for outcome dependent sampling design
Jongho Im1*, Wanseok Choi2 and BaekJun Jang1
1Statistics and Data Science, Yonsei University, Korea
2Biostatistics and Data Sciences, Boehringer Ingelheim Pharmaceuticals, USA

Abstract: Outcome dependent sampling (ODS) has been widely used to enhance the study efficiency in epidemiology or biomedical studies. We consider a biased two phase sampling design where the second phase samples are selected based on the outcome variable and the covariate x are only observed at the second phase. Many methods have been proposed by incorporating the estimated inclusion probabilities into the target score function of outcome model. In this paper, we propose an imputation method that is essentially implemented by data augmentation. The predictive distribution is nonparametrically estimated and then a Bayesian bootstrap method is used to generate imputed values. The proposed method employs Rubin's variance formula for variance estimation of imputation estimators. A limited simulation study shows that the proposed method performs well and is comparable to the previous methods.

Keywords: Approximate Bayesian bootstrap, Data Augmentation, Missing covariate, Informative sampling

Bayesian nonparametric adjustment of confounding
Chanmin Kim*
Department of Statistics, SungKyunKwan University, Korea

Abstract: In observational studies, confounder selection is a crucial task in estimation of causal effect of an exposure. Wang et al. (2012, 2015) propose Bayesian adjustment methods for confounding (BAC) to account for the uncertainty
in confounder selection by jointly fitting parametric models for exposure and outcome, in which Bayesian model averaging (BMA) is utilized to obtain the causal effect averaged across all potential models according to their posterior weights. In this work, we propose a Bayesian nonparametric approach to select confounders and estimate causal effects without assuming any model structures for exposure and outcome. With the Bayesian additive regression trees (BART) method, the causal model can capture complex data structure flexibly and select a subset of true confounders by specifying a common prior on the selection probabilities in both exposure and outcome models. The proposed model does not require a separate BMA process to average effects across many models as, in our method, selection of confounders and estimation of causal effects based on the selected confounders are processed simultaneously within each MCMC iteration. A set of extensive simulation studies demonstrates that the proposed method outperforms in a variety of situations.

**Keywords:** Causal Inference, BART, Confounder selection, BMA

---

**Multivariate point process models for microbiome image analysis**

Kyu Ha Lee*

Department of Nutrition, Department of Epidemiology, Department of Biostatistics, Harvard TH Chan School of Public Health, USA

**Abstract:** We investigate the spatial distribution of microbes to understand the role of biofilms in human and environmental health. Advances in spectral imaging technologies enable us to display how different taxa (e.g. species or genera) are located relative to one another and to host cells. However, most commonly used quantitative methods are limited to describing spatial patterns of bivariate data. Therefore, we propose a flexible multivariate spatial point process model that can quantify spatial relationships among the multiple taxa observable in biofilm images. We have developed an efficient computational scheme based on the Hamiltonian Monte Carlo algorithm, implemented in the R package. We applied the proposed model to tongue biofilm image data.

---

**Look before you leap: systematic evaluation of tree-based statistical methods in subgroup identification**

Xiaojing Wang*

Statistics, University of Connecticut, USA

**Abstract:** Subgroup analysis, as the key component of personalized medicine development, has attracted a lot of interest in recent years. While a number of exploratory subgroup searching approaches have been proposed, informative evaluation criteria and scenario-based systematic comparison of these methods are still underdeveloped topics. In this article, we propose two evaluation criteria in connection with traditional type I error and power concepts, and another criterion to directly assess recovery performance of the underlying treatment effect structure. Extensive simulation studies are carried out to investigate empirical performance of a variety of tree-based exploratory subgroup methods under the proposed criteria. A real data application is also included to illustrate the necessity and importance of method evaluation.

**Keywords:** Virtual twins
Scalable Bayesian high-dimensional local dependence learning

Kyoungjae Lee\textsuperscript{1*} and Lizhen Lin\textsuperscript{2}
\textsuperscript{1}Inha University, Korea\textsuperscript{2}Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, USA

Abstract: In this work, we propose a scalable Bayesian procedure for learning the local dependence structure in a high-dimensional model where the variables possess a natural ordering. The ordering of variables can be indexed by time, the vicinities of spatial locations, and so on, with the natural assumption that variables far apart tend to have weak correlations. Applications of such models abound in a variety of fields such as finance, genome associations analysis and spatial modeling. We adopt a flexible framework under which each variable is dependent on its neighbors or predecessors, and the neighborhood size can vary for each variable. It is of great interest to reveal this local dependence structure by estimating the covariance or precision matrix while yielding a consistent estimate of the varying neighborhood size for each variable. The existing literature on banded covariance matrix estimation, which assumes a fixed bandwidth cannot be adapted for this general setup. We employ the modified Cholesky decomposition for the precision matrix and design a flexible prior for this model through appropriate priors on the neighborhood sizes and Cholesky factors. The posterior contraction rates of the Cholesky factor are derived which are nearly or exactly minimax optimal, and our procedure leads to consistent estimates of the neighborhood size for all the variables. Another appealing feature of our procedure is its scalability to models with large numbers of variables due to efficient posterior inference without resorting to MCMC algorithms. Numerical comparisons are carried out with competitive methods, and applications are considered for some real datasets.

Keywords: Selection consistency, Optimal posterior convergence rate, Varying bandwidth

Fast and flexible estimation of effective migration surfaces

Wooseok Ha\textsuperscript{1*}, Joseph Marcus\textsuperscript{2}, Rina Barber\textsuperscript{3} and John Novembre\textsuperscript{4}
\textsuperscript{1}Department of Statistics, UC Berkeley, USA\textsuperscript{2}Bioinformatics and Data Science, GRAIL, USA\textsuperscript{3}Department of Statistics, University of Chicago, USA\textsuperscript{4}Department of Human Genetics, Department of Ecology and Evolution, University of Chicago, USA

Abstract: An important feature in spatial population genetic data is often “isolation-by-distance,” where genetic differentiation tends to increase as individuals become more geographically distant. Recently, Petkova et al. (2016) developed a statistical method called Estimating Effective Migration Surfaces (EEMS) for visualizing spatially heterogeneous isolation-by-distance on a geographic map. While EEMS is a powerful tool for depicting spatial population structure, it can suffer from slow runtimes. Here we develop a related method called Fast Estimation of Effective Migration Surfaces (FEEMS). FEEMS uses a Gaussian Markov Random Field in a penalized likelihood framework that allows for efficient optimization and output of effective migration surfaces. Further, the efficient optimization facilitates the inference of migration parameters per edge in the graph, rather than per node (as in EEMS). When tested with coalescent simulations, FEEMS accurately recovers effective migration surfaces with complex gene-flow histories, including those with anisotropy. Applications of FEEMS to population genetic data from North American gray wolves show it to perform comparably to EEMS, but with solutions obtained orders of magnitude faster. Overall, FEEMS expands the ability of users to quickly visualize and interpret spatial structure in their data.
Statistical inference for cluster trees
Jisu Kim1*, Yen-Chi Chen2, Sivaraman Balakrishnan3, Alessandro Rinaldo3 and Larry Wasserman3
1DataShape, Inria, France
2Statistics, University of Washington, USA
3Statistics & Data Science, Carnegie Mellon University, USA

Abstract: A cluster tree provides a highly-interpretable summary of a density function by representing the hierarchy of its high-density clusters. It is estimated using the empirical tree, which is the cluster tree constructed from a density estimator. This talk addresses how to quantify the uncertainty by assessing the statistical significance of topological features of an empirical cluster tree. To do this, I propose methods to construct and summarize confidence sets for the unknown true cluster tree. And I introduce how to prune some of the statistically insignificant features of the empirical tree, yielding interpretable and parsimonious cluster trees. Finally, I illustrate the proposed methods on a variety of examples data set.

Keywords: Cluster tree, Confidence set, Pruning, Kernel density estimator

Autologistic network model on binary data for disease progression study
Yei Eun Shin1*, Huiyan Sang2, Dawei Liu3, Toby Ferguson4 and Peter Song5
1Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, USA
2Department of Statistics, Texas A&M University, USA
3Global Analytics and Data Sciences, Biogen, USA
4Neurology Research and Early Clinical Development, Biogen, USA
5Department of Biostatistics, University of Michigan, USA

Abstract: We propose an autologistic network model on binary spatiotemporal data to study the spreading patterns of disease. The proposed model identifies an underlying network, without the pre-specification of neighborhoods based on proximity, that can have varying effects depending on the previous states. The model parameters are estimated by maximizing the penalized pseudolikelihood with bias-corrected, which can be adapted to the generalized linear model (GLM) framework, where we show the resulting estimators are asymptotically normal. We provide spatial-joint transition probabilities for predicting disease status in the next time interval. Simulation studies were conducted to evaluate the validity and performance of the proposed method. Examples are provided using the amyotrophic lateral sclerosis (ALS) patients’ data from EMPOWER Study.

Keywords: Amyotrophic lateral sclerosis disease, Bias-corrected LASSO, Network, Spatiotemporal modeling

Nonparametric Bayesian latent factor model for multivariate functional data with covariate dependency
Yeonseung Chung*
Department of Mathematical Sciences, Korea Advanced Institute of Science and Technology, Korea

Abstract: Nowadays, multivariate functional data are frequently encountered in many fields of science. While there exist a variety of methodologies for univariate functional clustering, the approach for multivariate functional clustering are less studied. Moreover, there is little research for the functional clustering methods incorporating additional covariate information. In this paper, we propose a Bayesian nonparametric sparse latent factor model for
covariate-dependent multivariate functional clustering. Multiple functional curves are represented by basis coefficients for splines, which are reduced to latent factors. Then, the factors and covariates are jointly modeled using a Dirichlet process (DP) mixture of Gaussians to facilitate a model-based covariate dependent multivariate functional clustering. The method is further extended to dynamic multivariate functional clustering to handle sequential multivariate functional data. The proposed methods are illustrated through a simulation study and applications to Canadian weather and air pollution data.

Keywords: Multivariate functional data, Model based clustering, Latent factors, Dirichlet process

Bayesian model selection for ultrahigh-dimensional doubly-intractable distributions
Jaewoo Park*
Department of Applied Statistics, Yonsei University, Korea

Abstract: Doubly intractable distributions commonly arise in many complex statistical models in physics, epidemiology, ecology, social science, among other disciplines. With an increasing number of model parameters, they often result in ultrahigh-dimensional posterior distributions; this is a challenging problem and is crucial for developing the computationally feasible approach. A particularly important application of ultrahigh-dimensional doubly intractable models is network psychometrics, which gets attention in item response analysis. However, its parameter estimation method, maximum pseudo-likelihood estimator (MPLE) combining with lasso certainly ignores the dependent structure, so that it is inaccurate. To tackle this problem, we propose a novel Markov chain Monte Carlo methods by using Bayesian variable selection methods to identify strong interactions automatically. With our new algorithm, we address some inferential and computational challenges: (1) likelihood functions involve doubly-intractable normalizing functions, and (2) increasing number of items can lead to ultrahigh dimensionality in the model. We illustrate the application of our approaches to challenging simulated and real item response data examples for which studying local dependence is very difficult. The proposed algorithm shows significant inferential gains over existing methods in the presence of strong dependence among items.

Keywords: Bayesian variable selection, Exponential random graph model, Network psychometrics, Doubly-intractable distributions

Post-processed posteriors for banded covariances
Kwangmin Lee*
Statistics, Seoul National University, Korea

Abstract: We consider Bayesian inference of banded covariance matrices and propose a post-processed posterior. The post-processing of the posterior consists of two steps. In the first step, posterior samples are obtained from the conjugate inverse-Wishart posterior which does not satisfy any structural restrictions. In the second step, the posterior samples are transformed to satisfy the structural restriction through a post-processing function. The conceptually straightforward procedure of the post-processed posterior makes its computation efficient and can render interval estimators of functionals of covariance matrices. We show that it has nearly optimal minimax rates for banded covariances among all possible pairs of priors and post-processing functions. Furthermore, we prove that, the expected coverage probability of the $(1 - \alpha)100\%$ highest posterior density region of the post-processed posterior is asymptotically $1 - \alpha$ with respect to a conventional posterior distribution. It implies that the highest posterior density region of the post-processed posterior is, on average, a credible set of a conventional posterior. The advantages of the post-processed posterior are demonstrated by a simulation study and a real data analysis.

Keywords: Bayesian, Covariance estimation, Minimax analysis
Adaptive Bayesian inference for current status data on a grid
Minwoo Chae*
Pohang University of Science and Technology

Abstract: We use the class of Wasserstein metrics to study asymptotic properties of posterior distributions. Our first goal is to provide sufficient conditions for posterior consistency. In addition to the well-known Schwartz's Kullback-Leibler condition on the prior, the true distribution and most probability measures in the support of the prior are required to possess moments up to an order which is determined by the order of the Wasserstein metric. We further investigate convergence rates of the posterior distributions for which we need stronger moment conditions. The required tail conditions are sharp in the sense that the posterior distribution may be inconsistent or contract slowly to the true distribution without these conditions. Our study involves techniques that build on recent advances on Wasserstein convergence of empirical measures. We apply the results to density estimation with a Dirichlet process mixture prior and conduct a simulation study for further illustration.

Keywords: Dirichlet process mixture, nonparametric Bayesian inference, posterior convergence rate, Wasserstein metrics

Resampling long-range dependent time series
Shuyang Bai*
Department of Statistics, University of Georgia, USA

Abstract: For time series exhibiting long-range dependence, inference through resampling is of particular interest since the asymptotic distributions are often difficult to determine statistically. On the other hand, due to the strong dependence and the non-standard scaling, designing versatile resampling strategies and establishing their validity is challenging. We shall introduce some progress on this direction.

Keywords: Long-range dependence, Resampling, Time Series

Robust test for structural instability in dynamic factor models
Changryong Baek1*, Byungsoo Kim2 and Junmo Song3
1Statistics, Sungkyunkwan University, Korea
2Statistics, Yeungnam University, Korea
3Statistics, Kyungpook National University, Korea

Abstract: In this paper, we consider a robust test for structural breaks in dynamic factor models. Our framework considers structural changes when the underlying high dimensional time series is contaminated by some outlying observations, which is typically observed in many real applications such as fMRI, economics and finance. We propose a test based on the robust estimation of vector autoregressive model for principal component factors using minimum density power divergence estimator. Simulations study shows excellent finite sample performance, higher powers while achieving good sizes in all cases considered. Our method is illustrated to resting state fMRI series to detect brain connectivity changes. It shows that brain connectivity indeed changes even in the resting state and this is not an artifact of outlier effects.

Keywords: High-dimensional time series, Dynamic factor models, Minimum density power divergence
**On scaling in high dimensions**
Gustavo Didier*
Mathematics, Tulane University, USA

**Abstract:** Scaling relationships have been found in a wide range of phenomena that includes coastal landscapes, hydrodynamic turbulence, the metabolic rates of animals and Internet traffic. For scale invariant systems, also called fractals, a continuum of time scales contributes to the observed dynamics, and the analyst's focus is on identifying mechanisms that relate the scales, often in the form of exponents. In this talk, we will look into the little explored topic of scale invariance in high dimensions, which is especially important in the modern era of "Big Data". We will discuss the role played by wavelets in the analysis of self-similar stochastic processes and visit recent contributions to the wavelet modeling of high- and multidimensional scaling systems.

[This is joint work with P. Abry (CNRS and ENS-Lyon), B.C. Boniece (Washington University in St Louis) and H. Wendt (CNRS and Université de Toulouse).]

**Keywords:** High dimensions, Scaling, Wavelets

---

**Thresholding and graphical local Whittle estimation**
Marie Düker¹*, Changryong Baek² and Vladas Pipiras¹
¹Department of Statistics and Data Science, Cornell University, USA
²Department of Statistics, Sungkyunkwan University, Korea
³Department of Statistics and Operations Research, UNC at Chapel Hill, USA

**Abstract:** The long-run variance matrix and its inverse, the so-called precision matrix, give, respectively, information about correlations and partial correlations between dependent component series of multivariate time series around zero frequency. This talk will present non-asymptotic theory for estimation of the long-run variance and precision matrices for high-dimensional Gaussian time series under general assumptions on the dependence structure including long-range dependence. The presented results for thresholding and penalizing versions of the classical local Whittle estimator ensure consistent estimation in a possibly high-dimensional regime. The key technical result is a concentration inequality of the local Whittle estimator for the long-run variance matrix around the true model parameters. In particular, it handles simultaneously the estimation of the memory parameters which enter the underlying model.

---

**Cotrending: testing for common deterministic trends in varying means model**
Vladas Pipiras*
University of North Carolina at Chapel Hill

**Abstract:** In a varying means model, the temporary evolution of a p-vector system is determined by p deterministic nonparametric functions superimposed by error terms, possibly dependent cross sectionally. The basic interest is in linear combinations across the p dimensions that make the deterministic functions constant over time. The number of such linearly independent linear combinations is referred to as a cotrending dimension, and their spanned space as a cotrending space. This work puts forward a framework to test statistically for cotrending dimension and space. Connections to principal component analysis and cointegration are also considered. Finally, a simulation study to assess the finite-sample performance of the proposed tests, and applications to several real data sets are also provided.
Asymptotically optimal control of FDR and related metrics for sequential multiple testing
Jay Bartroff*
Mathematics, University of Southern California, USA

Abstract: I will discuss asymptotically optimal multiple testing procedures for sequential data in the context of prior information on the number of false null hypotheses, for controlling FDR/FNR, pFDR/pFNR, and other metrics. These procedures are closely related to those proposed and shown by Song & Fellouris (2017) to be asymptotically optimal for controlling type 1 and 2 familywise error rates (FWEs). We show that by appropriately adjusting the critical values of the Song-Fellouris procedures, they can be made asymptotically optimal for controlling any multiple testing error metric that is bounded between multiples of FWE in a certain sense. In addition to FDR/FNR and pFDR/pFNR this includes other metrics like the per-comparison and per-family error rates, and the false positive rate. Our setup includes asymptotic regimes in which the number of null hypotheses approaches infinity.

Nearly optimal sequential detection of signals in correlated Gaussian noise
Alexander Tartakovsky1 and Grigory Sokolov2*
1Space Informatics Laboratory, Moscow Institute of Physics and Technology (MIPT), Russia
2Mathematics, Xavier University, USA

Abstract: Detecting an object in AR(p) noise assuming the intensity of the signal is not specified is a problem of interest to many practitioners. To this end we examine three procedures: (i) an adaptive version of the sequential probability ratio test (SPRT) built upon one-stage delayed estimators of the unknown signal intensity; (ii) the generalized SPRT; and (iii) the non-adaptive double SPRT (2-SPRT). The generalized SPRT has certain drawbacks in selecting thresholds to guarantee the upper bounds on error probabilities, but may appear to be slightly more efficient than the adaptive SPRT. However, simulations show that the loss in performance of the adaptive SPRT compared to the generalized SPRT is very minor, so-coupled with the error probability guarantee-the adaptive SPRT can be recommended for practical applications. And although the non-adaptive 2-SPRT is not asymptotically optimal for all signal strength values, it does offer benefits at the worst point in the indifference zone.

[Acknowledgement: The work of Alexander Tartakovsky was supported in part by the Russian Science Foundation Grant 18-19-00452 at the Moscow Institute of Physics and Technology.]

Keywords: Adaptive SPRT, Asymptotic optimality, Composite hypotheses, Sequential object detection

A unified approach for solving sequential selection problems
Alexander Goldenshluger1, Yaakov Malinovsky2* and Assaf Zeevi3
1Statistics, University of Haifa, Israel
2Mathematics and Statistics, University of Maryland, Baltimore County, USA
3Graduate School of Business, Columbia University, USA

Abstract: In this work we develop a unified approach for solving a wide class of sequential selection problems. This class includes, but is not limited to, selection problems with no-information, rank-dependent rewards, and considers both fixed as well as random problem horizons. We demonstrate that our approach allows exact and efficient
computation of optimal policies and various performance metrics thereof for a variety of sequential selection problems, several of which have not been solved to date.

**Keywords:** Sequential selection, Optimal stopping, Secretary problems

---

**Sequential change detection by optimal weighted l2 divergence**

Yao Xie* and Liyan Xie  
Georgia Institute of Technology

**Abstract:** We present a new non-parametric statistic, called the weighed l2 divergence, based on empirical distributions for sequential change detection. We start by constructing the weighed l2 divergence as a fundamental building block for two-sample tests and change detection. The proposed statistic is proved to attain the optimal sample complexity in the offline setting. We then study the sequential change detection using the weighed l2 divergence and characterize the fundamental performance metrics, including the average run length (ARL) and the expected detection delay (EDD). We also present practical algorithms to find the optimal projection to handle high-dimensional data and the optimal weights, which is critical to quick detection since, in such settings, there are not many post-change samples. Simulation results and real data examples are provided to validate the good performance of the proposed method.

**Keywords:** Change-point detection, Non-parametric statistics

---

**Detection of temporary disorders**

Michael Baron* and Sergey Malov  
1Mathematics and Statistics, American University, USA  
2Institute of Translational Biomedicine, St.-Petersburg State University, St.-Petersburg, Russia

**Abstract:** Change-point detection methods are proposed for the case of temporary failures, or transient changes, when an unexpected disorder is ultimately followed by an adjustment and return to the initial state. A known base distribution of the in-control state changes to different unknown distributions for unknown periods of time. Sequential and retrospective methods are proposed for the detection and estimation of each pair of change-points. Examples of similar problems are shown in quality and process control, energy finance, and statistical genetics, although the meaning of disorder and adjustment change-points is quite different in these applications.  

[Acknowledgment: The work of M. Baron is supported by U.S. National Science Foundation grant 1737960. The work of S. Malov is supported by the Russian Science Foundation grant 20-14-00072.]

**Keywords:** Change-point, Cusum, Transient change, False alarm

---

**OCS30 Stochastic Adaptive Optimization Algorithms and their Applications to Neural Networks**

**An adaptive strong order 1 method for SDEs with discontinuous drift coefficient**

Larisa Yaroslavtseva*  
Faculty of Computer Science and Mathematics, University of Passau, Germany

**Abstract:** In recent years, a number of results have been proven in the literature for strong approximation of stochastic differential equations (SDEs) with a drift coefficient that may have discontinuities in space. In many of these
results it is assumed that the drift coefficient satisfies piecewise regularity conditions and the diffusion coefficient is Lipschitz continuous and non-degenerate at the discontinuity points of the drift coefficient. For scalar SDEs of that type the best $L_p$-error rate known so far for approximation of the solution at the final time point is $3/4$ in terms of the number of evaluations of the driving Brownian motion and it is achieved by the transformed equidistant quasi-Milstein scheme. Recently in [1] it has been shown that for such SDEs the $L_p$-error rate $3/4$ can not be improved in general by no numerical method based on evaluations of the driving Brownian motion at fixed time points. In this talk we present a numerical method based on sequential evaluations of the driving Brownian motion, which achieves an $L_p$-error rate of at least $1$ in terms of the average number of evaluations of the driving Brownian motion for such SDEs.

[The talk is based on joint work with Thomas Müller-Gronbach (University of Passau).]

Reference

Keywords: Stochastic differential equations, Strong approximation, Discontinuous drift coefficient, Adaptive methods

Nonconvex optimization via TUSLA with discontinuous updating
Dongyoung Lim¹, Ariel Neufeld², Sotirios Sabanis¹ and Ying Zhang²*
¹School of Mathematics, University of Edinburgh, UK
²School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore

Abstract: We study the tamed unadjusted stochastic Langevin algorithm (TUSLA) proposed in Lovas et al. (2021) in the context of nonconvex optimization. In particular, we consider the case where the objective function of the optimization problem has a superlinear and discontinuous stochastic gradient. In such a setting, nonasymptotic error bounds are provided for the TUSLA algorithm in Wasserstein-1 and Wasserstein-2 distances. The latter result enables us to further derive theoretical guarantees for the expected excess risk. Numerical experiments are presented for synthetic examples where popular algorithms, e.g. ADAM, AMSGRAD, RMSProp, and SGD, fail to find the minimizer of the objective functions due to the superlinearity and the discontinuity of the stochastic gradients, while the TUSLA algorithm converges rapidly to the optimal solution. Moreover, an example in transfer learning is provided to illustrate the applicability of the TUSLA algorithm, and its simulation results support our theoretical findings.

Keywords: Nonconvex optimization, Nonasymptotic estimates, Discontinuous stochastic gradient, Superlinear gradient

Approximation of stochastic equations with irregular drifts
Konstantinos Dareiotis*
School of Mathematics, University of Leeds, UK

Abstract: In this talk we will discuss about the rate of convergence of the Euler scheme for stochastic differential equations with irregular drifts. Our approach relies on regularisation-by-noise techniques and more specifically, on the recently developed stochastic sewing lemma. The advantages of this approach are numerous and include the derivation of improved (optimal) rates and the treatment of non-Markovian settings. We will consider drifts in Hölder and Sobolev classes, but also merely bounded and measurable. The latter is the first and at the same time optimal quantification of a convergence theorem of Gyöngy and Krylov.

[This talk is based on joint works with Oleg Butkovsky, Khoa Lê, and Máté Gerencsér.]

Keywords: Approximation, Irregular drifts, Regularisation by noise
Neural SDEs: Deep generative models in the diffusion limit
Maxim Raginsky*
Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, USA

Abstract: In deep generative models, the latent variable is generated by a time-inhomogeneous Markov chain, where at each time step we pass the current state through a parametric nonlinear map, such as a feedforward neural net, and add a small independent Gaussian perturbation. In this talk, based on joint work with Belinda Tzen, I will discuss the diffusion limit of such models, where we increase the number of layers while sending the step size and the noise variance to zero. I will first provide a stochastic control formulation of sampling in such generative models. Then I will show how we can quantify the expressiveness of diffusion-based generative models. Specifically, I will prove that one can efficiently sample from a wide class of terminal target distributions by choosing the drift of the latent diffusion from the class of multilayer feedforward neural nets, with the accuracy of sampling measured by the Kullback-Leibler divergence to the target distribution.

Diffusion approximations and control variates for MCMC
Eric Moulines*
Centre de Mathématiques Appliquées, Ecole Polytechnique, France

Abstract: A new method is introduced for the construction of control variates to reduce the variance of additive functionals of Markov Chain Monte Carlo (MCMC) samplers. These control variates are obtained by minimizing the asymptotic variance associated with the Langevin diffusion over a family of functions. To motivate our approach, it is shown that the asymptotic variance of some well-known MCMC algorithms, including the Random Walk Metropolis and the (Metropolis) Unadjusted/Adjusted Langevin Algorithm, are well approximated by that of the Langevin diffusion. When applied to a class of linear control variates, it is established that the variance of the resulting estimators is smaller, for a given computational complexity, than the standard Monte Carlo estimator. Several examples of Bayesian inference problems support our findings.

Keywords: Control variates, Langevin diffusion, Markov Chain Monte Carlo

Multi-step reflection principle and barrier options
Hangsuck Lee¹, Gaeun Lee¹ and Seongjoo Song²*
¹Actuarial Science/Mathematics, Sungkyunkwan University, Korea
²Statistics, Korea University, Korea

Abstract: This paper examines a class of barrier options—multi-step barrier options, which can have any finite number of barriers of any level. We obtain a general, explicit expression of option prices of this type under the Black-Scholes model. Multi-step barrier options are not only useful in that they can handle barriers of different levels and time steps, but can also approximate options with arbitrary barriers. Moreover, they can be embedded in financial products such as deposit insurances based on jump models with simple barriers. Along the way, we derive multi-step reflection principle, which generalizes the reflection principle of Brownian motion.

Keywords: Barrier Options, Multi-step barrier, Reflection principle, Esscher transform

OCS31 BOK Contributed Session: Finance and Contemporary Issues
JULY 22 (THU), 11:30 - 12:00 KST
Change point analysis in Bitcoin return series: a robust approach
Junmo Song¹ and Jiwon Kang²
¹Department of Statistics, Kyungpook National University, Korea
²Department of Computer Science and Statistics, Jeju National University, Korea

Abstract: Over the last decade, Bitcoin has attracted a great deal of public interest and along with this, the Bitcoin market has grown rapidly. Its speculative price movements have also drawn the interest of many researchers as well as financial investors. Accordingly, numerous studies have been devoted to the analysis of Bitcoin, more exactly the volatility modelling of Bitcoin returns. In this study, we are interested in change point analysis in Bitcoin return data. Since Bitcoin returns have some outlying observations that can affect statistical inferences undesirably, we use a robust test for parameter change to locate some significant change points. We report some change points that are not detected by the existing tests and demonstrate that the model with parameter changes are better fitted to the data. Finally, we show that the model incorporating parameter change can improve the forecasting performance of Value-At-Risk.

Keywords: Bitcoin, Change point analysis, Outlying observations, Robust test

A self-normalization test for correlation matrix change
Ji Eun Choi¹* and Dong Wan Shin²
¹Statistics, Pukyong National University, Korea
²Statistics, Ewha Womans University, Korea

Abstract: We construct a new test for correlation matrix break based on the self-normalization method. The self-normalization test has practical advantage over the existing test: easy and stable implementation; not having the singularity issue and the bandwidth selection issue of the existing test; remedying size distortion problem of the existing test under (near) singularity, serial dependence, conditional heteroscedasticity or unconditional heteroscedasticity. This advantage is demonstrated experimentally by a Monte-Carlo simulation and theoretically by showing no need for estimation of complicated covariance matrix of the sample correlations. We establish the asymptotic null distribution and consistency of the self-normalization test. We apply the correlation matrix break tests to the stock log returns of the companies of 10 largest weight of the NASDAQ 100 index and to five volatility indexes for options on individual equities.

Keywords: Self-normalization

Volatility as a risk measure of financial time series: high frequency and realized volatility
Sun Young Hwang*
Department of Statistics, Sookmyung Women’s University, Korea

Abstract: Volatility as a risk measure is defined as a time varying variance process of return of an asset. The GARCH have been useful to model volatilities of various financial time series. This talk reviews standard volatility computations via GARCH models and then discusses recent issues such as multivariate volatility, realized volatility and high frequency volatility of financial time series. To illustrate, applications to various Korean financial time series are made.
CONTRIBUTED SESSION
A Sobolev space theory for SPDEs with space-time nonlocal operators
Junhee Ryu\(^1\), Kyeong-Hun Kim\(^1\) and Daehan Park\(^3\)
\(^1\)Mathematics, Korea University, Korea
\(^2\)SAARC, KAIST, Korea

Abstract: In this talk, we introduce an \( L_p \)-theory \((p \geq 2)\) for the semi-linear stochastic partial differential equation of the type
\[
\partial_t^\alpha u(t,x) = \phi(\Delta)u(t,x) + f(t,x,u) + \partial_t^\beta \sum_{k=1}^{\infty} \int_0^t g^k(s,x,u)dw^k_s,
\]
where the processes \( w^k_s, k \in \{1,2,\cdots\} \) are independent one-dimensional Wiener processes, \( \alpha \in (0,1), \beta < \alpha + \frac{1}{2} \) and \( \partial_t^\alpha \) and \( \partial_t^\beta \) denote the Caputo fractional derivatives of order \( \alpha \) and \( \beta \) respectively. Here, \( \phi \) is a Bernstein function satisfying the following condition: \( \exists \delta_0 \in (0,1] \) and \( c > 0 \) such that
\[
c(\frac{R}{r})^{\delta_0} \leq \frac{\phi(R)}{\phi(r)} \quad 0 < r < R < \infty.
\]
As an application, we also obtain an \( L_p \)-regularity theory for SPDEs driven by space-time white noise:
\[
\partial_t^\alpha u = \phi(\Delta)u + f(u) + \partial_t^\beta \int_0^t h(u)dB_t.
\]
Here, spatial dimension \( d \) satisfies \( d < 2\delta_0(2 - (2\beta - 1)\alpha^{-1}) \) with \( \delta_0 > 1/4 \). In particular, if \( \beta < (1 - 3/4\delta_0) + 1/2 \), then one can treat \( d = 1,2,3 \).
[This talk is based on a joint work with Kyeong-Hun Kim and Daehan Park.]

Keywords: Stochastic partial differential equations, Space-time nonlocal operators, Maximal \( L_p \)-regularity, Multi-dimensional space-time white noise

Improved Stability for Linear SPDEs Using Mixed Boundary/Internal Controls
Dan Goreac\(^*\) and Ionut Munteanu\(^2\)
\(^1\)Mathematics, University Shandong Weihai, Chine and University Gustave Eiffel, France, France
\(^2\)Mathematics, University Al I Cuza, Iasi, Romania

Abstract: This talk based on a joint work with I. Munteanu ("Al. I Cuza" Univ., Iasi) is motivated by the asymptotic stabilization of abstract Stochastic PDEs of linear type. As a first step, we exhibit an abstract contribution to the exact controllability (in a general \( L_p \)-sense, \( p > 1 \)) of a class of linear SDEs with general (but time-invariant) rank control coefficient in the noise term. Second, we illustrate on relevant frameworks of SPDEs, a way to drive exactly to 0 their unstable part (of dimension \( n \geq 1 \)) by using M internal (respectively N boundary) controls such that max \( (M, N) < n \). Some examples are presented as is the minimal gain for judicious control dimensions.

Keywords: Stochastic control, Exact controllability, (Stochastic) Partial differential equations, Asymptotic stability

Law of the large numbers and Central limit theorems for stochastic heat equations
Kunwoo Kim\(^*\) and Jaeyun Yi
Mathematics, POSTECH, Korea
Abstract: In this talk, we study limit theorems for time-dependent averages of the form
\[ X_t := \frac{1}{2L(t)} \int_{-L(t)}^{L(t)} u(t, x) \, dx , \quad t \to \infty , \]
where \( L(t) = \exp(\lambda t) \) and \( u(t, x) \) is the solution to a stochastic heat equation driven by space-time white noise. We provide the explicit conditions on \( \lambda \), which guarantee the weak and strong law of large numbers and also the central limit theorems for \( X_t \). These conditions on \( \lambda \) are related to the moment Lyapunov exponents and the full intermittency condition introduced by Carmona and Molchanov.

Keywords: Stochastic heat equations, Law of the large numbers, Central limit theorems, Intermittency

The stochastic heat equation with Lévy noise: Existence, moments and intermittency
Carsten Chong*
Department of Statistics, Columbia University, USA

Abstract: In this talk, we present results about existence, moments and large-time asymptotics of the solution to the stochastic heat equation driven by a Lévy space-time white noise, proved using a combination of decoupling techniques, point process methods and change-of-measure techniques. As one of the more surprising results, we show that the solution exhibits the phenomenon of intermittency for all exponents in all dimensions and for all non-Gaussian Lévy noises, which is fundamentally different to what is known in the Gaussian case. Moreover, we demonstrate that the behavior of the intermittency exponents in terms of a coupling constant depends critically on whether the Lévy noise is light- or heavy-tailed.

[This is based on joint work with Quentin Berger (Sorbonne) and Hubert Lacoin (IMPA).]

Keywords: SPDE, Stochastic heat equation, Lévy noise, Intermittency

How to detect a salami slicer: a stochastic controller-stopper game with unknown competition
Kristoffer Lindensjö*
Department of Mathematics, Stockholm University, Sweden

Abstract: A stochastic controller-stopper game for fraud detection is introduced. A process \( X_t = -\theta \Lambda_t + W_t \) represents the holdings of an account holder, where \( \theta \) is a Bernoulli random variable indicating whether a fraudster is active or not, and \( \theta \Lambda_t \) is the accumulated amount stolen by the fraudster; and \( W \) is a Brownian motion. The fraudster seeks to choose the process \( \Lambda \) that maximizes the expected discounted stolen amount
\[ \mathbb{E} \left[ \int_0^T e^{-rt} d\Lambda_s \, \left| \theta = 1 \right. \right] , \]
whereas the account holder — who observes only the path of \( X \) and does not know whether she is the victim of fraud or not — seeks to choose a stopping time \( \tau \) that deactivates the fraudster (at a cost \( M \)) in order to minimize the expected discounted cost of theft and deactivation
\[ \mathbb{E} \left[ \theta \int_0^\tau e^{-rt} d\Lambda_s + e^{-r\tau} M \mathbb{1}_{\{\tau < \infty\}} \right] . \]

Relying stochastic filtering theory we find explicit Nash equilibria of both pure and mixed type.
[Based on joint work with Erik Ekström (Uppsala U) and Marcus Olofsson (Umeå U).]

Keywords: Stochastic dynamic games, Filtering theory, Stochastic controller-stopper problems
Solving the selection-recombination equation: Ancestral lines and duality
Frederic Alberti*
Department of Mathematics, Bielefeld University, Germany

Abstract: The selection-recombination equation is a high-dimensional, nonlinear system of ordinary differential equations, which describe the evolution of the genetic type composition of a population under selection and recombination, in a law of large numbers regime. So far, explicit solutions have seemed out of reach; only in the special case of three loci, with selection acting on one of them, has an approximate solution been found, but without an obvious path to generalisation. We consider the case of an arbitrary number of neutral loci, linked to a single selected locus. In this setting, we investigate how the (random) genealogical structure of the problem can be succinctly encoded by a novel ‘ancestral initiation graph’, and how it gives rise to a recursive integral representation of the solution with a clear, probabilistic interpretation.

References

Keywords: Population genetics, Recombination, Stochastic processes, Recombination

Short time asymptotics for modulated rough stochastic volatility models
Barbara Pacchiarotti*
Dipartimento di Matematica, Università degli Studi di Roma "Tor Vergata", Italy

Abstract: In this paper, we establish a small time large deviation principle for log-price processes when the volatility is a function of a modulated Volterra process. With modulated process we mean a Volterra process with a self similar kernel multiplied by a slowly varying function. We also deduce short time asymptotics for implied volatility and for pricing.

Splitting methods for SDEs with locally Lipschitz drift. An illustration on the FitzHugh-Nagumo model
Massimiliano Tamborrino1*, Adeline Samson2, Evelyn Buckwar3 and Irene Tubikanec4
1Statistics, University of Warwick, UK
2Laboratoire Jean Kuntzmann, University Grenoble Alpes, France
3Institute for Stochastics / Centre for Mathematical Sciences, Johannes Kepler University Linz / Lund University, Austria
4Institute for Stochastics, Johannes Kepler University Linz, Austria

Abstract: In this talk, we construct and analyse explicit numerical splitting methods for a class of semilinear stochastic differential equations (SDEs) with additive noise, where the drift is allowed to grow polynomially and satisfies a global one-sided Lipschitz condition. The methods are proved to be mean-square convergent of order 1 and to preserve important structural properties of the SDE. In particular, first, they are hypoelliptic in every iteration step. Second, they are geometrically ergodic and have asymptotically bounded second moments. Third, they preserve oscillatory dynamics, such as amplitudes, frequencies and phases of oscillations, even for large time steps. Our results are
illustrated on the stochastic FitzHugh-Nagumo model (a well-known neuronal model describing the generation of spikes of single neurons at the intracellular level) and compared with known mean-square convergent tamed/truncated variants of the Euler-Maruyama method. The capability of the proposed splitting methods to preserve the aforementioned properties makes them applicable within different statistical inference procedures. In contrast, known Euler-Maruyama type methods commonly fail in preserving such properties, yielding ill-conditioned likelihood-based estimation tools or computationally infeasible simulation-based inference algorithms.

**Keywords:** Structure-preserving numerical methods, Hypoellipticity, Ergodicity, Mean-square convergence

**Simulation methods for trawl processes**

Dan Leonte*
Mathematics, PhD student at Imperial College London, UK

**Abstract:** Trawl processes are continuous-time, stationary and infinitely divisible processes which can describe a wide range of possible serial correlation patterns in data. This talk introduces a new algorithm for the efficient simulation of monotonic trawl processes. The algorithm accommodates any monotonic trawl shape and any infinitely divisible distribution described via the Lévy seed, requiring only access to samples from the distribution of the Lévy seed. Further, the computational complexity does not scale with the number of spatial dimensions of the trawl. We describe how the above method can be generalized to a simulation scheme for monotonic ambit fields via Monte Carlo methods.

**Keywords:** Lévy bases, Stochastic simulation, Trawl processes

**Stochastic optimal control of SDEs and importance sampling**

Han Cheng Lie*
Institute for Mathematics, University of Potsdam, Germany

**Abstract:** In applications that involve rare events, a common problem is to estimate the statistics of a functional with respect to a reference measure, where the reference measure is the law of the solution to a specific SDE. The presence of rare events motivates the approach of importance sampling by the change of drift technique. This leads to a stochastic optimal control problem, where the objective consists in the sum of the expectation of the functional of interest and a regularisation term that is proportional to the relative entropy or Kullback-Leibler divergence between the reference measure and the importance sampling measure. We analyse a class of gradient-based numerical methods for solving these stochastic optimal control problems, by computing derivatives of the individual terms in the objective, and by using this derivative information to analyse the convexity properties of the terms in the objective.

**Keywords:** Stochastic optimal control, Importance sampling, Stochastic differential equations, Numerical methods

**Opinion dynamics with Lotka-Volterra type interactions**

Michele Aleandri* and Ida Germana Minelli2
1LUISS Guido Carli, Italy
2Dipartimento di Ingegneria e Scienze dell’Informazione e Matematica, L’Aquila University, Italy

**Abstract:** We investigate a class of models for opinion dynamics in a population with two interacting families of individuals. Each family has an intrinsic mean field “Voter-like” dynamics which is influenced by interaction with the other family. The interaction terms describe a cooperative/conformist or competitive/nonconformist attitude of one family with respect to the other. We prove chaos propagation, ie, we show that on any time interval $[0, T]$, as the size of the system goes to infinity, each individual behaves independently of the others with transition rates driven by a
macroscopic equation. We focus in particular on models with Lotka-Volterra type interactions, ie, models with cooperative vs. competitive families. For these models, although the microscopic system is driven as to consensus within each family, a periodic behaviour arises in the macroscopic scale. In order to describe fluctuations between the limiting periodic orbits, we identify a slow variable in the microscopic system and, through an averaging principle, we find a diffusion which describes the macroscopic dynamics of such variable on a larger time scale.

**Keywords:** Averaging principle, Chaos propagation, Interacting particle systems, Scaling limits

---

**Parameter estimation for weakly interacting particle systems and stochastic McKean-Vlasov processes**

Louis Sharrock*, Nikolas Kantas¹, Grigorios Pavliotis¹ and Panos Parpas²

¹Department of Mathematics, Imperial College London, UK
²Department of Computing, Imperial College London, UK

**Abstract:** In this presentation, we consider the problem of parameter estimation for a fully observed McKean-Vlasov stochastic differential equation (MVSDE), and the associated system of weakly interacting particles. We begin by establishing consistency and asymptotic normality of the offline maximum likelihood estimator (MLE) of the interacting particle system (IPS) in the limit as the number of particles (N) tends to infinity. We then propose a recursive MLE for the MVSDE, which evolves according to a stochastic gradient ascent algorithm on the asymptotic log-likelihood of the IPS. Under suitable assumptions which guarantee exponential ergodicity and uniform-in-time propagation of chaos for the MVSDE and the IPS, we prove that this estimator converges in $L^1$ to the stationary points of the asymptotic log-likelihood of the MVSDE in the joint limit as $N$ and $t$ tend to infinity. Under the additional assumption of global strong concavity, we also demonstrate that our estimator converges in $L^2$ to the unique maximiser of the asymptotic log-likelihood of the MVSDE, and establish an $L^2$ convergence rate. Our results are demonstrated via several numerical examples of practical interest, including a linear mean field model, and a stochastic opinion dynamics model.

**Keywords:** Stochastic McKean-Vlasov Equation, Parameter Estimation, Maximum Likelihood, Stochastic Gradient Descent

---

**CLT for cyclic long-memory processes**

Andriy Olenko*

Mathematics and statistics, La Trobe University, Australia

**Abstract:** Cyclic long-memory stochastic processes are studied. Cyclic long-memory behavior attracted increasing attention in recent years due to its importance in finance, hydrology, cosmology, internet modelling, and other applications to data with non-seasonal cyclicity. However, various functionals of cyclic long-memory processes have complex asymptotic behavior that has not yet been fully understood and investigated. Spectral singularities at non-zero frequencies play an important role in investigating cyclic processes. The publication [1] introduced the generalized filtered method-of-moments approach to simultaneously estimate singularity location and long-memory parameters. The law of large numbers for the proposed estimators was proved. This talk discusses the central limit theorem for these simultaneous estimators. A wide class of Gegenbauer-type semi-parametric models is considered. Asymptotic normality of several functionals of the cyclic long-memory processes is proved. For the case when values of the functionals are outside the feasible region, we propose new adjusted estimators and investigate their

---
properties. It is shown that they have same asymptotic distributions as the corresponding ones in [1], but are computationally simpler. The methodology includes wavelet transformations as a particular case.

This presentation is based on recent joint results in [2] with A.Ayache, M.Fradon (University of Lille, France) and R. Nanayakkara (La Trobe University, Australia).

References

Keywords: Long-memory, Central limit theorem, Cyclic random process, Filter

Heat contents for time-changed killed Brownian motions
Hyunchul Park*
Mathematics, SUNY New Paltz, USA

Abstract: In this talk, we study various heat content with respect to time-changed killed Brownian motions. The time-change is given by either a large class of subordinators or inverse of the subordinators. When the time-change is given by inverse stable subordinators and the domain is smooth, we show that the spectral heat content has a complete asymptotic expansion which is similar to the case of Brownian motions.

This is a joint work with Kei Kobayashi (Fordham University, USA).

Keywords: Heat content, Spectral heat content, Subordinators, Inverse subordinators

Heat kernel bounds for nonlocal operators with singular kernels
Kyung-Youn Kim*
Mathematical Science, National Chengchi University, Rep of China (Taiwan)

Abstract: We prove sharp two-sided bounds of the fundamental solution for integro-differential operators of order $\alpha \in (0,2)$ that generate a $d$-dimensional Markov process. The corresponding Dirichlet form is comparable to that of $d$-independent copies of one-dimensional jump processes, i.e., the jumping measure is singular with respect to the $d$-dimensional Lebesgue measure.

This is joint work with Moritz Kassmann and Takashi Kumagai.

Keywords: Markov jump process, Heat kernel, Integro-differential operator

The full characterization of the expected supremum of infinitely divisible processes
Rafal Martynek*
Mathematics, University of Warsaw, Poland

Abstract: In this talk I will present the positive answer to the conjecture posed by M. Talagrand in “Regularity of Infinitely Divisible Processes” (1993) concerning two-sided bound of the expected suprema of such processes which does not require any additional assumption on the Levy measure associated with the process. It states that any infinitely divisible process can be decomposed into the part whose size is explained by the chaining method and the
other which is the positive process. The result relies highly on the Bednorz-Latała theorem characterizing suprema of Bernoulli processes and its recent reformulation due to Talagrand together with series representation due to Rosiński. I will also describe how the method of the proof leads to the positive settlement of two others conjectures of Talagrand. Namely, the Generalized Bernoulli Conjecture concerning selector processes and analogous result for empirical processes. These three results completes an important chapter of Talagrand’s program of understanding the suprema of random processes through chaining.

[The part of the talk concerning infinitely divisible processes is based on the joint work with W. Bednorz, while the part about selector and empirical processes was developed by M. Talagrand after we communicated him the initial result.]

**Keywords:** Infinitely divisible processes, Chaining method, Generalized Bernoulli Conjecture

---

**The e-property of asymptotically stable Markov-Feller operators**

Hanna Wojewódka-Ściążko* and Ryszard Kukulski

1Institute of Mathematics, University of Silesia in Katowice, Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Poland

2Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Poland

**Abstract:** We say that a regular Markov operator \( P \), with dual operator \( U \), has the e-property in the set \( R \) of functions if the family of iterates \( \left( U^n f \right)_{n \in \mathbb{N}} \) is equicontinuous for all \( f \in R \). Most often, \( R \) is assumed to be the set of all bounded Lipschitz functions, although it can be also the set of all bounded continuous functions, as in our paper [R. Kukulski and H. Wojewódka-Ściążko, Colloq. Math. 165, 269-283 (2021)]. In [S. Hille et al., Comptes Rendus Math. 355, 1247-1251 (2017)] it is shown that any asymptotically stable Markov-Feller operator with an invariant measure such that the interior of its support is non-empty has the e-property. We generalize this result. To be more precise, we prove that any asymptotically stable Markov-Feller operator has the e-property off a meagre set. Moreover, we propose an equivalent condition for the e-property of asymptotically stable Markov-Feller operators. Namely, we prove that an asymptotically stable Markov-Feller operator has the e-property if and only if it has the e-property at least at one point of the support of its invariant measure. Our results then naturally imply the main theorem of [S. Hille et al., Comptes Rendus Math. 355, 1247-1251 (2017)]. Indeed, if the interior of the support of an invariant measure of a Markov-Feller operator \( P \) is non-empty, then there exists at least one point in this support at which \( P \) has the e-property. This, in turn, implies that \( P \) has the e-property at any point. We also provide the example of an asymptotically stable Markov-Feller operator such that the set of points at which the operator fails the e-property is dense. The example shows that the main result of our paper is tight.

**Keywords:** Markov–Feller operator, Asymptotic stability, E-property, Equicontinuity

---

**On nonlinear filtering of jump diffusions**

Fabian Germ* and István Gyöngy

University of Edinburgh

**Abstract:** We consider a multi-dimensional signal and observation model, \( Z_t = (X_t, Y_t) \), which is a jump diffusion, i.e., the solution of an SDE driven by Wiener processes and Poisson martingale measures. The multi-dimensional “signal” \( X_t \) is not observable, and we are interested in classic questions about the mean square estimate of \( X_t \) for each time \( t \), given the “observations” \( Y_s \) for \( s \in [0,t] \). These questions were intensively studied for partially observable diffusion processes \( Z_t \) in various generality in the past, and a quite complete filtering theory for diffusion processes was
developed. Our aim is to contribute to recent studies in extending the nonlinear filtering theory of diffusion processes to that of jump diffusions. We allow the signal and observation noises to be correlated and the infinitesimal generator of $Z_t$ to be degenerate in the coordinate directions of the signal. First we present the filtering equations: the equations for the time evolution of the conditional distribution $P_t$ and of an unnormalised conditional distribution $Q_t$ of $X_t$, given the observations $Y_s$ for $s$ in $[0,t]$. These equations are (possibly) degenerate stochastic integro-differential equations. They are stochastic PDEs, often referred as the Kushner-Shiryayev and Zakai equations, respectively, in the special case when $Z_t$ is a diffusion process. Next we present new results on existence and uniqueness of the solutions to the filtering equations in $L^p$-spaces. Finally, using our results on the solutions to the filtering equations, we give conditions ensuring that the conditional density $p_t := dP_t / dx$, with respect to Lebesgue measure exists and belongs to an $L^p$-space for each $t>0$. Moreover, under quite general regularity conditions on the initial conditional density $p_0$ and on the coefficients of the SDE for $Z_t$, we prove that the process $p_t$ is a càdlàg process with values in Bessel potential spaces $H^s_p$ and in Slobodeckij spaces $W^a_p$.

**Keywords:** Nonlinear filtering, Poisson random measures, Lévy processes, Stochastic integro-differential equations

---

**Uniqueness and superposition of the distribution-dependent Zakai equations**

Huijie Qiao* and Meiqi Liu  
Department of Mathematics, Southeast University, China

**Abstract:** The work concerns the Zakai equations from nonlinear filtering problems of McKean-Vlasov stochastic differential equations with correlated noises. First, we establish the Kushner-Stratonovich equations, the Zakai equations and the distribution-dependent Zakai equations. And then, the pathwise uniqueness, uniqueness in joint law and uniqueness in law of weak solutions for the distribution-dependent Zakai equations are shown. Finally, we prove a superposition principle between the distribution-dependent Zakai equations and distribution-dependent Fokker-Planck equations. As a by-product, we give some conditions under which distribution-dependent Fokker-Planck equations have a weak solutions.

**Keywords:** McKean-Vlasov SDEs, The distribution-dependent Zakai equations, Distribution-dependent Fokker-Planck equations, A superposition principle

---

**Quadratic variation and quadratic roughness**

Purba Das*  
Mathematics, University of Oxford, UK

**Abstract:** We study the concept of quadratic variation of a continuous path along a sequence of partitions and its dependence with respect to the choice of the partition sequence. We define the concept of quadratic roughness of a path along a partition sequence and show that, for Hölder-continuous paths satisfying this roughness condition, the quadratic variation along balanced partitions is invariant with respect to the choice of the partition sequence. Typical paths of Brownian motion are shown to satisfy this quadratic roughness property almost-surely along any partition with a required step size condition. Using these results we derive a formulation of Föllmer’s pathwise integration along paths with finite quadratic variation which is invariant with respect to the partition sequence.

**Keywords:** Quadratic variation, Itô calculus, Local time, Brownian motion

---

**Eyring-Kramers formula for non-reversible metastable diffusion processes**

Jungkyoung Lee* and Insuk Seo  
Department of Mathematical Sciences, Seoul National University, Korea
Abstract: In this talk, we consider diffusion processes that admit a Gibbs invariant measure but are non-reversible. Such diffusion processes exhibit metastable behavior if the associated potential function owns multiple local minima. For this model, we provide a proof of the Eyring–Kramers formula which provides sharp asymptotics of the mean of the transition time from a local minimum to a deeper one. In particular, our work indicates that the metastable transitions of non-reversible processes are faster than that of reversible ones.

Keywords: Metastability, Diffusion process, Eyring-Kramers formula

CS07 SDEs and Fractional Brownian Motions

JULY 20 (TUE), 21:30 - 22:00 KST

Weak rough-path type solutions for singular Lévy SDEs
Helena Katharina Kremp* and Nicolas Perkowski
Mathematics, Freie Universität Berlin, Germany

Abstract: Since the works by Delarue, Diel and Cannizzaro, Chouk (in the Brownian noise setting), and our previous work, the existence and uniqueness of solutions to the martingale problem associated to multidimensional SDEs with additive $\alpha$-stable Lévy noise for $\alpha$ in $(1,2]$ and rough Besov drift of regularity $\beta$ in $((2-2\alpha)/3,0]$ is known. Motivated by the equivalence of probabilistic weak solutions to SDEs with bounded, measurable drift and solutions to the martingale problem, we define a (non-canonical) weak solution concept for singular Lévy diffusions, proving moreover equivalence to the martingale solution in both the Young (i.e. $\beta > (1-\alpha)/2$), as well as in the rough regime (i.e. $\beta > (2-2\alpha)/3$). This turns out to be highly non-trivial in the rough case and forces us to define certain rough stochastic sewing integrals involved. In particular, we show that the canonical weak solution concept (introduced also by Athreya, Butkovsky, Mytnik in the Young case), which is well-posed in the Young case, yields non-uniqueness of solutions in the rough case.

Keywords: Singular diffusions, Paracontrolled distributions, Equivalence of weak vs. martingale solutions, Stochastic sewing lemma

Functional limit theorems for approximating irregular SDEs, general diffusions and their exit times
Mikhail Urusov*
Faculty of Mathematics, University of Duisburg-Essen, Germany

Abstract: We propose a new approach for approximating one-dimensional continuous Markov processes in law. More specifically, we discuss the following results: (1) A functional limit theorem (FLT) for weak approximation of the paths of arbitrary continuous Markov processes; (2) An FLT for weak approximation of the paths and exit times. The second FLT has a stronger conclusion but requires a stronger assumption, which is essential. We propose a new scheme, called EMCEL, which satisfies the assumption of the second FLT and thus allows to approximate every one-dimensional continuous Markov process together with its exit times. The approach is illustrated by a couple of examples with peculiar behavior, including an irregular SDE, for which the corresponding Euler scheme does not converge even weakly, a sticky Brownian motion and a Brownian motion slowed down on the Cantor set.

[This is a joint work with Stefan Ankirchner and Thomas Kruse.]

Keywords: Functional limit theorem, Irregular SDE, Diffusion with sticky features, Exit time
Orlicz norm and concentration inequalities for beta-heavy tailed distributions
Emmanuel Gobet*
Applied Math Department, Ecole Polytechnique, France

Abstract: Understanding how sample statistical fluctuations impact prediction errors is crucial in learning algorithms. This is typically made by quantifying the probability that a sum of random variables deviates from its expectation by a certain threshold. The case of sub-Gaussian, or the sub-exponential random variables as well as the case of alpha-exponential tails have been largely covered by the literature (for example, via Bennett inequality and via Bernstein inequality...). In this work we focus on situations where the distributions have long tail (like log-normal or log-gamma distributions). In this setting, we establish a new Talagrand-type inequality about the Orlicz norm of the sum of independent random variables of this type, and some maximal inequality. The concentration inequalities then follow.

Keywords: Concentration inequalities, Heavy tail, Probability in Banach space

The Dickman–Goncharov distribution
Stanislav Molchanov¹ and Vladimir Panov²*
¹Department of Mathematics, University of North Carolina at Charlotte, USA
²International Laboratory of Stochastic Analysis and its Applications, HSE University, Russia

Abstract: In the 1930s and 40s, one and the same delay differential equation appeared in papers by two mathematicians, Karl Dickman and Vasily Goncharov, who dealt with completely different problems. Dickman investigated the limit value of the number of natural numbers free of large prime factors, while Goncharov examined the asymptotics of the maximum cycle length in decompositions of random permutations. The equation obtained in these papers defines, under a certain initial condition, the density of a probability distribution now called the Dickman–Goncharov distribution (this term was first proposed by A.Vershik in 1986). Recently, a number of completely new applications of the Dickman–Goncharov distribution have appeared in mathematics (random walks on solvable groups, random graph theory, and so on) and also in biology (models of growth and evolution of unicellular populations), finance (theory of extreme phenomena in finance and insurance), physics (the model of random energy levels), and other fields. Despite the extensive scope of applications of this distribution and of more general but related models, all the mathematical aspects of this topic (for example, infinite divisibility and absolute continuity) are little known even to specialists in limit theorems. My talk is mainly based on our survey [Molchanov S., Panov V. The Dickman–Goncharov distribution. Russian Mathematical Surveys. 2020. Vol. 75. No. 6. P. 1089-1132], which is intended to fill this gap. I'm going also to discuss several new results for the generalised Dickman-Goncharov distribution, which in the discrete case are closely related to the solution of the well-known Erdos problem for Bernoulli convolutions.

Keywords: Dickman–Goncharov distribution, Erdos problem, Random energy model, Vershik chain

Continuous scaled phase-type distributions
Jorge Yslas*
Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland

Abstract: In this talk, we study random variables characterized as the product of phase-type distributions and
continuous random variables. Under this construction, one can obtain closed-form formulas for the different functionals of the resulting models. We provide new results regarding the tail behavior of these distributions and show how an EM algorithm can be employed for maximum-likelihood estimation. Finally, we present several numerical examples with real insurance data sets.

**Keywords:** Heavy tails, Parameter estimation, Phase-type, Scale mixtures

---

**On eigenvalue distributions of large auto-covariance matrices**

Wangjun Yuan\(^1\) and Jianfeng Yao\(^2\)

\(^1\)Department of Mathematics, The University of Hong Kong, China

\(^2\)Department of Statistics and Actuarial Science, The University of Hong Kong, China

**Abstract:** In comparison to Hermitian random matrices, the study of random matrices without Hermitian or Unitary structure is more limited. The famous circular law, that the eigenvalue empirical measures of random matrices whose entries are i.i.d. centered complex random variables with unit variance converges almost surely to the uniform distribution on the unit disk, was posed in 1950s. Important breakthroughs were made by Bai, Girko, Tao and Vu, and the conjecture was finally proved by Tao et al. Auto-covariance matrices are important in statistics, especially in high-dimensional time series analysis. Let \(Y(k) = \frac{1}{n}(X_{i+1}X_i^* + \cdots + X_nX_{n-k}^*)\) be the time series observed at time \(1 \leq i \leq n\). We study the non-Hermitian matrix \(Y(k)\) is known as lag-\(k\) auto-covariance matrix of the time series. Recently, a similar model was studied in [1], where the matrix \(Z = Y(1) + \frac{1}{n}X_i^*X_i^*\) was considered. To obtained the limit of the sequence of eigenvalue empirical measures, [1] used the linearization technique, the small ball probability as well as the logarithmic potential. Although the two matrices \(Y(1)\) and \(Z\) differ only by a rank-one matrix, the limit eigenvalue empirical measure of \(Z\) does not imply anything a priori on the asymptotic properties of \(Y(1)\) since rank-one perturbations may destroy the limit completely. The linearization technique in [1] fails in our case. We design another auxiliary matrix to obtain the lower bound for the least singular value. The lower bound together with the small rank perturbations on the limit of singular values empirical measure leads to the limit of the sequence of eigenvalue empirical measures of \(Y(k)\). The results can be found in [2].

**References**


**Keywords:** Least singular value, Eigenvalue distribution, Large auto-covariance matrix, Girko’s Hermitization principle

---

**Linear spectral statistics of sequential sample covariance matrices**

Nina Dörnemann\(^*\) and Holger Dette

\(^*\)Faculty of Mathematics, Ruhr University Bochum, Germany

**Abstract:** Estimation and testing of a high-dimensional covariance matrix is a fundamental problem of statistical inference with numerous applications in biostatistics, wireless communications and finance. Linear spectral statistics are frequently used to construct tests for various hypotheses. In this work, we consider linear spectral statistics from a sequential point of view. To be more precise, we prove that the stochastic process corresponding to a linear spectral statistic of the sequential empirical covariance estimator converges weakly to a non-standard Gaussian process. As an application we use these results to develop a novel approach for monitoring the sphericity assumption in a high-
dimensional framework, even if the dimension of the underlying data is larger than the sample size. Compared to previous contributions in this field, the results of the present work are conceptually different, because the sequential parameter used in the definition of the process also appears in the eigenvalues. This “non-linearity” results in a substantially more complicated structure of the problem. In particular, the limiting processes are non-standard Gaussian processes, and the proofs of our results (in particular the proof of tightness) require an extended machinery, which has so far not been considered in the literature on linear spectral statistics. As a consequence, we provide a substantial generalization of the classical CLT for linear spectral statistics proven by Bai and Silverstein.

**Keywords:** Linear spectral statistics, Sequential processes, Sphericity, Sequential sample covariance matrix

### Couplings for Andersen dynamics and related piecewise deterministic Markov processes

Nawaf Bou-Rabee*
Rutgers University Camden

**Abstract:** Piecewise Deterministic Markov Processes (e.g. bouncy particle and zigzag samplers) have recently garnered increased research interest for their potential to rapidly sample high-dimensional probability distributions. However, there remain many open questions and challenges to understanding their mixing time and convergence properties. In this talk, I will highlight recent progress on Andersen dynamics: a PDMP that iterates between Hamiltonian flows and velocity randomizations of randomly selected particles. Various couplings of Andersen dynamics will be used to obtain explicit convergence bounds in a Wasserstein sense. The bounds are dimension free for not necessarily convex potentials with weakly interacting components on a high dimensional torus, and for strongly convex and gradient Lipschitz potentials on a Euclidean product space.

**Keywords:** Couplings, Piecewise Deterministic Markov Process, MCMC, High Dimension

### Measure Valued Processes Characterized by a Field of Reflecting Brownian Motions Arising from Certain Queuing Problems

Amarjit Budhiraja*
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, USA

**Abstract:** We study a class of queuing models in which the state of the system at any instant is given by a finite nonnegative Borel measure on the nonnegative real line which puts a unit atom at the remaining processing time of each job in system. The settings where the processing time distributions of jobs have bounded support or light tails have been investigated in previous works. In the current work we study the case where these distributions have finite second moments and regularly varying tails. By considering a parameter given in terms of the tails of processing time distributions, we consider a novel time, volume, and spatial scaling for the measure valued process and show that the scaled measure valued process converges in distribution (in the space of paths of measures). In a sharp contrast to results for bounded support and light tailed service time distributions, this time there is no state space collapse and the limiting random measures are not concentrated on a single atom. Nevertheless, the description of the limit is simple and given explicitly in terms of a certain random field of reflected Brownian motions. [This is joint work with Sayan Banerjee and Amber Puha.]

**Keywords:** Measure valued processes, Reflecting Brownian motions

---

**CS10 Reflecting Diffusion Processes, Stochastic Networks and Their Applications**

**JULY 20 (TUE), 22:30 - 23:00 KST**

---

**Measure Valued Processes Characterized by a Field of Reflecting Brownian Motions Arising from Certain Queuing Problems**

Amarjit Budhiraja*
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, USA

**Abstract:** We study a class of queuing models in which the state of the system at any instant is given by a finite nonnegative Borel measure on the nonnegative real line which puts a unit atom at the remaining processing time of each job in system. The settings where the processing time distributions of jobs have bounded support or light tails have been investigated in previous works. In the current work we study the case where these distributions have finite second moments and regularly varying tails. By considering a parameter given in terms of the tails of processing time distributions, we consider a novel time, volume, and spatial scaling for the measure valued process and show that the scaled measure valued process converges in distribution (in the space of paths of measures). In a sharp contrast to results for bounded support and light tailed service time distributions, this time there is no state space collapse and the limiting random measures are not concentrated on a single atom. Nevertheless, the description of the limit is simple and given explicitly in terms of a certain random field of reflected Brownian motions. [This is joint work with Sayan Banerjee and Amber Puha.]

**Keywords:** Measure valued processes, Reflecting Brownian motions
Asymptotic behavior of a critical fluid model for bandwidth sharing with general file size distributions
Yingjia Fu*
Mathematics, University of California San Diego, USA

Abstract: This work concerns the asymptotic behavior of solutions to a critical fluid model for a data communication network, where file sizes are generally distributed and the network operates under a fair bandwidth sharing policy, chosen from the family of (weighted) \( \alpha \)-fair policies introduced by Mo and Walrand. Solutions of the fluid model are measure-valued functions of time. Under law of large numbers scaling, Gromoll and Williams proved that these solutions approximate dynamic solutions of a flow level model for congestion control in data communication networks, introduced by Massoulié and Roberts. In a recent work, we proved stability of the strictly subcritical version of this fluid model under mild assumptions. In this talk, we study the asymptotic behavior (as time goes to infinity) of solutions of the critical fluid model, in which the nominal load on each network resource is less than or equal to its capacity and at least one resource is fully loaded. For this we introduce a new Lyapunov function, inspired by the work of Kelly and Williams, Mulvany et al. and Paganini et al. Using this, under moderate conditions on the file size distributions, we prove that critical fluid model solutions converge uniformly to the set of invariant states as time goes to infinity, when started in suitable relatively compact sets. We expect that this result will play a key role in developing a diffusion approximation for the critically loaded flow level model of Massoulié and Roberts. Furthermore, the techniques developed here may be useful for studying other stochastic network models with resource sharing.

Keywords: Critical fluid model, \( \alpha \)-fair bandwidth sharing, Asymptotic behavior, General file size distributions

Error bounds for the one-dimensional constrained langevin approximation for density dependent Markov chains
Felipe A. Campos* and Ruth J. Williams
Mathematics, University of California, San Diego, USA

Abstract: The stochastic dynamics of chemical reaction networks are often modeled using continuous-time Markov chains. However, except in very special cases, these processes cannot be analysed exactly and their simulation can be computationally intensive. An approach to this problem is to consider a diffusion approximation. The Constrained Langevin Approximation (CLA) is a reflected diffusion approximation for stochastic chemical reaction networks proposed by Leite & Williams. In this work, we extend this approximation to (nearly) density dependent Markov chains, when the diffusion state space is one-dimensional. Then, we provide a bound for the error of the CLA in a strong approximation. Finally, we discuss some applications for chemical reaction networks and epidemic models, illustrating these with examples.

Keywords: Diffusion approximation, Chemical reaction networks, Stochastic differential equation with reflection, Systems biology

Obliquely reflecting diffusions in nonsmooth domains: Some new uniqueness results
Cristina Costantini* and Thomas G. Kurtz
1Department of Economic Studies, University of Chieti-Pescara, Italy
2Department of Mathematics and Department of Statistics, University of Wisconsin-Madison, USA

Abstract: Exhaustive existence and uniqueness results are available for Brownian motion reflecting in a polyhedron with constant direction of reflection on each face (Varadhan and Williams, 1984; Dai and Williams, 1995) or in a smooth cone with radially constant direction of reflection (Kwon and Williams, 1991). Only partial results are available for
reflecting diffusions in nonsmooth domains with curved boundaries and varying directions of reflection, although these situations come up in applications (see, e.g., Kang, Kelly, Lee and Williams, 2009 or Kang and Williams, 2012). This talk will present some recent, published and unpublished, existence and uniqueness results. We consider semimartingale reflecting diffusions, characterized as solutions of Stochastic Differential Equations with Reflection (SDERs). We obtain existence and uniqueness of the solution in a piecewise smooth, 2-dimensional domain, with a varying direction of reflection on each “side”, under easily verifiable, geometrically meaningful conditions. In the case of a polygon with a constant direction of reflection on each side, our conditions coincide with Dai and Williams’. Moreover we allow for cusps (Costantini and Kurtz, 2018) and for situations where two “sides” meet smoothly but the direction of reflection is discontinuous. We also obtain existence and uniqueness in a d-dimensional domain with one singular point (such as a smooth cone or “horn”), with a varying direction of reflection, under similar assumptions. The keystone of our arguments is a new reverse ergodic theorem for nonhomogeneous, possibly killed, Markov chains (Costantini and Kurtz, 2021), which is used in combination with a result on existence of strong Markov solutions to SDERs (Costantini and Kurtz, 2019).

**Keywords:** Reflecting diffusions, Uniqueness, Ergodic theorem, Cusps

---

**CS11 Topics Related to KPZ Universality**

**Upper tail decay of KPZ models with Brownian initial conditions**

Patrik Ferrari\(^1\) and Balint Vető\(^2\)*

\(^1\)Institute for Applied Mathematics, Bonn University, Germany

\(^2\)Institute of Mathematics, Budapest University of Technology and Economics, Hungary

**Abstract:** We consider the limiting distribution of KPZ growth models with random but not stationary initial conditions introduced by Chhita, Ferrari and Spohn. The one-point distribution of the limit is given in terms of a variational problem. We deduce the right tail asymptotic of the distribution function. This gives a rigorous proof and extends the results obtained by Meerson and Schmidt.

**Keywords:** Kardar-Parisi-Zhang universality class, Growth models, Random initial conditions, Upper tail asymptotic

**Bijective matching between q-Whittaker and periodic Schur measures**

Matteo Mucciconi*

Physics, Tokyo Institute of Technology, Japan

**Abstract:** We report on a combinatorial construction that allows to relate marginal distributions of the q-Whittaker and the periodic Schur measures. The periodic Schur measure is a generalization of the Schur measure introduced by Borodin in 2006 and that models lozenge tilings in a cylindrical domain. Its free fermionic origin yields a nice mathematical structure and its correlations are determinantal. The q-Whittaker measure is another generalization of the Schur measure, which has found application in the rigorous description of KPZ models. Since mathematical properties of q-Whittaker polynomials are much more complicated than the Schur polynomials, the q-Whittaker measure is a more difficult object to handle. Using a bijective combinatorial approach we are able to relate the theories of Schur and q-Whittaker polynomials producing a remarkable correspondence between the two measures. Our arguments pivot around a combination of various theories, which had not yet been used in integrable probability, that include Kirillov-Reshetikhin crystals, Demazure modules, the Box-Ball system or the skew RSK correspondence. The talk is based on collaborations with Takashi Imamura and Tomohiro Sasamoto. Motivations and general ideas of our work are addressed by T. Imamura and applications to probabilistic systems are explained by T. Sasamoto.

**Keywords:** KPZ universality class, Determinantal point process, Combinatorics, Cauchy identities
A new approach to KPZ models by determinantal and Pfaffian measures
Tomohiro Sasamoto*
Physics, Tokyo Institute of Technology, Japan

Abstract: Recently we have established bijectively an identity which relates certain sums of q-Whittaker polynomials and skew Schur polynomials. More precisely we have found that a marginal of the q-Whittaker measure is equivalent to that of the periodic Schur measure. This enables us to study various models in the KPZ universality class by using methods associated with determinantal point processes. The main purpose of this talk is to explain this new approach to KPZ models. By imposing a symmetry on the bijection, we have also found an identity which relates sums which include single q-Whittaker polynomials and skew Schur polynomials. This is in fact related to the KPZ models in half-space and our identity allows us to study such models using Pfaffian measures. In particular we will establish the limiting distributions for models in half-space, which have not been achieved by other approaches.

[The talk is based on collaborations with Takashi Imamura and Matteo Mucciconi. Motivations and general ideas of our work are addressed by T. Imamura and the bijective proofs of identities are explained by M. Mucciconi.]

Keywords: KPZ universality class, Determinantal point process, Pfaffian point processes

Stochastic-uniform-approximations of Wasserstein barycenters
Florian Heinemann*
Institute for Mathematical Stochastics, Georg-August-University Göttingen, Germany

Abstract: Recently, optimal transport and more specifically the Wasserstein distance, have achieved renewed interest as they have been recognized as attractive tools in data analysis. Consequently, this also lead to an increasing interest in Fréchet means, or barycenters, with respect to that distance. These, so called, Wasserstein barycenters offer favorable geometric properties which lend itself well to many applications. However, even more than usual optimal transport, the barycenter problem suffers from a significant computational cost. To alleviate this issue, we propose a hybrid resampling method to approximate finitely supported Wasserstein barycenters on large-scale datasets, which can be combined with any exact solver. Nonasymptotic bounds on the expected error of the objective value as well as the barycenters themselves allow to calibrate computational cost and statistical accuracy. The rate of these upper bounds is shown to be optimal and independent of the underlying dimension, which appears only in the constants. Using a simple modification of the subgradient descent algorithm of Cuturi and Doucet, we showcase the applicability of our method on a myriad of simulated datasets, as well as a real-data example which are out of reach for state of the art algorithms for computing Wasserstein barycenters.

[This is joint work with Axel Munk and Yoav Zemel.]

Keywords: Optimal Transport, Wasserstein Distance, Wasserstein Barycenter, Resampling

Measuring dependence between random vectors via optimal transport
Gilles Mordant and Johan Segers*
LIChAM/ISBA, UCLouvain, Belgium

Abstract: To quantify the dependence between two random vectors of possibly different dimensions, we propose to rely on the properties of the 2-Wasserstein distance. We first propose two coefficients that are based on the Wasserstein distance between the actual distribution and a reference distribution with independent components.
The coefficients are normalized to take values between 0 and 1, where 1 represents the maximal amount of
dependence possible given the two multivariate margins. We then make a quasi-Gaussian assumption that yields two
additional coefficients rooted in the same ideas as the first two. These different coefficients are more amenable for
distributional results and admit attractive formulas in terms of the joint covariance or correlation matrix. Furthermore,
maximal dependence is proved to occur at the covariance matrix with minimal von Neumann entropy given the
covariance matrices of the two multivariate margins. This result also helps us revisit the RV coefficient by proposing a
sharper normalisation. The two coefficients based on the quasi-Gaussian approach can be estimated easily via the
empirical covariance matrix. The estimators are asymptotically normal and their asymptotic variances are explicit
functions of the covariance matrix, which can thus be estimated consistently too. The results extend to the Gaussian
copula case, in which case the estimators are rank-based. The results are illustrated through theoretical examples,
Monte Carlo simulations, and a case study involving electroencephalography data.

**Keywords:** Optimal transport, Bures-Wasserstein distance, RV coefficient, Gaussian copula

**Transportation duality and reverse functional inequalities for Markov kernels**

Nathaniel Eldredge\(^1\) and Fabrice Baudoin\(^2\)

\(^1\)Mathematical Sciences, University of Northern Colorado, USA

\(^2\)Department of Mathematics, University of Connecticut, USA

**Abstract:** Functional inequalities for a Markov semigroup \(P_t\), which may express its "smoothing" properties, can also
be studied in terms of the dual action of \(P_t\) on the space of probability measures. These can give rise to "contraction"
inequalities in terms of various distances between probability measures, such as the Wasserstein or Hellinger
distances. I will discuss results for the reverse Poincaré and reverse log Sobolev inequalities, which turn out to have
dual formulations to which they are actually equivalent. Applications to Markov processes include rates of
convergence to equilibrium, smoothness of transition densities, and quasi-invariance properties.

**Keywords:** Reverse Poincaré inequality, Reverse log Sobolev inequality, Optimal transport, Markov processes

---

**Universal phenomena for random constrained permutations**

Jacopo Borga\(^*\)

Institute of Mathematics, University of Zurich, Switzerland

**Abstract:** How do local/global constraints affect the limiting shape of random permutations? This is a classical
question that has received considerable attention in the last 15 years. In this talk we give an overview of some recent
results on this topic, mainly focusing on random pattern-avoiding permutations. We first introduce a notion of scaling
limit for permutations, called permutons. Then we present some recent results that highlight certain universal
phenomena for permuton limits of various families of pattern-avoiding permutations. These results will lead us to the
definition of three remarkable new limiting random permutons: the "biased Brownian separable permuton", the
"Baxter permuton" and the "skew Brownian permuton". We finally discuss some recent results that show how
permuton limits are useful to investigate the behaviour of certain statistics on random pattern-avoiding permutations,
such as the length of the longest increasing subsequence.

**Keywords:** Random constrained permutations, Scaling limits and permutons, Universality

---

**CS13 Random Structures**

JULY 21 (WED), 21:30 - 22:00 KST
The scaling limit of the strongly connected components of a uniform directed graph with an i.i.d. degree sequence
Serte Donderwinkel* and Zheneng Xie
Department of Statistics, University of Oxford, UK

Abstract: We consider the strongly connected components (SCC) of a uniform directed graph on \( n \) vertices with i.i.d. degree tuples distributed as \((D^-, D^+)\), with \( E[D^+] = E[D^-] = \mu \). We condition on the total in-degree and total out-degree to be equal. A phase transition for the emergence of a giant strongly connected component, that contains a positive proportion of the vertices, is known to occur at the critical value \( \frac{E[D^-] + E[D^+]}{\mu} + 1 \). We study the model at this critical value, and additionally, require that \( E[(D^-)^i(D^+)^j] < \infty \) for all \( i + j \leq 3 \), and for \( \{i,j\} = \{1,3\} \). We show that, under these conditions, the strongly connected components ranked by decreasing order of size and rescaled by \( n^{-1/3} \), converge in distribution to a sequence of finite strongly connected directed multigraphs with edge lengths which are either 3-regular or loops. The limit objects are in a 3-parameter family, which contains the scaling limit of the SCC in the directed Erdős-Rényi model at criticality as found in the work of Goldschmidt and Stephenson in 2019. This shows that the limit object is universal. We use a metric on the space of such multigraphs in which two multigraphs are close if there are compatible isomorphisms between their vertex and edge sets which roughly preserves the edge lengths. We use the product topology on the sequence of multigraphs. Our method involves depth-first exploration of the directed graph, resulting in a spanning forest, of which we study the limit under rescaling.

Keywords: Directed graphs, Configuration model, Scaling limits, Critical random graphs

Spherical principal curves
Jongmin Lee¹ and Jang-Hyun Kim²
¹Statistics, Seoul National University, Korea
²Computer Science, Seoul National University, Korea

Abstract: This paper presents a new approach for dimension reduction of data observed on spherical surfaces. Several dimension reduction techniques have been developed in recent years for non-Euclidean data analysis. As a pioneer work, Hauberg (2016) attempted to implement principal curves on Riemannian manifolds. However, this approach uses approximations to process data on Riemannian manifolds, resulting in distorted results. This study proposes a new approach to project data onto a continuous curve to construct principal curves on spherical surfaces. Our approach lies in the same line of Hastie and Stuetzle (1989) that proposed principal curves for data on Euclidean space. We further investigate the stationarity of the proposed principal curves that satisfy the self-consistency on spherical surfaces. The results on the real data analysis and simulation examples show promising empirical characteristics of the proposed approach.

Keywords: Dimensionality reduction, Feature extraction, Manifold, Principal curve

A multi-species Ehrenfest process and its diffusion approximation
Serena Spina*, Antonio Di Crescenzo and Barbara Martinucci
University of Salerno

Abstract: The celebrated Ehrenfest model is a Markov chain proposed to describe the diffusion of gas molecules in a
container. Our aim is to generalize this model by considering a multi-type Ehrenfest process on a star graph. The considered model results in a continuous-time stochastic process describing the dynamics of an evolutionary system that can accommodate \( N \) particles and is characterized by \( d \) evolution classes, represented with \( d \) semiaxis joined at the origin. The evolution of the stochastic process over each line evolves as a classical Ehrenfest model with suitable linear transition rates, moreover, after visiting the origin, the process can move toward any semiaxis with different rates, depending on the elements of a stochastic matrix. We investigate the dynamics of this process making use of a probability generating function-based approach. This leads to the determination of the transient transition probabilities (in closed form for a particular choice of the parameter), and of the asymptotic distribution, in general. In addition, we obtain some results on the asymptotic mean, variance, coefficient of variation for the process. We also consider a continuous approximation of the process, which leads to an Ornstein-Uhlenbeck diffusion process evolving on a spider-shaped continuous state space formed by \( d \) semiaxis of infinite length joined at the origin; the origin of the given domain constitutes the equilibrium point of the system. We determine the expression of the asymptotic probability distribution for each ray of the spider. Finally, we compare the discrete process with the diffusion process in order to show the goodness of the continuous approximation.

**Keywords:** Birth-death process, Diffusion process, Star graph, Gauss hypergeometric function

---

**Limit theorems for the realised semicovariances of multivariate Brownian semistationary processes**

Yuan Li*, Mikko Pakkanen and Almut Veraart
Mathematics, Imperial College London, UK

**Abstract:** In this talk we will introduce the realised semicovariance, which is resulted from the decomposition of the realised covariance matrix into components based on the signs of the returns, and study its in-fill asymptotic properties of multivariate Brownian semistationary (BSS) processes. The realised semicovariance is originally proposed in Bollerslev et al. (2020, Econometrica) where they worked on semimartingale settings. We extend their work to BSS processes, which are not necessarily semimartingales. More precisely, a weak convergence in the space of càdlàg functions endowed with the Skorohod topology for the realised semicovariance of a general Gaussian process with stationary increments is proved first. The methods are based on quantitative Breuer-Major theorems and on moment bound for sums of products of Gaussian vector’s functions. Furthermore, we demonstrate the corresponding stable convergence. Finally, a weak law of large numbers and a central limit theorem for the realised semicovariance of multivariate BSS processes are established. These results extend the limit theorems for the realised covariation to a version for the non-linear functionals.

**Keywords:** Realised semicovariance, Multivariate Brownian semistationary process, Central limit theory, Malliavin calculus

---

**A Yaglom type asymptotic result for subcritical branching Brownian motion with absorption**

Jiaqi Liu*
University of California, San Diego

**Abstract:** In this talk, we will consider a slightly subcritical branching Brownian motion with absorption, where particles move as Brownian motion with drift \(-\sqrt{2} + 2\epsilon\), undergo dyadic fission at rate 1, and are killed upon hitting the origin. We are interested in the asymptotic behaviors of the process conditioned on survival up to a large time \( t \) as the process approaches criticality. Results like this are called Yaglom type results. Specifically, we will talk about the long run expected number of particles conditioned on survival as the process approaches to being critical.

**Keywords:** Branching Brownian motion, Yaglom limit laws, Near critical phenomena
Coexistence of localized Gibbs measures and delocalized gradient Gibbs measures on trees
Florian Henning* and Christof Kuelske
Faculty of Mathematics, Ruhr University Bochum, Germany

Abstract: In statistical mechanics, Gibbs measures for a spin system (a stochastic process) indexed by a countable graph under the influence of an interaction potential describe equilibrium distributions. They are defined in terms of being compatible with the Gibbsian specification associated with the potential, a system of prescribed conditional distributions built from the potential. In case of an unbounded local spin space, existence of Gibbs measures does not directly follow from compactness arguments. In this talk we focus on the situation where the underlying graph is a regular tree, spins take values in the integers (or an integer lattice) and the interaction potential is spatially homogeneous and of gradient type, i.e., depends only on the difference of spins values at neighboring sites. We provide general conditions in terms of the relevant $p$-norms of the associated transfer operator $\Gamma$ (the exponential of the interaction potential) which ensure the existence of a countable family of spatially homogeneous Gibbs measures, describing localization at different heights. Next we prove existence of spatially homogeneous gradient Gibbs measures, describing increments of spin values along the edges of the tree. We construct these gradient Gibbs measures in terms of an edge-wise independent resampling process for $\mathbb{Z}^q$-valued Gibbs measures for a suitable transformed fuzzy transfer operator $Q^q$. Then we prove that they are delocalized. Finally, we show that the two conditions on $Q$ can be fulfilled at the same time, which then implies coexistence of both types of measures.

[The talk is based on joint work with Christof Kuelske, which is accepted for publication in the Annals of Applied Probability.]


Keywords: Gibbs measures, Gradient Gibbs measures, Regular tree, Heavy tails

Inhomogeneous gradient Gibbs measures on regular trees with homogeneous interactions
Christof Kuelske* and Florian Henning
Mathematics, Ruhr-University Bochum, Germany

Abstract: It is known that some statistical mechanics models with homogeneous interactions on regular lattices may admit inhomogeneous infinite-volume states. A famous example for this phenomenon are the Dobrushin-states for the Ising model which lack translation-invariance in three or more lattice dimensions. We investigate whether states which lack translation-invariance also exist on regular trees for $\mathbb{Z}^q$-valued spin models with nearest-neighbor gradient interactions. Our analysis includes the SOS-model and the discrete Gaussian, which are important models of mathematical statistical mechanics, where they are mostly studied on the lattice. We show that, under rather general assumptions on the interaction, such inhomogeneous gradient states do exist. Our proof uses probabilistic methods in close combination with dynamical systems methods. In a first part we extend the probabilistic approach of our earlier work ("Coexistence of localized Gibbs measures and delocalized gradient Gibbs measures on trees", to be published in the Annals of Applied Probability). This allows to draw a relation between the gradient Gibbs states we are aiming at, and the Gibbs states of certain internal $q$-state spin-models with discrete rotation symmetry, which holds also for inhomogeneous states. In a second part we investigate these $q$-spin models on the regular tree via their associated discrete dynamical systems. The proofs of existence and lack of translation invariance of infinite-volume gradient states are then specifically based on properties of the local pseudo-unstable manifold of the corresponding discrete dynamical systems of these internal models, around the free state, at large $q$.

Reference: arXiv:2102.11899, Existence of gradient Gibbs measures on regular trees which are not translation
Statistical mechanical model of adsorption at the surface interface contacting with an ideal gas
Changho Kim¹*, Alejandro Garcia² and John Bell³
¹Applied Mathematics, University of California, Merced, USA
²Physics, San Jose State University, USA
³Computational Research Division, Lawrence Berkeley National Laboratory, USA

Abstract: We develop a statistical mechanical model for an ideal gas interfaced with a lattice surface where adsorption and desorption of gas particles occur. While this type of model has been investigated, we revisit it for the development of a thermodynamically consistent particle-continuum hybrid model for stochastic simulations of gas-solid interfacial systems as described below. Following the Langmuir adsorption model, we assume that the mean adsorption rate is proportional to the mean impingement rate of gas particles onto the surface and the mean desorption rate is given as a function of surface temperature. As a result, thermodynamic equilibrium is expected to be established for a given pressure of the ideal gas and temperature of the system. By investigating the detailed balance conditions, we derive the equilibrium fluctuational properties of the ideal gas state and surface coverage. We consider several velocity models, including the spectacular reflection and thermal wall models, from which the velocity of each desorbed or colliding particle is drawn. Based on this statistical mechanical model, we find how the ideal gas state and surface coverage after a finite short time delta t can be updated using the adsorption and desorption counts during delta t. For the momentum and energy update, we confirm that the same thermodynamic equilibrium is established whether adsorption and desorption are considered. The resulting time update model provides essential information on how to construct a particle-continuum hybrid model, where positions of all adsorbed particles are tracked whereas only the aggregated information such as the total mass, momentum, and energy densities are tracked in the gas. We present preliminary simulation results of the particle-continuum hybrid simulation method and demonstrate the importance of using a thermodynamically consistent statistical mechanical model.

Keywords: Adsorption, Ideal gas, Statistical mechanics, Detailed balance
bound less than 1 for $\rho'(\cdot,1)$ cannot simply be deleted altogether from the theorem, even in the case of strict stationarity.

**Keywords:** Central limit theorems, Random fields, Strong mixing, Lindeberg condition

---

**On the law of the iterated logarithm and strong invariance principles in stochastic geometry**

Johannes Krebs*

Institute of Applied Mathematics, Heidelberg University, Germany

**Abstract:** The law of the iterated logarithm (LIL) dates back to the contributions of Khinchin (1924) and Kolmogorov (1929). Since then it has been recovered in various settings in probability theory and statistics. The nowadays more general statement was proved by Hartman and Wintner (1941). Let $(X_i)_{i \in \mathbb{N}}$ be independent and identically distributed with mean zero and unit variance. Then

$$\limsup_{n \to \infty} \frac{1}{n} \log \log n \sum_{i=1}^{n} X_i = 1 \quad \text{a.s.}$$

In this contribution we study the LIL for so-called stabilizing functionals $H$ in stochastic geometry. The idea of stabilizing functionals dates back to Lee (1997, 1999) and Penrose and Yukich (2001, 2003). The underlying point process $\Xi_n$ is either a homogeneous Poisson process observed on increasing sample windows or a coupled binomial process. Given a suitable assumption on the rate of stabilization and appropriate moment conditions, we show

$$\limsup_{n \to \infty} \frac{1}{n} \log \log n \sum_{i=1}^{n} X_i = 1 \quad \text{a.s.},$$

where the limiting variance $\sigma^2 = \lim_{n \to \infty} \frac{1}{n} \text{Var}(H(X_n))$ depends on the point process and is positive under mild conditions. Moreover, apart from recovering the LIL, we derive a strong invariance principle for the functional $H$ in a set-up which is a little more restrictive concerning the sample windows of the Poisson process.

**Keywords:** Law of the iterated logarithm, Stabilizing functionals, Stochastic geometry, Strong invariance principles

---

**Functional limit theorems for U-statistics**

Christian Döbler1, Mikolaj J. Kasprzak2* and Giovanni Peccati2

1Department of Mathematics, Heinrich Heine University Düsseldorf, Germany

2Department of Mathematics, University of Luxembourg, Luxembourg

**Abstract:** I will discuss a number of results obtained through three pieces of joined work with Christian Döbler and Giovanni Peccati. Firstly, I will talk about sequences of U-processes based on symmetric kernels of a fixed order that may depend on the sample size and present analytic sufficient conditions under which they converge to a linear combination of time-changed Brownian Motions. I will show how these conditions may be applied to deduce functional convergence of quadratic estimators in certain non-parametric models. Secondly, I will present quantitative bounds on the rate of functional convergence of vectors of weighted degenerate U-statistics to time-changed Brownian Motion, obtained via Stein’s method of exchangeable pairs. Finally, I will discuss a multivariate functional version of de Jong’s CLT, yielding that, given a sequence of vectors of degenerate U-statistics, the corresponding empirical processes on $[0,1]$ weakly converge in the Skorohod space as soon as their fourth cumulants in $t=1$ vanish asymptotically and a certain strengthening of the Lindeberg-type condition is verified.

**Keywords:** U-statistics, Functional limit theorems, Stein’s method, Non-parametric statistics

---

**Proving Liggett’s FCLT via Stein’s method**

Wasamon Jantai*

Department of Mathematics, Oregon State University, USA

**Abstract:** In 1990, A. D. Barbour extended Stein’s method for approximation by Gaussian processes, including
Brownian bridge. In this work, we rederive Liggett's functional central limit theorem (FCLT) using Barbour's approach.

**Keywords:** Stein's method, Functional central limit theorem, Brownian bridge, Exchangeability

---

**Covariance networks for functional data on multidimensional domains**

Soham Sarkar* and Victor Panaretos

Institute of Mathematics, Ecole Polytechnique Federale de Lausanne, Switzerland

**Abstract:** Covariance estimation is ubiquitous in functional data analysis. Yet, the case of functional observations over multidimensional domains introduces computational and statistical challenges, rendering the standard methods effectively inapplicable. To address this problem, we introduce Covariance Networks (CovNet) as a modeling and estimation tool. The CovNet model is universal – it can be used to approximate any covariance up to desired precision. Moreover, the model can be fitted efficiently to the data and its neural network architecture allows us to employ modern computational tools in the implementation. The CovNet model also admits a closed-form eigen-decomposition, which can be computed efficiently, without constructing the covariance itself. This facilitates easy storage and subsequent manipulation in the context of the CovNet. Moreover, we establish consistency of the proposed estimator and derive its rate of convergence. The usefulness of the proposed method is demonstrated using an extensive simulation study.

**Keywords:** FDA, Neural Network, Nonparametric Model, Universal Approximation

---

**Large dimensional sample covariance matrices with independent columns and diagonalizable simultaneously population covariance matrices**

Tianxing Mei*, Chen Wang and Jianfeng Yao

The University of Hong Kong

**Abstract:** We consider the limiting behavior of empirical spectral distribution (ESD) of a sample covariance matrix for independent but not necessary identical distributed samples with their corresponding population covariance matrices diagonalizable simultaneously asymptotically, when dimension of samples grows proportionally with the sample size. The existing works of different types of sample covariance matrices, including the weighted sample covariance matrix, the centered Gram matrix model and that of linear times series models with diagonalizable simultaneous coefficient matrices, can be covered by our approach. As applications, we obtain the existence and uniqueness of the limiting spectral distribution (LSD) of realized covariance matrix for a multidimensional diffusion process with its co-volatility process equipped with an anisotropic time-varying spectrum. Meanwhile, for a matrix-valued autoregressive model, we derive the common limiting spectral distribution of the sample covariance matrix for each matrix-valued observation when both row and column dimension are large.

**Keywords:** Large sample covariance matrix, Eigenvalue distribution, Realized covariance matrix, Matrix-valued autoregressive model

---

**Random surface covariance estimation by shifted partial tracing**

Tomas Masak*

Institut de Mathématiques, EPFL, Switzerland
Abstract: The problem of covariance estimation for replicated surface-valued processes is examined from the functional data analysis perspective. Considerations of statistical and computational efficiency often compel the use of separability of the covariance, even though the assumption may fail in practice. We consider a setting where the covariance structure may fail to be separable locally - either due to noise contamination or due to the presence of a non-separable short-range dependent signal component. That is, the covariance is an additive perturbation of a separable component by a non-separable but banded component. We introduce non-parametric estimators hinging on the novel concept of shifted partial tracing, enabling computationally efficient estimation of the model under dense observation. Due to the denoising properties of shifted partial tracing, our methods are shown to yield consistent estimators even under noisy discrete observation, without the need for smoothing. Further to deriving the convergence rates and limit theorems, we also show that the implementation of our estimators, including for the purpose of prediction, comes at no computational overhead relative to a separable model. Finally, we demonstrate empirical performance and computational feasibility of our methods in an extensive simulation study and on a real data set.

[This is a joint work with Victor M. Panaretos.]

Keywords: Separability, Bandedness, Functional data

Detection of outliers in compositional data on disabled people in the São Paulo State

Paulo Oliveira*

EESC, USP, Brazil

Abstract: Outliers are observations that, for some reason, differ from the other observations belonging to the data set. In univariate and bivariate data sets, outliers can be detected analyzing the scatter plot. Observations distant from the cloud formed by the data set are considered unusual. In multivariate data sets, the detection of outliers using graphics is more difficult because we have to analysis a couple of variables each time, which results is a long and less reliable process because we can find an observation that is unusual for one variable and not unusual for the others, masking the results. Compositional data are vectors, called compositions, whose components are all positive, it satisfies the sum equal one and has a Simplex space. The sum constraint induces the correlation between the components and this requires that the statistical methods for the analysis of datasets consider this fact. The theory for compositional data was developed mainly by Aitchison in the 1980s, and since then, several techniques and methods have been developed for compositional data modelling. Disabled Person is any person who presents loss or abnormality of a psychological or anatomical structure or function that generates incapacity for the performance of activities, that is, they have different characteristics from most people who are part of society and these characteristics make it difficult to their social inclusion. Disabilities can be permanent or temporary and limit the ability to perform one or more activities such as seeing, listening, walking and intellectual. It is characterized as a complex multidimensional experience and imposes several measurement challenges. Worldwide, disabled people have worse health prospects, lower levels of education, lower economic participation and higher poverty rates compared to people without disabilities. This is partly due to the fact that disabled people face barriers to access services that many of us have long considered guaranteed, such as health, education, employment. In statistical terms, were considered data from 3681111 respondents from the complete questionnaire of the IBGE (Statistics and Geography Brazilian Institute) census aggregated in 645 municipalities in the State of São Paulo, Brazil, considering as variables the 16 levels of disabled people with and the following methods were used to detect multivariate outliers' detection: 95% confidence ellipse based on the first two main components; The Forward Search; Based on MCD (Minimum Covariance Determinant). and finally; Based on MED (Mas Eigen Difference) for comparative study between outlier detection performance by different methods.
Keywords: Outlier data, Compositional data, Disabled people, IBGE census

Consistent change-point detection for general distributions
Florence Leonardi*
Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

Abstract: We propose a method based on regularized maximum likelihood for change point detection of general multivariate distributions under independent sampling. We show that the estimator is consistent and almost surely recovers the set of change-points under usual and easy to verify conditions. These conditions apply to a large variety of models, such as categorical or normal random variables and finite-state Markov chains. We also show that we can efficiently compute the estimator through a dynamic programming algorithm under a decomposable penalty term. [This is joint work with Lucas Prates de Oliveira.]

Keywords: Model selection, Regularized maximum likelihood, Structure estimation

Change point detection under linear model: Use of MOSUM approach
Joonpyo Kim1* and Haeran Cho2
1Department of Statistics, Seoul National University, Korea
2Institute for Statistical Science, University of Bristol, UK

Abstract: This talk presents a new detection method for structural change points based on a MOSUM approach under a piecewise linear model. Most existing methods focus on mean changes, assuming that the underlying model is piecewise constant. However, this stringent assumption cannot be applicable to many real-world processes, such as manufacturing. The proposed method significantly extends the scope of the change points structure by employing the linear regression model so that it is capable of identifying slope changes or smoothness of the processes beyond their mean changes. The proposed method is computationally efficient and easily used for real-time detection due to the inherent feature of the moving window approach. Furthermore, some theoretical properties of the proposed change point estimator are investigated. Results from the real data analysis and simulation examples show the promising empirical performance of the proposed method.

Keywords: Change Points, MOSUM, Piecewise Linear Model

Interval-censored least-squares regressions
Taehwa Choi* and Sangbum Choi
Department of Statistics, Korea University, Korea

Abstract: This article suggests the linear regression model under interval-censored data, where exact event times are unobserved but fall in observed censoring intervals. It is commonly arisen in longitudinal studies such as breast cosmesis data, where periodic monitoring is progressed to check the patient clinical status. Many of previous researches has been mostly focused on probability-based methods such as Cox and transformation models in terms of both theoretical and practical approaches. In contrast, there has not been received much attention on accelerated failure time model, despite direct interpretation on event time is possible. In this article, we generalize the Buckley-James method to explain the accelerated lifetime effects under the interval-censored data. Coupled with regression estimating procedure, a novel EM-algorithm for nonparametric likelihood estimation is devised for nuisance function parameter. Asymptotic behaviors are established, along with slower rate of convergence for nuisance function parameter due to absence of exact data. Simulation studies demonstrate the finite sample performance, and the
method is applied to the real data to illustrate the practical usage.

**Keywords:** Accelerated lifetime, Buckley-James method, Interval censoring, Nonparametric likelihood

---

**Estimation of multivariate generalized gamma convolutions through Laguerre expansions**

Oskar Laverty\(^1\)*, Véronique Maume-Deschamps\(^1\), Esterina Masiello\(^1\) and Didier Rullière\(^2\)

\(^1\)Institut Camille Jordan, Université Claude Bernard Lyon 1, France

\(^2\)Laboratoire Sciences Actuarielles et Financières, Université Claude Bernard Lyon 1, France

**Abstract:** The generalized gamma convolution class of distribution appeared in Thorin’s work while looking for the infinite divisibility of the log-Normal and Pareto distributions. Although these distributions have been extensively studied in the univariate case, the multivariate case and the dependence structures that can arise from it have received little interest in the literature. Furthermore, only one projection procedure for the univariate case was recently constructed, and no estimation procedure are available. By expending the densities of multivariate generalized gamma convolutions into a tensorized Laguerre basis, we bridge the gap and provide performant estimations procedures for both the univariate and multivariate cases. We provide some insights about performance of these procedures, and a convergent series for the density of multivariate gamma convolutions, which is shown to be more stable than Moschopoulos’s and Mathai’s univariate series. We furthermore discuss some examples.

---

**Copula-based Markov zero-inflated count time series models**

Mohammed Alqawba\(^1\)* and Nouro Diawara\(^2\)

\(^1\)Mathematics, Qassim University, Saudi Arabia

\(^2\)Mathematics and Statistics, Old Dominion University, USA

**Abstract:** Count time series data with excess zeros are observed in several applied disciplines. When these zero-inflated counts are sequentially recorded, they might result in serial dependence. Ignoring the zero-inflation and the serial dependence might produce inaccurate results. In this paper, Markov zero-inflated count time series models based on a joint distribution on consecutive observations are proposed. The joint distribution function of the consecutive observations is constructed through copula functions. First and second order Markov chains are considered with the univariate margins of zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), or zero-inflated Conway-Maxwell-Poisson (ZICMP) distributions. Under the Markov models, bivariate copula functions such as the bivariate Gaussian, Frank, and Gumbel are chosen to construct a bivariate distribution of two consecutive observations. Moreover, the trivariate Gaussian and max-infinitely divisible copula functions are considered to build the joint distribution of three consecutive observations. Likelihood based inference is performed and asymptotic properties are studied. To evaluate the estimation method and the asymptotic results, simulated examples are studied. The proposed class of models are applied to sandstorm counts example. The results suggest that the proposed models have some advantages over some of the models in the literature for modeling zero-inflated count time series data.

**Keywords:** Copula, Integer-valued time series, Zero-inflation
Bi-factor and second-order copula models for item response data
Sayed H. Kadhem* and Aristidis K. Nikoloulopoulos
School of Computing Sciences, University of East Anglia, UK

Abstract: Bi-factor and second-order models based on copulas are proposed for item response data, where the items can be split into non-overlapping groups such that there is a homogeneous dependence within each group. Our general models include the Gaussian bi-factor and second-order models as special cases and can lead to more probability in the joint upper or lower tail compared with the Gaussian bi-factor and second-order models. Details on maximum likelihood estimation of parameters for the bi-factor and second-order copula models are given, as well as model selection and goodness-of-fit techniques. Our general methodology is demonstrated with an extensive simulation study and illustrated for the Toronto Alexithymia Scale. Our studies suggest that there can be a substantial improvement over the Gaussian bi-factor and second-order models both conceptually, as the items can have interpretations of latent maxima/minima or mixtures of means in comparison with latent means, and in fit to data.

Keywords: Bi-factor model, Conditional independence, Second-order model, Truncated vines

Probabilistic principal curves on Riemannian manifolds
Seungwoo Kang*
Department of Statistics, Seoul National University, Korea

Abstract: This paper studies a new curve fitting approach that is useful for representation and dimension reduction of data on Riemannian manifolds. In this study, we extend the probabilistic formulation of the curve passing through the middle of data on Euclidean space by Tibshirani (1992) to Riemannian symmetric space. To this end, we define a principal curve based on a mixture model for observations and unobserved latent variables, and propose a new algorithm to estimate the principal curve for given data points on Riemannian manifolds using a series of procedures in ‘unrolling, unwrapping, and wrapping’ and EM algorithm. Some properties for justification of the estimation algorithm are further investigated. Results from numerical examples, including several simulation sets on hyperbolic space, sphere, special orthogonal group, and a real data example, demonstrate the promising empirical properties of the proposed probabilistic approach.

Keywords: Dimensionality reduction, EM algorithm, Principal curve, Riemannian manifold

The elastic information criterion for multicollinearity detection
Kimon Ntotsis¹*, Alex Karagrigoriou¹ and Andreas Artemiou²
¹Statistics and Actuarial-Financial Mathematics, University of the Aegean, Greece
²School of Mathematics, Cardiff University, UK

Abstract: When it comes to factor interpretation, multicollinearity is among the biggest issues that must be surmounted especially in this new era, of Big Data Analytics. Since even moderate size multicollinearity can prevent a proper interpretation, special diagnostics must be recommended and implemented for identification purposes. In this work, we propose the Elastic Information Criterion which is capable of capturing multicollinearity accurately and effectively without factor over-elimination. The performance in simulated and real numerical studies is demonstrated.

Keywords: Multicollinearity Detection, Elastic Net Regularization, Dimensionality Reduction, Coefficient Penalization
A bidimensional shock model driven by the space-fractional Poisson process
Alessandra Meoli* and Antonio Di Crescenzo
Department of Mathematics, Università degli Studi di Salerno, Italy

Abstract: We describe a competing risks model within the framework of bivariate random shock models, this being of great interest in reliability theory. Specifically, we assume that a system or an item fails when the sum of shocks of type 1 and of type 2 reaches a random threshold that takes values in the set of natural numbers. The two kinds of shock occur according to a bivariate space-fractional Poisson process, which is a time-changed bivariate homogeneous Poisson process where the time change is an independent stable subordinator. We obtain the failure densities, the survival function and other related quantities. In this way we generalize some results in the literature, which can be recovered when the index of stability characterizing the bivariate space-fractional Poisson process is equal to 1.

Keywords: Competing risks, Shock models, Fractional calculus

Bayesian and stochastic modeling of polysomnography data from children using pacifiers for improved estimation of the apnea-hypopnea index
Sujay Datta*
Department of Statistics, University of Akron (Ohio), USA

Abstract: Polysomnography is an overnight systematic procedure to collect physiological parameters during sleep. It is considered as gold-standard for diagnosing sleep-related disorders. It takes several days to score and interpret the raw data from this study and confirm a diagnosis of, say, Obstructive Sleep Apnea (OSA) – a potentially dangerous disorder. The presence of artifacts (anomalies created by a malfunctioning sensor) makes scoring even more difficult, potentially resulting in misdiagnosis. It is common to see airflow signal artifacts in infants that use a pacifier during sleep. The act of sucking on the pacifier causes artifacts in the oro-nasal sensor (thermistor) used to monitor airflow during respiration. The resulting inaccurate scoring leads to an under-estimation of the Apnea Hypopnea Index (AHI) — the basis for a formal OSA diagnosis. So researchers are now exploring two other information sources (blood oxygen saturation readings from a pulse-oximeter and occurrence of arousal events) to supplement the artifact-corrupted thermistor data. They first look for statistical association between the thermistor and the pulse-oximeter/arousal data and then statistically predict a modified AHI score using the latter whenever the former is corrupt. To our knowledge, no attempt of statistically modeling these three data-sources to bring out their association currently exists. This project aims at developing several competing probabilistic models for these data-types and then checking how strongly they bring out the association by applying them on archived data from the Akron Children’s Hospital. These modeling approaches are a significant statistical contribution to this important medical problem. After performing some statistical tests for association, the modeling approaches will include naïve Bayes, Beta-Binomial, correlated homogeneous Poisson processes and double-chain Markov models. The resulting improvement in AHI estimates is demonstrated using data-sets from a sample of non-pacifier users after artificially discarding part of their thermistor data (as if they were artifact-corrupted).

Keywords: Sleep apnea, Oro-nasal airflow, Bayes formula, Stochastic modeling
Asymmetric prior in wavelet shrinkage

Alex Rodrigo dos Santos Sousa*
Statistics, University of São Paulo, Brazil

Abstract: In bayesian wavelet shrinkage, the already proposed priors to wavelet coefficients are assumed to be symmetric around zero. Although this assumption is reasonable in many applications, it is not general. The present paper proposes the use of an asymmetric shrinkage rule based on the discrete mixture of a point mass function at zero and an asymmetric beta distribution as prior to the wavelet coefficients in a non-parametric regression model. Statistical properties such as bias, variance, classical and bayesian risks of the associated asymmetric rule are provided and performances of the proposed rule are obtained in simulation studies involving artificial asymmetric distributed coefficients and the Donoho-Johnstone test functions. Application in a seismic real dataset is also analyzed. In general, the asymmetric shrinkage rule outperformed classical symmetric rules both in simulation and real data application.

Keywords: Wavelet shrinkage, Nonparametric regression, Beta prior distribution, Bayesian statistics

Semiparametric Bayesian regression analysis of multi-typed matrix-variate responses

Inkoo Lee1* and Dipankar Bandyopadhyay2
1Department of Statistics, Rice University, USA
2Department of Biostatistics, Virginia Commonwealth University, USA

Abstract: Complex data such as tensor and multiple types of responses can be found in dental medicine. Dental hygienists measure triple biomarkers at 28 teeth and 6 tooth-sites for each participant. These data have challenging characteristics: 1) binary and continuous responses with skewness, 2) matrix-variate responses for each biomarker have heavy tails, 3) pattern for missing teeth is not random. To circumvent these difficulties, we propose a joint model of multiple types of matrix-variate responses via latent variables. The model accommodates skewness in continuous responses. This statistical framework incorporates exponential factor copula models to capture heavy-tail dependence and asymmetry. Since the number of existing teeth presents the magnitude of periodontal disease (PD) we model the missing mechanism. Our method also guarantees posterior consistency under suitable priors. We illustrate the substantial advantages of our method over alternatives through simulation studies and the analysis of PD data.

Keywords: Exponential factor copula models, Matrix-variate responses, Multiple response-types, Skewness

Bayesian phylogenetic inference of stochastic block models on infinite trees

Wenjian Liu*
Department of Mathematics and Computer Science, Queensborough Community College, CUNY, USA

Abstract: This talk involves a classification problem on a deep network, by considering a broadcasting process on an infinite communication tree, where information is transmitted from the root of the tree to all the vertices with certain probability of error. The information reconstruction problem on an infinite tree, is to collect and analyze massive data samples at the nth level of the tree to identify whether there is non-vanishing information of the root, as n goes to infinity. Its connection to the clustering problem in the setting of the stochastic block model, which has wide applications in machine learning and data mining, has been well established. For the stochastic block mode, an "information theoretically solvable but computationally hard" region, or say "hybrid-hard phase", appears whenever the reconstruction bound is not tight of the corresponding reconstruction on the tree problem. Inspired by the recently proposed $q_1 + q_2$ stochastic block model, we try to extend the classical works on the Ising model and the Potts model, by studying a general model which incorporates the characteristics of both Ising and Potts through...
different in-community and out-community transition probabilities, and rigorously establishing the exact conditions for reconstruction.

**Keywords:** Phase transition, Kesten-Stigum reconstruction bound, Markov random fields on trees, Distributional recursion, Nonlinear dynamical system

### Order-restricted Bayesian inference for the simple step-stress accelerated life tests
David Han* and Crystal Wiedner
Management Science and Statistics, The University of Texas at San Antonio, USA

**Abstract:** In this work, we investigate the order-restricted Bayesian estimation for a simple step-stress accelerated life tests. Based on the three-parameter gamma distribution as a conditional prior, we ensure that the failure rates increase as the stress level increases. In addition, its conjugate-like structure enables us to derive the exact joint posterior distribution of the parameters without a need to perform an expensive MCMC sampling. Upon these distributional results, several Bayesian estimators for the model parameters are suggested along with their individual/joint credible intervals. Through Monte Carlo simulations, the performance of our proposed inferential methods are assessed and compared. Finally, a real engineering case study for analyzing the reliability of a solar lighting device is presented to illustrate the methods developed in this work.

**Keywords:** Accelerated life tests, Bayesian analysis, Progressive censoring

### Bernstein - von Mises type theorem for a scale hyperparameter in Bayesian nonparametric inference
Natalia Bochkina¹* and Vladimir Spokoiny²
¹Mathematics, University of Edinburgh, UK
²Mathematics, Weierstrass Institute and Humboldt University, Germany

**Abstract:** We consider the problem of estimating a smooth function adaptively from a Bayesian perspective in a nonparametric regression model, observed either directly or indirectly. We consider the model in the sequence space, with a smoothing Gaussian prior on the unknown coefficients, and a hyperprior on the prior scale to achieve an adaptive estimator. We show, that under some conditions on the true function, such as self-similarity, the MAP estimator of the scale hyperparameter converges to its oracle value, and the posterior distribution of the scale can be approximated by a Gaussian distribution as the number of observations grows. As far as we are aware, it is the first result of Gaussian approximation of the posterior distribution of a hyperparameter. We will show that this can be interpreted as estimation of the scale parameter from data under model misspecification, and that in the considered setting the posterior variance and the Fisher information of the scale parameter are of different order. We will illustrate these results on an inverse problem with Volterra operator.

**Keywords:** Bernstein-von Mises theorem, Bayesian nonparametrics, Misspecified models

### Convergence of unadjusted Hamiltonian Monte Carlo for mean-field models
Katharina Schuh¹* and Nawaf Bou-Rabee²
¹Institute for Applied Mathematics, University Bonn, Germany
²Department of Mathematical Sciences, Rutgers University Camden, USA
Abstract: In this talk we study the unadjusted Hamiltonian Monte Carlo algorithm applied to high-dimensional probability distributions of mean-field type. We evolve dimension-free convergence and discretization error bounds in Wasserstein distance. These bounds require the discretization step to be sufficiently small, but do not require strong convexity of either the unary or pairwise potential terms present in the mean-field model. To handle high dimensionality, we use a particlewise coupling that is contractive in a complementary particlewise metric. [This talk is based on joint work with Nawaf Bou-Rabee.]
Keywords: Hamiltonian Monte Carlo, Couplings, Convergence to equilibrium, High-dimensional distributions

Nonparametric Bayesian volatility estimation for gamma-driven stochastic differential equations
Peter Spreij*
Korteweg-de Vries Institute, University of Amsterdam, Netherlands

Abstract: We study a nonparametric Bayesian approach to estimation of the volatility function of a stochastic differential equation (SDE) driven by a gamma process. The volatility function is assumed to be positive and Hölder continuous. We show that the SDE admits a weak solution, unique in law. The volatility function is modelled a priori as piecewise constant on a partition of the real line, and we specify a gamma prior on its coefficients. This leads to a straightforward procedure for posterior inference via the Gibbs sampler. We give the contraction rate of the posterior distribution in terms of the Hölder exponent and the sample size. [Joint work with Denis Belomestny, Shota Gugushvili, Moritz Schauer.]
Keywords: Gamma process, Nonparametric Bayesian estimation, Stochastic differential equation

Hamiltonian Monte Carlo in high dimensions
Andreas Eberle*
Institute for Applied Mathematics, University of Bonn, Germany

Abstract: Hamiltonian Monte Carlo (HMC) is a Markov Chain Monte Carlo method that is widely used in applications. It is based on a combination of Hamiltonian dynamics and momentum randomizations. The Hamiltonian dynamics is discretized, and the discretization bias can either be taken into account (unadjusted HMC) or corrected by a Metropolis-Hastings accept-reject step (Metropolis adjusted HMC). Despite its empirical success, until a few years ago there have been almost no convergence bounds for the algorithm. This has changed in the last years where approaches to quantify convergence to equilibrium based on coupling, conductance and hypocoercivity have been developed. In this talk, I will present the coupling approach, and show how it can be used to obtain an understanding of the dimension dependence for unadjusted HMC in several high dimensional model classes. I will also mention some open questions.
Keywords: Couplings of stochastic processes, Mixing time, Markov chain Monte Carlo, Hamiltonian dynamics

Statistical modelling of rainfall time series using ensemble empirical mode decomposition and generalised extreme value distribution
Willard Zvarevashe*
Mathematical Sciences, University of Zululand, South Africa
Abstract: The extreme rainfall patterns have direct and indirect effect on all earths spheres particularly the hydrosphere, biosphere and lithosphere. Therefore, an understanding of the extreme rainfall patterns is very important for future planning and management. In this study, using Western Cape (South African province) as a case study, the rainfall time series is decomposed into intrinsic mode functions (IMFs) using a data adaptive method, ensemble empirical mode decomposition. The IMFs are modelled using generalised extreme value distribution (GEVD). The model diagnosis and selection using QQ-Plot, PP-Plot and Akaike information criterion shows that the decomposed IMFs have better models than the original rainfall time series. The rainfall modelling using decomposed data may assist in future planning and further research by providing better predictions.

Keywords: Climate modelling, Generalised extreme value distribution, Time series analysis, Decomposition

Regularity of multifractional moving average processes with random Hurst exponent
Fabian Mies*, Dennis Loboda and Ansgar Steland
Institute of Statistics, RWTH Aachen University, Germany

Abstract: A recently proposed alternative to multifractional Brownian motion (mBm) with random Hurst exponent is studied, which we refer to as Itô-mBm. It is shown that Itô-mBm is locally self-similar. In contrast to mBm, its pathwise regularity is almost unaffected by the roughness of the functional Hurst parameter. The pathwise properties are established via a new polynomial moment condition similar to the Kolmogorov-Centsov theorem, allowing for random local Hölder exponents. Our results are applicable to a broad class of moving average processes where pathwise regularity and long memory properties may be decoupled, e.g. to a multifractional generalization of the Matérn process.

Keywords: Multifractional Brownian motion, Random Hölder exponent, Matérn process, Self-similarity

High-frequency instruments and identification-robust inference for stochastic volatility models
Md. Nazmul Ahsan1* and Jean-Marie Dufour2
1Economics, Concordia University, Canada
2Economics, McGill University, Canada

Abstract: We introduce a novel class of generalized stochastic volatility (GSV) models, which can utilize and relate many high-frequency realized volatility (RV) measures to the latent volatility. Instrumental variable methods are employed to provide a unified framework for GSV models’ analysis (estimation and inference). We study parameter inference problems in GSV models with nonstationary volatility and exogenous predictors in the latent volatility process. We develop identification-robust methods for joint hypotheses involving the volatility persistence parameter and the composite error’s autocorrelation parameter (or the noise ratio) and apply projection techniques for inference about the persistence parameter. The proposed tests include Anderson-Rubin-type tests, dynamic versions of the split-sample procedure, and point-optimal versions of these tests. For distributional theory, three sets of assumptions are considered: we provide exact tests and confidence sets for Gaussian errors, establish exact Monte Carlo test procedures for non-Gaussian errors (possibly heavy-tailed), and show asymptotic validity under weaker distributional assumptions. Simulation results show that the proposed tests outperform the asymptotic test regarding size and exhibit excellent power in empirically realistic settings. We apply our inference methods to IBM’s price and option data (2009-2013). We consider 175 different instruments (IV’s) spanning 22 classes and analyze their ability to describe the low-frequency volatility. The IV’s are compared based on the average length of confidence intervals, which are produced by the proposed tests. The superior instrument set mostly consists of 5-minute HF realized measures, and these IV’s produce confidence sets where the volatility persistence parameter lies roughly between 0.85 and 1.0. We
find RVs with higher frequency produce wider confidence intervals compared to RVs with slightly lower frequency, showing that these confidence intervals adjust to absorb market microstructure noise or discretization error. Further, when we consider irrelevant or weak IV’s (jumps and signed jumps), the proposed tests produce unbounded confidence intervals.

**Keywords:** Stochastic volatility, Realized variance, High frequency data, Identification robust test

---

**Robust Bayesian analysis of multivariate time series**

Yixuan Liu*

Statistics, The University of Auckland, New Zealand

**Abstract:** There is a surge in the literature of nonparametric Bayesian inference on multivariate time series over the last decade, many approaches consider modelling the spectral density matrix using the Whittle likelihood which is an approximation of the true likelihood and commonly employed for Gaussian time series. Meier et al. (2019) proposes a nonparametric Whittle likelihood procedure along with a Bernstein polynomial prior weighted by a Hermitian positive definite (Hpd) Gamma process. However, it is known that nonparametric techniques are less efficient and powerful than parametric techniques when the latter specify the parameters which model the observations perfectly. Therefore, Kirch et al. (2019) suggests a nonparametric correction to the parametric likelihood in the univariate case that takes the efficiency of parametric models and amends sensitivities through the nonparametric correction. Along with this novel likelihood, the Bernstein polynomial prior equipped with a Dirichlet process weight is employed. My current work is to extend the corrected Whittle likelihood procedure to the multivariate case, this will be done by combining the work of Meier et al. (2019) and Kirch et al. (2019). Precisely, the multivariate version of the corrected Whittle likelihood is proposed along with the Hpd Gamma process weighted Bernstein polynomial prior to implement Bayesian inference. A key study of this work is to prove the posterior consistency. In the talk, I will review the work done by Meier et al. (2019) and Kirch et al. (2019), then an introduction of the multivariate corrected Whittle likelihood procedure will be given.

**Keywords:** Bayesian, Nonparametric, Spectral Density

---

**Posterior consistency for the spectral density of non-Gaussian stationary time series**

Yifu Tang*, Claudia Kirch2, Jeong Eun Lee1 and Renate Meyer1

1Department of Statistics, The University of Auckland, New Zealand
2Institute for Mathematical Stochastics, Department of Mathematics, Otto-von-Guericke University, Germany

**Abstract:** Various nonparametric approaches for Bayesian spectral density estimation of stationary time series have been suggested in the literature, mostly based on the Whittle likelihood approximation. A generalization of this approximation has been proposed in Kirch et al. (2019) who prove posterior consistency for spectral density estimation in combination with the Bernstein-Dirichlet process prior for Gaussian time series. In this talk, I will talk about how to extend the posterior consistency result to non-Gaussian time series by employing a modified version of general consistency theorem of Shalizi (2009) for dependent data and misspecified models. As a special case, posterior consistency for the spectral density under the Whittle likelihood as proposed by Choudhuri, Ghosal and Roy (2004) is also extended to non-Gaussian time series.

**Keywords:** Bayesian consistency, Bayesian nonparametrics, Spectral density function, Stationary time series
ARMA models for zero inflated count time series
Vurukonda Sathish¹*, Siuli Mukhopadhyay² and Rashmi Tiwari²
¹Electrical Engineering, IIT Bombay, India
²Mathematics, IIT Bombay, India

Abstract: Zero inflation is a common nuisance while monitoring disease progression over time. This article proposes a new observation driven model for zero inflated and over-dispersed count time series. The counts given the past history of the process and available information on covariates are assumed to be distributed as a mixture of a Poisson distribution and a distribution degenerate at zero, with a time dependent mixing probability, \( \pi_t \). Since count data usually suffers from overdispersion, a Gamma distribution is used to model the excess variation, resulting in a zero inflated negative binomial (NB) regression model with the mean parameter \( \lambda_t \). Linear predictors with autoregressive and moving average (ARMA) type terms, covariates, seasonality, and trend are fitted to \( \lambda_t \) and \( \pi_t \) through canonical link generalized linear models. Estimation is done using maximum likelihood aided by iterative algorithms, such as Newton Raphson (NR) and Expectation and Maximization (EM). Theoretical results on the consistency and asymptotic normality of the estimators are given. The proposed model is illustrated using in-depth simulation studies and a dengue data set.

Keywords: Negative Binomial, EM algorithm, Overdispersion, Mixture Distribution

Time-series data clustering via thick pen transformation
Minji Kim*
Department of Statistics, Seoul National University, Korea

Abstract: Our ultimate goal is to cluster time-series data by suggesting a new similarity measure and an optimization algorithm. To illustrate, we propose a new time-series clustering method based on the Thick Pen Transformation (TPT) proposed by Fryzlewicz and Oh (2011), whose basic idea is to draw along the data with a pen of given thicknesses. The main contribution of this research is that we suggest a new similarity measure for time-series data based on the overlap or gap between the two thick lines after transformation. This method of applying TPT to measure the association exploits the strengths of the transformation; it is a multi-scale visualization technique that can be defined to provide some information on neighborhood values’ temporal trends. Moreover, we further suggest an efficient iterative clustering optimization algorithm appropriate for the proposed measure. Our main motivation is to cluster a large number of physical step count data obtained from a wearable device. In addition, comparative numerical experiments are performed to compare our method to some existing methods. Real data analysis and simulation studies suggest that the proposed method can be applied in general for time series data distributed on the same side along the axis, whose similarities are measurable in the form of a proportion of overlapping areas.

Keywords: Time-series clustering, Similarity measure, K-means algorithm, Thick pen transform

A nonparametric test for paired data
Grzegorz Wylupek*
Mathematics and Computer Science, Institute of Mathematics, University of Wrocław, Poland

Abstract: The paper proposes the weighted Kolmogorov-Smirnov type test for the two-sample problem when the data is paired. We derive the asymptotic distribution of the test statistic under the null model as well as prove the consistency of the related test under the general alternatives. The dependence of the asymptotic distribution of the
test statistic from the dependence structure of the data forces the usage of the wild bootstrap technique for the inference. The bootstrap version of the test controls the Type I error under the null model and works very well under the alternative. In the proofs, the main role play the empirical processes’ tools.

**Keywords:** Bootstrap, Empirical process, Matched pairs, Two-sample problem

---

**Inference for Generalized Multivariate Analysis of Variance (GMANOVA) models, under multivariate skew t distribution for modelling skewed and heavy-tailed data**

Sayantee Jana$^1$* and Jemila Hamid$^2$

$^1$Decision Sciences and Information Systems, Indian Institute of Management Nagpur, India
$^2$Mathematics and Statistics, University of Ottawa, Canada

**Abstract:** The most extensively used statistical model in practice, both in research and in practice, is the linear model, due to its simplicity and interpretability. Linear models are preferred, even when approximate, for both univariate and multivariate data, especially since, multivariate skewed models come with their own added complexity. Hence, researchers would not prefer to deliberately add extra layers of complexity by considering non-linear models. Generalized Multivariate Analysis of Variance (GMANOVA) models, is one such linear model useful for the analysis of longitudinal data, which is repeated measurements of a continuous variable, from several individuals across any ordered variable such as time, temperature, pressure etc. It consists of a bilinear structure which allows for comparison across between groups, while maintaining the temporal structure of the data, unlike the Multivariate Analysis of Variance (MANOVA) which does not allow for any temporal ordering or temporal correlation in the model. GMANOVA models are widely used in economics, social and physical sciences, medical research and pharmaceutical studies. However, despite financial data being time-varying, the traditional GMANOVA model has limited to no applications in finance, due to the skewed and volatile nature of such data. This in turn makes financial data the right candidate for Multivariate Skew t (MST) distribution, as it allows for outliers in the data to be modelled, due to its heavy tails. In fact, portfolio analysis including mutual funds, capital asset pricing are all modelled using elliptical distributions, especially multivariate t distribution. The classical GMANOVA model assumes multivariate normality, and hence inferential tools developed for the classical GMANOVA model, may not be appropriate for skewed and heavy-tailed data. In our study, first we explore the sensitivity of inferential tools developed under multivariate normality for skewed and volatile data, and then we develop inferential tools for the GMANOVA model under the MST distribution.

**Keywords:** GMANOVA model, Multivariate skew t (MST) distribution, Skewed and heavy-tailed data, Inference for multivariate volatile data

---

**Multiscale representation of directional scattered data: Use of anisotropic radial basis functions**

Junhyeon Kwon*

Department of Statistics, Seoul National University, Korea

**Abstract:** Spatial inhomogeneity along the one-dimensional curve makes two-dimensional data non-stationary. Curvelet transform, first proposed by Candes and Donoho (1999), is one of the most well-known multiscale methods to represent the directional singularity, but it has a limitation that the data needs to be observed on equally-spaced sites. On the other hand, radial basis function interpolation is widely used to approximate the underlying function from the scattered data. However, the isotropy of the radial basis functions lowers the efficiency of the directional representation. This research proposes a new multiscale method that uses anisotropic radial basis functions to efficiently represent the direction from the noisy scattered data in two-dimensional Euclidean space. Basis functions are orthogonalized across the scales so that each scale can represent global or local directional structure separately.
It is shown that the proposed method is remarkable for representing directional scattered data through the numerical experiments. Convergence property and practical issues in implementation are discussed as well.

**Keywords:** Anisotropic radial basis functions, Directional scattered data, Multiscale analysis, Nonparametric function estimation

---

**Replicability of statistical findings under distributional shift**  
Suyash Gupta*  
Statistics, Stanford University, USA

**Abstract:** Common statistical measures of uncertainty like p-values and confidence intervals quantify the uncertainty due to sampling, i.e. the uncertainty due to not observing the full population. In practice, populations change between locations and across time. This makes it difficult to gather knowledge that replicates across data sets. We propose a measure of uncertainty that quantifies the distributional uncertainty of a statistical estimand, that is, the sensitivity of the parameter under general distributional perturbations within a Kullback-Liebler divergence ball. We also propose measure to estimate the stability of estimators with respect to directional or variable-specific shifts. The proposed measures would help judge whether a statistical finding is replicable across data sets and give more accurate estimates of parameters under shifted distribution.

**Keywords:** Distributional robustness, Distributional shift, Replicability, Transfer learning

---

**Selection of graphical continuous Lyapunov models with Lasso**  
Philipp Dettling*  
Department of Mathematics, Technical University of Munich, Germany, Germany

**Abstract:** In some applications, multivariate data may be thought of as cross-sectional observations of temporal processes. The recently proposed graphical continuous Lyapunov models take this perspective in the context of a multi-dimensional Ornstein-Uhlenbeck process in equilibrium. Under a stability assumption, the equilibrium covariance matrix is determined by the continuous Lyapunov equation. Given a sample covariance matrix, a very natural approach to model selection is to obtain sparse solutions to the Lyapunov equation by means of ℓ1-regularization. We apply the primal-dual witness technique to give probabilistic guarantees for successful support recovery in this approach. The key assumption in this guarantee is an irrepresentability condition. As we demonstrate, the irrepresentability condition may be violated in subtle ways, particularly, for models with feedback loops.

**Keywords:** Graphical models, Lasso, Irrepresentability condition, Support recovery

---

**Identifiability of linear structural equation models with homoscedastic errors using algebraic matroids**  
Jun Wu* and Mathias Drton  
Department of Mathematics, Technical University of Munich, Germany

**Abstract:**
Abstract: We consider structural equation models (SEMs), in which every variable is a function of a subset of the other variables and a stochastic error. Each such SEM is naturally associated with a directed graph describing the relationships between variables. For the case of homoscedastic errors, recent work has proposed methods for inferring the graph from observational data under the assumption that the graph is acyclic (i.e., the SEM is recursive). In this work we study the setting of homoscedastic errors but allow the graph to be cyclic (i.e., the SEM to be non-recursive). Using an algebraic approach that compares matroids derived from the parameterizations of the models, we derive sufficient conditions for two simple directed graphs generating different distributions generically. Based on these conditions, we exhibit subclasses of graphs that allow for directed cycles, yet are generically identifiable. Our study is supplemented by computational experiments that provide a full classification of models given by simple graphs with up to 6 nodes.

Keywords: Algebraic matroids, Graphical model, Identifiability, Structural equation models

Convergence of stochastic gradient descent for Lojasiewicz-landscapes
Sebastian Kassing* and Steffen Dereich
Department of Stochastics, WWU Muenster, Germany

Abstract: In this talk we discuss almost sure convergence of Stochastic Gradient Descent (SGD) \((X_n)_{n \in \mathbb{N}}\) and Stochastic Gradient Flow (SGF) \((X_t)_{t \in \mathbb{R}^+}\) for a given target function \(F\). First, we give a simple proof for almost sure convergence of the target value \((F(X_n))\) (resp. \((F(X_t))\)) assuming that \(F\) admits a locally Hölder-continuous gradient \(f = DF\). This results entails convergence of the iterates \((X_n)\) (resp. \((X_t)\)) in the case where \(F\) does not posses a continuum of critical points. In a general non-convex setting with \(F\) possibly containing a rich set of critical points, convergence of the process itself is sometimes taken for granted, but actually is a non-trivial issue as there are solutions to the gradient flow ODE for \(F\) that stay in a compact set but do not converge. Using the Lojasiewicz-inequality we derive bounds on the step-sizes and the size of the perturbation in order to guarantee convergence of \((X_n)\) (resp. \((X_t)\)) for analytic target functions. Also, we derive the convergence rate under the assumptions that the loss function satisfies a particular Lojasiewicz-inequality. Last, we compare the results for SGD and SGF and discuss optimality of the assumptions.

Keywords: Stochastic gradient descent, Stochastic approximation, Lojasiewicz inequality, Deep learning

Simulated annealing-backpropagation algorithm on parallel trained maxout networks (SABPMAX) in detecting credit card fraud
Sheila Mae Golingay* and Joselito Magadia
School of Statistics, University of the Philippines - Diliman, Philippines

Abstract: Based on the Backpropagation (BP) artificial neural network algorithm, this study introduces the idea of combining Simulated Annealing (SA), a global searching algorithm and then proposes a new neural network algorithm: Simulated Annealing-Backpropagation Algorithm on Parallel Trained Maxout Networks (SABPMAX) algorithm. The proposed algorithm can improve the numerical stability and evaluation measures in detecting credit card fraud. It makes use of the global searching capability of SA and the precise local searching element of the backpropagation algorithm to improve the initial weights of the network towards improving detection of credit card fraud. Several models were made and tested using different fraud distributions. Furthermore, separate applications of BP algorithm and SABPMAX algorithm were compared. Numerical results show a higher accuracy rate, higher
The smoking gun: Statistical theory improves neural network estimates
Sophie Langer*
Mathematics, Technische Universität Darmstadt, Germany

Abstract: In this talk we analyze the $L_2$ error of neural network regression estimates with one hidden layer. Under the assumption that the Fourier transform of the regression function decays suitably fast, we show that an estimate, where all initial weights are chosen according to proper uniform distributions and where the weights are learned by gradient descent, achieves a rate of convergence of $1/\sqrt{n}$ (up to a logarithmic factor). Our statistical analysis implies that the key aspect behind this result is the proper choice of the initial inner weights and the adjustment of the outer weights via gradient descent. This indicates that we can also simply use linear least squares to choose the outer weights. We prove a corresponding theoretical result and compare our new linear least squares neural network estimate with standard neural network estimates via simulated data. Our simulations show that our theoretical considerations lead to an estimate with an improved performance. Hence the development of statistical theory can indeed improve neural network estimates.

Keywords: Rate of convergence, Gradient descent, Deep Learning

Stochastic block model for multiple networks
Tabea Rebafka*
LPSM, Sorbonne Université, France

Abstract: A model-based approach for the analysis of a collection of observed networks is considered. We propose to fit a stochastic block model to the data. The novelty consists in the analysis of not a single, but multiple networks. The major challenge resides in the development of a computationally efficient algorithm. Our method is an agglomerative algorithm based on the integrated classification likelihood criterion that performs simultaneously model selection and node clustering. Compared to the single-network context, an additional difficulty resides in the necessity to compare networks one to another and aggregate partial solutions. We propose a distance measure to compare stochastic block models and solve the label switching problem among graphs in a computationally efficient way.

Keywords: Multiple networks, Stochastic block model, Integrated classification likelihood

Deep neural networks for Faster nonparametric regression models
Mehmet Ali Kaygusuz1*, Abdullah Nuri Somuncuoğlu2 and Vilda Purutcuoğlu3
1Department of Statistics, Middle East Technical University, Turkey
2Department of Bioengineering, Middle East Technical University, Turkey
3Department of Statistics, and the department of Bioengineering, Middle East Technical University, Turkey

Abstract: Deep neural networks have been an attention in recent years due to it has huge success in applicational areas such as signal progressing, biological networks, and time series analysis. Schmidt-Hieber (2020) suggested Feedforward neural networks for Generalized additive models (GAMs) with sparsity and Re-Lu function. However, the over-parametrized problem can be challenging when the number of parameters exceeds the number of samples which is studied by Bauer and Kohler (2019). Therefore, we use the bootstrap methods to cope with this problem since
bootstrap methods (Efron(1979)) are computationally faster and reduced the variance. Specifically, we propose the Smooth bootstrap method (Sen et al. (2010)) which can be more appropriate for nonparametric regression while capturing the nonlinearity and the interaction between variables, resulting in better performance in bias-variance trade-off. In this study, when we combine bootstrap with multilayer neural network together with GAM's approaches, we also aim to optimize the model selection in GAMs via distinct model selection criteria, namely, consistent Akaike information criterion with Fisher matrix and information complexity (Bozdogan, 1987). We evaluate the performance of all suggested models in different dimensional protein-protein interaction network datasets and biomedical signal data in terms of various accuracy measures.

References

Keywords: Deep learning, Model selection, Nonparametric regression, Bootstrap

Generative model for fbm with deep ReLU neural networks
Michael Allouche*
CMAP, Ecole Polytechnique, France

Abstract: Over the last few years, a new paradigm of generative models based on neural networks have shown impressive results to simulate – with high fidelity – objects in high-dimension, while being fast in the simulation phase. In this work, we focus on the simulation of continuous-time processes (infinite dimensional objects) based on Generative Adversarial Networks (GANs) setting. More precisely, we focus on fractional Brownian motion, which is a centered Gaussian process with specific covariance function. Since its stochastic simulation is known to be quite delicate, having at hand a generative model for full path is really appealing for practical use. However, designing the architecture of such neural networks models is a very difficult question and therefore often left to empirical search. We provide a high confidence bound on the uniform approximation of fractional Brownian motion $B^H(t): t \in [0,1]$ with Hurst parameter $H$, by a deep-feedforward ReLU neural network fed with a $N$-dimensional Gaussian vector, with bounds on the network construction (number of hidden layers and total number of neurons). Our analysis relies, in the standard Brownian motion case ($H = 1/2$), on the Levy construction of $B^H$ and in the general fractional Brownian motion case ($H \neq 1/2$), on the Lemarié-Meyer wavelet representation of $B^H$. This work gives theoretical support to use, and guidelines to construct, new generative models based on neural networks for simulating stochastic processes. It may well open the way to handle more complicated stochastic models written as a Stochastic Differential Equation driven by fractional Brownian motion.

Keywords: Fractional Brownian motion, Gaussian process, neural networks, Generative models
Wild bootstrap for high-dimensional spatial data

Daisuke Kurisu1*, Kengo Kato2 and Xiaofeng Shao3
1Department of Industrial Engineering and Economics, Tokyo Institute of Technology, Japan
2Department of Statistics and Data Science, Cornell University, USA
3Department of Statistics, University of Illinois at Urbana-Champaign, USA

Abstract: This study establishes a high-dimensional CLT for the sample mean of p-dimensional spatial data observed over irregularly spaced sampling sites in $\mathbb{R}^d$, allowing the dimension $p$ to be much larger than the sample size $n$. We adopt a stochastic sampling scheme that can flexibly generate irregularly spaced sampling sites and include both pure increasing domain and mixed increasing domain frameworks. To facilitate statistical inference, we develop the spatially dependent wild bootstrap (SDWB) and justify its asymptotic validity in high dimensions by deriving error bounds that hold almost surely conditionally on the stochastic sampling sites. Our dependence conditions on the underlying random field cover a wide class of random fields such as Gaussian random fields and continuous autoregressive moving average random fields. Through numerical simulations and a real data analysis, we demonstrate the usefulness of our bootstrap-based inference in several applications, including joint confidence interval construction for high-dimensional spatial data and change-point detection for spatio-temporal data.

Keywords: Wild bootstrap, High-dimensional CLT, Irregularly spaced spatial data, Spatio-temporal data

Lifting scheme for streamflow data in river networks

Seoncheol Park*
Department of Information Statistics, Chungbuk National University, Korea

Abstract: In this presentation, we suggest a new multiscale method for analyzing water pollutant data located in river networks. The main idea of the proposed method is to adapt the conventional lifting scheme, reflecting the characteristics of streamflow data in the river network domain. Due to the complexity of the data domain structure, it is difficult to apply the lifting scheme to the streamflow data directly. To solve this problem, we propose a new lifting scheme algorithm for streamflow data that incorporates flow-adaptive neighborhood selection, flow proportional weight generation, and flow-length adaptive removal point selection. A nondecimated version of the proposed lifting scheme is also suggested. We will provide a simulation study and a real data analysis of water pollutant data observed on the Geum-River basin in South Korea.

Keywords: Lifting scheme, River network, Smoothing, Spatial modeling

Optimal designs for some bivariate cokriging models

Subhadra Dasgupta1*, Siuli Mukhopadhyay2 and Jonathan Keith3
1Department of Mathematics, IITB-Monash Research Academy, India
2Department of Mathematics, Indian Institute of Technology Bombay, India
3School of Mathematics, Monash University, Australia

Abstract: This article focuses on the estimation and design aspects of a bivariate collocated cokriging experiment. For a large class of covariance matrices a linear dependency criterion is identified, which allows the best linear unbiased estimator of the primary variable in a bivariate collocated cokriging setup to reduce to a univariate kriging...
estimator. Exact optimal designs for efficient prediction for such simple and ordinary reduced cokriging models, with one dimensional inputs are determined. Designs are found by minimizing the maximum and integrated prediction variance, where the primary variable is an Ornstein-Uhlenbeck process. For simple and ordinary cokriging models with known covariance parameters, the equispaced design is shown to be optimal for both criterion functions. The more realistic scenario of unknown covariance parameters is addressed by assuming prior distributions on the parameter vector, thus adopting a Bayesian approach to the design problem. The equispaced design is proved to be the Bayesian optimal design for both criteria. The work is motivated by designing an optimal water monitoring system for an Indian river.

**Keywords:** Cross-covariance, Equispaced designs, Exponential Covariance, Mean squared error of prediction

**Statistical inference for mean function of longitudinal imaging data over complicated domains**

Jie Li* and Qirui Hu
Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, Tsinghua University, China

**Abstract:** Motivated by longitudinal imaging data which possesses inherent spatial and temporal correlation, we propose a novel procedure to estimate its mean function. Functional moving average is applied to depict the dependence among temporally ordered images and flexible bivariate splines over triangulations are used to handle the irregular domain of images which is common in imaging studies. Both global and local asymptotic properties of the bivariate spline estimator for mean function are established with simultaneous confidence corridors (SCCs) as a theoretical byproduct. Under some mild conditions, the proposed estimator and its accompanying SCCs are shown to be consistent and oracle efficient as if all images were entirely observed without errors. The finite sample performance of the proposed method through Monte Carlo simulation experiments strongly corroborates the asymptotic theory. The proposed method is illustrated by analyzing two sea water potential temperature data sets.

**Keywords:** Bivariate splines, Spatiotemporal, Functional moving average, Simultaneous confidence corridor

**Gaussian linear dynamic spatio-temporal models and time asymptotics**

Suman Guha*
Department of Statistics, Presidency University, Kolkata, India

**Abstract:** Gaussian linear dynamic spatio-temporal models (LDSTMs) are linear gaussian state-space models for spatio-temporal data which contains deterministic or (and) stochastic spatio-temporal covariates besides the spatio-temporal response. They are extensively used to model discrete-time spatial time series data. The model fitting is carried out either by classical maximum likelihood approach or by calculating Bayesian maximum a posteriori estimate of the unknown parameters. While their finite sample behaviour is well studied, literature on their asymptotic properties is relatively scarce. Classical theory on asymptotic properties of maximum likelihood estimator for linear state-space models is not applicable as it hinges on the assumption of asymptotic stationarity of covariate processes, which is seldom satisfied by discrete-time spatial time series data. In this article, we consider a very general Gaussian LDSTM that can accommodate arbitrary spatio-temporal covariate processes which grow like power functions wrt. time in deterministic or (and) suitable stochastic sense. We show that under very minimal assumptions, any approximate MLE and Bayesian approximate MAPE of some of the unknown parameters and parametric functions are
strongly consistent. Furthermore, building upon the strong consistency theorems we also establish rate of convergence results for both approximate MLE and approximate MAPE.

**Keywords:** Gaussian LDSTM, Linear state space model, Spatio-temporal covariates, Spatial time series data

**High-dimensional spectral analysis**
Jonas Krampe* and Efstathios Paparoditis
University of Mannheim

**Abstract:** An important part of multivariate time series analysis is the spectral domain and here key quantities are here the spectral density matrix and the partial coherence. Under a high-dimensional set up, we present how inference can be derived for the partial coherence. Furthermore, we present a valid statistical test for the statistical hypothesis that the partial coherence is almost everywhere smaller than a given bound that includes testing whether the partial coherence is zero or not. Applications of this are among others the construction of graphical interaction models helpful for analyzing functional connectivity among brain regions. We illustrate our procedure by means of simulations and a real data application.

**Keywords:** Partial Coherence, Bootstrap, Testing, De-sparsified estimators

**Extreme value analysis for mixture models with heavy-tailed impurity**
Vladimir Panov and Ekaterina Morozova*
Laboratory of Stochastic Analysis and its Applications, HSE University, Russia

**Abstract:** While there exists a well-established theory for the asymptotic behaviour of maxima of the i.i.d. sequences, very few results are available for the triangular arrays, when the distribution can change over time. Typically, the papers on this issue deal with convergence to the Gumbel law or twice-differentiable distribution. The current research is a joint work with Vladimir Panov, available as a preprint on arXiv.org [2]. It contributes to the aforementioned problem by providing the extreme value analysis for mixture models with varying parameters, which can be viewed as triangular arrays. In particular, we consider the case of the heavy-tailed impurity, which appears when one of the components has a heavy-tailed distribution, and the corresponding mixing parameter tends to zero as the number of observations grows. We analyse two ways of modelling this impurity, namely, by the non-truncated regularly varying law and its upper-truncated version with an increasing truncation level. The set of possible limit distributions for maxima turns out to be much more diverse than in the classical setting, especially for a mixture with the truncated component, where it includes four discontinuous laws. In the latter case, the resulting limit depends on the asymptotic behaviour of the truncation point, which is shown to be related to the truncation regimes introduced in [1]. For practical purposes we describe the procedure of the application of the considered model to the analysis of financial returns.

**References**

**Keywords:** Extreme values, Triangular arrays, Heavy-tailed distributions, Mixture model
Robust geodesic regression
Ha-Young Shin* and Hee-Seok Oh
Department of Statistics, Seoul National University, Korea

Abstract: This study explores robust regression for data on Riemannian manifolds. Geodesic regression is the generalization of linear regression to a setting with a manifold-valued dependent variable and one or more real-valued independent variables. The existing work on geodesic regression uses the sum-of-squared errors to find the solution, but as in the classical Euclidean case, the least-squares method is highly sensitive to outliers. In this study, we use M-type estimators, including the L1, Huber and Tukey biweight estimators, to perform robust geodesic regression, and describe how to calculate the tuning parameters for the latter two. We show that, on compact symmetric spaces, all M-type estimators are maximum likelihood estimators, and argue for the overall superiority of the L1 estimator over the L2 and Huber estimators on high-dimensional manifolds and over the Tukey biweight estimator on compact high-dimensional manifolds. A derivation of the Riemannian Gaussian distribution on k-dimensional spheres is also included. Results from numerical examples, including analysis of real neuroimaging data, demonstrate the promising empirical properties of the proposed approach.

Keywords: Geodesic regression, Manifold statistics, M-type estimators, Robust regression

A multi-sigmoidal logistic model: statistical analysis and first-passage-time application
Antonio Di Crescenzo1, Paola Paraggio1*, Francisco Torres-Ruiz2 and Patricia Romàn-Romàn2
1Department of Mathematics, University of Salerno, Italy
2Department of Statistics and Operations Research, University of Granada, Spain

Abstract: Sigmoidal growth models are widely used in various applied fields, from biology to software reliability and economics. Usually, they describe dynamics in restricted environments. However, many real phenomena exhibit different phases, each one following a sigmoidal-type pattern. Stimulated by these more complex dynamics, many researchers investigate generalized versions of classical sigmoidal models characterized by several inflection points. Along these research lines, a generalization of the classical logistic growth model is considered in the present work, introducing in its expression a polynomial term. The model is described by a stochastic differential equation obtained from the deterministic counterpart by adding a multiplicative noise term. The resulting diffusion process, having a multi-sigmoidal mean, may be useful in the description of particular growth dynamics in which the evolution occurs by stages. The problem of finding the maximum likelihood estimates of the parameters involved in the definition of the process is also addressed. Precisely, the maximization of the likelihood function will be performed by means of meta-heuristic optimization techniques. Moreover, various strategies for the selection of the optimal degree of the polynomial will be provided. Further, the first-passage-time (FPT) problem is considered: an approximation of its density function will be obtained numerically, by means of the fptdApprox R-package. Finally, some simulated examples are presented.

Keywords: Multi-sigmoidal growth, Maximum likelihood estimation, First-passage-time, Simulation

Statistical inference for functional linear problems
Tim Kutta*, Holger Dette and Gauthier Dierickx
Faculty for Mathematics, Ruhr University Bochum, Germany
Abstract: In this talk we consider the linear regression model $Y = SX + e$ with functional regressors and responses. This model has attracted much attention in terms of estimation and prediction, but less is known with regard to statistical inference for the unobservable slope operator $S$. In this talk we discuss new inference tools to detect relevant deviations of the parameter $S$ from a hypothesized slope $S'$. As modes of comparison we consider the Hilbert-Schmidt norm $\|S - S'\|^2$ as well as the prediction error $E \|SX - S'X\|^2$. Our theory is based on the novel technique of “smoothness shifting”, which helps us to circumvent existing negative results on the weak convergence of estimators for $S$. In contrast to all related works the test statistic proposed converges at a rate of $N^{-1/2}$, permitting a fast detection of local alternatives. Furthermore, while most existing procedures rely on i.i.d. observations for Gaussian approximations, our test statistic converges even in the presence of dependence, quantified by phi- or strong mixing. Due to a self-normalization procedure, our approach is user friendly, computationally inexpensive and robust.

Keywords: Functional linear regression, Inverse problem, Prediction error

CS32 Statistical Modeling and Prediction
JULY 22 (THU), 22:30 - 23:00 KST

An evolution of the beta regression for non-monotone relations
Gloria Gheno*
Statistics, Ronin Institute, Italy

Abstract: The beta regression is based on the beta distribution or its reparameterizations, which are used to obtain a regression structure on the mean which is much easier to analyze and interpret. This method analyzes data whose value is within the range (0,1), such as rates, proportions or percentage and is useful for studying how the explanatory variables affect them. For the mean of the beta regression scholars continued to use the traditional link functions used for binary regressions, i.e. the logit, the probit and the complementary log-log. In this paper I propose a new link function for the parameter mean of a beta regression which has as its particular cases the logit, representing a traditional symmetric link function, and the gev (Generalized extreme value), introduced precisely because of its asymmetry. In the simplest form of the beta regression, the inverse of the link function, called response function, represents the link between the mean and the explanatory variables. In this paper the response function can become non-monotone and therefore the link function is calculated for intervals. This particularity has never been proposed so far in the literature although some scholars have found non-monotone relationships between the response variable and its explanatory variables. The parameters of the function are estimated by maximizing the likelihood function, using my modified version of the genetic algorithm. I compare my method with the one proposed by Cribari-Neto, in which the link function is decided a priori, using simulated data, so as to be able to compare which of the two methods is closest to the true values. My method is better because it is able to correctly determine the link function with which the data are simulated and to estimate the parameters with less error.

Keywords: Beta regression, Extreme values, Gev, Link function

Robust censored regression with l1-norm regularization
Jad Beyhum* and Ingrid Van Keilegom
ORSTAT, KU Leuven, Belgium

Abstract: This paper considers inference in a linear regression model with random right-censoring and outliers. The number of outliers can grow with sample size while their proportion goes to 0. We propose to penalize the estimator of Stute (1993) by the l1-norm. We derive rates of convergence and establish asymptotic normality. Our estimator has the same asymptotic variance as Stute’s estimator in the censored linear model without outliers. Tests and confidence
sets can therefore rely on the theory developed by Stute. The outlined procedure is also computationally advantageous, it amounts to solving a convex optimization program. We also propose a second estimator which uses the l1-norm penalized Stute estimator as a first step to detect outliers. It has similar theoretical properties but better performances in finite samples as assessed by simulations.

**Keywords:** Robust regression, l1-norm penalization, Censoring

---

**SPLVC modal regression with error-prone linear covariate**

Tao Wang*

Economics, University of California, Riverside, USA

**Abstract:** To broaden the scope of existing modal regressions, we in this paper propose two procedures, called B-splines-based procedure and stepwise-based procedure, to retrieve the estimates for a semiparametric partially linear varying coefficient (SPLVC) modal regression with error-prone linear covariate in which a linear covariate is not observed, but an ancillary variable is available. With B-splines-based procedure, varying coefficients are approximated through B-splines, and a deconvoluting kernel-based objective function is constructed straightly. For the stepwise-based procedure, by defining restricted regression mode via imposing a constructive condition on model format, a two-step method is developed in which the varying coefficients are concentrated out by applying the "correction for attenuation" methodology in mean regression to alter the original model to a reduced parametric modal regression. Consistency and asymptotic properties of the estimators for these two newly proposed procedures are investigated under mild conditions according to the tail behavior of the characteristic function of the error distribution, either ordinary smooth distribution or super smooth distribution. Bandwidth selection in theory and practice are explored. For comparison, we also develop the asymptotic theorems for the SPLVC modal estimators with B-splines approximation without covariate measurement error. Monte Carlo simulations are conducted to examine the finite sample performance of the estimators and a pseudo data analysis is presented to further illustrate the proposed estimation procedures.

**Keywords:** B-splines, Deconvoluting kernel, Error-prone linear covariate, Modal regression

---

**Regularized double machine learning in partially linear models with unobserved confounding**

Corinne Emmenegger* and Peter Bühlmann

D-MATH, ETH Zurich, Switzerland

**Abstract:** Double machine learning (DML) can be used to estimate the linear coefficient in a partially linear model with confounding variables. However, the standard DML estimator has a two-stage least squares interpretation and may yield overly wide confidence intervals. To address this issue, we present the regularization-selection regsDML method that leads to narrower confidence intervals but preserves coverage guarantees. We rely on DML to estimate nuisance parameters with arbitrary machine learning algorithms and combine it with a regularization and selection scheme. Our regsDML method is fully data driven and optimizes the estimated asymptotic mean squared error of the coefficient estimate. The regsDML estimator can be expected to converge at the parametric rate and to follow an asymptotic Gaussian distribution. Empirical examples demonstrate our theoretical and methodological developments. Software code for the regsDML method is available in the R-package dmlalg.

**Keywords:** Double machine learning, Endogenous partially linear model, Regularization, Two-stage least squares
An extended model for phylogenetic maximum likelihood based on discrete morphological characters
David Spade
Mathematical Sciences, University of Wisconsin–Milwaukee, USA

Abstract: Maximum likelihood is a common method of estimating a phylogenetic tree based on a set of genetic data. However, models of evolution for certain types of genetic data are highly flawed in their specification, and this misspecification can have an adverse impact on phylogenetic inference. Our attention here is focused on extending an existing class of models for estimating phylogenetic trees from discrete morphological characters. The main advance of this work is a model that allows unequal equilibrium frequencies in the estimation of phylogenetic trees from discrete morphological character data using likelihood methods. Possible extensions of the proposed model will also be discussed.

Keywords: Maximum Likelihood, Statistical Phylogenetics, Time Reversibility, Single-Nucleotide Variant

Combined linkage and association mapping integrating population-based and family-based designs using multinomial regression
Saurabh Ghosh and Sourojyoti Barick
Human Genetics Unit, Indian Statistical Institute, India

Abstract: Genetic association analyses yield higher powers compared to linkage analyses in identifying chromosomal regions harboring susceptibility genes modulating complex human disorders and correlated quantitative phenotypes. However, while population-based association designs suffer from the problem of population stratification that often results in inflated type I errors, linkage designs are based on families and are protected against such inflations. The models suggested in Multiphen (O’Reilly et al., 2012) and BAMP (Majumdar et al. 2015) provide an alternative to study population-based genotype-phenotype association by exploring the dependence of genotype on phenotype instead of the naturally arising dependence of phenotype on genotype. This reversal of the regression model, while has no impact on the inference on association, provides the flexibility of incorporating multiple phenotypes without the requirement of making any a priori assumptions on the correlation structure of the vector of phenotypes. Our aim is to investigate whether family based data can be included in addition to population level data in the framework of the BAMP (Binomial regression-based Association of Multivariate Phenotypes) model so as to develop a combined test for genetic linkage and association. The family-based regression model involves the conditional distribution of identity-by-state (i.b.s.) scores on the squared sib-pair phenotype differences. However, since the marginal distribution of i.b.s. counts do not follow a Binomial distribution, we propose a Trinomial Regression model for the linkage component of our combined test. Given that the marginal distributions of the response variables in the population-based and family-based designs are different, the combined test is constructed jointly on the estimated regression parameters corresponding to the two designs. The likelihood ratio test statistic asymptotically follows a mixture of two chi-squares distributions with one and two degrees of freedom respectively under the null. We carry out extensive simulations to evaluate the power of the proposed combined test.

Keywords: Genotype-phenotype association, Genetic epidemiology, Identity-by-state score, Likelihood ratio test
An alternative to intersection-union test for the composite null hypothesis used to identify shared genetic risk of disease outcomes
Debashree Ray*
Epidemiology, Johns Hopkins University, USA

Abstract: With a growing number of disease- and trait-associated genetic variants detected and replicated across genome-wide association studies (GWAS), scientists are increasingly noting the influence of individual variants on multiple seemingly unrelated traits—a phenomenon known as pleiotropy. Cross-phenotype association tests, applied on two or more traits, usually test the null hypothesis of no association of a variant with any trait. Rejection of this null can be due to association between the variant and a single trait, with no indication if the variant influences >1 trait. This problem can be formulated as a composite null hypothesis test for each variant. For two traits, a level-α two-parameter intersection-union test (IUT) can be used. However, for testing millions of variants at genome-wide significance threshold (α=5x10^{-8}), IUT is extremely conservative. In this talk, I will discuss a new statistical approach, PLACO—pleiotropic analysis under composite null hypothesis—to discover variants influencing risk of two traits using GWAS summary statistics (i.e., using estimated effect size, its standard error and p-value for each variant). PLACO uses the product of Z-statistics across two traits as test statistic for pleiotropy, the null distribution of which is derived in the form of a mixture distribution that allows for fractions of variants to be associated with none or only one of the traits. PLACO gives an approximate asymptotic p-value for association with both traits, avoiding estimation of nuisance parameters related to mixture proportions and variance components. Simulation studies demonstrate its well-controlled type I error, and massive power gain over IUT and alternative ad hoc methods typically used for testing pleiotropy. Finally, I will show application of PLACO to type 2 diabetes and prostate cancer genetics to explain their inverse association reported in many previous epidemiologic studies.

Keywords: Composite null hypothesis, Intersection-union test, Statistical genetics, Genome-wide association study

Efficient SNP-based heritability estimation using Gaussian predictive process in large-scale cohort studies
Saonli Basu*
Biostatistics, University of Minnesota, USA

Abstract: For decades, linear mixed models (LMM) have been widely used to estimate heritability in twin and family studies. Recently, with the advent of high throughput genetic data, there have been attempts to estimate heritability from genome-wide SNP data on a cohort of distantly related individuals. Fitting such an LMM in large-scale cohort studies, however, is tremendously challenging due to high dimensional linear algebraic operations. In this paper, we simplify the LMM by unifying the concept of Genetic Coalescence and Gaussian Predictive Process, and thereby greatly alleviating the computational burden. Our proposed approach PredLMM has much better computational complexity than most of the existing packages and thus, provides an efficient alternative for estimating heritability in large-scale cohort studies. We illustrate our approach with extensive simulation studies and use it to estimate the heritability of multiple quantitative traits from the UK Biobank cohort.

[This is joint work with Souvik Seal, Colorado School of Public Health, and Abhirup Datta, Johns Hopkins University.]
Keywords: Genetic Relationship Matrix, Gaussian Predictive Process
Data-adaptive groupwise test for genomic studies via the Yanai’s generalized coefficient of determination
Masao Ueki*
School of Information and Data Sciences, Nagasaki University, Japan

Abstract: In genomic studies, repeated univariate regression for each variable is utilized to screen useful variables. However, signals jointly detectable with other variables may be overlooked. Group-wise analysis for a pre-defined group is often developed, but the power will be limited if the knowledge is insufficient. A flexible data-adaptive test procedure is thus proposed for conditional mean applicable to a variety of model sequences that bridge between low and high complexity models as in penalized regression. The test is based on the model that maximizes a generalization of the Yanai’s generalized coefficient of determination by exploiting the tendency for the dimensionality to be large under the null hypothesis. The test does not require complicated null distribution computation, thereby enabling large-scale testing application. Numerical studies demonstrated that the proposed test applied to the lasso and elastic net had a high power regardless of the simulation scenarios. Applied to a group-wise analysis in real genome-wide association study data from Alzheimer’s Disease Neuroimaging Initiative, the proposal gave a higher association signal than the existing methods.

Ruin probabilities in the presence of risky investments and random switching
Konstantin Borovkov*
School of Mathematics and Statistics, The University of Melbourne, Australia

Abstract: We consider a reserve process where claim times form a renewal process, while between the claim times the process has the dynamics of geometric Brownian motion-type Itô processes with time-dependent random coefficients that are “reset” after each jump. Following the approach of Pergamenshchikov and Zeitoni (2006), we use the implicit renewal theory to obtain power-function bounds for the eventual ruin probability. In the special case of the gamma-distributed claim inter-arrival times and geometric Brownian motions with random coefficients, we obtain necessary and sufficient conditions for existence of Lundberg's exponent (ensuring the power function behaviour for the ruin probability). [Joint work with Roxanne He.]

Keywords: Ruin probability, Random switching, Stochastic volatility

Wasserstein convergence rates for random bit approximations of continuous Markov processes
Thomas Kruse1*, Stefan Ankirchner2, Wolfgang Loehr3 and Mikhail Urusov3
1Institute of Mathematics, University of Giessen, Germany
2Institute of Mathematics, University of Jena, Germany
3Faculty of Mathematics, University of Duisburg-Essen, Germany

Abstract: We determine the convergence speed of the EMCEL scheme for approximating one-dimensional continuous strong Markov processes. The scheme is based on the construction of certain Markov chains whose laws can be embedded into the process with a sequence of stopping times. Under a mild condition on the process’ speed measure we prove that the approximating Markov chains converge at fixed times at the rate of $1/4$ with respect to Wasserstein distance.
every \( p \)-th Wasserstein distance. For the convergence of paths, we prove any rate strictly smaller than \( 1/4 \). These results apply, in particular, to processes with irregular behavior such as solutions of SDEs with irregular coefficients and processes with sticky points. Moreover, we present several further properties of the EMCEL scheme and discuss its differences from the Euler scheme.

**Keywords:** Wasserstein distance, One-dimensional Markov process, Markov chain approximation, Rate of convergence

### The discrete membrane model on trees

Alessandra Cipriani\(^1\), Biltu Dan\(^2\)*, Rajat Subhra Hazra\(^3\) and Rounak Ray\(^4\)

\(^1\)DIAM, TU Delft, Netherlands
\(^2\)Department of Mathematics, Indian Institute of Science, India
\(^3\)Mathematical Institute, Leiden University, Netherlands
\(^4\)Mathematics and Computer Science, TU Eindhoven, Netherlands

**Abstract:** The discrete membrane model (MM) is a random interface model for separating surfaces that tend to preserve curvature. It is similar to the discrete Gaussian free field (DGFF) for which the most likely interfaces are those preserving mean height. However working with the two models presents some key differences. In particular, a lot of tools (electrical networks, random walk representation for the covariance) are available for the DGFF and lack in the MM. In this talk we will investigate a random walk representation for the covariance of the MM and by means of it will define and study the MM on regular trees. In particular, we will study the scaling limit of the maxima of the MM on regular trees.

**Keywords:** Membrane model, Gaussian free field

### On a two-server queue with consultation by main server with protected phases of service

Resmi Thekkiniyedath\(^1\)* and Lakshmy Balakrishna\(^2\)

\(^1\)Mathematics, KKTM Govt. College, Pullut, Kerala, India
\(^2\)Mathematics, Cochin University of Science and Technology, Cochin, Kerala, India

**Abstract:** This paper analyses a two-server queueing model with consultations given by the main server to the regular server. The main server not only serves customers but also provides consultation to the regular server with a pre-emptive priority over customers. The customers at the main server undergo interruptions during their service. The interruption is not allowed to a customer at the main server if the service is in any one of the protected phases of service. There are upper bounds for the number of interruptions to a customer at the main server and the number of consultations to the regular server during the service of a customer. A super clock also determines whether to allow further interruption to a customer at the main server or not. A threshold clock decides the restart or resumption of the services at the main and regular servers after each consultation. The arrival process and requirement of consultation follow mutually independent Poisson processes. The service times at the main server and the regular server are assumed to follow mutually independent phase type distributions. The stability condition is established and some performance measures are studied numerically.

**Keywords:** Main server, Regular server, Interruption, Protected phases
**CS35 Financial Data Analysis**

**JULY 22 (THU), 21:30 - 22:00 KST**

**Hedging portfolio for a degenerate market model**

İhsan Demirel¹*, Mine Caglar¹ and Ali Süleyman Üstünel²

¹Department of Mathematics, Koc University, Turkey
²Department of Mathematics, Bilkent University, Turkey

**Abstract:** The purpose of this talk is to derive the hedging portfolio in a financial market where prices-per-shares are governed by a stochastic equation with a singular volatility matrix. The main mathematical tools of the study are the representation with respect to a minimal martingale and Malliavin calculus for the functionals of a degenerate diffusion process, which have been established in recent studies. We use those developments to prove a version of the Hausmann-Bismut-Ocone type representation formula derived for these functionals under an equivalent martingale measure. Consequently, we derive the hedging portfolio as a solution to a system of linear equations. The uniqueness of the solution is achieved by a projection idea that lies at the core of the martingale representation. We apply our result to exotic options, whose value at maturity depends on the prices over the entire time horizon. This work is supported by Tubitak Project No. 118F403.

**Keywords:** Clark-Ocone formula, Replicating portfolio, Malliavin calculus, Degenerate diffusion

**An optimal combination of proportional - excess of loss reinsurance with random premiums**

Suci Sari¹*, Ikha Magdalena² and Khreshna Syuhada³

¹Statistics Research Division, Institut Teknologi Bandung, Indonesia
²Industrial and Financial Mathematics Research Group, Institut Teknologi Bandung, Indonesia

**Abstract:** In reinsurance policy, the insurer needs to pay a reinsurance premium to the reinsurer. Nonetheless, the insurer will get an insurance premium from the insurance policyholder. This research addresses choosing an optimal reinsurance policy using a more realistic model, i.e., individual risk model with random insurance premiums. Here, we assume that the insurance premiums are random, and the reinsurance premiums is charged by the expected value principle. We use minimizing the risk exposure of the insurer as optimization criteria. The insurer's risk exposure is measured by tail risk measure of the net cost of the insurer. To illustrate the applicability of our results, we consider an insurance company has two lines of business and derive the optimal reinsurance explicitly for Combination of Proportional and Excess of loss Reinsurance.

**Keywords:** Random premium, Net cost, Proportional, Excess of loss

**A novel inventory policy for imperfect items with stock dependent demand rate**

Praveen V P* and Manoharan M

Department of Statistics, University of Calicut, India

**Abstract:** Adoption of trade credit financing policy is prevalent in inventory management as an important strategy to increase profitability with a major motive of attracting new customers and also to avoid escalating price competition. We revisit an economic order quantity model under conditionally permissible delay in payments, fix a certain period to settling the account, during this period supplier charges no interest, but beyond this period interest is being charged. On the other hand, retailer can earn interest on the revenue generated during this period. Optimal inventory policy can be managed with explicitly specifying the demand for fresh produce to be a function of its freshness.
expiration date and displayed volume. Shortages are allowed and it is partially backlogged. Keeping this scenario in mind, an attempt has been made to formulate an inventory policy for imperfect items with permissible delay in payments and expiration date under freshness and stock dependent demand rate. The formulated model is illustrated through numerical examples to determine the effectiveness of the proposed model. Further, the effects of changes of different inventory parameters have been studied by a sensitivity analysis.

**Keywords:** Imperfect items, Permissible delay in payments, Stock dependent demand, Shortage with Partial backlogging

---

**Density deconvolution with non-standard error distributions: Rates of convergence and adaptive estimation**

Taeho Kim* and Alexander Goldenshluger
Statistics, University of Haifa, Israel

**Abstract:** It is a standard assumption in the density deconvolution problem that the characteristic function of the measurement error distribution is non-zero on the real line. While this condition is assumed in the majority of existing works on the topic, there are many problem instances of interest where it is violated. In this paper, we focus on non-standard settings where the characteristic function of the measurement errors has zeros, and study how zeros multiplicity affects the estimation accuracy. For a prototypical problem of this type, we demonstrate that the best achievable estimation accuracy is determined by the multiplicity of zeros, the rate of decay of the error characteristic function, as well as by the smoothness and the tail behavior of the estimated density. We derive lower bounds on the minimax risk and develop optimal in the minimax sense estimators. In addition, we consider the problem of adaptive estimation and propose a data-driven estimator that automatically adapts to unknown smoothness and tail behavior of the density to be estimated.

**Keywords:** Density Deconvolution, Non-standard Measurement Error, Minimax Risk, Zero Multiplicity

---

**Moments of the doubly truncated selection elliptical distributions: recurrence, existence and applications**

Christian Galarza*, Victor Lachos and Larissa Ávila

1Department of Mathematics, Escuela Superior Politécnica del Litoral, Ecuador
2Department of Statistics, University of Connecticut, USA
3Department of Statistics, Universidade Estadual de Campinas, Brazil

**Abstract:** We compute doubly truncated moments for the selection elliptical (SE) class of distributions, which includes some multivariate asymmetric versions of well-known elliptical distributions, such as, the normal, Student’s t, among others. We address the moments for doubly truncated members of this family, establishing neat formulation for high order moments as well as for its first two moments. We establish sufficient and necessary conditions for their existence. Further, we propose computational efficient methods to deal with extreme settings of the parameters, partitions with almost zero volume or no truncation. Applications and simulation studies are presented in order to illustrate the usefulness of the proposed methods.

**Keywords:** Truncated distributions, Truncated moments, Elliptical distributions, Higher moments
Characterization of probability distributions by a generalized notion of sufficiency and Fisher information
Atin Gayen* and M. Ashok Kumar
Mathematics, Indian Institute of Technology Palakkad, Kerala, India

Abstract: The notion of sufficiency introduced by Fisher is based on the usual likelihood function. This is useful particularly when the underlying model is exponential. We propose a generalized notion of principle of sufficiency based on two generalized likelihood functions, namely Basu et al. and Jones et al. likelihood functions that arise in robust inference. We find the specific form of the family of probability distributions that have a fixed number of sufficient statistics (independent of sample size) with respect to these likelihood functions. These distributions are of power-law form and are a generalization of the exponential family. Student distributions are a special case of this family. We also extend the concept of minimal sufficiency with respect to this generalized notion and find a minimal sufficient statistic for Student distributions. We observe that the generalized estimators of parameters of Student distributions are functions of the minimal sufficient statistics derived from this generalized notion. We finally show that these estimators are also efficient in the sense that variance of each of these estimators equals the variance given by the asymptotic normality result.

Keywords: Generalized likelihood function, Power-law model, Student distribution, Sufficient statistics
GMOTE: Gaussian-based minority oversampling technique for imbalanced classification adapting tail probability of outliers
Seung Jee Yang1*, Kyung-Joon Cha2
1Department of Applied Statistics and Institute for Convergence of Basic Science, Hanyang University, Korea
2Department of Mathematics and Institute for Convergence of Basic Science, Hanyang University, Korea

Abstract: Imbalanced data substantially affects the performance of the standard classification models. As a solution to these, oversampling methods have been proposed such as the synthetic minority oversampling technique (SMOTE). However, because methods such as SMOTE use linear interpolation to generate synthetic instances, the synthetic data space may appear similar to a polygon. Furthermore, oversampling methods generate synthetic outliers in minority classes. In this paper, we propose a Gaussian-based minority oversampling technique (GMOTE) with a statistical perspective for imbalanced datasets. The proposed method generates instances by using a Gaussian mixture model to avoid linear interpolation and to consider outliers. Motivated by the clustering-based multivariate Gaussian outlier score, we propose considering local outliers by calculating the tail probability of instances calculated using the Mahalanobis distance. Experiments were conducted on a representative set of benchmark datasets, and the GMOTE performance was compared with that of other methods. When GMOTE is combined with a classification and regression tree or support vector machine, it produces better accuracy and F1-score. Experimental results demonstrate this robust performance.

Acknowledgements: This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1A6A1A06046728).

Keywords: Imbalanced data, Oversampling, Gaussian mixture model, Tail probability

Exact inference for an exponential parameter under generalized progressive type II hybrid censored competing risk data
Subin Cho1*, Insang Hwang1 and Kyeongjun Lee2
1Statistics, Daegu University, Korea
2Division of Mathematics and Big Data Science, Daegu University, Korea

Abstract: Progressive censoring has the drawback that it might take a very long time to observe m-th failures and complete the life test. In this reason, generalized progressive type II censoring scheme was introduced. In addition, it is known that more than one risk factor may be present at the same time. In this paper, we discuss exact inference for competing risk model with generalized progressive type II hybrid censored exponential data. We derive the conditional moment generating function of the maximum likelihood estimators of scale parameters of exponential distribution and the resulting lower confidence bound under generalized progressive type II hybrid censoring scheme. From the example data, it can be seen that the PDF of MLE is almost symmetrical.

Keywords: Exact conditional distribution, Generalized progressive type II censoring scheme, Maximum likelihood estimator
Meta-analysis methods for multiple related markers: Applications to microbiome studies with the results on multiple α-diversity indices

Hyunwook Koh*
Department of Applied Mathematics and Statistics, The State University of New York, Korea

Abstract: Meta-analysis is a practical and powerful analytic tool that enables a unified statistical inference across the results from multiple studies. Notably, researchers often report the results on multiple related markers in each study (e.g., various α-diversity indices in microbiome studies). However, univariate meta-analyses are limited to combining the results on a single common marker at a time, whereas existing multivariate meta-analyses are limited to the situations where marker-by-marker correlations are given in each study. Thus, here we introduce two meta-analysis methods, namely, multi-marker meta-analysis (mMeta) and adaptive multi-marker meta-analysis (aMeta), to combine multiple studies throughout multiple related markers with no priori results on marker-by-marker correlations. mMeta is a statistical estimator for a pooled estimate and its standard error across all the studies and markers, whereas aMeta is a statistical test based on the test statistic of the minimum p-value among marker-specific meta-analyses. mMeta conducts both effect estimation and hypothesis testing based on a weighted average of marker-specific pooled estimates while estimating marker-by-marker correlations non-parametrically via permutations, yet its power is only moderate. In contrast, aMeta closely approaches the highest power among marker-specific meta-analyses, yet it is limited to hypothesis testing. While their applications can be broader, we illustrate the use of mMeta and aMeta to combine microbiome studies throughout multiple α-diversity indices. We evaluate mMeta and aMeta in silico and apply them to real microbiome studies on the disparity in α-diversity by the status of HIV infection.

Keywords: Multivariate Meta-Analysis, Random-Effects Meta-Analysis, Human Microbiome

Estimation for a nonlinear regression model with non-zero mean errors and an application to a biomechanical model

Hojun You*
Department of Statistics, Seoul National University, Korea

Abstract: We propose a modified least squares estimator for a nonlinear regression model with non-zero mean errors motivated by the head-neck position tracking application. A nonlinear regression with multiplicative errors can be handled under the framework of the proposed method. In addition, we assume temporal dependence in the errors. We propose not only the modified least squares procedure for parameter estimation, but also penalized least squares procedure for parameter estimation and selection at the same time. Asymptotic properties of the proposed estimators, especially local consistency and oracle property of the penalized least square estimator, are established under plausible assumptions imposed on the nonlinear function, errors, and a penalty function. A simulation study demonstrates that the proposed estimation performs well in both parameter estimation and selection with temporally correlated error. The analysis and comparison with the existing methods for head-neck position tracking data show better performance of the proposed method in terms of the variance accounted for (VAF).

Keywords: Multiplicative regression, Nonlinear regression, Oracle property, Strong mixing

Neural network-based clustering for ischemic stroke patients

Su Hoon Choi* and Joon-Tae Kim
1Department of Mathematics and Statistics, Chonnam National University, Korea
2Department of Neurology, Chonnam National University Medical School, Chonnam National University Hospital, Korea
Abstract: Finding similar clusters for stroke patients is important because it can lead to discovering new patterns and more effective ways to manage stroke. Although lifetime clustering is an important tool, it remains a relatively unexplored topic. In general, the degree of risk is classified using SPI-II, a traditional risk score to stratify the risk of recurrence of stroke. SPI-II is a verified and reliable stroke risk score. However, existing tools for predicting stroke outcome risks may have limitations because all possible variables cannot be considered. In this study, we compare several lifetime clustering methods including the deep lifetime clustering (DLC) method, which is a neural network-based clustering model. The performance of the clustering method on the real-world survival datasets of patients with ischemic stroke was evaluated. The SPI-II scores are grouped into three groups: low, medium, and high risk, based on previous studies. Accordingly, we conduct an analysis on three clustering in all methods. The metrics used to evaluate clusters obtained from the method are Concordance index, Brier score, and Log-rank score. An analysis was conducted on 7,650 patients out of data from the local comprehensive stroke center registry in patients with acute ischemic stroke. Compared to SPI-II stroke risk scores and other clustering methods, the DLC model performed much better clustering for all evaluation index. These results suggest that the DLC method may be useful for grouping stroke patients with similar outcome risks. Our study had an inherent limitation that it included only data from a single stroke center register in Republic of Korea. Therefore, further research with independent cohorts is likely to be required. Nevertheless, the neural network-based clustering method was first applied to stroke patients on real-world datasets.

Keywords: Ischemic stroke, Cluster model, Deep lifetime clustering

Principal component analysis of amplitude and phase variation in multivariate functional data
Soobin Kim* and Youngwook Kwon
Department of Statistics, Seoul National University, Korea

Abstract: In many situations, multivariate functional data have both phase and amplitude variations. A common approach is to remove phase variations using selected function aligning methods and then apply functional principal component analysis (FPCA) to the aligned functions, which contain only amplitude variations. To consider both types of variations, we propose an extension of FPCA for amplitude and phase variation to multivariate cases. The original functions are decomposed into amplitude functions and warping functions, and warping functions are transformed into square-integrable functions via a centered log-ratio transformation. Multivariate FPCA is then performed on each amplitude and phase component with data-adaptive weights to balance the variational effects. The proposed method demonstrates its usefulness through real data analysis with sea climate data in Korea.

Keywords: Multivariate functional data, Principal component analysis, Amplitude and phase variations, Climatological data

Clustering non-stationary advanced metering infrastructure data
Donghyun Kang* and Yaeji Lim
Department of Applied Statistics, Chung-Ang university, Korea

Abstract: We propose a clustering method for advanced metering infrastructure (AMI) data in Korea. As AMI data present non-stationarity, we consider time-dependent frequency domain principal components analysis and develop a new clustering method based on the time-varying eigenvectors. Our method provides a meaningful result that is different from the clustering results obtained by employing conventional methods, such as K-means and K-centres functional clustering. We further apply the clustering results to the evaluation of the electricity price system in South Korea, and validate the reform of the progressive electricity tariff system.

Keywords: Multivariate functional data, Principal component analysis, Amplitude and phase variations, Climatological data
**Keywords:** Feature-based clustering for electricity use time series data, A study for clustering method to generate typical load profiles for smart grid

---

**Poster I-2**

JULY 20 (TUE), 21:30 - 22:00 KST

**Geometrically adapted Langevin algorithm (GALA) for Markov Chain Monte Carlo (MCMC) simulations**

Mariya Mamajiwala*

Statistical Science, University College London, UK

**Abstract:** MCMC is a class of methods to sample from a given probability distribution. Of its myriad variants, the one based on the simulation of Langevin dynamics, which approaches the target distribution asymptotically, has gained prominence. The dynamics is specifically captured through a Stochastic Differential Equation (SDE), with the drift term given by the gradient of the log-likelihood function with respect to the parameters of the distribution. However, the unbounded variation of the noise (i.e. the diffusion term) tends to slow down the convergence, which limits the usefulness of the method. By recognizing that the solution of the Langevin dynamics may be interpreted as evolving on a suitably constructed Riemannian Manifold (RM), considerable improvement in the performance of the method can be realised. Specifically, based on the notion of stochastic development - a concept available in the differential geometric treatment of SDEs - we propose a geometrically adapted variant of MCMC. Unlike the standard Euclidean case, in our setting, the drift term in the modified MCMC dynamics is constrained within the tangent space of an RM defined through the Fisher information metric and the related connection. We show, through extensive numerical simulations, how such a mathematically tenable geometric restriction of the flow enables a significantly faster and accurate convergence of the algorithm.

**Keywords:** MCMC, Riemannian differential geometry

---

**Bayes estimation for the Weibull distribution under generalized adaptive hybrid progressive censored competing risks data**

Yeongjae Seong1,*, Hyein Koo1 and Kyeongjun Lee2

1Statistics, Daegu University, Korea

2Division of Mathematics and Big Data Science, Daegu University, Korea

**Abstract:** Adaptive progressive hybrid censoring schemes have become quite popular in reliability and lifetime-testing studies. However, the drawback of the adaptive progressive hybrid censoring scheme is that it might take a very long time in order to complete the life test. In this reason, generalized adaptive progressive hybrid censoring scheme was introduced. In this research, a competing risks model is considered under a generalized adaptive progressive hybrid censoring scheme. When the failure times are Weibull distributed, maximum likelihood estimates for the unknown model parameters are established where the associated existence and uniqueness are shown. An asymptotic distribution of the maximum likelihood estimators is used to construct approximate confidence intervals via the observed fisher information matrix. Moreover, Bayes point estimates and the highest probability density credible intervals of unknown parameters are also presented, and the Gibbs sampling technique is used to approximate corresponding estimates.

**Keywords:** Bayes estimation, Competing risks, Generalized adaptive progressive hybrid censoring
Large deviations of mean-field interacting particle systems in a fast varying environment
Sarath Yasodharan* and Rajesh Sundaresan
Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India

Abstract: We study large deviations of a "fully coupled" finite state mean-field interacting particle system in a fast varying environment. The empirical measure of the particles evolves in the slow time scale and the random environment evolves in the fast time scale. Our main result is the path-space large deviation principle for the joint law of the empirical measure process of the particles and the occupation measure process of the fast environment. This extends previous results known for two time scale diffusions to two time scale mean-field models with jumps. Our proof is based on the method of stochastic exponentials. We characterise the rate function by studying a certain variational problem associated with an exponential martingale.

Keywords: Mean-field interaction, Large deviations, Time scale separation, Metastability

Stochastic homogenisation of Gaussian fields
Leandro Chiarini* and Wioletta Ruszel
Department of Mathematics, Utrecht University, Netherlands

Abstract: In this poster we prove the convergence of a sequence of random fields that generalise the Gaussian Free Field and bi-Laplacian field. Such fields are defined in terms of non-homogeneous elliptic operators which will be sampled at random. Under standard assumptions of stochastic homogenisation, we identify the limit fields as the usual GFF and bi-Laplacian fields up to a multiplicative constant.

Keywords: Stochastic Homogenisation, Gaussian Fields

Concentration inequality for U-statistics for uniformly ergodic Markov chains, and applications
Quentin Duchemin*, Yohann De Castro and Claire Lacour
LAMA, Université Gustave Eiffel, France
Institut Camille Jordan, Centrale Lyon, France

Abstract: The non-asymptotic analysis of the tail behaviour of U-statistics in a dependent framework has been so far very little touched. This work presents a new concentration inequality for U-statistics of order 2 for uniformly ergodic Markov chains. We consider a uniformly ergodic Markov chain \( (X_i)_t \) on a general state space with invariant distribution \( \pi \) and transition kernel \( P \). We work with bounded functions, say \( h_{i,j} \) that are assumed to be \( \pi \)-canonical, a notion that naturally extends the canonical/degenerate property of the independent setting. We further assume that the transition kernel \( P \) of the chain can be upper-bounded, namely for some probability measure \( \nu \) and some constant \( c > 0 \), \( \sup_x P(x, \cdot) \leq c \nu(\cdot) \). This condition allows us to get back to a decoupled framework and we give several natural examples for which such condition holds. Under the previous assumptions, we provide a Bernstein-type concentration inequality for U-statistics of the form \( \sum_{i,j} h_{i,j}(X_i, X_j) \). Note that we recover the convergence rate obtained for the independent setting in the degenerate case. Using our new concentration inequality, we prove for the first time a non-asymptotic result of convergence of spectra for kernel matrices towards spectrum of integral operator that are not of positive-type (i.e., with positive and negative eigenvalues). Our result allows for a dependence of the kernels \( h_{i,j} \) with the indexes in the sums, which prevents the use of standard blocking tools. Hence our framework requires specific proof technics. Our approach is based on an inductive analysis where we use martingale techniques, uniform ergodicity and Nummelin splitting. Let us finally mention that our
assumptions typically hold for compact measurable spaces. That's why future work will include the extension of our result to milder assumptions to go beyond compactness, which would be of great interest for MCMC methods.

**Keywords:** U-statistics, Markov chains, Concentration inequality

---

**A Bayesian illness-death model to approach the incidence of recurrent hip fracture and death in elderly patients**

Fran Llopis-Cardona, Carmen Armero and Gabriel Sanfélix-Gimeno

1Health Services Research Unit, Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO), Spain

2Department of Statistics and Operational Research, Universitat de València, Spain

**Abstract:** Multi-state models are a wide class of stochastic processes models in which individuals can move between different states over time. These models are of special interest in survival analysis as they allow to deal with a wide range of complex scenarios. We focus on the so-called illness-death models, which include an initial state, an illness state, and a death state. They perform well with non-terminal diseases in which progression to death is a relevant outcome. We use an illness-death model to study the evolution of patients who have suffered a hip fracture. The dataset comes from the PREV2FO cohort and includes 34,491 patients aged 65 years and older who were discharged alive after a hospitalization for an osteoporotic hip fracture and followed until a recurrent hip fracture and death. Transition times, from the initial fracture to refracture and death, and from refracture to death, are modelled via Cox proportional hazards models with Weibull baseline hazard functions, the latter with regard to the time from fracture to refracture. For simplicity, we adjusted by covariates sex and age at discharge. We use a Bayesian approach to estimate the posterior distribution of the parameters of the model via Markov Chain Monte Carlo Methods (MCMC). Based on this distribution, we estimate posterior distributions for cumulative incidences of refracture and death, as well as transition probabilities which include free-event probability, probability of permanence at refracture state and the probability of death after refracture. We also estimate hazard ratios to assess the effect of covariates on each transition. Women and men were estimated to be at the same risk of refracture, but women showed higher incidences of refracture. This is due to the higher risks and incidences of death among men, preventing them from refractures. On the other hand, age was estimated to increase every hazard.

**Keywords:** Cause-specific hazard function, Cumulative incidence function, Multi-state model, Transition probability

---

**The contact process with two types of particles and priority: metastability and convergence in infinite volume**

Mariela Pentón Machado

Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil

**Abstract:** We consider a symmetric finite-range contact process on Z with two types of particles (or infections), which propagate according to the same supercritical rate and die (or heal) at rate 1. Particles of type 1 can occupy any site in \((-\infty, 0]\) that is empty or occupied by a particle of type 2 and, analogously, particles of type 2 can occupy any site in \([1, +\infty]\) that is empty or occupied by a particle of type 1. We prove that this system exhibits two metastable states: one with the two species and the other one with the family that survives the competition. In addition, we study the convergence of the process when it is defined in infinite volume.

**Keywords:** Contact process
A nonparametric instrumental approach to endogeneity in competing risks models
Jad Beyhum1*, Jean-Pierre Florens2 and Ingrid Van Keilegom1
1ORSTAT, KU Leuven, Belgium
2Toulouse School of Economics, Université Toulouse Capitole, France

Abstract: This paper discusses endogenous treatment models with duration outcomes, competing risks and random right censoring. The endogeneity issue is solved using a discrete instrumental variable. We show that the competing risks model generates a non-parametric quantile instrumental regression problem. The cause-specific cumulative incidence, the cause-specific hazard and the subdistribution hazard can be recovered from the regression function. A distinguishing feature of the model is that censoring and competing risks prevent identification at some quantiles. We characterize the set of quantiles for which exact identification is possible and give partial identification results for other quantiles. We outline an estimation procedure and discuss its properties. The finite sample performance of the estimator is evaluated through simulations. We apply the proposed method to the Health Insurance Plan of Greater New York experiment.

Keywords: Instrumental variable, Survival analysis, Competing risks, Endogeneity

Nonconstant error variance in generalized propensity score model
Doyoung Kim* and Chanmin Kim
Statistics, Sungkyunkwan University, Korea

Abstract: In observational study, the most salient challenge is to adjust for confounders to mimic randomized experiment. In the setting of more than two treatment levels, several generalized propensity score (GPS) models have been proposed to balance covariates among treatment groups. Those models assume some parametric forms for treatment variable distributions especially with constant variance assumption. With the existence of heteroskedasticity, the constant variance assumption might affect the existing propensity score methods and the causal effect of interest. In this paper, we propose a novel GPS method to handle non-constant variance in the treatment model by extending Xiao et al. (2020) with weighted least squares method. We conduct a set simulation studies and show that the proposed method outperforms in terms of covariate balance and low bias in causal effect estimates.

Keywords: Causal Inference, Continuous Treatment, Propensity Score, Observational Study

Causal mediation analysis with multiple mediators of general structures
Youngho Bae* and Chanmin Kim
Statistics, Sungkyunkwan University, Korea

Abstract: In assessing causal mediation effects, a challenge is that there can be more than one mediator on pathways from treatment to outcome. More precisely, we do not know exactly how many mediators are in the causal path and how they relate to each other. A few approaches have been proposed to estimate direct and indirect effects in the presence of two causally independent or dependent mediators. However, those methods cannot be generalized to settings of more than two mediators where causally independent and dependent mediators coexist. We propose a novel approach to identify direct and indirect effects under a general situation of multiple mediators: two causally dependent mediators (V,W) and one causally independent mediator (M). With our proposed sequential ignorability
assumption, the overall treatment effect can be decomposed into direct and mediator-specific indirect effects. A sensitivity analysis strategy is developed for testing the proposed identifying assumptions. We can try to apply this method to the pollination data. In other words, we may use this approach to estimate the effect of a particular emission control technology, that installed on power plants, on ambient pollution where power plant emissions are potential mediators.

**Keywords:** Causal inference, Effect decomposition, Sequential ignorability, Sensitivity analysis

---

**A fuzzy clustering ensemble based Mapper algorithm**

SungJin Kang* and Yaeji Lim
Department of Statistics, Chung-Ang University, Korea

**Abstract:** Mapper is a popular topological data analysis method to analyze structure of the complex high-dimensional dataset. Since Mapper algorithm can be applied to the clustering and feature selection with visualization, it is used in various fields such as biology, chemistry, etc. However, there are some resolution parameters to be chosen before applying the Mapper algorithm, and the results are sensitive to these selection. In this paper, we focus on the selection of the two resolution parameters, the number of intervals, and the overlapping percentage. We propose a new parameter selection method in Mapper based on ensemble technique. We generate multiple Mapper results under various parameters, and apply the fuzzy clustering ensemble method to combine the results. Three real data are considered to evaluate mapper algorithms including proposed one, and the results demonstrate the superiority of the proposed ensemble Mapper method.

**Keywords:** Mapper, Topological data analysis, Ensemble, Fuzzy clustering ensemble method

---

**Analysis of the association between suicide attempts and meteorological factors**

Seunghyeon Kim* and Eunsik Park
Department of Mathematics & Statistics, Chonnam National University, Korea

**Abstract:** Several studies indicate that there is an association between suicide and meteorological factors, particularly an increase in ambient temperature increases the risk of suicide. Although suicide attempts are highly likely to lead to suicide in the future, research on the relationship between suicide attempts and meteorological factors is not done much. We evaluated the association between suicide attempts and meteorological factors and examined gender and age differences. We studied 30,012 people who attempted suicide and hospitalized in the emergency room of medical institutions located in Seoul from January 1, 2014, to December 31, 2018. This information was provided by the National Emergency Department Information System data. Seven meteorological factors were studied: daily lowest temperature, highest temperature, average temperature, daily temperature difference, average relative humidity, sunshine duration, and average cloud cover in Seoul during the same period. Meteorological factors were categorized, and the daily Age-standardized Suicide Attempt rate (per 100,000) (ASDAR) was defined for each category. Subgroup analysis by gender and age was done to explore the association between meteorological factors and suicide attempts. From 2014 to 2018, the ASDAR was 61.3. The ASDAR for women was 69.3 and for men was 52.8, the highest suicide attempts by age in their 20s. In terms of the seven meteorological factors, suicide attempts increased as the lowest temperature, the highest temperature, the average temperature, and the relative humidity increased. Both genders showed an increase in suicide attempts as the lowest, the highest, the average temperature, and the relative humidity increased and showed the same trend in all ages except for women in their 20s. We found that the risk of suicide attempts increases as temperature and relative humidity increase. These results suggest that exposure to high temperatures can be a suicide attempt-inducing factor.

**Keywords:** Suicide attempt, Meteorological factor, Descriptive analysis
Spectral clustering with the Wasserstein distance and its application
Sanghun Jeong*, Mina Baek, Hojin Yang and Choongrak Kim
Department of Statistics, Pusan National University, Korea

Abstract: The advance of modern automatic devices can produce a massive number of samples from the population of the individual subject. Although this development allows us to access the entire distributional structure for the population of each individual subject, traditional approaches tend to focus on detecting the local feature to recognize the pattern of the data. In this project, we consider the pattern recognition problem classifying the subject specific distributions into a few categories after estimating the subject specific distributions. Suggested approach consists of three stages procedure including the probability density estimation, the dissimilarity computation, and the clustering computation. Specifically, we use the kernel density estimator for the subject specific distribution in the first stage. Then, we focus on the Wasserstein distance to account for the dissimilarity between these distributions while using the optimal transport map for distance. Finally, we use such a dissimilar measure to figure out the structure of the Laplacian graph and conduct the spectral clustering to deal with these distributions contained not in the Euclidean space but some nonlinear space. We will demonstrate the benefit of the spectral clustering with the Wasserstein distance through simulation studies, applying our suggested method to the real data.

Keywords: Distribution function, Eigenvalue, Kernel estimator, Laplacian

Robust covariance estimation for partially observed functional data
Hyunsung Kim*, Yaeji Lim and Yeonjoo Park
Chung-Ang University

Abstract: In recent years, applications have emerged that produce partially observed functional data, where each trajectory is collected over individual-specific subinterval(s) within the whole domain of interest. Robustness to atypical partially observed curves in the application is a practical concern, especially in the dimension reduction step through functional principal component analysis (FPCA). Existing studies implemented FPCA by applying smoothing techniques to estimate mean and covariance functions under irregular functional data structure, however, its estimation is easily affected by outlying curves with heavy-tailed noises or spikes. In this study, we investigate the robust method for the covariance estimation by using bounded loss function, and it enables us to obtain robust functional principal components under partially observed functional data. Using the functional principal scores, we reconstruct the missing parts of trajectories. Numerical experiments show that our method provides a stable and robust estimation when the data contain the atypical curves.

Keywords: partially observed functional data, Covariance estimation, Principal component analysis, Robustness

Fast Bayesian functional regression for non-Gaussian spatial data
Yeo Jin Jung*
Department of Statistics and Data Science, Yonsei University, Korea

Abstract: Functional generalized linear models (FGLM) have been widely used to study the relations between non-Gaussian response and functional covariates. However, most existing works assume independence among observations and therefore have limited applicability on correlated data. A particularly important example is functional data with spatial correlation, where we observe functions over spatial domains, such as the age population curve or temperature curve at each areal unit. In this paper, we extend FGLM by incorporating spatial random effects. However, such models have computational and inferential challenges. The high-dimensional spatial random effects cause the slow mixing of Markov chain Monte Carlo (MCMC) algorithms. Furthermore, spatial confounding can lead
to bias in parameter estimates and inflate their variances. To address these issues, we propose an efficient Bayesian method using a sparse reparameterization of high-dimensional random effects. Furthermore, we study an often-overlooked challenge in functional spatial regression: practical issues in obtaining credible bands of functional parameters and assessing whether they provide nominal coverage. We apply our methods to simulated and real data examples, including malaria incidence data and US COVID-19 data. The proposed method is fast while providing accurate functional estimates.

**Keywords:** Functional regression, Non-Gaussian spatial data, Markov chain Monte Carlo, Gaussian Markov random fields

---

**Busemann process and semi-infinite geodesics in Brownian last-passage percolation**

Evan Sorensen* and Timo Seppäläinen

Mathematics, University of Wisconsin-Madison, USA

**Abstract:** We prove the existence of semi-infinite geodesics for Brownian last-passage percolation (BLPP). Specifically, on a single event of probability one, there exist semi-infinite geodesics, started from every space-time point and traveling in every asymptotic direction. Properties of these geodesics include uniqueness for a fixed initial point and direction, non-uniqueness for fixed direction but random initial points, and coalescence of all geodesics traveling in a common, fixed direction. The semi-infinite geodesics are constructed from Busemann functions, whose existence was proved for fixed initial points and directions by Alberts, Rassoul-Agha, and Simper. We extend their result to a global process of Busemann functions and derive the joint distribution of Busemann functions for varying directions. From this joint distribution, we prove results about the geometry of the semi-infinite geodesics. More specifically, there exists a Hausdorff dimension 1/2 set of initial points, and to each point an associated direction, such that there are two semi-infinite geodesics in that direction whose only shared point is the initial point.

**Keywords:** Brownian last-passage percolation

---

**Application of kernel mean embeddings to functional data**

George Wynne*

Mathematics, Imperial College London, UK

**Abstract:** Kernel mean embeddings (KMEs) have enjoyed wide success in statistical machine learning over the past fifteen years. They offer a non-parametric method of reasoning with probability measures by mapping measures into a reproducing kernel Hilbert space. Much of the existing theory and practice has revolved around Euclidean data whereas functional data has received very little investigation. Likewise, in functional data analysis (FDA) the technique of KMEs has not been explored. This work proposes to bridge this gap in theory and practice. KMEs offer an alternative paradigm than the common practice in FDA of projecting data to finite dimensions. The KME framework can handle infinite dimensional input spaces, offers an elegant theory and leverages the spectral structure of functional data. Empirically, KMEs provide competitive performance against existing functional two-sample and goodness-of-fit tests. Finally, we discuss connections to empirical characteristic function based testing and functional depth techniques currently used in FDA.

**Keywords:** Functional Data Analysis, Kernel Methods, Testing
**SIR-based examination of the policy effects on the COVID-19 spread in U.S.**

David Han\(^1\)*, Macy K. Belle\(^1\), Mina Song\(^1\) and Aaron Mendlovitz\(^2\)

1Management Science and Statistics, The University of Texas at San Antonio, USA
2Mathematics, The University of Texas at San Antonio, USA

**Abstract:** Since the global outbreak of the novel COVID-19, many research groups have studied the epidemiology of the virus for short-term forecasts and to formulate the effective disease containment and mitigation strategies. The major challenge lies in the proper assessment of epidemiological parameters over time and of how they are modulated by the effect of any publicly announced interventions. Here we attempt to examine and quantify the effects of various (legal) policies/orders in place to mandate social distancing and to flatten the curve in each of the U.S. states. Through Bayesian inference on the stochastic SIR models of the virus spread, the effectiveness of each policy on reducing the magnitude of the growth rate of new infections is investigated statistically. This will inform the public and policymakers, and help them understand the most effective actions to fight against the current and future pandemics. It will aid the policy-makers to respond more rapidly (select, tighten, and/or loosen appropriate measures) to stop/mitigate the pandemic early on.

**Keywords:** Bayesian inference, COVID-19 pandemics, SIR compartmental models

**Cross-validation confidence intervals for test error**

Alexandre Bayle*

Statistics, Harvard University, USA

**Abstract:** This work develops central limit theorems for cross-validation and consistent estimators of its asymptotic variance under weak stability conditions on the learning algorithm. Together, these results provide practical, asymptotically-exact confidence intervals for k-fold test error and valid, powerful hypothesis tests of whether one learning algorithm has smaller k-fold test error than another. These results are also the first of their kind for the popular choice of leave-one-out cross-validation. In our real-data experiments with diverse learning algorithms, the resulting intervals and tests outperform the most popular alternative methods from the literature.

**Keywords:** Cross-validation, Asymptotic behavior, Algorithm comparison, Confidence intervals

**Comparison of quantile regression curves under different settings with censored data**

Lorenzo Tedesco*

Business and Economics Faculty, ORSTAT, Belgium

**Abstract:** The poster presents a new nonparametric test for conditional quantile curves equality when the outcome of interest, typically a duration, is subjected to right censoring. The test is based on a quantile regression estimation models and do not rely on distributional assumptions. Moreover, the proposed method holds for both dependent and independent samples. Consistency of the test and asymptotic results are also provided together with a bootstrap procedure which is intended to avoid density estimations in case of small sample sizes. The poster also includes a comparison with other methods and examples of application for both dependent and independent setting.

**Keywords:** Censored data, Survival analysis, Quantile regression
A penalized matrix normal mixture model for clustering matrix data
Jinwon Heo* and Jangsun Baek
Department of Mathematics & Statistics, Chonnam National University, Korea

Abstract: Along with the advance of technologies, matrix data such as medical/industrial images have been emerged in many practical fields. These data usually have high dimension and are not easy to be clustered due to its intrinsic correlated structure among rows and columns. Most approaches convert matrix data to multi-dimensional vector and apply conventional clustering methods to them, and hence suffer from extreme high-dimensionality problem as well as lack of interpretability of the correlated structure among row/column variables. Gao et. al.(2020) proposed a regularized mixture model for clustering matrix-valued data by imposing a sparsity structure for the mean signal of each cluster. We extend their approach by regularizing further on the covariance to cope with curse of dimensionality for images with large size. We propose a penalized matrix-normal mixture model with lasso-type penalty terms in both mean and covariance matrices, and then develop an expectation maximization algorithm to estimate the parameters. We apply the proposed method to simulated data as well as real data sets, and confirm its clustering accuracy performance over some conventional methods.

Keywords: Clustering, Matrix normal distribution, Expectation maximization algorithm, Penalized matrix-normal mixture model

Univariate and multivariate normality tests using an entropy-based transformation
Shahzad Munir* and Andrew Pua
Wang Yanan Institute for Studies in Economics, Xiamen University, China

Abstract: We introduce a new normality test which may be applied to univariate and multivariate IID or time series data. In the univariate case, the test is constructed by first applying a transformation based on the definition of entropy. The test only requires the estimation of the variance of transformed data; however, it is less sensitive to errors from estimating the kurtosis coefficient but is able to detect deviations from this higher-order moment. In the univariate case, we show that in a broad class of stationary processes, the proposed test statistic asymptotically follows a standard normal distribution, and does not require any kernel smoothing to consistently estimate the asymptotic variance of the proposed test. The extension to the multivariate case is also straightforward and allows for alternatives to diagnostic testing of vector autoregressive models.

Keywords: Entropy-based transformation, Kurtosis, Multivariate data, Non-IID data

Geum river network data analysis via weighted PCA
Seeun Park* and Kyusoon Kim
Statistics, Seoul National University, Korea

Abstract: Various measurements of water quality are collected at monitoring sites, spread throughout the river network. Monitoring this kind of dataset is critical for water quality evaluation and improvement, but the unique structure of the river network interrupts PCA, achieving accurate results due to autocorrelation among variables. In literature, Gallacher et al. (2017) introduced a weighted PCA that reflects the known spatiotemporal structure of the river network to adjust the autocorrelation. This study aims to apply the weighted PCA method to Geum River
network data in South Korea and improve the method itself. As a result, the weighted PCA successfully identified certain patterns in Geum River data that the conventional PCA cannot process. However, we believe that the weighted PCA method does not take into account the inhomogeneity on the covariance structure of the data, which might lead to inaccurate results in PCA. In fact, inhomogeneous covariance structures are found in Geum River data across regions and seasons. Therefore, our further plan is to improve the weighted PCA that can handle this problem due to the inhomogeneous structure.

**Keywords:** River network data, Flow direction, Weighted PCA, Inhomogeneous covariance structure

---

**Cauchy combination test with thresholding under arbitrary dependency structures**

Junsik Kim*
Statistics, Seoul National University, Korea

**Abstract:** Combining individual p-values to aggregate sparse and weak effects is a substantial interest in large-scale data analysis. The individual p-values or test statistics are often correlated, although many p-values combining methods are developed under i.i.d. assumption. The Cauchy combination test is a method to combine p-values for this arbitrary dependence structure, but in practice, type I error increases as the correlation increases. In this paper, we propose a global test that extends the Cauchy combination test by thresholding arbitrarily dependent p-values. Under an arbitrary dependence structure, we show that the tail probability of the proposed method is asymptotically equivalent to that of the Cauchy distribution. In addition, we show that the power of the proposed test achieves the optimal detection boundary asymptotically in a strong sparsity condition. Extensive simulation results show that the power of the proposed test is robust to correlation coefficients and more powerful under a sparse situation. As a case study, we apply the proposed test to GWAS of Inflammatory bowel disease (IBD).

**Keywords:** Combining p-values, Cauchy distribution, Global hypothesis testing, GWAS

---

**Control charts for monitoring linear profiles in the detection of network intrusion**

Daeun Kim1* and Sungim Lee2
1Statistics, Dankook University, Korea
2Information Statistics, Dankook University, Korea

**Abstract:** This study considers the problem of network intrusion detection. Sklavounos et al. (2017) proposed using EWMA charts for crucial characteristics as a network intrusion detection method. This paper expands on this idea and attempts to detect network intrusions by monitoring functional relationships between multiple features rather than a single feature. We consider profiles when the principal characteristic is functionally dependent on explanatory variables. Profile monitoring is then used to verify the stability of the functional relationship over time, which is widely applied in calibration applications. In particular, there has been much work on linear profiles. In this case, the stability of the profile is determined by monitoring statistics on the slope and intercept. We thus can consider Shewhart control charts or multivariate control charts. The previous studies assume that the explanatory variable has the same fixed value for each profile. Therefore, to consider the network intrusion problem, the explanatory variable should be expanded to the case observed differently for each profile. In this regard, we will evaluate the robustness of the existing control charts and determine whether the extended control charts effectively detect network intrusion. We perform real analysis using the NSL-KDD data, which is popular in evaluating the performance of network detection algorithms.

**Keywords:** Network intrusion detection, Profile monitoring, Control chart, NSL-KDD data
Benefits of international agreements as switching diffusions
Sheikh Shahnawaz*
Economics, California State University, USA

Abstract: We formulate a model to consider the dynamic stability of international agreements (IAs) such as those on disarmament, nuclear non-proliferation, the environment, or sovereign debt. An agreement is reached because all participants initially receive some benefit that is above a minimum threshold level, but the distribution of total benefits X (modeled as a stochastic process that solves a switching stochastic differential equation) varies beyond ratification. The agreement is sustained as long as participants receiving higher benefits transfer their surplus to those with slack but this comes at a cost alpha, which is a homogeneous continuous Markov chain. Under certain assumptions on uniqueness of the solution of our SDE and on alpha, we derive an optimal strategy that prolongs the life of the IA.

Keywords: Stochastic differential equations, Switching, Continuous-time Markov chains

Estimation of Hilbertian varying coefficient models
Young Kyung Lee¹, Byeong U. Park², Hyerim Hong*² and Dongwoo Kim²
¹Department of Information Statistics, Kangwon National University, Korea
²Department of Statistics, Seoul National University, Korea

Abstract: In this paper we discuss the estimation of a fairly general type of varying coefficient model. The model is for a response variable that takes values in a general Hilbert space and allows for various types of additive interaction terms in representing the effects of predictors. It also accommodates both continuous and discrete predictors. We develop a powerful technique of estimating the very general model. Our approach may be used in a variety of situations where one needs to analyze the relation between a set of predictors and a Hilbertian response. We prove the existence of the estimators of the model itself and of its components, and also the convergence of a backfitting algorithm that realizes the estimators. We derive the rates of convergence of the estimators and their asymptotic distributions. We also demonstrate via simulation study that our approach works efficiently, and illustrate its usefulness through a real data application.

Keywords: Hilbertian response, Varying coefficient model, Additive regression, Smooth backfitting

Duality for a class of continuous-time reversible Markov models
Freddy Palma*
Probability and Statistics, Fundación Universidad de las Américas Puebla, Mexico

Abstract: Using a conditional probability structure we build transition probabilities that drive appealing classes of reversible Markov processes. The mechanism used in such a construction allows to find a dual Markov process. This kind of duality is then used to compute the predictor operator of one process via its dual. In particular, we identify the dual of some non-conjugate models, namely the $\text{M}/\text{M}/\infty$ queue model and a simple birth, death and immigration process. Such duals ensures that the computation of the predictor operators can be done via finite sums.

Keywords: Duality for Markov, Reversible Markov process, Theory of queues