Abstracts

Genevera Allen
Rice University

Fast and Powerful Minipatch Ensemble Learning for Discovery and Inference

Enormous quantities of data are collected in many industries and disciplines; this data holds the key to solving critical societal and scientific problems. Yet, fitting models to make discoveries from this huge data often poses both computational and statistical challenges. In this talk, we propose a new ensemble learning strategy primed for fast, distributed, and memory-efficient computation that also has many statistical advantages. Inspired by random forests, stability selection, and stochastic optimization, we propose to build ensembles based on tiny subsamples of both observations and features that we term minipatches. While minipatch learning can easily be applied to prediction tasks similarly to random forests, this talk focuses on using minipatch ensemble approaches in unconventional ways: making data-driven discoveries and for statistical inference. Specifically, we will discuss using this ensemble strategy for structural graph learning on an enormous scale as well as for distribution-free and modelagnostic inference for both predictions and important features. Through real data examples from neuroscience and biomedicine, we illustrate the computational and statistical advantages of our minipatch ensemble learning approaches.

Back to Table of Contents

Andrew Barron
Yale University

Log Concave Coupling for Sampling from Neural Net Posterior Distributions

A framework for sampling from posterior distributions of parameters of artificial neural networks is presented. The idea is to couple the posterior with an auxiliary random variable such that both the forward distribution (of auxiliary variables given the parameters) and the reverse distribution (of parameters given the auxiliary variables) are fast to sample. Particularly useful is the case that the conditional distribution of the auxiliary variables given the parameters is Gaussian with independent coordinates. We show in our construction that the reverse distribution of parameters given the auxiliary variables is log-concave. This permits accurate computation of the score which is the gradient of the log of the density of the auxiliary random variables. Using this score as the drift function one may run a stochastic (Langevin) diffusion to sample from the auxiliary distribution. Then a draw from the log-concave conditional for the parameters permits sampling from their posterior distribution. Along with these algorithmic developments we present corresponding bounds for statistical risk and for online learning regret for predictions based on these neural network fits. This is based on joint work with Curtis McDonald of Yale University.

Back to Table of Contents

Thomas B. Berrett
University of Warwick

Nonparametric test of Missing Completely At Random

One of the most commonly-encountered discrepancies between real data sets and models hypothesised in theoretical work is that of missing data. When faced with incomplete data, the primary concern is to understand the relationship between the data-generating and missingness mechanisms. In the ideal situation, these two sources of randomness are independent, a setting known as Missing Completely At Random (MCAR), but this is often too restrictive in practice. In this talk I will discuss hypothesis tests of the MCAR assumption with material based on joint work with Richard Samworth (https://arxiv.org/abs/2205.08627) and Alberto Bordino (https://arxiv.org/abs/2401.05256).

It turns out that there are deep connections between this problem and ideas from copula theory and convex optimisation. Our methods in the first work are based on using linear programming to test the compatibility of distributions. In the second we draw connections with the matrix completion literature and thus develop tests based on semidefinite programming. In both cases our methods are more widely applicable than existing methods and, in cases that existing methods are applicable, we see strong empirical performance with comparable power.

Back to Table of Contents

Tony Cai
University of Pennsylvania

Federated Learning for Nonparametric Function Estimation: Framework and Optimality

Federated learning is a machine learning paradigm designed to tackle the challenges of data governance and privacy. It enables organizations (e.g., hospitals) to collaboratively train and enhance a shared global statistical model without sharing raw data externally. Instead, the learning process occurs locally at each participating entity, and only model characteristics, such as parameters and gradients, are exchanged, while preserving privacy.

In this talk, we consider statistical optimality for federated learning in the context of nonparametric regression. The setting we study is heterogeneous, encompassing varying sample sizes and differential privacy constraints across different servers. Within this framework, both global and pointwise estimation are considered, and optimal rates of convergence over the Besov spaces are established.

We propose distributed privacy-preserving estimation procedures and analyze their theoretical properties. The findings shed light on the delicate balance between accuracy and privacy preservation. In particular, we characterize the compromise not only in terms of the privacy budget but also concerning the loss incurred by distributing data within the privacy framework as a whole. This insight captures the folklore wisdom that it is easier to retain privacy in larger samples, and explores the differences between pointwise and global estimation under distributed privacy constraints.

Back to Table of Contents

Yuejie Chi
Carnegie Mellon University

Generative Modeling and Generative Prior

Diffusion models, which convert noise into new data instances by learning to reverse a Markov diffusion process, have become a cornerstone in contemporary generative modelling. While their practical power has now been widely recognized, the theoretical underpinnings remain far from mature. We first develop a suite of non-asymptotic theory towards understanding the data generation process of diffusion models in discrete time for both deterministic and stochastic samplers, assuming access to L2-accurate estimates of the (Stein) score functions. We further discuss how to provably accelerate the data generation without additional training, leveraging higher-order approximation. Last but not least, we advocate diffusion models as an expressive data prior in inverse problems, and introduces a plug-and-play method (Diffusion PnP) that alternatively calls two samplers, a proximal consistency sampler solely based on the forward model, and a denoising diffusion sampler solely based on the score functions of data prior. Performance guarantees and numerical examples will be demonstrated to illustrate the promise of our method.

Back to Table of Contents

Jianqing Fan
Princeton University

Distinguished Lecture Series in Statistic

Inferences on Mixing Probabilities and Ranking in Mixed-Membership Models

Network data is prevalent in numerous big data applications, including economics and health networks, where it is of prime importance to understand the latent structure of the network. In this paper, we model the network using the Degree-Corrected Mixed Membership (DCMM) model. In the DCMM model, for each node i, there exists a membership vector consisting of the weight that node i puts in community k. We derive novel finite-sample expansion for the weights, which allows us to obtain asymptotic distributions and confidence intervals of the membership mixing probabilities and other related population quantities. This fills an important gap on uncertainty quantification on the membership profile. We further develop a ranking scheme of the vertices based on the membership mixing probabilities on certain communities and perform relevant statistical inferences. A multiplier bootstrap method is proposed for ranking inference of individual member's profile with respect to a given community. The validity of our theoretical results is further demonstrated via numerical experiments in both real and synthetic data examples.

(Joint work with Sohom Bhattacharya and Jikai Hou)

Back to Table of Contents

Adel Javanmard
University of Southern California

Learning from Aggregate Responses

In many practical applications the training data is aggregated before being shared with the learner, in order to protect privacy of users' sensitive responses. In an aggregate learning framework, the dataset is grouped into bags of samples, where each bag is available only with an aggregate response, providing a summary of individuals' responses in that bag. In this talk, I will discuss some of the recent developments in this space, namely on loss construction and bagging schemes which improve the accuracy of the model, while providing privacy. In particular, I will show how priors can be used to inform bag construction and also present an iterative boosting algorithm which refines the prior via sample splitting.

Back to Table of Contents

Olga Klopp
ESSEC Business School

Denoising over network with application to partially observed epidemics

We introduce a novel approach to predict epidemic spread over networks using total variation (TV) denoising, a signal processing technique. The study proves the consistency of TV denoising with Bernoulli noise, extending existing bounds from Gaussian noise literature. The methodology is further extended to handle incomplete observations, showcasing its effectiveness. We show that application of 1-bit total variation denoiser improves the prediction accuracy of virus spread dynamics on networks.

Back to Table of Contents

Samory Kpotufe
Columbia University

Understanding the Benefits of Related Data

Modern machine applications require large amounts of high-dimensional data, often forcing practitioners to rely on imperfect but related data. However, the statistical benefits of such related data distributions are not yet well understood. Theoretical works have so far considered a number of metric and divergences on distributions, along with structural assumptions that might help explain beneficial aspects of imperfect but related data. Yet, a more unified theory remains elusive.

I will first present some recent results attempting to unify our understanding of divergences between distributions via certain moduli of continuity between excess risks; these are shown to recover many past proposals from the literature on distribution shifts. I will then briefly discuss some negative results (in so-called multi-task learning, and in model selection) which suggest that distribution-shifts are more structured in practice than captured by usual theoretical formalisms. An alternative formalism, suggested by common uses of neural networks, is that related training data may share nonlinear dimension-reduction spaces, often referred to as 'shared data representations'. I hope to discuss some new works in this direction, establishing the benefits of such shared representations in the context of kernel methods.

The talk is based on joint work with collaborators over the last few years, namely, G. Martinet, S. Hanneke, J. Suk, Y. Mahdaviyeh, N. Galbraith, A. Gretton, M. Zhu, D. Meunier.

Back to Table of Contents

Zehua Lai
The University of Texas at Austin

Central limit theorems for stochastic optimization

The classical Polyak-Juditsky stochastic gradient descent (SGD) estimator is widely used in modern optimization. Its limiting distribution and statistical efficiency are well-understood and can be used to perform statistical inference for model parameters. We consider two variant problems: statistical inference with Kiefer-Wolfowitz methods and statistical inference for contextual bandits, all under SGD setting. In those settings, we find new variance structures sensitive to learning parameters and discuss how to learn efficiently with those structures.

Back to Table of Contents

Hongzhe Li
University of Pennsylvania

Transfer learning and applications in genomics

This talk considers estimation and prediction of high-dimensional linear regression model for transfer learning, using samples from the target model as well as auxiliary samples from different but possibly related models. When the set of ``informative" auxiliary samples is known, an estimator and a predictor are proposed and their optimality is established. The optimal rates of convergence for prediction and estimation are faster than the corresponding rates without using the auxiliary samples. This implies that knowledge from the informative auxiliary samples can be transferred to improve the learning performance of the target problem. When sample informativeness is unknown, a data-driven procedure for transfer learning, called Trans-Lasso is proposed, and its robustness to non-informative auxiliary samples and its efficiency in knowledge transfer is established. A related method, Trans-CLIME is developed for estimation and inference of high-dimensional Gaussian graphical models with transfer learning. Several applications in genomics will be presented, including prediction of gene expressions using the GTEx data, estimation of tissue-specific gene regulatory networks, polygenetic risk score prediction using GWAS data, and proteomics-based cardiovascular disease risk prediction in patients with chronic kidney disease.

Back to Table of Contents

Xiang Li
University of Pennsylvania

A Statistical Framework of Watermarks for Large Language Models: Pivot, Detection Efficiency and Optimal Rules

Since ChatGPT was introduced in November 2022, embedding (nearly) unnoticeable statistical signals into text generated by large language models (LLMs), also known as watermarking, has been used as a principled approach to provable detection of AI-generated text from its human-written counterpart. In this talk, we introduce a general and flexible framework for reasoning about the statistical efficiency of watermarks and designing powerful detection rules. Inspired by the hypothesis testing formulation of watermark detection, our framework starts by picking a pivotal statistic of the text to test and a secret key---provided by the LLM only to the detector---to enable controlling the false positive rate (error that human written text is detected as AI-generated text). Next, this framework allows one to evaluate the statistical efficiency of watermark detection rules by obtaining a closed-form expression of the false negative rate (an error that AI-generated text is incorrectly classified as human-written text) in asymptotic. Our framework further reduces the problem of determining the optimal detection rule by solving a minimax optimization program. We apply this framework to two representative watermarks---one has been internally implemented at OpenAI---and obtained several findings that can be instrumental in guiding the practice of implementing the watermarks. In particular, we obtain optimal detection rules for these watermarks using the framework. These theoretically derived detection rules are demonstrated to be competitive and sometimes superior to existing detection approaches through numerical experiments.

Back to Table of Contents

Jonathan Scarlett
National University of Singapore

Recent Developments in High-Dimensional Estimation with Generative Priors

The problem of estimating an unknown vector (or image) from linear or non-linear measurements has a long history in statistics, machine learning, and signal processing. Classical studies focus on the "n >> p" regime (#measurements >> #parameters), and more recent studies handle the "n << p" regime by exploiting low-dimensional structure such as sparsity or low-rankness. Such variants are commonly known as compressive sensing. In this talk, I will review recent methods that move beyond these explicit notions of structure, and instead assume that the underlying vector is well-modelled by a data-driven generative model (e.g., produced by deep learning methods). I will focus primarily on theoretical developments, including upper and lower bounds on the sample complexity in terms of various properties of the generative model, such as its number of latent (input) parameters, its Lipschitz constant, and its width and depth in the special case of neural network models. If time permits, I will also discuss some developments regarding non-linear models, geometric properties of the relevant optimization landscapes, and methods for fully general probabilistic priors.

Back to Table of Contents

Ali Shojaie
University of Washington

Estimation and Inference for Networks of Multi-Experiment Point Processes

Modern high-dimensional point process data, especially those from neuroscience experiments, often involve observations from multiple conditions and/or experiments. Networks of interactions corresponding to these conditions are expected to share many edges, but also exhibit unique, condition-specific ones. However, the degree of similarity among the networks from different conditions is generally unknown. To address these needs, we propose a joint estimation procedure for networks of high-dimensional point processes that incorporates easy-to-compute weights in order to data-adaptively encourage similarity between the estimated networks. We also propose a powerful hierarchical multiple testing procedure for edges of all estimated networks, which accounts for the data-driven similarity structure of the multi-experiment networks. Compared to conventional multiple testing procedures, our proposed procedure greatly reduces the number of tests and results in improved power, while tightly controlling the family-wise error rate. Unlike existing procedures, our method is also free of assumptions on dependency between tests, offers flexibility on p-values calculated along the hierarchy, and is robust to misspecification of the hierarchical structure.

Back to Table of Contents

Pragya Sur
Harvard University

Spectrum-Aware Debiasing: High-Dimensional Inference beyond sub-Gaussian Covariates with Applications to Principal Components Regression

Debiasing methodologies have emerged as a powerful tool for statistical inference in high dimensions. Since its original introduction, the methodology witnessed a major advancement with the introduction of degrees-of-freedom debiasing in Bellec and Zhang (2019). While overcoming limitations of initial debiasing approaches, this updated method suffered a limitation—it relied on sub-Gaussian tails and independent, identically distributed samples. In this talk, we propose a novel debiasing formula that breaks this barrier by exploiting the spectrum of the sample covariance matrix. Our formula applies to a broader class of designs known as right rotationally invariant designs, which include some heavy-tailed distributions, as well as certain dependent data settings. Our correction term differs significantly from prior work but recovers the Gaussian-based formula as a special case. Notably, our approach does not require estimating the high-dimensional population covariance matrix yet can account for dependence among features and samples. We demonstrate the utility of our method for several statistical inference problems. As a by-product, our work also introduces the first debiased principal component regression estimator with formal guarantees in high dimensions. This is based on joint work with Yufan Li.

Back to Table of Contents

Yanshuo Tan
National University of Singapore

Generalization performance gaps between greedy and optimal regression trees in high dimensions

Regression trees and their ensembles are among the most popular and important machine learning models and are especially useful for high-dimensional data. Because empirical risk minimization (ERM) is computationally infeasible, these models are typically fitted using greedy algorithms. While these algorithms can lead to good models, they have been empirically observed to get stuck at local optima. In this talk, we provide the first theoretical understanding of this phenomenon. In particular, we formulate the Complete Signed Staircase Property (CSSP) for sparse regression functions, which characterizes the generalization of Breiman's CART, given binary features with the uniform measure: When the CSSP holds, generalization error obeys O(1/n) convergence once the sample size n is logarithmic in the ambient dimension p. When the CSSP does not hold, the generalization error is bounded below by a constant even when the sample size is exponential in p. Models fitted via ERM require only O(log(p)) samples under both cases, thereby establishing the first theoretical performance gap between greedy and optimal trees.

Back to Table of Contents

Miaoyan Wang
University of Wisconsin-Madison

Beyond Matrices: Nonparametric Tensor Estimation and Application

High-order tensor datasets present a ubiquitous challenge in applications of recommendation systems, neuroimaging, and social networks. Here we introduce provably guaranteed methods for estimating a likely high-rank signal tensor from such noisy observations. We propose two nonparametric models — sign-series model and latent variable model — that incorporates both high rank and low rank tensors, including simple hypergraphon models, and single index models. Our analysis establishes both statistical and computational limitations for signal tensor estimation. Excess risk bounds, estimation error rates, and sample complexities are established. Furthermore, we propose a polynomial-time spectral algorithm that provably achieves the optimal estimation rate. Notably, we show that a statistical computational gap only emerges for latent variable tensors of order 3 or higher. The efficacy of our approach is further showcased through numerical experiments and real world applications, demonstrating both theoretical soundness and practical merit.

Back to Table of Contents

Dong Xia
Hong Kong University of Science and Technology

Online Policy Learning and Inference by Matrix Completion

Making online decisions can be challenging when features are sparse and orthogonal to historical ones, especially when the optimal policy is learned through collaborative filtering. We formulate the problem as a matrix completion bandit (MCB), where the expected reward under each arm is characterized by an unknown low-rank matrix. The ε-greedy bandit and the online gradient descent algorithm are explored. Policy learning and regret performance are studied under a specific schedule for exploration probabilities and step sizes. A faster decaying exploration probability yields smaller regret but learns the optimal policy less accurately. We investigate an online debiasing method based on inverse propensity weighting (IPW) and a general framework for online policy inference. The IPW-based estimators are asymptotically normal under mild arm-optimality conditions. Numerical simulations corroborate our theoretical findings. Our methods are applied to the San Francisco parking pricing project data, revealing intriguing discoveries and outperforming the benchmark policy.

Back to Table of Contents

Yao Xie
Georgia Institute of Technology

Generative models for high-dimensional statistical inference

We consider the problem of learning a continuous probability density function from data--a fundamental problem in statistics known as density estimation. In this talk, I will present a new perspective of high-dimensional density estimation leveraging the recent advances of neural network-based generative models by formulating the problem of finding an (invertible) dynamic transport map from data distribution to a target distribution. Such a general formulation covers several cases, including (1) fixed parametric target distribution such as Gaussian, which is easy to sample from, such as in generative models; (2) fixed general target distribution represented by samples, a problem related to optimal transport; and (3) not fixed target distribution, which can be induced by a loss function, a problem arises from Wasserstein distributionally robust optimization (DRO). I will demonstrate the application of the framework for density ratio estimation and robust hypothesis testing.

Back to Table of Contents

Fanny Yang
ETH Zurich

Surprising phenomena of interpolating solutions in high dimensions

Interpolating models have recently gained popularity in the statistical learning community due to common practices in modern machine learning: complex models achieve good generalization performance despite interpolating high-dimensional training data. In this talk, I will present generalization bounds for high-dimensional linear models that interpolate data generated by a sparse ground truth for both regression and classification. First, we show that surprisingly, in the noiseless case, while minimizing the l1-norm achieves optimal rates for regression for hard-sparse ground truths, this adaptivity does not directly apply to the equivalent of max l1-margin classification. Further, for noisy observations, we prove how min-l1-norm interpolators and max-lp-margin classifiers can achieve minimax-optimal rates of 1/\sqrt{n} for p slightly larger than one, while for p=1 the rates are proportional to 1/sqrt(log(d/n)). I will explain how this is intuitively due to a "reverse" bias-variance trade-off arising only for interpolating solutions, and show how similar results can also be observed for nonlinear models.

Back to Table of Contents

Yannis Yatracos
Cyprus University of Technology

Statistical inference for Black-Box parameters generating data, and the Laplacian Fiducial distribution

Breiman (2001) urged statisticians to provide tools when data, **X**=s(**Y**) or **X**=s(θ,**Y**), but his suggestion was ignored; s is unknown Black-Box, parameter θεΘ, **Y** is random. However, computer scientists work with **X**=s(θ,**Y**), calling s learning machine. In this talk, statistical inference tools are presented for θ when **X**=s(θ), **Y** latent: a) The Empirical Discrimination Index (EDI), to detect a.s. θ-discrimination and identifiability. b) Matching estimates of θ with upper bounds on the errors in prob. that depend on the "massiveness" of Θ. c) For known stochastic model of **X**, Laplace's 1774 Principle is proved without Bayes rule, thus obtaining a unique Fiducial distribution and showing finally Laplace's and Fisher's intuitions were correct! For unknown **X**-model, an Approximate Fiducial distribution for θ is obtained. The tools are used in ABC, providing F-ABC, that includes all θ* drawn from a Θ-sampler, unlike the Rubin (1984) ABC-rejection method followed until now. Thus, when **X**=s(θ) and a cdf, F_θ, is assumed for **X**, a risk averse researcher can use instead the sampler, s, and a)-c), since F_θ and an assumed θ-prior may be wrong. Le Cam's Statistical Experiments that use $\{F_{\theta^*}, \theta^* \varepsilon \Theta\}$ are now extended to Data Generating Experiments using instead {s(θ*), θεΘ}, which allow "learning" cdfs F_{θ^*} with repeated "training" samples.

Back to Table of Contents

Yi Yu
University of Warwick

Rate Optimality and Phase Transition for User-Level Local Differential Privacy

Most of the literature on differential privacy considers the item-level case where each user has a single observation, but a growing field of interest is that of user-level privacy where each user holds multiple observations and wishes to maintain the privacy of their entire collection.

In this paper, we prove a general minimax lower bound, which shows that, for any locally private user-level estimation problem, the risk cannot be made to vanish for a fixed number of users even when each user holds an arbitrarily large number of observations. We then prove tight minimax lower and upper bounds for univariate and multidimensional mean estimation, sparse mean estimation, and non-parametric density estimation. In particular, we observe a phase-transition in the rate when the number of samples each user holds is sufficiently large relative to the number of users.

Further, in the case of (non-sparse) mean estimation and density estimation, we see that, up until the phase transition, the rate is the same as having an equivalent number of users in the item-level setting, matching similar behaviour seen in other user-level results from previous works. However different behaviour is observed in the case of sparse mean estimation wherein this problem is infeasible when the dimension exceeds the number of observations in the item-level setting, but is tractable in the user-level setting. The estimator recovers elements of the performance of the non-private estimator, which may be of independent interest for applications as an example of a high-dimensional problem that is feasible under local privacy constraints.

Back to Table of Contents

Ming Yuan
Columbia University

Tensor Methods in High Dimensional Data Analysis: Opportunities and Challenges

Large amount of multidimensional data represented by multiway arrays or tensors are prevalent in modern applications across various fields such as chemometrics, genomics, physics, psychology, and signal processing. The structural complexity of such data provides vast new opportunities for modelling and analysis, but efficiently extracting information content from them, both statistically and computationally, presents unique and fundamental challenges. Addressing these challenges requires an interdisciplinary approach that brings together tools and insights from statistics, optimization and numerical linear algebra among other fields. Despite these hurdles, significant progress has been made in the last decade. In this talk, I will review some of the key advancements and identify common threads among them, under several common statistical settings.

Back to Table of Contents

Renbo Zhao
University of Iowa, USA

Frank-Wolfe-Type Methods for Minimizing Log-Homogenous Self-Concordant Barriers

We present and analyze a new Frank–Wolfe method for minimizing Θ-log-homogenous self-concordant barriers, with applications including positron emission tomography, D-optimal design, TV-regularized Poisson image de-blurring, quantum state tomography. The iteration complexity of our method is essentially $O(\Theta^2/\in)$, which recovers that obtained by Khachiyan (1996) on the D-optimal design problem. In addition, we also present and analyse an away-step variant of our proposed Frank–Wolfe method, and we show the global linear convergence of this method. When specialized to the D-optimal design problem, this settles an open problem in Ahipasaoglu, Sun and Todd (2008).

Back to Table of Contents

Anru Zhang
Duke University

High-order Singular Value Decomposition in Tensor Analysis

The analysis of tensor data, i.e., arrays with multiple directions, is motivated by a wide range of scientific applications and has become an important interdisciplinary topic in data science. In this talk, we discuss the fundamental task of performing Singular Value Decomposition (SVD) on tensors, exploring both general cases and scenarios with specific structures like smoothness and longitudinality. Through the developed frameworks, we can achieve accurate denoising for 4D scanning transmission electron microscopy images; in longitudinal microbiome studies, we can extract key components in the trajectories of bacterial abundance, identify representative bacterial taxa for these key trajectories, and group subjects based on the change of bacteria abundance over time. We also showcase the development of statistically optimal methods and computationally efficient algorithms that harness valuable insights from high-dimensional tensor data, grounded in theories of computation and non convex optimization.

Back to Table of Contents

Emma Zhang
Emory University

Statistical Inference on Network Data: New Models and Algorithms

Network data play a fundamental role in characterizing complex systems, such as social interactions, local and global economies, neural connectivity in the brain, and gene-gene interactions. In this talk, I will discuss some of our recent work on algorithmic solutions for fast network community detection and new models for networks with textual edges. First, we describe a new approach for estimating blockmodels based on decoupling in the likelihood function. We apply this approach to the stochastic blockmodel and some of its variants and demonstrate the effectiveness of the proposed method for community detection in large-scale networks. Second, we propose a new model for analyzing networks with textual edges and investigate theoretical properties of the proposed estimator. The efficacy of our methods is demonstrated through simulations and analyses of several real-world data sets.

Back to Table of Contents