Abstracts

Table of Contents

Anish Agarwal
Columbia University, USA

Causal Matrix Completion

Matrix completion is the study of recovering an underlying matrix from a sparse subset of noisy observations. Traditionally, it is assumed entries of the matrix are "missing completely at random" (MCAR), i.e., each entry is revealed at random, independent of everything else, with uniform probability. This is likely unrealistic due to the presence of "latent confounders", i.e., unobserved factors that determine both the entries of the underlying matrix and the missingness pattern in the observed matrix. For example, in the context of movie recommender systems—a canonical application for matrix completion—a user who vehemently dislikes horror films is unlikely to ever watch horror films. In general, these confounders yield "missing not at random" (MNAR) data, which can severely impact any inference procedure.

We develop a formal causal model for MC through the language of potential outcomes, and provide novel identification arguments for a variety of causal estimands of interest. We combine a nearest neighbours approach for matrix completion —popularly known as collaborative filtering—with the synthetic controls approach for panel data, to design a simple two-step algorithm, which we call "synthetic nearest neighbours" (SNN) to estimate these causal estimands. We prove entry-wise finite-sample consistency and asymptotic normality of the SNN estimator for matrix completion with MNAR data. Importantly, we allow the revelation of an entry to be arbitrarily dependent across each other, and the minimum probability of observing an entry to be 0. As a special case, our results also provide entry-wise bounds for matrix completion with MCAR data. Across simulated and real data, we demonstrate the efficacy of our proposed estimator.

Back to Table of Contents

Xuan Bi
University of Minnesota, USA

Distribution-invariant differential privacy

Differential privacy is becoming one gold standard for protecting the privacy of publicly shared data. It has been widely used in biomedical sciences, data science, public health, information technology, and the U.S. decennial census. Nevertheless, to guarantee differential privacy, existing methods may unavoidably alter the conclusion of original data analysis, as privatization often changes the sample distribution. This phenomenon is known as the trade-off between privacy protection and statistical accuracy. In this work, we mitigate this trade-off by developing a distribution-invariant privatization (DIP) method to reconcile both high statistical accuracy and strict differential privacy. As a result, any downstream statistical or machine learning task yields essentially the same conclusion as if one used the original data. Numerically, under the same strictness of privacy protection, DIP achieves superior statistical accuracy across a wide range of simulation studies and real-world benchmarks.

Back to Table of Contents

Michael Celentano
University of California, Berkeley, USA

Challenges of the inconsistency regime: Novel debiasing methods for missing data models

In this talk, I will discuss semi-parametric estimation of the population mean when data is observed missing at random (MAR) in the "inconsistency regime," in which neither the outcome model nor the propensity/missingness model can be estimated consistently. I focus on a high-dimensional linear-GLM specification in which the number of confounders is proportional to the sample size. In the case $n > p$, past work has developed theory for the classical AIPW estimator in this model and established its variance inflation and asymptotic normality when the outcome model is fit by ordinary least squares. Ordinary least squares is no longer feasible in the case studied here, and I will demonstrate that a number of classical debiasing procedures become inconsistent. This challenge motivates the development and analysis of a novel procedure: we establish that it is consistent for the population mean under proportional asymptotics allowing for $n < p$. Providing such guarantees in the inconsistency regime requires a new debiasing approach that combines penalized M-estimates of both the outcome and propensity/missingness models in a non-standard way.

Back to Table of Contents

Michael Choi
National University of Singapore, Singapore

Markov chains + X: Markov chain entropy games and the geometry of their Nash equilibria

Consider the following two-person mixed strategy game of a probabilist against Nature with respect to the parameters (f,B,π), where f is a convex function satisfying certain regularity conditions, B is either the set {L_i}_{i=1}^{n} or its convex hull with each L_i being a Markov infinitesimal generator on a finite state space X and π is a given positive discrete distribution on X. The probabilist chooses a prior measure μ within the set of probability measures on B denoted by P(B) and picks a L∈B at random according to μ, whereas Nature follows a pure strategy to select M∈L(π), the set of π-reversible Markov generators on X. Nature pays an amount D_f(M||L), the f-divergence from L to M, to the probabilist. We prove that a mixed strategy Nash equilibrium always exists, and establish a minimax result on the expected payoff of the game. This also contrasts with the pure strategy version of the game where we show a Nash equilibrium may not exist. To find approximately a mixed strategy Nash equilibrium, we propose and develop a simple projected subgradient algorithm that provably converges with a rate of O(1/\sqrt{t}), where t is the number of iterations. In addition, we elucidate the relationships of Nash equilibrium with other seemingly disparate notions such as weighted information centroid, Chebyshev center and Bayes risk. This talk highlights the powerful interplay and synergy between modern Markov chains theory and geometry, information theory, game theory, optimization and mathematical statistics.

This is based on a joint work with Geoffrey Wolfer (RIKEN AIP), and the paper can be found in https://arxiv.org/abs/2310.04115.

Back to Table of Contents

Hai Dang Dau
Oxford University, UK

On using diffusion models for sampling

Diffusion models have been tremendously successful in generative modelling, but using them in sampling problems receives less attention. The fundamental reason is that, in this context, we only have access to the unnormalised density and do not have any training data. In this talk, we discuss a reverse KL minimisation approach by Vargas et al (2023). If time allows, we provide a sketch of our ongoing work drawing on the connection between sampling, diffusions, and sequential Monte Carlo methods.

Vargas F., Grathwohl W., Doucet A. (2023) Denoising Diffusion Samplers. ICLR

Back to Table of Contents

Jin-Hong Du
Carnegie Mellon University, USA

Simultaneous inference for generalized linear models with unmeasured confounders

Tens of thousands of simultaneous hypothesis tests are routinely performed in genomic studies to identify differentially expressed genes. However, due to unmeasured confounders, many standard statistical approaches may be substantially biased. This talk investigates the large-scale hypothesis testing problem for multivariate generalized linear models in the presence of confounding effects. Under arbitrary confounding mechanisms, we propose a unified statistical estimation and inference framework that harnesses orthogonal structures and integrates linear projections into three key stages. It begins by disentangling marginal and uncorrelated confounding effects to recover the latent coefficients. Subsequently, latent factors and primary effects are jointly estimated through lasso-type optimization. Finally, we incorporate projected and weighted bias-correction steps for hypothesis testing. Theoretically, we establish the identification conditions of various effects and non-asymptotic error bounds. We show effective Type-I error control of asymptotic z-tests as sample and response sizes approach infinity. Numerical experiments demonstrate that the proposed method controls the false discovery rate by the Benjamini-Hochberg procedure and is more powerful than alternative methods. By comparing single-cell RNA-seq counts from two groups of samples, we demonstrate the suitability of adjusting confounding effects when significant covariates are absent from the model.

Back to Table of Contents

Paromita Dubey
University of Southern California, USA

Two Sample Inference for Random Objects using Metric Profiles

Complex non-Euclidean data, also known as object data, have become standard fare in modern data science. They appear as networks, distributions, trees and so on. In this talk I will introduce a novel geometrical framework to distinguish between populations of random objects. The test statistic is based on the differences in the metric profiles of each observation with respect to their own population versus that obtained with respect to a potentially different population. I will describe the asymptotic behavior of the test statistic under the null hypothesis of no differences across the populations and study its power under contiguous alternatives close to the null. For approximating the critical value, we use a theoretically justified permutation scheme in practice. To make a convincing case, I will illustrate the performance of the test in a range of simulations for a large variety of metric spaces under challenging settings and on a real application with network valued data obtained from fMRI images.

Back to Table of Contents

Jean Feng
University of California, San Francisco, USA

Statistical tools for auditing machine learning algorithms across subgroups and time

Machine learning (ML) algorithms have the potential to derive insights from clinical data and improve patient outcomes. However, the performance of these highly complex systems often differs across patient subgroups and is sensitive to changes in the environment. Auditing these ML algorithms---both in pre-market and post-market settings---ensures their safety and effectiveness across diverse patient populations and over time. In this talk, we will explore how to audit for the fundamental safety requirement of strong calibration: for any subgroup, the average predicted probability should be close to its average event rate. Checking for strong calibration is challenging. Given the sheer number of possible subgroups, procedures are often underpowered after adjustment for multiple testing. Moreover, after a ML algorithm has been integrated into clinical practice, the ML algorithm modifies the medical decision-making process and becomes a major source of bias in the data. In this talk, we illustrate how to address these challenges using tools from changepoint detection and causal inference.

Back to Table of Contents

Swarnadip Ghosh
Radix Trading, LLC, USA

Backfitting for large scale crossed random effects regressions

Large-scale genomic and electronic commerce data sets often have a crossed random effects structure, arising from genotypes x environments or customers x products. Applying naive methods to handle such data often leads to inferences that lack generalizability. Regression models that effectively accommodate crossed random effects can be computationally intensive. Both generalized least squares and Gibbs sampling methods can become prohibitively expensive, with costs easily escalating as N^(3/2) (or more) for N observations. Papaspiliopoulos, Roberts, and Zanella (2020) introduced a collapsed Gibbs sampler that runs at O(N) cost, albeit under a highly strict sampling model. In contrast, we introduce a backfitting algorithm for computing a generalized least squares estimate. We demonstrate that it costs O(N) under considerably relaxed but still stringent sampling assumptions. Empirically, our backfitting algorithm demonstrates O(N) cost under even further relaxed assumptions. To illustrate its efficacy, we apply the new algorithm to a ratings dataset from Stitch Fix.

This is joint work with Trevor Hastie and Art Owen.

Back to Table of Contents

Richard Guo
University of Cambridge, UK

Harnessing Extra Randomness: Replicability, Flexibility and Causality

Many modern statistical procedures are randomized in the sense that the output is a random function of data. For example, many procedures employ data splitting, which randomly divides the dataset into disjoint parts for separate purposes. Despite their flexibility and popularity, data splitting and other constructions of randomized procedures have obvious drawbacks. First, two analyses of the same dataset may lead to different results due to the extra randomness introduced. Second, randomized procedures typically lose statistical power because the entire sample is not fully utilized.

To address these drawbacks, in this talk, I will study how to properly combine the results from multiple realizations (such as through multiple data splits) of a randomized procedure. I will introduce rank-transformed subsampling as a general method for delivering large sample inference of the combined result under minimal assumptions. I will illustrate the method with three applications: (1) a "hunt-and-test" procedure for detecting cancer subtypes using high-dimensional gene expression data, (2) testing the hypothesis of no direct effect in a sequentially randomized trial and (3) calibrating cross-fit "double machine learning" confidence intervals. For these problems, our method is able to derandomize and improve power. Moreover, in contrast to existing approaches for combining p-values, our method enjoys type-I error control that asymptotically approaches the nominal level. This new development opens up the possibility of designing procedures that explicitly randomize and derandomize: extra randomness is introduced to make the problem easier before being marginalized out.

This talk is based on joint work with Prof. Rajen Shah.

Back to Table of Contents

Qiyang Han
Rutgers University, USA

High dimensional asymptotics: some recent progress beyond Gaussian designs

The Convex Gaussian Min-Max Theorem (CGMT) has emerged as a prominent theoretical tool for analysing the precise stochastic behaviour of various statistical estimators in the so-called high dimensional proportional regime, where the sample size and the signal dimension are of the same order. A well recognized limitation of the existing CGMT machinery rests in its stringent requirement on the exact Gaussianity of the design matrix, therefore rendering the obtained precise high dimensional asymptotics largely a specific Gaussian theory in various important statistical models.

This talk provides a structural universality framework for a broad class of regularized regression estimators that is particularly compatible with the CGMT machinery. Here universality means that if a "structure" is satisfied by the regression estimator under a standard Gaussian design, then it will also be satisfied for a general non-Gaussian design with independent entries. In particular, we show that with a good enough delocalization bound for the regression estimator, any "structural property" that can be detected via the CGMT under a Gaussian design also carries over to a general design with independent entries. As a proof of concept, we demonstrate our new universality framework in two key examples of regression estimators, namely, the Lasso estimator and the Ridgeless interpolators. The key technical ingredient of our new framework relies on a set of new comparison inequalities for the optimum of a broad class of cost functions over arbitrary structure sets subject to ℓ_∞ constraints.

Back to Table of Contents

Yuchen Hu
Stanford University, USA

Switchback Experiments under Geometric Mixing

The switchback is an experimental design that measures treatment effects by repeatedly turning an intervention on and off for a whole system. Switchback experiments are a robust way to overcome cross-unit spillover effects; however, they are vulnerable to bias from temporal carryovers. In this paper, we consider properties of switchback experiments in Markovian systems that mix at a geometric rate. We find that, in this setting, standard switchback designs suffer considerably from carryover bias: Their estimation error decays as T^{-1/3} in terms of the experiment horizon T, whereas in the absence of carryovers a faster rate of T^{-1/2} would have been possible. We also show, however, that judicious use of burn-in periods can considerably improve the situation, and enables errors that decay almost as fast as T^{-1/2}. Our formal results are mirrored in an empirical evaluation.

Back to Table of Contents

Dongming Huang
National University of Singapore, Singapore

Sliced Inverse Regression with Large Structural Dimensions

The central space of a joint distribution ($\boldsymbol{X}$, Y) is the minimal subspace $\boldsymbol{S}$ such that $Y \perp\!\!\!\perp \boldsymbol{X} | P_S \boldsymbol{X}$ where P_S is the projection onto S. Sliced inverse regression (SIR) is one of the most popular methods for estimating the central space, but knowledge about its optimality is limited. We study properties of the SIR under a multiple index model $Y = f(\beta_1^T X, \beta_2^T X, \ldots, \beta_d^T X, \varepsilon)$, with a focus on the difficulty in estimating the central space when the structure dimension d is large (i.e., there is no constant upper bound on d). We establish a minimax lower bound for estimating the central space over a large class of high dimensional distributions, which characterizes the impact of the sample size n, the dimension of covariates p, and more importantly, the structure dimension d and the generalized signal-noise-ratio (gSNR). By matching the risk of the SIR with the lower bound, we conclude that SIR is minimax rate-optimal. Moreover, we demonstrate that as d increases, the gSNR tends to be extremely small, and our theory clarifies that this decay in signal strength is responsible for the empirically observed degradation in SIR performance. Our developed technical tools may also be of independent interest for analysing other central space estimation methods.

Back to Table of Contents

Cheng Li
National University of Singapore, Singapore

Bayesian fixed-domain asymptotics for covariance parameters in spatial Gaussian process models

Gaussian process models typically contain finite dimensional parameters in the covariance function that need to be estimated from the data. We study the Bayesian fixed-domain asymptotics for the covariance parameters in spatial Gaussian process models with an isotropic Matern covariance function, which has many applications in spatial statistics. For the model without nugget, we show that when the dimension of the domain is less than or equal to three, the microergodic parameter and the range parameter are asymptotically independent in the posterior. While the posterior of the microergodic parameter is asymptotically close in total variation distance to a normal distribution with shrinking variance, the posterior distribution of the range parameter does not converge to any point mass distribution in general. For the model with nugget, we derive new evidence lower bound and consistent higher-order quadratic variation estimators, which lead to explicit posterior contraction rates for both the microergodic parameter and the nugget parameter. We further study the asymptotic efficiency and convergence rates of Bayesian kriging prediction. All the new theoretical results are verified in numerical experiments and real data analysis.

Back to Table of Contents

Gen Li
Chinese University of Hong Kong, Hong Kong, China

Approximate message passing from random initialization with applications to Z2 synchronization

This talk is concerned with the problem of reconstructing an unknown rank-one matrix with prior structural information from noisy observations. While computing the Bayes-optimal estimator seems intractable in general due to its nonconvex nature, Approximate Message Passing (AMP) emerges as an efficient first-order method to approximate the Bayes-optimal estimator. However, the theoretical underpinnings of AMP remain largely unavailable when it starts from random initialization, a scheme of critical practical utility. Focusing on a prototypical model called Z2 synchronization, we characterize the finite-sample dynamics of AMP from random initialization, uncovering its rapid global convergence. Our theory --- which is non-asymptotic in nature --- is the first in this model to unveil the non-necessity of a careful initialization for the success of AMP.

Back to Table of Contents

Jinzhou Li
Stanford University, USA

Simultaneous false discovery proportion bounds via knockoffs and closed testing

We propose new methods to obtain simultaneous false discovery proportion bounds for knockoff-based approaches. We first investigate an approach based on Janson and Su's k-familywise error rate control method and interpolation. We then generalize it by considering a collection of k values, and show that the bound of Katsevich and Ramdas is a special case of this method and can be uniformly improved. Next, we further generalize the method by using closed testing with a multi-weighted-sum local test statistic. This allows us to obtain a further uniform improvement and other generalizations over previous methods. We also develop an efficient shortcut for its implementation. We compare the performance of our proposed methods in simulations and apply them to a data set from the UK Biobank.

Back to Table of Contents

Shuangning Li
Harvard University, USA

Random Graph Asymptotics for Treatment Effect Estimation under Network Interference

The network interference model for causal inference places all experimental units at the vertices of an undirected exposure graph, such that treatment assigned to one unit may affect the outcome of another unit if and only if these two units are connected by an edge. This model has recently gained popularity as means of incorporating interference effects into the Neyman--Rubin potential outcomes framework; and several authors have considered estimation of various causal targets, including the direct and indirect effects of treatment. In this paper, we consider large-sample asymptotics for treatment effect estimation under network interference in a setting where the exposure graph is a random draw from a graphon. When targeting the direct effect, we show that---in our setting---popular estimators are considerably more accurate than existing results suggest, and provide a central limit theorem in terms of moments of the graphon. Meanwhile, when targeting the indirect effect, we leverage our generative assumptions to propose a consistent estimator in a setting where no other consistent estimators are currently available. We also show how our results can be used to conduct a practical assessment of the sensitivity of randomized study inference to potential interference effects. Overall, our results highlight the promise of random graph asymptotics in understanding the practicality and limits of causal inference under network interference.

This is joint work with Stefan Wager.

Back to Table of Contents

Sifan Liu
Stanford University, USA

An Exact Sampler for Inference after Polyhedral Selection

The exploratory and interactive nature of modern data analysis often introduces selection bias and poses challenges to traditional statistical inference methods. To address selection bias, a common approach is to condition on the selection event. However, this often results in a conditional distribution that is intractable and requires Markov chain Monte Carlo (MCMC) sampling for inference. Notably, some of the most widely used selection algorithms yield selection events that can be characterized as polyhedra, such as the lasso for variable selection and the ε-greedy algorithm for multi-armed bandit problems. This work develops a method that is tailored for conducting inference following polyhedral selection. The proposed method transforms the variables constrained within a polyhedron into variables within a unit cube, allowing for exact sampling. Compared to MCMC, this method offers superior speed and accuracy. Furthermore, it facilitates the computation of maximum likelihood estimators based on selection-adjusted likelihoods. Numerical results demonstrate the enhanced performance of the proposed method compared to alternative approaches for selective inference.

Back to Table of Contents

Yuetian Luo
University of Chicago, USA

The Limits of Algorithm Evaluation and Comparison

Algorithm evaluation and comparison are frequently asked questions in machine learning and statistics. It has been suggested in the literature that evaluating or comparing algorithms without holdout data is a hard statistical problem. In this work, we show that these two tasks are indeed difficult. In particular, we prove that for a large class of ``black-box'' tests, any universally valid one is powerless in evaluating the prediction error of any algorithm or distinguishing any two algorithms when the ratio of the number of available samples and the target training size is small. This is in stark contrast with evaluating or comparing fitted models, where high power is achievable as long as we have a large holdout set. Thus, our result quantitatively shows that evaluating or comparing algorithms is significantly a harder problem. At the same time, cross-validation (CV) is a popular alternative to sample splitting in algorithm evaluation or comparison when the sample size is small. Given the nonexistence of a powerful valid test, it suggests that we must explore particular properties of the algorithms of interest. To address this need, we explore how could the stability of the algorithms help. We show that it is possible to powerfully evaluate or distinguish algorithms when they are extremely stable. However, we argue that this regime is not the most interesting one in the distribution-free setting, and complement this result by showing that, unfortunately, the power of any universally valid black-box test is again low for evaluating or comparing normally stable algorithms.

This is a joint work with Rina Foygel Barber.

Back to Table of Contents

Somabha Mukherjee
National University of Singapore, Singapore

High Dimensional Logistic Regression Under Network Dependence

The classical formulation of logistic regression relies on the independent sampling assumption, which is often violated when the outcomes interact through an underlying network structure, such as over a temporal/spatial domain or on a social network. This necessitates the development of models that can simultaneously handle both the network peer-effect (arising from neighbourhood interactions) and the effect of (possibly) high-dimensional covariates. In this talk, I will describe a framework for incorporating such dependencies in a high-dimensional logistic regression model by introducing a quadratic interaction term, as in the Ising model, designed to capture the pairwise interactions from the underlying network. The resulting model can also be viewed as an Ising model, where the node-dependent external fields linearly encode the high-dimensional covariates. We use a penalized maximum pseudo-likelihood method for estimating the network peer-effect and the effect of the covariates (the regression coefficients), which, in addition to handling the high-dimensionality of the parameters, conveniently avoids the computational intractability of the maximum likelihood approach. Our results imply that even under network dependence it is possible to consistently estimate the model parameters at the same rate as in classical (independent) logistic regression, when the true parameter is sparse and the underlying network is not too dense. Towards the end, I will talk about the rates of consistency of our proposed estimator for various natural graph ensembles, such as bounded degree graphs, sparse Erdos-Renyi random graphs, and stochastic block models, which follow as a consequence of our general results.

This is a joint work with Ziang Niu, Sagnik Halder, Bhaswar Bhattacharya and George Michailidis.

Back to Table of Contents

Yang Ni
Texas A&M University, USA

Causal Discovery from Multivariate Functional Data

Discovering causal relationship using multivariate functional data has received a significant amount of attention very recently. We introduce a functional linear structural equation model for causal structure learning. To enhance interpretability, our model involves a low-dimensional causal embedded space such that all the relevant causal information in the multivariate functional data is preserved in this lower-dimensional subspace. We prove that the proposed model is causally identifiable under standard assumptions that are often made in the causal discovery literature. To carry out inference of our model, we develop a fully Bayesian framework with suitable prior specifications and uncertainty quantification through posterior summaries. We illustrate the superior performance of our method over existing methods in terms of causal graph estimation through extensive simulation studies. We also demonstrate the proposed method using a brain EEG dataset.

Back to Table of Contents

Anant Raj
University of Illinois at Urbana-Champaign, USA

An algorithmic stability perspective on heavy-tails SGD

In this talk, we delve into the intricate relationships between heavy-tailed distributions, generalization error, and algorithmic stability in the realm of noisy stochastic gradient descent. Recent research has illustrated the emergence of heavy tails in stochastic optimization and their intriguing links to generalization error. However, these studies often relied on challenging topological and statistical assumptions. Empirical evidence has further challenged existing theory, suggesting that the relationship between heavy tails and generalization is not always monotonic. In response, we introduce novel insights, exploring the relationship between tail behaviour and generalization properties through the lens of algorithmic stability. Our analysis reveals that the stability of stochastic gradient descent (SGD) varies based on how we measure it, leading to interesting conclusions about its behaviour. Expanding upon these findings, we extend the scope to a broader class of objective functions, including non-convex ones. Leveraging Wasserstein stability bounds for heavy-tailed stochastic processes, our research sheds light on the non-monotonic connection between generalization error and heavy tails, offering a more comprehensive perspective

Back to Table of Contents

Yan Shuo Tan
National University of Singapore, Singapore

MDI+: A Flexible Random Forest-Based Feature Importance Framework

Mean decrease in impurity (MDI) is a popular feature importance measure for random forests (RFs). In this talk, we show that the MDI for a feature Xk in each tree in an RF is equivalent to the unnormalized R2 value in a linear regression of the response on the collection of decision stumps that split on Xk. We use this interpretation to propose a flexible feature importance framework called MDI+. Specifically, MDI+ generalizes MDI by allowing the analyst to replace the linear regression model and R2 metric with regularized generalized linear models (GLMs) and metrics better suited for the given data structure. Moreover, MDI+ incorporates additional features to mitigate known biases of decision trees against additive or smooth models. Extensive data-inspired simulations show that MDI+ significantly outperforms popular feature importance measures in identifying signal features. We also apply MDI+ to two real-world case studies on drug response prediction and breast cancer subtype classification and show that MDI+ extracts well-established predictive genes with significantly greater stability compared to existing feature importance measures.

Back to Table of Contents

Ye Tian
Columbia University, USA

Robust Unsupervised Multi-task Learning on Gaussian Mixture Models

Unsupervised learning has been widely used in many real-world applications. One of the simplest and most important unsupervised learning models is the Gaussian mixture model (GMM). In this work, we study the multi-task learning problem on GMMs, which aims to leverage potentially similar GMM parameter structures among tasks to obtain improved learning performance compared to single-task learning. We propose a multi-task GMM learning procedure based on the EM algorithm that not only can effectively utilize unknown similarity between related tasks but is also robust against a fraction of outlier tasks from arbitrary distributions. The proposed procedure is shown to achieve minimax optimal rate of convergence for both parameter estimation error and the excess mis-clustering error, in a wide range of regimes. Finally, we demonstrate the effectiveness of our methods through simulations and real data examples. To the best of our knowledge, this is the first work studying multi-task learning on GMMs with theoretical guarantees.

Back to Table of Contents

Denny Wu
University of Toronto, Canada

Understanding modern machine learning models through the lens of high-dimensional statistics

Modern machine learning tasks are often high-dimensional, due to the large amount of data, features, and trainable parameters. Mathematical tools such as random matrix theory have been developed to precisely study simple learning models in the high-dimensional regime, and such precise analysis can reveal interesting phenomena that are also empirically observed in deep learning. In this talk I will introduce two examples. First we consider the selection of regularization hyperparameters in the overparameterized regime. We establish a set of equations that rigorously describes the asymptotic generalization error of the ridge regression estimator, which leads to surprising findings including: (i) the optimal ridge penalty can be negative, (ii) regularization can suppress "multiple descent" in the risk curve. For the second part, we go beyond linear models and characterize the benefit of gradient-based representation (feature) learning in neural networks. By studying the performance of ridge regression on the trained features in a two-layer neural network, we prove that feature learning results in a considerable advantage over the initial random features model; this analysis also highlights the role of learning rate scaling in the early phase of gradient descent.

Back to Table of Contents

Jeff Wu
Georgia Institute of Technology, USA

Distinguished Lecture Series in Statistics

Statisticians at Work: Inspiration, Aspiration, Ambition

A key measure of the maturity and quality of a scientific community is how it judges and values accomplishments and (or versus) scholarship. To address this question, I will describe the motivation or drive for accomplishments and/or scholarship at three levels: inspiration, aspiration, ambition. They represent different (but not necessarily exclusive) mindsets or modi operandi. I will use several prominent examples in statistics history to explain or illustrate the acts of inspiration, aspiration, and ambition. They include: Pearson's arguments with Fisher and with Yule, some breakthrough work of Fisher, Neyman, Tukey, Box, Efron, etc. Then I will share some thoughts on what are good or bad mathematical statistics work. Throughout this talk, I will use the "lens" of inspiration, aspiration, and ambition in making my examinations, remarks and suggestions.

Back to Table of Contents

Yuchen Wu
Stanford University, USA

Posterior Sampling from the Spiked Models via Diffusion Processes

Sampling from the posterior is a key technical problem in Bayesian statistics. Rigorous guarantees are difficult to obtain for Markov Chain Monte Carlo algorithms of common use. In this paper, we study an alternative class of algorithms based on diffusion processes. The diffusion is constructed in such a way that, at its final time, it approximates the target posterior distribution. The stochastic differential equation that defines this process is discretized (using a Euler scheme) to provide an efficient sampling algorithm. Our construction of the diffusion is based on the notion of observation process and the related idea of stochastic localization. Namely, the diffusion process describes a sample that is conditioned on increasing information. An overlapping family of processes was derived in the machine learning literature via time-reversal.

We apply this method to posterior sampling in the high-dimensional symmetric spiked model. We observe a rank-one matrix $\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}$ corrupted by Gaussian noise, and want to sample $\boldsymbol{\theta}$ from the posterior. Our sampling algorithm makes use of an oracle that computes the posterior expectation of $\boldsymbol{\theta}$ given the data and the additional observation process. We provide an efficient implementation of this oracle using approximate message passing. We thus develop the first sampling algorithm for this problem with approximation guarantees.

Back to Table of Contents

Maoran Xu
Duke University, USA

Uncertainty quantification for varying dimensional parameters

Dimension reduction is now an essential step in high-dimensional inference. In Bayesian dimension reduction, it is difficult to define a low-dimensional measure without knowing the dimensionality, or when the dimensionality is a random number. For example, one may impose sparsity assumptions on vectors or low-rank assumptions on matrices without knowing the exact cardinality or rank, and a continuous prior in the full space cannot yield exact sparse or exact low-rank inference. In this talk, I will introduce the proximal prior, a novel framework for Bayesian modeling of varying dimensional parameters. By transforming a continuous variable in the high-dimensional space to a low-dimensional parameter with proximal mapping, we define an induced measure on varying dimensional space. This leads to a large class of new Bayesian models that can directly exploit the popular frequentist regularizations and their algorithms while providing a principled and probabilistic uncertainty estimation. This framework is well justified in the geometric measure theory and enjoys a convenient posterior computation. I will demonstrate the approach in several data applications, such as image segmentation and traffic network analysis.

Back to Table of Contents

Qi Xu
University of California, Irvine, USA

Estimation of Individualized Combination Treatment Rule

Individualized treatment rules (ITRs) have been widely applied in many fields such as precision medicine and personalized marketing. Beyond the extensive studies on ITRs with binary or multiple treatments, there is considerable interest in applying combination treatments to enhance the outcome. In this talk, I will introduce two estimation method of individualized combination treatment rule under the outcome regression and inverse probability weighting framework. Under the outcome regression framework, we propose a Double Encoder Model (DEM) which represents the treatment effects with two parallel neural network encoders. This model enables flexible choices of function bases of treatment effects, and improve the estimation efficiency via the parameter-sharing feature of the neural network. Under the inverse probability weighting framework, we target the same problem from multi-label classification perspective, and propose a novel non-convex loss function to replace the intractable 0-1 loss. The proposed method is Fisher-consistent regardless of the intensity level of interaction effects among treatments, and computationally tractable with a difference-of-convex algorithm. Our findings are corroborated by extensive simulation studies and real data examples.

Back to Table of Contents

Sheng Xu
Princeton University, USA

Maximum likelihood for high-noise group orbit estimation and cryo-EM

Motivated by applications to single-particle cryo-electron microscopy (cryo-EM), we study a problem of group orbit estimation where samples of an unknown signal are observed under uniform random rotations from a rotational group. In high-noise regime, we describe a stratification of the Fisher information eigenvalues according to transcendence degrees in the algebra of group invariants. We relate the critical points of the log-likelihood optimization landscape to those of a sequence of moment matching problems. Some examples including a simplified model of cryo-EM will be discussed.

Back to Table of Contents

Yaoming Zhen
The Chinese University of Hong Kong, China

Community Detection in General Hypergraph via Graph Embedding

Conventional network data has largely focused on pairwise interactions between two entities, yet multi-way interactions among multiple entities have been frequently observed in real-life hypergraph networks. In this talk, we will discuss a novel method for detecting community structure in general hypergraph networks, uniform or non-uniform. The proposed method introduces a null vertex to augment a non-uniform hypergraph into a uniform multi-hypergraph, and then embeds the multi-hypergraph in a low-dimensional vector space such that vertices within the same community are close to each other. The developed model is supported by a variety of simulation studies, real applications, as well as its asymptotic theories.

Back to Table of Contents

Wenzhuo Zhou
University of California, Irvine, USA

A Model-Agnostic Graph Neural Network for Integrating Local and Global Information

Graph Neural Networks (GNNs) have achieved promising performance in a variety of graph-focused tasks. Despite their success, existing GNNs suffer from two significant limitations: a lack of interpretability in results due to their black-box nature, and an inability to learn representations of varying orders. To tackle these issues, we propose a novel Model-agnostic Graph Neural Network (MaGNet) framework, which is able to sequentially integrate information of various orders, extract knowledge from high-order neighbours, and provide meaningful and interpretable results by identifying influential compact graph structures. In particular, MaGNet consists of two components: an estimation model for the latent representation of complex relationships under graph topology, and an interpretation model that identifies influential nodes, edges, and important node features. Theoretically, we establish the generalization error bound for MaGNet via empirical Rademacher complexity, and showcase its power to represent layer-wise neighbourhood mixing. We conduct comprehensive numerical studies using simulated data to demonstrate the superior performance of MaGNet in comparison to several state-of-the-art alternatives. Furthermore, we apply MaGNet to a real-world case study aimed at extracting task-critical information from brain activity data, thereby highlighting its effectiveness in advancing scientific research.

Back to Table of Contents

Yichen Zhu
Bocconi University, Italy

Fast Approximation and Optimal Contraction for Vecchia Gaussian Processes

Gaussian processes are widely applied in spatial statistics due to its flexibility and automatic uncertainty quantification, but its computation suffers from the $O(n^3)$ computational complexity. Vecchia approximations provide a scalable computational solution to Gaussian processes, but the absence of theoretical foundation leads to unclear guidance in specifying the underlying graphical model. We provide theoretical support for Vecchia approximations of Gaussian processes from two aspects: 1. Considering them as approximations of the original Gaussian processes, we bound rate of approximation in Wasserstein distance; 2. Directly analysing them as Gaussian process methods, we prove the posterior rate of contraction. Under Matern covariance function, such rate matches the minimax rate of regression when the truth belongs to some Sobolev space.

Back to Table of Contents