Statistical Foundations of Data Science and their Applications

A conference in celebration of Jianqing Fan's 60th Birthday

Princeton, May 8-10, 2023

Sponsorship & Support

Department of Operations Research & Financial Engineering (ORFE), Princeton University

Department of Statistics, Pennsylvania State University at University Park

Department of Data Sciences and Operations, University of Southern California (USC)

Department of Economics, Princeton University

Department of Mathematics, Imperial College London

Center for Statistics and Machine Learning (CSML), Princeton University

Program in Applied and Computational Mathematics (PACM), Princeton University

Bendheim Center for Finance (BCF), Princeton University

Center on Contemporary China, Princeton University

National Science Foundation (NSF)

Institute of Mathematical Statistics (IMS)

American Statistical Association (ASA)

International Chinese Statistical Association (ICSA)

Two Sigma

Hudson Bay Capital

Wizard Capital Research Limited, Hong Kong

Local Organizing Committee

- Matias D. Cattaneo (Chair), Princeton University
- Yacine Ait-Sahalia, Princeton University
- Rajita Chandak, Princeton University
- Sohom Bhattacharya, Princeton University
- Debarghya Mukherjee, Princeton University
- William Underwood, Princeton University
- Bingyan Wang, Princeton University
- Yuling Yan, Princeton University
- Mengxin Yu, Princeton University

Program Committee

- Runze Li (Chair), Pennsylvania State University
- Heather Battey, Imperial College London
- Jelena Bradic, University of California at San Diego
- Matias D. Cattaneo, Princeton University
- Ming-Yen Cheng, Hong Kong Baptist University
- Yingying Fan, University of Southern California
- Tracy Ke, Harvard University
- Dacheng Xiu, University of Chicago
- Wenyang Zhang, University of York (UK)

Banquet Activity Organizing Committee

- Tracy Ke (Chair), Harvard University
- Zhao Chen, Fudan University
- Yang Feng, New York University
- Runze Li, Pennsylvania State University
- Yue Niu, University of Arizona
- Xin Tong, University of Southern California
- Weichen Wang, University of Hong Kong
- Bingyan Wang, Princeton University
- Yuyan Wang, Google Inc
- Lucy Xia, Hong Kong University of Science and Technology
- Lingzhou Xue, Pennsylvania State University
- Yuling Yan, Princeton University
- Zhengjun Zhang, University of Chinese Academy of Sciences

Practical Information

Local Lodging & Visitor Links

- Lodging
- Visiting Princeton

Campus Parking & Transportation

All registrants should park in the Stadium Drive Garage

(Google Maps, Apple Maps) for the day with a parking permit, required between 8am and 5pm. **The link for the permit will be sent in an email to all registered participants who indicated they will be driving to campus.**

Guests can board the Tiger Transit Shuttle (Route #1 or #4) at the garage and be dropped off at McCosh walk stop (Robertson side) – below are the scheduled times of boarding and drop off, or see today's real-time bus location and schedule

From the Garage

If additional shuttle times are needed, please download the timetables or see the Transportation & Parking Services site for more information.

Download the Complete Timetable From the Garage

To the Garage

After the conference, guests should walk to McCosh Walk stop and wait on the Southbound side (Wooten side) – below are the times & bus numbers. Multiple buses stop at McCosh, but only #1 and #2 stop at the Stadium Dr garage – guests can confirm with the driver if unsure.

Download the Complete Timetable from McCosh to the Garage

Table of Contents

Schedule

Opening ceremony and all talks for all dates are located in 101 Friend Center

Day 1 (May 8, 2023)

8:00 am – 9:10 am Registration at Friend Center (pick up registration name badges and conference materials upon arrival to Friend Center)

9:10 am – 9:40 am Opening Ceremony

Matias Cattaneo (Chair, Local Organizing Committee),

Runze Li (Chair, Program Committee),

Sanjeev Kulkarni (ORFE/ECE, former Dean of the Faculty)

9:40 am – 10:05 am

Peter Bickel

Independence and functional dependence I

Xuming He (Chair)

10:05 am – 10:30 am

Kosuke Imai

Safe Policy Learning through Extrapolation: Application to Pre-trial Risk Assessment

Xuming He (Chair)

10:30 am – 10:55 am

Philip A. Ernst

New Frontiers in Statistical Inference for Stochastic Processes

Xuming He (Chair)

10:55 am – 11:35 am

Group Photo at Fountain of Freedom & Break

11:35 am – 12:35 pm Panels 1, 2, & 3

Panel 1: Frontier Research on Minimax estimation
Friend 004

Florentina Bunea, Harrison Zhou, Zongming Ma, Weijie Su, Bill Strawderman, Linda Zhao (Chair)

Panel 2: Meet Editors and Department Heads (Junior researcher Career development)
Computer Science 104

Gang Li, Annie Qu, Tian Zheng, Dylan Small, Ji Zhu, Liza Levina, Yazhen Wang, Edward George, Minge Xie (Chair)

Panel 3: Frontier research on optimization in data science
Friend 101

Steve Kou, Ethan Fang, Yuxin Chen, Zhaoran Wang, Yao Xie, Lingzhou Xue, Yang Ning, Yaqi Duan (chair)

12:35 pm – 2:00 pm

Lunch

2:00 pm – 2:25 pm

Yacine Ait-Sahalia

So Many Jumps, So Few News

Zhiliang Ying (Chair)

2:25 pm – 2:50 pm

Qiwei Yao

Autoregressive Networks

Zhiliang Ying (Chair)

2:50 pm – 3:15 pm

Jinchi Lv

SIMPLE-RP: Group Network Inference with Non-Sharp Nulls and Weak Signals

Zhiliang Ying (Chair)

3:15 pm – 3:30 pm

Break

3:30 pm – 3:55 pm

Irène Gijbels

Circular local likelihood regression

Jiayang Sun (Chair)

3:55 pm – 4:20 pm

Wenyang Zhang

A Flexible and Parsimonious Modelling Strategy for Clustered Data Analysis

Jiayang Sun (Chair)

4:20 pm – 4:45 pm

Ming-Yen Cheng

Inference for nonstationary time series with varying periodicity, a smooth trend and covariate effects

Jiayang Sun (Chair)

4:45 pm – 6:15 pm Poster

Zhengjun Zhang (Chair)

Location: Friend Center Upper Atrium

Day 2 (May 9, 2023)

8:30 am – 8:55 am

J. S. Marron

Object oriented data analysis

Raymond Carroll (Chair)

8:55 am – 9:20 am

Wolfgang Härdle

Robustifying Markowitz

Raymond Carroll (Chair)

9:20 am – 9:45 am

Enno Mammen

Random Planted Forest: a directly interpretable tree ensemble

Raymond Carroll (Chair)

9:45 am – 10:00 am

Break

10:00 am – 10:25 am

Richard Samworth

Optimal nonparametric testing of missing completely at random, and its connections to compatibility

Li-Shan Huang (Chair)

10:25 am – 10:50 am

Hui Zou

M-Optimal designs through the lens of modern optimization

Li-Shan Huang (Chair)

10:50 am – 11:15 am

Heather Battey

Inducement of population sparsity

Li-Shan Huang (Chair)

11:15 am – 11:30 am

Break

11:30 am – 12:30 pm Panels 4, 5, & 6

Panel 4: Data Science applications in real world
Computer Science 104

Hongtu Zhu, Rui Song, Emre Barut, Lei Qi, Igor Silin, Runlong Tang, Wei Dai (Chair)

Panel 5: Frontier research on business statistics and Econometrics
Friend 004

Zongwu Cai, Elynn Chen, Rong Chen, Weining Wang, Yan Yu, Emma Zhang, Yingying Li, Yongyi Guo (Chair)

Panel 6: Frontier research on machine learning
Friend 101

Edgar Dobriban, Xi Chen, Yuting Wei, Cong Ma, Krishna Balasubramanian, Marco Avella-Medina, Yuejie Chi, Lan Wang, Jelena Bradic, Haolei Weng (Chair)

12:30 pm – 2:00 pm

Lunch

2:00 pm – 2:25 pm

Hans Mueller

Distributional Regression with Optimal Transports

Jin-Ting Zhang (Chair)

2:25 pm – 2:50 pm

Jane-Ling Wang

Deep Learning for Partial Linear Cox Model

Jin-Ting Zhang (Chair)

2:50 pm – 3:15 pm

Chunming Zhang

New statistical learning method for independent component analysis in the presence of noise

Jin-Ting Zhang (Chair)

3:15 pm – 3:30 pm

Break

3:30 pm – 3:55 pm

Xihong Lin

Fast Distributed Principal Component Analysis for Large-Scale Federated Data

Jiancheng Jiang (Chair)

3:55 pm – 4:20 pm

Jianwen Cai

Feature screening for case-cohort studies with failure time outcome

Jiancheng Jiang (Chair)

4:20 pm – 4:45 pm

Kinh Truong

On estimating power spectral density for independent component analysis

Jiancheng Jiang (Chair)

4:45 pm – 5:00 pm

Jennifer Rexford (Provost, Princeton)

5:00 pm – 5:15 pm

Break

5:15 pm – 6:15 pm

Panels 7, 8, & 9

Panel 7: Frontier research on statistical genetics
Friend 004

Jun Liu, Haiyan Huang, Jessica Li, Hongyu Zhao (Chair)

Panel 8: Frontier research on biostatistics
Computer Science 104

Ying Lu, Yi Li, Quefeng Li, Ying Wei, Zhezhen Jin, Catherine Liu, Haibo Zhou (Chair)

Panel 9: Frontier Research on high-dimensional data modeling
Friend 101

Donggyu Kim, Kaizheng Wang, Yuan Ke, Yiqiao Zhong, Weichen Wang, Wenxin Zhou, Qiang Sun, Qiman Shao (Chair)

7:00 pm – 9:00 pm

Conference dinner

Ronnie Sircar (Chair, ORFE)

Location: Frick Chemistry Laboratory

Day 3 (May 10, 2023)

8:30 am – 8:55 am

Tony Cai

Optimal Statistical Estimation under Nonstatistical Constraints

Lucy Xia (Chair)

8:55 am – 9:20 am

Jiashun Jin

The Statistics Triangle

Lucy Xia (Chair)

9:20 am – 9:45 am

Per Mykland

Of Lean Cats, Unsupervised Learning, and Contiguity

Yingying Li (Chair)

9:45 am – 10:10 am

Marc Hallin

Inferential theory for generalized dynamic factor models

Yingying Li (Chair)

10:10 am – 10:30 am

Break

10:30 am – 10:55 am

Peter Bühlmann

Invariant!

Xin Tong (Chair)

10:55 am – 11:20 am

Ming Yuan

Statistical Optimality and Computational Tractability of ICA

Xin Tong (Chair)

11:20 am – 11:45 am

Xiao-Li Meng

From Gaussianity to Cauchyanity: A Worthwhile Trade-off for (Ultra) High-dimensional Inference?

Yang Feng (Chair)

11:45 am – 12:10 pm

David Donoho

ScreeNOT: Exact MSE-Optimal Singular Value Thresholding in Correlated Noise

Yang Feng (Chair)

12:10 pm – 12:30 pm

Closing remarks

12:30 pm – 2:00 pm

Lunch

Titles and Abstracts of Invited Talks

So Many Jumps, So Few News

Yacine Ait-Sahlia, Princeton University

This paper relates jumps in high frequency stock prices to firm-level, industry and macroeconomic news, in the form of machine-readable releases from Thomson Reuters News Analytics. We begin by examining the relationship from news to price jumps. We find that relevant new information, both idiosyncratic and systematic, gets incorporated quickly into prices, as economic theory suggests. However, in the reverse direction, from price jumps to news, the situation changes. Whereas we found that most relevant news lead to a jump, the vast majority of price jumps do not have identifiable public news that can explain them. We then analyze the various market microstructure features that lead to jumps without news.

Inducement of population sparsity

Heather Battey, Imperial College London, UK

The work on parameter orthogonalisation by Cox and Reid (1987) is presented as inducement of population-level sparsity. The latter is taken as a unifying theme for the talk, in which sparsity-inducing parameterisations or data transformations are sought. Three recent examples are framed in this light: sparse parameterisations of covariance models; construction of factorisable transformations for the elimination of nuisance parameters; and inference in high-dimensional regression. The solution strategy for the problem of exact or approximate sparsity inducement appears to be context specific and may entail, for instance, solving one or more partial differential equation, or specifying a parameterised path through transformation or parameterisation space.

Independence and functional dependence I

Peter Bickel, University of California at Berkeley, USA

Chatterjee (2019) (see also Dette et al., 2013) introduced a novel simple rank-based measure of dependence between X and Y real, which is 0 iff X and Y are independent, and 1 iff $Y = h(X)$ a.s. for some h. In a follow up of this work Azadkia and Chatterjee (2021) extended these results to conditional independence and applied them to variable selection. Subsequent work by Cao and Bickel (2020) and Shi, Drton and Han (2020) showed that, for testing independence, this statistic, when compared to classical procedures, such as those of Spearman or Blum, Kiefer and Rosenblatt (BKR) has no local power (ROC

area=1/2 asymptotically) for many plausible departures. In this lecture we will analyze this behaviour and its sources completely.

Invariant!

Peter Bühlmann, ETH Zürich, Switzerland

We take Jianqing Fan's recent work on EILLS (Environment Invariant Linear Least Squares) as a motivation and inspiration to discuss aspects of robustness in causality-driven nonlinear representation and machine learning.

Feature screening for case-cohort studies with failure time outcome

Jianwen Cai, University of North Carolina at Chapel Hill, USA

Case-cohort design has been demonstrated to be an economical and effective approach in large cohort studies when the measurement of some covariates on all individuals is expensive. Various methods have been proposed for case-cohort data when the dimension of covariates is smaller than sample size. However, limited work has been done for high-dimensional case-cohort data which are frequently collected in large epidemiological studies. We propose a variable screening method for ultrahigh-dimensional case-cohort data under the framework of proportional hazards model, which allows the covariate dimension increases with sample size at exponential rate. Our procedure enjoys the sure screening property and the ranking consistency under some mild regularity conditions. We further extend this method to an iterative version to handle the scenarios where some covariates are jointly important but are marginally unrelated or weakly correlated to the response. The finite sample performance of the proposed procedure is evaluated via both simulation studies and an application to a real data from the breast cancer study.

Optimal Statistical Estimation under Nonstatistical Constraints

Tony Cai, The Wharton School, University of Pennsylvania, USA

In the conventional statistical framework, a major goal is to develop optimal statistical procedures based on the sample size and statistical model. However, in many contemporary applications, non-statistical concerns such as privacy and communication constraints associated with the statistical procedures become crucial. This raises a fundamental question in data science: how can we make optimal statistical inference under these non-statistical constraints?

In this talk, we explore recent advances in differentially private learning and distributed learning under communication constraints in a few specific settings. Our results demonstrate novel and interesting phenomena and suggest directions for further investigation.

Inference for nonstationary time series with varying periodicity, a smooth trend and covariate effects

Ming-Yen Cheng, Hong Kong Baptist University

Traditional analysis of a periodic time series assumes its pattern remains the same over the entire time range. However, using ad hoc methods some recent empirical studies in climatology and other fields find the amplitude may change along with time and that has important implications. We develop a formal procedure to detect and estimate change-points in the periodic pattern. Often there is also a smooth trend, and sometimes the period is unknown and there can be other covariate effects. Based on a new model that takes into account all these, a three-step estimation procedure is proposed to estimate accurately the unknown period, change-points and varying amplitude in the periodic component, the trend and the covariate effects. First, we adopt penalized segmented least squares estimation for the unknown period with the trend and covariate effects approximated by B-splines. Then, given the period estimate, we construct a novel test statistic and use it in binary segmentation to estimate change-points in the periodic component. Finally, given the period and change-point estimates, we estimate the whole periodic component, trend and covariate effects using B-splines. Asymptotic results for the proposed estimators are derived, including consistency of the period and change-point estimators, and asymptotic normality of the estimated periodic sequence, trend and covariate effects. Simulation results demonstrate appealing performance of the new method, and empirical studies show its advantages.

ScreeNOT: Exact MSE-Optimal Singular Value Thresholding in Correlated Noise

David Donoho, Stanford University

Truncation of the singular value decomposition is a true scientific workhorse. But where to Truncate?

For 55 years the answer, for many scientists, has been to eyeball the scree plot, an approach which still generates hundreds of papers per year.

I will describe ScreeNOT, a mathematically solid alternative deriving from the many advances in Random Matrix Theory over those 55 years. Assuming a model of low-rank signal plus possibly correlated noise, and adopting an asymptotic viewpoint with number of rows proportional to the number of columns, we show that ScreeNOT has a surprising oracle property.

It typically achieves exactly, in large finite samples, the lowest possible MSE for matrix recovery, on each given problem instance – i.e. the specific threshold it selects gives exactly the smallest achievable MSE loss among all possible threshold choices for that noisy dataset and that unknown underlying true low rank model. The method is computationally efficient and robust against perturbations of the underlying covariance structure. (Joint work with Matan Gavish and Elad Romanov, Hebrew University)

New Frontiers in Statistical Inference for Stochastic Processes

Philip A. Ernst, Imperial College London

In 1926, G. Udny Yule considered the following problem: given a sequence of pairs of random variables $\{X_k, Y_k\}$ $(k = 1, 2, \ldots, n)$, and letting $X_i = S_i$ and $Y_i = S_i'$ where S_i and S_i' are the partial sums of two independent random walks, what is the distribution of the empirical correlation coefficient

$$\rho_n = \frac{\sum_{i=1}^n S_i S_i' - \frac{1}{n}(\sum_{i=1}^n S_i)(\sum_{i=1}^n S_i')}{\sqrt{\sum_{i=1}^n S_i^2 - \frac{1}{n}(\sum_{i=1}^n S_i)^2}\sqrt{\sum_{i=1}^n (S_i')^2 - \frac{1}{n}(\sum_{i=1}^n S_i')^2}}?$$

Yule empirically observed the distribution of this statistic to be heavily dispersed and frequently large in absolute value, leading him to call it "nonsense correlation." This unexpected finding led to his formulation of two concrete questions, each of which would remain open for more than ninety years: (i) Find (analytically) the variance of ρ_n as $n \to \infty$ and (ii): Find (analytically) the higher order moments and the density of ρ_n as $n \to \infty$. Ernst, Shepp, and Wyner (*The Annals of Statistics*, 2017) considered the empirical correlation coefficient

$$\rho := \frac{\int_0^1 W_1(t)W_2(t)dt - \int_0^1 W_1(t)dt \int_0^1 W_2(t)dt}{\sqrt{\int_0^1 W_1^2(t)dt - \left\{\int_0^1 W_1(t)dt\right\}^2}\sqrt{\int_0^1 W_2^2(t)dt - \left\{\int_0^1 W_2(t)dt\right\}^2}}$$

of two *independent* Wiener processes W_1, W_2, the limit to which ρ_n converges weakly, as was first shown by P.C.B. Phillips. Using tools from integral equation theory, Ernst et al. (2017) closed question (i) by explicitly calculating the second moment of ρ to be .240522. This talk begins where Ernst et al. (2017) leaves off. I shall explain how we finally succeeded in closing question (ii) by explicitly calculating all moments of ρ (up to order 16). This leads, for the first time, to an approximation to the density of Yule's nonsense correlation. I shall then proceed to explain how we were able to explicitly compute higher moments of ρ when the two independent Wiener processes are replaced by two correlated Wiener processes, two independent Ornstein-Uhlenbeck processes, and two independent Brownian bridges. I will conclude by stating a Central Limit Theorem for the case of two independent Ornstein-Uhlenbeck processes. This result shows that Yule's "nonsense correlation" is indeed not "nonsense" for stochastic processes which admit stationary distributions. Time permitting, I will speak about my most recent work on constructing statistical tests for testing independence of pairs of paths of *non-stationary* stochastic processes. (Joint work with L.C.G. Rogers at Cambridge and Quan Zhou at Texas A&M).

Circular local likelihood regression

Irène Gijbels, Department of Mathematics, KU Leuven, Belgium.

In this talk we present a general framework for estimation of regression models with circular covariates, where the conditional distribution of the response given the covariate is specified through a parametric model. The estimation of the conditional mean, or a transformation of it, is carried out nonparametrically, by maximizing the circular local likelihood. The methodology can be extended to joint estimation of conditional mean and dispersion. Important issues are to study bias and variance and to discuss selection of the concentration parameter(s). The generality of the approach is illustrated with several real-data examples, and simulations. (Joint work with Maria Alonso Pena and Rosa Maria Crujeiras Casais, University of Santiago de Compostella, Spain.)

Inferential theory for generalized dynamic factor models

Marc Hallin, Université libre de Bruxelles, Belgium

The so-called General Dynamic Factor Model (GDFM), where common shocks are loaded via filters, has many advantages over the more classical "static" Factor Model—where the loadings are matrices. While consistent estimation methods have been developed for the GDFM, no asymptotic distribution results (on the model of Bai, J. (2003) Inferential theory for factor models of large dimensions, Econometrica 71, 135–171) were available so far. Exploiting the duality between common shocks and dynamic loadings, we obtain such results and derive the asymptotic distributions of a class of estimators of the common shocks, the dynamic loadings (filters), the common components, and the impulse response functions. (Joint with Matteo Barigozzi, University of Bologna, Italy, and Paolo Zaffaroni, Imperial College, London)

Robustifying Markowitz

Wolfgang Karl Härdle, Humboldt University, Gemany

Markowitz mean-variance portfolios with sample mean and covariance as input parameters feature numerous issues in practice. They perform poorly out of sample due to estimation error, they experience extreme weights together with high sensitivity to change in input parameters. The heavy-tail characteristics of financial time series are in fact the cause for these erratic fluctuations of weights that consequently create substantial transaction costs. In robustifying the weights we present a toolbox for stabilizing costs and weights for global minimum Markowitz portfolios. Utilizing a projected gradient de- scent (PGD) technique, we avoid the estimation and inversion of the covariance operator as a whole and concentrate on robust estimation of the gradient descent increment. Using modern tools of robust statistics we construct a computationally efficient estimator with almost Gaussian properties based on median-of-means uniformly over weights. This robustified Markowitz approach is confirmed by empirical studies on equity markets. We demonstrate that robustified portfolios reach the lowest turnover compared to shrinkage- based and constrained portfolios while preserving or slightly improving out-of-sample performance.

The Statistics Triangle

Jiashun Jin, Carnegie Mellon University

In his Fisher's Lecture in 1996, Efron suggested that there is a philosophical triangle in statistics with "Bayesian", "Fisherian", and "Frequentist" being the three vertices, and most of the statistical methods can be viewed as a convex linear combination of the three philosophies. We collected and cleaned a data set consisting of the citation and bibtex (e.g., title, abstract, author information) data of 83,331 papers published in 36 journals in statistics and related fields, spanning 41 years. Using the data set, we constructed 21 co-citation networks, each for a time window between 1990 and 2015. We propose a dynamic Degree-Corrected Mixed- Membership (dynamic-DCMM) model, where we model the research interests of an author by a low-dimensional weight vector (called the network memberships) that evolves slowly over time. We propose dynamic-SCORE as a new approach to estimating the memberships. We discover a triangle in the spectral domain which we call the Statistical Triangle, and use it to visualize the research trajectories of individual authors. We interpret the three vertices of the triangle as the three primary research areas in statistics: "Bayes", "Biostatistics" and "Nonparametrics". The Statistical Triangle further splits into 15 sub- regions, which we interpret as the 15 representative sub-areas in statistics. These results provide useful insights over the research trend and behavior of statisticians.

Safe Policy Learning through Extrapolation: Application to Pre-trial Risk Assessment

Kosuke Imai: Harvard University, USA

Algorithmic recommendations and decisions have become ubiquitous in today's society. Many of these and other data-driven policies, especially in the realm of public policy, are based on known, deterministic rules to ensure their transparency and interpretability. For example, algorithmic pre-trial risk assessments, which serve as our motivating application, provide relatively simple, deterministic classification scores and recommendations to help judges make release decisions. How can we use the data based on existing deterministic policies to learn new and better policies? Unfortunately, prior methods for policy learning are not applicable because they require existing policies to be stochastic rather than deterministic. We develop a robust optimization approach that partially identifies the expected utility of a policy, and then finds an optimal policy by minimizing the worst-case regret. The resulting policy is conservative but has a statistical safety guarantee, allowing the policy-maker to limit the probability of producing a worse outcome than the existing policy. We extend this approach to common and important settings where humans make decisions with the aid of algorithmic recommendations. Lastly, we apply the proposed methodology to a unique field experiment on pre-trial risk assessment instruments. We derive new classification and recommendation rules that retain the transparency and interpretability of the existing instrument while potentially leading to better overall outcomes at a lower cost.

Fast Distributed Principal Component Analysis for Large-Scale Federated Data

Xihong Lin, Harvard University

Principal component analysis (PCA) is one of the most popular methods for dimension reduction. In light of the rapidly growing large-scale data in federated ecosystems, the traditional PCA method is often not applicable due to privacy protection considerations and large computational burden. Algorithms were proposed to lower the computational cost, but few can handle both high dimensionality and massive sample size under the distributed setting. In this paper, we propose the FAst DIstributed (FADI) PCA method for federated data when both the dimension d and the sample size n are ultra-large, by simultaneously performing parallel computing along d and distributed computing along n. Specifically, we utilize L parallel copies of p-dimensional fast sketches to divide the computing burden along d and aggregate the results distributively along the split samples. We present FADI under a general framework applicable to multiple statistical problems, and establish comprehensive theoretical results under the general framework. We show that FADI enjoys the same non-asymptotic error rate as the traditional PCA when Lp $\geq$ d. We also derive inferential results that characterize the asymptotic distribution of FADI, and show a phase-transition phenomenon as Lp increases. We perform extensive simulations to show that FADI substantially outperforms the existing methods in computational efficiency while preserving accuracy, and validate the distributional phase-transition phenomenon through numerical experiments. We apply FADI to the 1000 Genomes data to study the population structure. (Joint work with Shuting Shen and Junwei Liu at Harvard University)

SIMPLE-RC: Group Network Inference with Non-Sharp Nulls and Weak Signals

Jinchi Lv, University of Southern California.

Large-scale network inference with uncertainty quantification has important applications in natural, social, and medical sciences. The recent work of Fan, Fan, Han and Lv (2022) introduced a general framework of statistical inference on membership profiles in large networks (SIMPLE) for testing the sharp null hypothesis that a pair of given nodes share the same membership profiles. In real applications, there are often groups of nodes under investigation that may share similar membership profiles at the presence of relatively weaker signals than the setting considered in SIMPLE. To address these practical challenges, in this paper we propose a SIMPLE method with random coupling (SIMPLE-RC) for testing the non-sharp null hypothesis that a group of given nodes share similar (not necessarily identical) membership profiles under weaker signals. Utilizing the idea of random coupling, we construct our test as the maximum of the SIMPLE tests for subsampled node pairs from the group. Such technique reduces significantly the correlation among individual SIMPLE tests while largely maintaining the power, enabling delicate analysis on the asymptotic distributions of the SIMPLE-RC test. Our

method and theory cover both the cases with and without node degree heterogeneity. These new theoretical developments are empowered by a second-order expansion of spiked eigenvectors under the ℓ_∞-norm, built upon our work for random matrices with weak spikes. Our theoretical results and the practical advantages of the newly suggested method are demonstrated through several simulation and real data examples. This is a joint work with Jianqing Fan, Yingying Fan and Fan Yang.

Random Planted Forest: a directly interpretable tree ensemble

Enno Mammen, Heidelberg University, Gemany.

We introduce a novel interpretable and tree-based algorithm for prediction in a regression setting in which each tree in a classical random forest is replaced by a family of planted trees that grow simultaneously. The motivation for our algorithm is to estimate the unknown regression function from a functional ANOVA decomposition perspective, where each tree corresponds to a function within that decomposition. Therefore, planted trees are limited in the number of interaction terms. The maximal order of approximation in the ANOVA decomposition can be specified or left unlimited. If a first order approximation is chosen, the result is an additive model. In the other extreme case, if the order of approximation is not limited, the resulting model places no restrictions on the form of the regression function. We study the performance of the proposed estimators by extended simulations and we also develop theory for an idealised version of random planted forests. (Joint work with Munir Hiabu (Copenhagen) and Joseph T. Meyer (Heidelberg)

Object Oriented Data Analysis

J. S. Marron, University of North Carolina at Chapel Hill

The rapid change in computational capabilities has made Big Data a major modern statistical challenge. Less well understood is the rise of Complex Data as a perhaps greater challenge. Object Oriented Data Analysis (OODA) is a framework for addressing this, in particular providing a general approach to the definition, representation, visualization and analysis of Complex Data. The notion of OODA generally guides data analysis, through providing a useful terminology for interdisciplinary discussion of the many choices typically needed in modern complex data analyses. The main ideas are illustrated via several real data examples. Methods for analyzing sets of data objects that combine diverse data types, through understanding both joint and individual variation will be featured.

From Gaussianity to Cauchyanity: A Worthwhile Trade-off for (Ultra) High-dimensional Inference?

Xiao-Li Meng, Harvard University

Gaussianity plays a critical role in statistical inference due to the versatility of normal approximations. However, estimating and computing the covariance matrix required for normal approximation becomes challenging in high-dimensional inferences, particularly when the inference dimension is much larger than the data size. Drton and Xiao's (2016, Bernoulli) unexpected discovery and conjecture, later proven by Pillai and Meng (2016, Annals of Statistics), revealed the possibility of constructing many inferential pivotal quantitates for the normal mean that doesn't involve estimating or computing any covariance matrices. The price for this seemingly implausible construction is the replacement of the normal distribution with a standard Cauchy distribution, which has much heavier tails. However, Liu and Xie demonstrated in their 2020 JASA paper that eliminating the dependence on covariance matrices through Cauchyanity does not necessarily mean losing sharpness of inference, such as the power of the test. This talk provides an overview of this line of development and aims to encourage deeper and broader investigations into the hidden power of Cauchyanity as an additional tool for circumventing the curse of dimensionality created by arbitrary dependence structures.

Distributional Regression with Optimal Transports

Hans Mueller, University of California, Davis, USA

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 and associated regression problems are of broad interest. A recent approach is Wasserstein regression that utilizes tangent bundles of the Wasserstein metric space. The search for an intrinsic method motivated a novel transport algebra in the space of optimal transports, which can be harnessed for transport based distributional regression. This approach is illustrated with an autoregressive optimal transport model for distributional time series. (Joint work with Changbo Zhu, Notre Dame University)

Of Lean Cats, Unsupervised Learning, and Contiguity

Per Mykland, University of Chicago

This talk concerns principal component analysis (PCA) and singular value decomposition (SVD) in high dimensional high frequency data. We show that these analyses can reduce to a simpler form with the help of contiguity, thereby making it possible to conduct a deeper exploration of estimators. As an example of application, we focus on SVD based portfolios of financial data. Such portfolios offer the possibility to take the index concept further. They are also a class of implementable and publicly disclosable financial algorithms, which allow the lean cats (the customers) to keep more of their own funds. We show that contiguity can help to obtain hard-to-reach quantities such as how long to learn from SVD/PCA before unleashing estimated singular vectors into a trading strategy. (With work with Lan Zhang at University of Illinois Chicago)

Optimal nonparametric testing of Missing Completely At Random, and its connections to compatibility

Richard Samworth, Cambridge University, UK

Given a set of incomplete observations, we study the nonparametric problem of testing whether data are Missing Completely At Random (MCAR). Our first contribution is to characterise precisely the set of alternatives that can be distinguished from the MCAR null hypothesis. This reveals interesting and novel links to the theory of Fréchet classes (in particular, compatible distributions) and linear programming, that allow us to propose MCAR tests that are consistent against all detectable alternatives. We define an incompatibility index as a natural measure of ease of detectability, establish its key properties, and show how it can be computed exactly in some cases and bounded in others. Moreover, we prove that our tests can attain the minimax separation rate according to this measure, up to logarithmic factors.

On estimating power spectral density for independent component analysis

Kinh N. Truong, University of North Carolina at Chapel Hill

Independent Component Analysis (ICA) attempts to recover signals from a linear mixture of independent sources. Many existing algorithms have been implemented by modeling the marginal density functions of each source or latent random variable. Motivated by the spatial-temporal nature found in many applications, we approach the ICA problem by treating the sources as independent stochastic processes and develop several algorithms to estimate their features and the mixture parameters. The talk will highlight methods for modeling the color sources parametrically or non-parametrically. We will describe J. Fan's contribution to one of the areas in estimating the spectral density associated with the latent stochastic processes.

Deep Learning for Partial Linear Cox Model

Jane-Ling Wang, Univ. of California, Davis, USA

While deep learning approaches to survival data have demonstrated empirical success in applications, most of these methods are difficult to interpret and mathematical understanding of them is lacking. This paper studies the partially linear Cox model, where the nonlinear component of the model is implemented using a deep neural network. The proposed approach is flexible and able to circumvent the curse of dimensionality, yet it facilitates interpretability of the effects of treatment covariates on survival. We establish asymptotic theories of maximum partial likelihood estimators and show that our nonparametric deep neural network estimator achieves the minimax optimal rate of convergence (up to a poly-logarithmic factor). Moreover, we prove that the corresponding finite-dimensional estimator for treatment covariate effects is $\sqrt{n}$-consistent, asymptotically normal, and attains semiparametric efficiency. Extensive simulation studies and

analyses of two real survival datasets show the proposed estimator produces confidence intervals with superior coverage as well as survival time predictions with superior concordance to actual survival times. (Joint work with Qixian Zhong, Xiamen Univ. and Jonas Mueller, Clean Lab)

Autoregressive Networks

Qiwei Yao, London School of Economic, UK

We propose a first-order autoregressive model for dynamic network processes in which edges change over time while nodes remain unchanged. The model depicts the dynamic changes explicitly. It also facilitates simple and efficient statistical inference such as the maximum likelihood estimators which are proved to be (uniformly) consistent and asymptotically normal. The model diagnostic checking can be carried out easily using a permutation test. The proposed model can apply to any Erdös-Renyi network processes with various underlying structures. As an illustration, an autoregressive stochastic block model has been investigated in depth, which characterizes the latent communities by the transition probabilities over time. This leads to a more effective spectral clustering algorithm for identifying the latent communities. Inference for a change-point is incorporated into the autoregressive stochastic block model to cater for possible structure changes. The developed asymptotic theory as well as the simulation study affirm the performance of the proposed methods. Application with three real data sets illustrates both relevance and usefulness of the proposed models. If time permits, we also elucidate how the AR model can accommodate node heterogeneity, edge sparsity, transitivity and other stylized features in network data. (Joint work with Binyan Jiang and Jialiang Li.)

Statistical Optimality and Computational Tractability of ICA

Ming Yuan, Columbia University

Independent component analysis (ICA) is a powerful and general data analysis tool. Yet there is an increasing amount of empirical evidence that the classical methods for ICA are not well suited for modern applications, both computationally and statistically, where the effect of dimensionality is not negligible. We will investigate the optimal sample complexity and statistical performance for ICA, and how considerations of computational tractability may affect them. We will also introduce estimating procedures for ICA that are both statistically efficient and computationally tractable. Our development exploits the close connection between ICA and moment estimation and reveals a number of new insights for both problems.

New statistical learning method for independent component analysis in the presence of noise

Chunming Zhang, University of Wisconsin-Madison, USA

Independent Component Analysis (ICA) is a widely used unsupervised learning method in medical imaging and signal processing, aimed at extracting non-Gaussian independent components (ICs) from multi-dimensional data. However, existing optimization methods often recover ICs from observed signals in unrealistic noiseless settings, with limited theoretical guarantees. We propose a new framework for "noisy ICA" that tackles this challenge from different perspectives, inspired by the desire to identify latent components resembling neural sources of cortical origin from electroencephalography (EEG) recordings of brain activity. Our approach not only directly estimates ICs but also enables the estimation of the unknown number of latent ICs. We develop a computationally efficient algorithm that solves the non-convex and non-smooth optimization problem with guaranteed convergence. Furthermore, we prove that our estimator is consistent under mild conditions. Numerical simulations demonstrate that our approach outperforms some existing methods. Finally, we apply our method to EEG data and show that it can reveal brain source signals with improved quantity and quality.

A Flexible and Parsimonious Modelling Strategy for Clustered Data Analysis

Wenyang Zhang, The University of York, UK

Statistical modelling strategy is the key for success in data analysis. The trade-off between flexibility and parsimony plays a vital role in statistical modelling. In clustered data analysis, in order to account for the heterogeneity between the clusters, certain flexibility is necessary in the modelling, yet parsimony is also needed to guard against the complexity and account for the homogeneity among the clusters. In this talk, I will introduce a flexible and parsimonious modelling strategy for clustered data analysis. The strategy strikes a nice balance between flexibility and parsimony, and accounts for both heterogeneity and homogeneity well among the clusters, which often come with strong practical meanings. In fact, its usefulness has gone beyond clustered data analysis, it also sheds promising lights on transfer learning. An estimation procedure is developed for the unknowns in the resulting model, and asymptotic properties of the estimators are established. Intensive simulation studies are conducted to demonstrate how well the proposed methods work, and a real data analysis is also presented to illustrate how to apply the modelling strategy and associated estimation procedure to answer some real problems arising from real life.

M-Optimal designs through the lens of modern optimization

Hui Zou, University of Minnesota

The usual optimal designs are devised to minimize the variance of the least-squares estimator, although the notion of minimum variance can depend on the specific criterion. Modern regression methods are often biased and hence it is more appropriate to consider

a different optimality criterion for optimal design. We tackle this problem by proposing the M-optimal design and use ridge regression as a concrete example. We address technical issues with regard to the practical application and the inference theory for the corresponding estimator based on the M-optimal design. The promising performance of M-optimal design over the usual optimal designs is demonstrated with numerical studies.

Titles and Abstracts of Poster Presentations

A Nonstochastic Control Approach to Optimization

Xinyi Chen, Princeton University

Tuning optimizer hyperparameters, notably the learning rate to a particular optimization instance, is an important but nonconvex problem. Therefore iterative optimization methods such as hypergradient descent lack global optimality guarantees in general.

We propose an online nonstochastic control methodology for mathematical optimization. The choice of hyperparameters for gradient based methods, including the learning rate, momentum parameter and preconditioner, is described as feedback control. The optimal solution to this control problem is shown to encompass preconditioned adaptive gradient methods with varying acceleration and momentum parameters. Although the optimal control problem by itself is nonconvex, we show how recent methods from online nonstochastic control based on convex relaxation can be applied to compete with the best offline solution. This guarantees that in episodic optimization, we converge to the best optimization method in hindsight.

Geometric Exploration of Random Objects Through Optimal Transport

Yaqing Chen, Rutgers University

We propose new tools for the geometric exploration of data objects taking values in a general separable metric space. For a random object, we first introduce the concept of depth profiles. Specifically, the depth profile of a point in the metric space is the distribution of distances between the very point and the random object. Depth profiles can be harnessed to define transport ranks based on optimal transport, which capture the centrality of each element in the metric space with respect to the probability measure induced by the random object. We study the properties of transport ranks and show that they provide an effective device for detecting and visualizing patterns in samples of random objects. In particular, we establish the theoretical guarantees for the estimation of the depth profiles and the transport ranks for a wide class of metric spaces, followed by practical illustrations on distributional data comprising a sample of age-at-death distributions for different countries and compositional data for electricity generation for the U.S. states.

Adapted Sequential Variational Auto-encoder for modelling human contact dynamics in Germany during the COVID-19 pandemic

Yu Chen, Imperial College London

Characterising and understanding changes in human contact patterns are fundamental to disease modelling and forecasting, but also for characterising changing human behaviours. We are building AI models to capture contact patterns between two contacting individuals at population level by age, sex and other characteristics. Crucially, human contacts are highly structured into peer-based contacts, parent-child contacts, grandparent-child contacts, and further features. Consequently, we aim to identify the changes in age-specific contact patterns since COVID-19 at high resolution (by 1-year age bands), going far beyond the state-of-the-art approaches. We compare the latest sequential variational auto-encoder model and our Bayesian rate consistency model with Hilbert Space Gaussian Process approximation in terms of scalability and accuracy in longitudinal analysis. We combine these two models and present a novel sequential variational auto-encoder approach in the survey contact area which allows us to estimate the dynamics in the time x (sex-age) x (sex-age) space. We also apply our adapted sequential variational auto-encoder model to human contact data in Germany collected between May 2020 to April 2021 over 20 longitudinal survey waves.

Factor Modeling for Volatility

Yi Ding, University of Macau

We establish a framework to study the factor structure in stock variance un- der a high-frequency and high-dimensional setup. We prove the consistency of conducting principal component analysis on realized variances in estimating the factor structure. Moreover, based on strong empirical evidence, we propose a multiplicative volatility factor (MVF) model, where stock variance is represented by a common variance factor and a multiplicative lognormal idiosyncratic component. We further show that our MVF model leads to significantly improved volatility prediction. The favorable performance of the proposed MVF model is seen in both US stocks and global equity indices.

A New Geometric Two-Sample Test for Object Data

Paromita Dubey, University of Southern California

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 poster I will introduce a novel geometrical framework to distinguish between populations of random objects. The test statistic is based on the differences in the depth 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.

Approximate message passing from random initialization with applications to Z2 synchronization

Wei Fan, University of Pennsylvania

This paper 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 provides the first non-asymptotic characterization of AMP in this model without requiring either an informative initialization (e.g., spectral initialization) or sample splitting.

Mapping sleep's phenotypic and genetic links to the brain and heart: a systematic analysis of multimodal brain and cardiac images in the UK Biobank.

Zirui Fan, University of Pennsylvania

Sleep is essential for the health of the brain and heart, but a systematic analysis of the relationship between sleep and brain/heart and their genetic underpinnings is lacking. Using imaging features of organ (brain and heart) structures and functions as clinical endophenotypes, we present a systematic genetic investigation of sleep-brain/heart connections from over 40,000 subjects in the UK Biobank. We identified novel phenotypic and genetic links between sleep and a wide range of imaging traits, such as brain structures, white matter integrity, brain activities, as well as cardiac structures and functions. We prioritized imaging modalities and traits for specific sleep conditions, such as the resting brain function measures in the somatomotor network with narcolepsy. Sleep and brain/heart had overlapping genetic influences in 39 genomic loci, some of which showed evidence of shared causal genetic variants. In conclusion, large-scale imaging genetic data illuminate the implications of sleep on brain and cardiac health and their genetic links. An interactive web browser (www.ig4sleep.org) has been developed to facilitate exploring our results.

Integrated copula spectrum with applications to test for time-reversibility

Yuichi Goto, Kyushu University

The spectral density plays a pivotal role in time series analysis. Since the classical spectral density is defined as the Fourier transform of autocovariance functions, it fails to capture the distributional features. To overcome this drawback, we consider the spectral density based on copula and show the weak convergence of integrated copula spectra. This result combined with the subsampling procedure enables us to construct a test for time-reversibility. This poster is based on joint work with T. Kley (Georg-August-Univ. Gottingen), R. Van Hecke (Ruhr-Univ. Bochum), S. Volgushev (Univ. of Toronto), H. Dette (Ruhr-Univ. Bochum), and M. Hallin (Univ. libre de Bruxelles).

Contrastive learning: an expansion and shrinkage perspective

Yu Gui, University of Chicago

Contrastive learning is a popular learning framework that achieves remarkable empirical performance, especially when no or few labeled training examples are available. Yet, intriguing puzzles about the role of the projection layer and dimensional collapse phenomena are often reported but not fully understood. In this paper, we identify two major effects—expansion and shrinkage—that contrastive learning promotes. Our theoretical analysis is based on the Gaussian mixture model, which despite simplicity allows a systematic treatment. Our analysis reveals a rich phase transition phenomenon and characterizes generalization properties on downstream tasks, which closely match experimental results. Our expansion and shrinkage perspective is a step toward demystifying the empirical puzzles and has the potential to improve practice in self-supervised learning.

Semiparametric Modeling and Analysis for Longitudinal Network Data

Yinqiu He, University of Wisconsin

We propose a semiparametric latent space modeling approach to the analysis of longitudinal network data. The model consists of a stationary latent space component and a time-heterogeneous node-specific baseline component. We develop semiparametric efficient score equation for the latent space parameter, adjusting for the baseline nuisance component. The resulting estimation is achieved by a one-step Newton-Raphson updating and a suitably penalized log-likelihood function. We establish oracle error bounds for the estimators. We also address the identifiability issue from a quotient manifold point of view. The method is applied to the New York Citi Bike Dataset.

A Rank-Based Sequential Test of Independence

Michael Law, ETH Zurich

We consider the problem of independence testing for two univariate random variables in a sequential setting. By leveraging recent developments on safe, anytime valid inference,

we propose a test martingale that has uniform type I error control and derive explicit bounds on the finite sample performance of the test and the expected stopping time. We demonstrate the empirical performance of the procedure on synthetic and real data. Furthermore, since the proposed test is distribution free under the null hypothesis, we empirically simulate the gap due to Ville's inequality — the supermartingale analogue of Markov's inequality — that is commonly applied to control type I error in sequential analysis.

Data fission: splitting a single data point

James Leiner, Carnegie Mellon University

Suppose we observe a random vector X from some distribution P in a known family with unknown parameters. We ask the following question: when is it possible to split X into two parts $f(X)$ and $g(X)$ such that neither part is sufficient to reconstruct X by itself, but both together can recover X fully, and the joint distribution of $(f(X), g(X))$ is tractable? As one example, if $X = (X_1, \ldots, X_n)$ and P is a product distribution, then for any $m < n$, we can split the sample to define $f(X) = (X_1, \ldots, X_m)$ and $g(X) = (X_{m+1}, \ldots, X_n)$. Rasines and Young (2021) offers an alternative route of accomplishing this task through randomization of X with additive Gaussian noise which enables post-selection inference in finite samples for Gaussian distributed data and asymptotically for non-Gaussian additive models. In this paper, we offer a more general methodology for achieving such a split in finite samples by borrowing ideas from Bayesian inference to yield a (frequentist) solution that can be viewed as a continuous analog of data splitting. We call our method data fission, as an alternative to data splitting, data carving and p-value masking. We exemplify the method on a few prototypical applications, such as post-selection inference for trend filtering and other regression problems.

Post-Selection Confidence Inference with Information-Based Model Selection Criteria

Huiming Lin, Harvard University

Best subset variable selection is widely used in both classical and modern statistical inference. It has been increasingly recognized in recent literature that the classical inferential procedures based on the selected submodel lead to invalid confidence intervals in that the frequentist coverage is not attained. In the context of the Akaike information criterion (AIC), Hong et al. (2018) studied an under-coverage phenomenon in terms of overfitting, where the estimate of error variance under the selected submodel is smaller than that for the true model. Under-coverage is particularly troubling in selective inference as it points to inflated Type I errors that would invalidate significant findings. In this work, we delineate a complementary yet provably more deciding factor behind the incorrect coverage of classical confidence intervals under information-based model selection criteria, in terms of altered conditional sampling distributions of pivotal quantities. Resting

on selective techniques developed in other settings, our finite-sample characterization of the selection event under information criteria uncovers its geometry as a union of finitely many intervals on the real line, based on which we derive new confidence intervals with guaranteed coverage for known error variance. We also provide practical strategies to restrain the effect of estimation error of the variance with numerical confirmation in our finite-sample experiments. The proposed method is illustrated by an application to a US consumption dataset, in which we show the proposed post-selection inference arrives at different conclusions compared with the conventional confidence intervals, even when the selected model is the full model.

A Majorization-Minimization Gauss-Newton Method for 1-Bit Matrix Completion

Xiaoqian Liu, University of Texas MD Anderson Cancer Center

In 1-bit matrix completion, the aim is to estimate an underlying low-rank matrix from a partial set of binary observations. We propose a novel method for 1-bit matrix completion called MMGN. To solve the involved nonconvex rank-constrained optimization problem, we apply the majorization-minimization (MM) principle. This yields a sequence of standard low-rank matrix completion problems. We solve each of these sub-problems by a factorization approach that explicitly enforces the assumed low-rank structure and then apply a Gauss-Newton method. Numerical studies, as well as a real-data application, illustrate that in comparison to several existing methods, MMGN outputs more accurate estimates, is often significantly faster, and is less sensitive to the spikiness of the underlying matrix.

Shapley Curves: A Smoothing Perspective

Ratmir Miftachov, Humboldt-Universitat Zu Berlin

Originating from cooperative game theory, Shapley values have become one of the most widely used measures for variable importance in applied Machine Learning. However, the statistical understanding of Shapley values is still limited. In this paper, we take a nonparametric (or smoothing) perspective by introducing Shapley curves as a local measure of variable importance. We propose two estimation strategies and derive the consistency and asymptotic normality both under independence and dependence among the features. This allows us to construct confidence intervals and conduct inference on the estimated Shapley curves. The asymptotic results are validated in extensive experiments. In an empirical application, we analyze which attributes drive the prices of vehicles.

Aggregated functional data model applied on clustering and disaggregation of UK electrical load profiles

Camila P. E. de Souza, Western University

Understanding electrical energy demand at the consumer level plays an important role in planning the distribution of electrical networks and offering of off-peak tariffs, but observing individual consumption patterns is still expensive. On the other hand, aggregated load curves are normally available at the substation level. The proposed methodology separates substation aggregated loads into estimated mean consumption curves, called typical curves, including information given by explanatory variables. In addition, a model-based clustering approach for substations is proposed based on the similarity of their consumers' typical curves and covariance structures. The methodology is applied to a real substation load monitoring dataset from the United Kingdom and tested in eight simulated scenarios. This is joint work with Gabriel Franco and Nancy L. Garcia. Our manuscript has been recently accepted by the Journal of the Royal Statistical Society Series C (Applied Statistics), and a preprint is available at https://arxiv.org/pdf/2106.11448.pdf. Our proposed methodology is implemented as an R package called aggrmodel, currently available online at https://www.github.com/gabrielfranco89/aggrmodel.

Inference for Case Probability in High-dimensional Logistic Regression

Prabrisha Rakshit, Rutgers University

Labeling patients in electronic health records with respect to their statuses of having a disease or condition, i.e. case or control statuses, has increasingly relied on prediction models using high-dimensional variables derived from structured and unstructured electronic health record data. A major hurdle currently is a lack of valid statistical inference methods for the case probability. Here, considering high-dimensional sparse logistic regression models for prediction, we propose a novel bias-corrected estimator for the case probability, the conditional probability for being a case, through the development of linearization and variance enhancement techniques. We establish asymptotic normality of the proposed estimator for any loading vector in high dimensions. We construct a confidence interval for the case probability and propose a hypothesis testing procedure for patient case-control labelling. We demonstrate the proposed method via extensive simulation studies and application to real-world electronic health record data.

Learning from a Biased Sample

Roshni Sahoo, Stanford University

The empirical risk minimization approach to data-driven decision making assumes that we can learn a decision rule from training data drawn under the same conditions as the ones we want to deploy it in. However, in a number of settings, we may be concerned that our training sample is biased, and that some groups (characterized by either observable or unobservable attributes) may be under- or over-represented relative to the general population; and in this setting empirical risk minimization over the training set may fail

to yield rules that perform well at deployment. We propose a model of sampling bias called Γ-biased sampling, where observed covariates can affect the probability of sample selection arbitrarily much but the amount of unexplained variation in the probability of sample selection is bounded by a constant factor. Applying the distributionally robust optimization framework, we propose a method for learning a decision rule that minimizes the worst-case risk incurred under a family of test distributions that can generate the training distribution under Γ-biased sampling. We apply a result of Rockafellar and Uryasev to show that this problem is equivalent to an augmented convex risk minimization problem. We give statistical guarantees for learning a model that is robust to sampling bias via the method of sieves, and propose a deep learning algorithm whose loss function captures our robust learning target. We empirically validate our proposed method in simulations and a case study on ICU length of stay prediction.

Can machines learn weak signals?

Zhouyu Shen, University of Chicago

We study the asymptotical bahavior of Ridge and Lasso when the signal to noise ratio is sufficiently small. We found that Ridge can still learn with an appropriately selected tuning parameter while Lasso is worse than using zero as a predictor. In addition, when the data is independent and identically distributed, we prove that cross-validation is still reliable under weak signals and the out-of-sample R-squared can reflect the signal to noise ratio. Empirical results from six datasets support our theoretical results.

Causal Inference with Corrupted Data: Measurement Error, Missing Values, Discretization, and Differential Privacy

Rahul Singh, MIT

The US Census Bureau will deliberately corrupt data sets derived from the 2020 US Census in an effort to maintain privacy, suggesting a painful trade-off between the privacy of respondents and the precision of economic analysis. To investigate whether this trade-off is inevitable, we formulate a semiparametric model of causal inference with high dimensional corrupted data. We propose a procedure for data cleaning, estimation, and inference with data cleaning-adjusted confidence intervals. We prove consistency, Gaussian approximation, and semiparametric efficiency by finite sample arguments, with a rate of $n - 1/2$ for semiparametric estimands that degrades gracefully for nonparametric estimands. Our key assumption is that the true covariates are approximately low rank, which we interpret as approximate repeated measurements and validate in the Census. In our analysis, we provide nonasymptotic theoretical contributions to matrix completion, statistical learning, and semiparametric statistics. Calibrated simulations verify the coverage of our data cleaning-adjusted confidence intervals and demonstrate the relevance of our results for 2020 Census data.

Efficient shape constrained inference with applications in autocovariance sequence estimation

Hyebin Song, Penn State University

I will present a novel shape-constrained estimator of the autocovariance sequence resulting from a reversible Markov chain. A motivating application for studying this problem is the estimation of the asymptotic variance in central limit theorems for Markov chains. Our approach is based on the key observation that the representability of the autocovariance sequence as a moment sequence imposes certain shape constraints, which we can exploit in the estimation procedure. I will discuss the theoretical properties of the proposed estimator and provide strong consistency guarantees for the proposed estimator. Finally, I will empirically demonstrate the effectiveness of our estimator in comparison with other current state-of-the-art methods for Markov chain Monte Carlo variance estimation, including batch means, spectral variance estimators, and the initial convex sequence estimator.

Correlated Stochastic Block Models: Graph Matching and Community Recovery

Anirudh Sridhar, Princeton University

We consider the task of learning latent community structure from multiple correlated networks. First, we study the problem of learning the latent vertex correspondence between two edge-correlated stochastic block models, focusing on the regime where the average degree is logarithmic in the number of vertices. We derive the precise information-theoretic threshold for exact recovery: above the threshold there exists an estimator that outputs the true correspondence with probability close to 1, while below it no estimator can recover the true correspondence with probability bounded away from 0. As an application of our results, we show how one can exactly recover the latent communities using multiple correlated graphs in parameter regimes where it is information-theoretically impossible to do so using just a single graph.

Non-asymptotic minimax lower bounds for functional estimation under irregularity

Kenta Takatsu, Carnegie Mellon University

In a decision theoretic framework, the minimax lower bound provides the worst-case performance of estimators relative to given statistical models. A wide range of estimation procedures, both parametric and nonparametric, have been studied in relation to this lower bound. The Hájek–Le Cam local asymptotic minimax theorem is a well-known general risk lower bound; however, it has limitations as it only applies to the estimation of differentiable functionals under regular statistical models. We address this limitation by introducing a new risk lower bound with minimal regularity assumptions. The proposed lower bound does not require differentiability of functionals or regularity of statistical

models, extending the efficiency theory to a broader range of situations where standard approaches fail. Additionally, our bound provides a non-asymptotic constant, which can be used to evaluate the precise optimality of proposed estimators in a finite-sample sense. We demonstrate that our bound recovers many known results, including the local asymptotic minimax theorem, under appropriate regularity conditions. We also illustrate the use of our bound by deriving the risk lower bound for estimating the projection of a parameter onto a closed set and analyze the finite-sample efficiency of a maximum likelihood estimator.

Learning from Similar Linear Representations: Adaptivity, Minimaxity, and Robustness

Ye Tian, Columbia University

Representation multi-task learning (MTL) and transfer learning (TL) have achieved tremendous success in practice. However, the theoretical understanding of these methods is still lacking. Most existing theoretical works focus on cases where all tasks share the same representation, and claim that MTL and TL almost always improve performance. However, as the number of tasks grows, assuming all tasks share the same representation is unrealistic. Also, this does not always match empirical findings, which suggest that a shared representation may not necessarily improve single-task or target-only learning performance. In this paper, we aim to understand how to learn from tasks with similar but not exactly the same linear representations, while dealing with outlier tasks. With a known intrinsic dimension, we propose two algorithms that are adaptive to the similarity structure and robust to outlier tasks under both MTL and TL settings. Our algorithms outperform single-task or target-only learning when representations across tasks are sufficiently similar and the fraction of outlier tasks is small. Furthermore, they always perform no worse than single-task learning or target-only learning, even when the representations are dissimilar. We provide information-theoretic lower bounds to show that our algorithms are nearly minimax optimal in a large regime. We also propose an algorithm to adapt to the unknown intrinsic dimension. We conduct two simulation studies to verify our theoretical results.

Universality of Approximate Message Passing algorithms and tensor networks

Tianhao Wang, Yale University

Approximate Message Passing (AMP) algorithms provide a valuable tool for studying mean-field approximations and dynamics in a variety of applications. Although usually derived for matrices having independent Gaussian entries or satisfying rotational invariance in law, their state evolution characterizations are expected to hold over larger universality classes of random matrix ensembles. We develop several new results on AMP universality. For AMP algorithms tailored to independent Gaussian entries, we show that

their state evolutions hold over broadly defined generalized Wigner and white noise ensembles, including matrices with heavy-tailed entries and heterogeneous entrywise variances that may arise in data applications. For AMP algorithms tailored to rotational invariance in law, we show that their state evolutions hold over matrix ensembles whose eigenvector bases satisfy only sign and permutation invariances, including sensing matrices composed of subsampled Hadamard or Fourier transforms and diagonal operators. We establish these results via a simplified moment-method proof, reducing AMP universality to the study of products of random matrices and diagonal tensors along a tensor network. As a by-product of our analyses, we show that the aforementioned matrix ensembles satisfy a notion of asymptotic freeness with respect to such tensor networks, which parallels usual definitions of freeness for traces of matrix products.

Optimal Network Membership Estimation Under Severe Degree Heterogeneity

Jingming Wang, Harvard University

Real networks often exhibit severe degree heterogeneity. In this work, we are interested in studying the effect of degree heterogeneity on estimation of the underlying community structure. We consider the degree-corrected mixed membership model (DCMM) for a symmetric network with n nodes and K communities, where each node i has a degree parameter θ_i and a mixed membership vector π_i. The level of degree heterogeneity is captured by $F_n()$ – the empirical distribution associated with n (scaled) degree parameters. We first show that the optimal rate of convergence for the ℓ_1-loss of estimating π_i's depends on an integral with respect to $F_n()$. We call a method optimally adaptive to degree heterogeneity (in short, optimally adaptive) if it attains the optimal rate for arbitrary $F_n()$. Unfortunately, none of the existing methods satisfy this requirement. We propose a new spectral method that is optimally adaptive, the core idea behind which is using a pre-PCA normalization to yield the optimal signal-to-noise ratio simultaneously at all entries of each leading empirical eigenvector. Technically, we establish a new row-wise large deviation bound for eigenvectors of the regularized graph Laplacian.

Krylov-Bellman boosting: Super-linear policy evaluation in general state spaces

Eric Xia, MIT

We present and analyze the Krylov-Bellman Boosting (KBB) algorithm for policy evaluation in general state spaces. It alternates between fitting the Bellman residual using non-parametric regression (as in boosting), and estimating the value function via the least-squares temporal difference (LSTD) procedure applied with a feature set that grows adaptively over time. By exploiting the connection to Krylov methods, we equip this method with two attractive guarantees. First, we provide a general convergence bound that allows for separate estimation errors in residual fitting and LSTD computation. Consistent with our numerical experiments, this bound shows that convergence rates depend

on the restricted spectral structure, and are typically super-linear. Second, by combining this meta-result with sample-size dependent guarantees for residual fitting and LSTD computation, we obtain concrete statistical guarantees that depend on the sample size along with the complexity of the function class used to fit the residuals. We illustrate the behavior of the KBB algorithm for various types of policy evaluation problems, and typically find large reductions in sample complexity relative to the standard approach of fitted value iteration.

Root and Community Inference on Markovian Network Models

Min Xu, Rutgers University

We introduce the PAPER (Preferential Attachment Plus Erdos–Renyi) model for random networks, where we let a random network G be the union of a preferential attachment (PA) tree T and additional Erdos–Renyi (ER) random edges. The PA tree component captures the fact that real world networks often have an underlying growth/recruitment process where vertices and edges are added sequentially, while 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 early history, in particular the root node, of the unobserved growth process; the root node can be patient zero in a disease infection network or the source of fake news in a social media network. We propose an inference algorithm based on Gibbs sampling that scales to networks with 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 provide a new approach to community detection. This work is part of a discussion paper to appear in JRSSB.

Post-selection inference for e-value based confidence intervals

Ziyi Xu, Carnegie Mellon University

Suppose that one can construct a valid $(1-\delta)$-confidence interval (CI) for each of K parameters of potential interest. If a data analyst uses an arbitrary data-dependent criterion to select some subset S of parameters, then the aforementioned CIs for the selected parameters are no longer valid due to selection bias. We design a new method to adjust the intervals in order to control the false coverage rate (FCR). The main established method is the "BY procedure" by Benjamini and Yekutieli (JASA, 2005). Unfortunately, the BY guarantees require certain restrictions on the selection criterion and on the dependence between the CIs. We propose a natural and much simpler method which is valid under any dependence structure between the original CIs, and any (unknown) selection criterion, but which only applies to a special, yet broad, class of CIs. Our procedure reports $(1-\delta|S|/K)$-CIs for the selected parameters, and we prove that it controls the

FCR at δ for confidence intervals that implicitly invert *e-values*; examples include those constructed via supermartingale methods, via universal inference, or via Chernoff-style bounds on the moment generating function, among others. The e-BY procedure is admissible, and recovers the BY procedure as a special case via calibration. Our work also has implications for post-selection inference in sequential settings, since it applies at stopping times, to continuously-monitored confidence sequences, and under bandit sampling. We demonstrate the efficacy of our procedure using numerical simulations and real A/B testing data from Twitter.

Learning Gaussian Mixtures Using the Wasserstein-Fisher-Rao Gradient Flow

Yuling Yan, Princeton University

Gaussian mixture models form a flexible and expressive parametric family of distributions that has found applications in a wide variety of applications. Unfortunately, fitting these models to data is a notoriously hard problem from a computational perspective. Currently, only moment-based methods enjoy theoretical guarantees while likelihood-based methods are dominated by heuristics such as Expectation-Maximization that are known to fail in simple examples. In this work, we propose a new algorithm to compute the nonparametric maximum likelihood estimator (NPMLE) in a Gaussian mixture model. Our method is based on gradient descent over the space of probability measures equipped with the Wasserstein-Fisher-Rao geometry for which we establish convergence guarantees. In practice, it can be approximated using an interacting particle system where the weight and location of particles are updated alternately. We conduct extensive numerical experiments to confirm the effectiveness of the proposed algorithm compared not only to classical benchmarks but also to similar gradient descent algorithms with respect to simpler geometries. In particular, these simulations illustrate the benefit of updating both weight and location of the interacting particles.

Ranking Inferences Based on the Top Choice of Multiway Comparisons

Mengin Yu, Princeton University

This paper considers ranking inference of n items based on the observed data on the top choice among M randomly selected items at each trial. This is a useful modification of the Plackett-Luce model for M-way ranking with only the top choice observed and is an extension of the celebrated Bradley-Terry-Luce model that corresponds to M=2. Under a uniform sampling scheme in which any M distinguished items are selected for comparisons with probability p and the selected M items are compared L times with multinomial outcomes, we establish the statistical rates of convergence for underlying n preference scores using both l_2-norm and l_∞norm, with the minimum sampling complexity. In addition, we establish the asymptotic normality of the maximum likelihood estimator that allows us to construct confidence intervals for the underlying scores. Furthermore, we

propose a novel inference framework for ranking items through a sophisticated maximum pairwise difference statistic whose distribution is estimated via a valid Gaussian multiplier bootstrap. The estimated distribution is then used to construct simultaneous confidence intervals for the differences in the preference scores and the ranks of individual items. They also enable us to address various inference questions on the ranks of these items. Extensive simulation studies lend further support to our theoretical results. A real data application illustrates the usefulness of the proposed methods convincingly.

Safe Policy Learning under Regression Discontinuity Designs

Yi Zhang, Harvard University

The regression discontinuity (RD) design is widely used for program evaluation with observational data. The RD design enables the identification of the local average treatment effect (LATE) at the treatment cutoff by exploiting known deterministic treatment assignment mechanisms. The primary focus of the existing literature has been the development of rigorous estimation methods for the LATE. In contrast, we consider policy learning under the RD design. We develop a robust optimization approach to finding an optimal treatment cutoff that improves upon the existing one. Under the RD design, policy learning requires extrapolation. We address this problem by partially identifying the conditional expectation function of counterfactual outcome under a smoothness assumption commonly used for the estimation of LATE. We then minimize the worst case regret relative to the status quo policy. The resulting new treatment cutoffs have a safety guarantee, enabling policy makers to limit the probability that they yield a worse outcome than the existing cutoff. Going beyond the standard single-cutoff case, we generalize the proposed methodology to the multi-cutoff RD design by developing a doubly robust estimator. We establish the asymptotic regret bounds for the learned policy using the semi-parametric efficiency theory. Finally, we apply the proposed methodology to empirical and simulated data sets.

A Transformation Based Approach of High Dimensional Linear Hypothesis Testing Problem

Zhe Zhang, Pennsylvania State University

The inference of low-dimensional components and global testing in high-dimensional data analysis has been widely studied in recent years. However, the challenge of testing hypotheses in high-dimensional cases remains an open problem. In this paper, we propose a method for testing general linear hypotheses with high-dimensional loading matrices. Based on the work of Guo et al. (2016), we modify the well-known CQ-test, which is commonly used in location testing problems. The major challenge in our method is approximating the influence of high-dimensional nuisance parameters. We provide conditions under which our test statistic converges to the ideal version when nuisance parameters

are known in advance. Additionally, we establish asymptotic normality under the null hypothesis and propose a power enhancement technique by adding a specific term to the test statistic. Our simulation studies demonstrate that our method improves power without inflating type-I error rates when compared to maximum type statistics under different settings.

Smoothed Robust Phase Retrieval

Zhong Zheng, Pennsylvania State University

The phase retrieval problem with corruption plays an important role in many research fields, as it aims to recover the signal of interest from a set of quadratic measurements with infrequent but arbitrary corruptions. In the literature, robust phase retrieval based on the ℓ_1-loss has been shown to be effective for solving this problem. Despite recent developments in nonconvex algorithms and theoretical analyses, the essential geometric structure of the nonconvex robust phase retrieval is largely unknown to study spurious local solutions, even under the ideal noiseless setting. In this paper, we propose the smoothed robust phase retrieval (SRPR) based on a family of convolution-type smoothed loss functions, where gradients are available for optimization and theoretical analysis. Theoretically, we prove that SRPR enjoys a benign geometric structure with high probability: (1) under the noiseless situation, SRPR has no spurious local solutions, and the target signals are global solutions, which matches the existing result of the least-squares formulations; and (2) under the infrequent but arbitrary corruptions, we can characterize the stationary points of SRPR and prove its benign landscape, which is the first landscape analysis of phase retrieval with corruption in the literature to the best of our knowledge. Also, we prove the local linear convergence rate of gradient descent for solving SRPR under the noiseless situation. Moreover, numerical experiments on both simulated and real datasets are provided to demonstrate the better success rate and efficiency of SRPR.

Graphical Model Inference with Erosely Measured Data

Lili Zheng, Rice University

In this paper, we investigate the Gaussian graphical model inference problem in a novel setting that we call erose measurements, referring to irregularly measured or observed data. For graphs, this results in different node pairs having vastly different sample sizes which frequently arises in data integration, genomics, neuroscience, and sensor networks. Existing works characterize the graph selection performance using the minimum pairwise sample size, which provides little insights for erosely measured data, and no existing inference method is applicable. We aim to fill in this gap by proposing the first inference method that characterizes the different uncertainty levels over the graph caused by the erose measurements, named GI-JOE (Graph Inference when Joint Observations are Erose). Specifically, we develop an edge-wise inference method and an affiliated FDR control procedure, where the variance of each edge depends on the sample sizes associated

with corresponding neighbors. We prove statistical validity under erose measurements, thanks to careful localized edge-wise analysis and disentangling the dependencies across the graph. Finally, through simulation studies and a real neuroscience data example, we demonstrate the advantages of our inference methods for graph selection from erosely measured data.

Deflated HeteroPCA: Overcoming the curse of ill-conditioning in heteroskedastic PCA

Yuchen Zhou, University of Pennsylvania

This paper is concerned with estimating the column subspace of a low-rank matrix $\boldsymbol{X}^\star \in \mathbb{R}^{n_1 \times n_2}$ from contaminated data. How to obtain optimal statistical accuracy while accommodating the widest range of signal-to-noise ratios (SNRs) becomes particularly challenging in the presence of heteroskedastic noise and unbalanced dimensionality (i.e., $n_2 \gg n_1$). While the state-of-the-art algorithm HeteroPCA emerges as a powerful solution for solving this problem, it suffers from "the curse of ill-conditioning," namely, its performance degrades as the condition number of $\boldsymbol{X}^\star$ grows. In order to overcome this critical issue without compromising the range of allowable SNRs, we propose a novel algorithm, called Deflated-HeteroPCA, that achieves near-optimal and condition-number-free theoretical guarantees in terms of both ℓ_2 and $\ell_{2,\infty}$ statistical accuracy. The proposed algorithm divides the spectrum of $\boldsymbol{X}^\star$ into well-conditioned and mutually well-separated subblocks, and applies HeteroPCA to conquer each subblock successively. Further, an application of our algorithm and theory to two canonical examples — the factor model and tensor PCA — leads to remarkable improvement for each application.

Invited Speakers & Participants

Invited Speakers

Yacine Ait-Sahalia (Princeton University)

Heather Battey (Imperial College of London)

Peter Bickel (UC at Berkeley)

Peter Bühlmann (ETH Zürich)

Jianwen Cai (UNC at Chapel Hill)

Tony Cai (UPenn)

Ming-Yen Cheng (Hong Kong Baptist University)

David Donoho (Stanford)

Philip A. Ernst (Imperial College London)

Irène Gijbels (KU Leuven in Belgium)

Marc Hallin (Université libre de Bruxelles)

Wolfgang Härdle (Humboldt-Universität zu Berlin)

Kosuke Imai (Harvard)

Jiashun Jin (CMU)

Xihong Lin (Harvard)

Jinchi Lv (USC)

Enno Mammen (Heidelberg University, Germany)

J.S. Marron (UNC at Chapel Hill)

Xiao-Li Meng (Harvard)

Hans Mueller (UC at Davis)

Per Mykland (University of Chicago)

Richard Samworth (Cambridge)

Kinh Truong (UNC at Chapel Hill)

Jane-Ling Wang (UC at Davis)

Qiwei Yao (LSE)

Ming Yuan (Columbia)

Wenyang Zhang (University of York)

Chunming Zhang (UW at Madison)

Hui Zou (U. Minnesota)

Invited Participants

Caio Almeida (Princeton University)

Marco Avella-Medina (Columbia University)

Krishna Balasubramanian (UC Davis)

Emre Barut (Alexa AI)

Florentina Bunea (Cornell University)

Zongwu Cai (University of Kansas)

Raymond Carroll (Texas A & M University)

Yuxin Chen (University of Pennsylvania)

Rong Chen (Rutgers University)

Xi Chen (New York University)

Elynn Chen (New York University)

Yuejie Chi (Carnegie Mellon University)

Wei Dai (Dimensional Fund Advisor)

Edgar Dobriban (University of Pennsylvania)

Yaqi Duan (MIT)

Ethan Fang (Duke University)

Yang Feng (New York University)

Edward George (University of Pennsylvania)

Wenyan Gong (Two Sigma)

Yongyi Guo (Harvard University)

Xuming He (University of Michigan)

Jianhua Hu (Columbia University)

Li-Shan Huang (National Tsing Hua University)

Haiyan Huang (University of California at Berkeley)

Jianhua Huang (Chinese University of Hong Kong, Shenzhen)

Jiancheng Jiang (University of North Carolina at Charlotte)

Zhezhen Jin (Columbia University)

Yuan Ke (University of Georgia)

Donggyu Kim (KAIST)

Steve Kou (Boston University)

Liza Levina (University Michigan)

Gang Li (University of California at Los Angeles)

Quefeng Li (UNC at Chapel Hill)

Yingying Li (HKUST)

Jessica Li (UCLA)

Yi Li (University of Michigan)

Regina Liu (Rutgers University)

Catherine Liu (The Hong Kong Polytechnic University)

Jun Liu (Harvard University)

Ying Lu (Stanford University)

Cong Ma (University of Chicago)

Zongming Ma (University of Pennsylvania)

Ricardo Masini (Princeton University)

Marcelo C. Medeiros (The University of Illinois at Urbana-Champaign)

Yang Ning (Cornell University)

Yue Niu (University of Arizona)

Lei Qi (Wizard Capital Research)

Annie Qu (University of California at Irvine)

Qiman Shao (Southern University of Science and Technology)

Igor Silin (Two Sigma)

Dylan Small (University of Pennsylvania)

Rui Song (Amazon & NCSU)

Bill Strawderman (Rutgers University)

Weijie Su (University of Pennsylvania)

Qiang Sun (University of Toronto)

Jiayang Sun (George Mason University)

Runlong Tang (JP Morgan Chase & Co.)

Cheng Yong Tang (Temple University)

Francesca Tang (Princeton University)

Xin Tong (University of Southern California)

David Tyler (Rutgers University)

Lan Wang (University of Miami)

Kaizheng Wang (Columbia University)

Zhaoran Wang (Northwestern University)

Yazhen Wang (University of Wisconsin at Madison)

Weichen Wang (Univesrity of Hong Kong)

Weining Wang (University of York)

Yuting Wei (University of Pennsylvania)

Ying Wei (Columbia University)

Haolei Weng (Michigan State University)

Lucy Xia (HKUST)

Yao Xie (Georgia Institute of Technology)

Minge Xie (Rutgers University)

Lingzhou Xue (Penn State University)

Lirong Xue (Jane Street)

Zhuoran Yang (Yale University)

Jiawei Yao (Tower Research Capital)

Zhiliang Ying (Columbia University)

Yan Yu (University of Cincinnati)

Yao Zeng (University of Pennsylvania)

Cun-Hui Zhang (Rutgers University)

Jin-Ting Zhang (National University of Singapore)

Heping Zhang (Yale University)

Emma Zhang (University of Miami)

Zhengjun Zhang (University of Wisconsin)

Hongyu Zhao (Yale University)

Linda Zhao (University of Pennsylvania)

Tian Zheng (Columbia University)

Yiqiao Zhong (University of Wisconsin-Madison)

Haibo Zhou (UNC at Chapel Hill)

Harrison Zhou (Yale University)

Wenxin Zhou (UCSD)

Yinchu Zhu (Brandeis University)

Hongtu Zhu (University of North Carolina at Chapel Hill)

Ji Zhu (University of Michigan)