The 9th International Forum on Statistics

July 14-15, 2023

Abstract

School of Statistics, Renmin University of China
Center for Applied Statistics, Renmin University of China

Beijing · China

Contents

Plenary Session

From Polar cods to entropic central limit theorem

Speaker： Zhiming Ma (Chinese Academy of Sciences)

Abstract: Polar codes are the first class of channel codes which have been proved to reach Shannon capacity by martingale methods of probability theory. In this talk, I shall briefly review some probabilistic methods used in theoretic analysis of Polar cods, and report some of our resent results in this research direction, including entropic central limit theorem.

如何学习学习方法论？-- 兼论大模型的本质

Speaker： Zongben Xu (Xi'an Jiaotong University)

Abstract: 学习方法论是指导、管理学习者如何学习/完成学习任务的一般原则与方法学。在机器学习从人工化，走向自动化,迈向自主化的大趋势下，让机器学会人类的学习方法论，或者更严格地说，学会模拟学习方法论(Simulate Learning Methodology，SLeM)成为 AI 发展的必需，具有重大的科学意义和应用价值。本报告严格定义学习学习方法论问题，提出 SLeM 的元学习模式和“超参数化”求解方法，建立 SLeM 泛化性理论，并应用于多个机器学习自动化问题,展示其有效性。

我们说明：SLeM 是实现通用人工智能的主要途径,本质是学习从任务到方法的映照,数学上是无穷维空间上的机器学习问题。以 ChatGPT 为代表的大模型本质上正是在以“蛮力出奇迹”的方式实现 SLeM，而相比较而言，SLeM 的元学习模式则以“低维近似的方式”实现 SLeM。由此可见,SLeM 是非常值得关注和深入研究的新方向。

Fast Distributed Principal Component Analysis of Large-Scale Federated Data

Speaker： Xihong Lin (Harvard University)

Abstract: Principal component analysis (PCA) is one of the most popular methods for dimension reduction. In the 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 $\geqslant$ d. We also derive inferential results that characterize the asymptotic distribution of FADI, and show a phase-transition phenomenon as Lp increases. We also discuss estimation of the number of low ranks of a covariance matrix by Bulk Eigenvalue Matching Analysis (BEMA). 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.

This is joint work with Shuting Shen and Junwei Lu.

On Dynamics-Informed Blending of Machine Learning and Microeconomics

Speaker： Michael I. Jordan (University of California, Berkeley)

Abstract: Statistical decisions are often given meaning in the context of other decisions, particularly when there are scarce resources to be shared. Managing such sharing is one of the classical goals of microeconomics, and it is given new relevance in the modern setting of large, human-focused datasets, and in data-analytic contexts such as classifiers and recommendation systems. I'll discuss several recent projects that aim to explore the interface between machine learning and microeconomics, including leader/follower dynamics in strategic classification, a Lyapunov theory for matching markets with transfers, and the use of contract theory as a way to design mechanisms that perform statistical inference.

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

Speaker： Richard Samworth (University of Cambridge)

Abstract: 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. Our methodology does not require any complete cases to be effective, and is available in the R package MCARtest.

Distinguished Session

Generating Robust Evidence with Multi-institutional EHR Data

Speaker： Tianxi Cai (Harvard University)

Abstract: While clinical trials and cohort studies remain critical sources for studying disease progression and treatment response, they have limitations including the generalizability of the study findings to the real world, the limited ability to examine subgroup effects or test broader hypotheses, and the cost in performing these studies. In recent years, due to the increasing adoption of electronic health records (EHR) and the linkage of EHR with specimen bio-repositories and other research registries, integrated large datasets now open opportunities to generate real-world evidence (RWE). Generating reliable RWE with EHR studies, however, remain highly challenging due to heterogeneity across healthcare centers in their patient population and health dynamics. In addition, sharing detailed patient-level data across institutions remains infeasible due to privacy constraints. In this talk, I will discuss federated approaches to generating RWE using multi-institutional EHR data.

Fusion and i-Fusion (individualized Fusion) Learning

Speaker： Regina Y. Liu (Rutgers University)

Abstract: Advanced data collection technology nowadays has often made inferences from diverse data sources easily accessible. Fusion learning refers to combining inferences from multiple sources or studies to make a more effective overall inference than that from any individual source or study alone. We focus on the tasks: 1) Whether/When to combine inferences? 2) How to combine inferences efficiently? 3) How to combine inference to enhance an individual or target study?

We present a general framework for nonparametric and efficient fusion learning for inference on multi-parameters, which may be correlated. The main tool underlying this framework is the new notion of depth confidence distribution (depth-CD), which is developed by combining data depth, bootstrap and confidence distributions. We show that a depth-CD is an omnibus form of confidence regions, whose contours of level sets shrink toward the true parameter value, and thus an all-encompassing

inferential tool. The approach is shown to be efficient, general and robust. It readily applies to heterogeneous studies with a broad range of complex and irregular settings. This property also enables the approach to utilize indirect evidence from incomplete studies to gain efficiency for the overall inference. The approach will be shown with simulation studies and real applications in aircraft landing performance tracking and in financial forecasting.

This talk covers joint works with Dungan Liu (University of Cincinnati), Jieli Shen (Goldman Sachs) and Minge Xie (Rutgers University).

Quantum Machine Learning

Speaker： Yazhen Wang (University of Wisconsin-Madison)

Abstract: Quantum computation and quantum information have attracted great attention on multiple frontiers of scientific fields ranging from physics to chemistry and engineering, as well as from computer science to mathematics and statistics. As randomness and uncertainty are deeply rooted in quantum computing, statistics can play an important role in quantum computation, which in turn offers great potential to revolutionize computational statistics and data science. This talk will give a brief review on quantum computing and machine learning and then present a statistical learning problem and its quantum solution to illustrate the quantum advantage in statistics.

Spectral Learning for High Dimensional Tensors

Authors： Ming Yuan (Columbia University)

Abstract: Matrix perturbation bounds developed by Weyl, Davis, Kahan and Wedin and others play a central role in many statistical and machine learning problems. I shall discuss some of the recent progresses in developing similar bounds for higher order tensors. I will highlight the intriguing differences from matrices, and explore their implications in spectral learning problems.

Adaptive Inference in Sequential Experiments

Speaker： Cun-Hui Zhang (Rutgers University)

Abstract: Sequential data collection has emerged as a widely adopted technique for enhancing the efficiency of data gathering processes. Despite its advantages, such data collection mechanism often introduces complexities to the statistical inference procedure. For instance, the ordinary least squares estimator in an adaptive linear regression model can exhibit non-normal asymptotic behavior, posing challenges for accurate inference and interpretation. We propose a general method for constructing debiased estimator which remedies this issue. The idea is to make use of adaptive linear estimating equations. We establish theoretical guarantees of asymptotic normality, supplemented by discussions on achieving near-optimal asymptotic variance. A salient feature of our estimator is that in the context of multi-armed bandits, our estimator retains the non-asymptotic performance of the least square estimator while obtaining asymptotic normality property. Consequently, this work helps connect two fruitful paradigms of adaptive inference: a) non-asymptotic inference using concentration inequalities and b) asymptotic inference via asymptotic normality.

Genetic Studies of Human Brain Imaging Data

Speaker： Heping Zhang (Yale University)

Abstract: As an essential part of the central nervous system, white matter coordinates communications between different brain regions and is related to a wide range of neurodegenerative and neuropsychiatric disorders. Previous genome-wide association studies (GWAS) have uncovered loci associated with white matter microstructure. However, GWAS suffer from limited reproducibility and difficulties in detecting multi-single nucleotide polymorphism (SNP) and epistatic effects. In this study, we adopt the concept of super-variants, a combination of alleles in multiple loci, to account for potential multi-SNP effects. We perform super-variant

identification and validation for brain connectivity data and white matter fractional anisotropy phenotypes derived from diffusion tensor imaging. To increase reproducibility, we use several data sources. Our identified replicable super-variants contain genetic variants located in genes that have been related to brain structures, cognitive functions, and neuropsychiatric diseases. Our findings provide a better understanding of the genetic architecture underlying white matter microstructure.

This is a joint work with Shiying Wang, Wei Dai, Ting Li, Bingxin Zhao and Hongtu Zhu.

Invited Sessions

Session 1： Advances in Machine Learning Theories and Methods

2D-Shapley: A Framework for Fragmented Data Valuation

Authors： Zhihong Liu (Xi'an Jiaotong University)
Hoang Anh Just (Hoang Anh Just)
Xiangyu Chang (Xi'an Jiaotong University)
Xi Chen (New York University)
Ruoxi Jia (Virginia Tech)

Abstract: Data valuation—quantifying the contribution of individual data sources to certain predictive behaviors of a model—is of great importance to enhancing the transparency of machine learning and designing incentive systems for data sharing. Existing work has focused on evaluating data sources with the shared feature or sample space. How to evaluate fragmented data sources of which each only contains partial features and samples remains an open question. We propose a framework for fragmented data valuation and show its wide applications in this talk.

Consistent Selection of the Number of Groups in Panel Models via Sample-Splitting

Authors： Xuening Zhu (Fudan University)

Abstract: Group number selection is a key question for group panel data modelling. In this work, we develop a cross validation method to tackle this problem. Specifically, we split the panel data into a training dataset and a testing dataset on the time span. We first use the training dataset to estimate the parameters and group memberships. Then we apply the fitted model to the testing dataset and then the group number is estimated by minimizing certain loss function values on the testing dataset. We

design the loss functions for panel data models either with or without fixed effects. The proposed method has two advantages. First, the method is totally data-driven thus no further tuning parameters are involved. Second, the method can be flexibly applied to a wide range of panel data models. Theoretically, we establish the estimation consistency by taking advantage of the optimization property of the estimation algorithm. Experiments on a variety of synthetic and empirical datasets are carried out to further illustrate the advantages of the proposed method.

Efficient, Stable, and Analytic Differentiation of the Sinkhorn Loss

Authors: Yixuan Qiu (Shanghai University of Finance and Economics)
Haoyun Yin (Purdue University)
Xiao Wang (Purdue University)

Abstract: Optimal transport and the Wasserstein distance have become important statistical learning techniques and indispensable building blocks of modern deep generative models, but their computational costs greatly prohibit their applications in statistical machine learning models. Recently, the Sinkhorn loss, as an approximation to the Wasserstein distance, has gained massive popularity, and much work has been done for its theoretical properties. In this work, we consider the differentiation of the Sinkhorn loss, so that it can be used in a gradient-based learning framework. We first demonstrate issues of the widely-used Sinkhorn's algorithm, and show that the L-BFGS algorithm is a potentially better candidate for the forward pass. Then we derive an analytic form of the derivative of the Sinkhorn loss with respect to the input cost matrix, which results in an efficient backward algorithm. We rigorously analyze the convergence and stability properties of the advocated algorithms, and use various numerical experiments to validate the performance of the proposed methods.

Peer-Label Assisted Hierarchical Text Classification

Authors： Feifei Wang (Renmin University of China)

Abstract: Hierarchical text classification (HTC) is a challenging task, in which the

labels of texts can be organized into a category hierarchy. To deal with the HTC problem, many existing works focus on utilizing the parent-child relationships that are explicitly shown in the hierarchy. However, texts with a category hierarchy also have some latent relevancy among labels in the same level of the hierarchy. We refer to these labels as peer labels, from which the peer effects are originally utilized in our work to improve the classification performance. To fully explore the peer-label relationship, we develop a PeerHTC method. This method innovatively measures the latent relevancy of peer labels through several metrics and then encodes the relevancy with a Graph Convolutional Neural Network. We also propose a sample importance learning method to ameliorate the side effects raised by modelling the peer label relevancy. Our experiments on several standard datasets demonstrate the evidence of peer labels and the superiority of PeerHTC over other state-of-the-art HTC methods in terms of classification accuracy.

Session 2: Advance in Design and Analysis of Omic and Observational Studies

Sampling Bias: Impact and Tests

Authors: Jiayang Sun(George Mason University)

Abstract: In observational studies, data may come with sampling or selection bias. Inferential procedures that ignore the bias often fail when sampling bias occurs. Conversely, with no sampling bias, methods that account for it may be less efficient than the standard procedures. We present the development of two testing procedures to assess the null hypothesis of no bias against a general monotone bias or a smooth bias function, respectively. The first procedure involves a restricted, penalized, semi-parametric likelihood ratio test. The second procedure is a semi-parametric spline test. We provide asymptotic properties for relevant empirical processes and discuss the limiting distributions of our test statistics under the null hypothesis of no bias and certain local alternatives. If time permits, we will show some numerical studies and data applications to demonstrate the performance and practical utility of the proposed testing procedures.

Learning Directed Acyclic Graphs for Ligands and Receptors Based on Spatially Resolved Transcriptomic Analysis

Authors： Pei Wang(Icahn School of Medicine at Mount Sinai)

Abstract: Cell-to-cell communication relies on interactions between secreted ligands and cell-surface receptors, which create a highly connected signaling network through many ligand-receptor paths. The latest advance in spatial transcriptomic profiling provides unique opportunities to directly characterize ligand-receptor signaling networks that powers cell-cell communication. In this paper, we propose a novel statistical method to characterize the ligand-receptor interaction networks between adjacent tumor and stroma cells in ovarian tumors based on spatial transcriptomic data.

Learning Directed Acyclic Graphs with Mixed Variables

Authors: Jie Peng(University of California, Davis)
Pei Wang(Icahn School of Medicine at Mount Sinai)
Shrabanti Chowdhury(Icahn School of Medicine at Mount Sinai)

Abstract: In this talk, we will discuss a new tool -- DAGBagM-- to learn directed acyclic graphs (DAGs) with both continuous and binary nodes. DAGBagM allows for either continuous or binary nodes to be parent or child nodes. It employs a bootstrap aggregating strategy to reduce false positives in edge inference. At the same time, the aggregation procedure provides a flexible framework to robustly incorporate prior information on edges. We examine DAGBagM through simulation experiments. We also apply it to proteogenomic datasets from ovarian cancer studies for identifying protein biomarkers related to treatment response.

Error Rate 2.0 and Targeted Therapies for Personalized/Precision Medicine

Authors: Xinping Cui(University of California, Riverside)
Jason Hsu (The Ohio State University)

Song Zhai(Merck)

Abstract: Precision medicine is often perceived as an approach to healthcare that seeks to provide personalized treatment based on a patient's unique genetic makeup. Targeted therapies are becoming the most common precision medicine. The challenge of personalizing a targeted therapy is finding the patient subgroup that has enough drug targets in them for the targeted therapy to act upon. Most CDx (companion diagnostic test) approved by FDA used a single-SNP (or single-gene) biomarker to target patient subgroups for personalized medicine. However, if one single nucleotide polymorphism (SNP) affects efficacy of a treatment, then other linked SNPs may also "statistically" contribute to it. Therefore, ensuring control of Type I error rates is crucial in pharmaceutical drug development, as it directly impacts the rate of incorrect regulatory decisions. In this work, to select the best predictive SNP, we reformulated the SNP testing problem as a problem of Multiple Comparison with the Best and introduce two-layer error rate (Error Rate 2.0) control, where at the across SNPs layer, each test (of SNPi being the best) is at the 5% 1-sided level with no multiplicity control. We justify it by proving that the simultaneous MCB confidence intervals have rigorous LLN interpretation, regardless of whether ties (exact equalities) in the parameters are allowed or not, so long as we present the CIs as closed (not open) intervals. Similar ideas of two-layer error rate control presented in Ding et al. (2018) and Cui et al. (2023, et al.) will also be discussed. Our method is a step toward targeting a patient subgroup in a tailored drug development process.

Session 3: Advance in Machine Learning

A"Physical" Law of Data Separation in Deep Learning

Authors: Weijie Su(University of Pennsylvania)

Abstract: The remarkable abilities of large language models (LLMs) like ChatGPT and GPT-4, stem in part from their alignment with reward models trained on human preference rankings. In this talk, we ask how statistics can enrich LLM performance via this alignment. We spotlight a phenomenon we term 'reward collapse', where, during training, diverse responses receive similar reward distributions regardless of the prompts, thus limiting the

model's adaptability. We attribute this to the current alignment approach's inability to discern varied reward distributions.

Ocean 3D Variable Field Inversion Based on Improved Physics-Informed Neural Networks

Authors: Zhixi Xiong（Sun Yat-sen University）
Yukang Jiang（Sun Yat-sen University）
Wenfang Lu（Sun Yat-sen University）
Xueqin Wang（University of Science and Technology of China）
Ting Tian（Sun Yat-sen University）

Abstract: Using discrete, sparse, and uneven ocean variable observation data to invert continuous ocean variable fields is an essential task in oceanographic research. Most of the previous artificial intelligence inversion methods require time-series gridded data, which lack physical constraints and make it difficult to integrate multi-source data. However, the Physics-Informed Neural Networks (PINN) overcomes these limitations by seamlessly integrating physical mechanism equations and observation samples with a simple structure. The partial differential equation (PDE) embedded in PINN naturally brings strong interpretability.

To improve the neural network structure and training method of PINN, we embedded the Primitive Equations that can describe the movement of ocean currents and diffusion of temperature and salinity using observation data from the Argo project for ocean temperature and salinity and vector velocity reanalysis data from the Copernicus Marine Service of the European Union. We then conducted an inversion and analysis of the subsurface temperature field, salinity field, vector velocity field, and pressure field in the sea area near the central equator of the Pacific Ocean.

Finally, we compared the performance of PINN's inversion method with basic neural network and spatiotemporal data analysis, and our models showed superior performance.

Nonasymptotic Theory for Two-layer Neural Networks: Beyond the bias- variance trade-off

Authors: Huiyuan Wang（Peking University ）
Wei Lin (Peking University)

Abstract: Large neural networks have proved remarkably effective in modern deep learning practice, even in the over parametrized regime where the number of active parameters is much larger than the sample size. This contradicts the classical perspective that a machine learning model must trade off bias and variance for optimal generalization. To resolve this conflict, we present a nonasymptotic generalization theory for two-layer neural networks with ReLU activation function by incorporating scaled variation regularization. Interestingly, the regularizer is equivalent to ridge regression from the angle of gradient-based optimization, but plays a similar role to the group lasso in controlling the model complexity. By exploiting this "ridge–lasso duality," we obtain new prediction bounds for all network widths, which reproduce the double descent phenomenon. Moreover, the overparametrized minimum risk is lower than the underparametrized minimum risk when the signal is strong, and nearly attains the minimax optimal rate over a suitable class of functions. By contrast, we show that over parametrized random feature models suffer from the curse of dimensionality and thus are suboptimal.

Mining Consumer Complaints for Recall Management: A Topic Model for Decision Automation

Authors: Wen Shi (Central South University)
Yujie Qu (Central South University)
Jia Liu (HKUST, HK)

Abstract: Although consumer complaints have been acknowledged as the main trigger of product recalls in several recall-intensive sectors (such as cars, food, drinks, and pharmaceuticals), both firms and regulators lack human and technology resources to filter consumer complaints for further trend analysis. We develop a topic model, called the

hierarchically dual Pitman-Yor process (HDPYP), that can automatically process and analyze largescale consumer complaints and their associated recall statements. The HDPYP can extract defect topics, predict the importance of a consumer complaint, and anticipate the topic distribution of any resulting recall statement. This information can assist firms and regulators with important decisions related to issuing recalls and the appropriate wording in recall statements. We apply the HDPYP to consumer complaints and vehicle-recall data sets from the U.S. automobile industry. We demonstrate that the HDPYP provides an effective way to aggregate the textual information in consumer complaints so that the recall prediction accuracy of popular classification models can be improved significantly. We also show that the HDPYP can successfully identify the defect topics and the few important complaints that are worthy of being summarized in a recall statement. Such information can help improve the quality of the recall statements generated by either human or natural language generation tools.

Session 4： Advanced Genetic Analytics

Bi-level Structured Functional Analysis for Genome-wide Association Studies

Authors: Mengyun Wu (Shanghai University of Finance and Economics)
Fan Wang (Renmin University of China)
Yeheng Ge (Shanghai University of Finance and Economics)
Shuangge Ma (Yale School of Public Health)
Yang Li (Renmin University of China)

Abstract: Genome-wide association studies (GWAS) have led to great successes in identifying genotype–phenotype associations for complex human diseases. In such studies, the high dimensionality of single nucleotide polymorphisms (SNPs) often makes analysis difficult. Functional analysis, which interprets SNPs densely distributed in a chromosomal region as a continuous process rather than discrete observations, has emerged as a promising avenue for overcoming the high dimensionality challenges. However, the majority of the existing functional studies continue to be individual SNP based and are unable to sufficiently account for the intricate underpinning structures of SNP data. SNPs are often found in groups (e.g.,

genes or pathways) and have a natural group structure. Additionally, these SNP groups can be highly correlated with coordinated biological functions and interact in a network. Motivated by these unique characteristics of SNP data, we develop a novel bi-level structured functional analysis method and investigate disease-associated genetic variants at the SNP level and SNP group level simultaneously. The penalization technique is adopted for bi-level selection and also to accommodate the group-level network structure. Both the estimation and selection consistency properties are rigorously established. The superiority of the proposed method over alternatives is shown through extensive simulation studies. A type 2 diabetes SNP data application yields some biologically intriguing results.

A Network Approach to Compute Hypervolume under ROC Manifold for Multi-class Biomarkers

Authors: Jialiang Li (National University of Singapore)

Abstract: Computation of hypervolume under ROC manifold (HUM) is necessary to evaluate biomarkers for their capability to discriminate among multiple disease types or diagnostic groups. However the original definition of HUM involves multiple integration and thus a medical investigation for multi-class ROC analysis could suffer from huge computational cost when the formula is implemented naively. We introduce a novel graph-based approach to compute HUM efficiently in this paper. The computational method avoids the time-consuming multiple summation when sample size or the number of categories is large. We conduct extensive simulation studies to demonstrate the improvement of our method over existing R packages. We apply our method to two real biomedical data sets to illustrate its application.

Pathological Imaging-assisted Cancer Gene-environment Interaction Analysis

Authors: Kuangnan Fang ((Xiamen University))

Abstract: Gene-environment (G-E) interactions have important implications for cancer outcomes and phenotypes beyond the main G and E effects. Compared to main-effect-only analysis, G-E interaction analysis more seriously suffers from a lack of information caused by higher dimensionality, weaker signals, and other factors. It is also uniquely challenged by the "main effects, interactions'' variable selection hierarchy. Effort has been made to bring in additional information to assist cancer G-E interaction analysis. In this study, we take a strategy different from the existing literature and borrow information from pathological imaging data. Such data is a "byproduct'' of biopsy, enjoys broad availability and low cost, and has been shown as informative for modeling prognosis and other cancer outcomes/phenotypes in recent studies. Building on penalization, we develop an assisted estimation and variable selection approach for G-E interaction analysis. The approach is intuitive, can be effectively realized, and has competitive performance in simulation. We further analyze The Cancer Genome Atlas (TCGA) data on lung adenocarcinoma (LUAD). The outcome of interest is overall survival, and for G variables, we analyze gene expressions. Assisted by pathological imaging data, our G-E interaction analysis leads to different findings with competitive prediction performance and stability.

Genetic Underpinnings of Brain Structural Connectome for Young Adults

Authors: Yize Zhao (Yale University)
Changgee Chang (Indiana University)
Jingwen Zhang (Boston University)
Zhengwu Zhang (University of North Carolina)

Abstract: With distinct advantages in power over behavioral phenotypes, brain imaging traits have become emerging endophenotypes to dissect molecular contributions to behaviors and neuropsychiatric illnesses. Among different imaging features, brain structural connectivity (i.e., structural connectome) which summarizes the anatomical connections between different brain regions is one of the most cutting-edge while under-investigated traits; and the genetic influence on the structural connectome variation remains highly elusive. Relying on a landmark

imaging genetics study for young adults, we develop a biologically plausible brain network response shrinkage model to comprehensively characterize the relationship between high dimensional genetic variants and the structural connectome phenotype. Under a unified Bayesian framework, we accommodate the topology of brain network and biological architecture within the genome; and eventually establish a mechanistic mapping between genetic biomarkers and the associated brain sub-network units. An efficient expectation-maximization algorithm is developed to estimate the model and ensure computing feasibility. In the application to the Human Connectome Project Young Adult (HCP-YA) data, we establish the genetic underpinnings which are highly interpretable under functional annotation and brain tissue eQTL analysis, for the brain white matter tracts connecting the hippocampus and two cerebral hemispheres. We also show the superiority of our method in extensive simulations. Supplementary materials for this article are available online.

Session 5: Advanced in Complex Data Analysis

Optimal Reconciliation with Immutable Forecasts

Authors: Feng Li (Central University of Finance and Economics)

Abstract: The practical importance of coherent forecasts in hierarchical forecasting has inspired many studies on forecast reconciliation. Under this approach, base forecasts are produced for every series in the hierarchy and are subsequently adjusted to be coherent in a second reconciliation step. It is sometimes necessary or beneficial to keep forecasts of some variables unchanged after forecast reconciliation. In this paper, we formulate a reconciliation methodology that keeps forecasts of a pre-specified subset of variables “immutable”. (Published in European Journal of Operational Research (2013)).

Forecast Combinations: Modern Perspectives and Approaches

Authors: Yanfei Kang (Beihang University)

Xiaoqian Wang (Monash University)
Bohan Zhang (Beihang University)
Li Li (Beihang University)
Wei Cao (Beihang University)
Xixi Li (The University of Manchester)
Rob Hyndman (Monash University)
Fotios Petropoulos (The University of Bath)
Feng Li (Central University of Finance and Economics)
Anastasios Panagiotelis (The University of Sydney)

Abstract: Forecast combinations have flourished remarkably in the forecasting community and, in recent years, have become part of the mainstream of forecasting research and activities. Combination schemes have evolved from simple combination methods without estimation, to sophisticated methods involving time-varying weights, nonlinear combinations, correlations among components, and cross-learning. They include combining point forecasts, and combining probabilistic forecasts. They also include combining multiple forecasts derived from different methods for a given time series, combining the base forecasts of each series in a hierarchy, and aggregating forecasts computed on different perspectives of the same data. In this talk, I will start from classical forecast combination ideas and present modern perspectives of combining that can offer substantially improved forecasts on average: 1) combining multiple models: feature-based forecasting, 2) hierarchical forecasting: optimal reconciliation with immutable forecasts, and 3) wisdom of the data: improving forecasting by subsampling seasonal time series. This talk concludes with current research gaps and potential insights for future research on forecast combinations.

Accounting for Network Noise in Graph-guided Bayesian Modeling of High-dimensional -omics Data

Authors: Wenrui Li (University of Pennsylvania)
Changgee Chang (Indiana University)
Suprateek Kundu (University of Texas MD Anderson Cancer Center)
Qi Long (University of Pennsylvania)

Abstract: High-dimensional omics data offer great promise in advancing precision

medicine. Knowledge-guided statistical methods for analysis of omics data that can incorporate biological knowledge represented by graphs such as functional genomics have been shown to improve variable selection and predication accuracy and yield biologically more interpretable results, they typically use biological graph extracted from existing databases which is known to be incomplete and contain false edges. To address this issue, we propose a new knowledge-guided Bayesian modeling framework that treats the true biological graph as unknown or latent. Our model uses an adaptive structured shrinkage prior to incorporate the latent true biological graph to facilitate variable selection, and another set of priors motivated by the latent scale network model to connect two sources of noise-contaminated graph data, namely, biological graph extracted from a database and estimated covariance matrix for observed covariates, to the latent true graph. We develop an efficient MCMC algorithm for posterior sampling. We demonstrate the advantages of our model in simulations, and analysis of an AD genomics dataset.

Scaling Limit of DLA on a Long Line Segment

Authors: Eviatar B. Procaccia (Technion)
Yingxin Mu (Universität Leipzig)
Jiayan Ye (Dongguan University of Technology)
Yuan Zhang (Renmin University of China)

Abstract: In this talk, we show that the bulk of 2-dimensional DLA starting from a long line segment on the x-axis has a scaling limit. Such limit can be described as an infinite and stationary version of Diffusion Limited Aggregation (DLA) starting from the whole x-axis, which is invariant against horizontal translations. The main phenomenological difficulty is the multi-scale, non-monotone interaction of the DLA arms. We overcome this via a coupling scheme between the two processes and an intermediate DLA process with absorbing mesoscopic boundary segments. Our result allows to import results from the more amenable infinite stationary DLA process to the more physical finite aggregations. Based on joint work(s) with Yingxin Mu, Eviatar B. Procaccia, and Jiayan Ye.

Session 6： Advanced in Statistical Causal Inference and Applications

Interpretable Sensitivity Analysis for the Baron-Kenny Approach to Mediation with Unmeasured Confounding

Authors: Peng Ding (University of California Berkeley)

Abstract: Mediation analysis assesses the extent to which the treatment affects the outcome indirectly through a mediator and the extent to which it operates directly through other pathways. As the most popular method in empirical mediation analysis, the Baron-Kenny approach estimates the indirect and direct effects of the treatment on the outcome based on linear structural equation models. However, when the treatment and the mediator are not randomized, the estimates may be biased due to unmeasured confounding among the treatment, mediator, and outcome. Building on Cinelli and Hazlett (2020), we propose a sharp and interpretable sensitivity analysis method for the Baron-Kenny approach to mediation in the presence of unmeasured confounding. We first modify their omitted-variable bias formula to facilitate the discussion with heteroskedasticity and model misspecification. We then apply the result to develop a sensitivity analysis method for the Baron-Kenny approach. To ensure interpretability, we express the sensitivity parameters in terms of the partial R2's that correspond to the natural factorization of the joint distribution of the direct acyclic graph for mediation analysis. They measure the proportions of variability explained by unmeasured confounding given the observed variables. Moreover, we extend the method to deal with multiple mediators, based on a novel matrix version of the partial R2 and a general form of the omitted-variable bias formula. Importantly, we prove that all our sensitivity bounds are

attainable and thus sharp.

Strengthen Causal Inference by Leveraging Genetic Data

Authors： Can Yang (The Hong Kong University of Science and Technology)

Abstract: Inferring the causal relationship between a risk factor (exposure) and a complex trait of interest (outcome) is essential in biomedical research and social science. Mendelian Randomization (MR) is a valuable tool for inferring causal relationships among a wide range of traits using summary statistics from genome-wide association studies (GWASs). Existing summary-level MR methods often rely on strong assumptions, resulting in many false positive findings. To relax MR assumptions, ongoing research has been primarily focused on accounting for confounding due to pleiotropy. Here we show that sample structure is another major confounding factor, including population stratification, cryptic relatedness, and sample overlap. We propose a unified MR approach, MR-APSS, which (i) accounts for pleiotropy and sample structure simultaneously by leveraging genome-wide information; (ii) allows to include more genetic instruments with moderate effects to improve statistical power without inflating type I errors. We first evaluated MR-APSS using comprehensive simulations and negative controls, and then applied MR-APSS to study the causal relationships among a collection of diverse complex traits. The results suggest that MR-APSS can better identify plausible causal relationships with high reliability. In particular, MR-APSS can perform well for highly polygenic traits, such as psychiatric disorders and social traits, where the strengths of IVs tend to be relatively weak and existing summary-level MR methods for causal inference are vulnerable to confounding effects. This is a joint work with Xianghong Hu, Jia Zhao, Zhixiang Lin, Yang Wang, Heng Peng, Hongyu Zhao, and Xiang Wan.

Policy Learning with Asymmetric Utilities

Authors: Eli Ben-Michael(Carnegie Mellon University)

Kosuke Imai(Harvard University)
Zhichao Jiang (Sun Yat-sen University)

Abstract: Data-driven decision making plays an important role even in high stakes settings like medicine and public policy. Learning optimal policies from observed data requires a careful formulation of the utility function whose expected value is maximized across a population. Although researchers typically use utilities that depend on observed outcomes alone, in many settings the decision maker's utility function is more properly characterized by the joint set of potential outcomes under all actions. For example, the Hippocratic principle to "do no harm" implies that the cost of causing death to a patient who would otherwise survive without treatment is greater than the cost of forgoing life-saving treatment. We consider optimal policy learning with asymmetric utility functions of this form. We show that asymmetric utilities lead to an unidentifiable social welfare function, and so we first partially identify it. Drawing on statistical decision theory, we then derive minimax decision rules by minimizing the maximum regret relative to alternative policies. We show that one can learn minimax decision rules from observed data by solving intermediate classification problems. We also establish that the finite sample regret of this procedure is bounded by the mis-classification rate of these intermediate classifiers. We apply this conceptual framework and methodology to the decision about whether or not to use right heart catheterization for patients with possible pulmonary hypertension.

Uncertainty Quantification in Synthetic Controls with Staggered Treatment Adoption

Authors: Matias Cattaneo (Princeton University)
Yingjie Feng (Biostatistics, Tsinghua University)
Filippo Palomba (Princeton University)
Rocio Titiunik (Princeton University)

Abstract: We propose principled prediction intervals to quantify the uncertainty of a large class of synthetic control predictions (or estimators) in settings with staggered treatment adoption, offering precise non-asymptotic coverage probability guarantees. From a

methodological perspective, we provide a detailed discussion of different causal quantities to be predicted, which we call causal predictands, allowing for multiple treated units with treatment adoption at possibly different points in time. From a theoretical perspective, our uncertainty quantification methods improve on prior literature by (i) covering a large class of causal predictands in staggered adoption settings, (ii) allowing for synthetic control methods with possibly nonlinear constraints, (iii) proposing scalable robust conic optimization methods and principled data-driven tuning parameter selection, and (iv) offering valid uniform inference across post-treatment periods. We illustrate our methodology with an empirical application studying the effects of economic liberalization in the 1990s on GDP for emerging European countries. Companion general-purpose software packages are provided in Python, R and Stata.

Session 7: Aspects of Data Science; New and Old

Exact and Approximate Moment Derivation for Probabilistic Loops with Non-polynomial Assignments

Authors: Andrey Kofnov (TU Wien)
Efstathia Bura T (TU Wien)
Ezio Bartocci(TU Wien)

Abstract: Probabilistic programs (PPs) are modern tools to automate statistical modeling. They are becoming ubiquitous in AI applications, security/privacy protocols and stochastic dynamical system modeling. Many stochastic continuous-state dynamical systems can be modeled as probabilistic programs with nonlinear non-polynomial updates in non-nested loops. We present two methods, one approximate and one exact, to compute automatically and without sampling moment-based invariants for such probabilistic programs as a closed-form solution in loop iteration. The exact method applies to probabilistic programs with trigonometric and exponential updates and is embedded in the Polar tool. The approximate moment propagation method applies to any nonlinear random function as it exploits the theory of polynomial chaos expansion to approximate non-polynomial updates as the sum of orthogonal polynomials. This translates the dynamical system to a non-nested loop with polynomial updates, and thus renders it

conformable with the \textsc{Polar} tool that computes the moments of all orders of the state variables. We evaluate our methods on an extensive number of examples ranging from modeling monetary policy to several physical motion systems in uncertain environments. The experimental results demonstrate the advantages of our approach with respect to the current state-of-the-art.

Optimal Nonparametric Inference with Two-Scale Distributional Nearest Neighbors

Authors： Jinchi Lv (University of Southern California)

Abstract: In this work, we provide an in-depth technical analysis of the distributional nearest neighbors (DNN), based on which we suggest a bias reduction approach for the DNN estimator by linearly combining two DNN estimators with different subsampling scales, resulting in the novel two-scale DNN (TDNN) estimator. The two-scale DNN estimator has an equivalent representation of WNN with weights admitting explicit forms and some being negative. This is a joint work with Emre Demirkaya, Yingying Fan, Lan Gao, Patrick Vossler and Jingbo Wang.

Cointegrated Matrix Autoregressive Models

Authors: Zebang Li(Rutgers University)
Han Xiao (Rutgers University)

Abstract: We consider the cointegrated autoregressive models for matrix-valued time series. By assuming a bilinear form under the error correction model, we model the cointegration along the rows and columns of the matrix observations. Comparing with the approach of vectorizing the data and applying the VAR models, this model reduces the dimension significantly, and provides insights about the nature of the cointegration by preserving the matrix structure of the data. Both least squares and maximum likelihood estimation are studied with corresponding algorithms and asymptotics. We apply the proposed methodology to Fama-French portfolios and design a trading strategy based on the cointegration, which leads to higher returns than the market, especially during a bear

market.

Breiman's Samplers or Models? There is a Little but Important Difference: Models can be wrong!

Authors：Yannis Yatracos (Tsinghua University in Beijing)

Abstract: Breiman (2001, Statistical Science) urged statisticians to provide statistical tools when the data, X, is obtained from a sampler, f(θ,Y); f is known, parameter θεΘ, Y is random. Discussants of the paper, D. R. Cox and B. Efron, looked at the problem as X-prediction, surprisingly neglecting the statistical inference for θ, and disagreed with the main thrust of the paper. Consequently, mathematical statisticians ignored Breiman's suggestion! However, computer scientists work in Breiman's problem calling f(θ, Y) learning machine. In this talk, following Breiman's suggestion, statistical inference tools are presented for data X=f(θ,Y): a) the Empirical Discrimination Index (EDI), to detect θ-discrimination and identifiability, b) Matching estimates of θ with upper bounds on the errors that depend on the "massiveness" of Θ, c) an approximate posterior inclusive of all θ* drawn from a Θ-sampler, unlike the Rubin (1984, Annals of Statistics) rejection-method followed until now. The results in a) are unique in the literature (YY, 2023, JCGS). Mild assumptions are needed in b) and c). Unlike existing results that need often strong and unverifiable assumptions, the error rates in b) are independent of the data dimension, and when Θ is subset of R^m, m unknown, can be [m_n (log n)/n]^(1/2) in probability, with m_n increasing to infinity as slow as we wish. Approximate posteriors in c) are obtained for any data dimension. When X=f(θ,Y) and a c.d.f., F_θ, for X is assumed, one may be better off using the Sampler and a)-c) since F_θ may be wrong!

Session 8: Bio-Medical Data and Machine Learning: From Data, Model to Decision Making

KESER: Clinical Knowledge Extraction via Sparse Embedding

Regression with EHR data

Authors： Chuan Hong (Harvard Medical School)
Everett Rush (Oak Ridge National Lab)
Molei Liu (Harvard T.H. Chan School of Public Health)
Tianxi Cai (Harvard Medical School)

Abstract: Traditional data mining of EHR data often requires the use of patient-level data, which hinders the ability to share data across institutions. KESER a knowledge extraction pipeline via sparse embedding regression, which efficiently summarizes patient-level longitudinal EHR data into hospital-specific embedding data and enables the extraction of clinical knowledge based only on summary-level data. KESER bypasses the need for patient-level data in individual analyses providing a significant advance in enabling multi-center studies using EHR data.

STAARpipeline: an All-in-one Rare-variant Analysis Tool for Biobank-scale Whole-genome Sequencing Data

Authors: Zilin Li (Northeast Normal University)
Xihao Li (Harvard T.H. Chan School of Public Health)
Xihong Lin (Harvard T.H. Chan School of Public Health)

Abstract: Large-scale whole-genome sequencing (WGS) studies have enabled the analysis of rare variant associations with complex human diseases and traits. Variant set analysis is a powerful approach to studying rare variant associations. However, existing methods have limited ability to define the variant set in the genome, especially for the noncoding genome. We propose a computationally efficient and robust rare variant association-detection framework, STAARpipeline, to automatically annotate a WGS study and perform flexible rare variant association analysis, including gene-centric analysis and fixed-window and dynamic-window-based non-gene-centric analysis by incorporating variant functional annotations. In gene-centric analysis, STAARpipeline groups coding and noncoding variants based on functional categories of genes and incorporate multiple functional annotations. In non-gene-centric analysis, in addition to fixed-size sliding window analysis, STAARpipeline provides a data-adaptive-size dynamic window analysis. All these variant sets could be automatically defined and selected in STAARpipeline.

STAARpipeline also provides analytical follow-up of dissecting association signals independent of known variants via conditional analysis. We applied the STAARpipeline to analyze the total cholesterol in 30,138 samples from the NHLBI Trans-Omics for Precision Medicine (TOPMed) Program. All analyses scale well in computation time and memory. We discover several potentially new significant associations with lipids. In summary, STAARpipeline is a powerful and resource-efficient tool for association analysis of biobank-scale WGS studies.

Understanding Adaptive Gradient Methods in Training Neural Networks: Risk Bounds and Practical Implications

Authors: Difan Zou (The University of Hong Kong)
Yuan Cao (The University of Hong Kong)
Yuanzhi Li (Carnegie Mellon University)
Quanquan Gu (University of California, Los Angeles)

Abstract: Modern neural networks are often designed with an excessive number of parameters, which endow them with great expressive power. However, the prediction accuracy of these over-parameterized neural networks largely depends on the training algorithm employed. Specifically, recent empirical studies have shown that when compared to (stochastic) gradient descent, adaptive gradient methods can converge to a different solution with significantly worse test errors in various deep learning applications such as image classification. In this talk, I will present some recent results on the risk bounds of neural networks trained by adaptive gradient methods, and compare them with the corresponding risk bounds for vanilla gradient descent algorithm. Our analysis provides a theoretical explanation for the observed generalization gap between adaptive gradient methods and gradient descent. Moreover, I will also delve into the practical implications of our research in different applications, highlighting how the theoretical understanding can help improve the performance and efficiency of neural networks in certain tasks. Overall, this talk will provide valuable insights into the training of neural networks, and how different training algorithms can lead to vastly different outcomes.

Prediction of Cognitive Impairment Using Higher Order Item

Response Theory and Machine Learning Models

Authors: Lihua Yao (Northwestern University)
Emily Ho (Northwestern University)
Aaron Kaat (Northwestern University)
Richard Gershon (Northwestern University)

Abstract: Early detection of Cognitive impairment (CI) is very important for aged adults. The MyCog assessment uses two well-validated iPad-based measures from the NIH Toolbox for the Assessment of Neurological Behavior and Function Cognitive Battery (NIHTB-CB) that address two of the first domains to show CI: Picture Sequence Memory (PSM) which assesses episodic memory and Dimensional Change Card Sort (DCCS) measuring cognitive flexibility. 86 patients were administered MyCog assessments, and each was labeled of their CI diagnosis. Models were applied to predict patients CI status from the assessments.

Session 9: Data Science in Medical Application

Recent Advances and Applications of Digital Health in Stroke

Authors: Hongqiu Gu (Beijing Tiantan Hospital)

Abstract: Digital health has revolutionized healthcare delivery and holds immense potential for enhancing stroke care. Recent advancements in this field have led to the development of various tools and technologies aimed at improving stroke management. These include telemedicine, mobile health applications, wearable devices, and electronic health records. In this topic, I will provide a comprehensive overview of digital health, covering its conceptual foundations, the technologies it encompasses, and its applications in stroke research. Additionally, I will discuss areas of future research that need to be explored in order to further advance digital health's impact on stroke care.

AI for Medical Ultrasound: Imaging, Analysis and Generalization in the Wild

Authors: Xin Yang (Shenzhen University)

Abstract: Ultrasound is becoming a common imaging modality in clinics. However, limited by the image quality, field of view, large scales and user dependency of ultrasound scans, effective tools are highly desired in the clinic to improve the imaging, user consistency and diagnosis accuracy. Artificial intelligence sheds light on transforming the ultrasound examinations. This report will introduce our studies on intelligent ultrasound from four aspects: (1) online learning and adversarial learning for ultrasound reconstruction and simulation, (2) reinforcement learning and neural architecture search based 3D standard plane detection, (3) contrastive learning, self-supervised learning and weakly-supervised learning based high-dimensional ultrasound analysis, (4) style enhanced learning for robust ultrasound image segmentation.

Domain Adaptation for Data-Efficient Fundus Disease Recognition

Authors: Qijie Wei (Renmin University of China)

Abstract: Wide-field (WF) and ultra-wide-field (UWF) fundus imaging are playing an increasingly important role in fundus condition assessment and early diagnosis of retinal diseases. Compared to traditional color fundus photograhy (CFP), WF/UWF images have substantially larger field of view (FoV), making it possible for visualization of pathological alterations in peripheral retina. In this talk, we addresses the emerging task of recognizing multiple retinal diseases from WF/UWF fundus images. Due to the high cost of WF/UWF image collection and annotation, labeled WF/UWF images are in short supply. On the contrary, scale of labeled CFP data is much larger because of its long-lasting research. Thus, problem rises that how to exploit the existing large scale labeled CFP data for better diseases recognition in WF/UWF images. To resolve the problem, we propose Cross-domain Collaborative Learning that takes advantage of the mixup strategy succeed in unsupervised domain adaptation and employs self-attention to correct the intrinsic disparity between the FoV of CFP and WF/UWF images. Extensive experiments on multiple datasets covering both WF and UWF images show its advantages over a number

of competitive baselines

Adaptive Fusion of Radiomics and Deep Features for Lung Adenocarcinoma Subtype Recognition

Authors: Jing Zhou (School of Statistics, Renmin University of China)
Xiaotong Fu (School of Statistics, Renmin University of China)
Xirong Li (School of Information, Renmin University of China)
Ying Ji (Beijing Chao-Yang Hospital, Capital Medical University)

Abstract: The most common type of lung cancer, lung adenocarcinoma (LUAD), has been increasingly detected since the advent of low-dose computed tomography screening technology. In clinical practice, pre-invasive LUAD (Pre-IAs) should only require regular follow-up care, while invasive LUAD (IAs) should receive immediate treatment with appropriate lung cancer resection, based on the cancer subtype. However, prior research on diagnosing LUAD has mainly focused on classifying Pre-IAs/IAs, as techniques for distinguishing different subtypes of IAs have been lacking. In this study, we proposed a multi-head attentional feature fusion (MHA-FF) model for not only distinguishing IAs from Pre-IAs, but also for distinguishing the different subtypes of IAs. To predict the subtype of each nodule accurately, we leveraged both radiomics and deep features extracted from computed tomography images. Furthermore, those features were aggregated through an adaptive fusion module that can learn attention-based discriminative features. The utility of our proposed method is demonstrated here by means of real-world data collected from a multi-center cohort.

Session 10: Digital Economy, Digital Life and Statistical Measurement

Mechanism Analysis and Impact Measurement of the Platform Economy's Social Welfare Effect: An Empirical Study based on the

Questionnaire Survey of Merchants on MT Take-out Platform

Authors: Yuezhou Cai (Chinese Academy of Social Sciences)
Yuchen Gu (Fuzhou University of International Studies and Trade)

Abstract: As a typical model of the new digital economy and an important carrier of the integration of digital and real economy, the rapid rise of platform economy has played an important role in promoting economic circulation, making life convenient and safeguarding people's livelihood, and has become an infrastructure of systemic importance. However, with the rise of "oligarchic platform", the distribution of welfare has become more complex, posing new challenges to the government. Considering the important role of the platform, the goal of platform governance should be set as to enhance the welfare, promote fair distribution of welfare among various subjects, and ultimately achieve the healthy and orderly operation of the platform and the harmonious coexistence of all subjects. To this end, it is necessary to make scientific and accurate calculations on the social welfare impact of the platform economy in different stages.with the help of 9,132 valid questionnaires collected from an online questionnaire survey of more than 2 million registered merchants on MT take-out platform, and aggregating macro and micro data from other channels, we construct a Hicks compensation expenditure function and fit the supply and demand curve of the take-out platform to measure its welfare changes in 2020. We find that: 1) the take-out platform promotes merchants' business flow and profitability by enabling digital services. 2) Compared with traditional offline restaurants, online consumer welfare increases by about 27.79%, merchants' net and gross margins increase by about 1.5% and 6.5%, respectively, with the same transaction volume, and the platform and delivery riders also gain considerable benefits. 3) In 2020, MT platform's distribution of social welfare is close to the scenario of maximizing social welfare. The main reason is that the development of take-out platforms is still in the transitional period from " rapid development " to "stable maturity", and platforms are still competing fiercely for market shares through subsidies. This can indirectly prove the mechanism of welfare changes in different development stages of platform economy.

The Influence of Digital Finance Development on Bank Efficiency:

Evidence from China

Authors: Menggen Chen (School of Statistics, Beijing Normal University)
Qiao Zhang (Beijing Normal University)

Abstract: Digital finance has increased the accessibility of financial services and lowered their cost while also bringing a great challenge to the traditional mode of financial business. Based on the functional view of finance, a theoretical model including commercial banks, households, and enterprises is constructed to analyze the impact of digital finance on bank efficiency and explore its mechanisms from the liability and asset sides. In this paper, a three-dimensional framework including digital financial foundation, digital banking business and new financial services is constructed and a digital finance index is calculated to represent the development of digital finance at the city level. Then, using data on commercial banks from 2011 to 2020, this empirical study shows that the development of digital finance has strongly promoted the efficiency of China's commercial banks. These results also suggest that the influence of digital finance on the change in bank efficiency varies across different regions, scales, and types of ownership, among which high GDP regions, large-scale banks, and state-owned banks have a relatively strong effect on the improvement of efficiency. A further analysis of the mechanism shows that the development of digital finance affects the liability structure of banks, i.e., banks are usually inclined to have a smaller proportion of interbank liabilities as digital finance advances. At the same time, digital finance also changes the profitability of banks, which in turn affects their asset side. The underlying mechanism by which digital finance promotes bank efficiency is more closely connected to the strong optimization effect of digital finance on the liability side than to the weakening effect on the asset side.

Gig Economy in China: Measurement and Features

Authors: Jingping Li (School of Statistics, Renmin University of China)
Mingze Zhang (Renmin University of China)

Abstract: In the era of the digital economy, the gig economy is on the rise. However, official statistics of gig economy are basically in a blank state. Therefore, it is of great

importance to clarify the definition of gig workers and establish a statistical system of gig economy. In this research, we propose a double-dimension criterion to identify gig workers at first. Secondly, we devise the measurement of the employment and the value-added of gig economy after analyzing the existing statistical survey system, and design the corresponding questionnaire. At last, case studies are conducted to gain an understanding of the scale and characteristics of employment and value-added in China's gig economy. The main contributions of this study include: First, to clarify the connotation of gig workers, which provide a theoretical basis for gig economy statistics. Secondly, we propose a four-level gig economy statistical caliber, which lays a practical foundation for measuring the gig economy. Thirdly, we take a quantitative picture of China's gig economy, which provides a quantitative reference for relevant policy making.

Session 11： Financial Econometrics

Some Financial Applications of Dynamic Copulas

Authors: Ping Li (School of Statistics, Beihang Univeristy)
Jie Li (Beijing Technology and Business University)
Yingwei Han (China University of Geosciences Beijing)

Abstract: This talk will review the representative literatures on the application of dynamic copula models in finance, from the development of the dynamic copula model, several commonly-used dynamic copula models, and their applications in finance, such as financial risk management, pricing of credit derivatives, portfolio management, etc.

Disclosing and Cooling-Off: An Analysis of Insider Trading Rules

Authors: Jun Deng (University of International Business and Economics)
Huifeng Pan (University of International Business and Economics)
Hongjun Yan (DePaul University)
Liyan Yang (University of Toronto)

Abstract: We analyze two insider-trading regulations recently introduced by the U.S. Securities and Exchange Commission: advance disclosure and "cooling-off period." The former requires an insider to disclose trading plans at adoption, while the latter mandates a delay period before execution. Disclosure increases price efficiency but has mixed welfare implications. If the insider has large liquidity needs, in contrast to the conventional wisdom from "sunshine trading," disclosure can even reduce the welfare of all investors. A longer cooling-off period increases outside investors' welfare but decreases price efficiency. Its implication for the insider's welfare depends on whether the disclosure policy is already in place.

Limiting Behaviour of Realized Covariation in the Presence of Price Staleness

Author: Zhi Liu (Department of Mathematics, University of Macau)
Haibin Zhu (Jinan University)

Abstract: Considering the presence of systematic price staleness, we study the problem of estimating the integrated covariation of two semi-martingales. We propose a consistent estimator of the integrated covariation and establish a unified limiting theory, which includes several existing results as special cases. Our results demonstrate that the idiosyncratic price stalenesses appear in the limit of the standard realized covariation, but the systematic price staleness has only an impact on the second-order limiting behavior. Moreover, we find that price staleness makes the standard realized covariation closer to zero than that without price staleness. Hence it explains the well-known Epps effect appropriately. We conduct extensive Monte Carlo studies to assess the finite sample performance of the proposed theory, and some empirical applications to real high-frequency data are considered to illustrate our theory.

A Latent Space Model for Bipartite Networks with Applications in Interlocking Directorates in Chinese Companies

Authors: Yan Zhang (Xiamen University)
Feifei Wang (Renmin University)

Kuangnan Fang (Xiamen University)
Rui Pan (Central University of Finance and Economics)
Hansheng Wang (Peking University)

Abstract: Bipartite networks containing two types of nodes are commonly encountered in practice. To analyze bipartite networks, the latent space model is popularly used. With the increase in data availability, nodes in networks are often observed with nodal attributes, which provide fertile information for understanding the network structure. However, existing latent space models for dynamic bipartite networks often ignore nodal variables. To address this problem, we propose a latent space model for bipartite networks by incorporating information from both the covariates and the network structure. To reflect the evolution pattern of the network structure, we introduce two parameters representing the persistence effect. To estimate the model, we propose a computationally efficient algorithm using projected gradient descent. The theoretical properties are also established and validated through comprehensive simulation studies. Last, we analyze the dynamic bipartite network for the Chinese interlocking directorates from 2010 to 2020 using our proposed model.

Session 12: Frontiers in Financial Statistics

Individual-centered Partial Information in Social Networks

Authors: Xiao Han（University of Science and Technology）
Rachel Wang（University of Sydney）
Xin Tong（University of Southern California）

Authors： In statistical network analysis, we often assume either the full network is available or multiple subgraphs can be sampled to estimate various global properties of the network. However, in a real social network, people frequently make decisions based on their local view of the network alone. Here, we consider a partial information framework that characterizes the local network centered at a given individual by path length L and gives rise to a partial adjacency matrix. Under L=2, we focus on the problem of (global) community detection using the popular stochastic block model (SBM) and its

degree-corrected variant (DCSBM). We derive general properties of the eigenvalues and eigenvectors from the signal term of the partial adjacency matrix and propose new spectral-based community detection algorithms that achieve consistency under appropriate conditions. Our analysis also allows us to propose a new centrality measure that assesses the importance of an individual's partial information in determining global community structure. Using simulated and real networks, we demonstrate the performance of our algorithms and compare our centrality measure with other popular alternatives to show it captures unique nodal information. Our results illustrate that the partial information framework enables us to compare the viewpoints of different individuals regarding the global structure.

High-Dimensional Covariance Matrices Under Dynamic Volatility Models: Asymptotics and Shrinkage Estimation

Authors: Yi Ding (University of Macau)
Xinghua Zheng (HKUST)

Abstract: We study the estimation of the high-dimensional covariance matrix and its eigenvalues under dynamic volatility models. Data under such models have nonlinear dependency both cross-sectionally and temporally. We first investigate the empirical spectral distribution (ESD) of the sample covariance matrix under scalar BEKK models and establish conditions under which the limiting spectral distribution (LSD) is either the same as or different from the i.i.d. case. We then propose a time-variation adjusted (TV-adj) sample co- variance matrix and prove that its LSD follows the same Marcenko-Pastur law as the i.i.d. case. Based on the asymptotics of the TV-adj sample covariance matrix, we develop a consistent population spectrum estimator and an asymptotically optimal nonlinear shrinkage estimator of the unconditional covariance matrix. Based on joint work with Yi Ding

Estimating Efficient Frontier with All Risky Assets

Authors: Leheng Chen (HKUST)
Yingying Li (HKUST)
Xinghua Zheng (HKUST)

Abstract: We propose a method to estimate the efficient frontier with all risky assets under a high-dimensional setting. The method utilizes linear constrained LASSO based on an equivalent constrained regression representation of the mean-variance optimization. Under a mild sparsity assumption, we show that our estimator asymptotically achieves mean-variance efficiency. Extensive simulation and empirical studies are conducted to examine the performance of our proposed estimator. Based on joint work with Leheng Chen and Xinghua Zheng.

Currency Exchange Rate Predictability: the New Power of Bitcoin Prices

Authors: Wenjun Feng (Beijing Jiaotong University)
Zhengjun Zhang (University of Chinese Academy of Sciences)

Abstract: We show that Bitcoin prices have surprisingly predictive power for nominal currency exchange rates, both in-sample and out-of-sample. The predictability follows from the fact that Bitcoin prices are forward-looking: Bitcoin efficiently incorporates expectations of currency exchange rates and their drivers, as exchange rates serve as a fundamental of Bitcoin. We examine the Bitcoin-based exchange rate prediction model in the autoregressive distributed lag (ADL) specification and the error correction specification. Forecasts based on both specifications outperform different benchmarks for some of the exchange rates. The outperformance is most pronounced at the daily horizon using the ADL model. Bitcoin-based forex trading strategies generate Sharpe ratio gains relative to the US risk-free rate and the carry trade. Bitcoin returns incorporate extra knowledge of future interest rate differentials after controlling for lagged exchange rate movements. Our result is inspiring for currency market participants, given the well-documented difficulty in exchange rate prediction.

Session 13: Gradient Descent and Its Statistical Theory

Joint Feature Screening Incorporating Network Structure Among Responses

Authors: Xu Zhang (Huanan Normal University)
Xiangeng Fang (University of Michigan)
Sheng Xu (BeiGene)
Xuening Zhu (Fudan University)
Catherine Liu (Applied Mathematics, The Hong Kong Polytechnic University)

Abstract: The contemporary data acquisition technology makes it common to have wealth of ultrahigh dimensional data with auxiliary information inducing dependent structure among observations. When the dependency is decided by a network structure, it poses great challenges to developing new methodology and algorithms tackling models incorporating dependent responses under the ultrahigh-dimensional data setting. Feature screening is an essential step before elaborating delicate statistical analysis, whereas dependent response observations hinders the likelihood method to be employed, which consequently incurs loss of information among predictors. However, joint feature screening is known superior to independent screening based on a specific marginal utility although it provokes another challenge in relieving the computing burden. We are driven to propose a generalized autoregressive linear model incorporating structural information among responses and develop a highly efficient joint feature screening procedure. We address the computing problem by raising the pseudo-likelihood procedure making use of the conditional distribution relationship among responses. We present a feasible proximal gradient descent iterative computing algorithm that is adaptive from the hard-thresholding spirit and implemented for solving the maximum pseudo-likelihood optimization. Theoretical study demonstrates the effectiveness of the proposed method. The finite sample performance of the proposed method is assessed by simulation studies and illustrated by an empirical analysis of a dataset from the Chinese stock market.

Statistical Analysis of Fixed Mini-Batch Gradient Descent Estimator

Authors: Haobo Qi (School of Statistics, Beijing Normal University)

Feifei Wang (Renmin University of China)
Hansheng Wang (Peking University)

Abstract: We study here a fixed mini-batch gradient decent (FMGD) algorithm to solve optimization problems with massive datasets. In FMGD, the whole sample is split into multiple non-overlapping partitions. Once the partitions are formed, they are then fixed throughout the rest of the algorithm. For convenience, we refer to the fixed partitions as fixed mini-batches. Then for each computation iteration, the gradients are sequentially calculated on each fixed mini-batch. Because the size of fixed mini-batches is typically much smaller than the whole sample size, it can be easily computed. This leads to much reduced computation cost for each computational iteration. It makes FMGD computationally efficient and practically more feasible. To demonstrate the theoretical properties of FMGD, we start with a linear regression model with a constant learning rate. We study its numerical convergence and statistical efficiency properties. We find that sufficiently small learning rates are necessarily required for both numerical convergence and statistical efficiency. Nevertheless, an extremely small learning rate might lead to painfully slow numerical convergence. To solve the problem, a diminishing learning rate scheduling strategy can be used. This leads to the FMGD estimator with faster numerical convergence and better statistical efficiency. Finally, the FMGD algorithms with random shuffling and a general loss function are also studied.

Network Gradient Descent Algorithm for Decentralized Federated Learning

Authors: Shuyuan Wu (Shanghai University of Finance and Economics)
Danyang Huang (Renmin University of China)
Hansheng Wang (Peking University)

Abstract: We study a fully decentralized federated learning algorithm, which is a novel gradient descent algorithm executed on a communication-based network. For convenience, we refer to it as a network gradient descent (NGD) method. In the NGD method, only statistics (e.g., parameter estimates) need to be communicated, minimizing the risk of privacy. Meanwhile, different clients communicate with each other directly according to a carefully designed network structure without a central master. This greatly

enhances the reliability of the entire algorithm. Those nice properties inspire us to carefully study the NGD method both theoretically and numerically. Theoretically, we start with a classical linear regression model. We find that both the learning rate and the network structure play significant roles in determining the NGD estimator's statistical efficiency. The resulting NGD estimator can be statistically as efficient as the global estimator if the learning rate is sufficiently small and the network structure is well balanced, even if the data are distributed heterogeneously. Those interesting findings are then extended to general models and loss functions. Extensive numerical studies are presented to corroborate our theoretical findings. Classical deep learning models are also presented for illustration purposes.

An Asymptotic Analysis of Random Partition Based Minibatch Momentum Methods for Linear Regression Models

Authors: Yuan Gao (Guanghua School of Management, Peking University)
Xuening Zhu (Fudan University)
Haobo Qi (Peking University)
Guodong Li (University of Hong Kong)
Riquan Zhang (Shanghai University of International Business and Economics)
Hansheng Wang (Peking University)

Abstract: Momentum methods have been shown to accelerate the convergence of the standard gradient descent algorithm in practice and theory. In particular, the random partition based minibatch gradient descent methods with momentum (MGDM) are widely used to solve large-scale optimization problems with massive datasets. Despite the great popularity of the MGDM methods in practice, their theoretical properties are still underexplored. To this end, we investigate the theoretical properties of MGDM methods based on the linear regression models. We first study the numerical convergence properties of the MGDM algorithm and derive the conditions for faster numerical convergence rate. In addition, we explore the relationship between the statistical

properties of the resulting MGDM estimator and the tuning parameters. Based on these theoretical findings, we give the conditions for the resulting estimator to achieve the optimal statistical efficiency. Finally, extensive numerical experiments are conducted to verify our theoretical results.

Session 14: High Dimensional Methods for Estimation and Inference

High Dimensional Clustering via Latent Semiparametric Mixture Models

Authors: Lyuou Zhang (Shanghai University of Finance and Economics)
Lulu Wang (Gileads)
Wen Zhou (Colorado State University)
Boxiang Wang (Univerisity of Iowa)
Hui Zou (University of Minnesota)

Abstract: Cluster analysis is a fundamental task in machine learning. Several clustering algorithms have been extended to handle high-dimensional data by incorporating a sparsity constraint in the estimation of a mixture of Gaussian models. Though it makes

some neat theoretical analysis possible, this type of approach is arguably restrictive for many applications. In this work we propose a novel latent variable transformation mixture model for clustering in which we assume that after some unknown monotone transformations the data follows a mixture of Gaussians. Under the assumption that the optimal clustering admits a sparsity structure, we develop a new clustering algorithm named CESME for high-dimensional clustering. The use of unspecified transformation makes the model far more flexible than the classical mixture of Gaussians. On the other hand, the transformation also brings quite a few technical challenges to the model estimation as well as the theoretical analysis of CESME. We offer a comprehensive analysis of CESME including identifiability, initialization, algorithmic convergence, and statistical guarantees on clustering. In addition, the convergence analysis has revealed an interesting algorithmic phase transition for CESME, which has also been noted for the EM algorithm in literature. Leveraging such a transition, a data-adaptive procedure is developed and substantially improves the computational efficiency of CESME. Extensive numerical study and real data analysis show that CESME outperforms the existing high-dimensional clustering algorithms including CHIME, sparse spectral clustering, sparse K-means, sparse convex clustering, and IF-PCA.

Robust Statistical Inference for Large-dimensional Matrix-valued Time Series via Iterative Huber Regression

Authors: Yong He(Shandong University)
Xinbing Kong(Nanjing Audit Iniversity)
Dong Liu(Shanghai University of Finance and Economics)
Ran Zhao(Shandong University)

Abstract: Matrix factor model is drawing growing attention for simultaneous two-way dimension reduction of well-structured matrix-valued observations. This paper focuses on robust statistical inference for matrix factor model in the ``diverging dimension" regime. We propose an iterative Huber regression algorithm to estimate the factor loadings and factor scores. Theoretically, given the true dimensions of the factor matrices as a priori, we derive the convergence rates of the robust estimators for loadings, factors and

common components under finite second moment assumption of the idiosyncratic errors. In addition, the asymptotic distributions of the estimators are also derived under mild conditions. We also propose both a rank minimization and an eigenvalue-ratio method to estimate the pair of factor numbers robustly, which are proven to be consistent. Numerical studies confirm the iterative Huber regression algorithm is a practical and reliable approach for the estimation of matrix factor model, especially under the cases with heavy-tailed idiosyncratic errors . We also illustrate the practical usefulness of the proposed methods by two real datasets on financial portfolios and multinational macroeconomic indices of China. An R package ``HDMFA'' implementing the related robust matrix factor analysis methods in the literature is available on CRAN.

Multi-dimensional Domain Generalization with Low-rank Structures

Authors: Sai Li (Renmin University of China)
Linjun Zhang (Rutgers Univerisity)

Abstract: In health-related studies, certain sub-populations may be underrepresented, which presents a challenge for researchers seeking to understand the characteristics of these groups. In this work, we tackle this challenge in linear models by organizing the regression vectors of all the sub-populations into a tensor. We formulate the domain generalization problem as a tensor completion task, allowing us to learn about sub-populations with limited or no available data. Unlike previous studies in tensor completion, our model accounts for complex missing patterns and correlation structures. Our proposed method is supported by theoretical guarantees and numerical studies demonstrating its efficiency.

Adaptive False Discovery Rate Control with Privacy Guarantee

Authors: Zhanrui Cai (University of Hong Kong)

Abstract: Differentially private multiple testing procedures can protect the information of individual hypothesis tests while guaranteeing a small fraction of false discoveries. In this paper, we propose a differentially private adaptive FDR control method that can control the classic FDR metric exactly at a user-specified level α with privacy guarantee, which is a non-trivial improvement compared to the DP-BH method. Our analysis is based on two key insights: 1) a novel p-value transformation that preserves both privacy and the mirror conservative property, and 2) a mirror peeling algorithm that allows the construction of the filtration and application of the optimal stopping technique. Numerical studies demonstrate that the proposed DP-AdaPT performs better compared to the existing differentially private FDR control methods. Compared to the original AdaPT, it only incurs a small accuracy loss but also significantly reduces the computation cost.

Session 15: High-dimensional and Functional Learning

Matrix Estimation via Singular Value Shrinkage

Authors: Takeru Matsuda（University of Tokyo & RIKEN Center for Brain Science）

Abstract: In the estimation of a normal mean vector under the quadratic loss, the maximum likelihood estimator (MLE) is inadmissible and dominated by shrinkage estimators (e.g. James–Stein estimator) when the dimension is greater than or equal to three (Stein’s paradox). In particular, generalized Bayes estimators with respect to superharmonic priors (e.g. Stein’s prior) are minimax and dominate MLE. Note that a function is said to be superharmonic if its average value on a supersphere is always not greater than its value at the center.

In this talk, I will introduce recent studies on shrinkage estimation of matrices. First, we develop a superharmonic prior for matrices that shrinks singular values, which can be viewed as a natural generalization of Stein’s prior. This prior is motivated from the Efron–Morris estimator, which is an extension of the James–Stein estimator to matrices. The generalized Bayes estimator with respect to this prior is minimax and dominates MLE under the Frobenius loss. In particular, since it shrinks to the space of low-rank matrices, it attains large risk reduction when the unknown matrix is close to low-rank (e.g. reduced-rank regression). Next, we construct a theory of shrinkage estimation under the “matrix quadratic loss”, which is a matrix-valued loss function suitable for matrix

estimation. A notion of “matrix superharmonicity” for matrix-variate functions is introduced and the generalized Bayes estimator with respect to a matrix superharmonic prior is shown to be minimax under the matrix quadratic loss. The matrix-variate improper t-priors are matrix superharmonic and this class includes the above generalization of Stein’s prior. Applications include matrix completion and nonparametric estimation.

Error Analysis on Pre-taining and Fine-tuning in Based on Deep Sufficient and Invariant Representation Learning

Authors: Yuling Jiao（Wuhan Univeristy）

Abstract: In this talk, we will discuss an error analysis of pre-training and fine-tuning (PTFT) from the perspective of representation learning. Firstly, during pre-training, deep neural networks are used to learn a sufficiently invariant representation by utilizing a vast amount of data from multiple domains. Secondly, during the downstream task, we learn a simple task-dependent function with a small sample size. We propose a statistical model for PTFT and provide an end-to-end error bound to demonstrate the effectiveness of PTFT.

Robust Regularized Covariance Matrices

Authors: Mengxi Yi (Beijing Normal University)
David E. Tyler（Rutgers University ）
Klaus Nordhausen（University of Jyväskylä ）

Abstract: We introduce a class of regularized M-estimators of multivariate scatter and show, analogous to the popular spatial sign covariance matrix (SSCM), that they possess high breakdown points and bounded influence function. We also show that the SSCM can be viewed as an extreme member of this class. Unlike the SSCM, this class of estimators takes into account the shape of the contours of the data cloud when down-weighing observations. We also propose a median based cross validation criterion for selecting the tuning parameter for this class of regularized M-estimators. This cross validation criterion helps assure the resulting tuned scatter estimator is a good fit to the data as well as having a high breakdown point. A motivation for this new median based criterion is that when it is optimized over all possible scatter parameters, rather than only over the tuned candidates, it results in a new high

breakdown point affine equivariant multivariate scatter statistic.

Algriculture: Harnessing Statistical & Deep Learning for Precision Monitoring in Hydroponic Farms

Authors: Jishnu Mukhoti
Mohammed Alezzabi
Glen Wright Colopy (Wildfell Software LLC, Alesca Life Technologies)

Abstract: Indoor hydroponic farming offers a sustainable solution for the future of agriculture. Advantages include water and fertilizer efficiency, year-round productivity, reduced reliance on pesticides, and proximity to urban centers (to minimize transportation costs & spoilage waste). Despite these benefits, tasks like operational record keeping and plant health monitoring remain expensive and labor intensive. Our work combines (i) the predictive power of deep learning computer vision models with (ii) the explanatory power of statistical learning to provide useful automation in a real-world industrial setting.

Our plant image data comprises multispectral images to capture plant health signals across a wide range of wavelengths, including those beyond the visible spectrum. This enables the identification of early signs of plant health deterioration, for example, a loss in photosynthetic activity. These signals are often invisible to the human eye but allow for timely interventions before the onset of significant damage.

In our first use case, we fine-tuned a Vision Transformer (ViT), a state-of-the-art computer vision model introduced by Google Brain, to classify crop types at different locations, a task that would otherwise be manual. This allowed us to quickly and inexpensively train a highly accurate classification model tailored to our indoor growing environment (in which lighting comes solely from LED grow lights). This automated system drastically reduces the amount of manual labor & error rate of human record keeping of several farm operations.

In our second use case, we used statistical learning models to translate multispectral images into interpretable metrics of plant health. These unsupervised models were capable of picking up diurnal fluctuations of plant life and provides an early warning system for illnesses that are otherwise challenging to detect.

Combining these two approaches to machine learning allows us to exploit their respective strengths where each is most appropriate: deep learning for supervised

prediction of industrial processes & statistical learning for unsupervised explanation of scientific processes.

Session 16：Image and Text Data Analysis with Application

A Semiparametric Gaussian Mixture Model for Chest CT Based 3D Blood Vessel Reconstruction

Authors: Qianhan Zeng (Guanghua School of Management, Peking University)
Jing Zhou (Renmin University of China)
Ke Xu (University of International Business and Economics)
Xiao Wang (Qingdao University)

Abstract: Traditional 3D reconstruction based on computed tomography data relies on manual operations by experienced surgeons. We develop a novel semiparametric Gaussian mixture model for 3D blood vessel reconstruction. We propose a semiparametric Gaussian mixture model. A kernel-based expectation-maximization algorithm is developed to estimate the model, and a supporting asymptotic theory is established. A novel regression method is proposed for bandwidth selection, which outperforms the cross-validation-based method. In application, the 3D structures of blood

vessels are reconstructed automatically.

An Ensemble Deep Learning Model for Risk Stratification of Invasive Lung Adenocarcinoma Using Thin-Slice CT

Authors: Jing Zhou(Renmin University of China)

Abstract: Lung cancer has always been among the most frequently diagnosed cancers threatening people's health worldwide. Recently, with the rapid development of artificial intelligence and popularization of low-dose computed tomography (LDCT) in lung cancer screening, a number of studies focus on predicting benign and malignant lung tumors or diagnosing between pre-invasive and invasive lung tumors by CT images with advanced deep learning algorithm. However, seldom studies predict the invasive grades of lung adenocarcinoma. This is an important task since it will be helpful to design a more reasonable surgical mode (lobectomy or sublobar resection) before operation. We propose an ensemble multi-view 3D convolution neutral network (EMV-3D-CNN) model to comprehensively study the risk grades of lung adenocarcinoma. Our model achieves a state-of-art performance (91.3% AUC for diagnosis between benign and malignant, 92.5% AUC for diagnosis between pre-invasive and invasive, and 77.6% accuracy for diagnosis among risks of Grades 1,2,3) on 1,075 lung nodules (covering 627 CT trials) collected from three medical centers. We conducted six reader studies to evaluate the model performance. The results suggest that the EMV-3D-CNN model achieved equivalent or slightly higher performance compared with the doctors. Finally, for user-friendly access, the proposed model is also implemented as a web-based system (https://seeyourlung.com.cn). By uploading the full original CT images of DICOM format, our algorithm can give the risk grades of pulmonary nodules by specifying the center location of the target lung nodule.

"This Crime is Not That Crime" —— Classification and Evaluation of Four Common Crimes

Authors： Ke Xu (University of International Business and Economics)

Hangyu Liu (Peking University)
Fang Wang (Shandong University)
Hansheng Wang (Peking University)

Abstract: As the basis of criminal penalty, criminal conviction, integral to the protection of fundamental rights and freedom of people, constitutes the basis and the core issue of criminal trials. Based on the data published on China Judgments Online, we proposed two types of classification models to apply the data of four common crimes from China Judgments Online and expounded their applications in identifying “abnormal cases”, defined as wrongly sentenced cases in this paper. The two types of classification models we proposed are a two-stage model and two deep learning models. To construct the two-stage model, we first used three keyword-extraction models to extract the keywords and vectorize all the keywords, then used five classification models to build the two-stage model. For the deep learning models, we applied two different deep neural network models in the data to build the classifier. We then applied these two types of classification models to discover “abnormal cases” in two steps. In the first step, we applied the two-stage model to extract the “important words” that will significantly improve the probability of the two-stage model to classify cases into crimes of intentional injury. In the second step, we constructed a validation data set of cases whose verdicts are changed in the second instance rulings to test the “important words” extracted in first step and the ability of the two-stage model and the two deep learning models to discover “abnormal cases”. The results of this exercise show that: 1) “important words” extracted in the first step are often associated with “abnormal cases”; 2) these two types of classification models can effectively discover “abnormal cases”, but compared with the two deep learning models, the two-stage model (aka. TF-IDF & ANN, the combination of a keyword extraction model and a classic machine-learning model) is more capable of discovering “abnormal cases”.

A Geometrical Model with Stochastic Error for Abnormal Motion Detection of Portal Crane Bucket Grab

Authors: Baichen Yu (Guanghua School of Management, Peking University)

Xiao Wang (School of Statistics, Qingdao University)
Hansheng Wang (Guanghua School of Management, Peking University)

Abstract: Sea transportation is among the most important modes of transportation in the world, accounting for more than 80% of the volume of international trade in goods. Although there are multiple components that may impact sea transportation, sea port infrastructure, especially portal cranes plays a crucial role. Consequently, the safe and effective operation of portal cranes, including automatically monitoring the motions of a bucket grab, becomes a critically important issue of concern. To potentially address this issue, we have developed a novel approach to estimate the swing angle of the portal crane using video images generated by a surveillance camera installed on the fly-jib head as the input. Next, a spatial geometric model with stochastic error is developed. The model describes the geometric relationship between the signals observed on the image plane and the actual bucket grab motion. A statistical model is used to describe the stochastic motion behavior of the bucket grab along with a novel iterative algorithm to estimate the unknown parameters. This enables us to estimate the swing angle in a timely manner and generate alarm signal immediately. Numerical studies based on both simulated and real datasets are presented. We provide here a computer-vision based solution for automatic detection of abnormal motion for portal cranes. Our method can be used to safely guarantee the day-to-day operations of portal cranes for transferring freight between the port and cargo ships.

Session 17： Machine learning for Functional Data and Causal Inference

Distribution Estimation of Contaminated Data via DNN-based MoM-GANs

Authors: Huiming Zhang (Beihang University)
Fang Xie (BNU-HKBU United International College)
Lihu Xu (University of Macau)
Qiuran Yao (University of Macau)

Abstract: The traditional adversarial nets, for example, the generative adversarial network (GAN), are sensitive to contaminated data. In this talk, we develop a robust and deep neural network (DNN) method to estimate the learning distribution based on the adversarial nets and median-of-mean (MoM) approach, and it is called the MoM-GAN method. Theoretically, we obtain a non-asymptotic error bound for the DNN-based Wasserstein-1 MoM-GANs estimator measured by integral probability metrics with the H\"{o}lder function class. The derived finite-sample high-probability error-bounds concern the outlier proportion and the fraction of sane blocks. We give an algorithm for our proposed method and implement it through two real applications, which show that our proposed method outperforms Wasserstein GAN for contaminated data.

Design-based Theory for Lasso Adjustment in Randomized Block Experiments and Rerandomized Experiments

Authors: Ke Zhu (Tsinghua University)
Hanzhong Liu (Tsinghua University)
Yuehan Yang (Central University of Finance and Economics)

Abstract: Blocking, a special case of rerandomization, is routinely implemented in the design stage of randomized experiments to balance the baseline covariates. Regression adjustment is highly encouraged in the analysis stage to adjust for the remaining covariate imbalances. This study proposes a regression adjustment method based on Lasso to efficiently estimate the average treatment effect in randomized block experiments with high-dimensional covariates.

Comparison of the Slops in Functional Regression under Arbitrary

Transformations

Authors: Pratim Guha Niyogi (Johns Hopkins University, USA)
Subhra Sankar Dhar (IIT Kanpur, India)

Abstract: In scalar on function regression for two groups, we study whether the slope function of one group is the same as the slope function of another group up to an arbitrary transformation or not. In order to test it, we formulate a test statistic and derive the asymptotic distribution of the proposed test statistic. Moreover, extensive simulation study is carried out to demonstrate the performance of the proposed methodology.

Towards Trustworthy Explanation: On Causal Rationalization

Authors: Hengrui Cai (University of California Irvine)

Abstract: With recent advances in natural language processing, rationalization becomes an essential self-explaining diagram to disentangle the black box by selecting a subset of input texts to account for the major variation in prediction. Yet, existing association-based approaches on rationalization cannot identify true rationales when two or more snippets are highly inter-correlated and thus provide a similar contribution to prediction accuracy, so-called spuriousness. To address this limitation, we novelly leverage two causal desiderata, non-spuriousness and efficiency, into rationalization from the causal inference perspective. We formally define a series of probabilities of causation based on a newly proposed structural causal model of rationalization, with its theoretical identification established as the main component of learning necessary and sufficient rationales. The superior performance of the proposed causal rationalization is demonstrated on real-world review and medical datasets with extensive experiments compared to state-of-the-art methods.

Session 18： Mathematical Statistics under The Big Data Era

Estimation of Linear Functionals in High Dimensional Linear

Models: From Sparsity to Non-sparsity

Authors: Junlong Zhao (Beijing Normal University)

Abstract: High dimensional linear models are commonly used in practice. In many applications, one is interested in linear transformations $\bbeta^\top x$ of regression coefficients $\bbeta\in \mR^p$, where x is a specific point and is not required to be identically distributed as the training data. One common approach is the plug-in technique which first estimates $\bbeta$, then plugs the estimator in the linear transformation for prediction. Despite its popularity, estimation of $\bbeta$ can be difficult for high dimensional problems. Commonly used assumptions in the literature inc

Semiparametric Efficient G-estimation with Invalid Instrumental Variables

Authors: Baoluo Sun(NUS)
Zhonghua Liu (Columbia University)
Eric Tchetgen Tchetgen(UPenn)

Abstract: The instrumental variable method is widely used in the health and social sciences for identification and estimation of causal effects in the presence of potential unmeasured confounding. In order to improve efficiency, multiple instruments are routinely used, leading to concerns about bias due to possible violation of the instrumental variable assumptions. To address this concern, we introduce a new class of G-estimators that are guaranteed to remain consistent and asymptotically normal for the causal effect of interest provided that a set of at least γ out of K candidate instruments are valid, for $\gamma \leqslant$K set by the analyst ex ante, without necessarily knowing the identity of the valid and invalid instruments. We provide formal semiparametric efficiency theory supporting our results.

Both simulation studies and applications to the UK Biobank data demonstrate the superior empirical performance of our estimators compared to competing methods.

Design-based Theory for Cluster Rerandomization

Authors: Xin Lu (Tsinghua University)
Tianle Liu (Harvard University)
Hanzhong Liu (Industrial Engineering, Tsinghua University)
Peng Ding (University of California, Berkeley)

Abstract: Complete randomization balances covariates on average, but covariate imbalance often exists in finite samples. Rerandomization can ensure covariate balance in the realized experiment by discarding the undesired treatment assignments. Many field experiments in public health and social sciences assign the treatment at the cluster level due to logistical constraints or policy considerations. Moreover, they are frequently combined with rerandomization in the design stage. We define cluster rerandomization as a cluster-randomized experiment compounded with rerandomization to balance covariates at the individual or cluster level. Existing asymptotic theory can only deal with rerandomization with treatments assigned at the individual level, leaving that for cluster rerandomization an open problem. To fill the gap, we provide a design-based theory for cluster rerandomization. Moreover, we compare two cluster rerandomization schemes that use prior information on the importance of the covariates: one based on the weighted Euclidean distance and the other based on the Mahalanobis distance with tiers of covariates. We demonstrate that the former dominates the latter with optimal weights and orthogonalized covariates. Last but not least, we discuss the role of covariate adjustment in the analysis stage and recommend covariate-adjusted procedures that can be conveniently implemented by least squares with the associated robust standard errors.

Order-of-addition Experiments

Authors: Chunyan Wang (Renmin University of China)
Dennis K. J. Lin(Purdue University)

Abstract: In an order-of-addition (OofA) experiment, the response is a function of the addition order of components. The key objective of the OofA experiments is to find the optimal order of addition. The most popularly used model for OofA experiments is perhaps the pairwise ordering (PWO) model, which assumes that the response can be fully accounted by the pairwise ordering of components. Recently, Mee (2020) extended the PWO model by adding the interactions of PWO factors, to account for variations caused by the ordering of sets of three or more components, where the interaction term is defined by the multiplication of two PWO factors. This paper introduces a novel class of conditional PWO effect to study the interaction effect between PWO factors. The advantages of the proposed interaction terms are studied. Based on these conditional effects, a new model is proposed. The optimal order of addition can be straightforwardly obtained via the proposed model.

Session 19: MCMC and Clustering

Ranking Inferences Based on the Top Choice of Multiway Comparisons

Authors: Jianqing Fan (Princeton University)
Zhipeng Lou (Princeton University)
Weichen Wang (The University of Hong Kong)
Mengxin Yu (Princeton University)

Abstract: 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.

Snake Algorithm: A Rejection-Free Sampler for Binary Matrices with Fixed Margins

Authors: Zipei Nie (Lagrange Mathematics and Computing Research Center)
Guanyang Wang (Rutgers University)
Peng Zhang (Rutgers University)

Abstract: This talk presents a new algorithm for sampling binary matrices with fixed row and column sums, a problem with applications in network, ecology, differential privacy, and theoretical computer science. We introduce the 'Snake' algorithm, unlike existing methods, finds a random, swappable loop at every step. Our algorithm features a rejection-free design and scales better with the matrix's size, proving particularly efficient for high-dimensional and sparse matrices common in practical applications.

Proximal MCMC for Bayesian Inference of Constrained and

Regularized Estimation

Authors: Xinkai Zhou (Johns Hopkins University)
Eric C. Chi (Rice University)
Qiang Heng (North Carolina State University)
Hua Zhou (UCLA)

Abstract: This talk advocates proximal Markov Chain Monte Carlo (ProxMCMC) as a flexible and general Bayesian inference framework for constrained or regularized estimation. Originally introduced in the Bayesian imaging literature, ProxMCMC employs the Moreau-Yosida envelope for a smooth approximation of the total-variation regularization term, fixes nuisance and regularization parameters as constants, and relies on the Langevin algorithm for the posterior sampling. We extend ProxMCMC to the full Bayesian framework with modeling and data-adaptive estimation of all parameters including the regularization strength parameter. More efficient sampling algorithms such as the Hamiltonian Monte Carlo are employed to scale ProxMCMC to high-dimensional problems. Analogous to the proximal algorithms in optimization, ProxMCMC offers a versatile and modularized procedure for the inference of constrained and non-smooth problems. The power of ProxMCMC is illustrated on various statistical estimation and machine learning tasks. The inference in these problems is traditionally considered difficult from both frequentist and Bayesian perspectives.

Bayesian biclustering and its application in education data analysis

Authors: Weining Shen (UC Irvine)

Abstract: We propose a novel nonparametric Bayesian IRT model that estimates clusters at the question level, while simultaneously allowing for heterogeneity at the examinee level under each question cluster, characterized by the mixture of Binomial distributions. The main contribution of this work is threefold. First, we present our new model and demonstrate that it is identifiable under a set of conditions. Second, we show that our model can correctly identify question-level clusters asymptotically, and the parameters of interest that measure the proficiency of examinees in solving certain questions can be estimated at a $\sqrt{n}$ rate (up to a $\log$ term). Third, we present a tractable sampling algorithm to obtain valid posterior samples from our proposed model. We evaluate our model via a series of simulations as well as apply it to an English proficiency assessment

data set. This data analysis example nicely illustrates how our model can be used by test makers to distinguish different types of students and aid in the design of future tests.

Session 20: Modern Statistical Learning and Its Applications to Image Data Analysis

Wasserstein Generative Regeression

Authors: Tong Wang(The Chinese University of Hong Kong)
Shanshan Song(The Chinese University of Hong Kong)
Guohao Shen(The Hong Kong Polytechnic University)
Yuanyuan Lin(The Chinese University of Hong Kong)
Jian Huang (The Hong Kong Polytechnic University)

Abstract: In this paper, we propose a new and unified approach for nonparametric regression and conditional distribution learning. Our approach simultaneously estimates a regression function and a conditional generator using a generative learning framework, where a conditional generator is a function that can generate samples from a conditional distribution. The main idea is to estimate a conditional generator that satisfies the constraint that it produces a good regression function estimator. We use deep neural networks to model the conditional generator. Our approach can handle problems with multivariate outcomes and covariates, and can be used to construct prediction intervals. We provide theoretical guarantees by deriving non-asymptotic error bounds and the distributional consistency of our approach under suitable assumptions. We also perform numerical experiments with simulated and real data to demonstrate the effectiveness and superiority of our approach over some existing approaches in various scenarios.

Deep Kronecker Network

Authors: Long Feng (University of Hong Kong)
Guang Yang (City University of Hong Kong)

Abstract: We propose Deep Kronecker Network (DKN), a novel framework designed for analyzing medical imaging data, such as MRI, fMRI, CT, etc. Medical imaging data is different from general images in at least two aspects: i) sample size is usually much more limited, ii) model interpretation is more of a concern compared to outcome prediction. Due to its unique nature, general methods, such as convolutional neural network (CNN), are difficult to be directly applied. As such, we propose DKN, that is able to adapt to low sample size limitation and provide desired model interpretation. DKN is general in the sense that it not only works for both matrix and (high-order) tensor represented image data, but also could be applied to both discrete and continuous outcomes. DKN is built on a Kronecker product structure and implicitly imposes a piecewise smooth property on coefficients. Moreover, the Kronecker structure can be written into a convolutional form, so DKN also resembles a CNN, particularly, a fully convolutional network (FCN). Interestingly, DKN is also highly connected to the tensor regression framework proposed by Zhou et al. (2013), where a CANDECOMP/PARAFAC (CP) low-rank structure is imposed on tensor coefficients. We conduct both classification and regression analyses using real MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to demonstrate the effectiveness of DKN.

Ball Impurity: Measuring Heterogeneity in General Metric Spaces

Authors: Ting Li(Applied Mathematics,The Hong Kong Polytechnic University)

Abstract: Various domains, such as neuroimaging and network data analysis, have data in complex forms which do not process a Hilbert structure. We propose ball impurity, a general measure of heterogeneity among complex non-Euclidean objects. Our approach measures the difference between distributions in general metric spaces by generalizing the impurity degree in Hilbert spaces. The measure has properties analogous to triangular inequalities, is straightforward to compute and can be used for variable screening and tree models.

FPLS-DC: Functional partial least squares through distance covariance for imaging genetics

Authors: Wenliang Pan (Chinese Academy of Sciences)

Yue Shan(University of North Carolina at Chapel Hill)
Chuang Li(Sun Yat-sen University)
Shuai Huang(University of North Carolina at Chapel Hill)
Tengfei Li(University of North Carolina at Chapel Hill)
Yun Li(University of North Carolina at Chapel Hill)
Hongtu Zhu(University of North Carolina at Chapel Hill)

Abstract: Motivation: Imaging genetics integrates imaging and genetic techniques to examine how genetic variations influence the function and structure of organs like the brain or heart, providing insights into their impact on behavior and disease phenotypes. The use of organ-wide imaging endophenotypes has increasingly been employed to identify potential genes associated with complex disorders. However, analyzing organ-wide imaging data alongside genetic data presents two significant challenges: high dimensionality and complex relationships. To address these challenges, a proposed nonlinear inference framework aims to partially mitigate them. Results: We propose a functional partial least squares through distance covariance (FPLS-DC) framework for efficient whole-genome wide analyses of imaging phenotypes. It consists of two components: The first component utilizes the first FPLS-derived base function to reduce image dimensionality while screening genetic markers. The second component maximizes the distance correlation between genetic markers and projected imaging data, which is a linear combination of the first few FPLS-basis functions, using sequential quadratic programming. We efficiently approximate the null distribution of test statistics using a gamma approximation. Compared to existing methods, FPLS-DC offers computational and statistical efficiency for handling large-scale imaging genetics. In real-world applications, our approach successfully identifies novel Alzheimer's disease-related genetic variants and regions of interest, demonstrating its value as a statistical toolbox for imaging genetic studies. \\ Availability and implementation: The FPLS-DC method we propose opens up new research avenues and offers valuable insights for analyzing functional and high-dimensional data. Additionally, it serves as a useful tool for scientific analysis in practical applications within the field of imaging genetics research.

Session 21: Network Analysis and Spatial Autoregressive

Models

Testing Stochastic Block Models Via the Maximum Sampling Entry-Wise Deviation

Authors: Wei Lan (Southwestern University of Finance and Economics)

Abstract: The stochastic block model (SBM) has been widely used to analyze network data. Various goodness-of-fit tests have been proposed to assess the adequacy of model structures (see, e.g., Lei 2016 and Hu et al. 2021). To the best of our knowledge, however, none of the existing approaches are applicable for sparse networks in which the connection probability of any two communities is of order $O(n^{-1}\log n)$, and the number of communities is divergent. To fill this gap, we propose a novel goodness-of-fit test for the stochastic block model. The key idea is combining the test concept from Hu et al. (2021) with a sampling process that alleviates the negative impacts of network sparsity. We demonstrate theoretically that the proposed test statistic converges to the Type-I extreme value distribution under the null hypothesis regardless of the network structure. Accordingly, it can be applied to both dense and sparse networks. In addition, we obtain the asymptotic power against alternatives. Moreover, we introduce a bootstrap-corrected test statistic to improve the finite sample performance, recommend an augmented test statistic to increase the power, and extend the proposed test to the degree-corrected SBM. Simulation studies and two empirical examples with both dense and sparse networks indicate that the proposed method performs well.

Estimating Social Network Models with Missing Links

Authors: Arthur Lewbel (Boston College)
Xi Qu (Shanghai Jiao Tong University)
Xun Tang (Rice University)

Abstract: We propose an adjusted 2SLS estimator for social network models when

existing network links are missing from the sample at random (due, e.g., to recall errors by survey respondents, or lapses in data input). In the feasible structural form, missing links make all covariates endogenous and add a new source of correlation between the errors and endogenous peer outcomes (in addition to simultaneity), thus invalidating conventional estimators used in the literature. We resolve these issues by rescaling peer outcomes with estimates of missing rates and constructing instruments that exploit properties of the noisy network measures. We apply our method to study peer effects in household decisions to participate in a microfinance program in Indian villages. We find that ignoring missing links and applying conventional instruments would result in a sizeable upward bias in peer effect estimates.

Application of Functional Dependence to Spatial Econometrics

Authors: Zeqi Wu (Xiamen University)
Wen Jiang (Xiamen University)
Xingbai Xu (Xiamen University)

Abstract: This paper generalizes the concept of functional dependence from time series (Wu, 2005) and stationary random fields (El Machkouri, Volný and Wu, 2013) to non-stationary spatial processes. Within conventional settings in spatial econometrics, we define the concept of spatial functional dependence measure and establish a moment inequality, an exponential inequality, a law of large numbers, and a central limit theorem under it. We show that the dependent variables generated by some common spatial econometric models, including spatial autoregressive models and spatial panel data models, are functionally dependent under regular conditions. Furthermore, we investigate the properties of functional dependence measures under various transformations, which are useful in applications. Moreover, we compare spatial functional dependence with the spatial mixing and spatial near-epoch dependence proposed in Jenish and Prucha (2009, 2012), and illustrate its advantages.

Cryptocurrency Market Risk: a Network-centric Approach for Cryptocurrency Price trend Prediction

Authors: Wei DU (Renmin University of China)

Abstract: Predicting price trends of cryptocurrencies helps us understand the cryptocurrency market risk. Prior studies mainly investigate predictors such as historical trading data, macroeconomic development, and public interests in cryptocurrencies for price trend prediction while ignoring the predictive role of the relations between cryptocurrencies and systematic risk in the cryptocurrency market. In fact, the price movement of a cryptocurrency may be affected by those of other cryptocurrencies, thus, incorporating cryptocurrency interrelations can further improve the prediction performance. Therefore, we propose a novel end-to-end network-centric model for price trend prediction by utilizing a cryptocurrency network. A relation-wise graph attention network is proposed to extract network features. The effectiveness of the network model is validated using real-world cryptocurrency market data. The trading simulations for Bitcoin and portfolios reveal that our model obtains the highest profits. Our study provides insightful implications for investment decision support and risk understanding in the cryptocurrency market.

Session 22: New Machine Learning Paradigms in Biomedical Studies

Transfer Learning with Applications in Genomics

Authors: Hongzhe Li(University of Pennsylvania)

Abstract: This talk considers estimation and prediction of high-dimensional linear regression model for transfer learning, using samples from the target model as well as auxiliary samples from different but possibly related models. When the set of ``informative" auxiliary samples is known, an estimator and a predictor are proposed and their optimality is established. The optimal rates of convergence for prediction and estimation are faster than the corresponding rates without using the auxiliary samples. This implies that knowledge from the informative auxiliary samples can be transferred to improve the learning performance of the target problem. When sample informativeness is unknown, a data-driven procedure for transfer learning, called Trans-Lasso is proposed,

and its robustness to non-informative auxiliary samples and its efficiency in knowledge transfer is established. A related method, Trans-CLIME is developed for estimation and inference of high-dimensional Gaussian graphical models with transfer learning. Several applications in genomics will be presented, including prediction of gene expressions using the GTEx data and polygenetic risk score prediction using GWAS data. It is shown that Trans-Lasso and Trans-CLIME lead to improved performance in gene expression prediction in a target tissue by incorporating the data from multiple different tissues as auxiliary samples.

Robust Transfer Learning of Individualized Treatment Rules

Authors: Lu Tang(University of Pittsburgh)

Abstract: Causality-based individualized treatment rules (ITRs) are a steppingstone to precision medicine. To ensure unconfoundedness, ITRs are ideally derived from randomized experimental data, but the use cases of ITRs in the real-world data extend far beyond these controlled settings. It is of great interest to transfer knowledge learned from experimental data to real-world data, but hurdles remain. In this paper, we address two challenges in the transfer learning of ITRs. 1) In well-designed experiments, granular information crucial to decision making can be thoroughly collected. However, part of this may not be accessible in real-world decision-making. 2) Experimental data with strict inclusion criteria reflect a population distribution that may be very different from the real-world population data, leading to suboptimal ITRs. We propose a unified weighting scheme to learn a calibrated and robust ITR that simultaneously addresses the issues of covariate shift and missing covariates during prospective deployment, with a quantile-based approach to ensure worst-case safety under the uncertainty due to unavailable covariates. The performance of this method is evaluated in simulations and real-data applications.

ELSA: Efficient Label Shift Adaptation through the Lens of Semiparametric Models

Authors: Jiwei Zhao(University of Wisconsin-Madison)

Abstract: We study the domain adaptation problem with label shift in this work. Under the label shift context, the marginal distribution of the label varies across the training and testing datasets, while the conditional distribution of features given the label is the same. Traditional label shift adaptation methods either suffer from large estimation errors or require cumbersome post-prediction calibrations. To address these issues, we first propose a moment-matching framework for adapting the label shift based on the geometry of the influence function. Under such a framework, we propose a novel method named Efficient Label Shift Adaptation (ELSA), in which the adaptation weights can be estimated by solving linear systems. Theoretically, the ELSA estimator is root-n-consistent (n is the sample size of the source data) and asymptotically normal. Empirically, we show that ELSA can achieve state-of-the-art estimation performances without post-prediction calibrations, thus, gaining computational efficiency.

Deep Kernel Learning Based Gaussian Processes for Bayesian Image Regression Analysis

Authors: Jian Kang(University of Michigan)

Abstract: Regression models are widely used in neuroimaging applications to study complex associations between clinical variables and images. These models include scalar-on-image regression, image-on-scalar regression, and image-on-image regression. However, these models face challenges related to model interpretation, statistical inference, and prediction. To address these issues, we propose a Bayesian modeling framework that integrates deep neural networks (DNN) and Gaussian processes (GP) with kernel learning. We adopt GPs to represent observed images in a low dimensional vector space using the kernel decomposition approach. We construct the common covariance kernel of different GPs via DNNs, which can effectively learn important features from different types of images. We study the association among projections of images and variables of interests via regression models. We illustrate the advantages of the proposed framework over the state-of-the-art methods through extensive numerical experiments and analysis of fMRI data in the large-scale imaging studies.

Session 23: New Statistical Learning and Inferences in Data Science Applications

De-confounding Causal Inference Using Latent Multiple-mediator Pathways

Authors: Yubai Yuan (The Pennsylvania State University)
Annie Qiu (University of California Irvine)

Abstract: Causal effect estimation from observational data is one of the essential problems in causal inference. However, most estimation methods rely on the strong assumption that all confounders are observed, which is impractical and untestable in the real world. We develop a mediation analysis framework inferring the latent confounder for debiasing both direct and indirect causal effects. Specifically, we introduce generalized structural equation modeling that incorporates structured latent factors to improve the goodness-of-fit of the model to observed data, and deconfound the mediators and outcome simultaneously. One major advantage of the proposed framework is that it utilizes the causal pathway structure from cause to outcome via multiple mediators to debias the causal effect without requiring external information on latent confounders. In addition, the proposed framework is flexible in terms of integrating powerful nonparametric prediction algorithms while retaining interpretable mediation effects. In theory, we establish the nonparametric identification of both causal and mediation effects based on the proposed deconfounding method. Numerical experiments on both simulation settings and a normative aging study indicate that the proposed approach reduces the estimation bias of both causal and mediation effects.

Exploring the Causal Relationship between Geriatric Depression and Alzheimer's Disease

Authors: Yuexia Zhang (The University of Texas at San Antonio)
Yubai Yuan (The Pennsylvania State University)

Fei Xue (Purdue University)
Qi Xu (University of California, Irvine)
Annie Qu (University of California, Irvine)

Abstract: Depression and Alzheimer's Disease (AD) are both prevalent diseases in older adults. Using the data sets from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, we explore the causal relationship between geriatric depression and AD. We estimate the average treatment effect of geriatric depression on AD while controlling for ultrahigh-dimensional potential confounders, including DNA methylation. Moreover, we develop a novel causal mediation analysis approach to study mediation effects of potential mediators on the causal relationship between geriatric depression and AD. Based on the real data analysis results, we propose new prevention and treatment strategies for geriatric depression and AD through changing the selected confounders or mediators.

Statistical Inference for Maximin Effects: Identifying Stable Associations across Multiple Studies

Authors: Zijian Guo (Rutgers University of USA)

Abstract: Integrative analysis of data from multiple sources is critical to making generalizable discoveries. Associations that are consistently observed across multiple source populations are more likely to be generalized to target populations with possible distributional shifts. In this paper, we model the heterogeneous multi-source data with multiple high-dimensional regressions and make inferences for the maximin effect (Meinshausen, Bhlmann, AoS, 43(4), 1801--1830). The maximin effect provides a measure of stable associations across multi-source data. A significant maximin effect indicates that a variable has commonly shared effects across multiple source populations, and these shared effects may be generalized to a broader set of target populations. There are challenges associated with inferring maximin effects because its point estimator can have a non-standard limiting distribution. We devise a novel sampling method to construct valid confidence intervals for maximin effects. The proposed confidence interval attains a parametric length. This sampling procedure and the related theoretical analysis are of independent interest for solving other non-standard inference problems. Using genetic

data on yeast growth in multiple environments, we demonstrate that the genetic variants with significant maximin effects have generalizable effects under new environments.

Network Community Detection Using Higher-Order Structures

Authors: Xianshi Yu (University of Wisconsin)
Ji Zhu (University of Michigan)

Abstract: Many real-world networks commonly exhibit an abundance of subgraphs or higher-order structures, such as triangles and by-fans, surpassing what is typically observed in randomly generated networks (Milo et al., 2002). However, statistical models accounting for this phenomenon are limited, especially when community structure is of interest. This limitation is coupled with a lack of community detection methods that leverage subgraphs or higher-order structures. In this paper, we propose a novel community detection method that effectively incorporates these higher-order structures within a network. We also develop a finite-sample error bound for community detection accuracy under an edge-dependent network model, which includes both community and triangle structures. This error bound is characterized by the expected triangle degree, leading to the proposed method's consistency. To our knowledge, this is the first statistical error bound and consistency result considering a single network's community detection under a network model with dependent edges. Through simulations and a real-world data example, we demonstrate that our method reveals network communities otherwise obscured by methods that disregard higher-order structures.

Session 24: Probability and Statistical Methods in Science

Regularized Greedy Gradient Q-learning with Mobile Health Applications

Authors: Min Qian (Columbia University)

Abstract: Recent advance in health and technology has made mobile apps a viable

approach to delivering behavioral interventions in areas including physical activity encouragement, smoking cessation, substance abuse prevention, and mental health management. Due to the chronic nature of most of the disorders and heterogeneity among mobile users, delivery of the interventions needs to be sequential and tailored to individual needs. We operationalize the sequential decision making via a policy that takes a mobile user's past usage pattern and health status as input and outputs an app/intervention recommendation with the goal of optimizing the cumulative rewards of interest in an indefinite horizon setting. The vast majority of the literature in the field focuses on studying the convergence of the algorithms with infinite amount of data in computer science domain. Their performances in health applications with limited amount of data and high noise are yet to be explored. Technically the nature of sequential decision making results in an objective function that is non-smooth and non-convex in the model parameters. This poses theoretical challenges to the characterization of the asymptotic properties of the optimizer of the objective function, as well as computational challenges for optimization. This problem is especially exacerbated with the presence of high dimensional data in mobile health applications. We propose a regularized greedy gradient Q-learning (RGGQ) method to tackle this estimation problem. The optimal policy is estimated via an algorithm which synthesizes the proximal gradient method and the GGQ algorithm in the presence of regularization, and its asymptotic properties are established.

Bayesian Spatially Varying Weight Neural Networks with the Soft-Thresholded Gaussian Process Prior

Authors: Ben Wu (Renmin University of China)
Keru Wu(Duke University)
Jian Kang(University of Michigan)

Abstract: Deep neural networks (DNN) have been adopted in the scalar-on-image regression which predicts the outcome variable using image predictors. However, training DNN often requires a large sample size to achieve a good prediction accuracy and the model fitting results can be difficult to interpret. In this work, we construct a novel single-layer Bayesian neural network (BNN) with spatially varying weights for the scalar-on-image regression. Our goal is to select interpretable image regions and to

achieve high prediction accuracy with limited training samples. We assign the soft-thresholded Gaussian process (STGP) prior to the spatially varying weights and develop an efficient posterior computation algorithm based on stochastic gradient Langevin dynamics (SGLD). The BNN-STGP provides large prior support for sparse, piecewise-smooth, and continuous spatially varying weight functions, enabling efficient posterior inference on image region selection and automatically determining the network structures. We establish the posterior consistency of model parameters and selection consistency of image regions when the number of voxels/pixels grows much faster than the sample size. We compared our methods with state-of-the-art deep learning methods via analyses of multiple real data sets including the task fMRI data in the Adolescent Brain Cognitive Development (ABCD) study.

Data Augmentation MCMC for Bayesian Inference from Privatized Data

Authors: Nianqiao Ju (Purdue University)
Ruobin Gong (Rutgers University)
Jordan Awan (Purdue University)
Vinayak Rao (Purdue University)

Abstract: Differentially private mechanisms protect privacy by introducing additional randomness into the data. When the data analyst has access only to the privatized data, it is a challenge to perform valid statistical inference on parameters underlying the confidential data. Specifically, the likelihood function of the privatized data requires integrating over the large space of confidential databases and is typically intractable. For Bayesian analysis, this results in a posterior distribution that is doubly intractable, rendering traditional MCMC techniques inapplicable. We propose an MCMC framework to perform Bayesian inference from the privatized data, which is applicable to a wide range of statistical models and privacy mechanisms. Our MCMC algorithm augments the model parameters with the unobserved confidential data, and alternately updates each one conditional on the other. For the potentially challenging step of updating the confidential data, we propose a generic approach that exploits the privacy guarantee of the mechanism to ensure efficiency. We give results on computational complexity, acceptance rate, and mixing properties of our MCMC. This talk is based on joint work with Jordan Awan, Robin Gong, and Vinayak Rao (https://arxiv.org/abs/2206.00710, NeurIPS 2022).

Gender Differences in the Non-specific and Health-specific Use of Social Media Before and During the COVID-19 Pandemic: Trend Analysis Using HINTS 2017-2020 Data

Authors: Linglong Ye (Xiamen University)
Yang Chen (University of International Business and Economics)
Yongming Cai (University of International Business and Economics)
Yi-Wei Kao (Fu Jen Catholic University)
Yuanxin Wang (Minzu University of China)
Mingchih Chen (Fu Jen Catholic University)
Ben-Chang Shia (Fu Jen Catholic University)
Lei Qin (University of International Business and Economics)

Abstract: The use of social media has changed since the outbreak of coronavirus disease 2019 (COVID-19). However, little is known about the gender disparity in social media use for non-specific and health-specific issues before and during the COVID-19 pandemic. Based on a gender difference perspective, this study aimed to examine how the non-specific and health-specific uses of social media changed in 2017-2020. The data came from the Health Information National Trends Survey Wave 5 Cycle 1-4. This study included 10,426 participants with complete data. Compared to 2017, there were higher levels of general use in 2019 and 2020, and an increased likelihood of health-related use in 2020 was reported among the general population. Female participants were more likely to be non-specific and health-specific users than males. Moreover, the relationship of gender with general use increased in 2019 and 2020; however, concerning health-related use, it expanded in 2019 but narrowed in 2020. The COVID-19 global pandemic led to increased use of social media, especially for health-related issues among males. These findings further our understanding of the gender gap in health communication through social media, and contribute to targeted messaging to promote health and reduce disparities between different groups during the pandemic.

Session 25:Recent Advances in Analyzing Complex Structured

Data

Causal Structural Learning and Application in Epidemiology

Authors: Le Bao (Penn State)
Changcheng Li (Dalian University of Technology)
Runze Li (Penn State)
Songshan Yang (Renmin University of China)

Abstract: The Population-based HIV Impact Assessment (PHIA) is an ongoing project that conducts nationally representative HIV-focused surveys for measuring national and regional progress toward UNAIDS'90-90-90 targets, the primary strategy to end the HIV epidemic. We believe the PHIA survey offers a unique opportunity to better understand the key factors that drive the HIV epidemics in the most affected countries in sub-Saharan Africa. In this article, we propose a novel causal structural learning algorithm to discover important covariates and potential causal pathways for 90-90-90 targets. Existing constrained-based causal structural learning algorithms are quite aggressive in edge removal. The proposed algorithm preserves more information about important features and potential causal pathways. It is applied to the Malawi PHIA (MPHIA) data set and leads to interesting results. We further compare and validate the proposed algorithm using BIC and using Monte Carlo simulations, and show that the proposed algorithm achieves improvement in true positive rates in important feature discovery over existing algorithms.

Semi-supervised Estimation of Event Rate with Doubly-censored Survival Data

Authors: Yang Wang (Harvard University)
Qingning Zhou (University of North Carolina at Charlotte)
Xuan Wang (Harvard University)
Tianxi Cai(Harvard University)

Abstract: Electronic Health Record (EHR) has emerged as a valuable source of data for translational research. To leverage EHR data for risk prediction and subsequently clinical decision support, clinical endpoints are often time to onset of a clinical condition of interest.

Precise information on clinical event times are often not directly available and requires labor-intensive manual chart review to ascertain. In addition, events may occur outside of the hospital system, resulting in both left and right censoring or often termed as double censoring. On the other hand, proxies such as time to the first diagnostic code are readily available yet with varying degrees of accuracy. Using error-prone event times derived from these proxies can lead to biased risk estimates while only relying on manually annotated event times, which are typically only available for a small subset of patients, can lead to high variability. This signifies the need for semi-supervised estimation methods that can efficiently combine information from both the small subset of labeled observations and a large size of surrogate proxies. While semi-supervised estimation methods have been recently developed for binary and right-censored data, no methods currently exist in the presence of doubly censoring. This paper fills the gap by developing a robust and efficient Semi-supervised Estimation of Event rate with Doubly-censored Survival data (SEEDS) by leveraging a small set of gold standard labels and a large set of surrogate features. Under mild regularity conditions, we demonstrate that the proposed SEEDS estimator is consistent and asymptotically normal. Extensive simulation results illustrate that SEEDS performs well in finite samples and can be substantially more efficient compared to the supervised counterpart. We apply the SEEDS procedure to estimate the age-specific survival rate of type 2 diabetes (T2D) using EHR data from Mass General Brigham (MGB).

Distributed Nonparametric Regression via Prediction-Based Aggregation

Authors: Yikai Xu (Fudan University)
Zhao Chen (Fudan University)
Runze Li (Pennsylvania State University)

Abstract: Distributed statistical modelling is a powerful tool to tackle with modern massive dataset while protecting data privacy simultaneously. In this work, we propose a communication-efficient data-driven weighted aggregation procedure based on model prediction performance. Theoretically, we show our method is asymptotically optimal in the sense of achieving the lowest possible risk for a broad class of least squares estimator

(typically, B-spline nonparametric regression) and provide the limit of estimated weights. The superiority of our method is verified both from numerical experiments and a real data example.

Determination of the Effective Cointegration Rank in High-dimensional Time-series Predictive Regressions

Authors: Fuyi Fang (Zhejiang University)
Zhaoxing Gao (Zhejiang University)
Ruey Tsay (University of Chicago)

Abstract: This paper proposes a new approach to identifying the effective cointegration rank in high-dimensional unit-root (HDUR) time series from a prediction perspective using reduced-rank regression. For a HDUR process x_t and a stationary series y_t of interest, our goal is to predict future values of y_t using x_t and lagged values of y_t. The proposed framework consists of a two-step estimation procedure with theoretical guarantees. Simulated and real examples are used to illustrate the proposed method, and the empirical study suggests that the procedure fares well in predicting stock returns.

Session 26: Recent Advances in High Dimensional Statistics

Large-Scale Multiple Testing of Composite Null Hypotheses Under Heteroskedasticity

Authors: Bowen Gang (Fudan University)
Trambak Banerjee (University of Kansas)

Abstract: Heteroskedasticity poses several methodological challenges in designing valid and powerful procedures for simultaneous testing of composite null hypotheses. In particular, the conventional practice of standardizing or re-scaling heteroskedastic test

statistics in this setting may severely affect the power of the underlying multiple testing procedure. Additionally, when the inferential parameter of interest is correlated with the variance of the test statistic, methods that ignore this dependence may fail to control the type I error at the desired level. We propose a new Heteroskedasticity Adjusted Multiple Testing (HAMT) procedure that avoids data reduction by standardization, and directly incorporates the side information from the variances into the testing procedure. Our approach relies on an improved nonparametric empirical Bayes deconvolution estimator that offers a practical strategy for capturing the dependence between the inferential parameter of interest and the variance of the test statistic. We develop theory to show that HAMT is asymptotically valid and optimal for FDR control. Simulation results demonstrate that HAMT outperforms existing procedures with substantial power gain across many settings at the same FDR level. The method is illustrated on an application involving the detection of engaged users on a mobile game app.

Block-Diagonal Test for High-Dimensional Covariance Matrices

Authors: Jiayu Lai(Northeast Normal University)
Xiaoyi Wang(Beijing Normal University)
Kaige Zhao(Northeast Normal University)
Shurong Zheng(Northeast Normal University)

Abstract: The testing structure of a high-dimensional covariance matrix plays an important role in financial stock analyses, genetic series analyses, and many other fields. Testing that the covariance matrix is block-diagonal under the high-dimensional setting is a main focus of this paper. To tackle this problem, test procedures that are powerful under normality assumptions, two-diagonal block assumptions or sub-block dimensionality assumptions have been proposed in several existing studies. To relax these conditions, a test framework based on U-statistics is proposed in this paper, and the asymptotic distributions of those U-statistics are established under the null and alternative hypotheses. Moreover, another test approach is developed for alternatives with different sparsity levels. Finally, both a simulation study and real data analysis are conducted to show the performance of our proposed test procedures.

Robust Estimation of Number of Factors in High Dimensional Factor Modeling via Spearman's Rank Correlation Matrix

Authors: Jiaxin Qiu(The University of Hong Kong)
Zeng Li(Southern University of Science and Technology)
Jianfeng Yao(Chinese University of Hong Kong, Shenzhen)

Abstract: Determining the number of factors in high-dimensional factor modeling is essential but challenging, especially when the data are heavy-tailed. In this paper, we introduce a new estimator based on the spectral properties of Spearman's rank correlation matrix under the high-dimensional setting, where both dimension and sample size tend to infinity proportionally. Our estimator is applicable for scenarios where either the common factors or idiosyncratic errors follow heavy-tailed distributions. We prove that the proposed estimator is consistent under mild conditions. Numerical experiments also demonstrate the superiority of our estimator compared to existing methods, especially for the heavy-tailed case.

Session 27: Recent Advances of High-Dimensional Inference and Statistical Learning

Selecting the Number of Communities for Weighted Networks with Stepwise Variance Profile Scaling

Authors: Xiaodong Li (University of California, Davis)
Xiaohan Hu (University of California, Davis)

Abstract: This work aims to investigate how to select the number of communities for weighted networks without a full likelihood modeling. In addition to a degree-corrected

stochastic block model (DCSBM) for modeling the mean adjacency matrix, we also model the variance profile matrix in a parametric manner. By conducting sequential spectral clustering with an increasing number of communities, our stopping criterion is based on comparing the spectrum of a normalized residual matrix and an explicit threshold, while the normalization is based on the Sinkorn scaling of the estimated variance profile matrix. This procedure is shown to be consistent in estimating the true number of communities for a general class of weighted networks, provided the variance profile matrix can be consistently estimated when the candidate number of clusters is correct. Fast implementation of matrix scaling is also possible by exploiting the spectral structure of the DCSBM. Extensive numerical experiments on simulated and real network data also illustrate the competitive empirical properties of our proposed method.

Network Regression and Supervised Centrality Estimation

Authors: Junhui Cai (University of Notre Dame)
Dan Yang (The University of Hong Kong)
Wu Zhu (Tsinghua University)
Haipeng Shen (The University of Hong Kong)
Linda Zhao (University of Pennsylvania)

Abstract: The centrality in a network is often used to measure nodes' importance and model network effects on a certain outcome. Empirical studies widely adopt a two-stage procedure, which first estimates the centrality from the observed noisy network and then infers the network effect from the estimated centrality, even though it lacks theoretical understanding. We propose a unified modeling framework, under which we first prove the shortcomings of the two-stage procedure, including the inconsistency of the centrality estimation and the invalidity of the network effect inference. Furthermore, we propose a supervised centrality estimation methodology, which aims to simultaneously estimate both centrality and network effect. The advantages in both regards are proved theoretically and demonstrated numerically via extensive simulations and a case study in predicting currency risk premiums from the global trade network.

Empirical Bayes Estimation: When does g-modeling beat

f-modeling in theory (and in practice)?

Authors: Yandi Shen (University of Chicago)
Yihong Wu (Yale University)

Abstract: Empirical Bayes (EB) is a popular framework for large-scale inference that aims to find data-driven estimators to compete with the Bayesian oracle that knows the true prior. Two principled approaches to EB estimation have emerged over the years: f-modeling, which constructs an approximate Bayes rule by estimating the marginal distribution of the data, and g-modeling, which estimates the prior from data and then applies the learned Bayes rule. For the Poisson model, the prototypical examples are the celebrated Robbins estimator and the nonparametric MLE (NPMLE), respectively. It has long been recognized in practice that the Robbins estimator, while being conceptually appealing and computationally simple, lacks robustness and can be easily derailed by "outliers" (data points that were rarely observed before). In this talk we provide a theoretical justification for the superiority of NPMLE over Robbins for heavy-tailed data by considering priors with bounded pth moment previously studied for the Gaussian model. For the Poisson model with sample size n, assuming p>1 (for otherwise triviality arises), we show that the NPMLE with appropriate regularization achieves a total regret O(n^{3/(2p+1)}), which is minimax optimal within logarithmic factors. In contrast, the total regret of Robbins estimator (with similar truncation) is O(n^{3/(p+2)}) and hence suboptimal by a polynomial factor.

Session 28: Some Theories about Deep Neural Networks

Generalization Ability of Wide Neural Networks on R

Authors: Jianfa Lai (Tsinghua University)
Manyun Xu (Tsinghua University),
Rui Chen (Tsinghua University),
Qian Lin (Tsinghua University)

Abstract: We perform a study on the generalization ability of the wide two-layer ReLU neural network on $\mathbb{R}$. We first establish some spectral properties of the neural tangent kernel (NTK): $a)$ K_{d}, the NTK defined on $\mathbb{R}^{d}$, is positive

definite; $b)$ $\lambda_{i}(K_{1})$, the i-th largest eigenvalue of K_{1}, is proportional to i^{-2}. We then show that: $i)$ when the width $m\rightarrow\infty$, the neural network kernel (NNK) uniformly converges to the NTK; $ii)$ the minimax rate of regression over the RKHS associated to K_{1} is $n^{-2/3}$; $iii)$ if one adopts the early stopping strategy in training a wide neural network, the resulting neural network achieves the minimax rate; $iv)$ if one trains the neural network till it overfits the data, the resulting neural network can not generalize well. Finally, we provide an explanation to reconcile our theory and the widely observed ``benign overfitting phenomenon''.

ZooD: Exploiting Model Zoo for Out-of-Distribution Generalization

Authors: Qishi Dong (Hong Kong Baptist University)
Awais Muhammad (Kyung-Hee University)
Fengwei Zhou (Huawei Noah's Ark Lab)
Chuanlong Xie (Beijing Normal University)
Tianyang Hu (Huawei Noah's Ark Lab)
Yongxin Yang (Huawei Noah's Ark Lab)
Sung-Ho Bae (Kyung-Hee University)
Zhenguo Li (Huawei Noah's Ark Lab)

Abstract: Recent advances on large-scale pre-training have shown the great potential of leveraging a large set of Pre-Trained Models (PTMs) for improving Out-of-Distribution (OoD) generalization, for which the goal is to perform well on possible unseen domains after fine-tuning on multiple training domains. However, maximally exploiting a zoo of PTMs is challenging since fine-tuning all possible combinations of PTMs and hyperparameters is computationally prohibitive while accurate selection of PTMs requires tackling the possible data distribution shift for OoD tasks. In this work, we propose ZooD, a paradigm for PTMs ranking and ensemble with feature selection. Our proposed metric ranks PTMs by quantifying inter-class discriminability and inter-domain stability of the task data features extracted by the PTMs in a leave-one-domain-out cross-validation manner. The top-K ranked models are then aggregated for the target OoD task. To avoid accumulating noise induced by model ensemble, we propose an efficient variational EM algorithm to select informative features. We evaluate our paradigm on a diverse model zoo consisting of 35 models for various OoD tasks and demonstrate: (i) model ranking is better correlated with fine-tuning ranking than previous methods and up to 9859x faster than brute-force fine-tuning; (ii) OoD generalization outperforms the state-of-the-art methods and accuracy on most challenging task DomainNet is improved from 46.5% to

50.6%.

Optimality of Wide Neural Networks in Large Dimensions

Authors: Weihao Lu (Tsinghua University)
Haobo Zhang (Tsinghua University),
Manyun Xu (Tsinghua University),
Qian Lin (Tsinghua University)

Abstract: We investigate the generalization ability of a two-layer wide neural network for large dimensional data (where the sample size n is polynomially depending on the dimension d of the samples, i.e., $n\asymp d^{s}$ for some $s\geq 1$). We first characterize the lower bound and upper bound of the kernel regression through the metric entropy $\bar{\varepsilon}_{n}^{2}$ and the Mendelson complexity ε_{n}^{2} respectively. We then show that when $s=1$ or $s=3,7,11,\cdots$, the minimax rate of the excess risk of the neural tangent kernel(NTK) regression on $\bbS^{d}$ is $n^{-(s-1)/(2s)}$ provided the target function falls into the RKHS associated with the NTK. This result greatly extends our knowledge of the performance of kernel regression over large-dimension data (e.g., we now know that the NTK interpolation can not generalize when $n\asymp d$). Combined with the fact that the excess risk of a wide neural network can be fully characterized by the excess risk of the corresponding NTK regression, we know that the aforementioned claim also holds for the two-layer wide neural network.

Session 29: Special Invited Session for JDS

Iterative Connecting Probability Estimation for Networks

Authors: Yichen Qin (College of Business, University of Cincinnati)

Abstract: Estimating the probabilities of connections between vertices in a random network using an observed adjacency matrix is an important task for network data

analysis. Many existing estimation methods are based on certain assumptions on network structure, which limit their applicability in practice. Without making strong assumptions, we develop an iterative connecting probability estimation method based on neighborhood averaging. Starting at a random initial point or an existing estimate, our method iteratively updates the pairwise vertex distances, the sets of similar vertices, and connecting probabilities to improve the precision of the estimate. We propose a two-stage neighborhood selection procedure to achieve the trade-off between the smoothness of the estimate and the ability to discover local structure. The tuning parameters can be selected by cross-validation. We establish desirable theoretical properties for our method and further justify its superior performance by comparing it with existing methods in simulation and real data analysis.

Changepoint Detection in Preferential Attachment Networks

Authors: Daniel Cirkovic (Texas A&M University)
Tiandong Wang (Fudan University)
Xianyang Zhang (Texas A&M University)

Abstract: Generative, temporal network models play an important role in analyzing the dependence structure and evolution patterns of complex networks. Due to the complicated nature of real network data, it is often naive to assume that the underlying data-generative mechanism itself is invariant with time. Such observation leads to the study of changepoints or sudden shifts in the distributional structure of the evolving network. We propose a likelihood-based method to detect changepoints in undirected, affine preferential attachment networks, and establish a hypothesis testing framework to detect a single changepoint, together with a consistent estimator for the changepoint.

Dependence Model Assessment and Selection with DecoupleNets

Authors: Marius Hofert (The University of Hong Kong)
Avinash PRASAD (University of Waterloo)
Mu ZHU (University of Waterloo)

Abstract: Neural networks are suggested for learning a map from d-dimensional samples with any underlying dependence structure to multivariate uniformity in d' dimensions. This map, termed DecoupleNet, can be used for dependence model assessment and selection. For d' = 2, DecoupleNets allow for simple model assessment and selection, both numerically and graphically. The graphical approach allows one to identify in which regions of the domain, a candidate model does not provide an adequate fit. Applications to simulated and real world data illustrate the usefulness of DecoupleNets.

Model-based Doublet Detection in Single-cell Multi-omics Data

Authors: Wei Chen (University of Pittsburgh)

Abstract: In droplet-based single-cell experiments, doublets are generated when a droplet encapsulates more than one single cell and will bias downstream analysis. Most state-of-the-art computational methods adopt a semi-supervised machine learning approach, which artificially synthesizes doublets based on the observed data, and uses highly variable genes as machine learning features. We propose a novel model-based method for doublet identification, which explicitly models the generative process of doublet using compound Poisson distribution. Instead of relying on highly variable genes, our method makes use of stable genes that have similar expression levels across different cell types. Those stable genes provide valuable information for doublet detection, which is overlooked by the existing methods. Through simulated and real datasets with cell hashing as ground truth, we demonstrate that our method is superior or comparable to state-of-the-art methods and is particularly advantageous in doublet rate estimation, interpretability, and computational time. Our method and newly generated benchmarking datasets provide a valuable tool and resource for this fundamental problem in single-cell multiomics analysis.

Session 30: Statistical Methods for Healthcare and Biometrics

A Flexible Pseudo Outcome Regression Framework for Analyzing Treatment Heterogeneity in Survival Outcomes

Authors: Na Bo (University of Pittsburgh)

Ying Ding (Biostatistics, University of Pittsburgh)

Abstract: Estimating heterogeneous treatment effect plays a central role in personalized medicine as it provides informative guidelines in tailoring existing therapies for each patient to get the optimal treatment. In this project, we propose a flexible pseudo-outcome regression framework to estimate CATE in survival outcomes by using multi-step algorithms coupled with different machine learning methods. We perform comprehensive simulations under different randomized clinical trials (RCT) and observational study settings to evaluate six different pseudo-outcome approaches. We also apply the proposed methods to an RCT and an observational study to estimate CATE and make treatment recommendations.

Introducing the Specificity Score: a Measure of Causality Beyond P Value

Authors: Wang Miao (Center for Statistical Science, Peking University)

Abstract: There is considerable debate and doubt about the use of P value in scientific research in recent years, particularly after its use is banished in several prestigious journals. Much scientific research is concerned with uncovering causal associations, however, P value is mostly a measure of the significance of a statistical association, which could be biased from the causal association of interest and lead to false/trivial scientific discoveries particularly in the presence of unmeasured confounding. In this talk, I will introduce a score measuring the specificity of causal associations and a specificity score-based test about the existence of causal effects in the presence of unmeasured confounding. Under certain conditions, this approach has controlled type I error and power approaching unity for testing the null hypothesis of no causal effect. A visualization approach using a heatmap of specificity is proposed to communicate all specificity score/test information in a universal and effective manner. This approach only entails a rough idea on the broadness of the causal associations in sight, e.g., the maximum or upper-bound number of causes/outcomes of an outcome/treatment, but does not require to know exactly the exclusion of certain causal effects or the availability of auxiliary

variables. This approach is related to Hill's specificity criterion for causal inference, but I will discuss the difference from Hill's. This approach admits for joint causal discovery with multiple treatments and multiple outcomes, which is particularly suitable for gene expressions studies, Mendelian randomization and EHR studies. Identification and estimation will be briefly covered. Simulations are used for illustration and an application to a mouse obesity dataset detects potential active effects of genes on clinical traits that are relevant to metabolic syndrome.

Sparse Causal Mediation Analysis with Unmeasured Mediator-outcome Confounding

Authors: Kang Shuai(Peking University)
Lan Liu(University of Minnesota)
Yangbo He(Peking University)
Wei Li(Renmin University of China)

Abstract: Causal mediation analysis aims to investigate how an intermediary factor called mediator regulates the causal effect of a treatment on an outcome. With the increasing availability of measurements on a large number of potential mediators in various disciplines, methods for conducting mediation analysis with many or even high-dimensional mediators have been proposed. However, they often assume there is no unmeasured confounding between mediators and the outcome. This paper allows such confounding and provides an approach to address both identification and mediator selection problems under the structural equation modeling framework. The identification strategy involves constructing a pseudo proxy variable for unmeasured confounding based on a latent factor model for multiple mediators. Using this proxy variable, we then propose a partially penalized procedure to select important mediators which have nonzero effects on the outcome. The resultant estimates are consistent and the estimates of nonzero parameters are asymptotically normal. Simulation studies show advantageous performance of the proposed procedure over other existing methods. We finally apply our approach to genomic data and identify gene expressions that may actively mediate the effect of a genetic variant on mouse obesity.

Semiparametric Efficient Estimation of Genetic Relatedness with Machine Learning Methods

Authors: Xu Guo (Beijing Normal University)
Yiyuan Qian(Beijing Normal University)
Hongwei Shi(Beijing Normal University)
Weichao Yang (Beijing Normal University)
Niwen Zhou(Beijing Normal University)

Abstract: In this paper, we propose semiparametric efficient estimators of genetic relatedness between two traits in a model-free framework. Most existing methods require specifying certain parametric models involving the traits and genetic variants. However, the bias due to model misspecification may yield misleading statistical results. Moreover, the semiparametric efficient bounds for estimators of genetic relatedness are still lacking. In this paper, we develop semiparametric efficient estimators with machine learning methods and construct valid confidence intervals for two important measures of genetic relatedness: genetic covariance and genetic correlation, allowing both continuous and discrete responses. Based on the derived efficient influence functions of genetic relatedness, we propose a consistent estimator of the genetic covariance as long as one of genetic values is consistently estimated. The data of two traits may be collected from the same group or different groups of individuals. Various numerical studies are performed to illustrate our introduced procedures. We also apply proposed procedures to analyze Carworth Farms White mice genome-wide association study data.

Session 31: Statistical Methods for Network Data and Precision Health

A Model-Agnostic Graph Neural Network Integrating Heterogeneous Network Data

Authors: Wenzhuo Zhou (UCI)
Annie Qu (UC Irvine)
Keiland Cooper (UCI)
Norbert Fortin (UCI)
Babak Shalbaba (UCI)

Abstract: Graph Neural Networks (GNNs) have achieved promising performance in a variety of graph-focused tasks. Despite their success, the two major limitations of existing GNNs are the capability of learning various-order representations and providing interpretability of such deep learning-based black-box models. To tackle these issues, we propose a novel Model-agnostic Geometric Deep Network (MaGNet) framework. The proposed framework is able to extract knowledge from high-order neighbors, sequentially integrates information of various orders, and offers explanations for the learned model by identifying influential compact graph structures. In particular, MaGNet consists of two components: an estimation model for the latent representation of complex relationships under graph topology, and an interpretation model that identifies influential nodes, edges and important node features. Theoretically, we establish the generalization error bound for MaGNet via empirical Rademacher complexity, and showcase its power to represent the layer-wise neighborhood mixing. We conduct comprehensive numerical studies using both simulated data and a real-world case study on investigating the neural mechanisms of the Rat Hippocampus, demonstrating that the performance of MaGNet is competitive with state-of-the-art methods.

Estimating Cell-type-specific Gene Co-expression Networks from Bulk Gene Expression Data

Authors: Emma Jingfei Zhang (Emory University)

Abstract: Inferring and characterizing gene co-expression networks has led to important insights on the molecular mechanisms of complex diseases. Most co-expression analyses to date have been performed on gene expression data collected from bulk tissues with different cell type compositions across samples. As a result, the co-expression estimates

only offer an aggregate view of the underlying gene regulations and can be confounded by heterogeneity in cell type compositions, failing to reveal gene coordination that may be distinct across different cell types. In this paper, we describe a flexible framework for estimating cell-type-specific gene co-expression networks from bulk sample data, without making specific assumptions on the distributions of gene expression profiles in different cell types. We develop a novel sparse least squares estimator, referred to as CSNet, that is efficient to implement and has good theoretical properties. Using CSNet, we analyzed the bulk gene expression data from a cohort study on Alzheimer's disease and identified previously unknown cell-type-specific co-expressions among Alzheimer's disease risk genes, suggesting cell-type-specific disease pathology for Alzheimer's disease.

Fairness-adjusted Neyman-Pearson Classifiers

Authors: Ziqing Guo (HKUST)
Xin Tong (USC)
Lucy Xia (HKUST)

Abstract: We utilize a dual-focused Neyman-Pearson (NP) classification paradigm that seeks minimal type II error under simultaneous control over both type I error and fairness bias. Leveraging a LDA model, we develop a new oracle framework for dual-focused NP classification, which is a first of its kind. Our proposed finite-sample-based classifiers satisfy both the fairness constraint and type I error constraint with high probability at the population level. We also derive Oracle bounds on the excess type II error. Numerical and real data analyses demonstrate its superior performance.

Inference on Potentially Identified Subgroups in Clinical Trials

Authors: Shuoxun Xu (HKUST)
Xinzhou Guo (HKUST)

Abstract: When subgroup analyses are conducted in clinical trials with moderate or high dimensional covariates, we often need to identify candidate subgroups from the data and evaluate the potentially identified subgroups in a replicable way. The usual statistical

inference applied to the potentially identified subgroups, assuming the subgroups are just what we observe from the data, might suffer from bias issue when the regularity assumption that heterogeneity exists is violated. In this talk, we introduce a shift-based method to address nonregularity bias issue and combined with subsampling, develop a de-biased inference procedure for potentially identified subgroups. The proposed method is model-free and asymptotically efficient. We show that with appropriate noise added to the shift, the proposed method can be viewed as an asymmetric smoothing approach and achieve privacy protection while remaining valid and efficient. We demonstrate the merits of the proposed method by re-analyzing the ACTG 175 trial.

Session 32： Statistical Modeling for Complex Networks

Spectral Clustering for Heterophilic Stochastic Block Models with Dynamic Node Memberships

Authors: Jing Lei (Carnegie Mellon University)
Kevin Lin (University of Pennsylvania)

Abstract: We consider a time-ordered sequence of networks stemming from stochastic block models where nodes are gradually changing memberships over time and no network contains sufficient signal strength to recover its community structure. To estimate the time-varying community structure, we develop KD-SoS (kernel debiased sum-of-square), a method performing spectral clustering after a debiased sum-of-squared aggregation of adjacency matrices. Our theory demonstrates via a novel bias-variance decomposition that KD-SoS achieves consistent community detection of each network even when the networks are heterophilic, and do not require smoothness in the time-varying dynamics of between-community connectivities. We also prove the identifiability of aligning community structures across time based on how rapidly nodes change communities, and develop a data-adaptive bandwidth tuning procedure for KD-SoS. We demonstrate the utility and advantages of KD-SoS through simulations and a novel analysis on the time-varying dynamics in gene coordination in the human developing brain system.

Point of Interest Recommendation

Authors: Yiyuan Liu (Jiangxi University)
Ya Wang (SUSTech),
Bingyi Jing(SUSTech)

Abstract: With the rapid development of wireless communication technologies, location-based social networks, such as Foursquare and Gowalla, have become very popular. This makes it possible to mine user's preference on locations and provided favourite recommendations. However, check-in data is sparse, long-tail, temporal and sociability. In this talk, we consider recommendation system using tensor method for handling such types of data with various techniques. Experiments on a real check-in database show that the proposed method can offer better location recommendation.

A Dynamic Additive and Multiplicative Effects Network Model with Application to the United Nations Voting Behaviors

Authors: Bomin Kim (Freddie Mac)
Xiaoyue Niu (Penn State University)
David Hunter (Penn State University)
Xun Cao (Penn State University)

Abstract: Motivated by a study of United Nations voting behaviors, we introduce a regression model for a series of networks that are correlated over time. Our model is a dynamic extension of the additive and multiplicative effects network model (AMEN). In addition to incorporating a temporal structure, the model accommodates two types of missing data thus allows the size of the network to vary over time. We demonstrate via simulations the necessity of various components of the model. We apply the model to the United Nations General Assembly voting data from 1983 to 2014 to answer interesting research questions regarding international voting behaviors. In addition to finding important factors that could explain the voting behaviors, the model-estimated additive effects, multiplicative effects, and their movements reveal meaningful foreign policy positions and alliances of various countries.

Network Gradient Descent Algorithm for Decentralized Federated Learning

Authors: Shuyuan Wu (Peking University)
Danyang Huang (Remin University of China)
Hansheng Wang (Peking University)

Abstract: We study a fully decentralized federated learning algorithm, which is a novel gradient descent algorithm executed on a communication-based network. For convenience, we refer to it as a network gradient descent (NGD) method. In the NGD method, only statistics (e.g., parameter estimates) need to be communicated, minimizing the risk of privacy. Meanwhile, different clients communicate with each other directly according to a carefully designed network structure without a central master. This greatly enhances the reliability of the entire algorithm. Those nice properties inspire us to carefully study the NGD method both theoretically and numerically. Theoretically, we start with a classical linear regression model. We find that both the learning rate and the network structure play significant roles in determining the NGD estimator's statistical efficiency. The resulting NGD estimator can be statistically as efficient as the global estimator, if the learning rate is sufficiently small and the network structure is well balanced, even if the data are distributed heterogeneously. Those interesting findings are then extended to general models and loss functions. Extensive numerical studies are presented to corroborate our theoretical findings. Classical deep learning models are also presented for illustration purpose.

Session 33: Statistics and Machine Learning

Reinforcement Learning via Nonparametric Smoothing in a Continuous-Time Stochastic Setting with Noisy Data

Authors: Shang Wu (Fudan University)

Abstract: Reinforcement learning was developed mainly for discrete-time Markov decision processes. We establish a novel learning approach based on temporal-difference and nonparametric smoothing to solve reinforcement learning problems in a continuous-time setting with noisy data, where the true model to learn is governed by an ordinary differential equation, and data samples are generated from a stochastic differential equation that is considered as a noisy version of the ordinary differential equation. Continuous-time temporal-difference learning developed for deterministic models is unstable and in fact diverges when applied to data generated from stochastic models. Furthermore, because there are measurement errors or noises in the observed data, a new reinforcement learning framework is needed to handle the learning problems with noisy data. We show that the proposed learning approach has a robust performance for learning deterministic functions based on noisy data generated from stochastic models governed by stochastic differential equations. An asymptotic theory is established for the proposed approach, and a numerical study is carried out to solve a pendulum reinforcement learning problem and check the finite sample performance of the proposed method.

A General Framework for Treatment Effect Estimation in Semi-Supervised and High Dimensional Settings

Authors: Abhishek Chakrabortty (Texas A&M University)
Guorong Dai (Fudan University)
Eric Tchetgen Tchetgen (University of Pennsylvania)

Abstract: This work provides a general and complete understanding of semi-supervised (SS) causal inference for treatment effects. As prototype examples, we consider estimation of the average treatment effect and the quantile treatment effect in an SS setting, where in addition to a labeled data set on a response, a treatment indicator and a set of possibly high dimensional covariates, one also has a much larger unlabeled data set available without the response. Using these two data sets, we develop a family of SS estimators that are more robust and more efficient than their supervised counterparts based on the labeled data only. We show our estimators can be asymptotically normal whenever the propensity score in the model is correctly specified, without requiring

specific forms of the nuisance functions involved. This property is generally unattainable without the help of the massive unlabeled data. Further, under correct specification of all the nuisance functions, our estimators achieve semi-parametric efficiency. As an illustration of the nuisance estimation, we consider inverse-probability-weighting type kernel smoothing estimators involving possibly unknown transformations of high dimensional covariate, establishing novel results on their uniform convergence rates. The advantages of our methods over their supervised counterparts are validated by comprehensive numerical studies.

High-Dimensional Dynamic Pricing under Non-Stationarity: Learning and Earning with Change-Point Detection

Authors: Feiyu Jiang (Fudan University)
Zifeng Zhao (University of Notre Dame)
Yi Yu (University of Warwick)
Xi Chen (New York University)

Abstract: We consider a high-dimensional dynamic pricing problem under non-stationarity, where a firm sells products to T sequentially arrived consumers that behave according to an unknown demand model that may change at unknown locations over time. The demand model is assumed to be a sparse high-dimensional generalized linear model(GLM), allowing for a feature vector that encodes products and consumer information. To achieve optimal revenue(i.e. least regret), the seller needs to learn and exploit the unknown GLM model while at the same time monitoring for potential change-points (CP). To tackle such a problem, we first design a novel penalized likelihood based online CP detection algorithm for high-dimensional GLM, which is the first in the CP literature that achieves optimal minimax localization error rate for high-dimensional GLM. A CP detection augmented dynamic pricing policy named CPDP is further proposed, which achieves a regret of order O(slog(Td)sqrt{MT}), where s is the sparsity level and M is the number of CP. Somewhat surprisingly, this regret order can be independent of the magnitude of the change size. A matching lower bound is further provided to show the optimality of CPDP (up to logarithmic factors). In particular, the optimality w.r.t. the number of CP M is the first in the dynamic pricing literature, and is achieved via a novel accelerated exploration mechanism. Extensive simulation experiments and a real data

application on online lending illustrate the efficiency and practical value of the proposed policy.

Change Point Inference for High-dimensional Linear Models

Authors: Bin Liu (Fudan University)
Xinsheng Zhang (Fudan University)
Yufeng Liu (The University of North Carolina at Chapel Hill)

Abstract: In this article, we consider simultaneous change point detection and identification for high dimensional linear models. For change point detection, given any subgroup of variables, we propose a new method for testing the homogeneity of corresponding regression coefficients across the observations. Under some regularity conditions, the proposed new testing procedure controls the type I error asymptotically and is powerful against sparse alternatives and enjoys certain optimality. For change point identification, using the de-biased lasso process, an ``argmax"-based change point estimator is proposed which is shown to be consistent. To further improve the estimation accuracy of change point estimators, a novel two-step refitting-based algorithm is proposed. Moreover, combining with the binary segmentation technique, we further extend our new method for detecting and identifying multiple change points. Extensive simulation studies and an application to the Alzheimer's disease data analysis justify the validity of our new method.

Session 34: Statistics and Machine Learning for Compleex Data

Dynamic Models for Correcting Numerical Model Outputs

Authors: Yewen Chen (Sun Yat-sen University)
Xiaohui Chang (Oregon State University)

Hui Huang (Sun Yat-sen University)

Abstract: Numerical air quality models are pivotal for the prediction and assessment of air pollution, but numerical model outputs may be systematically biased. Several dynamic models are proposed to correct large-scale raw model outputs using data from other sources, including readings collected at ground monitoring networks and weather outputs from other numerical models.

Transfer DAG learning

Authors: Mingyang Ren (The Chinese University of Hong Kong)
Junhui Wang (The Chinese University of Hong Kong)

Abstract: Directed acyclic graph (DAG) has been widely employed to represent directional causal relations among collected nodes. Yet, the available data in one single study is often limited due to high acquisition costs. It remains an open question how to pool together heterogeneous data from relevant studies for better DAG reconstruction in the target study. In this work, we first introduce a novel set of structural similarity measures for DAG, which are beyond the parameter similarity focused by the existing research, and then present a transfer DAG learning framework by effectively leveraging information from various auxiliary studies. Our theoretical analysis shows substantial improvement in terms of DAG reconstruction in the target study, even without any auxiliary study with overall similarity. This is in sharp contrast to most existing transfer learning methods. The advantage of the proposed transfer DAG learning method is also supported by extensive numerical experiments on both synthetic data and a multi-site brain functional connectivity network data.

Transfer Learning by Optimal Model Averaging for Censored Data

Authors: Baihua He (University of Science andTechnology of China)
Jiping Wang (Yale University)
Shuangge Ma (Yale University)
Lixing Zhu (Beijing Normal University at Zhuhai)

Xinyu Zhang (Chinese Academy of Sciences)

Abstract: Transfer learning has gained significant attention in various domains, addressing the challenge of limited individual study data for prediction. In this paper, we develop a transfer learning approach with model averaging to predict censored responses in the main model. Specifically, several helper models are formulated with shared parameters from other datasets, and the optimal weights for the averaging procedure are derived by minimizing a delete-one cross-validation criterion. The proposed transfer learning allows the model framework to vary among helper models. We show that the proposed approach achieves the lowest prediction risk asymptotically when the main model is misspecified and attains model weight consistency when the main model is correctly specified. We further demonstrate that the risk of the proposed approach is no larger than the risks of the equal weighting approach and the pure model selection asymptotically, regardless of the correctness of the main model. We conduct extensive numerical studies to demonstrate the superior performances of the proposed procedure over the other existing methods and further show this using the Surveillance, Epidemiology, and End Results (SEER)-Medicare liver cancer data.

Latent Group Detection in Partially Functional Linear Regression Models

Authors: Wu Wang (University of Science andTechnology of China)
Ying Sun (King Abdullah University of Science and Technology)
Huixia Judy Wang (The George Washington University)

Abstract: We propose a functional partially linear regression model with latent group structures to accommodate the heterogeneous relationship between a scalar response and functional covariates. The proposed model is motivated by a salinity tolerance study of barley families, whose main objective is to detect salinity tolerant barley plants. Our model is flexible, allowing for heterogeneous functional coefficients while being efficient by pooling information within a group for estimation. We develop an algorithm in the spirit of the K-means clustering to identify latent groups of the subjects under study. We establish the consistency of the proposed estimator, derive the convergence rate and the asymptotic distribution, and develop inference procedures. We show by simulation studies

that the proposed method has higher accuracy for recovering latent groups and for estimating the functional coefficients than existing methods. The analysis of the barley data shows that the proposed method can help identify groups of barley families with different salinity tolerant abilities.

Session 35: Statistics in Finance and Risk Management

Research on Nonlinear State-dependent Conditional Risk Premium Estimation

Authors: Xiangyu Cui (Shanghai University of Finance and Economics)
Guan Zheng (Shanghai University of Finance and Economics)
Shi Yun (East China Normal University)

Abstract: coefficient functions of the basis functions using a semiparametric estimation method. In the second step, we obtain the final estimates of the coefficient functions of the basis functions using a weighted least squares estimation method with certain structural transformations. We demonstrate the consistency and asymptotic normality of the estimated parameters. This model can be extended to estimate the nonlinear state-dependent conditional risk premium in complex situations such as unbalanced panel data.

HAR-Itô Models and High-dimensional HAR Modeling for High-frequency Data

Authors: Huiling Yuan (East China Normal University)

Abstract: It is an important task to model realized volatilities for high-frequency data in finance and economics and, as arguably the most popular model, the heterogeneous autoregressive (HAR) model has dominated the applications in this area. However, this model suffers from three drawbacks: (i.) its heterogeneous volatility components are linear combinations of daily realized volatilities with fixed weights, which limit its flexibility for different types of assets, (ii.) it is still unknown what is the high-frequency probabilistic

structure for this model, as well as many other HAR-type models in the literature, and (iii.) there is no high-dimensional inference tool for HAR modeling although it is common to encounter many assets in real applications. To overcome these drawbacks, this paper proposes a multilinear low-rank HAR model by using tensor techniques, where a data-driven method is adopted to automatically select the heterogeneous components. In addition, HAR- Itô models are introduced to interpret the corresponding high-frequency dynamics, as well as those of other HAR-type models. Moreover, non-asymptotic properties of the high-dimensional HAR modeling are established, and a projected gradient descent algorithm with theoretical justifications is suggested to search for estimates. Theoretical and computational properties of the proposed method are verified by simulation studies, and the necessity of using the data-driven method for heterogeneous components is illustrated in real data analysis.

Robust Backtests for Expected Shortfall

Authors: Zaichao Du (Fudan University)
Xuhui Wang (Shanghai Lixin)

Abstract: distribution, which itself is stronger than the assumption of the correct specification of ES. In this paper we aim to propose some simple and powerful unconditional and conditional backtests for ES that only require the reported VaR, ES and realized losses. Besides, we allow situations where the institutions change their portfolio weights. We establish the asymptotic properties of the tests, and investigate their finite sample performance.

Random Distortion Risk Measures

Authors: Xin Zang (Beijing Jiaotong University)
Fan Jiang (Peking University)
Chenxi Xia (Peking University)
Jingping Yang (Peking University)

Abstract: This paper presents a random risk measure, named as the random distortion risk measure. The random distortion risk measure is a generalization of the traditional deterministic distortion risk measure by randomizing the deterministic distortion function and the risk distribution respectively, where a stochastic distortion is introduced to

randomize the distortion function, and a sub-sigma-algebra is introduced for illustrating the influence of the known information on the risk distribution. Some theoretical properties of the random distortion risk measure are provided, such as normalization, conditional positive homogeneity, conditional comonotonic additivity, monotonicity in stochastic dominance order, and continuity from below, and method for specifying the stochastic distortion and the sub-sigma-algebra is provided. Based on some stochastic axioms, the representation theorem of the random distortion risk measure is proved. For considering the randomization of a given deterministic distortion risk measure, some families of random distortion risk measures are introduced with the stochastic distortions constructed from Poisson process, Brownian motion and Dirichlet process respectively, and numerical analysis is carried out for showing the influence of the stochastic distortion and the risk distribution by focusing on the sample mean, variance, skewness, kurtosis, and the tail behavior of the random distortion risk measures.

Session 36: Statistics in Pharmaceutical Industry

Be More Efficient? - Weighted Parametric Multiple Test Procedures in Group Sequential Design (WPGSD)

Authors: Keaven Anderson (Merck & Co., Inc.)
Zifang Guo (Merck & Co., Inc.)
Jing Zhao (Merck & Co., Inc.)
Yujie Zhao(Merck & Co., Inc.)
Qi Liu (Merck & Co., Inc.)
Linda Z. Sun (Merck & Co., Inc.)

Abstract: Contemporary clinical trials are getting more complex and usually with multiple primary objectives. Multiple primary objectives resulting in tests with known correlations include evaluating 1) multiple experimental treatment arms, 2) multiple populations, 3) the combination of multiple arms and multiple populations. Group sequential design (GSD) is widely used in such clinical trials in which correlated tests of multiple hypotheses are used. In this presentation, we extend the framework of the weighted parametric multiple test procedure from fixed designs with a single analysis per objective to a GSD setting where different objectives may be assessed at the same or different times, each in a group

sequential fashion. Pragmatic methods for design and analysis of weighted parametric group sequential design (WPGSD) under closed testing procedures are proposed to maintain the strong control of family-wise Type I error rate (FWER) when correlations between tests are incorporated. This results in the ability to relax testing bounds compared to designs not fully adjusting for known correlations, increasing power or allowing decreased sample size. We illustrate the proposed methods using clinical trial examples and conduct a simulation study to evaluate the operating characteristics.

Network and Covariate Adjusted Response-Adaptive Design for Binary Response

Authors: Hao Mei (Renmin University of China)
Jiaxin Xie (Renmin University of China)
Yichen Qin (University of Cincinnati)
Yang Li (Renmin University of China)

Abstract: Randomization is a distinguishing feature of clinical trials for unbiased assessment of treatment efficacy. With a growing demand for more flexible and efficient randomization schemes and motivated by the idea of adaptive design, in this article we propose the network and covariate adjusted response-adaptive design (NCRD) that can concurrently manage three challenges: 1) maximizing benefits of a trial by assigning more patients to the superior treatment group randomly; 2) balancing social network ties across treatment arms to eliminate potential network interference; and 3) ensuring balance of important covariates, such as age, gender, and other potential confounders. We conduct simulation with different network structures and a variety of parameter settings. It is observed that the NCRD outperforms four alternative randomization designs in solving the above-mentioned problems and has a comparable power for detecting true difference between treatment groups. In addition, we conduct real data analysis to implement the new design in two clinical trials. Compared to equal randomization (the original design utilized in the trials), the NCRD slightly increases power, largely increases the percentage of patients assigned to the better-performing group, and significantly improves network and covariate balances. It is also noted that the advantages of the NCRD are augmented when the sample size is small and the level of network interference is high. In summary, the proposed NCRD assists researchers in conducting clinical trials with high-quality and high-efficiency.

Unleash the Power of Data - Statistics in the Pharmaceutical Industry: Today and Tomorrow

Authors: Hong Tian(BeiGene USA, Inc.)

Abstract: Statistics is an important discipline to enable the fit-for-purpose evidence generation in the pharmaceutical industry. Through rigorous statistical training, professionals acquire the necessary mindset for effective decision-making processes. This presentation aims to offer a comprehensive overview of the diverse roles statisticians can assume throughout a product's life cycle in the pharmaceutical industry, utilizing real-life examples. Furthermore, it will emphasize the key competencies that need to be cultivated today to ensure continuous success in the future.

Immortal Time Bias in Real World Data Analysis

Authors: Yang Zhao (Nanjing Medical University)

Abstract: TBD

Session 37: Subgroup Analyses and Identification, Graphical Model and Differential Privacy

Best Subset Selection: Theory and Algorithm

Authors: Xueqin Wang (University of Science and Technology of China)

Abstract: Best subset selection has been one of the most famous classical problems in Statistical Science. It aims to find a small subset of predictors but still gives adequate prediction accuracy in a linear regression model, but it is computationally intractable. Many relaxed regularization approaches and approximate algorithms have been proposed. In this talk, we review the development of the best subset selection from both theoretical and algorithmic aspects, including the very recent mixed integer optimization(MIO) approach introduced by Bertsimas et al. (2016) and our newly developed sequencing-splicing-selecting (SSS) algorithm. We show that, for the first time, our algorithm has a stable solution: the oracle estimator of the true parameters with probability one under mild conditions. We also provide a C++ implementation of the algorithm using the Rcpp interface. Finally, we demonstrate through numerical experiments based on enormous simulations and real datasets that the new method has competitive performance compared to state-of-the-art schemes for best subset selection purposes in terms of statistical properties and computational efficiency.

Bayesian Variable Selection on Structured Logistic-Mixture Models for Subgroup Analysis

Authors: Juan Shen (Fudan University)

Abstract: Subgroup analysis has emerged as an important tool to identify unknown subgroup memberships in the presence of heterogeneity. However, majority of the existing work focused on the low dimensional scenario where only a few candidate variables are considered for modeling the subgroup membership while in reality covariates can be high dimensional such as in gene expression datasets. In this paper, we propose a structured mixture model along with a Bayesian variable selection approach for identifying predictive and prognostic variables separately in the high dimensional setting. We achieve selection of predictive and prognostic variables, and estimation of the treatment effect in the selected subgroup simultaneously. We establish theoretical properties by showing strong variable selection consistency of our proposed method and demonstrate its performance using simulation studies. We perform subgroup analysis of

an AIDS treatment data using the proposed method and identify prognostic and predictive variables associated with subgroups having differential treatment effects.

Chain Graph Models: Identifiability, Estimation and Asymptotics

Authors: Junhui Wang (Chinese University of Hong Kong)
Ruixuan Zhao (City University of Hong Kong)
Haoran Zhang (Chinese University of Hong Kong)

Abstract: In this talk, we consider a flexible chain graph (CG) model, which admits both undirected and directed edges in one graph and thus can encode much more diverse relations among objects. We first establish the identifiability conditions for the CG model through a low rank plus sparse matrix decomposition, where the sparse matrix implies the sparse undirected edges within each chain component and the low rank matrix implies the presence of hub nodes with multiple children or parents. On this ground, we develop an efficient estimation method for reconstructing the CG structure, which first identifies the chain components via estimated undirected edges, determines the causal ordering of the chain components, and eventually estimates the directed edges among the chain components. Its theoretical properties will be discussed in terms of both asymptotic and finite-sample probability bounds on model estimation and graph reconstruction. The advantage of the proposed method is also demonstrated through extensive numerical experiments on both synthetic data and the Standard \& Poor's 500 index data.

Differentially Private Data Release for Mixed-type Data via Latent Factor Models

Authors: Yanqing Zhang (Yunnan University)
Qi Xu (University of California Irvine)
Niansheng Tang (Yunnan University)
Annie Qu(University of California Irvine)

Abstract: Differential privacy is a particular data privacy preserving technology which can publish synthetic data or statistical analysis with a minimum disclosure of private information of individual record. The tradeoff between privacy-preserving and utility

guarantee is always a challenge for differential privacy technology, especially for synthetic data generation. In this paper, we propose a differential private synthetic data algorithm for mixed-type data with correlation based on latent factor models. The proposed method can add a relatively small amount of noise to synthetic data under the same level of privacy protection while capturing correlation information.

Moreover, the proposed algorithm can generate synthetic data preserving the same data type as mixed-type original data, which greatly improves the utility of synthetic data. The key idea of our method is to partially perturb the factor matrix to construct a synthetic data generation model, and to utilize link functions to ensure consistency of synthetic data type with original data.The proposed method can generate privacy-preserving synthetic data at low computation cost even when the original data is high-dimensional. In theory, we establish differentially private properties of the proposed method. Our numerical studies also demonstrate superb performance of the proposed method on the utility guarantee of the privacy-preserving data released.

Session 38: Subsampling Methods for Massive Data Analysis

Distributed Logistic Regression for Massive Data with Rare Events

Authors: Xuetong Li (Peking University)
Zhuxue Ning(Fudan University)
Hansheng Wang(Peking University)

Abstract: Large-scale networks are commonly encountered in practice (e.g., Facebook and Twitter) by researchers. In order to study the network interaction between different nodes of large-scale networks, the spatial autoregressive (SAR) model has been popularly employed. Despite its popularity, the estimation of a SAR model on large-scale networks remains very challenging. On the one hand, due to policy limitations or high collection costs, it is often impossible for independent researchers to observe or collect all network information. On the other hand, even if the entire network is accessible, estimating the SAR model using the quasi-maximum likelihood estimator (QMLE) could be computationally infeasible due to its high computational cost. To address these challenges, we propose here a subnetwork estimation method based on QMLE for the SAR model. By using appropriate sampling methods, a subnetwork, consisting of a

much-reduced number of nodes, can be constructed. Subsequently, the standard QMLE can be computed by treating the sampled subnetwork as if it were the entire network. This leads to a significant reduction in information collection and model computation costs, which increases the practical feasibility of the effort. Theoretically, we show that the subnetwork-based QMLE is consistent and asymptotically normal under appropriate regularity conditions. Extensive simulation studies, bj

A Selective Introduction to Design-inspired Subsampling Methods

Authors: Jun Yu(Beijing institute of Technology)

Abstract: Subsampling focuses on selecting a subsample that can efficiently sketch the information of the original data in terms of statistical inference. It provides a powerful tool in big data analysis and gains the attention of data scientists in recent years. In this talk, some state-of-the-art subsampling methods inspired by statistical design will be introduced. Two important types of designs, namely optimal design, and space-filling design, have shown their great potential in subsampling for different objectives. The relationships between experimental designs and the related subsampling approaches will also be discussed.

A Sequential Addressing Subsampling Method for Massive Data Analysis under Memory Constraint

Authors: Yingqiu Zhu (University of International Business and Economics)
Rui Pan (Central University of Finance and Economics)
Baishan Guo(Meta AI)
Xuening Zhu(Fudan University)
Hansheng Wang(Peking University)

Abstract: The emergence of massive data in recent years brings challenges to automatic statistical inference. Data may be too numerous to be read into memory as a whole. Accordingly, new sampling techniques are needed to sample data from a hard drive. In this paper, we propose a sequential addressing subsampling (SAS) method that can sample data directly from the hard drive. The proposed SAS method is time saving in

terms of addressing cost compared to that of random addressing subsampling (RAS). Estimators based on the SAS subsamples are constructed, and their properties are studied.

Optimal Subsampling Bootstrap for Massive Data

Authors: Yingying Ma (Beihang Universiy)
Chenlei Leng(Warwick University)
Hansheng Wang(Peking University)

Abstract: The bootstrap is widely used for statistical inference. However, the vanilla version of bootstrap is no longer feasible computationally for modern massive dataset. Several improvements to the bootstrap method have been made, which assess the quality of estimators by subsampling the full dataset before resampling the subsamples. Naturally, the performance of these modern subsampling methods is influenced by tuning parameters. Our framework provides closed-form solutions for the optimal hyperparameter for subsampled bootstrap at no or little extra time cost. The results are promising.

第九届中国人民大学统计国际论坛