



浙江大学 数据科学研究中心  
Center for Data Science  
ZHEJIANG UNIVERSITY

**2023**

**Hangzhou International Conference on  
Frontiers of Data Science**

INNOVATION WITH HIGH DIMENSIONAL STATISTICS AND MACHINE LEARNING

August 20-22, 2023

**Organized by:** Center for Data Science, Zhejiang University  
**Operated by:** Hangzhou Qizhen Exhibition Service Co., Ltd





# Contents

|  |    |
|--|----|
| Instruction .....  | 1  |
| Program.....   | 2  |
| Plenary Lectures.....  | 4  |
| Abstracts .....  | 6  |
| Challenges in the analysis of big data <i>Zhezhen Jin</i> .....  | 6  |
| De novo Protein Design Based on Deep Learning <i>Ting Wei</i> .....  | 7  |
| Federated Learning of Multi-Institutional Electronic Health Records Data <i>Tianxi Cai</i> .....                               | 8  |
| On PC Adjustments for High-Dimensional Association Studies <i>Rajarshi Mukherjee</i> .....                                     | 9  |
| Sparse Kronecker Network Decomposition: A General Framework of Signal Region Detection<br><i>Long Feng</i> .....               | 10 |
| Two-way Node Popularity Model for Directed and Bipartite Networks <i>Ting Li</i> .....   | 11 |
| Inference on Potentially Identified Subgroups in Clinical Trials <i>Xinzhou Guo</i> .....                                      | 12 |
| Structure Learning via unstructured kernel-based M-estimation <i>Xin He</i> .....  | 13 |
| Distributionally Robust Machine Learning with Multi-source Data <i>Zhenyu Wang</i> .....                                       | 14 |
| Environment Invariant Linear Least Squares <i>Cong Fang</i> .....  | 15 |
| Introducing the specificity score: a measure of causality beyond P value <i>Wang Miao</i> .....                                | 16 |
| Parameter-Transfer Learning by Semiparametric Model Averaging <i>Xinyu Zhang</i> .....   | 17 |
| Malaria protection due to sickle haemoglobin depends on parasite genotype - but why? <i>Gavin Band</i><br>.....                | 18 |
| Hyperuniformity in urban systems <i>Lei Dong</i> .....   | 19 |
| Peacekeeping Loss Gradient: Assessing UN Peacekeeping Operations Effectiveness (1997-2020)<br><i>Andre Python</i> .....        | 20 |
| Ensemble methods for testing a global null <i>Yaowu Liu</i> .....  | 21 |
| Doubly Inhomogeneous Reinforcement Learning <i>Chengchun Shi</i> .....   | 22 |
| Identifying Temporal Pathways using Biomarkers in the Presence of Latent Non-Gaussian<br>Components <i>Shanghong Xie</i> ..... | 23 |
| Mediation analysis with the mediator and outcome missing not at random <i>Fan Yang</i> .....                                   | 24 |
| A Generative Approach to Learning a Conditional Distribution <i>Jian Huang</i> .....   | 25 |
| Optimal Policy Evaluation Using Kernel-Based Temporal Difference Methods <i>Yaqi Duan</i><br>.....                             | 26 |
| Statistical Learning and Matching <i>Xiaowu Dai</i> .....  | 27 |
| Post-Episodic Reinforcement Learning Inference <i>Ruohan Zhan</i> .....  | 28 |
| Uniform Inference for Nonlinear Endogenous Treatment Effects with High-Dimensional<br>Covariates <i>Ziwei Mei</i> .....        | 29 |
| Robust Instrumental Variable Regression in Genetics with GWAS Summary Data <i>Haoran Xue</i><br>.....                          | 30 |
| Strengthen causal inference with genetic data <i>Can Yang</i> .....  | 31 |

|  |                      |    |
|--|----------------------|----|
| Robust Inference for GMM with Possibly Nonsmooth Moments   | <i>Seojeong Lee</i>  | 32 |
| Analysis of microbiome compositions: Testing hypotheses on unobservable absolute abundance                                   | <i>Tao Wang</i>      | 33 |
| Semiparametric efficient estimation of genetic relatedness with machine learning methods                                     | <i>Xu Guo</i>        | 34 |
| Dimension Reduction for Extreme Regression via Contour Projection  | <i>Jing Zeng</i>     | 35 |
| Sparse causal mediation analysis with unmeasured mediator-outcome confounding  | <i>Wei Li</i>        | 36 |
| Design-based theory for cluster rerandomization  | <i>Hanzhong Liu</i>  | 37 |
| Long-term causal inference under persistent confounding via data combination   | <i>Xiaojie Mao</i>   | 38 |
| Identification and estimation of treatment effects on long-term outcomes in clinical trials with external observational data | <i>Peng Wu</i>       | 39 |
| Multi-task Additive Models for Robust Estimation and Automatic Structure Discovery   | <i>Yingjie Wang</i>  | 40 |
| Collaborative Learning under Non-stationary Heterogeneous Environments   | <i>Tao Lin</i>       | 41 |
| Towards Understanding the Generalization of Graph Neural Networks  | <i>Yong Liu</i>      | 42 |
| Statistical Inference for Segmented Regression Models  | <i>Han Yan</i>       | 43 |
| Statistical Inference for Maximin Effects: Identifying Stable Associations across Multiple Studies                           | <i>Zijian Guo</i>    | 44 |
| High-Dimensional Covariance Matrices Under Dynamic Volatility Models: Asymptotics and Shrinkage Estimation                   | <i>Xinghua Zheng</i> | 45 |
| Integrative conformal p-values for out-of-distribution testing with labeled outliers   | <i>Wenguang Sun</i>  | 46 |
| Continuous-Time Stochastic Setting with Noisy Data   | <i>Shang Wu</i>      | 47 |
| Robust and Efficient Case-Control Studies with Contaminated Case Pools: A Unified M-Estimation Framework                     | <i>Guorong Dai</i>   | 48 |
| Ranking and Selection in Large-Scale Inference of Heteroscedastic Units  | <i>Bowen Gang</i>    | 49 |
| Statistical Methods for Dynamic Risk Prediction with Longitudinal Risk Factors   | <i>Lihui Zhao</i>    | 50 |
| Multiple Instance Learning for Ordinal Outcomes  | <i>Menggang Yu</i>   | 51 |
| A dynamic RMST model supporting individual life expectancy prediction  | <i>Zheng Chen</i>    | 52 |
| Bootstrap Cross-Validations  | <i>Lu Tian</i>       | 53 |
| Optimal Clustering by Lloyd Algorithm for Low-Rank Mixture Model   | <i>Dong Xia</i>      | 54 |
| TBA  | <i>Cynthia Rush</i>  | 55 |
| Pseudo-Labeling for Kernel Ridge Regression under Covariate Shift  | <i>Kaizheng Wang</i> | 56 |
| A Sequential Addressing Subsampling Method for Massive Data Analysis under Memory Constraint                                 | <i>Yingqiu Zhu</i>   | 57 |
| Distributed Estimation and Inference for Spatial Autoregression Model with Large Scale Networks                              | <i>Yimeng Ren</i>    | 58 |
| Statistical Analysis of Fixed Mini-Batch Gradient Descent Estimator  | <i>Haobo Qi</i>      | 59 |
| Network Gradient Descent Algorithm for Decentralized Federated Learning  | <i>Shuyuan Wu</i>    | 60 |

|  |                     |    |
|--|---------------------|----|
| Estimation error of gradient descent in deep regressions   | <i>Yuling Jiao</i>  | 61 |
| Distributed Semi-Supervised Sparse Statistical Inference   | <i>Xiaojun Mao</i>  | 62 |
| Two Phases of Scaling Laws for kNN Classifiers   | <i>Pengkun Yang</i> | 63 |
| Deep Nonlinear Sufficient Dimension Reduction  | <i>Zhou Yu</i>      | 64 |
| Controlling the False Discovery Rate in Structural Sparsity: Split Knockoffs                                 | <i>Yuan Yao</i>     | 65 |
| Deep Reinforcement Learning for Online Assortment Customization: A Data-Driven Approach                      | <i>Yao Wang</i>     | 66 |
| Misspecification Analysis of High-Dimensional Random Effects Models for Estimation of Signal-to-Noise Ratios | <i>Xiaodong Li</i>  | 67 |
| Semiparametric estimation for dynamic networks with shifted connecting intensities                           | <i>Shizhe Chen</i>  | 68 |
| Dynamic Modeling for Multivariate Functional and Longitudinal Data   | <i>Qixian Zhong</i> | 69 |
| Deep Nonparametric Inference for Conditional Hazard Function   | <i>Wen Su</i>       | 70 |
| A Semiparametric Gaussian Mixture Model for Chest CT-based 3D Blood Vessel Reconstruction                    | <i>Qianhan Zeng</i> | 71 |
| A Geometrical Model with Stochastic Error for Abnormal Motion Detection of Portal Crane Bucket Grab          | <i>Baichen Yu</i>   | 72 |
| Mixture Conditional Regression with Ultrahigh Dimensional Text Data for Estimating Extralegal Factor Effects | <i>Jiixin Shi</i>   | 73 |
| An Ensemble Deep Learning Model for Risk Stratification of Invasive Lung Adenocarcinoma Using Thin-Slice CT  | <i>Jing Zhou</i>    | 74 |
| An asymptotic analysis of random partition based minibatch momentum methods for linear regression models     | <i>Yuan Gao</i>     | 75 |
| Uniform design motivated basis selection methods for smoothing spline regression                             | <i>Jun Yu</i>       | 76 |
| Optimal Subsampling Bootstrap for Massive Data   | <i>Yingying Ma</i>  | 77 |
| Distributed Logistic Regression for Massive Data with Rare Events  | <i>Xuetong Li</i>   | 78 |
| Deep Nonparametric Inference for Conditional Hazard Function   | <i>Xingqiu Zhao</i> | 79 |
| Regularized t distribution: definition, properties and applications  | <i>Tiejun Tong</i>  | 80 |
| Bayesian Estimation of Rates of Transitions between Marital Statuses   | <i>Junni Zhang</i>  | 81 |
| Adaptive Distributed Learning with Privacy Preserving for Online Diagnosis Platform                          | <i>Shao-Bo Lin</i>  | 82 |
| Generalization Analysis of Triplet Learning via Algorithmic Stability  | <i>Jun Chen</i>     | 83 |
| Stochastic Gradient Methods: Stability and Implicit Regularization   | <i>Yunwen Lei</i>   | 84 |
| Venue Map  |                     | 85 |
| Transportation   |                     | 87 |
| Dinning  |                     | 88 |
| Index of authors   |                     | 89 |



# Instruction

## 2023 杭州数据科学前沿国际研讨会

2023 Hangzhou International Conference on Frontiers of Data Science

August 20 -- August 22, 2023;

Xixi Hotel, Hangzhou

杭州西溪宾馆

### Programming Committee

|              |                                 |
|--------------|---------------------------------|
| Jianfei Cai  | Monash University               |
| Tony Cai     | University of Pennsylvania      |
| Tianxi Cai   | Harvard University              |
| Lu Tian      | Stanford University             |
| Yazhen Wang  | University of Wisconsin-Madison |
| Ming Yuan    | Columbia University (Chair)     |
| Heping Zhang | Yale University                 |

### Local Organizing Committee

|              |                             |
|--------------|-----------------------------|
| Yifan Cui    | Zhejiang University (Chair) |
| Junhong Lin  | Zhejiang University         |
| Wei Luo      | Zhejiang University         |
| Andre Python | Zhejiang University         |
| Wenguang Sun | Zhejiang University         |

### Organizer:

Center for Data Science, Zhejiang University;

### Official Website:

<https://www.zjuyh.com/data2023/rb>

# Program

| Aug 21, Monday |  |  |  |  |
|----------------|--|--|--|--|
| 08:30-08:45    | Opening Ceremony   |  | Dongwan Hall (董湾厅)   | Chair: Yifan Cui   |
| 08:45-09:45    | Plenary Lecture<br>Martin Wainwright   | Title: Challenges with<br>Covariate Shift: What<br>is it and what to do it?                  | Dongwan Hall (董湾厅)   | Chair: Tony Cai  |
| 09:45-10:20    | Group Photo, Tea Break   |  |  |  |
| 10:20-12:00    | <b>Leveraging Large<br/>Scale Data For<br/>Discovery</b>                           | <b>Statistical Learning<br/>with Applications to<br/>Image and Network<br/>Data Analysis</b> | <b>Recent advances in<br/>multi-source<br/>learning</b>                  |  |
|                | Dongwan Hall (董湾厅)   | Meishu Hall (梅墅厅)  | Meizhu Hall (梅竹厅)  |  |
|                | Chair: Andre Python  | Chair: Long Feng   | Chair: Zijian Guo  |  |
|                | Organizer: Tianxi Cai  | Organizer: Long Feng   | Organizer: Zijian Guo  |  |
|                | <b>Speakers:</b><br>Zhezhen Jin<br>Ting Wei<br>Tianxi Cai<br>Rajarshi Mukherjee    | <b>Speakers:</b><br>Long Feng<br>Ting Li<br>Xinzhou Guo<br>Xin He                            | <b>Speakers:</b><br>Zhenyu Wang<br>Cong Fang<br>Wang Miao<br>Xinyu Zhang |  |
| 12:00-13:30    | Lunch  |  |  |  |
| 13:30-15:10    | <b>Applied statistics<br/>and machine<br/>learning to tackle<br/>global issues</b> | <b>New inferential<br/>tools for data<br/>science</b>  | <b>Statistical Learning</b>  | <b>Robust causal<br/>inference</b>   |
|                | Dongwan Hall A<br>(董湾厅 A)  | Dongwan Hall B<br>(董湾厅 B)  | Meishu Hall<br>(梅墅厅)   | Meizhu Hall<br>(梅竹厅)   |
|                | Chair: Andre Python  | Chair: Yifan Cui   | Chair: Dong Xia  | Chair: Mengchu Zheng   |
|                | Organizer: Andre Python  | Organizer: Yifan Cui   | Organizer: Ming Yuan   | Organizer: Zijian Guo  |
|                | <b>Speakers:</b><br>Gavin Band<br>Lei Dong<br>Andre Python                         | <b>Speakers:</b><br>Yaowu Liu<br>Chengchun Shi<br>Shanghong Xie<br>Fan Yang                  | <b>Speakers:</b><br>Jian Huang<br>Yaqi Duan<br>Xiaowu Dai<br>Ruohan Zhan | <b>Speakers:</b><br>Ziwei Mei<br>Haoran Xue<br>Can Yang<br>Seojeong Lee    |
| 15:10-15:30    | Tea Break  |  |  |  |
| 15:30-17:10    | <b>Statistical Analysis<br/>for Complex Data</b>                                   | <b>Recent<br/>developments in<br/>causal inference</b>                                       | <b>Machine learning<br/>theory</b>                                       | <b>New Inference Tools<br/>for Data Science<br/>Applications</b>           |
|                | Dongwan Hall A<br>(董湾厅 A)  | Dongwan Hall B<br>(董湾厅 B)  | Meishu Hall<br>(梅墅厅)   | Meizhu Hall<br>(梅竹厅)   |
|                | Chair: Wei Luo   | Chair: Yifan Cui   | Chair: Junhong Lin   | Chair: Mengchu Zheng   |
|                | Organizer: Wei Luo   | Organizer: Yifan Cui   | Organizer: Junhong Lin   | Organizer: Zijian Guo  |
|                | <b>Speakers:</b><br>Tao Wang<br>Xu Guo<br>Jing Zeng                                | <b>Speakers:</b><br>Wei Li<br>Hanzhong Liu<br>Xiaojie Mao<br>Peng Wu                         | <b>Speakers:</b><br>Yingjie Wang<br>Tao Lin<br>Yong Liu                  | <b>Speakers:</b><br>Han Yan<br>Zijian Guo<br>Xinghua Zheng<br>Wenguang Sun |
| 18:00          | Banquet: Dongwan Hall  |  |  |  |

| Aug 22, Tuesday |   |   |  |  |
|-----------------|---|---|--|--|
| 09:00-10:00     | Plenary Lecture<br>Song Xi Chen   | Title: Statistical Inference for Decentralized Federated Learning         | Dongwan Hall(董湾厅)  | Chair: Wenguang Sun  |
| 10:00-10:20     | Tea Break   |   |  |  |
| 10:20-12:00     | <b>Statistical Machine Learning</b>                                       | <b>The development and inference for complex prediction algorithm</b>     | <b>Statistics and Machine Learning II</b>                      |  |
|                 | Dongwan Hall (董湾厅)  | Meishu Hall (梅墅厅)   | Meizhu Hall (梅竹厅)  |  |
|                 | Chair: Shang Wu   | Chair: Lu Tian  | Chair: Xiaowu Dai  |  |
|                 | Organizer: Yazhen Wang  | Organizer: Lu Tian  | Organizer: Ming Yuan   |  |
|                 | <b>Speakers:</b><br>Shang Wu<br>Guorong Dai<br>Bowen Gang                 | <b>Speakers:</b><br>Lihui Zhao<br>Menggong Yu<br>Zheng Chen<br>Lu Tian    | <b>Speakers:</b><br>Dong Xia<br>Cynthia Rush<br>Kaizheng Wang  |  |
| 12:00-13:30     | Lunch   |   |  |  |
| 13:30-15:10     | <b>Distributed Methods and its Statistical Theory</b>                     | <b>Recent Developments in Machine Learning Research</b>                   | <b>Statistics and Machine Learning I</b>                       | <b>Modern statistical methods for longitudinal and survival data</b> |
|                 | Dongwan Hall A (董湾厅 A)  | Dongwan Hall B (董湾厅 B)  | Meishu Hall (梅墅厅)  | Meizhu Hall (梅竹厅)  |
|                 | Chair: Shuyuan Wu   | Chair: Wei Luo  | Chair: Yao Wang  | Chair: Yifan Cui   |
|                 | Organizer: Heping Zhang   | Organizer: Wei Luo  | Organizer: Ming Yuan   | Organizer: Yifan Cui   |
|                 | <b>Speakers:</b><br>Yingqiu Zhu<br>Yimeng Ren<br>Haobo Qi<br>Shuyuan Wu   | <b>Speakers:</b><br>Yuling Jiao<br>Xiaojun Mao<br>Pengkun Yang<br>Zhou Yu | <b>Speakers:</b><br>Yuan Yao<br>Yao Wang<br>Xiaodong Li        | <b>Speakers:</b><br>Shizhe Chen<br>Qixian Zhong<br>Wen Su            |
| 15:10-15:30     | Tea Break   |   |  |  |
| 15:30-17:10     | <b>Statistical Methods for Image and Text Data</b>                        | <b>Subsampling Methods for Massive Data Analysis</b>                      | <b>Applied statistics</b>                                      | <b>Advances in Statistical Machine Learning</b>                      |
|                 | Dongwan Hall A (董湾厅 A)  | Dongwan Hall B (董湾厅 B)  | Meishu Hall (梅墅厅)  | Meizhu Hall (梅竹厅)  |
|                 | Chair: Jing Zhou  | Chair: Xuotong Li   | Chair: Tiejun Tong   | Chair: Yao Wang  |
|                 | Organizer: Heping Zhang   | Organizer: Heping Zhang   | Organizer: Ming Yuan   | Organizer: Yao Wang  |
|                 | <b>Speakers:</b><br>Qianhan Zeng<br>Baichen Yu<br>Jiaxin Shi<br>Jing Zhou | <b>Speakers:</b><br>Yuan Gao<br>Jun Yu<br>Yingying Ma<br>Xuotong Li       | <b>Speakers:</b><br>Xingqiu Zhao<br>Tiejun Tong<br>Junni Zhang | <b>Speakers:</b><br>Shao-Bo Lin<br>Jun Chen<br>Yunwen Lei            |





**Martin Wainwright**

**Massachusetts Institute of Technology (MIT)**

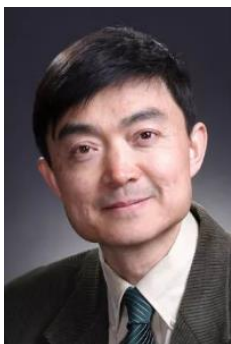
Martin Wainwright is the Cecil H. Green Professor in Electrical Engineering and Computer Science and Mathematics at MIT, and affiliated with the Laboratory for Information and Decision Systems and Statistics and Data Science Center.

Professor Wainwright is broadly interested in statistics, machine learning, information theory and optimization. He has received a number of awards and recognition including an Alfred P. Sloan Foundation Fellowship, best paper awards from the IEEE Signal Processing Society, the IEEE Communications Society, and the IEEE Information Theory and Communication Societies, the Medallion Lectureship and Award from the Institute of Mathematical Statistics, and the COPSS Presidents' Award from the Joint Statistical Societies. He was a Section Lecturer with the International Congress of Mathematicians in 2014 and received the Blackwell Award from the Institute of Mathematical Statistics in 2017. He has co-authored several books, including on graphical models with Michael Jordan, on sparse statistical modeling with Trevor Hastie and Rob Tibshirani, and a solo-authored book on high dimensional statistics.

### **Title: Challenges with Covariate Shift: What is it and what to do it?**

**Abstract:** A central problem in statistics and machine learning is prediction: learning how to predict responses based on observed features/covariates. Prediction is very well-understood under the canonical assumption of test and training distributions being identical. What if they differ, as has been documented in many applications? In this talk, we give an overview of some research that characterizes fundamental limits, and provides some computationally efficient algorithms for prediction under covariate shift.

Based on joint work with Cong Ma (Chicago), Reese Pathak (Berkeley/MIT), and Lin Xiao (Meta Research).



**Song Xi Chen**  
**Peking University**

Song Xi Chen (Chinese: 陈松蹊), Chair Professor of Peking University; Member of the Chinese Academy of Sciences; Fellow of the Institute of Mathematical Statistics, the American Statistical Association and the American Association for the Advancement of Science; Elected Member of the International Statistical Institute; Scientific Secretary of the Bernoulli Society.

Professor Chen is internationally renowned for his contributions to several prominent research areas, including inference for high dimensional data; environmental modeling and assessment; empirical likelihood; statistical and machine learning; inference for stochastic processes.

### **Title: Statistical Inference for Decentralized Federated Learning**

**Abstract:** This paper considers decentralized Federated Learning under heterogeneous distributions among distributed clients or data blocks {for the M-estimation}. The mean square error and consensus error across the estimators from different clients via the decentralized stochastic gradient descent algorithm is derived. The asymptotic normality of the Polyak-Ruppert averaged estimator in the decentralized distributed setting is attained, which shows that its statistical efficiency comes at a cost as it is more restrictive on the number of clients than that in the distributed M-estimation. To overcome the restriction, a one-step estimator is proposed which permits a much larger number of clients while still achieving the same efficiency as the original PR-averaged estimator {in the non-distributed setting}. The confidence regions based on both the PR-averaged estimator and the proposed one-step estimator are constructed to facilitate statistical inference for decentralized Federated Learning. The effect of the sparseness of the decentralized connection network on the statistical property of the one-step estimator is also derived.

A joint work with Jia Gu, a fifth year PhD student at PKU

# Abstracts

## Challenges in the analysis of big data

**Zhezhen Jin**

*Department of Biostatistics, Mailman School of Public Health, Columbia University*

*E-mail: zj7@cumc.columbia.edu*

Abstract: Analysis of large data is challenging due to its size and computational issues. In this talk, the issues and challenges will be discussed and subsampling methods will be presented. In particular, we will discuss a perturbation subsampling approach based on independent and identically distributed stochastic weights for the analysis of large data. We justify the method based on optimizing convex objective functions by establishing asymptotic consistency and normality for the resulting estimators.

# De novo Protein Design Based on Deep Learning

Ting Wei

*Shanghai Jiao Tong University*

*E-mail: weitinging@sjtu.edu.cn*

**Abstract:** Protein design is central to nearly all protein engineering problems, as it can enable the creation of proteins with new biological function, such as improving the catalytic efficiency of enzymes. One key facet of protein design, fixed-backbone protein sequence design, seeks to engineer new sequences that will conform to a prescribed protein backbone structure. Nonetheless, existing sequence design methods present limitations, such as reduced sequence diversity and shortcomings in experimental validation of the designed protein function, inadequacies that obstruct the goal of functional protein design. To overcome these hurdles, we initially developed the Graphormer-based Protein Design (GPD) model. This model deploys the Transformer on a graph-based representation of 3D protein structures of 31 thousands natural proteins and supplements it with Gaussian noise and a sequence random mask applied to node features, thereby enhancing sequence recovery and diversity. In order to boost experimental success rates, functional filtering strategies focusing on structure folding, solubility, and function were implemented. We executed the "sequence design - functional filtering - functional experiment" process on CalB hydrolase. The outcome of these experiments demonstrated a notable increase in the specific activity of the designed protein, improving by a factor of 1.7 compared to the CalB wild type.



# Federated Learning of Multi-Institutional Electronic Health Records Data

**Tianxi Cai**

*Department of Biomedical Informatics, Harvard University*

*E-mail : tcai@hsph.harvard.edu*

**Abstract:** The wide adoption of electronic health records (EHR) systems has led to the availability of large clinical datasets available for discovery research. EHR data, linked with bio-repository, is a valuable new source for deriving real-world, data-driven prediction models of disease risk and progression. Yet, they also bring analytical difficulties especially when aiming to leverage multi-institutional EHR data. Synthesizing information across healthcare systems is challenging due to heterogeneity and privacy. Statistical challenges also arise due to high dimensionality in the feature space. In this talk, I'll discuss analytical approaches for mining EHR data with a focus on transfer learning and federated learning. These methods will be illustrated using EHR data from Mass General Brigham and Veteran Health Administration.

# On PC Adjustments for High-Dimensional Association Studies

Rajarshi Mukherjee

*Harvard T.H. Chan School of Public Health*

*E-mail: ram521@mail.harvard.edu*

**Abstract:** We consider the effect of Principal Component (PC) adjustments while inferring the effects of variables on outcomes. This is motivated by the EIGENSTRAT procedure in genetic association studies where one performs PC adjustment to account for population stratification. We consider simple statistical models to obtain an asymptotically precise understanding of when such PC adjustments are supposed to work. We also verify these results through extensive numerical experiments. These results are based on joint work with Sohom Bhattacharya (Stanford University) and Rounak Dey (Harvard T.H.Chan School of Public Health).

# **Sparse Kronecker Network Decomposition: A General Framework of Signal Region Detection**

**Long Feng**

*University of Hong*

*E-mail: Konglfeng@hku.hk*

**Abstract:** This paper aims to present the first Frequentist framework on signal region detection in high-resolution and high-order image regression problems. Image data and scalar-on-image regression are intensively studied in recent years. However, most existing studies on such topics focussed on outcome prediction, while the research on region detection is rather limited, even though the latter is often more important. In this paper, we develop a general framework named Sparse Kronecker Product Decomposition (SKPD) to tackle this issue. The SKPD framework is general in the sense that it works for both matrices and tensors represented image data. Our framework includes one-term, multi-term, and nonlinear SKPDs. We propose nonconvex optimization problems for one-term and multi-term SKPDs and develop path-following algorithms for the nonconvex optimization. Under a Restricted Isometric Property, the computed solutions of the path-following algorithm are guaranteed to converge to the truth with a particularly chosen initialization even though the optimization is nonconvex. Moreover, the region detection consistency could also be guaranteed. The nonlinear SKPD is highly connected to shallow convolutional neural networks (CNN), particularly to CNN with one convolutional layer and one fully-connected layer. Effectiveness of SKPD is validated by real brain imaging data in the UK Biobank database.

# Two-way Node Popularity Model for Directed and Bipartite Networks

Ya Wang<sup>1</sup>, Ting Li<sup>2</sup>, Jiangzhou Wang<sup>3,\*</sup>, Bing-Yi Jing<sup>1,\*</sup>

<sup>1</sup> *Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen 518055, China*

<sup>2</sup> *Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong 999077, China*

<sup>3</sup> *College of Mathematics and Statistics, Institute of Statistical Sciences, Shenzhen University, Shenzhen 518060, China*  
*E-mail:tingeric.li@polyu.edu.hk*

**Abstract:** In recent years, there has been extensive research on community detection in directed and bipartite networks. However, these studies often fail to consider the popularity of nodes in different communities, which is a common phenomenon in real-world networks. To address this issue, we propose a new probabilistic framework called the Two-Way Node Popularity Model (TNPM). We also introduce the Rank-One Approximation Algorithm (ROA) for model fitting and community structure identification, and provide a comprehensive theoretical analysis. Additionally, we propose the Two-Stage Divided Cosine Algorithm (TSDC) to handle large-scale networks more efficiently. Our proposed methods offer multi-folded advantages in terms of estimation accuracy and computational efficiency, as demonstrated through extensive numerical studies. We apply our methods to two real-world applications, uncovering interesting findings.



# Inference on Potentially Identified Subgroups in Clinical Trials

Shuoxun Xu<sup>1</sup>, Xinzhou Guo<sup>1\*</sup>

*The Hong Kong University of Science and Technology*

*E-mail: xinzhoug@ust.hk*

**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.

# Structure Learning via unstructured kernel-based M-estimation

Xin He

*Shanghai University of Finance and Economics*

*E-mail: he.xin17@mail.shufe.edu.cn*

**Abstract:** In statistical learning, identifying underlying structures of true target functions based on observed data plays a crucial role to facilitate subsequent modeling and analysis. Unlike most of those existing methods that focus on some specific settings under certain model assumptions, this paper proposes a general and novel framework for recovering true structures of target functions by using unstructured M-estimation in a reproducing kernel Hilbert space (RKHS). The proposed framework is inspired by the fact that gradient functions can be employed as a valid tool to learn underlying structures, including sparse learning, interaction selection and model identification, and it is easy to implement by taking advantage of the nice properties of the RKHS. More importantly, it admits a wide range of loss functions, and thus includes many commonly used methods, such as mean regression, quantile regression, likelihood-based classification, and margin-based classification, which is also computationally efficient by solving convex optimization tasks. The asymptotic results of the proposed framework are established within a rich family of loss functions without any explicit model specifications. The superior performance of the proposed framework is also demonstrated by a variety of simulated examples and a real case study.

# Distributionally Robust Machine Learning with Multi-source Data

Zhenyu Wang<sup>1</sup>, Peter Buhlmann<sup>2</sup>, Zijian Guo<sup>1</sup>

<sup>1</sup>*Department of Statistics, Rutgers University, USA*

<sup>2</sup>*Seminar of Statistics, ETH Zurich, Switzerland*

*E-mail: zw425@stat.rutgers.edu*

*buhlmann@stat.math.ethz.ch*

*zijguo@stat.rutgers.edu*

**Abstract:** Classical machine learning methods may lead to poor prediction performance when the target distribution differs from the source populations. This paper utilizes data from multiple sources and introduces a group distributionally robust prediction which is defined to optimize {an} adversarial reward {about explained variance} {with respect to a class} of target distributions. Compared to classical empirical risk minimization, the proposed robust prediction model improves the prediction accuracy for target populations with distribution shifts. We show that the {our} {group} distributionally robust prediction model is identified as the weighted average of the source populations' conditional outcome models. We leverage this key identification result to robustify arbitrary machine learning algorithms, including, for example, random forests, boosting, and deep neural networks. In particular, we devise a novel bias-corrected estimator and establish its convergence rate. Our proposal can be seen as a distributionally robust federated learning approach that is computationally efficient, satisfies some privacy constraints, and has a nice interpretation of the importance of different sources for predicting {at a given target covariate distribution}. We demonstrate on simulated and real data the performance of our proposed group distributional robust method with random forests and deep neural networks as base-learning algorithms.

# Environment Invariant Linear Least Squares

Cong Fang

*Peking University*

*E-mail: fangcong@pku.edu.cn*

**Abstract:** We consider a multiple environment linear regression model, in which data from multiple experimental settings are collected. The joint distribution of the response variable and covariate may vary across different environments, yet the conditional expectation of  $y$  given the unknown set of important variables are invariant. Such a statistical model is related to the problem of endogeneity, transfer learning, and causal inference. We construct a novel environment invariant linear least squares (EILLS) objective function, a multiple-environment version of linear least squares that leverages the above conditional expectation invariance structure together with the heterogeneity among different environments to determine the true parameter. Our proposed method is applicable under the minimal structural assumption. We establish non-asymptotic error bounds on the estimation error for the EILLS estimator in the presence of endogenous variables. Moreover, we further show that the sparsity penalized EILLS estimator can achieve variable selection consistency in high-dimensional regimes. These non-asymptotic results demonstrate the sample efficiency of the EILLS estimator and its capability to circumvent the curse of endogeneity in an algorithmic manner without any prior structural knowledge.



# Introducing the specificity score: a measure of causality beyond P value

Wang Miao

*Peking University*

*E-mail: mwfy@pku.edu.cn*

**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.

# Parameter-Transfer Learning by Semiparametric Model Averaging

Xinyu Zhang

*Academy of Mathematics and Systems Science, Chinese Academy of Sciences*

*E-mail: xinyu@amss.ac.cn*

**Abstract:** In this article, we focus on prediction of a target model by transferring the information of source models. To be flexible, we use semiparametric additive frameworks for the target and source models. Inheriting the spirit of parameter-transfer learning, we assume that different models possibly share common knowledge across parametric components that is helpful for the target predictive task. Unlike existing parameter-transfer approaches, which need to construct auxiliary source models by parameter similarity with the target model and then adopt a regularization procedure, we propose a frequentist model averaging strategy with a J-fold cross-validation criterion so that auxiliary parameter information from different models can be adaptively transferred through data-driven weight assignments. The asymptotic optimality and weight convergence of our proposed method are built under some regularity conditions. Extensive numerical results demonstrate the superiority of the proposed method over competitive methods.

# **Malaria protection due to sickle haemoglobin depends on parasite genotype - but why?**

**Gavin Band**

*University of Oxford*

*E-mail: gavin.band@well.ox.ac.uk*

**Abstract:** Host genetic factors can confer resistance against malaria. Does this lead to evolutionary adaptation of parasite populations? In this talk, Dr. Band will introduce his recently-published work that found a strong association between sickle haemoglobin in the host and three regions of the parasite genome based on the analysis of Gambian and Kenyan children affected by severe malaria caused by *Plasmodium falciparum*.

# Hyperuniformity in urban systems

Lei Dong

*Peking University*

*E-mail: leidong@pku.edu.cn*

**Abstract:** Quantifying the spatial pattern of human settlements is fundamental to understanding the complexity of urban systems. By analyzing the spatial distributions of settlements in diverse regions, we find that urban systems encode a 'hyperuniform' order (exhibiting small density fluctuations), an intriguing pattern had been identified in physical and biological systems, but had rarely evidenced in socio-economic systems. We develop a simple model that shows how urban systems evolve into a hyperuniformity through matching growth and competition mechanisms. These results empirically and theoretically provide insights into the self-organization of cities, and reveal the hidden universality shared among social, physical, and biological systems.

# Peacekeeping Loss Gradient: Assessing UN Peacekeeping Operations Effectiveness (1997-2020)

Andre Python

*Zhejiang University*

*E-mail: apython@zju.edu.cn*

**Abstract:** Since 1948, the United Nations (UN) have deployed military troops and civilians from more than 120 countries to prevent and combat conflict across the world. Over the last decade, peacekeeping operations have been reduced, which has led to growing concerns about the impacts on world peace. To assess the effectiveness of UN peacekeeping operations on reducing conflict, we introduce peacekeeping loss gradient, a metric that considers proximity to peacekeeping operations, the associated troop size, and the number of operations. We apply machine learning algorithms to discover the relationship between peacekeeping loss gradient and conflict predicted a month ahead within half-degree grid cells across all countries that encountered peacekeeping operations. The results suggest a U-shape relationship between peacekeeping loss gradient and conflict, which appears invariant to various types of conflict. This brings evidence that the number of peacekeeping operations, the size of the deployed troops, and proximity to them may jointly contribute to reduce conflict risk. Our study provides a detailed description of the impact of the location and troop capacity of peacekeeping operations on promoting and maintaining peace worldwide.

# Ensemble methods for testing a global null

Yaowu Liu

*Southwestern University of Finance and Economics*

*E-mail: yaowuliu615@gmail.com*

**Abstract:** Testing a global null is a canonical problem in statistics and has a wide range of applications. In view of the fact that no uniformly most powerful test exists, prior and/or domain knowledge are commonly used to focus on a certain class of alternatives to improve the testing power. However, it is generally challenging to develop tests that are particularly powerful against a certain class of alternatives. In this paper, motivated by the success of ensemble learning methods for prediction or classification, we propose an ensemble framework for testing that mimics the spirit of random forests to deal with the challenges. Our ensemble testing framework aggregates a collection of weak base tests to form a final ensemble test that maintains strong and robust power. We apply the framework to four problems about global testing in different classes of alternatives arising from Whole Genome Sequencing (WGS) association studies. Specific ensemble tests are proposed for each of these problems, and their theoretical optimality is established in terms of Bahadur efficiency. Extensive simulations and an analysis of a real WGS dataset are conducted to demonstrate the type I error control and/or power gain of the proposed ensemble tests.

# Doubly Inhomogeneous Reinforcement Learning

Chengchun Shi

*London School of Economics and Political Science*

*E-mail: C.Shi7@lse.ac.uk*

**Abstract:** We study reinforcement learning (RL) in doubly inhomogeneous environments under temporal non-stationarity and subject heterogeneity. In a number of applications, it is commonplace to encounter datasets generated by system dynamics that may change over time and population, challenging high-quality sequential decision making. Nonetheless, most existing RL solutions require either temporal stationarity or subject homogeneity, which would result in sub-optimal policies if both assumptions were violated. To address both challenges simultaneously, we propose an original algorithm to determine the "best data chunks" that display similar dynamics over time and across individuals for policy learning, which alternates between most recent change point detection and cluster identification. Our method is general, and works with a wide range of clustering and change point detection algorithms. It is multiply robust in the sense that it takes multiple initial estimators as input and only requires one of them to be consistent. Moreover, by borrowing information over time and population, it allows us to detect weaker signals and has better convergence properties when compared to applying the clustering algorithm per time or the change point detection algorithm per subject. Empirically, we demonstrate the usefulness of our method through extensive simulations and a real data application.

# Identifying Temporal Pathways using Biomarkers in the Presence of Latent Non-Gaussian Components

Shanghong Xie

*Southwestern University of Finance and Economics*

*E-mail: shanghongxie@gmail.com*

**Abstract:** Time series data collected from a network of random variables are useful for identifying temporal pathways among the network nodes. Observed measurements may contain multiple sources of signals and noises, including Gaussian signals of interest and non-Gaussian noises including artifacts, structured noise, and other unobserved factors (e.g., genetic risk factors, disease susceptibility). Existing methods including vector autoregression (VAR) and dynamic causal modeling do not account for unobserved non-Gaussian components. Furthermore, existing methods cannot effectively distinguish contemporaneous relationships from temporal relations. In this work, we propose a novel method to identify latent temporal pathways using time series biomarker data collected from multiple subjects. The model adjusts for the non-Gaussian components and separates the temporal network from the contemporaneous network. Specifically, an independent component analysis (ICA) is used to extract the unobserved non-Gaussian components, and residuals are used to estimate the contemporaneous and temporal networks among the node variables based on method of moments. The algorithm is fast and can easily scale up. We derive the identifiability and the asymptotic properties of the temporal and contemporaneous networks. We demonstrate superior performance of our method by extensive simulations and an application to a study of attention-deficit/hyperactivity disorder (ADHD), where we analyze the temporal relationships between brain regional biomarkers. We find that temporal network edges were across different brain regions, while most contemporaneous network edges were bilateral between the same regions and belong to a subset of the functional connectivity network.



# Mediation analysis with the mediator and outcome missing not at random

Fan Yang

*Tsinghua University*

*E-mail: yangfan1987@tsinghua.edu.cn*

**Abstract:** Mediation analysis is widely used for investigating direct and indirect causal pathways through which an effect arises. However, many mediation analysis studies are challenged by missingness in the mediator and outcome. In general, when the mediator and outcome are missing not at random, the direct and indirect effects are not identifiable without further assumptions. In this work, we study the identifiability of the direct and indirect effects under some interpretable missing not at random mechanisms. We evaluate the performance of statistical inference under those assumptions through simulation studies and illustrate the proposed methods via the National Job Corps Study.

# A Generative Approach to Learning a Conditional Distribution

**Jian Huang**

*Department of Applied Mathematics, The Hong Kong Polytechnic University*

*E-mail: j.huang@polyu.edu.hk*

**Abstract:** Conditional distribution is a fundamental quantity in statistics and machine learning that provides a full description of the relationship between a response and a predictor. There is a vast literature on conditional density estimation. A common feature of the existing methods is that they seek to estimate the functional form of the conditional density. We propose a deep generative approach to learning a conditional distribution by estimating a conditional generator, so that a random sample from the target conditional distribution can be obtained by transforming a sample from a simple reference distribution. The conditional generator is estimated nonparametrically using neural networks by matching appropriate joint distributions. There are several advantages of the proposed generative approach over the classical methods for conditional density estimation, including: (a) there is no restriction on the dimensionality of the response or the predictor, (b) it can handle both continuous and discrete type predictors and responses, and (c) it is easy to obtain estimates of the summary measures of the underlying conditional distribution. We show that the proposed conditional learning approach can mitigate the curse of dimensionality under a low-dimensional data support assumption. We also conduct extensive numerical experiments to validate the proposed method and using several benchmark datasets, including the California housing, MNIST, and CelebA datasets, to illustrate its applications in conditional sample generation, visualization of multivariate data, conformal prediction, image generation and image reconstruction.

# Optimal Policy Evaluation Using Kernel-Based Temporal Difference Methods

Yaqi Duan

*New York University*

*E-mail: yaqid22@gmail.com*

**Abstract:** We study non-parametric methods for estimating the value function of an infinite-horizon discounted Markov reward process (MRP). We establish non-asymptotic bounds on the statistical error of a kernel-based least-squares temporal difference (LSTD) estimate, which can be understood either as a non-parametric instrumental variables method, or as a projected approximation to the Bellman fixed point equation. Our analysis imposes no assumptions on the transition operator of the Markov chain, but rather only conditions on the reward function and population-level kernel LSTD solutions. Using empirical process theory and concentration inequalities, we establish a non-asymptotic upper bound on the error with explicit dependence on the effective horizon  $H = (1 - \gamma)^{-1}$  of the Markov reward process, the eigenvalues of the associated kernel operator, as well as the instance-dependent

variance of the Bellman residual error. In addition, we prove minimax lower bounds over sub-classes of MRPs, which shows that our rate is optimal in terms of the sample size  $n$  and the effective horizon  $H$ . Whereas existing worst-case theory predicts cubic scaling ( $H^3$ ) in the effective horizon, our theory reveals that there is in fact a much wider range of scalings, depending on the kernel, the stationary distribution, and the variance of the Bellman residual error. Notably, it is only parametric and near-parametric problems that can ever achieve the worst-case cubic scaling.

# Statistical Learning and Matching

**Xiaowu Dai**

*Department of Statistics and Data Science and Department of Biostatistics, UCLA*

*E-mail: dai@stat.ucla.edu*

**Abstract:** We study the problem of decision-making in the setting of a scarcity of shared resources when the preferences of agents are unknown a priori and must be learned from data. Taking the two-sided matching market as a running example, we focus on the decentralized setting, where agents do not share their learned preferences with a central authority. Our approach is based on the representation of preferences in a reproducing kernel Hilbert space, and a learning algorithm for preferences that accounts for uncertainty due to the competition among the agents in the market. Under regularity conditions, we show that our estimator of preferences converges at a minimax optimal rate. Given this result, we derive optimal strategies that maximize agents' expected payoffs and we calibrate the uncertain state by taking opportunity costs into account. We also derive an incentive-compatibility property and show that the outcome from the learned strategies has a stability property. Finally, we prove a fairness property that asserts that there exists no justified envy according to the learned strategies.

# Post-Episodic Reinforcement Learning Inference

Ruohan Zhan

*Hong Kong University of Science and Technology*

*E-mail: rhzhan@ust.hk*

**Abstract:** We consider estimation and inference with data collected from episodic reinforcement learning (RL) algorithms, i.e. adaptive experimentation algorithms that at each period (aka episode) interact multiple times in a sequential manner with a single treated unit. Our goal is to be able to evaluate counterfactual adaptive policies after data collection and to estimate structural parameters such as dynamic treatment effects, which can be used for credit assignment (e.g., what was the effect of the first period action on the final outcome). Such parameters of interest can be framed as solutions to moment equations, but not minimizers of a population loss function, leading to Z-estimation approaches in the case of static data. However, such estimators fail to be asymptotically normal in the case of adaptive data collection. We propose a re-weighted Z-estimation approach with carefully designed adaptive weights to stabilize the episode-varying estimation variance, which results from the nonstationary policy that typical episodic RL algorithms invoke. We identify proper weighting schemes to restore the consistency and asymptotic normality of the re-weighted Z-estimators for target parameters, which allows for hypothesis testing and constructing uniform confidence regions for target parameters of interest. Primary applications include dynamic treatment effect estimation and dynamic off-policy evaluation. This is joint work with Vasilis Syrgkanis.

# Uniform Inference for Nonlinear Endogenous Treatment Effects with High-Dimensional Covariates

Qingliang Fan<sup>1</sup>, Zijian Guo<sup>2</sup>, Ziwei Mei<sup>1</sup>, and Cun-Hui Zhang<sup>2</sup>

<sup>1</sup>*Department of Economics, The Chinese University of Hong Kong*

<sup>2</sup>*Department of Statistics, Rutgers University*

*E-mails: michaelqfangmail.com (Q.Fan)*

*zijguostat.rutgers.edu (Z.Guo)*

*zwmei@link.cuhk.edu.hk (Z.Mei).*

*czhang@stat.rutgers.edu (C-H.Zhang)*

**Abstract:** Nonlinearity and endogeneity are common in empirical studies with observational data. This paper proposes new estimation and inference procedures for nonparametric treatment effect functions with potentially high-dimensional covariates. One innovation of this paper is the uniform confidence band of the marginal effect function defined as the derivative of the nonlinear treatment function, which is essential in the policy-relevant decision-making process. The asymptotic honesty of the confidence band is verified in theory. Simulation studies and an empirical study of air pollution and migration show the validity of our procedures.

# Robust Instrumental Variable Regression in Genetics with GWAS Summary Data

**Haoran Xue**

*City University of Hong Kong*

*E-mail: xuexx268@umn.edu*

**Abstract:** Instrumental variable regression is widely applied in genetics to discover putative causal factors for complex traits and diseases. Due to the widespread pleiotropy, genetic variants being used as IVs might be invalid, leading to false conclusions. We propose a robust and efficient inferential method to account for both hidden confounding and some invalid IVs via two-stage constrained maximum likelihood. We first develop the proposed method with individual-level data, then extend it both theoretically and computationally to GWAS summary data. Our numerical results demonstrate its wider applicability and superior finite-sample performance over the standard 2SLS.

# Strengthen causal inference with genetic data

Can Yang

*The Hong Kong University of Science and Technology*

*E-mail: macyang@ust.hk*

**Abstract:** 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 1) accounts for pleiotropy and sample structure simultaneously by leveraging genome-wide information; and 2) allows the inclusion of more genetic variants with moderate effects as instrument variables (IVs) 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, where the IV strengths tend to be relatively weak and existing summary-level MR methods for causal inference are vulnerable to confounding effects.



# Robust Inference for GMM with Possibly Nonsmooth Moments

Byunghoon Kang<sup>1</sup>, Seojeong Lee<sup>2</sup>

*<sup>1</sup>Lancaster University*

*Email: b.kang1@lancaster.ac.uk*

*<sup>2</sup>Seoul National University*

*Email: s.jay.lee@snu.ac.kr*

**Abstract:** This paper develops an asymptotic distribution theory for the Generalized Method of Moments (GMM) estimator when the moment condition is nonsmooth and misspecified. Our results extend existing theories for nonsmooth GMM to allow moment misspecification. Under misspecification, the conventional GMM variance estimators are inconsistent, and we show how to consistently estimate the true asymptotic variance for valid inference. Our results also extend the existing theories for the misspecified-GMM setup which assume the moment functions to be twice continuously differentiable. Detailed analyses of quantile regression with endogeneity under the location-scale model are provided to illustrate the application of the general results in the paper. Simulation evidence shows that our methods provide robust inference under misspecification.

# Analysis of microbiome compositions: Testing hypotheses on unobservable absolute abundance

Tao Wang

*Shanghai Jiao Tong University*  
*E-mail: neowangtao@sjtu.edu.cn*

**Abstract:** Analysis of compositions of microbiomes (ANCOM) compares the absolute abundances of microbes between two or more ecosystems using relative abundances in specimens derived from these ecosystems. Despite its impressive performance, there are two drawbacks to ANCOM. First, with  $K$  microbes it requires fitting  $K(K-1)/2$  models for log-ratios of counts, and so can be computationally intensive. Second, it does not output P-values for microbes detected as differentially abundant. We propose a fast implementation of ANCOM, fastANCOM, that fits only  $K$  models for log-transformed counts. fastANCOM provides P-values to declare statistical significance and outputs log fold changes of abundance between groups. fastANCOM compares favorably with existing differential abundance testing methods.

# Semiparametric efficient estimation of genetic relatedness with machine learning methods

**Xu Guo**

*School of Statistics, Beijing Normal University*

*E-mail: xustat12@bnu.edu.cn*

**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.

# Dimension Reduction for Extreme Regression via Contour Projection

LiuJun Chen<sup>1</sup>, Jing Zeng<sup>1</sup>

*International Institute of Finance, School of Management, University of Science and Technology of China, Hefei, China*

*E-mail: ljchen22@ustc.edu.cn, zengjxl@ustc.edu.cn*

**Abstract:** In a variety of regression problems, the central interest is to infer the extremes of the response given a set of predictors. The high dimensionality and heavy-tailedness of predictors are two primary sources hindering the application of the classical tools in inferring the conditional extremes. In this paper, we introduce the central extreme subspace (CES), whose existence and uniqueness are guaranteed under fairly mild conditions. By projecting the predictor onto the CES, all necessary information for inferring the conditional extremes is retained, which elegantly solves the high dimensionality issue. We project the elliptically-contoured predictor onto an elliptical contour and propose three estimators based on the contour-projected predictor to recover the CES, which are shown to be robust against the heavy-tailed data. Under mild assumptions, the proposed estimators achieve favorable consistency results. On the one hand, our proposal contributes a valuable addition to the toolkit of extreme regression. On the other hand, the proposed method expands the realm of dimension reduction. We further demonstrate the effectiveness of our proposal through extensive simulation studies and an application on the Chinese stock market data.

# Sparse causal mediation analysis with unmeasured mediator-outcome confounding

Kang Shuai<sup>1</sup>, Lan Liu<sup>2</sup>, Yangbo He<sup>1</sup>, and Wei Li<sup>3\*</sup>

<sup>1</sup>*School of Mathematical Sciences, Peking University*

<sup>2</sup>*School of Statistics, University of Minnesota at Twin City*

<sup>3</sup>*Center for Applied Statistics and School of Statistics, Renmin University of China*

*E-mail: weilistat@ruc.edu.cn*

**Abstract:** Causal mediation analysis aims to investigate how an intermediary factor, called a 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, these methods often assume there is no unmeasured confounding between mediators and the outcome. This paper allows for 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 propose a partially penalized procedure to select important mediators that have nonzero effects on the outcome. The resultant estimates are consistent, and the estimates of nonzero parameters are asymptotically normal. Simulation studies demonstrate advantageous performance of the proposed procedure over other existing methods. Finally, we apply our approach to genomic data and identify gene expressions that may actively mediate the effect of a genetic variant on mouse obesity.

# Design-based theory for cluster rerandomization

Hanzhong Liu

*Center for Statistical Science, Department of Industrial Engineering, Tsinghua University*

*E-mail: lhz2016@tsinghua.edu.cn*

**Joint work with Xin Lu, Tianle Liu and Peng Ding**

**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 refer to 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.

# Long-term causal inference under persistent confounding via data combination

**Xiaojie Mao**

*School of Economics and Management, Tsinghua University*

*E-mail: maoxj@sem.tsinghua.edu.cn*

**Abstract:** We study the identification and estimation of long-term treatment effects when both experimental and observational data are available. Since the long-term outcome is observed only after a long delay, it is not measured in the experimental data, but only recorded in the observational data. However, both types of data include observations of some short-term outcomes. In this paper, we uniquely tackle the challenge of persistent unmeasured confounders, i.e., some unmeasured confounders that can simultaneously affect the treatment, short-term outcomes and the long-term outcome, noting that they invalidate identification strategies in previous literature. To address this challenge, we exploit the sequential structure of multiple short-term outcomes, and develop three novel identification strategies for the average long-term treatment effect. We further propose three corresponding estimators and prove their asymptotic consistency and asymptotic normality. We finally apply our methods to estimate the effect of a job training program on long-term employment using semi-synthetic data. We numerically show that our proposals outperform existing methods that fail to handle persistent confounders.

# Identification and estimation of treatment effects on long-term outcomes in clinical trials with external observational data

Peng Wu

*Beijing Technology and Business University*

*E-mail: pengwu@btbu.edu.cn*

**Abstract:** In biomedical studies, estimating drug effects on chronic diseases requires a long follow-up period, which is difficult to meet in randomized clinical trials (RCTs). The use of a short-term surrogate to replace the long-term outcome for assessing the drug effect relies on stringent assumptions that empirical studies often fail to satisfy. Motivated by a kidney disease study, we investigate the drug effects on long-term outcomes by combining an RCT without observation of long-term outcomes and an observational study in which the long-term outcome is observed but unmeasured confounding may exist. Under a mean exchangeability assumption weaker than the previous literature, we identify the average treatment effects in the RCT and derive the associated efficient influence function and semiparametric efficiency bound. Furthermore, we propose a locally efficient doubly robust estimator and an inverse probability weighted (IPW) estimator. The former attains the semiparametric efficiency bound if all the working models are correctly specified, which may be hard to achieve due to the intertwined working models. While the latter has a simpler form and requires much fewer model specifications. The IPW estimator using estimated propensity scores is more efficient than that using true propensity scores and achieves the semiparametric efficient bound in the case of discrete covariates and surrogates with finite support. Both estimators are shown to be consistent and asymptotically normally distributed. Extensive simulations are conducted to evaluate the finite-sample performance of the proposed estimators. We apply the proposed methods to estimate the efficacy of oral hydroxychloroquine on renal failure in a real-world data analysis.



# Multi-task Additive Models for Robust Estimation and Automatic Structure Discovery

Yingjie Wang<sup>1</sup>, Hong Chen<sup>2</sup>

<sup>1</sup>*China University of Petroleum (East China);*

<sup>2</sup>*Huazhong Agricultural University*

*E-mails: yingjiawang@upc.edu.cn;*

*chenh@mail.hzau.edu.cn*

**Abstract:** Additive models have attracted much attention for high-dimensional regression estimation and variable selection. However, the existing models are usually limited to the single-task learning framework under the MSE criterion, where the utilization of variable structure depends heavily on a priori knowledge among variables. For high-dimensional observations in real environment, the learning performance of previous methods may be degraded seriously due to the complex non-Gaussian noise and the insufficiency of a prior knowledge on variable structure. To tackle this problem, we propose a new class of additive models, called Multi-task Additive Models (MAM), by integrating the mode-induced metric, the structure-based regularizer, and additive hypothesis spaces into a bilevel optimization framework. Our approach does not require priori knowledge of variable structure and suits for high-dimensional data with complex noise. Some theoretical analysis and empirical evaluations are established for the proposed MAM. We also report some recent related works, e.g., Huber additive models for time series analysis.

# Collaborative Learning under Non-stationary Heterogeneous Environments

**Tao Lin**

*School of Engineering, Westlake University*

*E-mail: lintao@westlake.edu.cn*

**Abstract:** Federated Learning is an emerging learning paradigm that collaboratively optimizes neural network models on the edge without sharing private data. Despite some recent research efforts on improving federated learning on heterogeneous data, learning and inference on time-evolving heterogeneous data in real-world scenarios have not been well studied. In this talk, I will briefly introduce our recent attempts in this area, including robust federated learning under diverse data distribution shifts and test-time robust personalization for inference.

# Towards Understanding the Generalization of Graph Neural Networks

Yong Liu

*Renmin University of China*

*E-mail: liuyonggsai@ruc.edu.cn*

**Abstract:** networks (GNNs) are the most widely adopted model in graph representation learning. Despite their extraordinary success in real-world applications, understanding their working mechanism by theory is still on primary stage. In this topic, we move towards this goal from the perspective of generalization. To be specific, we establish high probability bounds of generalization gap and gradients under transductive setting with consideration of stochastic optimization. The theoretical results reveal the architecture specific factors affecting the generalization gap. Experimental results on benchmark datasets show the consistency between theoretical results and empirical evidence. Our results provide new insights in understanding the generalization of GNNs.

# Statistical Inference for Segmented Regression Models

Han Yan

*Peking University*

*E-mail: hanyan@stu.pku.edu.cn*

**Abstract:** Segmented regression models are attractive for their flexibility and interpretability as compared to the global parametric and the nonparametric models, and yet are challenging in both estimation and inference. We consider a four-regime segmented model for temporally dependent data with two segmenting boundaries depending on multivariate covariates with non-diminishing boundary effects. A mixed integer quadratic programming algorithm is formulated to facilitate the least square estimation to both the regression and the boundary coefficients. The rates of convergence and the asymptotic distributions of the least square estimators are obtained, which show differential convergence rates and limiting distributions between the regression and the boundary coefficients. Estimation and testing for degenerated segmented models with less than four segments are also considered with a testing procedure to decide if neighboring segments can be merged. Numerical simulations and a case study on air pollution in Beijing are conducted to demonstrate the proposed model and the inference results. In particular, it shows that the segmented models with three or four regimes are suitable for the modeling of the meteorological effects on the  $PM_{2.5}$  concentration.

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

Zijian Guo

*Rutgers University*

*E-mail:zijguo@stat.rutgers.edu*

**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, *Bernoulli*, 18(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.

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

Xinghua Zheng

*HKUST*

*E-mail: xzheng@ust.hk*

**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 covariance 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.

# Integrative conformal p-values for out-of-distribution testing with labeled outliers

Wenguang Sun

*Zhejiang University*

*E-mail: wgsun@zju.edu.cn*

**Abstract:** We present novel conformal inference methods for out-of-distribution testing that leverage side information from labeled outliers, which are commonly underutilized or even discarded by conventional conformal p-values. Blending inductive and transductive conformal inference strategies in a principled way, our methods are computationally efficient and can automatically take advantage of the most powerful model from a collection of one-class and binary classifiers. Then, we study how to control the false discovery rate in multiple testing with a conditional calibration strategy. Simulations with synthetic and real data show the proposed integrative conformal p-values outperforms existing methods.

# Continuous-Time Stochastic Setting with Noisy Data

Shang Wu

*Fudan University*

*E-mail: shangwu@fudan.edu.cn*

**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.



# Robust and Efficient Case-Control Studies with Contaminated Case Pools: A Unified M-Estimation Framework

Guorong Dai<sup>1</sup>, Jinbo Chen<sup>2</sup>

<sup>1</sup>*Department of Statistics and Data Science, School of Management, Fudan University*

<sup>2</sup>*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania*

*E-mails: guorongdai@fudan.edu.cn*

*jinboche@pennmedicine.upenn.edu*

**Abstract:** We consider a general M-estimation problem based on contaminated case-control data, including the primary and secondary analyses of case-control studies as special examples. The case pool contains ineligible patients who should be excluded from the study if known, but the true status of an individual in the case pool is unclear except in a small subset. Through imputing the possibly unobserved status variable with a function of all available relevant predictors, followed by an appropriate debiasing procedure, we exploit the whole sample to develop a family of robust and efficient estimators, eliminating bias from the case contamination. The imputation function can be constructed using any reasonable regression or machine learning approaches. Our estimators are always root-n-consistent and asymptotically normal regardless of the imputation function's limit. Further, we explore relaxation of requirements on the imputation function. We show even without any assumption on its convergence properties, our estimators are still root-n-consistent while asymptotic normality can be achieved by a sample-splitting variant. Results of this type, which are entirely free of convergence assumptions on the nuisance estimators, can be pursued for other problems involving nuisance functions as well, so our analysis approach should be of independent interest. The finite-sample superiority of our method is demonstrated by comprehensive simulation studies. We also apply our method to analyze sepsis-related death based on a real data set from electronic health records.

# Ranking and Selection in Large-Scale Inference of Heteroscedastic Units

**Bowen Gang**

*Fudan University*

*E-mail: bgang@fudan.edu.cn*

**Abstract:** The allocation of limited resources to a large number of potential candidates presents a pervasive challenge. In the context of ranking and selecting top candidates from heteroscedastic units, conventional methods often result in over-representations of subpopulations, and this issue is further exacerbated in large-scale settings where thousands of candidates are considered simultaneously. To address this challenge, we propose a new multiple comparison framework that incorporates a modified power notion to prioritize the selection of important effects and employs a novel ranking metric to assess the relative importance of units. We develop both oracle and data-driven algorithms, and demonstrate their effectiveness in controlling the error rates and achieving optimality. We evaluate the numerical performance of our proposed method using simulated and real data. The results show that our framework enables a more balanced selection of effects that are both statistically significant and practically important, and results in an objective and relevant ranking scheme that is well-suited to practical scenarios.

# Statistical Methods for Dynamic Risk Prediction with Longitudinal Risk Factors

Lihui Zhao

*Northwestern University*

*E-mail: lihui.zhao@northwestern.edu*

**Abstract:** Cardiovascular disease (CVD) is a leading cause of morbidity and mortality. CVD risk prediction plays a central role in clinical CVD prevention strategies, by aiding decision making for lifestyle modification and to match the intensity of therapy to the absolute risk of a given patient. Various CVD risk factors have been identified and used to construct multivariate risk prediction algorithms. However, these algorithms are generally based on the risk factors measured at a single time. Since risk factors like blood pressure are regularly collected in clinical practice, and electronic medical records are making longitudinal data on these risk factors available to clinicians, dynamic prediction of CVD risk on a real-time basis using the history of CV risk factors will likely improve the precision of personalized CVD risk prediction. We will present statistical methods to build dynamic CVD risk prediction models using repeated measured risk factor levels. The pooled data from multiple community-based CVD cohorts will be used for illustration.

# Multiple Instance Learning for Ordinal Outcomes

Menggang Yu

*University of Wisconsin – Madison*

*E-mail: menggang.yu@gmail.com*

**Abstract:** The multiple instance learning (MIL) literature, where instances of data are naturally grouped into collections (“bags”) with a label, but where the labels on the instances themselves are unobserved, has primarily focused on binary classification. We propose a new SVM-based approach for learning from ordinal MIL data called Ordinal Multiple Instance Support Vector Machines (OMI-SVM). The method is motivated by a new assumption that the instance with the highest ordinal label defines the bag class. This “max-ordinal” assumption has justifiable connections to several applications. We also propose a deep learning approach. We explore many combinations of popular and applicable approaches through a large statistical experiment designed to detect their performance across data sets from several problem applications. From leveraging the findings of this experiment, we gain additional insight into the motivating breast cancer biomarker application.

# A dynamic RMST model supporting individual life expectancy prediction

Zheng Chen<sup>1\*</sup>, Zijing Yang<sup>1</sup>, Chengfeng Zhang<sup>1</sup>, Yawen Hou<sup>2</sup>

<sup>1</sup>*Department of Biostatistics, School of Public Health, Southern Medical University, Guangzhou, China;*

<sup>2</sup>*Department of Statistics and Data Science, School of Economics, Jinan University, Guangzhou, China*

*E-mail: zchen@smu.edu.cn*

**Abstract:** In clinical follow-up studies with a time-to-event end point, the difference in the restricted mean survival time (RMST) is a suitable substitute for the hazard ratio (HR). However, the RMST only measures the survival of patients over a period of time from the baseline and cannot reflect changes in life expectancy over time. Based on the RMST, we study the conditional restricted mean survival time (cRMST) by estimating life expectancy in the future according to the time that patients have survived, reflecting the dynamic survival status of patients during follow-up. In this paper, we introduce the estimation method of cRMST, the statistical inference concerning the difference between two cRMSTs (cRMSTd), and the establishment of the robust dynamic prediction model using the landmark method. Simulation studies are conducted to evaluate the statistical properties of these methods. The results indicate that the estimation of the cRMST is accurate, and the dynamic RMST model has high accuracy in coefficient estimation and good predictive performance. In addition, an example of patients with chronic kidney disease who received renal transplantations is employed to illustrate that the dynamic RMST model can predict patients' expected survival times from any prediction time, considering the time-dependent covariates and time-varying effects of covariates.

# Bootstrap Cross-Validations

Lu Tian

*Stanford University*

*E-mail: lutian@stanford.edu*

**Abstract:** Cross-validation is a widely used technique for evaluating the performance of prediction models. It helps avoid the optimism bias in error estimates, which can be significant for models built using complex statistical learning algorithms. However, since the cross-validation estimate is a random value dependent on observed data, it is essential to accurately quantify the uncertainty associated with the estimate. This is especially important when comparing the performance of two models using cross-validation, as one must determine whether differences in error estimates are a result of chance fluctuations. Although various methods have been developed for making inferences on cross-validation estimates, they often have many limitations, such as stringent model assumptions. This paper proposes a fast bootstrap method that quickly estimates the standard error of the cross-validation estimate and produces valid confidence intervals for a population parameter measuring average model performance. Our method overcomes the computational challenge inherent in bootstrapping the cross-validation estimate by estimating the variance component using a random effects model. It is just as flexible as the cross-validation procedure itself. To showcase the effectiveness of our approach, we employ comprehensive simulations and real data analysis across three diverse applications.

# Optimal Clustering by Lloyd Algorithm for Low-Rank Mixture Model

Dong Xia

*Hong Kong University of Science and Technology*

*E-mail: madxia@ust.hk*

**Abstract:** We investigate the computational and statistical limits in clustering matrix-valued observations. We propose a low-rank mixture model (LrMM), adapted from the classical Gaussian mixture model (GMM) to treat matrix-valued observations, which assumes low-rankness for population center matrices. A computationally efficient clustering method is designed by integrating Lloyd algorithm and low-rank approximation. Once well-initialized, the algorithm converges fast and achieves an exponential-type clustering error rate that is minimax optimal. Meanwhile, we show that a tensor-based spectral method delivers a good initial clustering. Comparable to GMM, the minimax optimal clustering error rate is decided by the separation strength, i.e, the minimal distance between population center matrices. By exploiting low-rankness, the proposed algorithm is blessed with a weaker requirement on separation strength. Unlike GMM, however, the statistical and computational difficulty of LrMM is characterized by the signal strength, i.e, the smallest non-zero singular values of population center matrices. Evidences are provided showing that no polynomial-time algorithm is consistent if the signal strength is not strong enough, even though the separation strength is strong. The performance of our low-rank Lloyd algorithm is further demonstrated under sub-Gaussian noise. Intriguing differences between estimation and clustering under LrMM are discussed. The merits of low-rank Lloyd algorithm are confirmed by comprehensive simulation experiments. Finally, our method outperforms others in the literature on real-world datasets.

**TBA**

**Cynthia Rush**

*Columbia University*

*E-mail: cynthia.rush@columbia.edu*

**Abstract:** TBA



# Pseudo-Labeling for Kernel Ridge Regression under Covariate Shift

**Kaizheng Wang**

*Columbia University*

*E-mail:kw2934@columbia.edu*

**Abstract:** We develop and analyze a principled approach to kernel ridge regression under covariate shift. The goal is to learn a regression function with small mean squared error over a target distribution, based on unlabeled data from there and labeled data that may have a different feature distribution. We propose to split the labeled data into two subsets and conduct kernel ridge regression on them separately to obtain a collection of candidate models and an imputation model. We use the latter to fill the missing labels and then select the best candidate model accordingly. Our non-asymptotic excess risk bounds show that in quite general scenarios, our estimator adapts to the structure of the target distribution as well as the covariate shift. It achieves the minimax optimal error rate up to a logarithmic factor. The use of pseudo-labels in model selection does not have major negative impacts.

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

Rui Pan<sup>1</sup>, Yingqiu Zhu<sup>2\*</sup>, Baishan Guo<sup>3</sup>, Xuening Zhu<sup>4</sup> and Hansheng Wang<sup>5</sup>

<sup>1</sup> *School of Statistics and Mathematics, Central University of Finance and Economics,*

<sup>2</sup> *School of Statistics, University of International Business and Economics,*

<sup>3</sup> *Meta AI,*

<sup>4</sup> *School of Data Science, Fudan University,*

<sup>5</sup> *Guanghua School of Management, Peking University*

*E-mail: inqzhu@uibe.edu.cn*

**Abstract:** The emergence of massive data in recent years brings challenges to automatic statistical inference. This is particularly true if the data are 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 newly proposed SAS method is time saving in terms of addressing cost compared to that of the random addressing subsampling (RAS) method. Estimators (e.g., the sample mean) based on the SAS subsamples are constructed, and their properties are studied. We conduct a series of simulation studies to verify the finite sample performance of the proposed SAS estimators. The time cost is also compared between the SAS and RAS methods. An analysis of the airline data is presented for illustration purpose.

# Distributed Estimation and Inference for Spatial Autoregression Model with Large Scale Networks

Yimeng Ren

*Fudan University*

*E-mail: ymren22@m.fudan.edu.cn*

**Abstract:** The rapid growth of online network platforms generates large-scale network data and it poses great challenges for statistical analysis using the spatial autoregression (SAR) model. In this work, we develop a novel distributed estimation and statistical inference framework for the SAR model on a distributed system. We first propose a distributed network least squares approximation (DNLSA) method. This enables us to obtain a one-step estimator by taking a weighted average of local estimators on each worker. Afterwards, a refined two-step estimation is designed to further reduce the estimation bias. For statistical inference, we utilize a random projection method to reduce the expensive communication cost. Theoretically, we show the consistency and asymptotic normality of both the one-step and two-step estimators. In addition, we provide theoretical guarantee of the distributed statistical inference procedure. The theoretical findings and computational advantages are validated by several numerical simulations implemented on the Spark system. Lastly, an experiment on the Yelp dataset further illustrates the usefulness of the proposed methodology.

# Statistical Analysis of Fixed Mini-Batch Gradient Descent Estimator

Haobo Qi

*Beijing Normal University*  
*E-mail: qihaobo\_gsm@pku.edu.com*

**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

Shuyuan Wu<sup>1</sup>, Danyang Huang<sup>1</sup>, Hansheng Wang<sup>1</sup>

*Shanghai University of Finance and Economics*

*E-mail: shuyuan.w@pku.edu.cn*

**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.

# Estimation error of gradient descent in deep regressions

Yuling Jiao

*Wuhan University*

*E-mail:yulingjiaomath@whu.edu.cn*

**Abstract:** To achieve a theoretical understanding of deep learning, it is necessary to consider the approximation, generalization, and optimization errors. In recent years, there have been significant advancements in the literature regarding each or two of these errors. However, there have been few works that simultaneously analyze all three errors. This is due to the gap that exists between the optimization and generalization errors in over-parameterized regimes. In this work, we attempt to bridge this gap by establishing consistency between the outputs of gradient descent and the true regression function in the over-parameterized scenario. Our research offers a feasible perspective for a more comprehensive understanding of the theory behind deep learning.

# Distributed Semi-Supervised Sparse Statistical Inference

**Xiaojun Mao**

*Shanghai Jiao Tong University*

*E-mail: [maoxj@sjtu.edu.cn](mailto:maoxj@sjtu.edu.cn)*

**Abstract:** This paper is devoted to studying the semi-supervised sparse statistical inference in a distributed setup. An efficient multi-round distributed debiased estimator, which integrates both labeled and unlabelled data, is developed. We will show that the additional unlabeled data helps to improve the statistical rate of each round of iteration. Our approach offers tailored debiasing methods for  $M$ -estimation and generalized linear model according to the specific form of the loss function. Our method also applies to a non-smooth loss like absolute deviation loss. Furthermore, our algorithm is computationally efficient since it requires only one estimation of a high-dimensional inverse covariance matrix. We demonstrate the effectiveness of our method by presenting simulation studies and real data applications that highlight the benefits of incorporating unlabeled data.

# Two Phases of Scaling Laws for kNN Classifiers

Pengkun Yang

*Tsinghua University*

*E-mail: yangpengkun@tsinghua.edu.cn*

**Abstract:** A scaling law refers to the observation that the test performance of a model improves as the number of training data increases. A fast scaling law implies that one can solve machine learning problems by simply boosting the data and the model sizes. Yet, in many cases, the benefit of adding more data can be negligible. In this work, we study the rate of scaling laws of nearest neighbor classifiers. We show that a scaling law can have two phases: in the first phase, the generalization error depends polynomially on the data dimension and decreases fast; whereas in the second phase, the error depends exponentially on the data dimension and decreases slowly. Our analysis highlights the complexity of the data distribution in determining the generalization error. When the data distributes benignly, our result suggests that nearest neighbor classifier can achieve a generalization error that depends polynomially, instead of exponentially, on the data dimension.



# Deep Nonlinear Sufficient Dimension Reduction

Zhou Yu

*East China Normal University*

*E-mail: zyu@stat.ecnu.edu.cn*

**Abstract:** Linear sufficient dimension reduction, as exemplified by sliced inverse regression, has seen substantial development in the past thirty years. However, with the advent of more complex scenarios, nonlinear dimension reduction has become a more general topic that gains considerable interest recently. This article introduces a novel method for nonlinear sufficient dimension reduction, utilizing the generalized martingale difference divergence measure in conjunction with deep neural networks. The optimal solution of the objective function is shown to be unbiased at the general level of  $\sigma$ -fields. And two optimization schemes considered, based on the fascinating deep neural networks, exhibit higher efficiency and flexibility compared to the classical eigendecomposition of linear operators. Moreover, we systematically investigate the slow rate and fast rate for the estimation error based on advanced U-process theory. Remarkably, the fast rate is nearly minimax optimal. The effectiveness of the deep nonlinear sufficient dimension reduction methods is demonstrated through simulations and real data analysis.

# Controlling the False Discovery Rate in Structural Sparsity: Split Knockoffs

Yuan Yao

*HKUST*

*E-mail:yuany@ust.hk*

**Abstract:** Controlling the False Discovery Rate (FDR) in a variable selection procedure is critical for reproducible discoveries, which receives an extensive study in sparse linear models. However, it remains largely open in the scenarios where the sparsity constraint is not directly imposed on the parameters, but on a linear transformation of the parameters to be estimated. Examples include total variations, wavelet transforms, fused LASSO, and trend filtering, etc. In this paper, we propose a data adaptive FDR control in this transformational or structural sparsity setting, the Split Knockoff method. The proposed scheme exploits both variable and data splitting. The linear transformation constraint is relaxed to its Euclidean proximity in a lifted parameter space, yielding an orthogonal design for improved power and orthogonal Split Knockoff copies. To overcome the challenge that exchangeability fails due to the heterogeneous noise brought by the transformation, new inverse supermartingale structures are developed for provable FDR control. Simulation experiments show that the proposed methodology achieves desired FDR and power. An application to Alzheimer's Disease study is provided that atrophy brain regions and their abnormal connections can be discovered based on a structural Magnetic Resonance Imaging dataset (ADNI). This is a joint work with CAO, Yang and SUN, Xinwei.

# Deep Reinforcement Learning for Online Assortment Customization: A Data-Driven Approach

Yao Wang

*Xi'an Jiaotong University*  
*E-mail: yao.s.wang@gmail.com*

**Abstract:** When a retailer has limited inventory and is operating on a periodic selling schedule, it is important to have a variety of products available for each customer. To maximize revenue over the long term, an optimal assortment policy is required that takes into account the complex purchasing behaviors of customers whose arrival order and preferences are unknown. By analyzing historical customer arrival and transaction data, we propose a data-driven approach for dynamic assortment planning. To address the challenge of online assortment customization, we utilize a Markov decision process (MDP) framework and employ a model-free deep reinforcement learning (DRL) approach to learn a policy that is nearly optimal. Our method involves using a specialized deep learning model called Gated-DNN to create assortments while adhering to constraints, and a modified version of the Advantage Actor-Critic (A2C) algorithm to adjust the parameters of the Gated-DNN model. The updates to the model's parameters are done by interactions with a simulated environment built from historical sequences of customer arrivals. The feedback we receive from simulated customers can take any form and should match the historical transaction data as closely as possible to ensure the effectiveness of the policy we learn. To evaluate the effectiveness of our approach, we conduct simulations using both a synthetic data set generated with a pre-determined customer type distribution and ground-truth choice model, as well as a real-world data set. Our extensive experiments demonstrate that our approach produces significantly higher long-term revenue compared to existing methods and remains robust under various conditions. We also demonstrate that our approach can be easily adapted to a more general problem that includes reusable products, where customers return purchased items after a period of time.

# Misspecification Analysis of High-Dimensional Random Effects Models for Estimation of Signal-to-Noise Ratios

Xiaodong Li

*UC-Davis*

*E-mail:xdgli@ucdavis.edu*

**Abstract:** Estimation of signal-to-noise ratios and residual variances in high-dimensional linear models has various important applications including, e.g. heritability estimation in bioinformatics. One commonly used estimator, usually referred to as REML, is based on the likelihood of the random effects model, in which both the regression coefficients and the noise variables are respectively assumed to be i.i.d Gaussian random variables. In this paper, we aim to establish the consistency and asymptotic distribution of the REML estimator for the SNR, when the actual coefficient vector is fixed, and the actual noise is heteroscedastic and correlated, at the cost of assuming the entries of the design matrix are independent and skew-free. The asymptotic variance can be also consistently estimated when the noise is heteroscedastic but uncorrelated. Extensive numerical simulations illustrate our theoretical findings and also suggest some assumptions imposed in our theoretical results are likely to be further relaxed.

# Semiparametric estimation for dynamic networks with shifted connecting intensities

Shizhe Chen

*Department of Statistics, University of California, Davis*

*E-mail:szdchen@ucdavis.edu*

**Abstract:** Stochastic block models are widely used to analyze random networks, where nodes are clustered based on similar connecting probabilities. In many applications, the connecting intensities are subject to node-wise time shifts. Failing to account for the unknown time shifts may result in unidentifiability or misclustering. In this project, we propose a stochastic block model that incorporates the unknown time shifts in dynamic networks. We establish the conditions that guarantee the identifiability of cluster memberships of nodes and representative connecting intensities across clusters. Using methods for shape invariant models, we propose computationally efficient semiparametric estimation procedures to simultaneously estimate time shifts, cluster memberships, and connecting intensities. We illustrate the performance of the proposed procedures via extensive simulation experiments. We further apply the proposed method on a neural data set to reveal distinct roles of neurons during motor circuit maturation in zebrafish.

# Dynamic Modeling for Multivariate Functional and Longitudinal Data

Siteng Hao<sup>1,#</sup>, Qixian Zhong<sup>2,#</sup>, Shu-Chin Lin<sup>1</sup>, Jane-Ling Wang<sup>1,\*</sup>

<sup>1</sup>*Department of Statistics, University of California, Davis, CA 95616, USA*

<sup>2</sup>*Department of Statistics and Data Science, School of Economics and Wang Yanan*

<sup>#</sup>*Contributed equally*

*E-mail: janelwang@ucdavis.edu*

**Abstract:** Dynamic interactions among several stochastic processes are common in many scientific fields. It is crucial to model these interactions to understand the dynamic relationship of the corresponding multivariate processes with their derivatives and to improve predictions. In reality, full observations of the multivariate processes are not feasible as measurements can only be taken at discrete locations or time points, and often only sparingly and intermittently in longitudinal studies. This results in multivariate longitudinal data that are measured at different times for different subjects. We propose a time-dynamic model to handle multivariate longitudinal data by modeling the derivatives of multivariate processes using the values of these processes. Starting with a linear concurrent model, we develop methods to estimate the regression coefficient functions, which can accommodate irregularly measured longitudinal data that are possibly contaminated with noise. Our approach can also be applied to settings when the observational times are the same for all subjects. We establish the convergence rates of our estimators with phase transitions and further illustrate our model through a simulation study and a real data application.

# Deep Nonparametric Inference for Conditional Hazard Function

Wen Su<sup>1</sup>, Kin-Yat Liu<sup>2</sup>, Guosheng Yin<sup>1</sup>, Jian Huang<sup>3</sup>, Xingqiu Zhao<sup>3,\*</sup>

<sup>1</sup>*City University of Hong Kong*

<sup>2</sup>*The Chinese University of Hong Kong*

<sup>3</sup>*The Hong Kong Polytechnic University*

*Email: xingqiu.zhao@polyu.edu.hk*

**Abstract:** We propose a novel deep learning approach to nonparametric statistical inference for the conditional hazard function of survival time with right-censored data. We use a deep neural network (DNN) to approximate the logarithm of a conditional hazard function given covariates and obtain a DNN likelihood-based estimator of the conditional hazard function. Such an estimation approach grants model flexibility and hence relaxes structural and functional assumptions on conditional hazard or survival functions. We establish the consistency, convergence rate, and functional asymptotic normality of the proposed estimator. Subsequently, we develop new one-sample tests for goodness-of-fit evaluation and two-sample tests for treatment comparison. Both simulation studies and real application analysis show superior performances of the proposed estimators and tests in comparison with existing methods.

# A Semiparametric Gaussian Mixture Model for Chest CT-based 3D Blood Vessel Reconstruction

**Qianhan Zeng**

*Guanghua School of Management, Peking University*

*E-mail: helenology@stu.pku.edu.cn*

**Abstract:** Computed tomography (CT) has been a powerful diagnostic tool since its emergence in the 1970s. Its highly detailed and high-resolution results contribute significantly to medical screening, especially for early stage lung cancer detection. Based on CT data, the three-dimensional (3D) structures of human internal organs and tissues (e.g., blood vessels) can be reconstructed using professional software. 3D reconstruction is extremely beneficial for surgical operations and can serve as a vivid medical teaching example. However, traditional 3D reconstruction relies on manual operation by experienced surgeons, which is time-consuming, subjective, and requires substantial experience. To address this problem, we develop a novel semiparametric Gaussian mixture model for 3D blood vessel reconstruction. We theoretically extend the classical Gaussian mixture model by allowing both the component-wise mean and variance to nonparametrically vary according to voxel positions. 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 is then compared with the traditional cross-validation-based method. The regression method outperforms the cross-validation method in both computational and statistical efficiency. In application, the 3D structures of blood vessels are successfully reconstructed in a fully automatic manner.



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

Baichen Yu<sup>1</sup>, Xiao Wang<sup>1</sup>, Hansheng Wang<sup>1</sup>

*East China Normal University*  
*E-mail: baichen.yu@stu.ecnu.edu.cn*

**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 an alarm signal immediately. Numerical studies based on both simulated and real datasets are presented. We provide here a computer-vision based solution for the automatic detection of abnormal motion for portal cranes. Our method can be used to guarantee the day-to-day safe operations of portal cranes for transferring freight from the port to cargo ships, and vice versa.

# Mixture Conditional Regression with Ultrahigh Dimensional Text Data for Estimating Extralegal Factor Effects

Jiixin Shi<sup>1</sup>, Fang Wang<sup>2\*</sup>, Yuan Gao<sup>3</sup>, Xiaojun Song<sup>1</sup>, Hansheng Wang<sup>1</sup>

<sup>1</sup>*Guanghua School of Management, Peking University, Beijing, China*

<sup>2</sup>*Data Science Institute, Shandong University, Jinan, China*

<sup>3</sup>*School of Statistics and KLATASDS-MOE, East China Normal University, Shanghai, China*

*E-mail: jxshi0stat@gmail.com*

**Abstract:** Testing judicial impartiality is a problem of fundamental importance in empirical legal studies, for which standard regression methods have been popularly used to estimate the extralegal factor effects. However, those methods cannot handle control variables with ultrahigh dimensionality, such as found in judgment documents recorded in text format. To solve this problem, we develop a novel mixture conditional regression (MCR) approach, assuming that the whole sample can be classified into a number of latent classes. Within each latent class, a standard linear regression model can be used to model the relationship between the response and a key feature vector, which is assumed to be of a fixed dimension. Meanwhile, ultrahigh dimensional control variables are then used to determine the latent class membership, where a Naive Bayes type model is used to describe the relationship. Hence, the dimension of control variables is allowed to be arbitrarily high. A novel expectation-maximization algorithm is developed for model estimation. Therefore, we are able to estimate the interested key parameters as efficiently as if the true class membership were known in advance. Simulation studies are presented to demonstrate the proposed MCR method. A real dataset of Chinese burglary offenses is analyzed for illustration purpose.

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

Jing Zhou<sup>#1</sup>, Bin Hu<sup>#2</sup>, Wei Feng<sup>3</sup>, Zhang Zhang<sup>4</sup>, Xiaotong Fu<sup>1</sup>, Handie Shao<sup>1</sup>, Hansheng Wang<sup>5</sup>, Longyu Jin<sup>3</sup>, Siyuan Ai<sup>6</sup> and Ying Ji<sup>2\*</sup>

<sup>1</sup> Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China.

<sup>2</sup> Department of Thoracic Surgery, Beijing Institute of Respiratory Medicine and Beijing Chao-Yang Hospital, Capital Medical University, Beijing, China.

<sup>3</sup> Department of Cardiothoracic Surgery, The Third Xiangya Hospital of Central South University, Changsha, China

<sup>4</sup> Department of Thoracic Surgery, Changsha Central Hospital, Changsha, China.

<sup>5</sup> Guanghua School of Management, Peking University, Beijing, China.

<sup>6</sup> Department of Thoracic Surgery, Beijing LIANGXIANG Hospital, Beijing, China.

E-mail: zhoujing\_89@126.com

**Abstract:** Lung cancer screening using computed tomography (CT) has increased the detection rate of small pulmonary nodules and early-stage lung adenocarcinoma. It would be clinically meaningful to accurate assessment of the nodule histology by CT scans with advanced deep learning algorithms. However, recent studies mainly focus on predicting benign and malignant nodules, lacking of model for the risk stratification of invasive adenocarcinoma. We propose an ensemble multi-view 3D convolutional neural network (EMV-3D-CNN) model to study the risk stratification of lung adenocarcinoma. We include 1075 lung nodules ( $\leq 30$  mm and  $\geq 4$  mm) with preoperative thin-section CT scans and definite pathology confirmed by surgery. Our model achieves a state-of-art performance of 91.3% and 92.9% AUC for diagnosis of benign/malignant and pre-invasive/invasive nodules, respectively. Importantly, our model outperforms senior doctors in risk stratification of invasive adenocarcinoma with 77.6% accuracy [i.e., Grades 1, 2, 3]). It provides detailed predictive histological information for the surgical management of pulmonary nodules. Finally, for user-friendly access, the proposed model is implemented as a web-based system (<https://seeyourlung.com.cn>).

# An asymptotic analysis of random partition based minibatch momentum methods for linear regression models

Yuan Gao

*Guanghua School of Management, Peking University*

*E-mail: yuan\_gao96@126.com*

**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.

# Uniform design motivated basis selection methods for smoothing spline regression

Jun Yu

*Beijing Insititute of Technology*

*E-mail: yujunbeta@bit.edu.cn*

**Abstract:** Fitting a smoothing spline model on a large-scale dataset is daunting due to the high computational cost. The basis selection methods for smoothing spline calculation are regarded as an efficient way to deal with the large-scale dataset. The key to success is to force a non-parametric function in an infinite-dimensional functional space to reside in a relatively small and finite-dimensional model space without the loss of too much prediction accuracy. Space-filling basis selection is proven more efficient among various basis selection methods since the dimension of its model space is smaller than others. In this talk, we illustrate two efficient space-filling basis selection methods for smoothing spline calculation. The key idea is to adopt a uniform design to the large-scale dataset and use projective uniformity to improve the statistical efficiency when the underlying response surface is not isomorphic. It is proved that the illustrated estimator has the same convergence rate as the full-basis estimator. Compared with the standard approach, the proposed method significantly reduce the computational cost.

# Optimal Subsampling Bootstrap for Massive Data

Yingying Ma<sup>1</sup>, Chenlei Leng<sup>2</sup>, Hansheng Wang<sup>3</sup>

<sup>1</sup>*Beihang University*

<sup>2</sup>*University of Warwick*

<sup>3</sup>*Peking Univerisity*

*E-mail: mayingying@buaa.edu.cn,*

*C.Leng@warwick.ac.uk,*

*hansheng@pku.edu.cn*

**Abstract:** The bootstrap is a widely used procedure for statistical inference because of its simplicity and attractive statistical properties. However, the vanilla version of bootstrap is no longer feasible computationally for many modern massive datasets due to the need to repeatedly resample the entire data. Therefore, several improvements to the bootstrap method have been made in recent years, 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 such as the size of subsamples, the number of subsamples, and the number of resamples per subsample. In this article, we develop a novel hyperparameter selection methodology for selecting these tuning parameters. Formulated as an optimization problem to find the optimal value of some measure of accuracy of an estimator subject to computational cost, our framework provides closed-form solutions for the optimal hyperparameter values for subsampled bootstrap, subsampled double bootstrap and bag of little bootstraps, at no or little extra time cost. Using the mean square errors as a proxy of the accuracy measure, we apply our methodology to study, compare and improve the performance of these modern versions of bootstrap developed for massive data through numerical study. The results are promising.

# Distributed Logistic Regression for Massive Data with Rare Events

Xuetong Li

*Peking University*

*E-mail: 2001110929@stu.pku.edu.cn*

**Abstract:** Large-scale rare events data are commonly encountered in practice. To tackle the massive rare events data, we propose a novel distributed estimation method for logistic regression in a distributed system. For a distributed framework, we face the following two challenges. The first challenge is how to distribute the data. In this regard, two different distribution strategies (i.e., the RANDOM strategy and the COPY strategy) are investigated. The second challenge is how to select an appropriate type of log-likelihood function so that the best asymptotic efficiency can be achieved. Then, the under-sampled (US) and inverse probability weighted (IPW) types of log-likelihood functions are considered. Our results suggest that the COPY strategy together with the IPW log-likelihood function is the best solution for distributed logistic regression with rare events. The finite sample performance of the distributed methods is demonstrated by simulation studies and a real-world Swedish Traffic Signs dataset.

# Deep Nonparametric Inference for Conditional Hazard Function

Wen Su<sup>1</sup>, Kin-Yat Liu<sup>2</sup>, Guosheng Yin<sup>1</sup>, Jian Huang<sup>3</sup>, Xingqiu Zhao<sup>3,\*</sup>

<sup>1</sup>*City University of Hong Kong*

<sup>2</sup>*The Chinese University of Hong Kong*

<sup>3</sup>*The Hong Kong Polytechnic University*

*Email: xingqiu.zhao@polyu.edu.hk*

**Abstract:** We propose a novel deep learning approach to nonparametric statistical inference for the conditional hazard function of survival time with right-censored data. We use a deep neural network (DNN) to approximate the logarithm of a conditional hazard function given covariates and obtain a DNN likelihood-based estimator of the conditional hazard function. Such an estimation approach grants model flexibility and hence relaxes structural and functional assumptions on conditional hazard or survival functions. We establish the consistency, convergence rate, and functional asymptotic normality of the proposed estimator. Subsequently, we develop new one-sample tests for goodness-of-fit evaluation and two-sample tests for treatment comparison. Both simulation studies and real application analysis show superior performances of the proposed estimators and tests in comparison with existing methods.



# Regularized t distribution: definition, properties and applications

Tiejun Tong

*Department of Mathematics, Hong Kong Baptist University*

*E-mail: tongt@hkbu.edu.hk*

**Abstract:** For gene expression data analysis, an important task is to identify genes that are differentially expressed between two or more groups. Nevertheless, as biological experiments are often measured with a relatively small number of samples, how to accurately estimate the variances of gene expression becomes a challenging issue. To tackle this problem, we introduce a regularized t distribution and derive its statistical properties including the probability density function and the moment generating function.

The noncentral regularized t distribution is also introduced for computing the statistical power of hypothesis testing. For practical applications, we apply the regularized t distribution to establish the null distribution of the regularized statistic, and then formulate it as a regularized t-test for detecting the differentially expressed genes. Simulation studies and real data analysis show that our regularized t-test performs much better than the Bayesian t-test in the “limma” package, in particular when the sample sizes are small.

# Bayesian Estimation of Rates of Transitions between Marital Statuses

**Junni Zhang**

*National School of Development, Peking University*

*E-mail: zjn@nsd.pku.edu.cn*

**Abstract:** Transitions between marital statuses (e.g., first-time marriage, remarriage of divorced individuals, remarriage of widowed individuals, and divorce of married individuals) not only affect the proportions of groups with different marital statuses in the population, but also affect fertility rates and the size and structure of households. Therefore, estimating rates of transitions of marital statuses is important. In survey data on transitions between marital statuses, after cross-classification by age, year and other dimensions, sample sizes are often rather small, making direct estimation of transition rates unreliable. We use Bayesian methods to address this challenge, and illustrate our approach with Chinese and US data.

# Adaptive Distributed Learning with Privacy Preserving for Online Diagnosis Platform

Shao-Bo Lin

*Xi'an Jiaotong University*  
*E-mail: sblin1983@gmail.com*

**Abstract:** In this talk, we propose a novel adaptive distributed learning system based on divide-and-conquer and local average regression for prediction and privacy preservation simultaneously. Different from the classical distributed learning strategy whose algorithmic parameters and patterns are given by the central agent, our approach provides autonomy to each local agent in terms of parameter selection, algorithm designation and data perturbation. Such an adaptive manner significantly enhances the privacy preservation of the system. Our theoretical results demonstrate that the novel adaptive distributed learning system does not degrade the prediction performance of classical systems via presenting optimal learning rates in the framework of statistical learning theory. Our theoretical assertions are verified via numerous numerical experiments including both toy simulations and real data study. In our analysis, the new system also admits a certain perturbation of the test data via showing an almost comparable accuracy to that of the original data, which provides a realistic possibility for protecting privacy from both training and testing sides.

# Generalization Analysis of Triplet Learning via Algorithmic Stability

Jun Chen<sup>1</sup>, Hong Chen<sup>2</sup>

*Huazhong Agricultural University*

*E-mails: cj850487243@163.com;*

*chenh@mail.hzau.edu.cn*

**Abstract:** Triplet learning algorithms have shown competitive performance on individual-level fine-grained tasks, e.g. face recognition and person re-identification. However, their foundations of statistical learning theory are far less understanding. In this talk, we will introduce the stability-based generalization analysis for triplet learning implemented by stochastic gradient descent (SGD) and regularized risk minimization (RRM) respectively. In addition, we also discuss generalization bounds of federated zeroth-order (FedZO) algorithms under non-convex and heavy-tailed conditions.

# Stochastic Gradient Methods: Stability and Implicit Regularization

Yunwen Lei

*The University of Hong Kong*

*E-mail: leiyw@hku.hk*

**Abstract:** Stochastic gradient methods (SGM) such as SGD have found various applications in solving minimization and minimax optimization problems in machine learning. An important problem regarding SGM is how the models produced by SGM would generalize to testing examples. In this talk, I will present our recent results on the generalization analysis of SGM from the perspective of algorithmic stability. We introduce a new algorithmic stability concept to relax the existing restrictive assumptions and improve the existing generalization bounds. Our results show how the implicit regularization can be achieved by tuning the step size and the number of passes.

# Venue Map

## Hangzhou Xixi Hotel

(No.803 Wener West Road, Xihu District)

Located at northern part of Xixi Wetland, Hangzhou Xixi Hotel is a 30-minute cab drive from Hangzhou Railway Station and a 1-hour drive from Hangzhou Xiaoshan International Airport. Free Wi-Fi is available within the entire hotel.

Hangzhou Xixi Hotel All units here are equipped with an air conditioning, a flat-screen cable TV and an in-room safe. A mini bar, bottled water and an electric kettle are available. Guests can enjoy the stunning scenery from the window. Hangzhou Xixi Hotel Free toiletries, slippers and hair dryer are found in attached bathroom. A tour desk can arrange a sightseeing tour or rent a car for guests. Currency exchange, laundry and baggage storage are provided at 24-hour front desk.

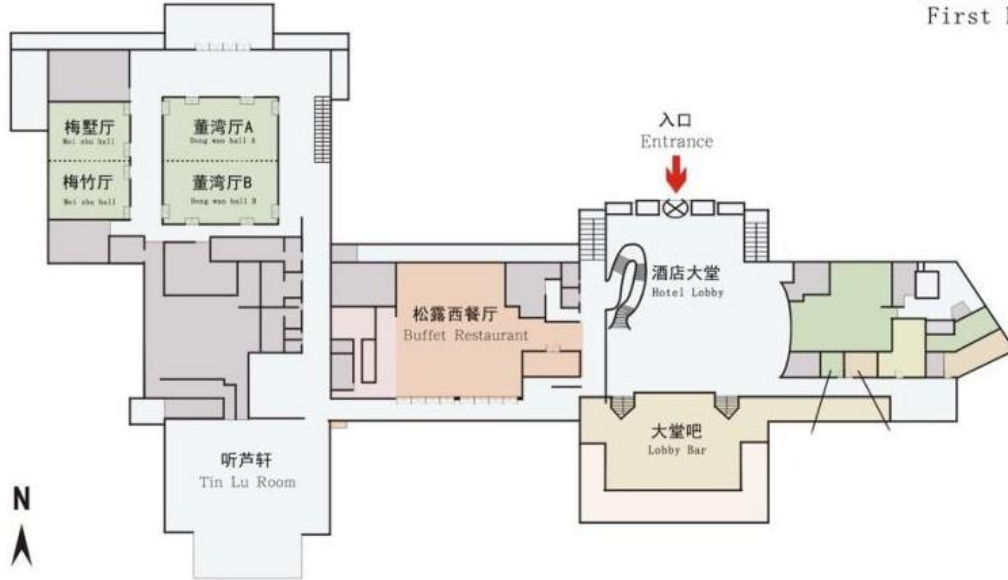


# 主楼平面图

## General Layout of Main Building

### 一楼示意图

First Floor



# Transportation


Hangzhou Xixi Hotel can be reached by taxi or public transport from Hangzhou East Railway Station, Hangzhou Railway Station and Hangzhou Xiaoshan International Airport. Details are as follows:




## Transportation options:

### Line one:


 Hangzhou East Railway Station -- Hangzhou Xixi Hotel

 Subway Line 19 → 297M | 17km | 41minutes

 17 km | 43 minutes | ¥ 66

### Line two:


 Hangzhou Railway Station -- Hangzhou Xixi Hotel

 Subway Line 5 → Subway Line 2 → 297M | 17.5 km | 1hour 2minutes

 18 km | 39minutes | ¥ 67

### Line three:

 Hangzhou Xiaoshan International Airport -- Hangzhou Xixi Hotel

 Subway Line 19 → 297M | 42 kilometers | 1 hour 25 minutes

 50.4 km | 1 hour 2 minutes | ¥ 167



## Dinning

| Date   | Time        | Meal    | Venue  |
|--------|-------------|---------|--|
| Aug 20 | 17:30-21:00 | Dinner  | Buffet Restaurant (Truffle)<br>松露西餐厅<br>1 <sup>st</sup> floor of the main building |
| Aug 21 | 12:00-13:30 | Lunch   | Buffet Restaurant (Truffle)<br>松露西餐厅<br>1 <sup>st</sup> floor of the main building |
| Aug 21 | 18:00-20:00 | Banquet | Dongwan Hall<br>董湾厅  |
| Aug 22 | 12:00-13:30 | Lunch   | Buffet Restaurant (Truffle)<br>松露西餐厅<br>1 <sup>st</sup> floor of the main building |



## Index of authors

| Name          | Affiliation  | E-mail                      | Page |
|---------------|--|-----------------------------|------|
| Andre Python  | Zhejiang University                                  | apython@zju.edu.cn          | 20   |
| Baichen Yu    | East China Normal University                         | baichen.yu@stu.ecnu.edu.cn  | 72   |
| Bowen Gang    | Fudan University                                     | bgang@fudan.edu.cn          | 49   |
| Can Yang      | Hong Kong University of Science and Technology       | macyang@ust.hk              | 31   |
| Chengchun Shi | The London School of Economics and Political Science | c.shi7@lse.ac.uk            | 22   |
| Cong Fang     | Peking University                                    | fangcong@pku.edu.cn         | 15   |
| Cynthia Rush  | Columbia University                                  | cynthia.rush@columbia.edu   | 55   |
| Dong Xia      | HKUST  | madxia@ust.hk               | 54   |
| Fan Yang      | Tsinghua University                                  | yangfan1987@tsinghua.edu.cn | 24   |
| Gavin Band    | University of Oxford                                 | gavin.band@well.ox.ac.uk    | 18   |
| Guorong Dai   | Fudan University                                     | guorongdai@fudan.edu.cn     | 48   |
| Han Yan       | Peking University                                    | hanyan@stu.pku.edu.cn       | 43   |
| Hanzhong Liu  | Tsinghua University                                  | lh2016@tsinghua.edu.cn      | 37   |
| Haobo Qi      | Beijing Normal University                            | qihaobo_gsm@pku.edu.cn      | 59   |
| Haoran Xue    | University of Minnesota                              | xuexx268@umn.edu            | 30   |
| Jian Huang    | The Hong Kong Polytechnic University                 | j.huang@polyu.edu.hk        | 25   |
| Jiixin Shi    | Sun Yat-Sen University                               | shijx9@mail2.sysu.edu.cn    | 73   |
| Jing Zeng     | University of Science and Technology of China        | zengjxl@ustc.edu.cn         | 35   |
| Jing Zhou     | Renmin University of China                           | zhoujing_89@126.com         | 74   |
| Jun Chen      | Huazhong Agricultural University                     | cj850487243@163.com         | 83   |
| Jun Yu        | Beijing Institute of Technology                      | yujunbeta@bit.edu.cn        | 76   |
| Junni Zhang   | Peking University                                    | zjn@nsd.pku.edu.cn          | 81   |
| Kaizheng Wang | Columbia University                                  | kw2934@columbia.edu         | 56   |
| Lei Dong      | Peking University                                    | leidong@pku.edu.cn          | 19   |
| Lihui Zhao    | Northwestern University                              | lihui.zhao@northwestern.edu | 50   |
| Long Feng     | University of Hong Kong                              | lfeng@hku.hk                | 10   |

| Name               | Affiliation                                      | E-mail                      | Page |
|--------------------|--|-----------------------------|------|
| Lu Tian            | Stanford University                              | lutian@stanford.edu         | 53   |
| Menggang Yu        | University of Wisconsin at Madison               | meyu@biostat.wisc.edu       | 51   |
| Peng Wu            | Beijing Technology and Business University       | pengwu@btbu.edu.cn          | 39   |
| Pengkun Yang       | Tsinghua University                              | yangpengkun@tsinghua.edu.cn | 63   |
| Qianhan Zeng       | Peking University                                | helenology@stu.pku.edu.cn   | 71   |
| Qixian Zhong       | Xiamen University                                | qxzhong@xmu.edu.cn          | 69   |
| Rajarshi Mukherjee | Harvard T.H. Chan School of Public Health        | ram521@mail.harvard.edu     | 9    |
| Ruohan Zhan        | Hong Kong University of Science and Technology   | rhzhan@ust.hk               | 28   |
| Seojeong Lee       | Seoul National University                        | s.jay.lee@snu.ac.kr         | 32   |
| Shang Wu           | Fudan University                                 | shangwu@fudan.edu.cn        | 47   |
| Shanghong Xie      | Southwestern University of Finance and Economics | xiesh@swufe.edu.cn          | 23   |
| Shao-Bo Lin        | Xi'an Jiaotong University                        | sblin1983@gmail.com         | 82   |
| Shizhe Chen        | UC-Davis   | szdchen@ucdavis.edu         | 68   |
| Shuyuan Wu         | Shanghai University of Finance and Economics     | shuyuan.w@pku.edu.cn        | 60   |
| Tao Lin            | Xihu University                                  | lintao@westlake.edu.cn      | 41   |
| Tao Wang           | Shanghai Jiao Tong University                    | neowangtao@sjtu.edu.cn      | 33   |
| Tianxi Cai         | Harvard University                               | tcai@hsph.harvard.edu       | 8    |
| Tiejun Tong        | HKBU   | tongt@hkbu.edu.hk           | 80   |
| Ting Li            | The Hong Kong Polytechnic University             | tingeric.li@polyu.edu.hk    | 11   |
| Ting Wei           | Shanghai Jiao Tong University                    | weitinging@sjtu.edu.cn      | 7    |
| Wang Miao          | Peking University                                | mwfy@pku.edu.cn             | 16   |
| Wei Li             | Renmin University of China                       | weilistat@ruc.edu.cn        | 36   |
| Wen Su             | City University of Hong Kong                     | w.su@cityu.edu.hk           | 70   |
| Wenguang Sun       | Zhejiang University                              | wgsun@zju.edu.cn            | 46   |
| Xiaodong Li        | UC-Davis   | xdgli@ucdavis.edu           | 67   |
| Xiaojie Mao        | Tsinghua University                              | maoxj@sem.tsinghua.edu.cn   | 38   |
| Xiaojun Mao        | Jiao Tong University                             | maoxj@sjtu.edu.cn           | 62   |
| Xiaowu Dai         | UCLA   | daixiaowu0925@gmail.com     | 27   |

| Name          | Affiliation   | E-mail                     | Page |
|---------------|---|----------------------------|------|
| Xin He        | Shanghai University of Finance and Economics                            | he.xin17@mail.shufe.edu.cn | 13   |
| Xinghua Zheng | HKUST   | xhzheng@ust.hk             | 45   |
| Xingqiu Zhao  | The Hong Kong Polytechnic University                                    | xingqiu.zhao@polyu.edu.hk  | 79   |
| Xinyu Zhang   | Academy of Mathematics and Systems Science, Chinese Academy of Sciences | xinyu@amss.ac.cn           | 17   |
| Xinzhou Guo   | Hong Kong University of Science and Technology                          | xinzhoug@ust.hk            | 12   |
| Xu Guo        | Beijing Normal University   | liushengjunyi@163.com      | 34   |
| Xuetong Li    | Peking University   | 2001110929@stu.pku.edu.cn  | 78   |
| Yao Wang      | Xi'an Jiaotong University   | yao.s.wang@gmail.com       | 66   |
| Yaowu Liu     | Southwestern University of Finance and Economics                        | liuyw@swufe.edu.cn         | 21   |
| Yaqi Duan     | New York University   | yaqid22@gmail.com          | 26   |
| Yimeng Ren    | Fudan University  | ymren22@m.fudan.edu.cn     | 58   |
| Yingjie Wang  | China University of Petroleum (East China)                              | yingjiawang@upc.edu.cn     | 40   |
| Yingqiu Zhu   | University of International Business and Economics                      | inqzhu@uibe.edu.cn         | 57   |
| Yingying Ma   | Beihang University  | mayingying_11@163.com      | 77   |
| Yong Liu      | Renmin University of China  | liuyonggsai@ruc.edu.cn     | 42   |
| Yuan Gao      | Peking University   | yuan_gao96@126.com         | 75   |
| Yuan Yao      | HKUST   | yuany@ust.hk               | 65   |
| Yuling Jiao   | Wuhan University  | yulingjiaomath@whu.edu.cn  | 61   |
| Yunwen Lei    | The University of Hong Kong   | leiyw@hku.hk               | 84   |
| Zheng Chen    | Southern Medical University   | zchen@smu.edu.cn           | 52   |
| Zhenyu Wang   | Rutgers University  | zw425@stat.rutgers.edu     | 14   |
| Zhezhen Jin   | Columbia University   | zj7@cumc.columbia.edu      | 6    |
| Zhou Yu       | East China Normal University  | zyu@stat.ecnu.edu.cn       | 64   |
| Zijian Guo    | Rutgers University  | zijguo@stat.rutgers.edu    | 44   |
| Ziwei Mei     | Chinese University of Hong Kong   | zwmei@link.cuhk.edu.hk     | 29   |