2026 Statistical Learning and Data Science Conference

Inference and Intelligence

Conference Dates and Location: November 1–3, 2026, Brooklyn, NY
New York Marriott at the Brooklyn Bridge

Sessions and Abstracts

Contents

Venue Floor Plan **4**

Parallel Sessions 1 | 10:30 AM–12:00 PM, November 1 **5**

Session 85: Trustworthy Learning and Inference at Scale . . . 5
Session 68: Causal Intelligence: Modern Perspectives at the Intersection of Statistical Inference and AI . . . 6
Session 24: Modeling Structure, Time, and Uncertainty in Medical Imaging Analysis . . . 7
Session 28: Recent Advancement in Network Analysis . . . 8
Session 29: Advances in Nonparametric Learning and Inference for Complex Data . . . 9
Session 09: Advances in High-Dimensional Statistics and Random Matrix Theory . . . 10
Session 44: Robust Inference & Learning Strategies in Modern Data Science . . . 11
Session 17: Causal and Statistical Methods for Genomic Data . . . 13
Session 11: Topics in Online Statistical Inference and Learning . . . 14
Session 65: Statistical Challenges in Data Science: Privacy, Explanation Errors, and Dynamic Ratings . . . 15
Session 70: Statistical Analysis of LLM and Representation Learning . . . 16
Session 52: New Statistical Insights on Learning and Selection . . . 17
Session 81: Causal and Interpretable Learning for Heterogeneous and Longitudinal Effects in Clinical Studies . . . 18

Parallel Sessions 2 | 01:30 PM–03:00 PM, November 1 **19**

Session 63: Statistical Foundations of Modern Generative and Decision Models . . . 19
Session 30: Modern Statistical Learning and Representation Methods for Complex Biomedical Data . . . 20
Session 64: Emerging Advances in Analysis of Complex Data . . . 21
Session 32: Causal Inference and Decision-Making on Networks . . . 22
Session 79: Recent Advances in Empirical Bayes: Theory, Applications, and Methods . . . 23
Session 20: Reliable Discovery in Complex Data: Advances in High-Dimensional and Interpretable Clustering . . . 24
Session 62: Structure Learning and Inference in High-Dimensional Dynamic Systems . . . 26
Session 69: From Spatial Patterns to Biological Intelligence: Statistical Inference for Spatial Omics . . . 27
Session 101: Advances in Transfer Learning . . . 28
Session 53: Advances in Machine Learning for Neuroimaging and Behavioral Data in Mental Health Research . . . 29
Session 77: Novel Methods in Learning Dependent and Dynamical Data . . . 30

Parallel Sessions 3 | 03:30 PM–05:00 PM, November 1 **32**

Session 95: Statistical Foundations and Interpretability for LLM Evaluation and Reasoning . . . 32
Session 66: Statistical Learning for Causal Inference: From Estimation to Decision-Making . . . 33
Session 43: Large Language Models and Statistical Foundations . . . 34
Session 67: New Frontiers in High-Dimensional Graphical Modeling . . . 35
Session 39: Statistical Learning and Inference for Metric-Space-Valued Data . . . 36
Session 51: Advances in Random Matrix Theory and Its Statistical Applications . . . 37
Session 34: Statistical Inference under Data Perturbation, Structure, and Heterogeneity . . . 38
Session 04: Analysis of Omic Data, and AI and Data-Driven Science . . . 39

Session 89: Equilibrium-Aware A/B Testing . . . 40
Session 48: Integrative Statistical Learning for Complex Biomedical Data . . . 41
Session 42: Impactful Applications in Statistics and Data Science in Aging . . . 42

Parallel Sessions 4 | 10:30 AM–12:00 PM, November 2 **44**

Session 94: Statistical Foundations for Reliable and Interpretable AI . . . 44
Session 86: Machine Learning for Causal Inference and Econometrics . . . 44
Session 46: Statistical Principles and Optimization for High-Dimensional Learning and AI . . . 45
Session 56: Novel Network Models and Model Selection . . . 46
Session 50: Uncertainty, Dependence, and Structure in Modern Data Analysis . . . 47
Session 35: Statistical Inference and Learning in High-Dimensional Structured Models . . . 49
Session 14: Reliable Modeling and Prediction in Complex Data Analysis . . . 50
Session 19: Emerging Methods and Applications in Single-Cell Analysis . . . 51
Session 21: Learning from Neuroimaging Data: Statistical and AI Methods for Brain Aging . . . 52
Session 03: Mediation Analysis and Causal Inference in Biomedical Research . . . 53
Session 83: Statistical Methods for Dissecting Heterogeneity and Decision-Making in Precision Medicine Research 54
Session 57: Advances in Statistical Learning for Data Integration . . . 56
Session 23: The Interplay Between Statistics and Data-Driven Decision-Making . . . 57

Parallel Sessions 5 | 01:30 PM–03:00 PM, November 2 **58**

Session 96: When Does Explainable AI Actually Open the Black Box . . . 58
Session 38: Statistical Methods for AI: Attribution, Alignment, Representation Learning, and Generative Evaluation 59
Session 16: Causality and AI in Science . . . 61
Session 26: Recent Advances in Statistical Network Modeling and Inference . . . 61
Session 71: Advancement in Nonparametric Method in Complex Data . . . 62
Session 08: Conformal Prediction with Partially Observed Data . . . 63
Session 49: Advances in Statistical Inference for Complex Data . . . 65
Session 73: AI/ML for Biomedical Research . . . 66
Session 75: Causal AI and Representation Learning . . . 67
Session 61: AI/ML and Advanced Statistical Methods in Clinical Development . . . 68
Session 58: Advances in Methodology and Theory for Network Analysis . . . 68

Parallel Sessions 6 | 03:30 PM–05:00 PM, November 2 **70**

Session 82: Artificial Intelligence in Rare Disease Therapeutics: Integrating Drug Repurposing, Hybrid Trial Design, and Real-World Evidence . . . 70
Session 31: From Statistical Inference to Production AI: Industry Case Studies in Causality and Intelligent Systems 71
Session 72: Methods and Theory for Machine Learning: Estimation and Prediction . . . 71
Session 55: Statistical-Computational Gaps in Network and Tensor Data . . . 72
Session 12: Modern Perspectives in Bayesian Statistics . . . 74
Student Paper Award Session . . . 75
Session 10: Statistical Advances in the Analysis of Embeddings, Networks, and Graphs . . . 75
Session 74: Enhanced Causal Inference and Clinical Trials with AI/ML . . . 76
Session 47: High-Dimensional Data in Neuroimaging and AI . . . 77
Session 106: From Data to Structure and Inference . . . 78
Session 103: Recent Methodological Advances in Causal Inference . . . 79

Parallel Sessions 7 | 10:30 AM–12:00 PM, November 3 **80**

Session 37: Valid Uncertainty Quantification in Modern Statistical Learning . . . 80
Session 91: AI Foundations: Statistical Principles for Modern Learning and Large-Scale Models . . . 81

Session 40: Data Attribution in Statistical Science and AI . . . 83
Session 59: Modern Network Analysis and Applications Across Disciplines . . . 84
Session 84: Analyzing Complex-Structured Data with Latent Geometry and Heterogeneity . . . 85
Session 54: Learning from Real-World Health Data: Methods for Imbalance, Heterogeneity, Equity, and Real-World Evidence Translation . . . 85
Session 22: Causal Mechanisms and Inference in Complex Data . . . 86
Session 88: Advances in Methods for Leveraging External Information in Precision Medicine . . . 87
Session 36: Network Interference, Spillover and Temporal Effects . . . 89
Session 107: Next-Generation Decision Intelligence: From Underwriting and Uplift Modeling to LLM-Driven Statistical Experiment Lifecycle Automation and Clinical Assessment . . . 90
Session 60: Robust Causal Inference Under Real-World Complications: Extrapolation, High Dimensions, and Missing Data . . . 90
Session 27: Advances in Statistical Learning for Real-World Evidence . . . 92
Session 92: Advances in Graphical Models for Causality and Missingness . . . 92

Parallel Sessions 8 | 01:30 PM–03:00 PM, November 3 **94**

Session 78: AI-Augmented Design and Analysis in Clinical Trials and Scientific Research: Methods, Applications, and Responsible Implementation . . . 94
Session 105: AI Safety and Evaluation . . . 94
Session 15: Distribution-Free Statistical Inference for AI . . . 95
Session 98: Robustness Bayesian Methods with Machine Learning Applications . . . 96
Session 76: Improving Inference in Nonparametric Models . . . 97
Session 18: Causal and Machine Learning Methods in Health Research . . . 98
Session 01: Neural Networks in Dimension Reduction and Causal Inference . . . 99
Session 33: Embeddings and Dynamics in Complex Networks . . . 100
Session 45: Statistical Learning and Network Analysis for Complex Biological Data . . . 101
Session 90: Mechanism Design for Data Science and AI . . . 102
Session 06: Statistical Learning in Biomedical Studies . . . 104
Session 41: Innovative Methods in Statistics and Data Science in Aging . . . 105
Session 97: Transparent Machine Learning Through Visualization and Human-Centered Design . . . 106

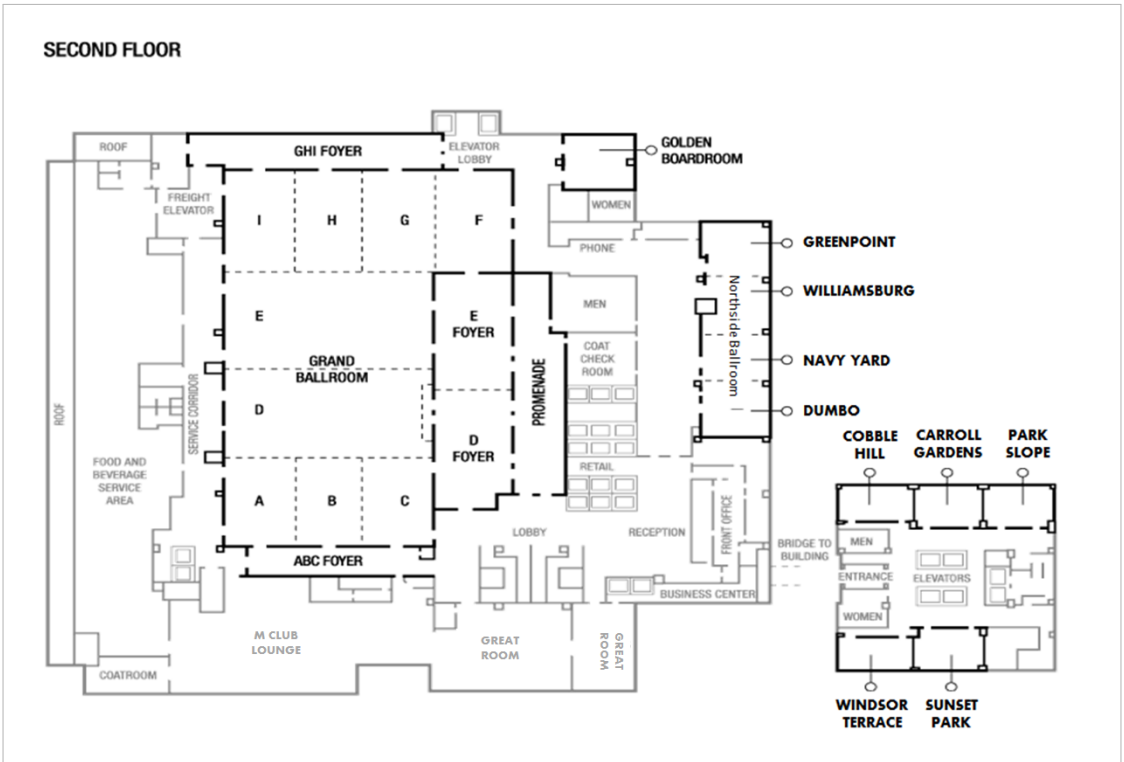
Parallel Sessions 9 | 03:30 PM–05:00 PM, November 3 **107**

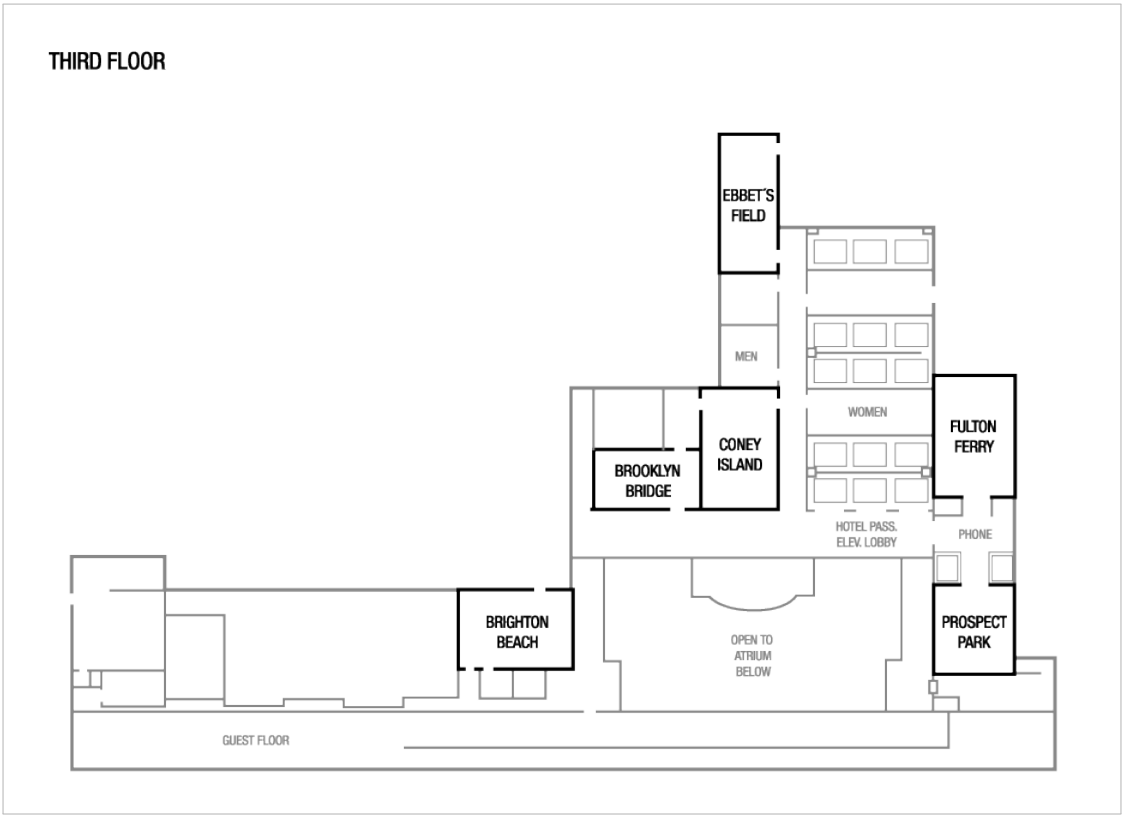
Session 87: Human-AI Collaboration . . . 107
Session 100: The Role of the Rashomon Effect in Responsible AI . . . 108
Session 07: New Statistical Tools for High-Dimensional Biomedical Data Analysis . . . 109
Session 02: Learning, Inference and Decision Making in Complex Structures . . . 111
Session 102: Making Sense of Multivariate Data . . . 112
Session 104: AI-Driven Clinical Intelligence: Integrating Generative Models, Predictive Analytics & Clinical Decision Making . . . 113
Session 13: Navigating the AI Transformation in Biostatistics: Opportunities, Methods, and Evolving Practice . . . 113
Session 80: Spatial Causal Inference . . . 114
Session 99: Digital Twins and Synthetic Data for Clinical Research . . . 115
Session 25: Statistical Machine Learning for Complex Data . . . 117
Session 05: Statistical Learning in Complex System . . . 118
Session 93: Causal Inference in Complex Real-World Data Applications . . . 119

Poster Session | 06:30 PM–08:00 PM, November 1 **120**

Venue Floor Plan

New York Marriott at the Brooklyn Bridge





Salon DE / Plenary comprises sections D and E of the Grand Ballroom; Salon C is section C.

Parallel Sessions 1: Sunday, November 1, 2026, 10:30 am–12:00 pm

Session 85: Trustworthy Learning and Inference at Scale

Location: Salon C

Organizer: Yinqiu He (University of Wisconsin-Madison)

Chair: Yinqiu He (University of Wisconsin-Madison)

Presenters: Yichi Zhang; Yanyuan Ma; Yang Feng; Debarghya Mukherjee

Random-walk debiased inference for contextual ranking model with application in large language model evaluation

Yichi Zhang, University of Georgia

Yichi Zhang (University of Georgia), Alexandre Belloni (University of Chicago), Ethan X. Fang (Duke University), Junwei Lu (Harvard University), Xiaoan Xu (Duke University)

We propose a debiased inference framework to infer the ranking structure in the contextual Bradley- Terry-Luce (BTL) model. For the pairwise item comparison, we introduce a novel random-walk debiased estimator to efficiently aggregate the estimating functions of different item pairs. To approximate the nuisance preference score functions in the debiased estimator, we further introduce a nonparametric maximum likelihood-based method that can leverage many loss-minimization method, e.g., the spline regression and deep neural networks. With decently estimated nuisance functions, our debiased estimator yields a tractable distribution, and achieves the semiparametric efficiency lower bound asymptotically. We further extend our method for the multiple hypotheses testings by incorporating the multiplier bootstrap techniques, and adapting it to accommodate the distributional shift of contextual variables. We provide thorough numerical studies to validate the statistical properties of our method, and showcase its applicability in evaluating large language models based on human preferences under different contexts.

Several aspects in label shift

Yanyuan Ma, Pennsylvania State University

Yanyuan Ma

We provide an introduction to label shift problems. In the context of discrete response, we study the importance weights confidence set problem by a paradigm shift from traditional inversion-based inference to a direct matrix constraint framework. We use this framework to characterize a joint confidence region and extract marginal intervals via linear programming, deriving provably tighter bounds for importance weights while maintaining exact finite-sample validity. In the context of continuous response, we study the estimation and inference of a general target population characteristic by developing doubly and singly robust estimators as well as the efficient estimator. Many ongoing and future developments will be discussed too.

Federated transfer learning with differential privacy

Yang Feng, New York University

Mengchu Li, Ye Tian, Yang Feng, Yi Yu

Federated learning has emerged as a powerful framework for analysing distributed data, yet two challenges remain pivotal: heterogeneity across sites and privacy of local data. In this paper, we address both challenges within a federated transfer learning framework, aiming to enhance learning on a target data set by leveraging information from multiple heterogeneous source data sets while adhering to privacy constraints. We rigorously formulate the notion of federated differential privacy, which offers privacy guarantees for each data set without assuming a trusted central server. Under this privacy model, we study four statistical problems: univariate mean estimation, low-dimensional linear regression, high-dimensional linear regression, and M-estimation. By investigating the minimax rates and quantifying the cost of privacy, we show that federated differential privacy is an intermediate privacy model between the well-established local and central models of differential privacy. Our analyses account for data heterogeneity and privacy, highlighting the fundamental costs associated with each factor and the benefits of knowledge transfer in federated learning.

PDE-constrained inverse problems at the root(n) rate via debiased physics-informed neural networks

Debarghya Mukherjee, Boston University

Yves Atchade (Boston University); Debarghya Mukherjee (Boston University)

We study the problem of estimating unknown parameters in PDE-constrained inverse problems from noisy observations, where the PDE solution is approximated using Physics-Informed Neural Networks (PINNs). While PINNs have demonstrated remarkable empirical success, existing estimators often inherit the slow nonparametric convergence rate of the neural-network solution, leading to biased and statistically inefficient inference for the finite-dimensional parameters of interest. To address this, we propose a two-step debiased estimation procedure that combines neural-network-based nonparametric estimation with an influence-function-based bias correction. By eliminating the first-order sensitivity of the estimator to errors in the nuisance function, our procedure yields a root(n) consistent and asymptotically normal estimator without requiring undersmoothing of the neural network component. We further extend this framework to Bayesian inference by replacing the original likelihood with a debiased quasi-likelihood and establish a Bernstein-von Mises theorem showing that the resulting posterior contracts at the root(n) rate with an asymptotic covariance matching that of the frequentist estimator. As a by-product of our analysis, we establish near-minimax optimal

convergence rates for estimating a nonparametric regression function and its derivatives in Sobolev spaces using neural networks. Extensive numerical experiments corroborate our theoretical findings and demonstrate the necessity of the proposed debiasing procedure for valid statistical inference in PDE-constrained inverse problems.

Session 68: Causal Intelligence: Modern Perspectives at the Intersection of Statistical Inference and AI

Location: Williamsburg

Organizer: Honglang Wang (Indiana University Indianapolis)

Chair: Kun Zhang (Carnegie Mellon University)

Presenters: Rajesh Ranganath; Kun Zhang; Ruoqi Yu; Michele Santacatterina

Causality and competing events with autoregressive transformers

Rajesh Ranganath, New York University

Yoav Wald (Technion), Xiang Gao (NYU), Sumit Chopra (NYU), Juho Lee (KAIST), Rajesh Ranganath (NYU)

Modeling interventions that include timing is an important task in many domains, such as treatments in healthcare, transactions in finance, and others. However, estimating these interventional distributions is challenging because standard next-token predictors, typically implemented using transformers, do not naturally support altering the timing of specific future events. We develop Competing Event Models (CEMs), an autoregressive generative approach that simplifies the estimation of interventions on both what interventions are applied and when. Drawing from the competing risks literature, CEMs model independent probability distributions for latent times of each event type, where the earliest realized time determines the next event. We show that this approach enables straightforward interventions on event timing and efficient sampling, and, through appropriate design choices, can also model concurrent events.

Causal representation learning and causal generative AI

Kun Zhang, Carnegie Mellon University

Kun Zhang

As a core pillar of science and engineering, causality is transforming our approach to modern machine learning and artificial intelligence. Uncovering the causal process underlying observed data naturally helps answer 'why' and 'what-if' questions, informs optimal decision-making, and enables adaptive prediction. In many scenarios, observed variables, such as image pixels and questionnaire responses, are often reflections of the underlying hidden causal variables rather than being causal variables themselves. Causal representation learning aims to reveal the underlying hidden causal variables and their relations. In this talk, we show how the modularity property of causal systems makes it possible to recover the underlying causal representations from observational data with identifiability guarantees. We further demonstrate how identifiable causal representation learning can directly benefit generative AI, using image generation / editing and extrapolative data generation as illustrative examples.

Using a two-parameter sensitivity analysis framework to efficiently combine randomized and non-randomized studies

Ruoqi Yu, University of Illinois Urbana-Champaign

Ruoqi Yu (University of Illinois Urbana-Champaign), Bikram Karmakar (University of Wisconsin–Madison), Jessica Vandeleest (University of California Davis), Eleanor Bimla Schwarz (University of California San Francisco)

Causal inference is vital for informed decision-making across fields such as biomedical research and social sciences. Randomized controlled trials (RCTs) are considered the gold standard for internal validity of inferences, whereas observational studies (OSs) often provide the opportunity for greater external validity. However, both data sources have inherent limitations preventing their use for broadly valid statistical inferences: RCTs may lack generalizability due to their selective eligibility criterion, and OSs are vulnerable to unobserved confounding. This paper proposes an innovative approach to integrate RCT and OS that borrows the other study's strengths to remedy each study's limitations. The method uses a novel triplet matching algorithm to align RCT and OS samples and a new two-parameter sensitivity analysis framework to quantify internal and external validity biases. This combined approach yields causal estimates that are more robust to hidden biases than OSs alone and provides reliable inferences about the treatment effect in the general population. We apply this method to investigate the effects of lactation on maternal health using a small RCT and a long-term observational health records dataset from the California National Primate Research Center. This application demonstrates the practical utility of our approach in generating scientifically sound and actionable causal estimates.

Automatic, debiased, and invariant counterfactual generation under general interventions

Michele Santacatterina, New York University

Raphael C. Kim (Cornell University); Jingsen Zhu (Cornell University); Ramin Zabih (Cornell University); Michele Santacatterina (New York University)

Generative models for counterfactual outcomes have great potential to support decision-making under complex interventions, but existing approaches are limited by unstable estimation, poor generalization across environments, and bias from nuisance model misspecification. We introduce ADIGen, a framework for automatic, debiased, and invariant counterfactual generation under general interventions, including high-dimensional interventions and outcomes. ADIGen

combines Riesz regression to avoid unstable density-ratio estimation, causal invariance to improve generalization under distribution shift, and orthogonal statistical learning to obtain doubly robust guarantees against nuisance model misspecification. We provide excess-risk bounds showing that ADIGen controls counterfactual risk under general interventions, with a product-bias nuisance remainder and an invariant risk bound across environments. We then extend this framework to multiple, interacting objects with a joint intervention, and apply ADIGen to counterfactual world modeling. In contrast to standard statistical settings, the joint outcome is modeled natively without the need for exposure mappings or direct/indirect effect decompositions.

Session 24: Modeling Structure, Time, and Uncertainty in Medical Imaging Analysis

Location: Fulton Ferry

Organizer: (Din) Ding-Geng Chen (Arizona State University)

Chair: (Din) Ding-Geng Chen (Arizona State University)

Presenters: G M Fahad Bin Mostafa; Ruiwen Zhou; Hao Wang; Ding-Geng Chen

Estimation of statistical uncertainty for joint CCTA cardiovascular structure segmentation under distributional shift

G M Fahad Bin Mostafa, Arizona State University

Fahad Mostafa (Arizona State University)

Deep learning offers a promising approach for automated Coronary Computed Tomography Angiography (CCTA) analysis, including cardiovascular structure segmentation and disease-related classification. However, its reliability may degrade under distributional shifts arising from scanner protocols, image quality, contrast enhancement, reconstruction settings, and anatomical variability. We propose a statistical uncertainty quantification framework for robust CCTA analysis under dataset shift. Using high-resolution three-dimensional CCTA images with pixel-level annotations, the framework supports supervised segmentation and downstream cardiovascular disease classification.CCTA images are split using stratified k-fold training, validation, and testing, with controlled distributional shifts introduced to assess robustness. An uncertainty-aware deep learning model, based on Bayesian inference, Monte Carlo dropout, or deep ensembles, jointly estimates segmentation masks, class probabilities, and predictive uncertainty. A data-adaptive uncertainty threshold separates high-confidence predictions from uncertain or potentially out-of-distribution cases. Performance is assessed using segmentation, classification, robustness, and calibration criteria, including Dice similarity, Hausdorff distance, sensitivity, specificity, area under the ROC curve, Brier score, and expected calibration error. At inference, high-confidence predictions are accepted for automated analysis, while uncertain cases are referred for expert review. By integrating uncertainty estimation, calibration, dataset-shift evaluation, and selective prediction, the proposed framework provides a statistically principled approach for developing reliable CCTA-based clinical decision support systems.

A temporal graph neural network for time-to-Alzheimer's disease prediction using longitudinal neuroimaging data

Ruiwen Zhou, Washington University in St. Louis

Shui Mauser (Departments of Neuroscience and Mathematics, College of Wooster); Lei Liu (Division of Biostatistics, Washington University School of Medicine in St. Louis); Ran Chen (Department of Statistics and Data Science, Washington University School of Medicine in St. Louis); Jason Hassenstab (Department of Neurology, Washington University School of Medicine in St. Louis); Shu Jiang (Department of Surgery, Washington University School of Medicine in St. Louis); Guoqiao Wang (Department of Neurology, Washington University School of Medicine in St. Louis); Chengjie Xiong (Division of Biostatistics, Washington University School of Medicine in St. Louis); Linying Zhang (Division of Biostatistics, Washington University School of Medicine in St. Louis); Ruiwen Zhou (Division of Biostatistics, Washington University School of Medicine in St. Louis)

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by gradual structural brain changes that precede clinical symptoms. The Alzheimer's Disease Neuroimaging Initiative (ADNI) provides longitudinal structural MRI (sMRI) data—cortical thickness, surface area, and volumetric measures across brain regions—that serve as important biomarkers for tracking early AD progression. However, predicting time-to-AD conversion from these data is challenging due to inter-subject variability, complex spatial dependencies among brain regions, and temporal heterogeneity across visits. Traditional and deep learning classifiers often overlook censoring and fail to jointly model spatiotemporal patterns of neurodegeneration. We propose a temporal graph neural network (GNN) framework for time-to-Alzheimer's prediction that integrates anatomical structure with longitudinal trajectories. Our approach combines landmark analysis to handle censoring with a GNN that encodes structural relationships among brain regions, along with LSTM/Transformer-based temporal modules that capture both local dynamics and long-range progression. These hybrid architectures jointly model spatial organization and evolving sMRI patterns, producing individualized and clinically meaningful time-to-event risk predictions. The proposed framework offers a flexible and powerful tool for early detection of AD and may improve the identification of high-risk individuals in longitudinal neuroimaging studies.

Multi-model augmentation for robust retinal disease classification

Hao Wang, Arizona State University

Hao Wang (Arizona State University), Wenhui Zhu (Arizona State University), Xuanzhao Dong (Arizona State University), Yanxi Chen (Arizona State University), Xin Li (Arizona State University), Peijie Qiu (Washington University in St. Louis), Xiwen Chen (Morgan Stanley),

Vamsi Krishna Vasa (Arizona State University), Yujian Xiong, Oana M. Dumitrascu (Mayo Clinic), Abolfazl Razi (Clemson University), and Yalin Wang (Arizona State University)

Large-scale multimodal foundation models have demonstrated strong performance across diverse tasks in recent years, yet their dominance has shifted attention away from task-specific modeling, particularly in medical applications where reliable learning systems often require specialized designs to address domain-specific challenges. We propose an efficient training framework for robust retinal disease classification based on multiple lightweight CNN architectures. Our approach mitigates overfitting and dataset bias while improving robustness to domain shift by integrating diverse data augmentation strategies, ensemble learning across models with varying complexities, and weakly supervised multi-task learning under missing labels. The resulting fusion model achieves improved generalization in data-scarce settings while balancing computational efficiency and representational capacity.

Robust PCA with truncated weighted nuclear norm and adaptive histogram equalization for medical image enhancement

Ding-Geng Chen, Arizona State University

Ding-Geng Chen and Habte Tadesse LIKASSA

In this talk, we will present a RPCA method that addresses existing limitations in retinal image enhancement by integrating a Truncated Weighted Nuclear Norm (TWNN) with Adaptive Histogram Equalization (AHE). In this approach, AHE improves local contrast and preserves fine retinal details, while TWNN strengthens RPCA's ability to separate essential structural components from noise. The method is formulated as an optimization problem and efficiently solved using the Alternating Direction Method of Multipliers (ADMM), ensuring scalability and computational efficiency for large datasets. Extensive experiments on synthetic and publicly available retinal datasets, including EyeQ, DRIVE, STARE, and Kaggle, demonstrate that the proposed method substantially improves image fidelity, clearly reveals fine blood vessels, and outperforms existing techniques. We have further implemented this approach in a user-friendly application, enabling rapid enhancement of degraded retinal images for clinical use. A patent has been filed with ASU Innovation, highlighting the translational potential of this technology. This framework establishes a robust methodological foundation for next-generation high-dimensional medical imaging analytics and accurate machine learning based disease detection, bringing us closer to real-time, clinically actionable retinal image analysis.

Session 28: Recent Advancement in Network Analysis

Location: Sunset Park

Organizer: Tracy Ke (Harvard University)

Chair: Tracy Ke (Harvard University)

Presenters: Liza Levina; Jiashun Jin; Qiwei Yao; Tracy Ke

Comparing groups of networks

Liza Levina, University of Michigan

Liza Levina (University of Michigan)

Network-structured data arise in many applications; joint analysis of multiple networks is especially relevant to applications like neuroimaging, where each network corresponds to a patient's brain connectome. Models for multiple networks have typically focused on estimating their shared structure. Two-sample tests have also been developed, testing some version of the hypothesis of two samples of networks being indistinguishable across all node pairs (the global null). However, scientifically relevant hypotheses rarely take this form: for example, in neuroimaging it is rarely of interest to compare whole brains of patients and healthy controls, and the focus is more often on a particular brain region. Beyond comparisons, it is also of interest to estimate structures that correspond to a specific subgroup of networks, for example, for patients with a certain trait. One could always do that using just the subgroup of interest, but using all available samples allows us to better estimate structures that are shared by all, which in turn helps separate out the structure associated with a trait. This talk will introduce two methods that help address these challenges: mesoscale testing on networks, which conducts formal hypothesis testing on a subset of edges (like a brain region) while leveraging the rest of the network to increase power; and group MultiNeSS, a method that takes a sample of networks and estimates structures that are shared by all, specific to groups, or just unique to an individual network. In both cases, we leverage the assumption of low-rank expectation of adjacency matrices which has been observed widely in practice. Based on joint work with Peter MacDonald, Alexander Kagan, and Ji Zhu.

Title TBA

Jiashun Jin, Carnegie Mellon University

Abstract TBA

An adversarial approach for goodness-of-fit tests and its application in dynamic network models

Qiwei Yao, London School of Economics and Political Science

Jinyuan Chang (Southwestern University of Finance and Economics); Chengchun Shi (London School of Economics); Mingcong Wu (Southwestern University of Finance and Economics); Qiwei Yao (London School of Economics); Xinyang Yu (London School of Economics)

Diagnostic checking for goodness-of-fit is one of the important and routine steps in building a statistical model. The most frequently used approach for checking the goodness-of-fit is the residual analysis in the context of regression analysis. However for many statistical models there exist no natural residuals, including some dynamic network models especially with dependent edges. We propose an adversarial approach in this paper for checking the goodness-of-fit: we generate a synthetic sample from the fitted model and construct a classifier to classify the original sample and the synthetic sample into two different classes. The hardness of the classification is then taken as a measure for the goodness-of-fit. We adopt an appropriate measure for the 'hardness' and construct the optimal multi-layer perceptron (MLP) classifier which minimizes the hardness measure. An estimated hardness measure for this optimal MLP classifier is taken as our test statistic. Via a sample splitting scheme, we have shown that the proposed test is asymptotically normal under the null hypothesis and has the power converging to 1 if the null hypothesis does not hold. Based on the same adversarial framework, we also construct a bootstrap version of the goodness-of-fit test based on a random forest classifier. It is specifically designed for the high-dimensional applications in the sense that the number of observations is small in relation to the size/dimension of the data. The bootstrap resampling automatically adapts to the possible serial dependence in the data. The proposed methods are further illustrated via both simulation and some real data applications.

The gold rush in AI4Math

Tracy Ke, Harvard University

Yuanchuan Guo (Harvard); Jiashun Jin (SEU); Bingcheng Sui (Harvard)

Recent advances in artificial intelligence have sparked a gold rush in mathematical research. Yet beyond striking individual achievements, how much do we know about AI's actual use and its influence on mathematical practice? This talk presents two complementary studies and highlights the role of statistics in turning a rapidly evolving phenomenon into questions that can be measured and investigated.

First, we analyze 32,944 arXiv mathematics submissions posted between March and August 2026. Using authors' explicit disclosures, we document a sharp increase in substantive AI assistance and examine its distribution across mathematical fields, research activities, institutions, and countries.

Second, we introduce ProofTrace, a framework for investigating structural signatures associated with AI-assisted mathematics. We represent each paper as a result-dependency graph, linking mathematical results and substantive intermediate conclusions through their direct dependencies. Statistical analysis of these representations reveals differences in the organization of arguments, proof placement, citation patterns, and dependency structure. These features also support detection that is robust to AI rewording.

Session 29: Advances in Nonparametric Learning and Inference for Complex Data

Location: Golden Boardroom

Organizer: Zhao Ren (University of Pittsburgh)

Chair: Zhao Ren (University of Pittsburgh)

Presenters: Wen Zhou; Heejong Bong; Tianyu Zhang; Lang Zeng

Adaptive inference for stratified directed network effects with an application to asymmetric partisan polarization

Wen Zhou, New York University

Pengfei Tian (Tsinghua University); Cassandra Handan-Nader (New York University); Yuan Zhang (The Ohio State University); Fan Yang (Tsinghua University)

U.S. legislatures exhibit persistent and evolving partisan polarization, which is often studied using voting networks. Bill cosponsorship networks, which link legislators through jointly supported bills, provide a richer relational alternative that includes a larger set of bills and allows for the study of asymmetric polarization through directed links. However, such networks pose substantial challenges for inference, since directionality renders standard parametric models overly restrictive and unreliable. Statistically, inference on directed network effects can be based on network U-statistics, yet their degeneracy may be indeterminate. These difficulties are further amplified by the bi-strata structure induced by party affiliation: dependence patterns differ across strata, invalidating methods such as cluster-based bootstrap and forcing the coexistence of distinct degeneracy regimes across unequally sized strata. To address these, we show that a broad class of network effect estimators in bi-strata directed networks can be reformulated as two-sample network U-statistics, yielding a unified nonparametric framework. We identify three regimes of network effect estimators: degenerate, partially degenerate, and nondegenerate, derive their asymptotics within a single framework, and propose a data-adaptive variance estimator that yields a unified test with Berry–Esseen bounds in each regime, providing finite-sample guarantees. Applied to U.S. Congressional cosponsorship networks from 2003 to 2023, our method reveals that Republican minorities were more likely to be dissatisfied with Democratic majority proposals than the reverse, and that Democratic majorities drove a novel decline in polarization during the 116th House and 117th Senate by backing minority party proposals.

Heterogeneous treatment effects under network interference: a nonparametric approach based on node connectivity

Heejong Bong, Purdue University

Colin B. Fogarty (University of Michigan); Elizaveta Levina (University of Michigan); Ji Zhu (University of Michigan)

In network settings, interference between units makes causal inference more challenging as outcomes may depend on the treatments received by others in the network. Typical estimands in network settings focus on treatment effects aggregated across individuals in the population. We propose a framework for estimating node-wise counterfactual means, allowing for more granular insights into the impact of network structure on treatment effect heterogeneity. We develop a doubly robust and non-parametric estimation procedure, KECENI (Kernel Estimator of Causal Effect under Network Interference), which offers consistency and asymptotic normality under network dependence. The utility of this method is demonstrated through an application to microfinance data, revealing the node-wise impact of network characteristics on treatment effects.

Model selection consistency of rolling validation

Tianyu Zhang, University of California, Santa Barbara

Tianyu Zhang (University of California, Santa Barbara); Julian Braganza (Carnegie Mellon University); Jing Lei (Carnegie Mellon University)

Cross-validation is one of the most commonly applied model selection criteria. However, it is interesting that CV may fail in many important settings, including variable selection in parametric linear models; It cannot choose the better-performing model even with a diverging sample size. Formally similar to leave-one-out CV, rolling-validation (RV) is a selection criterion that naturally emerges in streaming data analysis. This method examines model quality through its prediction accuracy on the next observed sample. We discovered that RV can achieve correct model selection even in the settings where CV fails (while remaining consistency when CV succeeds). A close investigation enhanced our understanding of what constitutes signal and noise in model selection tasks.

Nonparametric probabilistic index model

Lang Zeng, The Ohio State University

Lang Zeng (University of Pittsburgh); Weijing Tang (Carnegie Mellon University); Wenjie Wang (Eli Lilly and Company); Zhao Ren (University of Pittsburgh); Lu Mao (University of Wisconsin-Madison); Ying Ding (University of Pittsburgh)

Many scientific applications involve ranking objects according to predicted scores derived from observed covariates, such as identifying patients at higher risk of disease progression or screening molecules with superior therapeutic efficacy. Motivated by applications with complex covariates, such as images and RNA sequences, we propose the nonparametric probabilistic index model (NPIM), a neural network-based extension of the classical probabilistic index model (PIM), to directly model the conditional probability of outcome ordering between two objects. Compared with classical PIM, NPIM relaxes parametric assumptions on covariate effects and learns a flexible real-valued score function that captures complex covariate effects on the ordering structure of outcomes. We establish consistency and convergence rates for the proposed neural network M-estimator under composite smoothness assumptions on the true score function. NPIM is broadly applicable to continuous, ordinal, and composite outcomes whenever a meaningful ordering can be defined. To improve interpretability, we introduce a partially linear extension combining interpretable linear effects for selected covariates with a flexible nonparametric component for the remaining covariates. We further extend the framework to time-to-event outcomes through nonparametric win-fraction regression. Simulation studies and real-data applications demonstrate the utility of NPIM for rank prediction, subgroup identification, and evaluating covariate effects on outcome improvement.

Session 09: Advances in High-Dimensional Statistics and Random Matrix Theory

Location: Carroll Gardens

Organizer: Joshua Cape (University of Wisconsin-Madison)

Chair: Joshua Cape (University of Wisconsin-Madison)

Presenters: Rishabh Dudeja; Michael J. Feldman; Dmitriy (Tim) Kunisky; Elad Romanov

High-dimensional asymptotics of differentially private PCA

Rishabh Dudeja, University of Wisconsin–Madison

Youngjoo Yun (University of Wisconsin–Madison); Rishabh Dudeja (University of Wisconsin–Madison)

In differential privacy, random noise is introduced to privatize summary statistics of a sensitive dataset before releasing them. The noise level determines the privacy loss, which quantifies how easily an adversary can detect a target individual's presence in the dataset using the published statistic. Most privacy analyses provide non-asymptotic upper bounds on the privacy loss which hold uniformly across all datasets. Sometimes, these bounds can be pessimistic on a given dataset. In such cases, it can be useful to complement these privacy bounds with sharp privacy characterizations that quantify a mechanism's exact privacy loss on a given dataset. With this goal, we study differentially private principal component analysis (PCA), where the goal is to privatize the leading principal components of a dataset with n samples and p features. We analyze the exponential mechanism and provide sharp asymptotic characterizations of its utility and privacy loss in the high-dimensional limit ($p \to \infty$). In this limit, we show that detecting a target individual's presence using privatized principal components is asymptotically equivalent to distinguishing between two Gaussians with different means, where the mean difference depends on certain spectral properties of the dataset. Our analysis combines the hypothesis-testing formulation of privacy guarantees proposed by Dong, Roth, and Su (2022) with Le Cam's contiguity arguments.

Random matrix theory for geographic population genetics

Michael J. Feldman, New York University

Michael J. Feldman, Saharon Rosset

Principal component analysis is a central tool in genetics, where the leading PCs of genetic data often reveal geographic population structure. A striking example is "Genes Mirror Geography within Europe" by Novembre et al. (2008): plotting the first two PC scores produced an approximate map of Europe, although the data contained no explicit geographical information. Although the eigenvalue spectra of such genetic data resemble those of spiked covariance models, it remains unclear how this spectral structure emerges from migration and genetic drift. We investigate this connection through the Wright–Fisher stepping-stone model with migration and mutation, a standard model in population genetics. Specifically, we characterize the following spectral properties in high-dimensional regimes: (1) The empirical spectral distribution converges to a Marchenko–Pastur-type law. (2) Under suitable conditions, outlier eigenvalues separate from the bulk and carry information about the migration process. (3) For nearest-neighbor migration on a two-dimensional grid, the leading eigenvectors are axis-oriented and recover the "Genes Mirror Geography" spatial organization. Classical random matrix theory does not apply directly: the data is discrete, and its dependence structure does not have the standard separable covariance form. The model also exhibits distinct spectral phenomena across several high-dimensional scaling regimes. This is joint work with Saharon Rosset.

Pseudoindependence in spiked matrix models with applications to multi-spectral algorithms

Dmitriy (Tim) Kunisky, Johns Hopkins University

Shujing Chen (Johns Hopkins University); Dmitriy Kunisky (Johns Hopkins University)

Spiked matrix models describe noisy observations of low-rank signals. I will present results characterizing the fluctuations of the leading eigenvectors of such models around these hidden signals. In particular, for a rank-one signal and in the supercritical regime where the leading eigenvector of the noisy matrix is correlated with the signal vector, I will show that the fluctuations around this correlated part are universal and depend only on the first two moments of the noise entries, provided that the signal is sufficiently delocalized. In particular, these fluctuations are asymptotically Gaussian for Wigner-like noise matrices, resembling the case of noise drawn from the Gaussian orthogonal or unitary ensembles. A more general version of this analysis also applies to collections of several spiked matrices. In particular, even highly dependent collections of spiked matrices with uncorrelated Wigner-like noise can have leading eigenvectors with asymptotically independent Gaussian fluctuations, a phenomenon I will refer to as "pseudoindependence". Finally, I will describe the ramifications of pseudoindependence for spectral algorithms in statistical settings, showing that in some cases it can be advantageous to build spectral estimators from the leading eigenvectors of several deterministic transformations of an observed matrix rather than just from the leading eigenvector of the observed matrix itself.

On spectral aggregation in distributed PCA

Elad Romanov, Texas A&M University

Elad Romanov (Texas A&M University); Michael J. Feldman (New York University)

We study communication-efficient distributed principal component analysis (PCA) under the high-dimensional spiked covariance model. In a one-shot procedure, each machine transmits a collection of local empirical principal components (PCs) to a central server, which uses them to estimate the population spike directions.

We first consider the low-communication regime, in which each machine transmits finitely many leading PCs. A local PC is informative if and only if its local signal-to-noise ratio exceeds the Baik–Ben Arous–Péché (BBP) threshold. A spectral aggregation scheme with optimally chosen weights has recently been considered in the literature. We show that this scheme is in fact optimal in a considerably stronger sense: it achieves the optimal asymptotic risk among all estimators based solely on the transmitted PCs, spectral or otherwise.

Our main focus is the high-communication regime, in which each machine transmits a number of PCs proportional to the ambient dimension. In this regime, local machines may operate below the BBP threshold, so that every fixed collection of local PCs is asymptotically decorrelated from the population spike. Nevertheless, a macroscopic collection of PCs can retain nontrivial information in aggregate. For spectral aggregation, we derive exact asymptotic formulas for the estimation accuracy and characterize the resulting accuracy–communication tradeoff.

In particular, transmitting subleading PCs can yield substantially better accuracy for a given communication budget than naive data integration schemes, such as having each local machine transmit a subset of its raw observations.

Session 44: Robust Inference & Learning Strategies in Modern Data Science

Location: Coney Island

Organizer: Chi-Kuang Yeh (Georgia State University)

Chair: Chi-Kuang Yeh (Georgia State University)

Presenters: Pingbo Hu; Cong Jiang; Zeyu Bian; Chi-Kuang Yeh

Deep learning for continuous-time dynamic survival prediction with time-varying covariates

Pingbo Hu, Columbia University

Pingbo Hu (Columbia University); Yifei Sun (Columbia University)

In clinical risk prediction, event risk is often reassessed as additional patient information is observed over time. We study landmark-based dynamic survival prediction with time-varying covariates. At a landmark time, the target is the landmark survival function: the future survival probability conditional on being event-free at that landmark and on the covariate history observed up to that time. Classical landmark supermodels estimate this target by fitting a single Cox-type model over stacked landmark datasets, which rely on proportional hazards and use only covariate snapshots at the landmark. Existing deep survival methods typically discretize event time or are not designed to directly estimate the landmark survival function. We propose a continuous-time deep landmark model for estimating this target under right censoring. A sequence encoder summarizes the full covariate history available at each landmark, and a shared hazard head maps this summary, the landmark time, and the elapsed time to a landmark hazard. This shared parameterization pools information across landmarks while allowing flexible dependence on covariate history. The corresponding landmark survival function is obtained by integrating the hazard over time. For estimation, we introduce a counting-process-based M-estimation loss. We show that the true landmark hazard is the unique optimizer of the population criterion, establishing Fisher consistency. Experiments on three real-world datasets benchmark our method against Cox-type landmark models and deep survival baselines. Simulations assess the proposed methods under known data-generating mechanisms.

Evaluating causal indirect effects when mediators are left-censored by assay limit of quantification

Cong Jiang, SUNY Downstate Health Sciences University

Michael D. Hughes (Harvard Chan School of Public Health); Nima S. Hejazi (Harvard Chan School of Public Health)

Causal mediation analysis is essential for disentangling the mechanisms by which investigational therapeutic and preventive agents impact clinical outcomes. However, the measurement of biological mediators is often subject to left-censoring by technical measurement limitations, most commonly an assay's limit of quantification. This form of censoring can pose severe challenges for both identification and estimation of causal mediation estimands, particularly when the censoring mechanism is deterministic and the resulting missingness is missing not at random (MNAR) or nonignorable. Motivated by the question of assessing the role of viral RNA in the action mechanism of monoclonal antibody therapies for COVID-19 in the Accelerating COVID-19 Therapeutics and Vaccine (ACTIV)-2 platform trial, we develop a semi-parametric framework for estimation of the natural direct and indirect effects when the mediator of interest is partially subject to this form of left-censoring. Our proposed strategy combines fractional imputation with a semi-parametric EM algorithm to flexibly estimate key components of the factorized data likelihood. Applying the proposed strategy to circumvent the left-censoring, we discuss both traditional plug-in and asymptotically efficient estimators of the direct and indirect effect estimands, introducing a data-adaptive m-out-of-n bootstrap for robust inference under the imputation procedure. We demonstrate in numerical experiments that our approach significantly reduces bias and allows for reliable inference. An application to data from the ACTIV-2 platform trial confirms that monoclonal antibody therapies reduce the risk of hospitalization and death due to COVID-19, while suggesting that changes in viral RNA mediate only a modest proportion of the overall treatment effect.

Double fairness policy learning: integrating action fairness and outcome fairness in decision-making

Zeyu Bian, Florida State University

Zeyu Bian (Florida State University)

Fairness has recently received significant attention in machine learning, driven by ethical concerns surrounding algorithmic applications. However, fairness in policy learning, where the goal is to make fair decisions while simultaneously maximizing expected reward, has received less attention. Most existing work on fairness in policy learning focuses primarily on action fairness. However, many applications require fairness in both the actions taken and their resulting outcomes. This paper tackles the challenge of double fairness in policy learning by jointly accounting for three objectives: action fairness, outcome fairness, and value maximization. We propose a novel framework that integrates fairness directly into a multi-objective policy-learning problem. Since commonly used linear scalarization may fail to recover the full Pareto fairness set under non-convex policy classes, we adopt lexicographic weighted Tchebycheff scalarization, which does not require convexity. Our framework is flexible and accommodates various commonly used fairness notions. We establish theoretical guarantees for the proposed method and demonstrate favorable fairness–value trade-offs relative to competing approaches. Finally, we apply our method to a motor third-party liability insurance dataset and an entrepreneurship training dataset. In both applications, our proposed method substantially reduces action and outcome unfairness relative to unconstrained policy optimization, with only a modest reduction in estimated value.

Auditing LLM-as-a-judge with positive and unlabelled learning

Chi-Kuang Yeh, Georgia State University

Chi-Kuang Yeh (Georgia State University); Zilong Zhang (Georgia State University); Noah Yi-Ting Hung (Georgia State University); Weiyi He (Michigan State University); Junxi Zhang (Concordia University); Lei Ding (University of Manitoba)

Large language models (LLMs) are increasingly used as judges to evaluate generated responses, but their decisions may exhibit systematic biases and disagree with human preferences. This talk presents joint work with my student

and collaborators on auditing LLM judges through positive and unlabelled (PU) learning under limited human supervision. Judge decisions verified to agree with human preferences are treated as positive examples, while unaudited decisions remain unlabelled. I briefly introduce a semiparametric double exponential tilting framework to motivate the challenges of selective labelling, before focusing on two auditing approaches. The first uses partial optimal transport in a fixed representation space to identify human-consistent comparisons and correct biased judgements without retraining the judge. By matching only part of the unlabelled pool to reliable positives, it accommodates the mixture of consistent and inconsistent decisions among unaudited comparisons. The second adaptively refines confidence assessments and prioritizes uncertain comparisons for further human review. Experiments examine agreement with human preferences, robustness to evaluation biases, and the effective use of limited human feedback. Together, these methods address a practical question: which automated judgements should we trust, correct, or further verify?

Session 17: Causal and Statistical Methods for Genomic Data

Location: Park Slope

Organizer: Zhonghua Liu (Columbia University)

Chair: Zhonghua Liu (Columbia University)

Presenters: Hongyu Zhao; Hongzhe Li; Jingyi Jessica Li; Haiyan Huang

Spatial co-expression and cell-cell communication inference from spatially resolved transcriptomics data

Hongyu Zhao, Yale University

Jia Zhao (Yale), Xinning Shan (Yale), Gefei Wang (Yale), Tinyi Chu (Yale), Chen Lin (Yale), Rui Chang (Yale), Hongyu Zhao (Yale)

Cell-cell communication is fundamental to tissue organization, homeostasis, and disease progression. Recent advances in spatial transcriptomics provide unprecedented opportunities to systematically characterize ligand-receptor interactions directly within intact tissues. However, robust inference of spatial ligand-receptor interactions remains challenging because intrinsic features of spatial transcriptomics data, including spatial autocorrelation, variation in total molecular counts, and measurement errors, can induce spurious spatial co-expression and inflate false-positive rates. Most existing methods do not adequately account for these confounding factors, limiting the reliability of inferred cellular communication. Here, we present a statistical method for spatially constrained co-expression and ligand-receptor interaction inference that jointly models spatial autocorrelation, variation in total molecular counts, measurement errors, and spatial proximity constraints. Our method combines efficient moment-based parameter estimation with analytical hypothesis testing, enabling fast and statistically rigorous inference without restrictive distributional assumptions. Through extensive simulations, real-data permutation experiments, and biologically motivated negative-control analyses across different spatial transcriptomics platforms, we show that most existing methods presented inflated false-positive rates, whereas our method achieved well-calibrated inference, robust false-positive control, and improved detection power. Its applications to high-resolution MERFISH and CosMx datasets from intestinal inflammation and non-small cell lung cancer further highlight its biological utility in disease contexts.

Confounder-robust causal discovery and inference in Perturb-seq using proxy and instrumental variables

Hongzhe Li, University of Pennsylvania

Hongzhe Li (University of Pennsylvania); Kwangmoon Park (University of Pennsylvania)

Emerging single-cell technologies that combine CRISPR-based genetic perturbations with single-cell RNA sequencing, such as Perturb-seq, offer unprecedented opportunities to uncover cause-and-effect relationships among genes. Nonetheless, Perturb-seq experiments are subject to unobserved factors that, if not properly handled, can severely bias the inferred causal relationships between genes. These latent factors may arise not only from intrinsic molecular features of the regulatory elements, but also from unmeasured genes omitted due to cost-constrained experimental designs. Although methods for analyzing large-scale Perturb-seq data are rapidly maturing, approaches that explicitly account for such unobserved confounders when inferring causal gene networks are still lacking. Here, we propose a novel approach to accurately reconstruct causal gene networks from Perturb-seq data even when important confounders are missing. Our framework leverages proxy and instrumental variable strategies to exploit the rich information embedded in the perturbations, enabling unbiased estimation of the underlying directed acyclic graph (DAG) of gene expression. Applications to both comprehensive synthetic data and real CRISPR interference experiments in K562 cells demonstrate that our method outperforms baseline approaches that lack principled adjustments for unmeasured confounding, yielding more accurate and biologically relevant recovery of the true causal DAGs.

Causal interpretation of pooled single-cell perturbation screens

Jingyi Jessica Li, Fred Hutch Cancer Center / University of Washington

Siyuan Chen (Fred Hutch Cancer Center); Qingyang Wang (University of California, Los Angeles); Jingyi Jessica Li (Fred Hutch Cancer Center)

Pooled single-cell perturbation screens, such as CRISPRi Perturb-seq, combine experimental gene perturbation with single-cell RNA-seq readouts to study gene regulation at high throughput. However, even when the perturbation is experimental, the treatment effect of interest (the causal

effect of perturbing a target gene on the expression of a downstream gene) is not directly observed. A key distinction is between gRNA presence and the realized perturbation in each cell. Different analysis methods represent the realized perturbation differently and can therefore yield different treatment effect estimates and interpretations. Cell state creates an additional challenge: baseline cell state, such as cell-cycle state, may affect both the realized perturbation and its outcome, acting as a confounder, but is usually unobserved. In contrast, cell state inferred from the post-perturbation transcriptome may already be affected by the perturbation, so covariate adjustment may introduce collider bias. Using a CRISPRi Perturb-seq dataset, we compare commonly used methods and find substantial differences in estimated treatment effects across methods and inferred cell states. These results highlight the need to clearly define the treatment, realized perturbation, outcome, cell state, and inferential target, and to use sensitivity analyses when key quantities are unobserved.

Computational and AI-driven design of random heteropolymers as protein mimics

Haiyan Huang, University of California, Berkeley

Shuni Li (University of California, Berkeley); Dylan Webb (University of California, Berkeley); Yun Zhou (University of California, Berkeley); Zhiyuan Ruan (University of California, Berkeley); Haotian Chen (University of California, Berkeley); Ting Xu (University of California, Berkeley); Haiyan Huang (University of California, Berkeley)

Synthetic random heteropolymers (RHPs), composed of a predefined set of monomers, offer a promising strategy for creating protein mimicking materials with tailored biochemical functions. When designed appropriately, RHPs can replicate protein behavior, enabling applications in drug delivery, therapeutic protein stabilization, biosensing, tissue engineering, and medical diagnostics. However, designing RHPs that achieve specific biological functions in a time- and cost-effective manner remains a major challenge. In this talk, I will review this problem and discuss several successful efforts we have made to address it, using statistical, computational, and AI approaches. These include a generalized semi-hidden markov model (GSHMM) and a modified variational autoencoder (VAE) within a semi-supervised framework, designed to capture the structures of critical chemical features as well as individual RHP sequence patterns. These studies highlight the potential of computational approaches to accelerate the rational design of RHPs for a wide range of biological, medical, and healthcare applications.

Session 11: Topics in Online Statistical Inference and Learning

Location: Windsor Terrace

Organizer: Zhimei Ren (University of Pennsylvania)

Chair: Zhimei Ren (University of Pennsylvania)

Presenters: Ying Jin; Isaac Gibbs; Kaizheng Wang; Jiadong Liang

Conformal annotation with randomized expert review: application to online cell-type labeling

Ying Jin, University of Pennsylvania

Xuhe Zhang (University of Pennsylvania), Zhi Huang (University of Pennsylvania), Ying Jin (University of Pennsylvania)

Artificial intelligence (AI) models can accelerate cell-type annotation in digital pathology, but safely scaling their use requires determining—with error control—which model-generated labels can be accepted without expert review. This is challenging in ongoing annotation workflows, where cell morphology and type composition may change across tissue regions and slides, making one-time calibration unreliable. We propose *Conformal Annotation with Randomized Expert review under Distribution shift* (CARED), a distribution-free framework that leverages randomized expert audit to achieve long-run error control for AI-assisted annotation under distribution shift. For each incoming batch, CARED uniformly select a subset of cases for expert review and uses only the acquired labels to calibrate later decisions. Among unreviewed cases, those with singleton prediction sets safely skip expert labeling, while larger sets remain reliable candidate lists for downstream resolution. We prove that this workflow controls long-run cumulative error over both the full stream and the cells outside of the audit sample, without stationarity of exchangeability assumptions. We further enhance the framework by introducing class-conditional guarantee to protect against important but rare cell types, as well as a utility module to target more automatic resolutions, smaller candidate sets, and greater biological coherence, without compromising validity. Experiments on pancreas, lung, and ovary annotation streams, together with a stream constructed from real expert annotations, show that CARED controls the cumulative errors while a substantial fraction of cells—up to 60%—bypass expert review.

Calibrated multi-level quantile forecasting

Isaac Gibbs, University of Toronto

Tiffany Ding (University of California, Berkeley); Isaac Gibbs (University of Toronto); Ryan J. Tibshirani (University of California, Berkeley)

We will discuss an online method that guarantees calibration of quantile forecasts at multiple quantile levels simultaneously. In this work, a sequence of quantile forecasts is said to be calibrated provided that its α-level predictions are greater than or equal to the target value at

an α fraction of time steps, for each level α. Our procedure, called the multi-level quantile tracker (MultiQT), is lightweight and wraps around any point or quantile forecaster to produce adjusted quantile forecasts that are guaranteed to be calibrated, even against adversarial distribution shifts. Critically, it does so while ensuring that the quantiles remain ordered, e.g., the 0.5-level quantile forecast will never be larger than the 0.6-level forecast. Moreover, the method has a no-regret guarantee, implying it will not degrade the performance of the existing forecaster (asymptotically), with respect to the quantile loss. In our experiments, we find that MultiQT significantly improves the calibration of real forecasters in epidemic and energy forecasting problems, while leaving the quantile loss largely unchanged or slightly improved.

MINTS: minimalist Thompson sampling

Kaizheng Wang, Columbia University

Kaizheng Wang

The Bayesian paradigm offers principled tools for sequential decision-making under uncertainty, but its reliance on a probabilistic model for all parameters can hinder the incorporation of complex structural constraints. We introduce a minimalist Bayesian framework that places a prior only on the location of the optimum, while eliminating nuisance parameters through profile likelihood. This yields a generalized posterior that naturally accommodates structural constraints. As a direct instantiation, we develop MINimalist Thompson Sampling (MINTS). For multi-armed bandits with mean constraints, we establish near-optimal non-asymptotic regret guarantees and sharp almost-sure asymptotic regret characterizations. In particular, MINTS attains the classical Lai–Robbins constant in the unstructured setting and automatically adapts to unimodal structure, achieving the sharp constant determined only by the immediate neighbors of the optimal arm.

Optimal training-conditional regret for online conformal prediction

Jiadong Liang, University of Pennsylvania

Jiadong Liang (University of Pennsylvania); Zhimei Ren (University of Pennsylvania); Yuxin Chen (University of Pennsylvania)

We study online conformal prediction for non-stationary data streams subject to unknown distribution drift. While most prior work studied this problem under adversarial settings and/or assessed performance in terms of gaps of time-averaged marginal coverage, we instead evaluate performance through training-conditional cumulative regret. We specifically focus on independently generated data with two types of distribution shift: abrupt change points and smooth drift. When non-conformity score functions are pretrained on an independent dataset, we propose a split-conformal–style algorithm that leverages drift detection to adaptively update calibration sets, which provably achieves minimax-optimal regret. When non-conformity scores are instead trained online, we develop a full-conformal–style algorithm that again incorporates drift detection to handle non-stationarity; this approach relies on stability—rather than permutation symmetry—of the model-fitting algorithm, which is often better suited to online learning under evolving environments. We establish non-asymptotic regret guarantees for our online full conformal algorithm, which match the minimax lower bound under appropriate restrictions on the prediction sets. Numerical experiments corroborate our theoretical findings.

Session 65: Statistical Challenges in Data Science: Privacy, Explanation Errors, and Dynamic Ratings

Location: DUMBO

Organizer: Li-Pang Chen (National Chengchi University)

Chair: Li-Pang Chen (National Chengchi University)

Presenters: Robert Hsuan-Fu Hua; Yidan Shi; Trang Bui

Rating of players by Laplace approximation and dynamic modeling

Robert Hsuan-Fu Hua, Florida State University

Robert Hsuan-Fu Hua (Florida State University); Ruby Chiu-Hsing Weng (National Chengchi University)

The Elo rating system is a simple and widely used method for calculating players' skills from paired comparisons data. Many have extended it in various ways. Yet the question of updating players' variances remains to be further explored. In this paper, we address the issue of variance update by using the Laplace approximation for posterior distribution, together with a random walk model for the dynamics of players' strengths, and a lower bound on players' variances. The random walk model is motivated by the Glicko system, but here we assume nonidentically distributed increments to take care of player heterogeneity. Experiments on men's professional matches showed that the prediction accuracy slightly improves when the variance update is performed. They also showed that new players' strengths may be better captured with the variance update.

A deep learning framework for multistate model for heterogeneous Alzheimer's disease progression and multi-omics risk pathways

Yidan Shi, NYU Langone Health

Yidan Shi (NYU Langone), Ziyue Wang (NYU Langone), Yang Li (NYU Langone), Keenan Walker (NIH), Josef Coresh (NYU Langone), Yongzhao Shao (NYU Langone)

Alzheimer's disease and related dementias (ADRD) exhibit substantial heterogeneity in clinical presentation, underlying biology, and rates of disease progression, yet most existing approaches treat dementia subtypes as mutually exclusive and fail to account for mixed-pathology processes. In this work, we propose a deep learning–enhanced multistate modeling framework to

characterize heterogeneous ADRD progression by integrating longitudinal clinical measures with high-dimensional multi-omics data from the Atherosclerosis Risk in Communities (ARIC) Study. Disease progression is modeled as a finite mixture of multistate Markov processes, allowing individuals to belong probabilistically to latent dementia subtypes with distinct transition dynamics and risk factor effects. Deep learning methods are used to learn low-dimensional representations of genomic and proteomic features, which are incorporated into a multistate time-to-event framework to model subtype-specific risk. This approach enables estimation of individualized disease trajectories, identification of subtype-specific molecular risk pathways, and more accurate characterization of heterogeneous ADRD progression, with the goal of supporting more precise epidemiologic inference and personalized intervention strategies.

Error analysis and correction of Shapley-value-based black-box explanations

Trang Bui, University of Saskatchewan

Ningsheng Zhao (Harvard University), Jiayuan Yu (Amazon), Trang Bui (University of Saskatchewan), Krzysztof Dzieciolowski (Daesys)

An increasingly popular method to explain a black-box machine learning model is Shapley value feature attribution, in which the importance of a feature to a model is quantified taking into account its relationship with other features. However, current methods to calculate the Shapley value feature attribution suffer from possible errors, which result in nonsensible explanations. My collaborators and I develop a framework to analyze such errors and propose novel methods for error correction. We demonstrate the performance of our methods using simulated and real-world datasets.

Session 70: Statistical Analysis of LLM and Representation Learning

Location: Greenpoint

Organizer: Yang Ning (Cornell University)

Chair: Yang Ning (Cornell University)

Presenters: Florentina Bunea; Marten Wegkamp; Young Hyun Cho; Inbeom Lee

Shared-feature ensemble MoE

Florentina Bunea, Cornell University

Florentina Bunea (Cornell University)

Modern representation-learning methods routinely produce high-dimensional embeddings of complex data. While these embeddings provide rich representations of the underlying observations, their induced distributions can themselves be highly complex. This motivates simple and interpretable statistical models that extract shared features across collections of samples. We introduce a Shared-Feature Ensemble Mixture of Experts (MoE), in which a common feature dictionary captures salient regions of the embedding space and sample-specific mixtures describe how these shared features are expressed across samples. Specifically, the model comprises a collection of shared softmax components, each defining a probability distribution over the common feature dictionary, and each sample is represented by a sample-specific mixture of these components. Lifting the resulting feature probabilities to continuous distributions centered at the shared feature locations yields a flexible sample-level representation while preserving a parsimonious structure shared across the ensemble.

Once the shared feature dictionary is identified, estimation reduces to fitting an ensemble of softmax mixtures with common atoms and sample-specific mixing weights. We develop an Expectation-Maximization (EM) algorithm for this problem and establish end-to-end theoretical guarantees. For initialization, we estimate the low-dimensional subspace spanned by the softmax mixture atoms, from which suitable initializations can be generated. When the ensemble exhibits additional structure, however, we show that iterative estimation may not be necessary. Specifically, we develop a procedure for screening the ensemble for near-pure mixtures and, for samples identified as arising from such mixtures, estimate their corresponding dominant atoms. We identify regimes in which these estimates are sufficiently accurate for direct estimation of the softmax mixture components, thereby turning a computationally costly iterative procedure into a screening-and-estimation procedure. When further refinement is desired, the same dominant-atom estimates provide high-quality warm starts for EM.

The EM-algorithm and the method of moments in softmax mixture models

Marten Wegkamp, Cornell University

Xin Bing (Department of Statistical Sciences, University of Toronto); Florentina Bunea (Department of Statistics and Data Science, Cornell University); Jonathan Niles-Weed (Courant Institute and Center for Data Science, New York University); Marten Wegkamp (Department of Statistics and Data Science, Cornell University)

Softmax Mixture Models (SMMs) are discrete K-component mixture models for the probabilities of selecting one of p candidate feature vectors in heterogeneous populations and are widely used in econometrics and scientific applications. Related softmax mixture mechanisms also appear in modern LLM architectures. We provide a theoretical and methodological study of SMMs, focusing on the Expectation-Maximization (EM) algorithm and the Method of Moments (MoM). We show that EM recovers the mixture atoms at the parametric rate, up to logarithmic factors, after O(log N) iterations, provided atom separation is at least of order log K. This improves on separation conditions in existing analyses of EM for high-dimensional Gaussian mixtures. We also develop MoM procedures for parameter and subspace estimation. Although MoM parameter estimates converge more slowly than EM and

can deteriorate with K, they provide provable warm starts for EM and are useful for small K. For general K, we estimate the atom subspace via MoM and recommend running EM from multiple random initializations within this subspace. Finally, as $p \to \infty$, we show that SMMs approximate mixtures of exponential tilts of the feature distribution, yielding asymptotic identifiability.

From black-box AI workflows to principled human–AI collaboration

Young Hyun Cho, Harvard University

Young Hyun Cho (Harvard University); Will Wei Sun (Purdue University)

Artificial intelligence (AI) is increasingly operating as an agent, with multiple AI components working together to solve a given task. As these systems rapidly improve, there is growing interest in how they can be integrated into professional work alongside human experts. We study this question through an iterative generate–verify workflow, in which AI repeatedly generates candidate outputs, evaluates them, and either continues working or returns an answer. A key challenge is understanding how much such a workflow can reliably contribute to human work and what uncertainty remains in its outputs. This is particularly difficult because modern AI systems are largely black-box, making their internal behavior difficult to model directly. We use human expert evaluation before deployment to calibrate the workflow. A human expert evaluates outputs from pilot tasks according to the quality expected for the underlying work, and these evaluations determine which outputs the AI can release, how long the workflow can continue before escalating a task to a human expert, and how often released outputs fall below the human expert's quality standard.

Beyond identifiability: learning causal representations with few environments and finite samples

Inbeom Lee, University of Chicago

Inbeom Lee (University of Chicago); Tongtong Jin (University of Chicago); Bryon Aragam (University of Chicago)

We provide explicit, finite-sample guarantees for learning causal representations from data with a sublinear number of environments. Causal representation learning seeks to provide a rigorous foundation for the general representation learning problem by bridging causal models with latent factor models in order to learn interpretable representations with causal semantics. Despite a blossoming theory of identifiability in causal representation learning, estimation and finite-sample bounds are less well understood. We show that causal representations can be learned with only a logarithmic number of unknown, multi-node interventions, and that the intervention targets need not be carefully designed in advance. Through a careful perturbation analysis, we provide a new analysis of this problem that guarantees consistent recovery of (a) the latent causal graph, (b) the decoder matrix and representations, and (c) unknown intervention targets. Our approach is easily implemented and demonstrated through experiments.

Session 52: New Statistical Insights on Learning and Selection

Location: Navy Yard

Organizer: Rong Ma (Harvard University)

Chair: Rong Ma (Harvard University)

Presenters: Yaqi Duan; Lihua Lei; Lucas Janson

Toward a statistical perspective on LLM posttraining: preference sampling and gradient reweighting

Yaqi Duan, New York University

Yaqi Duan (NYU); Yunzhen Feng (NYU); Julia Kempe (NYU, Meta); Ariel Kwiatkowski (Meta); Kunhao Zheng (Meta); Parag Jain (Meta); Anthony Hartshorn (Meta)

Posttraining—the process of finetuning large language models (LLMs) with human preference or verifiable feedback—has rapidly evolved into a central problem in LLM development, yet its principles remain largely heuristic. This talk explores how statistical thinking can provide structure and rigor to this stage of learning. I will present two studies as steps toward formulating a statistical perspective on posttraining. The first, PILAF, views human preference collection as an optimal experimental design problem, deriving sampling strategies that maximize reward-model information under budget constraints. The second, LENS, reinterprets reinforcement learning with verifiable rewards (RLVR) as a likelihood-based estimation problem, showing how confidence-weighted corrections on negative responses recover gradients otherwise lost in standard policy optimization. Together, these results illustrate how classical statistical reasoning can strengthen the foundations of posttraining data collection and policy-update procedures, advancing efficiency, stability, and theoretical clarity.

Compound selection decisions: an almost SURE approach

Lihua Lei, Stanford University

Jiafeng Chen (Stanford University), Lihua Lei (Stanford University), Timothy Sudijono (University of Pennsylvania), Liyang Sun (University College London & CEMFI), Tian Xie (University of Melbourne)

This paper proposes methods for compound selection decision problems in a Gaussian sequence model. Inspired by Stein's unbiased risk estimate (SURE), we introduce ASSURE, a family of estimators for welfare, defined as the expected utility of a selection decision rule. ASSURE enables robust evaluation of selection decisions. For empirical Bayes decisions derived from random-effects models of unknown parameters, ASSURE estimates their welfare even when the models are misspecified. Optimizing ASSURE-estimated welfare over a pre-specified class of decision rules further borrows strength across noisy

estimates and yields decision rules with favorable regret properties. These regret properties are again robust to potential misspecification of random effects models, thereby robustifying empirical Bayes methods. We apply ASSURE to selecting Census tracts for economic mobility, identifying discriminating firms, and evaluating p-value decision rules in A/B testing.

Chiseling: powerful and valid subgroup selection via interactive machine learning

Lucas Janson, Stanford University

Nathan Cheng (Harvard University); Asher Spector (Flapping Airplanes); Lucas Janson (Stanford University)

In regression and causal inference, controlled subgroup selection aims to identify, with inferential guarantees, a subgroup (defined as a subset of the covariate space) on which the average response or treatment effect is above a given threshold. E.g., in a clinical trial, it may be of interest to find a subgroup with a positive average treatment effect. However, existing methods either lack inferential guarantees, heavily restrict the search for the subgroup, or sacrifice efficiency by naive data splitting. We propose a novel framework called chiseling that allows the analyst to interactively refine and test a candidate subgroup by iteratively shrinking it. The sole restriction is that the shrinkage direction only depends on the points outside the current subgroup, but otherwise the analyst may leverage any prior information or machine learning algorithm. Despite this flexibility, chiseling controls the probability that the discovered subgroup is null (e.g., has a non-positive average treatment effect) under minimal assumptions: for example, in randomized experiments, this inferential validity guarantee holds under only bounded moment conditions. When applied to a variety of simulated datasets and a real survey experiment, chiseling identifies substantially better subgroups than existing methods with inferential guarantees. This is joint work with Nathan Cheng and Asher Spector and is available at https://arxiv.org/abs/2509.19490.

Session 81: Causal and Interpretable Learning for Heterogeneous and Longitudinal Effects in Clinical Studies

Location: Brighton Beach

Organizer: Yue Shentu (Merck & Co)

Chair: Jinchun Zhang (Merck & Co)

Presenters: Zachary Izzo; Min Qian; Iván Díaz

Subgroup discovery with the Cox model

Zachary Izzo, NEC Labs America

Zachary Izzo (NEC Labs America); Iain Melvin (NEC Labs America)

We study the problem of subgroup discovery for survival analysis, where the goal is to find an interpretable subset of the data on which a Cox model is highly accurate. We examine why existing quality functions are insufficient for this problem and introduce two technical innovations: the expected prediction entropy (EPE), a novel metric for evaluating survival models that predict hazard functions, and the conditional rank statistics (CRS), which quantifies individual point deviation from a subgroup's survival time distribution. We study the EPE and CRS theoretically and show they address problems with existing metrics. We then introduce eight algorithms for Cox subgroup discovery. Our main algorithm leverages both the EPE and CRS and has theoretical correctness guarantees in well-specified settings. Empirical evaluation on synthetic and real data confirms our theory, showing our methods recover ground-truth subgroups in well-specified cases and achieve better model fit than naively fitting the Cox model to the entire dataset.

Online learning for personalized policy using longitudinal mobile health data

Min Qian, Columbia University

Wanwan Liang; Min Qian

With the increasing focus on improving personal health and fitness using smart devices and wearables, it is crucial to create a mobile clinical decision support system. In this work, we consider the development of personalized policies that allow different intervention recommendations for individuals with the same observed features. Personalized policy represents a paradigm shift from one decision rule for all users to an individualized decision rule for each user. Aiming to optimize the expected rewards, we propose using a generalized linear mixed modeling framework where population effects and individual deviations from the population effects are modeled as fixed and random effects, respectively, and synthesized to form the personalized policy. We introduce a contextual bandit algorithm to learn the personalized policies. This approach is theoretically justified using a regret bound and illustrated using a mobile health data example.

Computationally and statistically efficient estimation of time-smoothed counterfactual curves

Iván Díaz, New York University Grossman School of Medicine

Herbert P. Susmann, Nicholas T. Williams, Richard Liu, Jessica G. Young, Iván Díaz

Longitudinal causal inference is concerned with defining, identifying, and estimating the effect of a time-varying intervention on a time-varying outcome that is indexed by a follow-up time. In an observational study, Robins's generalized g-formula can identify causal effects induced by a broad class of time-varying interventions. Various methods for estimating the generalized g-formula have been posed for different outcome types, such as a failure event indicator by a specified time (e.g. mortality by 5 year follow-up), as well as continuous or dichotomous/multi-valued outcomes measures at a specified time (e.g. blood pressure in mm/hg or an

indicator of high blood pressure at 5-year follow-up). Multiply-robust, data-adaptive estimators leverage flexible nonparametric estimation algorithms while allowing for statistical inference. However, extant methods do not accommodate time-smoothing when multiple outcomes are measured over time, which can lead to substantial loss of precision. We propose a novel multiply-robust estimator of the generalized g-formula that accommodates time-smoothing over numerous available outcome measures. Our approach accommodates any intervention that can be described as a Longitudinal Modified Treatment Policy, a flexible class suitable for binary, multi-valued, and continuous longitudinal treatments. Our method produces an estimate of the effect curve: the causal effect of the intervention on the outcome at each measurement time, taking into account censoring and non-monotonic outcome missingness patterns. In simulations we find that the proposed algorithm outperforms extant multiply-robust approaches for effect curve estimation in scenarios with high degrees of outcome missingness and when there is strong confounding. We apply the method to study longitudinal effects of union membership on wages.

Parallel Sessions 2: Sunday, November 1, 2026, 1:30 pm–3:00 pm

Session 63: Statistical Foundations of Modern Generative and Decision Models

Location: Salon C

Organizer: Grace Yi (University of Western Ontario)

Chair: Grace Yi (University of Western Ontario)

Presenters: Bei Jiang; Annie Qu; Nhat Ho

A data-augmented predictive framework for generating fair synthetic labels with controlled fairness-faithfulness trade-offs

Bei Jiang, University of Alberta

Junxin Zhang; Enze Shi; Naisyin Wang; Linglong Kong and Bei Jiang

Algorithmic decision systems are increasingly used in socially sensitive domains, raising concerns about bias inherited from historical data. Fair synthetic data generation provides a pre-processing strategy for bias mitigation, but existing methods often rely on black-box generative models whose tuning parameters have limited theoretical interpretation. We propose fDA, a Data-Augmented predictive framework for generating Fair synthetic labels. The framework combines a fairness-enforcing model, which specifies a fair reference conditional law satisfying the desired fairness constraint, with a faithfulness-preserving model, which generates auxiliary variables from observed labels to retain controlled original-label information. Synthetic labels are sampled from the predictive distribution induced by jointly modeling these components, coupled with a tuning mechanism. For continuous labels, the Gaussian working specification yields explicit calibration of the effective noise level to a target population upper bound on unfairness. For both continuous and ordinal labels, the predictive distribution is fully specified for each tuning value, enabling empirical calibration of the achieved fairness–faithfulness trade-off. Theoretically, when the auxiliary variable becomes fully informative, synthetic labels converge to the original labels in probability and distribution; when it becomes non-informative, the mechanism reduces to the fair reference law. Experiments on simulated and real datasets show interpretable trade-offs and improved faithfulness over GAN-based baselines.

Syntax-guided diffusion language models with user-integrated personalization

Annie Qu, University of California, Santa Barbara

Ruqian Zhang, Yijiao Zhang, Juan Shen, Zhongyi Zhu and Annie Qu

Large language models have achieved remarkable progress in generating human-like text, yet their outputs often lack structural diversity, limiting personalized expression. Recent advances in diffusion models offer new opportunities for text generation, surpassing the limitations of autoregressive paradigms. Building on the idea of hierarchical modeling, we develop a syntax-guided diffusion language model that incorporates latent syntactic variables to enhance text quality, diversity, and personalized conditioning. The proposed framework decomposes the text distribution into a syntactic prior and a conditional text distribution, enabling interpretable control over both syntactic and lexical components. We design a cascaded diffusion process for syntax-to-text generation and extend it to a generalized noncascaded formulation with a unified attention mechanism. To achieve fine-grained personalization, we introduce shared-latent representations that integrate information across users to capture style-specific lexical and syntactic patterns, supporting zero-shot inference. Extensive experiments demonstrate that the proposed approach consistently improves multiple empirical measures of fluency, diversity, and stylistic fidelity. Further qualitative analyses highlight its interpretability and flexibility in learning personalized patterns.

Statistical foundation of mixture of experts

Nhat Ho, The University of Texas at Austin

Huy Nguyen; Dung Le; Fanqi Yan; Alessandro Rinaldo; Nhat Ho

Mixture-of-experts architectures have become an important mechanism for scaling modern AI systems, particularly large language models. By routing each input to only a small subset of experts, sparse MoE models can substantially increase model capacity while keeping computational cost manageable. Beyond this computational motivation, sparse routing also raises a fundamental statistical question: when do the learned experts correspond to stable and recoverable functional modules?

In this talk, I will discuss a statistical theory of expert recovery in dense and sparse mixture-of-experts models. A central challenge is that accurate prediction does not necessarily imply that the individual experts are well determined. When several similar experts are simultaneously active, different expert decompositions can induce nearly indistinguishable predictive distributions, making expert specialization, interpretation, fine-tuning, and replacement potentially unstable.

Our theory shows that the difficulty of expert recovery is governed not by the total number of experts in the model, but by the number of similar experts that are simultaneously active for a given input. Top-K routing reduces this local ambiguity, while Top-1 routing removes the within-input mixture ambiguity for the selected expert. These results provide a statistical perspective on sparse routing beyond computational efficiency and clarify its role in supporting stable specialization and modularity in modern AI systems.

Session 30: Modern Statistical Learning and Representation Methods for Complex Biomedical Data

Location: Williamsburg

Organizer: Hai Shu (New York University, Department of Biostatistics)

Chair: Taehyo Kim (New York University)

Presenters: Ruizhi Yuan; Zhe Fei; Zhengwu Zhang; Taehyo Kim

Learning where to look: pathologist-inspired multi-field-of-view evidence retrieval for cell type classification in H&E

Ruizhi Yuan, University of Pittsburgh

Ruizhi Yuan(University of Pittsburgh); Chongyue Zhao(University of Pittsburgh); Tianhao Liu(Tsinghua University); Qian Wang(University of Pittsburgh); Zeqiu Yu(University of Virginia); Lu Tang(University of Pittsburgh); Heng Huang(University of Maryland); Wei Chen(University of Pittsburgh)

Accurate classification of cell types in H&E whole-slide images requires integrating evidence across scales, from cell morphology to the local micro-environment and a broader tissue context. However, existing models rely on a fixed field-of-view that frequently misses decisive out-of-view evidence or dilutes local evidence with irrelevant surroundings. We therefore propose a mixture-of-experts (MoE) framework that retrieves multi-field-of-view evidence via query-guided attention pooling, allowing the model to select the right visual cues across views for cell type classification. To ensure that experts effectively collaborate, we deploy an uncertainty-aware router and train the model with a responsibility-driven objective, assigning per-sample credit based on ground truth support to drive complementary specialization across the experts. We validate on public and in-house spatial-transcriptomic-derived datasets of over 400k cell-centered patches, covering lung and skin tissues and diverse cell types, where our method outperforms strong baselines. Finally, the learned pooling weights provide intuitive evidence maps; a small pilot review by a practicing pathologist suggests that these highlighted regions show qualitative agreement with human pathological reasoning.

Interpretable causal mediation analysis with graph autoencoders

Zhe Fei, University of California, Riverside

Xin Shen, Zhe Fei, Yehua Li

Causal Mediation analysis is essential for understanding how treatment effects propagate through intermediate variables. However, existing methods often rely on restrictive modeling assumptions ill-suited for longitudinal data with many interacting mediators, correlated measurement errors, and mixed variable types. We propose Causal Graph Autoencoders (CGAE), an interpretable, graph-based framework for deciphering causal pathways in longitudinal data with multivariate mediators. CGAE represents treatment, multiple mediators, and outcome within a structural causal model and learns the underlying directed acyclic graph. Causal effects are derived through the learned causal graph, enabling decomposition into total, direct, indirect, and mediator-specific effects. The CGAE framework extended to accommodate binary outcomes by defining effects on the log-odds scale with population-level integration; we develop both Monte Carlo estimation and an efficient probit-based approximation to compute these causal estimands. We evaluate CGAE in extensive simulations, compare it with existing methods, and illustrate its use with data from the Study of Women's Health Across the Nation (SWAN), a decade-long longitudinal study on midlife women.

Generative modeling of point process with application to brain structural connectivity analysis

Zhengwu Zhang, University of North Carolina at Chapel Hill

Zhanhao Yan, Lexin Li, Zhengwu Zhang

Brain structural connectivity is increasingly represented by the continuous cortical locations of white-matter fiber endpoints, yielding spatial point process on a complex product manifold. Generative modeling of such point process can characterize population variability, generate realistic synthetic cohorts, and quantify how connectivity varies with phenotypic traits. However, generative modeling of high-dimensional non-Euclidean point process remains challenging: classical point process models do not readily scale to such domains, while recent deep generative approaches operate largely as black boxes and provide limited support for interpretable population-level analysis. In this article, we propose a modular and statistically principled framework for generative modeling of point process on complex manifold. Specifically, we first

construct an interpretable reduced-rank representation through geometry-aware smoothing, spectral expansion, and principal component analysis. We then refine the resulting coefficients via penalized maximum likelihood evaluated on the raw point observations. We next train a flow matching model entirely in the refined coefficient space, thereby separating geometric representation and statistical estimation from generative learning. Finally, we decode the generated coefficients into valid spatial densities, from which synthetic point patterns can be sampled. We evaluate the empirical performance on standard point process benchmarks, and show that our method achieves competitive generative fidelity while retaining an explicit and interpretable density representation. We then perform an in-depth analysis of the Human Connectome Project data, and show that our approach generates realistic structural connectivity, captures biologically meaningful population variation, and recovers neuroanatomical sex differences that agree with an empirical contrast computed from the observed data.

A false discovery rate control method using a fully connected hidden Markov random field for neuroimaging data

Taehyo Kim, New York University

Taehyo Kim (New York University); Qiran Jia (University of Southern California); Mony J. de Leon (Weill Cornell Medicine); Hai Shu (New York University)

False discovery rate (FDR) control methods are essential for voxel-wise multiple testing in neuroimaging data analysis, where hundreds of thousands or even millions of tests are conducted to detect brain regions associated with disease-related changes. Classical FDR control methods (e.g., BH, q-value, and LocalFDR) assume independence among tests and often lead to high false non-discovery rates (FNR). Although various spatial FDR control methods have been developed to improve power, they still fall short of jointly addressing three major challenges in neuroimaging applications: capturing complex spatial dependencies, maintaining low variability in both false discovery proportion (FDP) and false non-discovery proportion (FNP) across replications, and achieving computational scalability for high-resolution data. To address these challenges, we propose fcHMRF-LIS, a powerful, stable, and scalable spatial FDR control method for voxel-wise multiple testing. It integrates the local index of significance (LIS)-based testing procedure with a novel fully connected hidden Markov random field (fcHMRF) designed to model complex spatial structures using a parsimonious parameterization. We develop an efficient expectation–maximization algorithm incorporating mean-field approximation, the Conditional Random Fields as Recurrent Neural Networks (CRF-RNN) technique, and permutohedral lattice filtering, reducing the time complexity from quadratic to linear in the number of tests. Extensive simulations demonstrate that fcHMRF-LIS achieves accurate FDR control, lower FNR, reduced variability in FDP and FNP, and a higher number of true positives compared to existing methods. Applied to an FDG-PET dataset from the Alzheimer's Disease Neuroimaging Initiative, fcHMRF-LIS identifies neurobiologically relevant brain regions and offers notable advantages in computational efficiency.

Session 64: Emerging Advances in Analysis of Complex Data

Location: Fulton Ferry

Organizer: Wenqing He (University of Western Ontario)

Chair: Wenqing He (University of Western Ontario)

Presenters: Erica EM Moodie; Yinjun Zhao; Eric D. Kolaczyk; Wenqing He

Leveraging partial adherence to improve estimation of adaptive treatment strategies

Erica EM Moodie, McGill University

Chloe Si (McGill University); David A Stephens (McGill University)

Precision medicine is focused on optimizing sequential decision-making; in the statistical literature, this is formalized through dynamic treatment regimes. A flexible approach to estimation relies on inverse probability weighting estimators, however these can exhibit high variability. A key reason for this inefficiency is due to the estimator's requirement of strict alignment between observed and candidate treatment strategies which results in discarding substantial information from individuals with only slight deviations from the candidate treatment strategies. I will introduce novel approaches to relaxing the strict alignment requirement, resulting in estimators that are asymptotically consistent with the target estimand and offer good finite sample properties.

Joint modeling of multimodal mixed-type longitudinal data for dynamic subtyping

Yinjun Zhao, Columbia University

Yinjun Zhao (Columbia University); Seonjoo Lee (Columbia University); Yuanjia Wang (Columbia University)

Longitudinal subtyping requires methods that integrate heterogeneous measurements while distinguishing individual developmental trajectories from changes in subtype membership. We propose Dynamic Mixed INtegrative Data Subtyping (D-MINDS), a latent-variable model that jointly analyzes binary symptoms and continuous cognitive and neuroimaging measures. Item-specific loadings connect these outcomes to a shared multidimensional representation, subject-specific random intercepts and slopes describe individual trajectories, and a hidden Markov chain captures transitions between latent subtypes. An approximate expectation-conditional-maximization procedure combines Pólya–Gamma working updates, forward–backward state smoothing, and closed-form conditional parameter updates. We establish population identifiability and

provide a canonical representation for comparing fitted models. Simulations show improved parameter recovery as the number of items increases and lower held-out classification error than the evaluated competing methods. An application to $11{,}867$ youth across three visits of the Adolescent Brain Cognitive Development Study integrates 29 psychiatric symptoms with 34 cognitive, behavioral, and neural timescale measures. The analysis identifies two states distinguished primarily by attention symptoms; 19.5% of participants change state at least once. The more symptomatic state is associated with poorer school grades and greater reported substance use. D-MINDS provides a framework for relating multimodal latent dimensions to changes in subtype membership during development.

Hybrid GNN–tree models for low-data optical property prediction

Eric D. Kolaczyk, McGill University

Michael Montemurri (McGill University); Shambhavi Tannir (Boston University); Eric D. Kolaczyk (McGill University)

Deep learning has emerged as a powerful tool for in silico molecular property prediction, yet many experimental workflows operate in data-limited regimes where graph neural networks (GNNs) may exhibit instability and poor generalization. In this work, we investigate hybrid modeling strategies that decouple molecular representation learning from downstream regression for optical property prediction of organic molecules. Across multiple datasets, we compare end-to-end directed message passing neural networks (D-MPNNs) with hybrid architectures that pair frozen D-MPNN embeddings with gradient-boosted tree regressors. In small- and medium-sized data regimes, the hybrid models reduce peak absorption wavelength prediction error relative to the original neural prediction head, particularly under scaffold-based distribution shift, without requiring additional quantum-chemical descriptors or further end-to-end optimization. We investigate the origins of these gains from perspectives of variance decomposition, transfer learning, and explainable ML.

Boosting methods for interval censored data with regression and classification

Wenqing He, University of Western Ontario

Yuan Bian (University of Western Ontario); Grace Yi(University of Western Ontario); Wenqing He (University of Western Ontario)

Boosting has gained significant interest across both machine learning and statistical communities. Traditional boosting algorithms, designed for fully observed random samples, often struggle with real-world problems, particularly with interval censored data. This type of data is common in survival analysis where exact event times are unobserved but fall within known intervals. Effective handling of such data is crucial in fields like medical research, reliability engineering, and social sciences. In this work, we introduce novel nonparametric boosting methods for regression and classification tasks with interval censored data. Our approaches leverage censoring unbiased transformations to adjust loss functions and impute transformed responses while maintaining model accuracy. Implemented via functional gradient descent, these methods ensure scalability and adaptability. We rigorously establish their theoretical properties. Our proposed methods not only offer a robust framework for enhancing predictive accuracy in domains where interval censored data are common but also complement existing work, expanding the applicability of existing boosting techniques. Empirical studies demonstrate robust performance across various finite-sample scenarios, highlighting the practical utility of our approaches.

Session 32: Causal Inference and Decision-Making on Networks

Location: Sunset Park

Organizer: Emma Jingfei Zhang (Emory University)

Chair: Emma Zhang (Emory University)

Presenters: Christopher Harshaw; Yubai Yuan; Yuan Zhang; Yinqiu He

On the impossibility of specification testing of interference models based on exposure mappings

Christopher Harshaw, Columbia University

Chao Gao (UChicago); Christopher Harshaw (Columbia); Fredrik Sävje (Uppsala); Yitan Wang (Yale)

Researchers use interference models based on exposure mappings to facilitate estimation of causal effects in randomized experiments with interference. To test the veracity of such models, researchers can use specification tests that aim to detect departures from the stipulated model. However, existing tests suffer from poor power and are often unable to detect important model violations. The main result in this paper is to show that the specification testing problem for exposure mapping models is inherently difficult, and the poor power of existing tests is inescapable. In particular, the worst-case Type I and Type II error rates must sum to one for any specification test of such models, ruling out the existence of a uniformly consistent test. This is the worst-case overall error rate achieved by a naive test that discards all data and arbitrarily rejects the null at random; the testing problem is in this sense impossible. This negative result holds true for all exposure mappings, all sample sizes, for uniformly bounded outcomes, and for alternatives that are maximally separated from the null. While some tests can detect some type of departures from the null model, there will always be relevant departures from the null that are undetectable. Informative specification tests must therefore restrict the alternative model against which they seek to attain power for, beyond the restrictions imposed by the exposure mappings alone. We illustrate this by providing a uniformly consistent test for differentiating no-interference from a network-linear-in-means model.

Double machine learning for heterogeneous spillover

effect

Yubai Yuan, Pennsylvania State University

Yubai Yuan (Pennsylvania State University); Yuanchen Wu (Pennsylvania State University)

Many significant scientific research studies, such as those in epidemiology, and social sciences, involve estimating the effects of treatments, exposures, or interventions on populations where interference between units exists. The influence of one unit's treatment on other units, mediated through network interactions, is commonly referred to as spillover effects. Although the ubiquitous applications and theoretical significance, traditional causal inference methods cannot directly estimate spillover effects due to the interference over network. In addition, the spillover effect is typically heterogeneous in reality. Therefore, it is critical to appropriately specify the exposure mapping and its cross-unit variation. Inspired by the successful application of attention mechanisms in graph neural network, we propose a new method for causal inference on networks that enables researchers to estimate heterogeneous treatment and spillover effects. The proposed method provides interpretable structures for causal estimands and allows for the use of machine learning methods to estimate nuisance components in the models without relying on strong parametric assumptions. Compared to existing methods, the new framework employs a data-adaptive approach to estimate exposure mapping, rather than requiring the functional form to be known a priori.

Conformal network link prediction with false discovery rate control under unstructured missingness

Yuan Zhang, The Ohio State University

Wenqin Du (USC); Wanteng Ma (Upenn); Dong Xia (HKUST); Yuan Zhang (Ohio State); Wen Zhou (NYU)

We propose a new method for predicting multiple missing links in partially observed networks while controlling the false discovery rate (FDR), a largely unresolved challenge in network analysis. The main difficulty lies in handling complex dependencies and unknown missing patterns. We introduce conformal link prediction, a distribution-free procedure grounded in the exchangeability structure of weighted graphon models. Our approach constructs conformal p-values via a novel multi-splitting strategy that restores exchangeability within local test sets, thereby ensuring valid row-wise FDR control, even under unknown missing mechanisms. To achieve FDR control across all missing links, we further develop a new aggregation scheme based on e-values, which accommodates arbitrary dependence across network predictions. Our method requires no assumptions on the missing rates, applies to weighted, unweighted, undirected, and bipartite networks, and enjoys finite-sample theoretical guarantees. Extensive simulations and a real-world data study confirm the effectiveness and robustness of the proposed approach.

Bridging theory and practice: statistical inference for latent space models of networks

Yinqiu He, University of Wisconsin–Madison

Yuang Tian (The Chinese University of Hong Kong); Jiajin Sun (Florida State University); Yinqiu He (University of Wisconsin–Madison)

Networks encode relationship between entities in complex systems, from social connections between people to interactions among genes. To study such data, latent space models offer a flexible way by representing each entity through unobserved attributes that help explain its connections. For these models to support reliable scientific conclusions, we need both accurate parameter estimates and valid measures of uncertainty. However, statistical theory often considers an ideal maximum likelihood estimator, whereas practical analyses use outputs from non-convex iterative algorithms. This yields a fundamental question: when can the solutions produced by algorithms in practice relate to the oracle maximum likelihood estimators? In this talk, I will present a unified framework that connects statistical inference theory with practical computation for latent space models of networks. I will first show how to relax restrictive assumptions on the model's spectral structure, extending the applicability of existing asymptotic theory. I will then introduce adaptive criteria and analytical tools that allow us to study computational algorithms without requiring knowledge of the true model parameters. Together, these results build a foundation for translating statistical inferential theory into reliable uncertainty quantification in practical network analysis.

Session 79: Recent Advances in Empirical Bayes: Theory, Applications, and Methods

Location: Golden Boardroom

Organizer: Yanjun Han (New York University)

Chair: Yanjun Han (New York University)

Presenters: Yandi Shen; Nikolaos Ignatiadis; Tudor Manole; Jacob Shkrob

Empirical Bayes for high-dimensional linear models with random effects

Yandi Shen, Carnegie Mellon University

Zhou Fan (Yale University), Yandi Shen (Carnegie Mellon University), Haoyu Wang (Yale University), Yihong Wu (Yale University)

Empirical Bayes (EB) provides a powerful framework for learning and adapting to latent structure in data. In classical sequence models, where each observation is independently associated with a latent parameter, EB theory and methodology are well understood. However, modern datasets often involve more complex designs in which latent parameters and observations interact in structured ways, raising new statistical and computational challenges.

In this talk, we present recent progress on empirical Bayes inference in high-dimensional settings, focusing on the canonical linear model with random effects. We highlight

two complementary approaches—one based on maximum likelihood and the other on the method of moments—that address both estimation of random effects and posterior inference. We also discuss accompanying information-theoretic lower bounds that characterize the fundamental limits of these problems.

Stein's unbiased risk estimate and Hyvärinen's score matching

Nikolaos Ignatiadis, University of Chicago

Sulagna Ghosh; Nikolaos Ignatiadis; Frederic Koehler; Amber Lee (University of Chicago)

Given a collection of observed signals corrupted with Gaussian noise, how can we learn to optimally denoise them? This fundamental problem arises in both empirical Bayes and generative modeling. In empirical Bayes, the predominant approach is via nonparametric maximum likelihood estimation (NPMLE), while in generative modeling, score matching (SM) methods have proven very successful. In our setting, Hyvärinen's implicit SM is equivalent to another classical idea from statistics—Stein's Unbiased Risk Estimate (SURE). Revisiting SURE minimization, we establish, for the first time, that SURE achieves nearly parametric rates of convergence of the regret in the classical empirical Bayes setting with homoscedastic noise. We also prove that SURE-training can achieve fast rates of convergence to the oracle denoiser in a commonly studied misspecified model. Our empirical results demonstrate the superior performance of SURE-training over NPMLE under misspecification. Collectively, our findings advance SURE/SM as a strong alternative to the NPMLE for empirical Bayes problems in both theory and practice.

Minimax optimality of the NPMLE for Gaussian mixture models

Tudor Manole, Stanford University

Shayan Hundrieser (Yale University), Tudor Manole (Stanford University)

The nonparametric maximum likelihood estimator (NPMLE) is a classical object of study in the field of Gaussian mixture modeling. Beyond applications in which Gaussian mixtures are themselves of primary interest, the NPMLE also serves as an important primitive in various downstream statistical procedures, such as g-modeling estimators in Gaussian empirical Bayes. A question that arises in such applications is to characterize the convergence rate of the NPMLE under the Hellinger distance. In this talk, we present upper bounds for this problem under various moment assumptions on the underlying mixing measures, showing in each case that the NPMLE is minimax optimal. The main technical contribution of our work is a sharp characterization of the local Hellinger bracketing entropy of the Gaussian mixture density family. This talk is based on joint work with Shayan Hundrieser (Yale University).

An empirical Bayes perspective on heteroskedastic mean estimation

Jacob Shkrob, New York University

Yanjun Han, Abhishek Shetty

Towards understanding the fundamental limits of estimation from data of varied quality, we study the problem of estimating a mean parameter from heteroskedastic Gaussian observations where the variances are unknown and may vary arbitrarily across observations. While a simple linear estimator with known variances attains the smallest mean squared error, estimation without this knowledge is challenging due to the large number of nuisance parameters. We propose a simple and principled approach based on empirical Bayes: model the observations as if they were i.i.d. from a normal scale mixture and compute the profile maximum likelihood estimator (MLE) for the mean, treating the nonparametric mixing distribution as nuisance. Our result shows that this estimator achieves near-optimal error bounds across various heteroskedastic models in the literature. In particular, for the subset-of-signals problem where an unknown subset of observations has small variance, our estimator adaptively achieves the minimax rate for all signal sizes, including the sharp phase transition, without any tuning parameters. One of our key technical steps is a sharper metric entropy bound for normal scale mixtures, obtained via Chebyshev approximations on a transformed polynomial basis. This approach yields an improved polylogarithmic, rather than polynomial, dependence on the variance ratio, which could be of independent interest.

Session 20: Reliable Discovery in Complex Data: Advances in High-Dimensional and Interpretable Clustering

Location: Carroll Gardens

Organizer: Ardavan Yazdanbakhsh (City College of New York)

Chair: Boxiang Wang (University of Iowa)

Presenters: Genevera Allen; Yuqi Gu; Hui Zou; Linjun Zhang

Inference for clustering: conformal sets for cluster labels

Genevera Allen, Columbia University

YoonHaeng Hur (Columbia University); Anirban Nath (Columbia University); Genevera Allen (Columbia University)

While clustering is ubiquitously used across science and industry, uncertainty in cluster assignments is rarely quantified with rigorous guarantees. We propose a novel conformal inference framework for clustering that returns confidence sets for cluster labels. The key challenge is that labels are unobserved and estimated from data, so naively using deterministic cluster labels can violate exchangeability and induce severe under-coverage. To address this, we propose split conformal clustering with stochastic labels, which samples labels from soft cluster labels, fits a soft classifier to predict these stochastic labels,

and calibrates conformal scores to construct confidence sets for cluster labels at any query point. We establish a finite-sample lower bound on marginal coverage that reveals how under-coverage is controlled by two properties of the clustering algorithm: consistency of estimated soft labels and replace-one stability. Under mild conditions, we prove asymptotic coverage and verify these conditions for correctly specified parametric mixture models. Simulations for mixture models show that our method attains target coverage with informative set sizes, validating our theoretical results. Applications to clustering cell types in single-cell RNA-seq data demonstrate the practical utility and interpretability of our approach to quantifying cluster label uncertainty.

Covariance projected spectral clustering for high-dimensional anisotropic mixtures: practical algorithm and optimal rates

Yuqi Gu, Columbia University

Chengzhu Huang (Columbia University); Yuqi Gu (Columbia University)

In mixture models, anisotropic noise within each cluster is widely present in real-world data. This work investigates both computationally efficient procedures and fundamental statistical limits for clustering in high-dimensional anisotropic mixtures. We propose a new clustering method, Covariance Projected Spectral Clustering (COPO), which adapts to a wide range of dependent noise structures. We first project the data onto a low-dimensional space via eigen-decomposition of a diagonal-deleted Gram matrix. Our central methodological idea is to sharpen clustering in this embedding space by a covariance-aware reassignment step, using quadratic distances induced by estimated projected covariances. Through a novel row-wise analysis of the subspace estimation step in weak-signal regimes, which is of independent interest, we establish tight performance guarantees and algorithmic upper bounds for COPO, covering both Gaussian noise with flexible covariance and general noise with local dependence. To characterize the fundamental difficulty of clustering high-dimensional anisotropic Gaussian mixtures, we further establish two distinct and complementary minimax lower bounds, each highlighting different covariance-driven barriers. Our results show that COPO attains minimax-optimal misclustering rates in Gaussian settings. Extensive simulation studies across diverse noise structures, along with a real data application, demonstrate the superior empirical performance of our method.

A latent semiparametric mixture model for robust heterogeneity learning in high dimensions

Hui Zou, Johns Hopkins University

Lyuou Zhang (Shanghai University of Finance and Economics), Boxiang Wang (University of Iowa), Lulu Wang (Gilead Sciences), Wen Zhou (New York University) and Hui Zou (Johns Hopkins University)

Learning heterogeneity from high-dimensional data has become a fundamental task across many data-driven fields. Cluster analysis is a widely used unsupervised tool for this purpose, and many existing methods have been adapted to high-dimensional settings through sparsity constraints. In contrast to algorithmic approaches, model-based clustering offers clean theoretical guaranties and computational simplicity, but can be restrictive in practice due to its reliance on specific parametric distributions. We propose a novel latent semiparametric mixture model for clustering, in which the observed data follow a Gaussian mixture after latent coordinate-wise monotone transformations. Furthermore, assuming a sparse optimal clustering rule in the latent space, we develop a new method, *CESME*, tailored for high-dimensional clustering. By allowing unspecified transformations, our framework becomes substantially more flexible and robust than traditional model-based approaches. The clustering error of CESME is shown to approach the optimal clustering error at a rate matching that obtained when the true transformations are known. Extensive simulations and real data analysis confirm that CESME outperforms a broad range of existing high-dimensional clustering methods, including those designed for single-cell RNA sequencing data.

Personalizing black-box models for nonparametric regression with minimax optimality

Linjun Zhang, Rutgers University

Sai Li (Tsinghua University); Linjun Zhang (Rutgers University)

Recent advances in large-scale models, including deep neural networks and large language models, have substantially improved performance across a wide range of learning tasks. The widespread availability of such pre-trained models creates new opportunities for data-efficient statistical learning, provided they can be effectively integrated into downstream tasks. Motivated by this setting, we study few-shot personalization, where a pre-trained black-box model is adapted to a target domain using a limited number of samples. We develop a theoretical framework for few-shot personalization in nonparametric regression and propose algorithms that can incorporate a black-box pre-trained model into the regression procedure. We establish the minimax optimal rate for the personalization problem and show that the proposed method attains this rate. Our results clarify the statistical benefits of leveraging pre-trained models under sample scarcity and provide robustness guarantees when the pre-trained model is not informative. We illustrate the finite-sample performance of the methods through simulations and an application to the California housing dataset with several pre-trained models.

Session 62: Structure Learning and Inference in High-Dimensional Dynamic Systems

Location: Coney Island

Organizer: Mladen Kolar (USC & MBZUAI)

Chair: Paromita Dubey (USC)

Presenters: Kuang-Yao Lee; Junwei Lu; Guo Yu; Mladen Kolar

Graphical modeling of multivariate Cox process

Kuang-Yao Lee, Temple University

Jiehuan Sun (UIC), Kuang-Yao Lee (Temple), Tianxi Cai (Harvard), and Lexin Li (University of California, Berkeley)

A central question in the analysis of multivariate temporal point process data that are simultaneously recorded for multiple subjects is to model the inter-dependency among different point processes. In this work, we propose a novel hidden point process graphical model to tackle this question. Built on a log Gaussian multivariate Cox process, our model hinges on the random intensity functions to wholly determine the point process distribution, and the conditional dependency among the intensity functions characterizes the network's edges. Our approach thus distinguishes itself from most existing vectorial or functional graphical models, by dealing with the entirely unobserved, and non-Hilbertian, random intensity functions. We introduce a precision operator, and establish the one-to-one correspondence between the zero element of this operator and the conditional independence of the intensity function. We develop a hard thresholding-based estimation procedure, and establish its asymptotic consistency. We investigate the empirical performance of the method through both simulations and the analysis of an electronic health records (EHR) data, which infers relatedness of different EHR encounters and reconstructs the healthcare knowledge network for patients with type 2 diabetes.

Preference inference for large language models

Junwei Lu, Harvard T.H. Chan School of Public Health

Nan Lu (University of Pennsylvania); Ethan X. Fang (Duke University); Junwei Lu (Harvard University)

Reinforcement Learning from Human Feedback (RLHF) has become a pivotal paradigm in artificial intelligence to align large models with human preferences. In this paper, we propose a novel statistical framework to simultaneously conduct the online decisionmaking and statistical inference on the optimal model using human preference data based on dynamic contextual information. Our approach introduces an efficient decision strategy that achieves both the optimal regret bound and the asymptotic distribution of the estimators. We will also talk about its application to Lean formalization for statistical theory.

Convex additive Gaussian graphical regression

Guo Yu, University of California, Santa Barbara

Ruobin Liu; Guo Yu

Gaussian graphical models (GGMs) are widely used to recover the conditional independence structure among random variables. Recent work has sought to incorporate auxiliary covariates to improve estimation, particularly in applications such as co-expression quantitative trait locus (eQTL) studies, where both gene expression levels and their conditional dependence structure may be influenced by genetic variants. Existing approaches to covariate-adjusted GGMs either restrict covariate effects to the mean structure or lead to nonconvex formulations when jointly estimating the mean and precision matrix. In this paper, we propose a convex framework that simultaneously estimates the covariate-adjusted mean and precision matrix via a natural parametrization of the multivariate Gaussian likelihood. The resulting formulation enables joint convex optimization and yields improved theoretical guarantees under high-dimensional scaling, where the sparsity and dimension of covariates grow with the sample size. We support our theoretical findings with numerical simulations and demonstrate the practical utility of the proposed method through a reanalysis of an eQTL study of glioblastoma multiforme and an analysis of diet on the human gut microbiome.

Uncovering hidden dynamics: inference in high-dimensional Markov-switching ODEs

Mladen Kolar, USC & MBZUAI

Katherine Tsai (University of California, Berkeley); Mladen Kolar (USC & MBZUAI); Sanmi Koyejo (Stanford University)

Complex systems often evolve continuously while switching between hidden regimes, each with its own network of interactions. How can we recover these networks and their switching dynamics from noisy, discrete observations? In this talk, I will present a framework for inference in high-dimensional Markov-switching ordinary differential equations. The model combines a latent continuous-time Markov chain with sparse, nonlinear dynamics, allowing directed interactions to change across regimes. Our estimation procedure uses wavelet smoothing to reconstruct latent trajectories and a regularized expectation-maximization algorithm to estimate regime-specific networks and transition rates. I will discuss convergence guarantees, finite-sample error bounds, and conditions for graph recovery, highlighting the challenges posed by temporal dependence, hidden states, and trajectory reconstruction. Simulations illustrate the method's ability to recover switching dynamics and network structure. An application to resting-state fMRI explores differences in brain-state transitions between children with ADHD and typically developing children.

Session 69: From Spatial Patterns to Biological Intelligence: Statistical Inference for Spatial Omics

Location: Park Slope

Organizer: Honglang Wang (Indiana University Indianapolis)

Chair: Wenpin Hou (Duke University)

Presenters: Dylan Cable; Wenpin Hou; Ying Ma; Lulu Shang

Empirical-Bayes spatial priors for cell-type identification

Dylan Cable, University of Michigan

Tom Liu (University of Michigan) and Dylan Cable (University of Michigan)

In spatial transcriptomics, spatial context can help identify cell types when cellular evidence is weak. We introduce spatialTV, an empirical Bayes model which learns a spatial cell-type prior by penalized marginal likelihood and combines it with per-location likelihoods for posterior calls. We develop an estimator using alternating FISTA and linear programming, with conditional stability and objective-gap bounds. Simulations with sharp and gradual spatial priors show improved identification over likelihood alone, concentrated at low read depth. Paired experiments hold focal reads fixed and show how informative surroundings improve prior recovery and individual calls. In cerebellum, hippocampus and a supplementary kidney applications, posterior calls retain within-domain heterogeneity and change chiefly where local likelihoods are weak.

Modeling continuous spatial dynamics in complex tissues

Wenpin Hou, Duke University

Wenpin Hou (Duke University, Columbia University), Won Eui Hong (Columbia University), Zhicheng Ji (Duke University)

Spatially resolved molecular data provide new opportunities to characterize how biological states vary continuously across complex tissues. Most existing methods focus on detecting discrete spatial domains based on similarities in gene expression and spatial location. However, continuous processes, such as cellular development and differentiation, may span multiple cell types and tissue regions and therefore cannot be fully represented by discrete domains. We developed SpatialCD, a graph-based computational method that dissects tissues into continuous spatial domains. Each domain represents a distinct developmental process transitioning across multiple cell types, unlike conventional spatial domains, which often correspond to individual cell types or discrete tissue regions. SpatialCD also identifies genes with differential temporal patterns across these domains, highlighting potential drivers of spatial heterogeneity in developmental processes. Applications to multiple datasets demonstrate its ability to resolve complex tissue structures and identify genes with distinct spatiotemporal patterns. I will also discuss broader computational strategies for modeling continuous spatial dynamics across molecular modalities, including emerging approaches for predicting spatial epigenomic states. Together, these methods provide a flexible framework for investigating how transcriptional and epigenetic programs vary across space and dynamic biological processes.

From cell types to pathways: scalable and interpretable models for spatial omics

Ying Ma, Brown University

Ying Ma

Recent advances in spatial omics technologies enable molecular profiling of tissues at increasingly high resolution and across multiple modalities, creating new opportunities to study cellular organization, tissue heterogeneity, and molecular function in their native spatial context. At the same time, these data present substantial computational challenges in cell type annotation, multimodal integration, uncertainty quantification, and spatially resolved functional analysis. In this talk, I will present recent statistical and machine learning methods developed in our lab for analyzing complex spatial omics data. Our work addresses problems including reference-based cell type annotation with novel population discovery, integration of spatial multi-omics data, and spatial gene set and pathway activity analysis. Across these problems, we focus on developing methods that are scalable, interpretable, and capable of preserving spatial and modality-specific structure. Together, these approaches illustrate how computational modeling can help translate increasingly complex spatial measurements into interpretable insights into tissue organization and biological function.

Prototype-driven fusion of pathology and spatial transcriptomics for interpretable survival prediction

Lulu Shang, MD Anderson Cancer Center

Lihe Liu, Xiaoxi Pan, Yinyin Yuan, Lulu Shang

Whole slide images (WSIs) enable weakly supervised prognostic modeling via multiple instance learning (MIL). Spatial transcriptomics (ST) preserves in situ gene expression, providing a spatial molecular context that complements morphology. As paired WSI-ST cohorts scale to population level, leveraging their complementary spatial signals for prognosis becomes crucial; however, principled cross-modal fusion strategies remain limited for this paradigm. To this end, we introduce PathoSpatial, an interpretable end-to-end framework integrating co-registered WSIs and ST to learn spatially informed prognostic representations. PathoSpatial uses task-guided prototype learning within a multi-level experts architecture, adaptively orchestrating unsupervised within-modality discovery with supervised cross-modal aggregation. By design, PathoSpatial substantially strengthens interpretability while maintaining discriminative ability. We evaluate PathoSpatial on a triple-negative breast cancer

cohort with paired ST and WSIs. PathoSpatial delivers strong and consistent performance across five survival endpoints, achieving superior or comparable performance to leading unimodal and multimodal methods. PathoSpatial inherently enables post-hoc prototype interpretation and molecular risk decomposition, providing quantitative, biologically grounded explanations, highlighting candidate prognostic factors. We present PathoSpatial as a proof-of-concept for scalable and interpretable multimodal learning for spatial omics-pathology fusion.

Session 101: Advances in Transfer Learning

Location: Windsor Terrace

Organizer: Maryclare Griffin (University of Massachusetts Amherst)

Chair: Maryclare Griffin (University of Massachusetts Amherst)

Presenters: Donatello Telesca; Nathan Wycoff; Elynn Chen; Ye Tian

Bayesian transfer learning for sparse regression

Donatello Telesca, University of California, Los Angeles

Donatello Telesca, Parsa Jamshidian

We develop gBLAST, a Bayesian transfer learning method for high-dimensional sparse generalized regression that selectively borrows information from multiple external datasets. The method jointly models target regression coefficients and source-specific deviations, using global-local shrinkage priors to distinguish sources whose effects are compatible with the target population from those requiring larger departures. Data augmentation provides a conditionally Gaussian representation of the likelihood, enabling efficient Bayesian computation in high-dimensional settings. Source compatibility is represented through a latent mixture distribution for source-specific global shrinkage parameters, yielding posterior probabilities that quantify the extent to which each source is compatible with the target. We study the estimation, prediction, source-classification, and interval-estimation properties of the proposed method through simulation under varying degrees of source-target similarity. We also apply Logistic BLAST to genomic data from multiple prostate cancer cohorts to predict late genitourinary toxicity following stereotactic body radiation therapy. The results demonstrate the potential of the proposed framework to improve prediction and inference by borrowing information across related but heterogeneous datasets while accounting for uncertainty in source compatibility.

Formal Bayesian transfer learning via the total risk prior

Nathan Wycoff, University of Massachusetts Amherst

Nathan Wycoff (University of Massachusetts Amherst), Ali Arab (Georgetown University), Lisa Singh (Georgetown University)

In analyses with severe data-limitations, augmenting the target dataset with information from ancillary datasets in the application domain, called source datasets, can lead to significantly improved statistical procedures. Existing methods for such transfer learning struggle to deal with situations where the source datasets are limited and not guaranteed to be well-aligned with the target dataset. A typical strategy is to use the empirical loss minimizer on the source data as a prior mean for the target parameters. Our key conceptual contribution is to use a risk minimizer conditional on source parameters instead. This allows us to construct a single joint posterior for all parameters from the source datasets as well as the target dataset. As a consequence, we benefit from full Bayesian uncertainty quantification and can perform model averaging via Gibbs sampling over indicator variables governing the inclusion of each source dataset. We show how a particular instantiation of our prior leads to a Bayesian Lasso in a transformed coordinate system and discuss computational techniques to scale our approach to moderately sized datasets. We discuss connections between the Maximum a Posteriori estimate associated with our approach and the recently proposed Trans-Lasso method and demonstrate that the MAP estimator MSE-dominates the Trans-Lasso in the normal means setting when there is no regularization on the source datasets. Finally, we perform numerical experiments finding that our approach provides superior predictive performance relative to frequentist and Bayesian baselines on a genetics application, especially when the source data are limited.

What makes an RL sample transferable? A statistical theory of transfer Q-learning

Elynn Chen, New York University

Elynn Chen (New York University)

In business, healthcare, and education, sequential decisions must often be learned for a target population with little data, while abundant data exist from related populations. Transfer learning is the natural remedy, and in regression the recipe is well understood: pool source and target samples, then correct for a low-complexity difference. This talk asks whether the same recipe works for reinforcement learning (RL), and answers: only sometimes. In a Markov decision process the regression response at each stage is not the observed reward but a Bellman target that embeds the source's future value under the source's transition. Pooling therefore introduces a bias that is absent from supervised transfer and that persists even when reward functions are identical across populations; it vanishes only at the final stage of a finite horizon. This observation leads to a precise notion of a transferable RL sample and to a re-weighted targeting procedure that restores Bellman alignment: source responses are re-targeted to the target's value function and re-weighted by a transition density ratio, cascading backward through the horizon.

We establish convergence rates for the resulting transfer

Q-learning estimator with deep neural network function approximation, expressed through two "knobs" of task discrepancy—the reward gap and the transition density ratio—and show exactly when source data accelerate learning over single-task rates. The same principle extends to stationary MDPs with iterative Q-learning, to online RL via one-step Bellman alignment, and to composite MDPs with transferable transition components. Experiments on synthetic environments and an ICU sepsis cohort from MIMIC-III confirm the theory.

Related papers: Deep Transfer Q-Learning for Offline Non-Stationary Reinforcement Learning. J. Chai, E. Chen, J. Fan. arXiv:2501.04870. https://arxiv.org/abs/2501.04870 Transfer Q-Learning for Finite-Horizon Markov Decision Processes. E. Chen, S. Li, M. I. Jordan. Electronic Journal of Statistics, 2025. https://doi.org/10.1214/25-EJS2459 Data-Driven Knowledge Transfer in Batch Q-Learning. E. Chen, X. Chen, W. Jing. Journal of the American Statistical Association, 2026. https://doi.org/10.1080/01621459.2025.2603731 One-Step Bellman Alignment Enables Provably Efficient Transfer in Online RL. E. Chen, E. Zhang, J. Chai, Y. Yan. arXiv:2601.21924. https://arxiv.org/abs/2601.21924 Transition Transfer Q-Learning for Composite Markov Decision Processes. J. Chai, E. Chen, L. Yang. ICML, 2025. https://proceedings.mlr.press/v267/chai25b.html

Cross-population PRS prediction: transfer learning with sparse and correlated structures

Ye Tian, Pennsylvania State University

Ye Tian (Pennsylvania State University), Yueqian Jing (Yale University), Hongyu Zhao (Yale University)

Cross-population polygenic risk score (PRS) prediction has recently received substantial attention for their potential to improve prediction across diverse populations, particularly when leveraging a large population dataset to improve prediction in a smaller population. Many new cross-population PRS methods are proposed each year, yet the theoretical understanding of this problem and these methods remains very limited. In this paper, we study this problem through a high-dimensional sparse linear regression framework, in which each population has both population-specific independent signals and shared signals that are correlated through genetic correlation. Our theory characterizes how optimal estimation and prediction performance is governed by genetic correlation, sample sizes, signal variances, and noise variances. We show that the Bayes posterior mean with plugged-in prior estimates and hierarchical Bayes estimators achieve the optimal performance up to logarithmic factors under certain conditions, whereas some transfer learning methods based on global penalization generally do not.

Session 53: Advances in Machine Learning for Neuroimaging and Behavioral Data in Mental Health Research

Location: DUMBO

Organizer: Yuanjia Wang (Columbia University)

Chair: Yuan Bian (Columbia University)

Presenters: Xingche Guo; Shanghong Xie; Seonjoo Lee; Haochang Shou

Statistical modeling for human behavioral task data in computational psychiatry

Xingche Guo, University of Connecticut

Xingche Guo (University of Connecticut); Yuan Bian (Columbia University); Donglin Zeng (University of Michigan); Yuanjia Wang (Columbia University)

Human behavioral tasks offer a window into reward learning, attention, and decision-making in psychiatric disorders. Translating observed choices and response times into interpretable cognitive characteristics, however, requires models that account for individual differences and changing strategies. Motivated by major depressive disorder, this talk presents a progression of statistical approaches for understanding these behavioral processes. I will discuss how reinforcement learning models can be extended to distinguish engaged decision-making from lapses, how drift-diffusion models incorporate response times and the speed–accuracy trade-off, and how shared latent factors connect information across multiple tasks. Applications to the probabilistic reward and Flanker tasks illustrate how these approaches reveal differences in engagement and decision dynamics between individuals with depression and healthy controls. I will conclude with opportunities for integrating behavioral tasks and neuroimaging data. Together, these developments highlight the role of statistical modeling in characterizing cognitive heterogeneity and advancing understanding of mental disorders through computational psychiatry.

Dynamic system for modelling disease progression with temporal registration

Shanghong Xie, University of South Carolina

Shanghong Xie (University of South Carolina); Zexi Cai (Columbia University); Yuanjia Wang (Columbia University)

Neurodegenerative diseases, such as Alzheimer's disease, are chronic and progressive, and the heterogeneity of disease progression and absence of a common reference timeline complicate the analysis of longitudinal cognitive decline. We propose a dynamic disease progression model based on ordinary differential equations that describes latent disease dynamics while allowing an external event, such as treatment initiation, to alter the progression mechanism. Subject-specific initial conditions account for heterogeneity in baseline disease state, and covariate-dependent temporal registration functions align individual trajectories to a common internal time scale.

Both parametric and nonparametric forms of the ordinary differential equations are considered. Variational inference jointly estimates the unknown initial conditions, progression dynamics and registration functions, and the resulting estimators are consistent, with a convergence rate established when the progression parameters are Euclidean. Simulations show accurate parameter estimation and prediction of future marker trajectories, while temporal registration substantially reduces error in the estimated time scale. In an application to the Alzheimer's Disease Neuroimaging Initiative, the model captures the longitudinal progression of multiple cognitive markers. The estimated registration functions indicate more rapid cognitive deterioration in females than in males.

Inference for explainable AI with longitudinal data using gated recurrent units with decay (GRU-D)

Seonjoo Lee, Columbia University

Linda Lin (Columbia University), Hyun Kim (Columbia University), Seonjoo Lee (Columbia University)

Variable importance scores are commonly used in explainable artificial intelligence (XAI), but most implementations provide only heuristic rankings rather than rigorous statistical inferences. Previous work proposed inference methods for deep learning and machine learning. However, these are not directly applicable to longitudinal deep learning models, particularly Gated Recurrent Units with Decay (GRUD) for irregularly sampled longitudinal outcomes with missingness. We construct permutation-based hypothesis tests for the contribution of individual predictors to the conditional distribution of the binary outcome over time. Simulation experiments show that the tests achieve low type I error and high power for detecting truly influential predictors. We apply the method to Alzheimer's Disease Neuroimaging Initiative FDG PET data to study how regional brain metabolism predicts the longitudinal course of neuropsychiatric symptoms. The results isolate a subset of metabolically vulnerable regions whose inferred importance is for NPS symptoms.

Personalized trajectory modeling and network enrichment testing in neuroimaging

Haochang Shou, University of Pennsylvania

Vasiliki Tassopoulou (University of Pennsylvania); Noah Hillman (University of Pennsylvania); Christos Davatzikos (University of Pennsylvania); Taki Shinohara (University of Pennsylvania); Haochang Shou (University of Pennsylvania)

Understanding how brain markers change over time and how the brain relates to behavior are crucial for capturing dynamics in brain development, disease progression, and treatment effects. In this talk, we will first introduce a personalized deep kernel regression framework for forecasting longitudinal brain biomarkers. Estimating individual brain trajectories is complicated by biological variability, scanner inconsistencies, and sparse or irregular measurements. Our approach integrates a population-level model that utilizes Deep Kernel Gaussian Process modeling and high dimensional predictors at baseline trained from a large, diverse cohort with a subject-specific model, combined through Adaptive Shrinkage Estimation. We benchmark our methods and demonstrate superior predictive performance, and validation across three external neuroimaging studies confirms robustness across clinical contexts. Next we introduce NETDOM, a permutation-based approach for testing whether brain-behavior associations are meaningfully enriched within a region of interest. We show that NETDOM properly controls the type I error rate while achieving greater statistical power when enrichment occurs in a subset of in-network locations. Applied to two large-scale neurodevelopmental cohorts, NETDOM effectively detects enriched associations between structural and functional brain measures and neurocognitive performance.

Session 77: Novel Methods in Learning Dependent and Dynamical Data

Location: Greenpoint

Organizer: Wen Zhou (New York University)

Chair: Wen Zhou (New York University)

Presenters: Zhao Ren; Zhengyuan Zhu; Jiawei Ge; Hai Shu

Statistical inference for large Potts models

Zhao Ren, University of Pittsburgh

Fan Yang (University of Pittsburgh); Yinan Lin (Chongqing Normal University); Wen Zhou (New York University); Zhao Ren (University of Pittsburgh)

The Potts model is a fundamental graphical model for multistate categorical data, yet statistical inference for edge interactions, represented by matrix-valued parameters, in high-dimensional settings remains largely unexplored. This paper develops a likelihood-based inference framework for high-dimensional Potts models. Building on penalized node-wise multinomial regression, we propose a block-wise debiasing procedure that constructs one-step corrected estimators via an orthogonalized score matrix derived from the likelihood. The resulting estimators are shown to be asymptotically normal, enabling valid chi-square tests with asymptotic control of type I error and nontrivial power against local alternatives. To facilitate large-scale inference, we further develop an FDR-controlled multiple testing procedure and a simultaneous testing method for collections of interaction blocks. The multiple testing procedure is shown to asymptotically control the FDR, while the simultaneous test is calibrated via a bootstrap approximation for maxima of dependent quadratic forms, supported by Gaussian approximation theory. Our analysis introduces new technical tools, including a quadratic-to-quadratic comparison result and anti-concentration bounds for maxima of chi-square-type statistics. The effectiveness of the proposed methods is demonstrated through simulation studies.

Online change-point detection using foundation probabilistic forecasting models

Zhengyuan Zhu, Iowa State University

Alan Moore (Iowa State University), Zhengyuan Zhu (ISU), Lynna Chu (ISU)

Online change-point detection is a fundamental problem in sequential monitoring, where the goal is to detect distributional changes in a time series as quickly as possible while controlling false alarms. We propose a general framework for online change-point detection that leverages foundation probabilistic forecasters, such as Chronos-2, which can capture complex temporal structure including trend, seasonality, heteroskedasticity, and nonlinear dependence. Our approach transforms each incoming observation relative to its predictive quantile forecast into an approximately standard Gaussian monitoring statistic, which can then be used in classical sequential detection procedures such as CUSUM and MOSUM. To reduce adaptation of the forecaster to post-change observations, we introduce a buffer between the forecasting context window and the forecast target. Because this design can induce serial dependence in the monitoring statistics, we develop a parametric boot- strap procedure to calibrate critical thresholds and control false alarms in practice. Extensive simulations across data-generating processes involving non-normality, trend, seasonality, heteroskedasticity, and autoregressive dependence show that the proposed method achieves competitive and often superior detection performance relative to benchmark procedures, while requiring no explicit specification of a parametric pre-change model.

Covariates-adjusted mixed-membership estimation: a novel network model with optimal guarantees

Jiawei Ge, University of California, Berkeley

Jianqing Fan (Princeton University), Jiawei Ge (University of California, Berkeley), Jikai Hou (Morgan Stanley)

This talk addresses the problem of mixed-membership estimation in networks, where the goal is to efficiently estimate the latent mixed-membership structure from the observed network. Recognizing the widespread availability and valuable information carried by node covariates, we propose a novel network model that incorporates both community information, as represented by the Degree-Corrected Mixed Membership (DCMM) model, and node covariate similarities to determine connections.

We investigate the regularized maximum likelihood estimation (MLE) for this model and demonstrate that our approach achieves optimal estimation accuracy for both the similarity matrix and the mixed-membership, in terms of both the Frobenius norm and the entrywise loss. Since directly analyzing the original convex optimization problem is intractable, we employ nonconvex optimization to facilitate the analysis. A key contribution of our work is identifying a crucial assumption that bridges the gap between convex and nonconvex solutions, enabling the transfer of statistical guarantees from the nonconvex approach to its convex counterpart. Importantly, our analysis extends beyond the MLE loss and the mean squared error (MSE) used in matrix completion problems, generalizing to all the convex loss functions. Consequently, our analysis techniques extend to a broader set of applications, including ranking problems based on pairwise comparisons.

Finally, simulation experiments validate our theoretical findings, and real-world data analyses confirm the practical relevance of our model.

Decomposition of common and distinctive latent factors for multi-view high-dimensional data

Hai Shu, New York University

Hai Shu (New York University); Hongtu Zhu (University of North Carolina at Chapel Hill)

Modern biomedical studies often collect multi-view data, such as multi-modal imaging and/or multi-omics data. A typical approach to the joint analysis of multiple high-dimensional data views is to decompose each view's data matrix into three parts: a low-rank common-source matrix generated by common latent factors of all data views, a low-rank distinctive-source matrix generated by distinctive latent factors of the corresponding data view, and an additive noise matrix. Existing decomposition methods often focus on the uncorrelatedness between the common latent factors and distinctive latent factors, but inadequately address the equally necessary uncorrelatedness between distinctive latent factors from different data views. We propose a novel decomposition method, called Decomposition of Common and Distinctive Latent Factors (D-CDLF), to effectively achieve both types of uncorrelatedness. Consistent estimators of our D-CDLF method are established with good finite-sample numerical performance. The superiority of D-CDLF over state-of-the-art methods is also corroborated through both simulation studies and an application to multi-view data from the Alzheimer's Disease Neuroimaging Initiative.

Parallel Sessions 3: Sunday, November 1, 2026, 3:30 pm–5:00 pm

Session 95: Statistical Foundations and Interpretability for LLM Evaluation and Reasoning

Location: Salon C

Organizer: Yuqi Gu (Columbia University)

Chair: Yuqi Gu (Columbia University)

Presenters: Yiqun Chen; Sang Truong; Zhiyu Xu; Zhuoran Yang

Valid and efficient model evaluation with LLM-as-a-judge

Yiqun Chen, Johns Hopkins University

Yiqun Chen (Johns Hopkins University); Sizhu Lu (University of California, Berkeley); Sijia Li (University of California, Los Angeles); Moran Gu (Johns Hopkins University); Shengyi Li (Johns Hopkins University)

Reliable model evaluation is increasingly bottlenecked by the scarcity of high-quality expert annotations. A common solution is to rely on surrogate labels such as large language model (LLM)-generated labels in place of ground-truth annotations. When the labels used to compute evaluation metrics are imperfect, naively relying on these labels can lead to systematic bias and underestimated uncertainty in model evaluation.

We propose a unified statistical framework for unbiased model evaluation using noisy labels and a limited budget of ground-truth verifications. Our approach leverages the semiparametric efficiency theory and can be applied to commonly used metrics including accuracy, F1, correlation, and the Adjusted Rand Index. We demonstrate that correcting for label noise can be formulated as a conditional expectation estimation problem, enabling us to derive estimators that are both consistent and asymptotically efficient. We validate our framework through experiments on real-world datasets, demonstrating its effectiveness in correcting bias and properly quantifying uncertainty in LLM-based evaluation pipelines.

Statistical foundations of AI measurement science for large language models

Sang Truong, Stanford University

Sang Truong (Stanford University)

Evaluating large language models requires statistical methods that account for both limited evaluation budgets and imperfections in benchmark questions. Average benchmark scores can obscure differences in question difficulty, while flawed questions can undermine the interpretation of model performance. In this talk, I present two complementary approaches to these challenges. First, I discuss model-based evaluation, which uses item response theory and content-based prediction of question difficulty to support efficient capability estimation and adaptive testing. Second, I describe a framework for benchmark revision that analyzes model response patterns to identify potentially problematic questions for expert review. This approach makes explicit the measurement assumptions underlying aggregate scores and uses departures from their statistical implications to guide further investigation. Together, these studies illustrate how latent variable modeling and statistical diagnostics can improve the efficiency and reliability of language model evaluation, while clarifying the assumptions on which benchmark-based conclusions depend.

Latency-response theory model: evaluating large language models via response accuracy and chain-of-thought length

Zhiyu Xu, Department of Statistics, Columbia University

Zhiyu Xu (Columbia University); Jia Liu (Columbia University); Yixin Wang (University of Michigan); Yuqi Gu (Columbia University)

The proliferation of Large Language Models (LLMs) necessitates valid evaluation methods to provide guidance for both downstream applications and actionable future improvements. The Item Response Theory (IRT) model with Computerized Adaptive Testing has recently emerged as a promising framework for evaluating LLMs via their response accuracy. Beyond simple response accuracy, LLMs' chain of thought (CoT) lengths serve as a vital indicator of their reasoning ability. To leverage the CoT length information to assist LLM evaluation, we propose the **La**tency-**R**esponse **T**heory (LaRT) model, which jointly models both the response accuracy and CoT length by introducing a key correlation parameter between the latent ability and the latent speed. We derive an efficient stochastic approximation Expectation-Maximization algorithm for parameter estimation. We establish rigorous identifiability results for the latent ability and latent speed parameters to ensure the statistical validity of their estimation. Through both theoretical asymptotic analyses and simulation studies, we demonstrate LaRT's advantages over IRT in terms of superior estimation accuracy and shorter confidence intervals for latent trait estimation. To evaluate LaRT in real data, we collect responses from diverse LLMs on popular benchmark datasets. We find that LaRT yields different LLM rankings than IRT and outperforms IRT across multiple key evaluation metrics including predictive power, item efficiency, ranking validity, and LLM evaluation efficiency.

On the mechanism and dynamics of modular addition: Fourier features, lottery ticket, and grokking

Zhuoran Yang, Yale University

Jianliang He, Leda Wang, Siyu Chen, Zhuoran Yang

We present a comprehensive analysis of how two-layer neural networks learn features to solve the modular addition task. Our work provides a full mechanistic interpretation of the learned model and a theoretical explanation of its training dynamics. While prior work has

identified that individual neurons learn single-frequency Fourier features and phase alignment, it does not fully explain how these features combine into a global solution. We bridge this gap by formalizing a diversification condition that emerges during training when overparametrized, consisting of two parts: phase symmetry and frequency diversification. We prove that these properties allow the network to collectively approximate a flawed indicator function on the correct logic for the modular addition task. While individual neurons produce noisy signals, the phase symmetry enables a majority-voting scheme that cancels out noise, allowing the network to robustly identify the correct sum. Furthermore, we explain the emergence of these features under random initialization via a lottery ticket mechanism. Our gradient flow analysis proves that frequencies compete within each neuron, with the "winner" determined by its initial spectral magnitude and phase alignment. From a technical standpoint, we provide a rigorous characterization of the layer-wise phase coupling dynamics and formalize the competitive landscape using the ODE comparison lemma. Finally, we use these insights to demystify grokking, characterizing it as a three-stage process involving memorization followed by two generalization phases, driven by the competition between loss minimization and weight decay.

Session 66: Statistical Learning for Causal Inference: From Estimation to Decision-Making

Location: Williamsburg

Organizer: Qilu Yu (NIH, National Center for Complementary and Integrative Health)

Chair: Qilu Yu (NIH, National Center for Complementary and Integrative Health)

Presenters: Minxin Lu; Nikki L. B. Freeman; Foroogh Shamsi; Yunxun Wang

Causal inference for individualized treatment decisions

Minxin Lu, AstraZeneca Pharmaceuticals LP

Minxin Lu (AstraZeneca Pharmaceuticals LP)

Randomized trials and observational studies typically report average treatment effects, yet clinical decisions are made for individual patients. This presentation offers an accessible introduction to how causal inference supports precision medicine by shifting the target from the average treatment effect to the conditional average treatment effect and, ultimately, to individualized treatment rules. We review the potential outcomes framework, the key identification assumptions, and the distinction between prognostic and effect-modifying variables. We will show machine-learning approaches for estimating treatment effect heterogeneity, including meta-learners and causal forests, and how these estimates translate into treatment rules whose value can be evaluated against one-size-fits-all rules. We also illustrate common pitfalls and close with practical challenges for real-world applications and a checklist for responsible translation into clinical decision-making.

Uncertainty propagation and quantification for individualized treatment decisions: a Bayesian approach

Nikki L. B. Freeman, Duke University

Nikki L. B. Freeman (Duke University); Sophia Yazzourh (McGill); Emmanuel Rockwell (University of North Carolina at Chapel Hill); Michael R. Kosorok (University of North Carolina at Chapel Hill)

In this presentation, we will highlight Bayesian machine learning for individualized decision-making and its use for principled uncertainty propagation and quantification. After this general overview, we will present three related works: Bayesian outcome weighted learning, Bayesian backward outcome weighted learning, and a BART-based approach to treatment policy mediation analysis. Emphasis will be on the connective tissue among the papers, particularly how they relate to optimal treatment rule estimation and uncertainty quantification.

Covariate adjustment using targeted learning: methods and software tools

Foroogh Shamsi, Novo Nordisk

Foroogh Shamsi (Novo Nordisk); Benedikt Sommer (Novo Nordisk); Klaus Kähler Holst (Novo Nordisk)

Covariate adjustment can improve the precision of treatment effect estimates in randomized clinical trials, increasing power and potentially reducing sample size requirements. Targeted learning provides a framework for incorporating flexible outcome models, including machine learning, into population-level treatment effect estimation and inference. Simulation allows their performance to be examined across diverse trial scenarios, including variations in design, population characteristics, and modeling assumptions. We introduce carts, an R package we developed to streamline simulation-based assessment of covariate adjustment. The framework is organized around three components: the data-generating process, estimation, and inference, enabling investigators to explore different trial desing scenarios and analysis strategies. It supports power and sample size calculations, assessment of estimator performance, and sensitivity analyses across a range of trial designs and outcome types. Through integration with the targeted R package, carts supports targeted learning estimators alongside conventional analysis methods. Its modular structure also accommodates user-defined models and estimators, allowing investigators to tailor simulations to their research questions.

Causal AI/ML for real-world evidence: bridging the rigor gap between RWE and RCTs in pharma

Yunxun Wang, Bayer U.S. LLC

Yunxun Wang (Bayer)

Regulatory agencies (FDA, EMA, and PMDA) are increasingly recognizing that real-world evidence (RWE) serves not only as a supplement to randomized controlled trials (RCTs) but also as supportive evidence for new drug application (NDA) or supplemental NDA (sNDA) submissions. The pharmaceutical industry faces a critical challenge: Generating regulator-grade RWE requires establishing robust fit-for-purpose methodologies and mitigating inherent biases within real-world data (RWD) to achieve scientific rigor comparable to RCTs. Furthermore, in the LLM era, innovative causal artificial intelligence (AI) and machine learning (ML) methodologies apply to RWE studies and show promise to control bias and confounding by fully utilizing rich information in RWD compared with those traditional methodologies. This discussion, anchored in the FDA and EMA jointly released Guiding Principles for Good AI and Machine Learning Practice in Drug Development, investigates how causal AI/ML methodologies are emerging as critical tools for closing this gap, transforming RWE from supportive evidence to regulatory-grade decision-making. A roadmap for adopting causal ML in RWE operations that fulfills both scientific and regulatory rigor, enabling cross-functional collaboration between AI/ML developers and decision makers, will be given.

Session 43: Large Language Models and Statistical Foundations

Location: Fulton Ferry

Organizer: Yan Sun (New Jersey Institute of Technology)

Chair: Yan Sun (New Jersey Institute of Technology)

Presenters: Jiawei (Joe) Zhou; Enric Boix-Adsera; Shuai Zhang; Yue Xing

Rethinking tokens in language models for efficient and adaptable inference

Jiawei (Joe) Zhou, Stony Brook University

Jiawei Zhou (Stony Brook University)

Language models typically process and generate text as sequences of discrete tokens. While this interface has enabled remarkable progress, it also shapes the computational cost of inference and how models access and use information. This talk explores how reconsidering the representation and use of tokens can enable more efficient and adaptable language model inference while preserving task performance. Drawing on recent work in decoding—including Chunk-Distilled Language Modeling (CD-LM)—input compression, and visual representations of text, I will discuss opportunities to reduce computation by changing how models consume and produce language, as well as to incorporate external knowledge during generation. These approaches revisit the foundational mechanisms through which language models represent, process, and generate text, opening new possibilities for balancing efficiency, information preservation, and adaptability. The talk will highlight technical insights and empirical tradeoffs, alongside broader questions about how language should be represented and processed in large-scale models.

Specialization in neural networks: from theory to architecture design

Enric Boix-Adsera, University of Pennsylvania

Enric Boix-Adsera

We begin with a theoretical question: what are feedforward layers actually doing in an LLM? Our investigations reveal a robust, general principle: feedforward layers operation on activations that are secretly specialized; they compute interactions between sparse features in activation space. This theoretical insight allows us to design a new LLM architecture that empirically shrinks the RAM footprint of a 3B-parameter model by a factor of 6x, while preserving performance and inference speed.

Towards theoretical foundations of foundation models: feature learning, optimization dynamics, and generalization

Shuai Zhang, New Jersey Institute of Technology

Shuai Zhang

Foundation models have rapidly become a central paradigm in modern AI, driving breakthroughs across language, vision, and multimodal learning. Despite their empirical success, many fundamental questions remain, including why these large-scale neural networks learn effective representations and why they generalize well despite extreme overparameterization. This talk discusses recent theoretical progress toward understanding foundation models within the feature learning framework. In this framework, data are modeled as a mixture of latent features, and learning dynamics guide models toward particular solutions through implicit preferences that lead to strong generalization.

Impact of positional encoding: clean and adversarial Rademacher complexity for transformers under in-context regression

Yue Xing, Michigan State University

Weiyi He (Michigan State University); Yue Xing (Michigan State University)

Positional encoding (PE) is a core architectural component of Transformers, yet its impact on the Transformer's generalization and robustness remains unclear. In this work, we provide the first generalization analysis for a single-layer Transformer under in-context regression that explicitly accounts for a completely trainable PE module. Our result shows that PE systematically enlarges the generalization gap. Extending to the adversarial setting, we derive the adversarial Rademacher generalization bound. We find that the gap between models with and without PE is magnified under attack, demonstrating that PE amplifies the vulnerability of models. Our bounds are empirically validated by a simulation study. Together, this work

establishes a new framework for understanding the clean and adversarial generalization in ICL with PE.

Session 67: New Frontiers in High-Dimensional Graphical Modeling

Location: Sunset Park

Organizer: Jing Ma (Fred Hutchinson Cancer Center)

Chair: Jing Ma (Fred Hutchinson Cancer Center)

Presenters: Armeen Taeb; Y. Samuel Wang; Yang Ni; George Michailidis

Integrating multi-source local information for global undirected graph reconstruction via integer programming

Armeen Taeb, University of Washington

Armeen Taeb (University of Washington), Nir Keret (University of Washington), Ali Shojaie (University of Washington)

Graphical models are widely used to represent conditional independence structures, but most methods assume access to a single dataset in which all variables are jointly observed. In many modern applications, data are instead fragmented across sources, involve different subsets of variables, and may not be directly shareable. We develop a general framework for reconstructing a shared global graph from multiple heterogeneous and partially observed sources. Our approach allows both the observed variable sets and source-specific distributions to differ, requiring only that the underlying graph is shared. It can integrate either raw data or estimated source graphs, making it applicable under privacy and data-access constraints. We formulate reconstruction through graphical projections and develop mixed-integer optimization methods and scalable approximations for several classes of graphical models. Simulations and applications show that our integrative approach accurately reconstructs the global graph and can substantially improve estimation of the source-specific projected graphs relative to methods that analyze each source separately.

Estimation of functional graphical models with unobserved confounding

Y. Samuel Wang, Cornell University

Filippo Fiocchi (Cornell University); Y. Samuel Wang (Cornell University)

Multivariate functional data is gathered in many settings including neuroscience, finance, and atmospheric science. Oftentimes, the scientist is interested in recovering a graphical model for the underlying stochastic process, and various procedures have been proposed for estimating both directed and undirected graphs for functional data. However, the existing work assumes that the process of interest is observed without latent confounding. This assumption is violated in many interesting settings, and the resulting misspecification may yield an unreliable and inconsistent estimate for the graph of the unconfounded process of interest. In this work, we propose a method which, under certain conditions, consistently estimates a graphical model for the process of interest (both directed and undirected) despite unobserved confounding.

Multivariable bidirectional Mendelian randomization via Bayesian directed cyclic graphical models with correlated errors

Yang Ni, The University of Texas at Austin

Bitan Sarkar (Texas A&M University); Yuchao Jiang (Texas A&M University); Tian Ge (Massachusetts General Hospital); Yang Ni (The University of Texas at Austin)

Mendelian randomization (MR) is a pivotal tool in genetics, genomics, and epidemiology, leveraging genetic variants as instrumental variables to infer causal relationships between exposures and outcomes. Traditional MR methods, while powerful, often rely on stringent assumptions such as the absence of feedback loops, which are frequently violated in complex biological networks. In addition, many popular MR approaches focus on only two variables (i.e., one exposure and one outcome), whereas our motivating applications of gene regulatory networks have many variables. In this article, we introduce a novel Bayesian framework for multivariable MR that concurrently addresses unmeasured confounding and feedback loops. Central to our approach is a sparse conditional cyclic graphical model with a sparse error variance-covariance matrix. Two structural priors are employed to enable the modeling and inference of causal relationships as well as latent confounding structures. Our method is designed to operate effectively with summary-level data, facilitating its application in contexts where individual-level data are inaccessible, e.g., due to privacy concerns. It can also account for horizontal pleiotropy, under which we establish the sufficient identifiability conditions. Through extensive simulations and applications to the GTEx and OneK1K data, we demonstrate the superior performance of our approach in recovering biologically plausible causal relationships in the presence of possible feedback loops and unmeasured confounding. Using posterior samples, we further quantify uncertainty in inferred network motifs by computing their posterior probabilities.

Joint learning of panel VAR models with low rank and sparse structure

George Michailidis, University of California, Los Angeles

George Michailidis (University of California, Los Angeles)

Panel vector auto-regressive (VAR) models are widely used to capture the dynamics of multivariate time series across different subpopulations, where each subpopulation shares a common set of variables. In this work, we propose a panel VAR model with a shared low-rank structure, modulated by subpopulation-specific weights, and complemented by idiosyncratic sparse components. To ensure parameter identifiability, we impose structural constraints that lead to a nonsmooth, nonconvex optimization problem. We develop a multi-block Alternating Direction Method of

Multipliers (ADMM) algorithm for parameter estimation and establish its convergence under mild regularity conditions. Furthermore, we derive consistency guarantees for the proposed estimators under high-dimensional scaling. The effectiveness of the proposed modeling framework and estimators is demonstrated through experiments on both synthetic data and a real-world neuroscience data set.

Session 39: Statistical Learning and Inference for Metric-Space-Valued Data

Location: Golden Boardroom

Organizer: Mladen Kolar (USC & MBZUAI)

Chair: Mladen Kolar (USC & MBZUAI)

Presenters: Yidong Zhou; Zhenhua Lin; Paromita Dubey; Adam Quinn Jaffe

Fréchet geodesic boosting

Yidong Zhou, University of Minnesota

Yidong Zhou (University of Minnesota); Su I Iao (Waymo); Hans-Georg Müller (University of California, Davis)

Gradient boosting has become a cornerstone of machine learning, enabling base learners such as decision trees to achieve exceptional predictive performance. While existing algorithms primarily handle scalar or Euclidean outputs, increasingly prevalent complex-structured data, such as distributions, networks, and manifold-valued outputs, present challenges for traditional methods. Such non-Euclidean data lack algebraic structures such as addition, subtraction, or scalar multiplication required by standard gradient boosting frameworks. To address these challenges, we introduce Fréchet geodesic boosting (FGBoost), a novel approach tailored for outputs residing in geodesic metric spaces. FGBoost leverages geodesics as proxies for residuals and constructs ensembles in a way that respects the intrinsic geometry of the output space. Through theoretical analysis, extensive simulations, and real-world applications, we demonstrate the strong performance and adaptability of FGBoost, showcasing its potential for modeling complex data.

Neural Wasserstein two-sample tests

Zhenhua Lin, National University of Singapore

Xiaoyu Hu (Xi'an Jiaotong University); Zhenhua Lin (National University of Singapore)

The two-sample homogeneity testing problem is fundamental in statistics and becomes particularly challenging in high dimensions, where classical tests can suffer substantial power loss. We develop a learning-assisted procedure based on the projection Wasserstein distance. The method is motivated by the observation that there often exists a low-dimensional projection under which the two high-dimensional distributions differ. In practice, we learn the projection directions via manifold optimization and a witness function using deep neural networks. To adapt to unknown projection dimensions and sparsity levels, we aggregate a collection of candidate statistics through a max-type construction, avoiding explicit tuning while potentially improving power. We establish the validity and consistency of the proposed test and prove a Berry-Esseen type bound for the Gaussian approximation. In particular, under the null hypothesis, the aggregated statistic converges to the absolute maximum of a standard Gaussian vector, yielding an asymptotically pivotal (distribution-free) calibration that bypasses resampling. Simulation studies and a real-data example demonstrate the strong finite-sample performance of the proposed method.

Minimax optimal conditional paired sample tests via distance profiles

Paromita Dubey, University of Southern California

Bilol Banerjee (National University of Singapore); Paromita Dubey (University of Southern California)

We study the problem of comparing the conditional distributions of two random variables X and Y given a confounding variable Z, where all variables take values in general metric spaces. We introduce a new discrepancy measure that quantifies conditional distributional differences through joint distance profiles and develop a consistent estimator of this quantity. Building on this estimator, we construct a conditional paired sample test that avoids local smoothing and, therefore, remains simple, broadly applicable, and powerful against general alternatives. We further propose an algorithm to calibrate our tests, which is both Pitman efficient and minimax-rate optimal against a large class of alternatives. Through simulations, we demonstrate that the proposed method outperforms existing approaches, and we illustrate its practical utility by analyzing real datasets.

Phylogenetic inference and the stickiness of Fréchet means, via precise asymptotics of an embedded random walk

Adam Quinn Jaffe, Columbia University

Adam Quinn Jaffe (Columbia University)

A well-known phenomenon in statistical analyses of populations of phylogenetic trees in the Billera-Holmes-Vogtmann space is that the topology of the Fréchet mean tree can contain multifurcations (i.e., internal nodes with more than two children), which raises the practical question of whether this reflects a population-level branching structure (hard polytomy) or merely sampling variability in the data (soft polytomy). This is an instance of the more general phenomenon of "stickiness" in non-Euclidean statistics, whereby the sample Fréchet mean in certain non-positively curved stratified spaces becomes permanently trapped in a lower-dimensional stratum. In this work, we identify a particular multidimensional random walk embedded within the Fréchet mean process, and we show that the time at which stickiness occurs is determined by the largest

last-passage time above zero of the coordinates of this random walk. Using this representation, we develop a fully nonparametric procedure for estimating the probability that trifurcations in a sample Fréchet mean tree will bifurcate at some future time if more observations are collected. Lastly, we apply our methodology to a problem in phylogenetics where we consider whether an observed trifurcation in the species tree of primates, glires, and tree shrews is genuinely trifurcated at the population level.

Session 51: Advances in Random Matrix Theory and Its Statistical Applications

Location: Carroll Gardens

Organizer: Rong Ma (Harvard University)

Chair: Rong Ma (Harvard University)

Presenters: Alden Green; William Leeb; Yichen Hu; Tavor Z. Baharav

The high-dimensional asymptotics of principal components regression

Alden Green, Johns Hopkins University

Alden Green (Johns Hopkins University); Elad Romanov (Texas A&M University)

We study principal components regression (PCR) in an asymptotic high-dimensional setting, where the number of data points is proportional to the dimension. We derive exact limiting formulas for estimation and prediction risk, which depend in a complicated manner on the eigenvalues of the population covariance, the alignment between the population PCs and the true signal, and the number of selected PCs. A key challenge in the high-dimensional setting stems from the fact that the sample covariance is an inconsistent estimate of its population counterpart, so that sample PCs may fail to fully capture potential latent low-dimensional structure in the data. We demonstrate this point through several case studies, including that of a spiked covariance model.

Matrix estimation in the high-noise, high-dimensional regime

William Leeb, University of Minnesota, Twin Cities

William Leeb

This talk will describe selected methods based on spectral shrinkage, and the accompanying theory, for estimating low-rank matrices in the high-noise, high-dimensional regime.

Bias-corrected multiplier bootstrap inference for edge eigenvalues of large covariance matrices

Yichen Hu, University of California, Davis

Xiucai Ding (University of California, Davis); Yichen Hu (University of California, Davis); Jiahui Xie (University of California, Davis)

Inference for spectral edges of large covariance matrices is a fundamental problem in high-dimensional statistics, but it is challenging because the largest non-spiked sample eigenvalues fluctuate on the Tracy–Widom scale, requiring accurate estimation of both the deterministic spectral edge and its scaling constant under general unknown population covariance structures. We propose a bias-corrected multiplier bootstrap procedure for inference on the bulk spectral edge. The method introduces a calibrated multiplier perturbation that regularizes the edge fluctuation to a slightly larger scale where Gaussian approximation becomes tractable, together with a data-driven recentering step that corrects the bootstrap-induced shift of the deterministic edge. We show that the largest few non-spiked bootstrap eigenvalues are conditionally asymptotically Gaussian after bias correction and rescaling, leading to asymptotically valid confidence intervals whose length is only slightly larger than the Tracy–Widom scale $n^{-2/3}$, up to a polynomial factor. The same framework also provides a threshold-free estimator of the number of spikes, or equivalently a data-driven cutoff for the scree plot. Numerical studies and real-data applications demonstrate strong finite-sample performance, particularly in settings with weak spikes and heterogeneous bulk spectra.

Stacked SVD or SVD stacked? A random matrix theory perspective on data integration

Tavor Z. Baharav, Broad Institute of MIT and Harvard

Tavor Z. Baharav (Eric and Wendy Schmidt Center, Broad Institute); Phillip B. Nicol (Department of Data Science, Dana-Farber Cancer Institute); Rafael A. Irizarry (Department of Data Science, Dana-Farber Cancer Institute); Rong Ma (Department of Biostatistics, Harvard T.H. Chan School of Public Health)

Modern data analysis increasingly requires identifying shared latent structure across multiple high-dimensional datasets. A commonly used model assumes that the data matrices are noisy observations of low-rank matrices with a shared singular subspace. In this case, two primary methods have emerged for estimating this shared structure, which vary in how they integrate information across datasets. The first approach, termed Stack-SVD, concatenates all the datasets, and then performs a singular value decomposition (SVD). The second approach, termed SVD-Stack, first performs an SVD separately for each dataset, then aggregates the top singular vectors across these datasets, and finally computes a consensus amongst them. While these methods are widely used, they have not been rigorously studied in the proportional asymptotic regime, which is of great practical relevance in today's world of increasing data size and dimensionality. We derive exact expressions for the asymptotic performance and phase transitions of these two methods and develop optimal weighting schemes to further improve both methods. Our analysis reveals that while neither method uniformly dominates the other in the unweighted case, optimally weighted Stack-SVD dominates optimally weighted SVD-Stack when the low rank signal is fully shared across the datasets. We then extend our analysis to handle multiple, partially shared components per dataset and

demonstrate that SVD-Stack can yield improved performance without requiring estimation of subspace alignment. Finally, provide practical algorithms for estimating optimal weights from data, offering theoretical guidance for method selection in practical data integration problems. Extensive numerical simulations and semi-synthetic experiments on genomic data corroborate our theoretical findings.

Session 34: Statistical Inference under Data Perturbation, Structure, and Heterogeneity

Location: Coney Island

Organizer: Elynn Chen (New York University)

Chair: Elynn Chen (New York University)

Presenters: Xiaodong Li; Xiaoqian Liu; Kai Tan; Shurong Lin

Inference on canonical directions with partially paired data

Xiaodong Li, University of California, Davis

Chau Tran, Xiaodong Li

Canonical correlation analysis (CCA) is a classical method for studying linear associations between two data views. Inference on canonical directions helps interpret these associations and identify the variables that contribute to them. In this talk, we consider hypothesis testing for canonical directions when paired observations are supplemented by unpaired observations from either view. The central challenge is to use the additional marginal information while accounting for its dependence on the paired data used to estimate the canonical directions. We introduce coupled augmentation functions that combine marginal information from both views and use the resulting corrections to construct a one-step estimator. Under the null hypothesis, we establish consistency and asymptotic normality when the dimension satisfies (d=o(n)), where (n) is the number of paired observations. These results yield a sign-invariant Wald test for variable contributions to canonical directions. Numerical studies demonstrate improved inferential performance over methods using only paired observations.

Transfer learning for robust structured regression with bi-level source detection

Xiaoqian Liu, University of California, Riverside

Haoming Shi (Rice University); Hengrui Luo (Rice University); Yang Feng (New York University); Xiaoqian Liu (University of California, Riverside)

High-dimensional data in modern applications, such as COVID-19 mortality, often span multiple domains. Leveraging auxiliary information from source domains to improve performance in a target domain motivates the use of transfer learning. However, a practical issue that has been overlooked is data contamination, which induces heterogeneity and can significantly degrade transfer learning performance. To address this challenge, we propose a novel approach that tackles transfer learning under data contamination within a structured regression setting. By employing the robust L2E criterion, we develop the TransL2E method that accounts for contamination in both target and source data while effectively transferring relevant information. Beyond robust estimation, TransL2E introduces a data-driven bi-level source detection mechanism, operating at both individual and cohort levels, which possesses multiple advantages over existing source detection approaches. Comprehensive simulation studies and a real data application demonstrate the superior performance of TransL2E in both robust estimation and structure recovery in the presence of data limitation and contamination.

When does dropout help? Consistent risk estimation and phase transitions in linear regression

Kai Tan, The Ohio State University

Kai Tan (The Ohio State University)

We study generalization-error estimation for dropout-regularized linear regression when the feature dimension p is proportional to the sample size n. Classical generalized cross-validation need not be consistent for dropout, even under isotropic designs: for every nontrivial retention rate, we identify a proportional regime with strictly positive asymptotic bias. We propose two corrections: an *initial* estimator using the population covariance and a fully data-driven *final* estimator. We prove non-asymptotic $\sqrt{n}$-consistency under Gaussian linear models and asymptotic consistency for non-Gaussian designs and nonlinear responses through a best linear predictor decomposition. Under isotropic designs, we derive a closed-form deterministic equivalent for risk and a sharp phase transition at $\gamma_c = (3+\sqrt{5})/2$. Below this threshold, some nontrivial dropout rate improves upon the no-dropout baseline; at or above it, the no-dropout endpoint minimizes risk along the dropout path. For anisotropic designs, we develop a data-driven boundary-sign diagnostic for the local effect near this endpoint, while global tuning minimizes the proposed risk estimator. Our proofs combine Stein's formula with random-matrix resolvent analysis. Simulations under linear and nonlinear models corroborate the theory and show strong finite-sample accuracy.

Hypothesis testing under differential privacy via binning-aggregation

Shurong Lin, Pennsylvania State University

Shurong Lin (Pennsylvania State University); Yuxuan Wang (Pennsylvania State University); Aleksandra Slavković (Pennsylvania State University)

Differential privacy limits how much a statistical analysis can reveal about any individual, providing rigorous privacy protection by introducing additional randomness that also creates challenges for statistical inference. We develop an inference workflow for differentially private linear regression, including hypothesis testing and

aggregate-level diagnostics, based on a recent binning-aggregation approach. Using the asymptotic normality of the private estimator, we construct Wald-type tests for individual and joint regression hypotheses. Because these tests use the same privacy-protected summaries used for estimation, they require no additional privacy loss. Through simulations, we evaluate Type I error and power across privacy levels, sample sizes, error variances, and covariate distributions. The proposed tests provide reliable error control and improved power relative to an existing private testing benchmark. We also illustrate multivariable testing and private diagnostics on real data, providing a unified workflow for estimation, inference, and model assessment under differential privacy.

Session 04: Analysis of Omic Data, and AI and Data-Driven Science

Location: Park Slope

Organizer: Heping Zhang (Yale University School of Public Health)

Chair: Qiao Liu (Yale University School of Public Health)

Presenters: Pei Wang; Minge Xie; Ying Wei; Shikun Wang

Network inference and integrative omics study

Pei Wang, Icahn School of Medicine at Mount Sinai

Pei Wang (Icahn School of Medicine at Mount Sinai)

In contemporary proteogenomic cancer studies, a single tumor is often analyzed with more than ten different proteogenomic profiling strategies, ranging from genetics, transcriptomics, proteomics, and post-translational modifications. In this talk, I will discuss the principal needs and challenges of multi-omics integration, with a focus on application of network models in proteogenomic studies. I will illustrate how these approaches illuminate regulatory effects among genes and proteins. Analyses of recent CPTAC proteogenomic datasets illustrate practical workflows, tool usage, and performance considerations in real-world data contexts.

Model-free inference for high-dimensional binary classification using Repro samples

Minge Xie, Rutgers University

Minge Xie

In high-dimensional settings, when the underlying data-generating regression structure is completely unknown and unspecified, conducting statistical inference is a particularly challenging task. In this talk, we develop a model-free inference approach for high-dimensional binary classification, leveraging a class of (likely misspecified) sparsity generalized models. Our method builds on the so-called repro samples framework, which utilizes artificial samples that are generated by mimicking the true data or noises. The proposed approach facilitates inference by targeting both the model support (discrete model parameter) and the regression coefficients (continuous model parameter) of the oracle model, defined as the working model closest to the unknown true model. The method offers three key advantages: (1) it is fully model-free, requiring neither correct model specification nor sparsity of the true model; (2) it constructs a candidate set of influential covariates with guaranteed coverage under weak signal conditions; and (3) it provides confidence sets for any linear transformation of the oracle coefficients. When the oracle model coincides with the true underlying model, the inference results directly apply to the true model, allowing simultaneous quantification of uncertainty in (discrete) model selection and (continuous) parameter estimation. Simulation studies illustrate the effective performance of the proposed method. An application to single-cell RNA-seq immune response data demonstrates that it identifies key genes, offering new insights into cellular immune response mechanisms.

Joint estimation and grouping of quantile association curves across multiple responses

Ying Wei, Columbia University

Tianying Wang; Yanyuan Ma; Ying Wei

Large biobanks provide an opportunity to study how associations with related traits vary across age and across their distributions. We propose a quantile regression procedure that jointly groups responses and estimates their shared varying-coefficient curves on conditional percentile scales. An information criterion selects both the number of groups and their membership from grouped B-spline fits. We establish consistent group recovery, curve convergence rates, and feasible pointwise inference after group selection as the spline dimension grows. The analysis distinguishes true curve differences from improvements due to separate spline approximations of a shared curve. Errors in the estimated percentile scales partly cancel in comparisons of pooled and split fits of a shared curve, but can still contribute to the asymptotic variance of the curve estimators. In 127,040 UK Biobank women aged 40 to 70, we grouped nine metabolic traits by their associations with a type 2 diabetes polygenic risk score and found seven groups at the median, compared with three and four at the lower and upper quantiles. The method identifies shared tail associations among traits with distinct median associations, allowing their information to be pooled where the curves agree.

New analytical methods for spatial transcriptomics data

Shikun Wang, Columbia University

Shikun Wang (Columbia); Ming Yuan (Columbia); Yiping Wang (Columbia); Benjamin Izar (Columbia); Anil Rustgi (Columbia); Jianhua Hu (Columbia)

Identifying cell subgroups and detecting gene differential expressions from recently blooming single-cell RNA-sequencing (scRNA-seq) and spatial transcriptomics (ST) is a topic of increasing interest to biomedical researchers because spatial landscapes reveal how gene expression varies across tissue samples. The statistical

challenges to identifying cell subgroups and detecting gene differential expressions are multifold; the gene expression in individual cells is high dimensional and zero-inflated, associated with the spatial information, with technical noise and batch effects due to variations in experimental conditions. We will present our recent method developments to address these challenges and, therefore, improve detection accuracy over existing popular pipelines.

Session 89: Equilibrium-Aware A/B Testing

Location: Windsor Terrace

Organizer: Lihua Lei (Stanford University)

Chair: Lihua Lei (Stanford University)

Presenters: David Holtz; Timothy Sudijono; Jizhou Liu; Zahra Khanalizadeh

Reducing symbiosis bias through better A/B tests of recommendation algorithms

David Holtz, Columbia Business School

Jennifer Brennan (Google Research), Yahu Cong (University of California, Berkeley), Yiwei Yu (Columbia Business School), Lina Lin (Google), Yajun Peng (Google), Changping Meng (Google), Ningren Han (Google), Jean Pouget-Abadie (Google Research), David M. Holtz (Columbia Business School)

It is increasingly common in digital environments to use A/B tests to compare the performance of recommendation algorithms. However, such experiments often violate the stable unit treatment value assumption (SUTVA), particularly SUTVA's "no hidden treatments" assumption, due to the shared data between algorithms being compared. This results in a novel form of bias, which we term "symbiosis bias," where the performance of each algorithm is influenced by the training data generated by its competitor. In this paper, we investigate three experimental designs–cluster-randomized, data-diverted, and user-corpus co-diverted experiments–aimed at mitigating symbiosis bias. We present a theoretical model of symbiosis bias and simulate the impact of each design in dynamic recommendation environments. Our results show that while each design reduces symbiosis bias to some extent, they also introduce new challenges, such as reduced training data in data-diverted experiments. We further validate the existence of symbiosis bias using data from a large-scale A/B test conducted on a global recommender system, demonstrating that symbiosis bias affects treatment effect estimates in the field. Our findings provide actionable insights for researchers and practitioners seeking to design experiments that accurately capture algorithmic performance without bias in treatment effect estimates introduced by shared data.

Compound selection decisions: an almost SURE approach

Timothy Sudijono, Department of Statistics and Data Science, Wharton School, University of Pennsylvania

Jiafeng Chen (Stanford); Lihua Lei (Stanford); Timothy Sudijono (UPenn); Liyang Sun (UCL); Tian Xie (UCL)

This paper proposes methods for compound selection decision problems in a Gaussian sequence model. Inspired by Stein's unbiased risk estimate (SURE), we introduce ASSURE, a family of estimators for welfare, defined as the expected utility of a selection decision rule. ASSURE enables robust evaluation of selection decisions. For empirical Bayes decisions derived from random-effects models of unknown parameters, ASSURE estimates their welfare even when the models are misspecified. Optimizing ASSURE-estimated welfare over a pre-specified class of decision rules further borrows strength across noisy estimates and yields decision rules with favorable regret properties. These regret properties are again robust to potential misspecification of random effects models, thereby robustifying empirical Bayes methods. We apply ASSURE to selecting Census tracts for economic mobility, identifying discriminating firms, and evaluating p-value decision rules in A/B testing.

Randomization inference for two-sided market experiments

Jizhou Liu, Peking University

Azeem Shaikh; Panos Toulis

Randomized experiments are increasingly employed in two-sided markets, such as buyer–seller platforms, to evaluate the effects of marketplace interventions. These experiments must reflect the underlying two-sided market structure in their design and can therefore be challenging to analyze. In this paper, we develop a randomization inference framework for outcomes from two-sided experiments, with a focus on testing and inference for two-sided spillover effects. Our approach is finite-sample valid under sharp null hypotheses. Regarding weak null hypotheses, we find that the commonly used Neyman-style studentization does not universally ensure asymptotic validity, and we document how it depends on the specific formulation of the null. We then propose a two-way variance estimator for studentization that restores asymptotic validity. We further propose methods to improve testing power by exploiting the two-sided structure of the problem, which we validate empirically. We demonstrate our methods through a series of simulation studies and an applied example from a network experiment in micro-lending.

Ranking as experimental design under marketplace interference

Zahra Khanalizadeh, University of Washington

Mahyar Habibi (Lyft); Zahra Khanalizadeh (University of Washington); Negar Ziaeian (University of Warwick)

Online marketplaces frequently run experiments in environments where users choose from a ranked list of items. In these settings, items compete for users' limited attention and demand, creating interference among items within a list: Treating any item can affect the demand for

others, biasing estimates from item-level A/B tests. In addition, a key consideration in some marketplace experiments is preserving platform coherency across treatment status and item availability. This requirement can rule out experimental designs such as user-level A/B tests when they expose different users to different treatment states. We propose Two-Sided Prioritized Ranking (TSPR) to estimate the total average treatment effect in such settings. TSPR exploits position bias in ranked search results to create variation in treatment exposure without compromising coherency. TSPR randomizes both users and items and reorders ranked lists, prioritizing treated items for one group of users and untreated items for the other. All users see the same items with consistent treatment status, but differ in exposure to treatment as they pay disproportionate attention across ranks. In semi-synthetic simulations based on Expedia hotel search data, TSPR outperforms baseline coherency-preserving designs by reducing estimation bias and providing sufficient statistical power. More broadly, the approach illustrates how ranking itself can serve as an experimental-design lever in online marketplaces.

Session 48: Integrative Statistical Learning for Complex Biomedical Data

Location: DUMBO

Organizer: Guanqun Cao (Michigan State University)

Chair: Guanqun Cao (Michigan State University)

Presenters: Grace Yi; Yuan Huang; Peter X.K. Song; Jiawei Shan

The hidden distortion of measurement error in ultrahigh-dimensional data

Grace Yi, University of Western Ontario

Yifan Sun (University of Western Ontario); Grace Yi (University of Western Ontario)

While standard statistical methodologies often rely on the tacit assumption that observed data are error-free, real-world data frequently suffer from various forms of measurement error arising from data collection procedures, reporting inaccuracies, privacy-preserving mechanisms, and intrinsic variability. This talk examine how measurement error distorts statistical inference, particularly in ultrahigh-dimensional settings, where it can induce complex and nonstandard effects. The discussion focuses on a function-on-scalar regression framework and highlights distinct impacts of different measurement error processes in ultrahigh- versus finite-dimensional settings. Both theoretical and numerical results will be presented.

Mediation analysis for sparse and irregularly spaced longitudinal outcomes

Yuan Huang, Yale University

Rui Ren (Yale University); Haoyi Yang (Pennsylvania State University), Qian Xiao (University of Texas Health Sciences Center); Lingzhou Xue (Pennsylvania State University); Yuan Huang (Yale University)

Mediation analysis has become a widely used method for identifying the pathways through which an independent variable influences a dependent variable via intermediate mediators. However, limited research addresses the case where mediators are high- dimensional and the outcome is represented by sparse, irregularly spaced longitudinal data. To address these challenges, we propose a mediation analysis approach for scalar exposures, high-dimensional mediators, and sparse longitudinal outcomes. This approach effectively identifies significant mediators by addressing two key issues: (i) the underlying correlation structure within the sparse and irregular cognitive measurements, and (ii) adjusting mediation effects to handle the high-dimensional set of candidate mediators. In the MrOS Sleep study, our primary objective is to explore lipid pathways that may mediate the relationship between rest-activity rhythms and longitudinal cognitive decline in older men. Our findings suggest a potential mechanism involving rest-activity rhythms, lipid metabolites, and cognitive decline, and highlight significant mediators identified through multiple testing procedures.

Integrating epigenetic clocks using neural networks to estimate biological age

Peter X.K. Song, University of Michigan

Yahui Zhang (University of Michigan); Peter X.K. Song (University of Michigan)

The advent of DNA methylation (DNAm) technology has provided a rich source of omics data to help evaluate epigenetic age and age acceleration. Such aging markers be estimated through predictive models known as epigenetic clocks that are derived by high-dimensional data analytics. So far, there are hundreds of epigenetic clocks available in the literature but there lacks consensus which ones should be used in practice. We proposed a new model averaging approach to integrating multiple epigenetic clocks into a meta-clock via neural networks. Our approach is built upon a nonlinear model aggregation, which is different from the classical linear combination of models. This talk will highlight theoretical justification, algorithmic implementation, illustrated through real-world data applications.

SADA: safe and adaptive aggregation of multiple black-box predictions in semi-supervised learning

Jiawei Shan, University of Wisconsin–Madison

Jiawei Shan (UW-Madison); Zhifeng Chen (UW-Madison); Yiming Dong (UW-Madison); Yazhen Wang (UW-Madison); Jiwei Zhao (UW-Madison)

Semi-supervised learning (SSL) arises in practice when labeled data are scarce or expensive to obtain, while large quantities of unlabeled data are readily available. With the growing adoption of machine learning techniques, it has become increasingly feasible to generate multiple

predicted labels using a variety of models and algorithms, including deep learning, large language models, and generative AI. In this paper, we propose a novel approach that safely and adaptively aggregates multiple black-box predictions of uncertain quality for both inference and prediction tasks. Our method provides two key guarantees: (i) it never performs worse than using the labeled data alone, regardless of the quality of the predictions; and (ii) if any one of the predictions (without knowing which one) perfectly fits the ground truth, the algorithm adaptively exploits this to achieve either a faster convergence rate or the semiparametric efficiency bound. We demonstrate the effectiveness of the proposed algorithm through small-scale simulations and two real-data analyses with distinct scientific goals. A user-friendly R package, `sada`, is provided to facilitate practical implementation.

Session 42: Impactful Applications in Statistics and Data Science in Aging

Location: Greenpoint

Organizer: Jaime Lynn Speiser (Wake Forest University School of Medicine)

Chair: Panpan Zhang (Vanderbilt University Medical Center)

Presenters: Sarah B Peskoe; Marissa C. Ashner; Wenhui (Vivian) Zhang; Jaime Lynn Speiser

Recovering smarter: data driven definitions of physical resilience in aging

Sarah B Peskoe, Duke University School of Medicine

Sarah B Peskoe (Duke University School of Medicine); Marissa Ashner (Duke University School of Medicine); Virginia B Kraus (Duke University School of Medicine); Janet L Huebner (Duke University School of Medicine); Corey B Simon (Duke University School of Medicine); Katherine Hall (Duke University School of Medicine, Durham Veterans Administration Medical Center); Cathleen Colon-Emeric (Duke University School of Medicine); Heather Whitson (Duke University School of Medicine)

Quantifying physical resilience, or the capacity to recover after an acute health stressor, is essential for advancing aging research and identifying individuals whose recovery exceeds or falls short of expectations. This talk introduces two complementary, data driven approaches for defining resilience: (1) recovery trajectories derived from longitudinal performance, and (2) the Expected Recovery Differential (ERD), which measures how much better or worse an individual recovers relative to a personalized predicted outcome.

The trajectory based approach characterizes the shape, speed, and magnitude of recovery using latent class trajectory analysis. In contrast, the ERD focuses on individualized recovery by integrating demographic, clinical, functional, psychosocial, and stressor specific predictors into penalized regression models to estimate expected recovery, then comparing expected with observed outcomes.

Using data from older adults undergoing total knee arthroplasty in the PRIME KNEE study, we applied both approaches to repeated measures of pain, function, and physical activity over six months. The methods aligned most strongly at the extremes: participants with clearly high or low resilience trajectories typically also had notably better or worse ERD scores. However, among individuals with moderate trajectories, ERD values varied widely, revealing distinctions not captured by trajectory classification alone. Trajectories illuminate overall recovery patterns, whereas ERD contextualizes how impressive or concerning those patterns are relative to personalized expectations.

Together, these approaches capture complementary dimensions of resilience and improve our ability to characterize heterogeneity in recovery. Their combined use strengthens the foundation for biomarker discovery, mechanistic research, and the development of targeted interventions to support resilience in aging populations.

An exploratory stability selection (ESS) framework for robust predictor discovery in aging research

Marissa C. Ashner, Duke University School of Medicine

Marissa C. Ashner (Duke University School of Medicine); Virginia B. Kraus (Duke University School of Medicine); Heather E. Whitson (Duke University School of Medicine); Corey B. Simon (Duke University School of Medicine); Janet L. Huebner (Duke University School of Medicine); Akshay Bareja (Duke University School of Medicine); Chelsea R. Perfect (Duke University School of Medicine); Laura Pietrosimone (Duke University School of Medicine); Katherine S. Hall (Duke University School of Medicine); Cathleen S. Colón-Emeric (Duke University School of Medicine); Sarah B. Peskoe (Duke University School of Medicine)

Identifying signals that predict physical resilience, defined as one's ability to regain function following a health stressor, is essential for advancing precision approaches to recovery. High-dimensional datasets hold tremendous promise but pose analytic challenges due to correlation, distributed signals, instability, and sensitivity to analytic choices. The complexity of these data requires strategies that prioritize transparency and stability in variable selection. We present Exploratory Stability Selection (ESS), a resampling-based statistical framework designed for hypothesis-generating predictor discovery. ESS is an ensemble-style variable selection technique that integrates multiple resampling strategies and sparsity levels, enabling exploration of robustness and context-dependence across data perturbations. We demonstrate the utilization of ESS using the PRIME-KNEE study, which examines physical resilience in older adults undergoing elective total knee arthroplasty. ESS analyses were applied to clinical-only, plasma biomarker-only, and combined predictor sets to evaluate the stability and competitiveness of candidate variables associated with the probability of a highly resilient recovery trajectory for pain interference. ESS retains configuration-level results and summarizes stability metrics across configurations to provide insight into the

subsequent prioritization and validation of candidate predictors. This framework is well-suited for hypothesis-generating variable selection problems common to exploratory resilience research and other aging-related applications with complex, multi-domain predictor sets.

Untangling aging with multimorbidity: a nursing data science perspective

Wenhui (Vivian) Zhang, Center for Data Science, Nell Hodgson Woodruff School of Nursing, Emory University

Wenhui (Vivian) Zhang, PhD, MS, Center for Data Science, Nell Hodgson Woodruff School of Nursing, Emory University; Na You, MPH, Rollins School of Public Health, Emory University; Fan Fei, MPH, Rollins School of Public Health, Emory University; Chenyu Zhong, MPH, Rollins School of Public Health, Emory University; Rumshah Iftikhar, BDS, MPH candidate, Rollins School of Public Health, Emory University; Maytal Salomon, Nell Hodgson Woodruff School of Nursing, Emory University; Chaeeun Kim, BS, MPH candidate, Rollins School of Public Health, Emory University; Yilin Zhang, MPH, Rollins School of Public Health, Emory University; Runming Guo, MPH, Rollins School of Public Health, Emory University; Abha Namjoshi, MBBS, MPH, Rollins School of Public Health; Vicki Stover Hertzberg, PhD, FASA, Center for Data Science, Nell Hodgson Woodruff School of Nursing, Emory University

Multimorbidity is complex to study and intervene, especially when the chronic conditions increase functional limitations and decrease life expectancy. Arthritis and diabetes are both common and disabling comorbidities. Multimorbidity typologies are categorized by resource utilization, symptoms, and self-management concordance. Multiple conceptual frameworks and statistical or machine learning methods have been developed to untangle this complexity. A combined nursing data science perspective will support model explainability and further clinical translations. With a nursing data science perspective, we will discuss our past and ongoing efforts to untangle aging with multimorbidity. First, we will discuss our cross-sectional survey study on the self-regulation (SR) strategies of selection, optimization and compensation (SOC) in people aging with arthritis and multimorbidity. Second, we will discuss our study based on Nell data from Emory Healthcare electronic health records. We will introduce ongoing analysis and results using visualizations and machine learning that explored arthritis multimorbidity typologies and trajectories and the impacts of social determinants. Furthermore, we will discuss a recently funded project on diabetes multimorbidity data harmonization and an audio diary mobile application to collect data from older adults with diabetes and multimorbidity, and an ongoing study on delirium and dementia interrelationships that could connect multimorbidity, delirium, and dementia.

Benchmarking statistical and machine learning methods for predicting time to mobility limitation in older adults

Jaime Lynn Speiser, Wake Forest University School of Medicine

Jaime Speiser (Wake Forest); Michelle Shardell (Maryland)

Background: Time to event models are central to aging research, yet guidance on selecting among traditional statistical and machine learning (ML) methods remains limited, particularly for real world aging cohorts characterized by long follow up, censoring, and heterogeneous risk trajectories. Mobility limitation is a clinically important aging outcome with such complexities, providing a relevant test case for methodological benchmarking. Methods: We conducted a systematic benchmark comparison of 14 statistical and ML time-to-event methods using harmonized data from four longitudinal U.S. aging cohorts (N = 14,582). The outcome was time to incident mobility limitation, defined by self-reported walking difficulty or gait speed < 0.8 m/s. Models were evaluated using 4-fold cross validation repeated 100 times within cohorts. Performance metrics captured complementary dimensions of predictive quality: discrimination (C index), accuracy and calibration (Brier score, integrated log loss, calibration index), and computational efficiency (training time). Prespecified decision rules synthesized results across metrics to identify preferable methods. Results: Across pooled analyses, multiple methods demonstrated comparable discrimination, including penalized Cox models, tree-based ensembles, and boosting approaches. Elastic net Cox regression achieved the most favorable overall trade off, combining high discrimination, good calibration, low prediction error, and the fastest training times. When computational efficiency was deprioritized, accelerated oblique random survival forests, conditional inference forests, and random survival forests were also competitive. Conclusions: In aging cohort data, regularized Cox models, particularly elastic net, may be preferable compared to other methods. These findings support penalized regression as a strong default strategy for time-to-event modeling in gerontologic research and highlight the importance of joint evaluation of discrimination, calibration, and feasibility.

Parallel Sessions 4: Monday, November 2, 2026, 10:30 am–12:00 pm

Session 94: Statistical Foundations for Reliable and Interpretable AI

Location: Salon C

Organizer: Xiwei Tang (University of Texas at Dallas)

Chair: Xiwei Tang (University of Texas at Dallas)

Presenters: Jie Ding; Hengrui Luo; Jacob Bien; Xuan Bi

The future of AI scientists: emerging directions and fundamental challenges

Jie Ding, University of Minnesota

Jie Ding (University of Minnesota)

Modern AI systems are rapidly evolving from passive tools into agentic systems that can reason, learn, collaborate, and interact with human scientists in the research process. This talk discusses emerging directions in AI scientist systems, drawing inspiration from human-centered principles such as lifelong learning, teaching, collaboration, objective alignment, and auditing. The goal is to identify fundamental research challenges and stimulate new perspectives on building AI systems that can operate reliably, adaptively, and transparently in complex real-world research settings.

Asymptotic optimism for tensor regression models with applications to neural network compression

Hengrui Luo, Rice University

Haoming Shi, Eric C. Chi, Hengrui Luo

We study rank selection for low-rank tensor regression under random covariates design. Under a Gaussian random-design model and some mild conditions, we derive population expressions for the expected training–testing discrepancy (optimism) for both CP and Tucker decomposition. We further demonstrate that the optimism is minimized at the true tensor rank for both CP and Tucker regression. This yields a prediction-oriented rank-selection rule that aligns with cross-validation and extends naturally to tensor-model averaging. We also discuss conditions under which under- or over-ranked models may appear preferable, thereby clarifying the scope of the method. Finally, we showcase its practical utility on a real-world image regression task and extend its application to tensor-based compression of neural network, highlighting its potential for model selection in deep learning.

LLM-informed Gaussian process regression

Jacob Bien, University of Southern California

Farhad de Sousa (University of Southern California); Arash Amini (University of California Los Angeles); Jacob Bien (University of Southern California)

Large language models contain a vast amount of human knowledge. Leveraging this information to improve prediction performance, especially when data sets are small, is a tempting goal. However to do so one must overcome several key challenges: (a) How do we decide how much to trust the LLM versus the data at hand? (b) How do we ensure that data remains private in situations where data cannot be inputted into an LLM? (c) How do we decide what to ask of the LLM to improve predictive performance? (d) How can we audit the influence that the LLM's information is having on the predictions? We develop a method built on Gaussian process regression that addresses these challenges.

Observation-level watermarking and detection for tabular data

Xuan Bi, University of Minnesota

Dongyu Cui (University of Minnesota); Xuan Bi (University of Minnesota)

With the development of generative AI, watermark techniques have been widely used to detect the authenticity of AI-generated data and protect the rights of users and creators. While it is already well applied in data types including imaging and text data, watermarking tabular data is still under-explored. Existing methods primarily focus on numerical data, leaving discrete, categorical, and mixed data less studied. In this work, we propose STAMP (Single-observation Tabular Attribution and Marking Procedure), a novel framework for watermarking tabular data that can accommodate and preserve a wide range of distributions. We also develop a corresponding detection mechanism, which can reliably identify watermarks even when the sample size is as small as one. We establish theoretical guarantees for asymptotic consistency and detection accuracy. Finally, through extensive simulation studies and two real-data applications, we demonstrate that the proposed method is effective and robust to subsetting, while maintaining data fidelity and a high detection rate.

Session 86: Machine Learning for Causal Inference and Econometrics

Location: Williamsburg

Organizer: Lihua Lei (Stanford University)

Chair: Lihua Lei (Stanford University)

Presenters: Rahul Singh; Elena Manresa; Xiaohong Chen; Sokbae Lee

Uniform inference for kernel instrumental variable regression

Rahul Singh, Harvard University

Marvin Lob (QuantCo); Rahul Singh (Harvard University); Suhas Vijaykumar (UC San Diego)

Instrumental variable regression is a foundational tool for causal analysis across the social and biomedical sciences. Recent advances use kernel methods to estimate nonparametric causal relationships, with general data types, while retaining a simple closed-form expression.

Empirical researchers ultimately need reliable inference on causal estimates; however, uniform confidence sets for the method remain unavailable. To fill this gap, we develop valid and sharp confidence sets for kernel instrumental variable regression, allowing general nonlinearities and data types. Computationally, our bootstrap procedure requires only a single run of the kernel instrumental variable regression estimator. Theoretically, it relies on the same key assumptions. Overall, we provide a practical procedure for inference that substantially increases the value of kernel methods for causal analysis.

Title TBA

Elena Manresa, Princeton University

Abstract TBA

Adaptive hypothesis testing for nonparametric quantile IV models in high dimensions

Xiaohong Chen, Yale University

Chunrong Ai (CUHKSZ); Xiaohong Chen (Yale Univ); Zeqi Wu (Renmin Univ)

We propose a new adaptive test of equality and inequality restrictions on structural functions in nonparametric quantile instrumental variables models, allowing for many endogenous variables and conditional instrumental variables. The test is based on a diagonal-deleted quadratic form in restricted quantile residuals. An augmented instrument sieve combines directions guided by a structural dictionary with a baseline space that approximates conditional-moment violations. We select the range of sieve dimensions by a random exponential scan and use Bonferroni-adjusted chi-squared critical values. Implementation requires neither an unrestricted structural estimator nor conditional-density estimation or bootstrap critical values. Under the stated regularity and null-calibration conditions, the test controls size uniformly over the composite null and has power against alternatives separated in the conditional-moment norm. We establish a minimax lower bound for this separation criterion and an adaptive upper bound that matches it up to a log-log factor. Additional inverse-link conditions yield sufficient separation rates in structural L^2 distance. Random-feature neural-network and Random-tensor product B-spline implementations accommodate models with many regressors and instruments. Simulation studies examine additivity and monotonicity restrictions in quantile and mean IV models, including multi-product demand. Empirical applications test Slutsky restrictions in gasoline demand and functional-form and price-monotonicity restrictions in cereal demand.

SLIM: stochastic learning and inference in overidentified models

Sokbae Lee, Columbia University

Xiaohong Chen (Yale University); Min Seong Kim (University of Connecticut); Sokbae Lee (Columbia University); Myung Hwan Seo (Hong Kong University of Science and Technology); Myunghyun Song (Columbia University)

We propose SLIM (Stochastic Learning and Inference in overidentified Models), a scalable stochastic approximation framework for nonlinear GMM. SLIM forms iterative updates from independent mini-batches of moments and their derivatives, producing unbiased directions that ensure almost-sure convergence. It requires neither a consistent initial estimator nor global convexity and accommodates both fixed-sample and random-sampling asymptotics. We further develop an optional second-order refinement achieving full-sample GMM efficiency and inference procedures based on random scaling and plug-in methods, including plug-in, debiased plug-in, and online versions of the Sargan-Hansen J-test tailored to stochastic learning.

Session 46: Statistical Principles and Optimization for High-Dimensional Learning and AI

Location: Fulton Ferry

Organizer: Quefeng Li (UNC Chapel Hill)

Chair: Quefeng Li (UNC Chapel Hill)

Presenters: Yufeng Liu; Iris Zhang; Xingye Qiao; Xiaowu Dai

Structured learning of individualized decision rules with complex options

Yufeng Liu, University of Michigan

Yufeng Liu, University of Michigan

Advances in medical research and clinical practice have greatly expanded the range of treatment options available for many diseases. While this increased flexibility creates opportunities for personalized care, it also poses significant challenges for learning optimal individualized decision rules (IDRs), particularly when the number of treatment options is large. In this talk, I will present a unified framework for learning IDRs by exploiting latent structure among treatment options. By identifying and leveraging similarities and shared patterns across treatments, the proposed approach improves statistical efficiency and leads to more accurate, stable, and interpretable decision rules. I will discuss the theoretical properties of the framework and illustrate its performance through simulation studies and real-data applications. The results demonstrate the potential of structured learning approaches to enhance personalized decision-making in complex treatment settings.

FASTER: feature alignment and structured transfer via efficient regularization

Iris Zhang, New York University

Iris Zhang & Yang Feng (New York University)

Transfer learning has been shown to substantially improve estimation and prediction when target data are limited, but most existing methods assume that the source and target domains share the same feature space. In many practical

settings, however, datasets only partially overlap in their covariates. We propose Faster (Feature Alignment and Structured Transfer via Efficient Regularization), a two-stage framework for high-dimensional generalized linear models with partially overlapping feature spaces, in the regime where the total feature dimension may grow much faster than any individual sample size. In the first stage, we estimate a sparse mapping from each source feature space to the target feature space using covariance information assembled from all available paired observations across domains; we consider both a Lasso-based approach and an alternative based on graphical Lasso. In the second stage, the aligned source datasets are combined with the target data through a structured transfer estimator, followed by a target-only debiasing step. We establish a high-probability guarantee for sparse recovery in the alignment step and derive an error bound for the final estimator. Simulation studies under linear and logistic models show that Faster outperforms target-only and shared-feature transfer baselines, while remaining competitive with an oracle benchmark that has access to complete data. Applied to NHANES data for predicting elevated blood pressure, Faster achieves consistently lower misclassification rates relative to target-only and shared-feature baselines.

Conformal prediction under limited feedback: from bandit with OOD detection to diluted bandit

Xingye Qiao, Binghamton University

Zhou Wang (Binghamton University); Xingye Qiao (Binghamton University)

Reliable uncertainty quantification is critical in high-stakes settings, but standard set-valued prediction assumes access to true labels, an assumption that fails when feedback is weak, as in many deployed interactive systems. This talk presents two methods for building rigorous prediction sets under such restricted feedback. The first method performs set-valued prediction with out-of-distribution detection under bandit feedback: the learner first probes with a single predicted label and observes only whether that pull was correct, then later delivers a prediction set as its actual output. Using an estimate of the unobserved true label and a loss that balances OOD detection against set size, it achieves prescribed per-class recall with a regret that converges at nearly a square-root rate in the number of rounds. The second method addresses a setting where the bandit feedback is even weaker: the learner presents a randomized prediction set and learns only whether the true label fell somewhere within it, witho

Prediction-powered conditional inference

Xiaowu Dai, University of California, Los Angeles

Yang Sui (University of California, Los Angeles); Jin Zhou (University of California, Los Angeles); Hua Zhou (University of California, Los Angeles); Xiaowu Dai (University of California, Los Angeles)

We study prediction-powered conditional inference in the setting where labeled data are scarce, unlabeled covariates are abundant, and a black-box machine-learning predictor is available. The goal is to perform statistical inference on conditional functionals evaluated at a fixed test point, such as conditional means, without imposing a parametric model for the conditional relationship. Our approach combines localization with prediction-based variance reduction. First, we introduce a reproducing kernel-based localization method that learns a data-adaptive weight function from covariates and reformulates the target conditional moment at the test point as a weighted unconditional moment. Second, we incorporate machine-learning predictions through a correction-based decomposition of this localized moment, yielding a prediction-powered estimator and confidence interval that reduce variance when the predictor is informative while preserving validity regardless of predictor accuracy. We establish nonasymptotic error bounds and minimax-optimal convergence rates for the resulting estimator, prove pointwise asymptotic normality with consistent variance estimation, and provide an explicit variance decomposition that characterizes how machine-learning predictions and unlabeled covariates improve statistical efficiency. Numerical experiments on simulated and real datasets demonstrate valid conditional coverage and substantially sharper confidence intervals than alternative methods.

Session 56: Novel Network Models and Model Selection

Location: Sunset Park

Organizer: Keith Levin (University of Wisconsin-Madison)

Chair: Keith Levin (University of Wisconsin-Madison)

Presenters: Jing Lei; MInh Tang; Srijan Sengupta; Ji Zhu

Least squares inference for data with network dependency

Jing Lei, Carnegie Mellon University

Haeun Moon (Seoul National University); Kehui Chen (University of Pittsburgh); Jing Lei (Carnegie Mellon University)

We study the statistical inference problem concerning regression coefficients in a classical linear regression model using least squares estimates. The analysis is conducted under circumstances where network dependency exists across units in the sample. Neglecting the dependency among regression error terms may lead to biased estimation of the asymptotic variance, resulting in spurious association. In this paper, we first establish a central limit theorem for the ordinary least squares estimate, with a verifiable dependence condition alongside corresponding neighborhood growth conditions. Subsequently, we propose a consistent estimator for the asymptotic variance of the estimated coefficients, which employs a data-driven method to balance the bias-variance trade-off. We find that the optimal tuning depends on the linear hypothesis under consideration and must be chosen adaptively. The presented theory and methods are illustrated and

supported by numerical experiments and data examples.

Eigenvector fluctuations and limit results for random graphs with infinite rank kernels

MInh Tang, North Carolina State University

Minh Tang (North Carolina State University); Joshua Cape (University of Wisconsin–Madison)

In this talk we present results on the behavior of the leading eigenvectors for independent edge undirected random graphs generated from a general latent position model whose link function is possibly infinite rank and also possibly indefinite. We first derive uniform error bounds in the two-to-infinity norm as well as row-wise normal approximations for the leading sample eigenvectors. We then build on these results to tackle two graph inference problems, namely (i) entrywise bounds for graphon estimation and (ii) testing for the equality of latent positions, the latter of which is achieved by proposing a rank-adaptive test statistic that converges in distribution to a weighted sum of independent chi-square random variables under the null hypothesis.

Network cross-validation and model selection via subsampling

Srijan Sengupta, North Carolina State University

Sayan Chakrabarty (University of Michigan); Srijan Sengupta (North Carolina State University); Yuguo Chen (University of Illinois Urbana-Champaign)

Complex and larger networks are becoming increasingly prevalent in scientific applications in various domains. Although a number of models and methods exist for such networks, cross-validation on networks remains challenging due to the unique structure of network data. In this paper, we propose a general cross-validation procedure called NETCROP (NETwork CRoss-Validation using Overlapping Partition). The key idea is to divide the original network into multiple subnetworks with a shared overlap part, producing training sets consisting of the subnetworks and a test set with the node pairs between the subnetworks. This train-test split provides the basis for a network cross-validation procedure that can be applied on a wide range of model selection and parameter tuning problems for networks. The method is computationally efficient for large networks as it uses smaller subnetworks for the training step. We provide methodological details and theoretical guarantees for several model selection and parameter tuning tasks using NETCROP. Numerical results demonstrate that NETCROP performs accurate cross-validation on a diverse set of network model selection and parameter tuning problems. The results also indicate that NETCROP is computationally much faster while being often more accurate than the existing methods for network cross-validation.

Statistical inference for network latent space models with edge covariates

Ji Zhu, University of Michigan

Feifan Jiang (University of Michigan); Gongjun Xu (University of Michigan); Ji Zhu (University of Michigan)

Latent space models (LSMs) provide a powerful framework for analyzing network data by embedding nodes in a latent space. Incorporating covariate information via edge covariates offers an important generalization that strengthens both the interpretability and practical utility of the model. However, we show that coefficient estimates for edge covariate effects obtained through maximum likelihood estimation exhibit asymptotic bias due to high-order geometric effects and errors in latent variable estimation. To address this issue, we propose a plug-in bias-correction estimator that enables asymptotically valid and unbiased statistical inference for the effects of edge covariates. We establish theoretical guaranties, including consistency and asymptotic normality, under various network structures. Extensive simulations and real-world data examples demonstrate that our method effectively reduces estimation bias and improves the accuracy of inference. Our findings contribute to the statistical methodology of LSMs by providing a principled framework for unbiased parameter estimation in network models with edge covariates.

Session 50: Uncertainty, Dependence, and Structure in Modern Data Analysis

Location: Golden Boardroom

Organizer: Yufeng Liu (University of Michigan)

Chair: Hang Zhou (University of North Carolina)

Presenters: Ziping Xu; Weining Shen; José A. Sánchez Gómez; Kai Zhang

Statistical inference for misspecified contextual bandits

Ziping Xu, University of North Carolina at Chapel Hill

Yongyi Guo (Department of Statistics, University of Wisconsin–Madison) and Ziping Xu (School of Data and Information Sciences, University of North Carolina at Chapel Hill)

Contextual bandit algorithms have transformed modern experimentation by enabling real-time adaptation for personalized treatment and efficient use of data. Yet these advantages create challenges for statistical inference due to adaptivity. A fundamental property that supports valid inference is policy convergence, meaning that action-selection probabilities converge in probability given the context. Convergence ensures replicability of adaptive experiments and stability of online algorithms. In this paper, we highlight a previously overlooked issue: widely used algorithms such as LinUCB may fail to converge when the reward model is misspecified, and such non-convergence creates fundamental obstacles for statistical inference. This issue is practically important, as misspecified models—such as linear approximations of complex dynamic system—are often employed in real-world adaptive experiments to balance bias and variance.

Motivated by this insight, we propose and analyze a broad class of algorithms that are guaranteed to converge even under model misspecification. Building on this guarantee, we develop a general inference framework based on an inverse-probability-weighted Z-estimator (IPW-Z) and establish its asymptotic normality with a consistent variance estimator. Simulation studies con- firm that the proposed method provides robust and data-efficient confidence intervals, and can outperform existing approaches that exist only in the special case of offline policy evaluation. Taken together, our results underscore the importance of designing adaptive algorithms with built-in convergence guarantees to enable stable experimentation and valid statistical inference in practice.

Statistical analysis of 3D path data in Minecraft

Weining Shen, University of California, Irvine

Weining Shen

Understanding spatial navigation and memory formation is critical to exploring how humans learn and adapt in complex environments. To investigate these processes, we conducted an experiment using the Minecraft Memory and Navigation Task, collecting detailed 3D path data in a virtual open-world setting. Statistically, we developed a novel methodology to convert complex high-dimensional 3D movement data into functional representations, enabling standardized comparisons and analyses across participants and environments. We applied techniques such as functional clustering and regression to identify navigation patterns and their relationships with cognitive map development and memory retention. Our analysis uncovered two significant insights: first, participants who adopted moderately exploratory behaviors during training demonstrated superior retention of object locations; second, inefficient navigation strategies were strongly linked to poorer spatial memory and navigation performance. These findings highlight the effectiveness of our methodology in advancing the study of navigation behaviors and cognitive processes in dynamic 3D environments.

Quantifying moderator effects on high-dimensional multi-subject time series

José A. Sánchez Gómez, University of California, Riverside

José Á. Sánchez Gómez (UCR); Holly P. O'Rourke (UCR); Gene Brewer (UCR)

Researchers across many areas use time-series data to estimate relationships among variables over time. These relationships can be modeled via Granger-causal networks (GCNs), where nodes correspond to variables and directed edges indicate lagged temporal relationships among them. When time-series data are gathered over multiple subjects, GCN models must account for the presence of shared and subject-specific edges across the GCNs. While existing methods can estimate GCNs on multi-subject data, they often lack interpretable explanations for the source of the observed heterogeneity. In the present paper, we propose the moderator vector autoregressive model (MOD-VAR) for interpretable estimation of GCNs in multi-subject time-series data. To this end, we decompose the GCNs into shared, moderator-dependent, and subject-specific components. The use of moderator-dependent components can enhance both the quantification and interpretability of heterogeneous GCNs. Our theoretical results establish that MOD-VAR improves multi-subject GCN estimation in high-dimensional regimes with substantial heterogeneity. Furthermore, extensive numerical simulations demonstrate that MOD-VAR outperforms existing methods in terms of both GCN estimation and forecasting in a variety of settings. Applying MOD-VAR to a resting-state fMRI dataset, we recover moderating effects of age and locus coeruleus structural integrity on signaling within and between the salience, frontoparietal, and default mode networks.

BELIEF in dependence

Kai Zhang, University of North Carolina at Chapel Hill

Benjamin Brown (University of North Carolina at Chapel Hill); Kai Zhang (University of North Carolina at Chapel Hill); Xiao-Li Meng (Harvard University)

Two linearly uncorrelated binary variables must be also independent because non-linear dependence cannot manifest with only two possible states. This inherent linearity is the atom of dependency constituting any complex form of relationship. Inspired by this observation, we develop a framework called binary expansion linear effect (BELIEF) for understanding arbitrary relationships with a binary outcome. Models from the BELIEF framework are easily interpretable because they describe the association of binary variables in the language of linear models, yielding convenient theoretical insight and striking Gaussian parallels. With BELIEF, one may study generalized linear models (GLM) through transparent linear models, providing insight into how the choice of link affects modeling. For example, setting a GLM interaction coefficient to zero does not necessarily lead to the kind of no-interaction model assumption as understood under their linear model counterparts. Furthermore, for a binary response, maximum likelihood estimation for GLMs paradoxically fails under complete separation, when the data are most discriminative, whereas BELIEF estimation automatically reveals the perfect predictor in the data that is responsible for complete separation. We explore these phenomena and provide related theoretical results. We also provide preliminary empirical demonstration of some theoretical results. This is joint work with Benjamin Brown and Xiao-Li Meng

Session 35: Statistical Inference and Learning in High-Dimensional Structured Models

Location: Carroll Gardens

Organizer: Marianna Pensky (University of Central Florida)

Chair: Marianna Pensky (University of Central Florida)

Presenters: Joshua Cape; Joshua Agterberg; Olga Klopp; Nalin Arora

Extreme value theory for singular subspace estimation in the matrix denoising model

Joshua Cape, University of Wisconsin–Madison

Junhyung Chang (University of Wisconsin–Madison); Joshua Cape (University of Wisconsin–Madison)

This talk considers fine-grained singular subspace estimation in the matrix denoising model where a deterministic low-rank signal matrix is additively perturbed by a stochastic matrix of Gaussian noise. We establish that the maximum Euclidean row norm (i.e., the two-to-infinity norm) of the aligned difference between the leading sample and population singular vectors approaches the Gumbel distribution in the large-matrix limit, under suitable signal-to-noise conditions and after appropriate centering and scaling. We apply this novel asymptotic distributional theory to test hypotheses of low-rank signal structure encoded in the leading singular vectors and their corresponding principal subspace. We provide de-biased estimators for the corresponding nuisance signal singular values and show that the proposed plug-in test statistic has desirable properties. Notably, compared to using the Frobenius norm subspace distance, our test statistic based on the two-to-infinity norm empirically has higher power to detect structured alternatives that differ from the null in only a few matrix entries or rows. Our main results are obtained by a novel synthesis of and technical analysis involving row-wise matrix perturbation analysis, extreme value theory, saddle point approximation methods, and random matrix theory. Our contributions complement the existing literature for matrix denoising focused on minimax estimation, mean squared error analysis, unitarily invariant distances between subspaces, component-wise asymptotic distributional theory, and row-wise uniform error bounds. Numerical simulations illustrate our main results and demonstrate the robustness properties of our testing procedure to non-Gaussian noise distributions.

A high-dimensional statistical theory for convex and nonconvex matrix sensing

Joshua Agterberg, University of Illinois Urbana-Champaign

Joshua Agterberg (University of Illinois Urbana-Champaign), Rene Vidal (University of Pennsylvania)

Matrix sensing, or trace regression, is a problem wherein one wishes to estimate a low-rank matrix from linear measurements perturbed with noise. A number of existing works have studied both convex and nonconvex approaches to this problem, establishing minimax error rates when the number of measurements is sufficiently large relative to the rank and dimension of the low-rank matrix, though a precise comparison of these procedures still remains unexplored. In this work we provide a high-dimensional statistical analysis for symmetric low-rank matrix sensing observed under Gaussian measurements and noise. Our main result describes a novel phenomenon: in this statistical model and in an appropriate asymptotic regime, the behavior of any local minimum of the nonconvex factorized approach (with known rank) is approximately equivalent to that of the matrix hard-thresholding of a corresponding matrix denoising problem, and the behavior of the convex nuclear-norm regularized least squares approach is approximately equivalent to that of matrix soft-thresholding of the same matrix denoising problem. Here “approximately equivalent” is understood in the sense of concentration of Lipchitz functions. As a consequence, the nonconvex procedure uniformly dominates the convex approach in mean squared error.

Semi-supervised learning on graphs with GNNs

Olga Klopp, ESSEC Business School

J.Chen(Xiamen University), C. Donnat (Chicago University), Olga Klopp (Essec business school), Johannes Schmidt-Hieber (University of Twente)

We study semi-supervised node prediction on graphs where responses arise from a graph-aware feature operator followed by a smooth regression map. Within a class combining skip-connected GCN propagation with a fully connected ReLU network, we (i) derive an oracle inequality for population risk under random label masks that separates approximation and estimation error and exposes dependence on the labeled fraction, covering numbers, and a receptive-field constant; (ii) show skip connections exactly represent multi-hop polynomial filters, mitigating over-smoothing; (iii) give covering-number bounds; and (iv) quantify robustness of our algorithm. These results link classical graph regularization and modern GNN design.

Statistical inference and learning in high-dimensional structured models

Nalin Arora, Cornell University

Nalin Arora (Cornell University), Himel Mallick (Cornell University)

Multimodal artificial intelligence (AI) has advanced considerably over the past decade, making multimodal data integration increasingly central to biomedical research. However, most existing analytic approaches remain limited to association-based, single-modality modeling, and there is currently no unified framework for causal mediation analysis across biological layers.

To address this gap, we introduce a population-scale multimodal mediation framework for characterizing causal relationships across multiomics data. The framework accommodates both parallel mediation structures (e.g., X

$\rightarrow M_1 \rightarrow Y$ and $X \rightarrow M_2 \rightarrow Y$) and sequential pathways (e.g., $X \rightarrow M_1 \rightarrow M_2 \rightarrow Y$), enabling systematic decomposition of cross-modal interactions and their influence on downstream outcomes.

We present Zentangler, a unified multimodal AI engine that integrates early, intermediate, and late fusion strategies, and supports continuous, binary, survival, and multiclass outcomes. This engine is embedded within a counterfactual mediation framework, allowing decomposition of total effects into natural direct and indirect components across complex mediation pathways.

Across large-scale microbiome multiomics datasets, including iHMP and FINRISK, Zentangler outperforms existing single-modality approaches and reveals biologically meaningful cross-modal relationships. Synthetic benchmarking further demonstrates its improved accuracy and robustness. The method is implemented as an open-source R/Bioconductor package, available at: https://github.com/himelmallick/Zentangler/

Session 14: Reliable Modeling and Prediction in Complex Data Analysis

Location: Coney Island

Organizer: Boxiang Wang (University of Iowa)

Chair: Boxiang Wang (University of Iowa)

Presenters: Archer Yi Yang; Yan Sun; Wei Qian; Kun Chen

Multivariate conformal selection

Archer Yi Yang, McGill University / Mila

Tian Bai, Yue Zhao, Xiang Yu, Archer Yang

Selecting high-quality candidates from large datasets is critical in applications such as drug discovery, precision medicine, and alignment of large language models (LLMs). While Conformal Selection (CS) provides rigorous uncertainty quantification, it is limited to univariate responses and scalar criteria. To address this issue, we propose Multivariate Conformal Selection (mCS), a generalization of CS designed for multivariate response settings. Our method introduces regional monotonicity and employs multivariate nonconformity scores to construct conformal p-values, enabling finite-sample False Discovery Rate (FDR) control. We present two variants: mCS-dist, using distance-based scores, and mCS-learn, which learns optimal scores via differentiable optimization. Experiments on simulated and real-world datasets demonstrate that mCS significantly improves selection power while maintaining FDR control, establishing it as a robust framework for multivariate selection tasks.

MultiRisk: multiple risk control via iterative score thresholding

Yan Sun, New Jersey Institute of Technology

Sunay Joshi (University of Pennsylvania), Yan Sun (New Jersey Institute of Technology), Hamed Hassani (University of Pennsylvania), Edgar Dobriban (University of Pennsylvania)

As generative AI systems are increasingly deployed in real-world applications, regulating multiple dimensions of model behavior has become essential. We focus on test-time filtering: a lightweight mechanism for behavior control that compares performance scores to estimated thresholds, and modifies outputs when these bounds are violated. We formalize the problem of enforcing multiple risk constraints with user-defined priorities, and introduce two efficient dynamic programming algorithms that leverage this sequential structure. The first, MULTIRISK-BASE, provides a direct finite-sample procedure for selecting thresholds, while the second, MULTIRISK, leverages data exchangeability to guarantee simultaneous control of the risks. Under mild assumptions, we show that MULTIRISK achieves nearly tight control of all constraint risks. The analysis requires an intricate iterative argument, upper bounding the risks by introducing several forms of intermediate symmetrized risk functions, and carefully lower bounding the risks by recursively counting jumps in symmetrized risk functions between appropriate risk levels. We evaluate our framework on a three-constraint Large Language Model alignment task using the PKU-SafeRLHF dataset, where the goal is to maximize helpfulness subject to multiple safety constraints, and where scores are generated by a Large Language Model judge and a perplexity filter. Our experimental results show that our algorithm can control each individual risk at close to the target level.

On fairness-assisted sequential variable selection

Wei Qian, University of Delaware

Zeyu Chen; Wei Qian; Bintong Chen (University of Delaware)

In high-dimensional modeling, sequential variable selection is an important technique for interpretable and efficient model construction. We propose a new stepwise sparsity learning method for generalized linear models that integrates fairness considerations into the sequential variable selection process. We further introduce a model averaging framework under fairness constraints, which aggregates a collection of sparse model candidates generated via the sequential variable selection method to produce asymptotically fair estimators and achieves near-optimal estimation risk in flexible scenarios. The proposed approach is demonstrated through an application to a community crime dataset to showcase its practical utility in predictive modeling.

Contrastive and atomic regularization for feature aggregation on trees

Kun Chen, University of Connecticut

Kun Chen

In supervised learning, features organized in a known hierarchy can be aggregated to reduce dimensionality and improve model estimation, but different branches may require different levels of resolution. Motivated by two complementary views of tree-guided feature aggregation,

we propose CARAT, Contrastive and Atomic Regularization for Aggregation on Trees. One component encourages equal effects within branches by shrinking their internal contrasts; the other regularizes shared effect sizes through their least costly representation using subtree aggregates. Together, these components use the hierarchy to learn local resolution while estimating the effects that remain. Our theory explains how branch size and neighboring effects influence shrinkage and gives conditions for recovering the underlying homogeneous branches. Numerical studies show favorable estimation and prediction performance in sparse rare-feature and correlated settings.

Session 19: Emerging Methods and Applications in Single-Cell Analysis

Location: Park Slope

Organizer: Guanyu Hu (Michigan State University)

Chair: Guanyu Hu (Michigan State University)

Presenters: Hao Feng; Huiyan Sang; Chang Su; Guanao Yan

Detecting differential DNA methylation dynamics in single cell

Hao Feng, The University of Texas Health Science Center at Houston

Hao Feng (The University of Texas Health Science Center at Houston)

Recent advancements in single-cell DNA methylation (scDNAm) sequencing technologies have enabled the profiling of epigenetic landscapes at unprecedented resolution, offering insights into cellular heterogeneity, differentiation and evolution. Trajectory inference, which orders cells along pseudotime, allows researchers to track genomics changes across continuous cell states and identify key loci exhibiting differential methylation. However, no methods currently exist to model methylation changes along pseudotime in scDNAm data. Here, we present a hierarchical Bayesian framework for scDNAm data analysis. Our method, named mist (methylation inference for single-cell along trajectory), models stage-specific biological variations, identifies genomic features with significant methylation changes along pseudotime, and performs Differential Methylation (DM) analysis across phenotypical groups. mist is publicly available as an R/Bioconductor package at https://bioconductor.org/packages/mist.

Consistent Bayesian local spatial feature selection with application to spatial multimodal omics

Huiyan Sang, Texas A&M University

Kun Huang (Texas A&M University); Xiyu Peng (Texas A&M University); Huiyan Sang (Texas A&M University); Ligang Lu (Shell Inc.)

Motivated by a high-dimensional regression problem in spatial multimodal omics (SMO), we propose a Bayesian framework for local spatial feature selection, where a random domain partition prior is introduced to divide the spatial domain into several contiguous clusters with flexible shapes and an unknown number of clusters, conditional on which a local feature selection prior is imposed within each cluster. The notion of "feature" is general and may include both covariates and functional bases, allowing the framework to perform both local variable selection and local basis selection, the latter being essential for adaptively approximating spatially varying functions with localized characteristics. We derive coupled hyperparameter conditions linking domain partition and local feature selection priors, under which the consistency theory and posterior contraction rates of both the domain partition and feature selection are established. We develop an efficient informed reversible jump Markov chain Monte Carlo algorithm to address the computational challenges encountered in joint posterior sampling of domain partitions and selected features. Simulation studies demonstrate the effectiveness of the proposed model and algorithm, highlighting its advantages over existing methods. The application of our model to an SMO dataset reveals biologically meaningful spatial patterns within breast cancer tissue.

Error-in-variables graphical model for gene regulatory network inference with single-cell data

Chang Su, Emory University

Weiwei Wu (Emory University); Jingfei Zhang (Emory University); Chang Su (Emory University)

Gene regulatory networks (GRNs) capture direct regulatory relationships between transcription factors (TF) and their target genes, providing insights into the biological processes underlying complex diseases. Gene regulation is highly heterogeneous across cell types, and recent advances in single-cell technologies have enabled the inference of cell-type-specific GRNs. However, inferring direct associations between TFs and target genes remains challenging due to substantial technical noise in single-cell data, which may lead to inaccurate network estimates and invalid statistical inference. As a result, existing methods often yield GRNs with low reproducibility across datasets. To this end, we develop a high-dimensional graphical model to infer direct associations among genes. It accounts for technical noise through an error-in-variables design and enables rigorous statistical tests via a debiased statistical inference procedure. This model scales to high-dimensional networks and achieves computationally efficiency by employing a penalized node-wise regression framework. Preliminary results demonstrate improved accuracy in GRN estimation, effective control of false positives, and increased statistical power. Application to Alzheimer's disease (AD) data reveals TF-driven GRNs that govern AD risk genes in microglia. We will further discuss extensions to integrating multimodal data and testing differential GRNs across conditions.

scIsoSim: realistic simulation with isoform ground truths for benchmarking single-cell isoform and

splicing quantification methods

Guanao Yan, Michigan State University

Guanao Yan; Weijian Wang; Zhiyin Liu; Jingyi Jessica Li

Alternative splicing generates multiple RNA isoforms from a single gene. However, quantifying isoforms and alternative splicing (AS) at single-cell resolution is hindered by the lack of ground truths in experimental data. To address this gap, we present scIsoSim, a simulator that generates realistic single-cell RNA-seq (scRNA-seq) reads from isoforms, for which ground-truth structures and abundances are specified. Supporting both plate-based (Smart-seq2) and droplet-based (10x Genomics) technologies, scIsoSim mimics read sequences and splicing patterns observed in experimental data. Our benchmarking reveals that, for Smart-seq2 data, bulk RNA-seq isoform quantification methods, such as Salmon and Kallisto, remain applicable with high accuracy, with Salmon exhibiting superior computational efficiency. For 10x Genomics data, Scasa, a scRNA-seq isoform quantification method, has low accuracy due to data sparsity. Our benchmarking also evaluates single-cell AS quantification methods, with BRIE achieving higher precision and outrigger yielding higher recall. As a simulation framework with isoform ground truths, scIsoSim guides method selection for single-cell isoform and AS analyses and supports the benchmarking of future methods.

Session 21: Learning from Neuroimaging Data: Statistical and AI Methods for Brain Aging

Location: Windsor Terrace

Organizer: Jun Yan (University of Connecticut)

Chair: Panpan Zhang (Vanderbilt University Medical Center)

Presenters: Marina Vannucci; Shuo Chen; Martin Lindquist; Chenxi Yuan

Bayesian brain network mediation analysis of longitudinal Alzheimer's disease data

Marina Vannucci, Rice University

Marina Vannucci (Rice University); Janet Fu (Rice University); Suprateek Kundu (UT MD Anderson Cancer Center)

Mediation analysis provides critical insights into how exposures influence outcomes through intermediate variables, or mediators. In this study, we examine mediation effects of complex-structured data represented as whole-brain connectivity networks derived from resting state fMRI data on Alzheimer's Disease subjects and healthy controls. Our modeling approach incorporates low-dimensional latent scales representations of the high-dimensional connectivity matrices, while preserving node-level characteristics and facilitating the identification of key mediating brain regions. It also accommodates flexible non-linear relationships between the mediators and the outcomes of interest. Using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), we perform cross-sectional and longitudinal analyses to investigate the effects of several factors known to affect cognitive decline.

Systematic prediction bias in ML/AI regression and biological aging clocks

Shuo Chen, University of Maryland

Hwiyoung Lee (Georgia Southern University); Zhenyao Ye (University of Maryland); Shuo Chen (University of Maryland)

Aging clocks aim to estimate an individual's underlying physiological and molecular age, distinct from chronological age. Because biological age cannot be measured directly, aging clocks are computed indices derived from high-throughput multi-omics and imaging data using techniques such as machine learning (ML) and artificial intelligence (AI). Aging clocks derived from standard ML/AI regression models are known to exhibit systematic prediction bias that shrinks predicted values toward the mean, leading to underestimation in older participants and overestimation in younger participants. Here, we elaborate on the causes of this bias inherited from ML/AI regression models and emphasize that it can compromise downstream analyses, for example, by inducing biased association estimates between accelerated aging clocks and risk factors. We further present ML regression strategies that automatically remove systematic bias during aging-clock construction and ensure unbiased inference in downstream analyses.

Latent directions of mediation analysis

Martin Lindquist, Johns Hopkins University

Martin Lindquist (Johns Hopkins University)

Mediation analysis investigates the role of intermediate variables that lie in the path between a treatment and an outcome variable. Classical methods assume a single scalar mediator. Many modern studies, however, measure high-dimensional mediators, such as brain region volumes, gene expression panels, or curves recorded over time, where no single feature captures the pathway. We propose latent direction mediation, a framework that finds the weighted combination of features that best transmits the treatment effect to the outcome. Because this combined score is never directly observed, we show that it can be eliminated algebraically, so estimation requires only the observed treatment, outcome, and feature data. Finding the strongest mediating direction then reduces to a standard eigenvalue problem, which is fast and straightforward to solve. The same framework handles both multivariate features and functional data, where the estimated direction becomes an interpretable weight curve over time. We apply the approach to data from an fMRI study of thermal pain.

Recovery of missing neuroimaging using diffusion models with application to Alzheimer's disease

Chenxi Yuan, New Jersey Institute of Technology

Chenxi Yuan (Department of Informatics, Ying Wu College of

Computing, New Jersey Institute of Technology); Nicholas J. Tustison (Department of Radiology and Medical Imaging, School of Medicine, University of Virginia); Rebecca A. Hubbard (Department of Biostatistics, Epidemiology & Informatics, Perelman School of Medicine, University of Pennsylvania); Kristin A. Linn (Department of Biostatistics, Epidemiology & Informatics, Perelman School of Medicine, University of Pennsylvania)

Missing data is a significant challenge in medical research. In longitudinal studies of Alzheimer's disease (AD) where structural magnetic resonance imaging (MRI) is collected from individuals at multiple time points, participants may miss a study visit or drop out. Additionally, technical issues such as participant motion in the scanner may result in unusable imaging data at designated visits. Such missing data may hinder the development of high-quality imaging-based biomarkers. To address the problem of missing MRI data in studies of AD, we introduced a novel 3D diffusion model specifically designed for imputing missing structural MRI. The model generates a whole-brain image conditional on a single structural MRI observed at a past visit or conditional on one past and one future observed structural MRI relative to the missing observation. Our 3D diffusion model can impute missing structural MRI data at a single designated visit and outperforms alternative methods for imputing whole-brain images that are missing from longitudinal trajectories.

Session 03: Mediation Analysis and Causal Inference in Biomedical Research

Location: DUMBO

Organizer: Heping Zhang (Yale University School of Public Health)

Chair: Ying Wei (Columbia University)

Presenters: Mengling Liu; Qiao Liu; Zhonghua Liu; Haiqun Lin and Weiyi Xia

Mediator selection in linear and nonlinear mediation analysis via adaptive measurement errors infusion

Mengling Liu, New York University Grossman School of Medicine

Chen Liang (New York University Grossman School of Medicine); Wen Zhou (NYU School of Global Public Health); Mengling Liu (New York University Grossman School of Medicine)

Mediation analysis often involves multiple mediators, and selection among them is challenging when pathway signals are weak, mediator-outcome relationships are nonlinear, or mediator-exposure and mediator-mediator interactions exist; most existing methods rely on coefficient-based regularization under linear models. We propose a mediator selection framework based on measurement error modeling (MEM), which quantifies mediator importance by how much exposure-mediator and mediator-outcome pathway fit degrades under controlled infusion of artificial measurement error. Linear MEM distributes a constrained error-precision budget across mediators under a joint Gaussian criterion, selecting each pathway separately and retaining mediators selected in both; kernel MEM extends this to nonlinear mediator-outcome relationships by replacing outcome-model error precisions with coordinate-specific inverse bandwidths. In simulations, linear MEM was competitive with established linear-mediation methods, with an advantage under correlated mediator residuals, while kernel MEM substantially outperformed linear methods and a kernel-ridge importance benchmark when mediators entered the outcome model through quadratic or interaction terms. Applied to a DNA methylation study of childhood trauma and cortisol stress reactivity, MEM selected eight (linear) and three (kernel) loci, several previously reported. The MEM framework offers a practical post-screening approach to mediator selection under linear and nonlinear mediator-outcome relationships, yielding mediators supported by both pathways.

AI-powered Bayesian generative modeling approach for causal inference in observational studies

Qiao Liu, Yale University

Qiao Liu (Yale University); Wing Hung Wong (Stanford University)

Causal inference in observational studies with high-dimensional covariates presents significant challenges. We introduce CausalBGM, an AI-powered Bayesian generative modeling approach that captures the causal relationship among covariates, treatment, and outcome. The core innovation is to estimate the individual treatment effect (ITE) by learning the individual-specific distribution of a low-dimensional latent feature set (e.g., latent confounders) that drives changes in both treatment and outcome. This individualized posterior representation yields estimates of the individual treatment effect (ITE) together with well-calibrated posterior intervals while mitigating confounding effect. CausalBGM is fitted through an iterative algorithm to update the model parameters and the latent features until convergence. This framework leverages the power of AI to capture complex dependencies among variables while adhering to the Bayesian principles. Extensive experiments demonstrate that CausalBGM achieves superior or competitive performance compared to state-of-the-art methods, particularly in scenarios with high-dimensional covariates and large-scale datasets. By addressing key limitations of existing methods, CausalBGM emerges as a robust and promising framework for advancing causal inference in a wide range of modern applications. The code and tutorials for CausalBGM are available at https://github.com/liuq-lab/bayesgm.

Identification and inference for structural accelerated failure time models via instrumental variable interactions

Zhonghua Liu, Columbia University

Qiushi Bu (Columbia), Wen Su (CityU of Hong Kong), Xinyu Zhang (Chinese Academy of Sciences), Xingqiu Zhao (Hong Kong PolyU), Zhonghua Liu (Columbia)

We consider causal inference for right-censored

time-to-event outcomes in the presence of unmeasured confounding. Within a structural accelerated failure time model, we allow candidate instrumental variables to violate the exclusion and independence restrictions underlying classical instrumental variable methods. We show that the causal acceleration parameter can nevertheless be identified through centered interactions among mutually independent candidate instruments, provided that at least one such interaction is associated with the exposure. To accommodate right censoring, we construct augmented inverse probability censoring weighted moment functions that preserve these interaction-based identifying restrictions in the observed data. The resulting moments are Neyman orthogonal with respect to nuisance functions and possess a double robustness property. We combine potentially numerous and weak interaction moments using generalized empirical likelihood, and establish consistency and asymptotic normality under many-weak-moment asymptotics. We also develop an overidentification test for assessing the compatibility of the interaction-based moment restrictions with the observed data. Both simulation studies and an application to UK Biobank data demonstrate the practical value of this method.

A covariate-adjusted extension of the Bolck–Croon–Hagenaars method for latent class mediation analysis

Haiqun Lin (Rutgers University) and **Weiyi Xia** (Johnson & Johnson)

Weiyi Xia (Johnson & Johnson), Olga Jarrin, Haiqun Lin

Mediation analysis has been a highly valuable statistical method in many research and practice and fields. The mediation analysis is used to assess whether an exposure exerts its effect, on an outcome of interest through an intermediate variable (the mediator). Some studies require assessing multiple correlated observed variables for their mediation effects and a latent class mediator can be conveniently used to summarize the mediation effect of the multiple variables. The class with the highest posterior probability is assigned as the discrete latent class mediator. However, the misclassification of the latent class mediator often biases the effect of the mediation. In such a situation, an adjustment method called the Bolck–Croon–Hagenaars (BCH) approach, is typically used for adjusting the misclassification error when we assess the association between the latent class and an outcome of interest. The BCH method reweights the data by the inverse of the misclassification probabilities. This method accounts for the misclassification error from the latent class assignment. In mediation analysis, the exposure variable is a predictor for the latent class mediator as well as for the distal outcome, the BCH adjustment has not been done in this situation in the previous literature. This work proposes to modify the current BCH misclassification adjustment formula to accommodate predictors for the latent class membership.

Session 83: Statistical Methods for Dissecting Heterogeneity and Decision-Making in Precision Medicine Research

Location: Greenpoint

Organizer: Yuanjia Wang (Columbia University)

Chair: Yinjun Zhao (Columbia University)

Presenters: Judy Zhong; Annie J. Lee; Yuan Chen; Yuan Bian

Potential bias in efficacy estimates due to intercurrent treatment discontinuations

Judy Zhong, Weill Cornell Medicine

Bjorn Redfors (Weill Cornell Medicine); Jacky Choi (Weill Cornell Medicine); Mario Gaudino (Weill Cornell Medicine); Harmony R. Reynolds (Weill Cornell Medicine); Judy Zhong (Weill Cornell Medicine)

Aims: Treatment discontinuation is a common intercurrent event in randomized clinical trials (RCTs) that can bias treatment effect estimates, yet it remains under recognized and inconsistently addressed. Antithrombotic trials provide a relevant case, as bleeding-related discontinuation is common and more often occur in patients at increased risk also for ischaemic events. Methods and results: We reviewed large antithrombotic RCTs published in The New England Journal of Medicine (2020–2025) to assess reporting and analytic handling of treatment discontinuation. We also conducted simulations to examine how discontinuation patterns influence efficacy estimates under three scenarios: discontinuation only in the treatment arm, in both arms with similar patterns, and in both arms but outcome-associated only in the treatment arm. Treatment policy (intention-to-treat) and while-on-treatment strategies were compared within the ICH E9(R1) framework. Among fourteen reviewed antithrombotic RCTs, about one-third of patients discontinued treatment in trials with more than 6 months follow-up, with higher imbalance in open-label studies. Over half of discontinuations were safety-related, yet timing and temporary interruptions were rarely reported. Most reported both ITT and per-protocol (PP) analyses, with PP assuming non-informative censoring and yielding larger efficacy estimates. In simulations, treatment policy and while-on-treatment strategies were biased when discontinuation was outcome-associated, while Simon–Makuch curves helped identify prognostic discontinuation differing by treatment arm. Conclusion: Discontinuation is common, prognostically meaningful, and often oversimplified in analysis, threatening the validity of RCT findings. Explicitly addressing discontinuation as an intercurrent event within the estimand framework, along with stratified outcome curves, can improve interpretation and support more robust trial conclusions.

Integrative outcome-guided clustering through statistical learning of longitudinal clinical trajectories and high-dimensional omics

Annie J. Lee, Columbia University

Boyi Hu (The Department of Neurology, Vagelos College of Physicians and Surgeons, Columbia University); Yuanjia Wang (Department of Biostatistics, Columbia University); Annie J. Lee (The Department of Neurology, Vagelos College of Physicians and Surgeons, Columbia University; The Gertrude H. Sergievsky Center, Columbia University; The Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University)

Background: Integrating longitudinal outcomes with high-dimensional molecular data is a fundamental challenge in statistical learning, particularly for uncovering latent heterogeneity in progressive diseases. Most existing clustering and mixture models either ignore outcome dynamics or treat temporal trajectories and molecular features in separate stages, limiting interpretability and inferential validity. We propose Temporal Profile-Guided Subtyping Using High-Dimensional Omics Data (TPClust), a supervised, semi-parametric clustering method that jointly models longitudinal clinical trajectories and high-dimensional omics data within a unified statistical learning framework.

Methods: TPClust estimates latent cluster membership probabilities using multinomial logistic regression with structured sparsity-inducing regularization to enable feature selection in high-dimensional molecular data. Longitudinal outcomes are modeled using cluster-specific semi-parametric regression with B-spline bases, allowing flexible inference of nonlinear, time-varying effects of covariates. Model parameters are estimated via an expectation–maximization algorithm, and uncertainty in time-varying effects is quantified using bootstrap-based confidence intervals.

Results: Simulation studies showed that TPClust provides valid inference for time-varying effects and achieves robust high-dimensional feature selection compared with static mixture models. Applied to 1,020 participants from the Religious Orders Study and Memory and Aging Project, integrating longitudinal cognitive trajectories with prefrontal cortex transcriptomics and adjusting for demographic and vascular risk factors, TPClust identified four distinct aging clusters with markedly different cognitive trajectories, risk factor sensitivities, and molecular signatures, including intermediate subgroups not captured by conventional approaches. Cluster-specific differences were further supported by multimodal neuropathological, proteomic, and epigenomic analyses.

Conclusions: TPClust provides a robust method for integrative, outcome-guided clustering of longitudinal clinical and high-dimensional molecular data, bridging semi-parametric modeling, structured regularization, and latent variable inference. By uncovering coherent latent structure and subtype-specific molecular signatures, the method supports precise disease stratification and biomarker discovery, and informs the development of targeted interventions in complex longitudinal and multimodal data settings.

Learning locally interpretable individualized treatment rules for black-box decision models

Yuan Chen, Memorial Sloan Kettering Cancer Center

Yasin Khadem Charvadeh (Memorial Sloan Kettering Cancer Center); Katherine Panageas (Memorial Sloan Kettering Cancer Center); Yuan Chen (Memorial Sloan Kettering Cancer Center)

Individualized treatment rules (ITRs) aim to optimize healthcare by tailoring treatment decisions to patient-specific characteristics. Existing methods typically rely on either interpretable but inflexible models or highly flexible black-box approaches that sacrifice interpretability; moreover, most impose a single global decision rule across patients. We introduce the Locally Interpretable Individualized Treatment Rule (LI-ITR) method, which combines flexible machine learning models to accurately learn complex treatment outcomes with locally interpretable approximations to construct subject-specific treatment rules. LI-ITR employs variational autoencoders to generate realistic local synthetic samples and learns individualized decision rules through a mixture of interpretable experts. Simulation studies show that LI-ITR accurately recovers true subject-specific local coefficients and optimal treatment strategies. An application to precision side-effect management in breast cancer illustrates the necessity of flexible predictive modeling and highlights the practical utility of LI-ITR in estimating optimal treatment rules while providing transparent, clinically interpretable explanations.

Integrative learning of individualized treatment rules from multiple studies with partially overlapping treatments

Yuan Bian, Columbia University

Yuan Bian (Columbia University); Donglin Zeng (University of Michigan); Hyun-Joon Yang (Stanford University); Leanne M. Williams (Stanford University); Yuanjia Wang (Columbia University)

An individualized treatment rule (ITR) tailors treatments to a patient's specific characteristics. However, randomized controlled trials (RCTs) are often underpowered to detect the treatment effect heterogeneity needed for reliable ITR estimation. To address this limitation, there is growing interest in leveraging information from multiple studies to improve statistical power and support individualized decision-making. A key challenge in this context is that available RCTs may not evaluate the same set of treatments. In this paper, we propose an integrative learning framework that synthesizes evidence across multiple RCTs that share a common comparator but differ in their alternative treatment arms. Our method integrates information through a regularized weighted misclassification risk function and adaptively determines the contribution of each study to the ITRs of the others. We rigorously study the excess risk of the resulting estimator. Simulation studies demonstrate that the proposed approaches improve the estimation of both value and benefit functions. We illustrate the utility of our methodology using data from two landmark studies of

major depressive disorder: the Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care study and the International Study to Predict Optimized Treatment in Depression study, both of which include a selective serotonin reuptake inhibitor as a common treatment arm. We find that the separate learning method outperforms one-size-fits-all methods, and our integrative methods further improve performance.

Session 57: Advances in Statistical Learning for Data Integration

Location: Navy Yard

Organizer: Jing Ma (Fred Hutchinson Cancer Center)

Chair: Jing Ma (Fred Hutchinson Cancer Center)

Presenters: Yue Wang; Gen Li; Shulei Wang; Jing Ma

RoDPAA: a distance-based framework for integrative analysis of heterogeneous biological data

Yue Wang, University of Colorado Anschutz Medical Campus

Yanyan Xu (University of Colorado Anschutz Medical Campus), Yue Wang (University of Colorado Anschutz Medical Campus)

Joint dimension reduction is widely used for integrating multiview biological data. Most existing methods define joint structure through identical latent subspaces and operate directly with Euclidean geometry. These assumptions may be restrictive for heterogeneous biological data, where shared biological mechanisms often induce complementary rather than identical representations and non-Euclidean similarity measures are commonly used. We propose RoDPAA (Robust Distance-based Principal-Angle Analysis), a distance-based multiview dimension reduction framework that generalizes existing approaches in two ways. RoDPAA defines shared information through statistically significant principal-angle alignment, allowing partial information sharing between data views, and constructs latent representations from user-specified distance matrices to accommodate modality-specific data geometry. A permutation-based inference procedure identifies significant shared directions. Simulations show that RoDPAA accurately recovers shared structure under partial information sharing and heterogeneous data geometries while remaining competitive under classical settings. In host–microbiome integration across multiple inflammatory bowel disease (IBD) cohorts, RoDPAA identifies reproducible shared host–microbiome structure that is missed by existing methods. Therefore, RoDPAA provides a flexible and statistically principled framework for multiview dimension reduction by integrating distance-based data representation with principal-angle geometry. It broadens the applicability of joint dimension reduction to heterogeneous biological data while encompassing classical methods as special cases.

DELIVER: deep latent factor-adjusted variable selection in multi-omics regression

Gen Li, University of Michigan

Zongrui Dai (University of Michigan); Gen Li (University of Michigan)

The integration of multi-omics data provides a novel view of biological systems by combining diverse molecular modalities. However, identifying biomarkers across these modalities is frequently confounded by unobserved latent factors that induce complex, nonlinear dependencies, destabilizing traditional variable selection. While existing factor-based approaches attempt to mitigate this confounding, they predominantly rely on linear assumptions, which often fail to capture the complex dependencies in biological systems. In this article, we present DELIVER, a deep learning framework for latent factor disentanglement. By combining Deep Canonical Correlation Analysis (DeepCCA) with a Belted and Ensembled Neural Network (BENN), DELIVER learns a flexible shared latent representation and extracts purified individual components via nonlinear conditional expectation. This process decouples shared structural variation from modality-specific signals, enabling stable variable selection. Extensive simulations demonstrate that DELIVER outperforms factor-based models and lasso regression in both selection stability and coefficient estimation. Furthermore, application to the ORIGINS microbiome dataset across multiple taxonomic levels illustrates DELIVER's ability to accurately identify correlated biomarkers across phylogeny, offering novel insights into high-dimensional integrative omics research.

Microbiome data integration via shared dictionary learning

Shulei Wang, University of Illinois Urbana-Champaign

Bo Yuan (University of Illinois Urbana-Champaign); Shulei Wang (University of Illinois Urbana-Champaign)

Data integration is a powerful tool for facilitating a comprehensive and generalizable understanding of microbial communities and their association with outcomes of interest. However, integrating data sets from different studies remains a challenging problem because of severe batch effects, unobserved confounding variables, and high heterogeneity across data sets. We propose a new data integration method called MetaDICT, which initially estimates the batch effects by weighting methods in causal inference literature and then refines the estimation via novel shared dictionary learning. Compared with existing methods, MetaDICT can better avoid the overcorrection of batch effects and preserve biological variation when there exist unobserved confounding variables, data sets are highly heterogeneous across studies, or the batch is completely confounded with some covariates. Furthermore, MetaDICT can generate comparable embedding at both taxa and sample levels that can be used to unravel the hidden structure of the integrated data and improve the integrative analysis. Applications to synthetic and real microbiome data sets demonstrate the robustness

and effectiveness of MetaDICT in integrative analysis. Using MetaDICT, we characterize microbial interaction, identify generalizable microbial signatures, and enhance the accuracy of outcome prediction in two real integrative studies, including an integrative analysis of colorectal cancer metagenomics studies and a meta-analysis of immunotherapy microbiome studies.

Network-based integration of multi-view data

Jing Ma, Fred Hutchinson Cancer Center

Jing Ma (Fred Hutchinson Cancer Center)

Multi-view data arise when complementary sets of features are measured on the same observational units. A fundamental problem in multi-view integration is to identify associations between features across views while accounting for the large number of candidate relationships and their underlying structure. These cross-view associations can be represented as a bipartite network, with nodes corresponding to features from two views and edges representing their associations. Standard multiple-testing procedures, such as the Benjamini–Hochberg procedure, do not exploit higher-order organization in this network and can lose power when associations exhibit structured patterns.

We develop a network-based framework that learns and leverages latent structure across data views to improve detection of cross-view associations while controlling the false discovery rate (FDR). We model the bipartite association network using a stochastic block model, allowing features within each view to cluster into modules with distinct patterns of cross-view connectivity. A variational expectation–maximization algorithm learns this structure, which is then incorporated into multiple testing to improve power while retaining FDR control. The learned modules also provide an interpretable summary of cross-view organization. We demonstrate the method through simulations and an application integrating microbiome and metabolomics data in a study of bacterial vaginosis.

Session 23: The Interplay Between Statistics and Data-Driven Decision-Making

Location: Brighton Beach

Organizer: Zhimei Ren (University of Pennsylvania)

Chair: Ying Jin (University of Pennsylvania)

Presenters: Bryce McLaughlin; Yuxuan Han; Shuangning Li; Wanteng Ma

Inference-aware policy optimization

Bryce McLaughlin, Boston College

Hamsa Bastani (University of Pennsylvania); Osbert Bastani (University of Pennsylvania); Bryce McLaughlin (Boston College)

There has been a surge of recent interest in automatically learning policies to target treatment decisions based on rich individual covariates. In addition, practitioners want confidence that the learned policy has better performance than the incumbent policy according to downstream policy evaluation. However, evaluating these learned policies is challenging: model-based evaluation methods are subject to bias from the *winner's curse*—an issue where the policy learning procedure exploits the model's prediction errors—while probability weight-based evaluation methods can suffer excessive variance. We propose a novel strategy called *inference-aware policy optimization*, which modifies policy optimization to account for how the policy will be evaluated downstream. Specifically, it optimizes not only for the estimated objective value, but also for the chances that the estimate of the policy's improvement passes a significance test during downstream policy evaluation. We mathematically characterize the Pareto frontier of policies according to the tradeoff of these two goals. Based on our characterization, we design a policy optimization algorithm that estimates the Pareto frontier using machine learning models; then, the decision-maker can select the policy that optimizes their desired tradeoff, after which policy evaluation can be performed on the test set as usual.

Towards optimal inventory control under censored demand: a biased SAA approach

Yuxuan Han, New York University

Yuxuan Han (New York University); Xiaoyu Fan (New York University); Jiawei Zhang(New York University); Zhengyuan Zhou (New York University)

We study data-driven multi-period lost-sales inventory control under censored demand, where a stockout reveals only that demand exceeded the stocking level. We develop a unified, model-based framework for policy learning from censored data, built on a new cost decomposition for base-stock policies and a biased sample-average approximation (SAA) approach. The cost decomposition allows us to characterize the coverage condition under which censored observations are informative enough for sample-efficient policy learning. Guided by this characterization, we design two biased SAA algorithms: an upper-biased one that achieves near-optimal sample complexity under the offline coverage condition, and a lower-biased one that actively generates the required coverage and achieves near-optimal regret online. More broadly, this biased SAA approach provides a general principle for implementing pessimism and optimism under censored feedback, which may be of independent interest.

Choosing the better bandit algorithm under data sharing: when do A/B experiments work?

Shuangning Li, University of Chicago

Shuangning Li (UChicago); Chonghuan Wang (TTIC); Jingyan Wang (UT Dallas)

We study A/B experiments that are designed to compare the performance of two recommendation algorithms. Prior work has observed that the stable unit treatment value assumption (SUTVA) often does not hold in large-scale recommendation systems, and hence the estimate for the

global treatment effect (GTE) is biased. Specifically, units under the treatment and control algorithms contribute to a shared pool of data that subsequently train both algorithms, resulting in interference between the two groups. In this paper, we investigate when such interference may affect our decision making on which algorithm is better. We formalize this insight under a multi-armed bandit framework and theoretically characterize when the sign of the difference-in-means estimator of the GTE under data sharing aligns with or contradicts the sign of the true GTE. Our analysis identifies the level of exploration versus exploitation as a key determinant of how data sharing impacts decision making, and we propose a detection procedure based on ramp-up experiments to signal incorrect algorithm comparison in practice.

Optimal regret–inference tradeoffs in bandits: a geometric characterization

Wanteng Ma, University of Pennsylvania

Wanteng Ma (UPenn); T. Tony Cai (UPenn); Yuting Wei (UPenn)

In online decision making problems, regret minimization and uncertainty quantification place competing demands on the same policy: low regret forces sampling to concentrate on near-optimal actions, while short confidence intervals require information along inferential directions that may incur large regret. Despite growing evidence of this tension, it has not been formally understood. This paper provides the first unified characterization of the tradeoff between regret control and statistical inference under the general stochastic bandit framework. We first establish a geometric lower bound showing that, for any nonanticipating policy and any valid confidence interval for a linear functional of the reward function, the product between expected regret and expected confidence interval length is bounded below by a *regret-priced separation coefficient*, determined by the local geometry of a family of perturbations at a base instance; the coefficient makes explicit the price of information, namely the rate at which regret is converted into information about the target, that governs the lower bound. Instantiating this mechanism yields sharp tradeoffs across a wide range of models, including two-armed and K-armed bandits, fixed-action evaluation, linear bandits, with its generalized version applying to nonparametric bandits with Hölder-smooth rewards, in each case with matching upper bounds attained by algorithms that track the whole optimal frontier.

Parallel Sessions 5: Monday, November 2, 2026, 1:30 pm–3:00 pm

Session 96: When Does Explainable AI Actually Open the Black Box

Location: Salon C

Organizer: Srikar Katta (Duke University)

Chair: Lesia Semenova (Rutgers University)

Presenters: Foster Provost; Zachary T. Rewolinski; Giles Hooker; Jon Donnelly

Prompt-counterfactual explanations for generative AI behavior

Foster Provost, NYU Stern School of Business

Sofie Goethals (University of Antwerp); Foster Provost (New York University); João Sedoc (New York University)

As applications integrate generative AI, organizations increasingly need to be able to understand and interpret AI system behavior. I'll examine a key question: what is it about the input – the prompt – that causes an LLM-based generative AI system to produce output that exhibits specific characteristics, such as toxicity, negative sentiment, or political bias. I adapt a common technique from Explainable AI: counterfactual explanations. I'll explain why traditional counterfactual explanations cannot be applied directly to generative AI systems, due to differences in how generative AI systems function. I'll then propose a flexible framework that adapts counterfactual explanations to non-deterministic, generative AI systems in scenarios where downstream classifiers can reveal key characteristics of their outputs. Finally, I'll demonstrate the production of counterfactual explanations for generative AI systems with three case studies, examining different output characteristics. The case studies show that these Prompt-Counterfactual Explanations can streamline prompt engineering to suppress undesirable output characteristics and can enhance red-teaming efforts to uncover additional prompts that elicit undesirable outputs. This work lays a foundation for prompt-focused interpretability in generative AI, critical as generative models are entrusted with higher-stakes tasks and subject to emerging regulatory requirements for transparency and accountability.

Local MDI+: local feature importances for tree-based models

Zachary T. Rewolinski, University of California, Berkeley

Zhongyuan Liang (University of California, Berkeley); Zachary T. Rewolinski* (University of California, Berkeley); Abhineet Agarwal (University of California, Berkeley); Tiffany M. Tang (University of Notre Dame); Bin Yu (University of California, Berkeley)*

Tree-based ensembles such as random forests remain the go-to for tabular data over deep learning models due to their prediction performance and computational efficiency. These advantages have led to their widespread deployment

in high-stakes domains, where interpretability is essential for ensuring trustworthy predictions. This has motivated the development of popular local (i.e. sample-specific) feature importance (LFI) methods such as LIME and TreeSHAP. However, these approaches rely on approximations that ignore the model's internal structure and instead depend on potentially unstable perturbations. These issues are addressed in the global setting by MDI+, a global feature importance method which combines tree-based and linear feature importances by exploiting an equivalence between decision trees and least squares on a transformed node basis. However, the global MDI+ scores are not able to explain predictions when faced with heterogeneous individual characteristics. To address this gap, we propose Local MDI+ (LMDI+), a novel extension of the MDI+ framework that quantifies feature importances for each particular sample. Across twelve real-world benchmark datasets, LMDI+ outperforms existing baselines at identifying instance-specific predictive features, yielding an average 10% improvement in predictive performance when using only the selected features. It further demonstrates greater stability by consistently producing similar instance-level feature importance rankings across repeated model fits with different random seeds. Ablation experiments show that each component of LMDI+ contributes to these gains, and that the improvements extend beyond random forests to gradient boosting models. Finally, we show that LMDI+ enables local interpretability use cases by identifying closely matched counterfactuals for each classification benchmark and discovering homogeneous subgroups in a case study using a commonly-used housing dataset.

Inference for regulatory assessments of machine learning

Giles Hooker, University of Pennsylvania

Giles Hooker (University of Pennsylvania); Fei Huang (University of New South Wales)

The use of machine learning in high-stakes decisions about people has lead to the development of a literature around fairness assessments for these models. However, specific government regulation has been somewhat slower to develop. Using retail insurance as a case study, this talk surveys the current landscape of regulation along with methodological proposals for conducting audits of insurance products. We examine challenges in translating current interpretability metrics into inference about fairness, including translating current procedures into fairness metrics, specifying appropriate uncertainty quantification, and accounting for imputing protected attributes. We propose an audit framework and procedures and list a set of statistical challenges needed to develop a thorough assessment process.

Towards reliable model reliance: local conditional permutation importance via conditional density estimation

Jon Donnelly, Johns Hopkins University

Yixiao Wang (Duke University); Srikar Katta (Duke University); Jon Donnelly (Johns Hopkins University); Emanuele Borgonovo (Bocconi University); Cynthia Rudin (Duke University)

Variable importance (VI) is critical for interpreting feature usage in machine learning models. In general, VI metrics measure how much a black-box model's output changes when you replace the values of specific features of interest with randomized values through perturbation. However, common VI methods can misrepresent this reliance in two ways: (i) by averaging importance across observations into a single global value, which can wash out local variation in variable importance; and (ii) by ignoring conditional dependence between features, which can create unrealistic counterfactual samples, thereby evaluating the model in regions unsupported by the observed data distribution. We address both challenges by presenting a new model-agnostic VI metric and estimator, Local Conditional Variable Importance (LCVI). LCVI measures feature importance for individual observations by perturbing the features of interest with plausible values drawn from their conditional distribution given the remaining observed features. We theoretically establish consistency and a convergence rate, derive valid posterior credible regions, and reveal deep connections to classical VI metrics. Simulated and real-world examples show that LCVI captures local heterogeneity, identifies interaction effects, avoids importance leakage across inactive regions.

Session 38: Statistical Methods for AI: Attribution, Alignment, Representation Learning, and Generative Evaluation

Location: Williamsburg

Organizer: Yuan Zhang (Ohio State University)

Chair: Yuan Zhang (Ohio State University)

Presenters: Siqi (Cindy) Zeng; Zhehang Du; Weijing Tang; Yuliang Xu

Convex dataset selection for post-training via portfolio optimization

Siqi (Cindy) Zeng, University of Illinois Urbana-Champaign

S. Zeng (UIUC), C. Jung (Meta), R. Li (Meta), Z. Kang (Meta), M. Li (Meta), N. Noorshams (Meta), Z. Wang (Meta), F. Peng (Meta), H. Zhao (UIUC), X. Feng (Meta)

Improving LLM performance on downstream tasks sometimes requires leveraging auxiliary datasets during post-training. In practice, however, developers face constraints on compute, labeling, and licensing costs that preclude using all available data, necessitating principled dataset-level selection. These constraints are increasingly shaped by dataset marketplaces, where data acquisition is governed by budgets and negotiation. We study dataset valuation as a subset selection problem during LLM post-training. Our goal is to identify and weight auxiliary datasets so as to maximize target task performance given

constrained budgets. We first show that commonly used gradient alignment scores provide a reasonable yet incomplete valuation signal, as they ignore redundancy among datasets. To address this, we propose a scalable convex dataset-level valuation method based on kernel mean matching (KMM) in gradient space, which jointly accounts for alignment with the target task and redundancy across auxiliary datasets. Through extensive experiments across diverse post-training settings and tasks, we show that our approach consistently outperforms existing valuation baselines, achieving stronger performance with low computational overhead. Our results position dataset valuation as a practical decision tool for post-training data selection in market-constrained large language model settings.

The Newton-Muon optimizer: Muon is secretly a Newton method

Zhehang Du, University of Pennsylvania

Zhehang Du (University of Pennsylvania); Weijie Su (University of Pennsylvania)

The Muon optimizer has received considerable attention for its strong performance in training large language models, yet the theoretical mechanism underlying its matrix-gradient orthogonalization remains mysterious. In this paper, we introduce a surrogate model that not only reveals an implicit design principle behind Muon, but more importantly leads to a new optimizer. In the same spirit as the derivation of Newton's method, the surrogate approximates the loss as a quadratic function of the perturbation to a weight matrix W using only three matrices: the gradient G, an output-space curvature matrix H, and the data matrix Z that stacks the layer inputs. By minimizing this surrogate in one step and adopting a certain isotropic assumption on the weights, we obtain the closed-form update rule (up to momentum and weight decay) $W \leftarrow W - \eta \cdot \mathrm{polar}(G(ZZ^\top)^{-1})$, where η is the learning rate and $\mathrm{polar}(X) = UV^\top$ if $X = USV^\top$ is a compact singular value decomposition. This new optimization method, which we refer to as Newton–Muon, shows that standard Muon can be interpreted as an implicit Newton-type method that neglects the right preconditioning induced by the input second moment. Empirically, on a reproduction of the earliest publicly released Modded-NanoGPT speedrun configuration using Muon for GPT-2 pretraining, Newton–Muon reaches the target validation loss in 6% fewer iteration steps and reduces wall-clock training time by about 4%.

Towards scalable game-theory-based attribution

Weijing Tang, Carnegie Mellon University

Ziqi Liu (Carnegie Mellon University), Kiljae Lee (The Ohio State University), Yuan Zhang (The Ohio State University), Weijing Tang (Carnegie Mellon University)

Shapley values and related probabilistic values provide a flexible model-agnostic framework for attributing the behavior of a black-box system to data points, features, or other components, with a wide range of applications from model explanation and data valuation to credit assignment and compensation. Their practical use, however, is limited by a fundamental computational challenge: exact evaluation requires considering exponentially many coalitions, while even a single coalition evaluation may involve expensive model training or inference. In this talk, I will present a unified first-order perspective on Monte Carlo estimation of probabilistic values. Although existing estimators are motivated by seemingly different strategies, many share a common error structure that reveals how coalition sampling and surrogate modeling jointly determine statistical efficiency. This perspective leads to an Efficiency-Aware Surrogate-adjusted Estimator (EASE) that directly minimizes an explicit first-order approximation to the mean squared error. We demonstrate that EASE consistently improves estimation accuracy compared to existing estimators under the same utility-evaluation budget.

Distributional evaluation of generative models via relative density ratio

Yuliang Xu, Johns Hopkins University

Yuliang Xu (Johns Hopkins University); Yun Wei (University of Texas Dallas); Li Ma (University of Chicago)

We propose a function-valued evaluation metric for generative models based on the relative density ratio (RDR) designed to characterize distributional differences between real and generated samples. As an evaluation metric, the RDR function preserves ϕ-divergence between two distributions, enables sample-level evaluation that facilitates downstream investigations of feature-specific distributional differences, and has a bounded range that affords clear interpretability and numerical stability. Function estimation of the RDR is achieved efficiently through optimization on the variational form of ϕ-divergence. We provide theoretical convergence rate guarantees for general estimators based on M-estimator theory, as well as the convergence rate of neural network-based estimators when the true ratio is in the anisotropic Besov space. We demonstrate the power of the proposed RDR-based evaluation through numerical experiments on MNIST, CelebA64, and the American Gut project microbiome data. We show that the estimated RDR enables not only effective overall comparison of competing generative models, but also a convenient way to reveal the underlying nature of goodness-of-fit. This enables one to assess support overlap, coverage, and fidelity while pinpointing regions of the sample space where generators concentrate and revealing the features that drive the most salient distributional differences.

Session 16: Causality and AI in Science

Location: Fulton Ferry

Organizer: Zhonghua Liu (Columbia University)

Chair: Zhonghua Liu (Columbia University)

Presenters: Xihong Lin; Nilanjan Chatterjee; Kathryn Roeder

Title TBA

Xihong Lin, Harvard University

Abstract TBA

Improving Mendelian randomization analysis by instrument borrowing from auxiliary outcome traits

Nilanjan Chatterjee, Johns Hopkins University

Anagh Chattopadhyay (Johns Hopkins University), Nilanjan Chatterjee (Johns Hopkins University)

Mendelian randomization (MR) uses genetic variants as instruments to infer causal effects, but remains vulnerable to bias and power loss when instruments are invalid because of pleiotropy. We hypothesize that closely related outcome traits share many of their valid and invalid instruments for a given exposure, so that borrowing instruments across such traits can make MR more robust. We introduce a coheterogeneity statistic that measures the alignment of pleiotropic deviations across two outcomes, and establish its consistency and asymptotic normality when the number of instruments grows with the GWAS sample size, yielding valid confidence intervals for selecting auxiliary traits. We then extend MR-Mode and MR-PRESSO to IB-Mode and IB-PRESSO, which analyze a primary and an auxiliary outcome jointly. In simulations, instrument borrowing increased power and reduced type I error inflation, with gains growing with the overlap in valid instruments. Across 46 positive- and negative-control hypotheses using Million Veteran Program data, IB-Mode best controlled false positives while retaining power, and found no robust evidence that vitamin D affects cardiometabolic traits

From prediction to inference: debiasing and confounder control in spatial transcriptomics

Kathryn Roeder, Carnegie Mellon University

Lorenzo Testa (Carnegie Mellon University); Jing Lei (Carnegie Mellon University); Kathryn Roeder (Carnegie Mellon University)

Spatial transcriptomics has transformed the study of tissue biology by measuring gene expression directly within intact tissues, preserving the spatial organization of cells and their local microenvironments. But current technologies trade off spatial resolution against transcriptome coverage. Deep learning methods increasingly close this gap by reconstructing unmeasured genes from matched single-cell RNA-sequencing references, extending analysis to the full transcriptome. Yet reconstructed expression cannot be reliably compared across conditions: existing analyses treat predictions as if directly measured and ignore latent spatial variation from tissue architecture and cellular organization, inflating false discoveries. We introduce TIDEST, a framework for differential expression analysis on reconstructed spatial transcriptomes. Our method calibrates reconstructed expression using the measured genes and adjusts for latent spatial variation before comparing groups, providing a statistically principled basis for discovery. Across realistic tissue simulations, TIDEST controls false discoveries where existing methods fail, while preserving power. Applied to mouse brain, glioblastoma and breast cancer tissues, it recovers biologically meaningful differential expression – established cortical-layer, tumor-state and invasion markers – missed or mischaracterized by conventional analyses. Overall, TIDEST turns increasingly accurate reconstructions into trustworthy biological discovery.

Session 26: Recent Advances in Statistical Network Modeling and Inference

Location: Sunset Park

Organizer: Tianxi Li (University of Minnesota)

Chair: Tianxi Li (University of Minnesota)

Presenters: Min Xu; Michael Schweinberger; Fangzheng Xie; Chen-Wei Hua

Statistical problems on randomly growing networks

Min Xu, Rutgers University

Min Xu (Rutgers University); Jianxiang Wang (Rutgers University); Harry Crane (Rutgers University)

Many real-world networks arise from a growth process in which vertices and edges are added sequentially. The preferential attachment (PA) model captures this process by connecting each new node to a random subset of existing nodes, with probabilities that depend on their degrees. PA networks exhibit high edge sparsity, heavy-tailed degree distributions, and hub and chain subgraphs, which are features that static models such as the stochastic block model struggle to reproduce.

Supposing we observe only a single snapshot of the final network, we consider two statistical problems. The first problem is estimating the early growth history of the network–in particular the root nodes–which yields a statistically principled method to assess node importance, with strong theoretical guarantees. The second problem is estimating the community structure in cases where multiple communities are formed from independent growth processes. This leads to a new approach to community recovery based on identifying and clustering anchor nodes.

Causal inference in connected populations with contagion

Michael Schweinberger, Department of Statistics, The Pennsylvania State University

Subhankar Bhadra (Pennsylvania State University); Michael

Schweinberger (Pennsylvania State University)

Causal inference in connected populations is complicated by contagion and other real-world processes inducing dependence among outcomes. We address a gap in the literature on causal inference under contagion: while there is a growing body of work on estimating causal effects under contagion, little is known about how contagion impacts causal effects and inference. We provide insight into how contagion impacts causal effects and inference based on closed-form expressions for causal effects under contagion. These closed-form expressions reveal that the effects of interventions, spillover, and contagion are intertwined even in the simplest possible settings, and that contagion can decrease or increase causal effects. We discuss statistical implications, including asymptotic bias of model-based estimators ignoring dependence among outcomes due to contagion, violations of neighborhood exposure assumptions underlying design-based estimators of the total causal effect of interventions under unrestricted contagion, and possible remedies.

Uncertainty quantification for mixed membership in multilayer networks with degree heterogeneity using Gaussian variational inference

Fangzheng Xie, Indiana University

Fangzheng Xie (Indiana University); Hsin-Hsiung Huang (University of Central Florida)

Analyzing multilayer networks is central to understanding complex relational measurements collected across multiple conditions or over time. A pivotal task in this setting is to quantify uncertainty in community structure while appropriately pooling information across layers and accommodating layer-specific heterogeneity. Building on the multilayer degree-corrected mixed-membership (ML-DCMM) model, which captures both stable community membership profiles and layer-specific vertex activity levels, we propose a Bayesian inference framework based on a spectral-assisted likelihood. We then develop a computationally efficient Gaussian variational inference algorithm implemented via stochastic gradient descent. Our theoretical analysis establishes a variational Bernstein–von Mises theorem, which provides a frequentist guarantee for using the variational posterior to construct confidence sets for mixed memberships. We demonstrate the utility of the method on a U.S. airport longitudinal network, where the procedure yields robust estimates, natural uncertainty quantification, and competitive performance relative to state-of-the-art methods.

Conformal inference for multiple source identification in network diffusion

Chen-Wei Hua, University of Minnesota

Chen-Wei Hua(University of Minnesota); Xuan Bi(University of Minnesota); Tianxi Li(University of Minnesota)

Diffusion source identification (DSI) is crucial for localizing the origins of spreading processes in networks, with broad applications in infectious disease control, rumor detec- tion and innovation stimulation. Given a network and an observed snapshot of binary infection statuses, the objective is to identify the nodes that most likely initiated the diffusion. While recent deep-learning-based approaches offer promising methods of es- timating node-wise infection probabilities, they often fail to provide valid uncertainty quantification or guaranteed accuracy. Specifically, existing inference methods are pre- dominantly tailored for single-source scenarios, leaving multi-source identification as a salient open challenge. In this work, we propose a principled inference framework that bridges predictive mod- eling with conformal prediction. We first employ graph neural networks and alternative predictive models to estimate node-wise source probabilities from infection snapshots. We then apply conformal prediction as a downstream procedure to convert these scores into statistically valid confidence sets of candidate source nodes. This approach provides a unified framework that naturally scales to multi-source diffusion settings. Experiments on benchmark networks demonstrate that our framework substantially outperforms pre- vious methods in single-source settings, while offering a more robust path for multi-source identification.

Session 71: Advancement in Nonparametric Method in Complex Data

Location: Golden Boardroom

Organizer: Wen Zhou (New York University)

Chair: Xiwei Tang (UT Dallas)

Presenters: Xiwei Tang; Jiayu Li; Fei Xue; Boxiang Wang

Low-rank pairwise ranking with sequence-induced dependence and graph sparsification

Xiwei Tang, University of Texas at Dallas

Yizhe He (UT Dallas); Haowen Zhou (U of Virginia); Jiayi Wang (UTDallas)

Preference learning is a core supervision paradigm in modern AI systems, from1 recommendation and retrieval to human-feedback-driven ranking. In practice, pair-2 wise labels are often obtained by projecting a user's ranking over an exposed item3 set into pairwise comparisons. Such labels are dependent, and the all-pairs projec-4 tion produces a quadratic number of comparisons whose statistical value may be5 limited by the shared underlying ranking. We study this phenomenon in a low-rank6 user–item matrix factorization model using a pairwise logistic pseudo-likelihood.7 We first establish a non-asymptotic recovery guarantee under independent pair8 sampling with heterogeneous observation probabilities, providing a benchmark for9 pairwise low-rank preference learning. We then analyze sequence-based feedback10 generated from exposed item sets under a Plackett–Luce model, quantifying how11 sequence-induced dependence affects recovery when all induced comparisons are12 used. Motivated by this

analysis, we propose two linear-size comparison sparsifi-13 cation schemes. A user-stratified matching scheme retains disjoint comparisons14 within each exposed set and admits finite-sample analysis: it reduces the number of15 comparisons from quadratic to linear in the exposure-set size while preserving the16 leading statistical error rate of the full-pair estimator. A random-order chain retains17 adjacent comparisons along an auxiliary item order sampled independently of the18 observed ranking and admits the same leading-order finite-sample recovery bound.19 A separate local equal-utility Godambe-information analysis compares matching,20 chain, star, and all-pairs beyond their upper-bound orders. Expanded synthetic,21 MovieLens, and Jester experiments show metric- and budget-dependent tradeoffs:22 chain is strongest under the fixed synthetic comparison-update budget, whereas23 all-pairs is strongest after longer optimization, and sparse designs use substantially24 fewer retained comparisons.

Dual-channel tensor neural networks: finite-sample theory and conformal structure selection

Jiayu Li, New York University

Elynn Chen (New York University); Jiayu Li (New York University); Jian Pei (Duke University); Zheshi Zheng (University of Michigan)

Tensor-valued data arise naturally in neuroimaging, genomics, climate science, and spatiotemporal networks. Existing approaches either impose a single low-rank structure, which can miss localized signal, or treat the tensor as a long vector, which discards its multiway geometry. We propose a Dual-Channel Tensor Neural Network (DC-TNN) that decomposes each tensor input into a low-rank core and a sparse refinement, processed through coupled neural channels. The framework is structure-agnostic, accommodating CP, Tucker, and tensor-train cores within a single architecture. We establish non-asymptotic risk bounds for the DC-TNN estimator that decompose into network approximation, core estimation, and refinement-selection terms, and show that the effective dimension is determined by the core rank and refinement sparsity rather than the ambient tensor size. For inference, we develop a structure-aware conformal ROC procedure that calibrates within the core-refinement latent space and produces ROC and AUC confidence bands with finite-sample, distribution-free coverage. Building on this, we propose a conformal structure selector, to our knowledge the first distribution-free procedure for choosing among candidate tensor decompositions with finite-sample validity. Simulations and a protein data analysis demonstrate competitive predictive accuracy, reliable uncertainty quantification, and consistent recovery of the tensor structure.

Federated learning for feature mismatch

Fei Xue, Purdue University

Huayu Tang (Purdue University), Yuying Lu (Columbia University), Rui Duan (Harvard University)

We consider a framework that involves multiple sites with privacy constraints and heterogeneity, where different sites collect data from independent subjects, and site-specific, mismatched feature sets (e.g., hospitals with differing equipment and workflows). Most existing methods assume either aligned covariates or overlapping cohorts, and thus are inapplicable when both samples are disjoint and the covariates are mismatched. We propose Fed-PATE, a three-stage framework: local estimation at each site using its own observed features; a cross-site projection aggregation that recovers a consensus full-feature coefficients among similar sites; and a penalized transfer estimator that feeds the aggregated information back to each site. To accommodate heterogeneity and prevent negative transfer, we introduce a weighting scheme for model selection that adapts to inter-site similarity. Our proposed approach shows a lower prediction risk in theoretical analyses. Simulations show significant reductions in prediction MSE and robustness to inter-site heterogeneity. In a real-world application to Serum Creatinine (SCr) prediction using EHR variables, Fed-PATE improves performance over existing approaches for the majority of sites.

Quantile-based discriminant analysis for high-dimensional imbalanced classification

Boxiang Wang, University of Iowa

Qian Tang, Yuwen Gu, Boxiang Wang

I will present the quantile-based discriminant analysis (QuanDA), a new method for addressing class imbalance in high-dimensional classification problems. Inspired by the least squares view of linear discriminant analysis, QuanDA employs the quantile regression, enhanced with elastic-net regularization, for feature selection and classification. Unlike traditional approaches to discriminant analysis, such as the linear, quadratic, and Fisher discriminant analyses, QuanDA does not rely on heavy model assumptions, thus significantly increasing the robustness of the classification. I will present comprehensive simulation studies and benchmark tests to demonstrate that QuanDA achieves competitive performance relative to other leading classification methods that deal with class imbalance.

Session 08: Conformal Prediction with Partially Observed Data

Location: Carroll Gardens

Organizer: Matteo Sesia (University of Southern California)

Chair: Vladimir Svetnik (Merck & Co.)

Presenters: Rebecca Farina; Max Sampson; Zhimei Ren; Matteo Sesia

Doubly robust prediction sets for right-censored time-to-event outcomes

Rebecca Farina, Carnegie Mellon University

Rebecca Farina (Carnegie Mellon University), Eric J. Tchetgen

Tchetgen (University of Pennsylvania), Arun Kumar Kuchibhotla (Carnegie Mellon University)

Our objective is to construct well-calibrated prediction sets for a time-to-event outcome subject to right-censoring with guaranteed coverage. Inspired by modern conformal inference, our approach avoids the need for a well-specified parametric or semiparametric survival model. Unlike existing conformal methods for survival data, which assume Type-I censoring with fully observed censoring times, we consider the more common right-censoring setting in which only the censoring time or only the event time is observed, whichever comes first. Under a standard conditional independence censoring condition, we propose and analyze several lower prediction bounds for the survival time of a future observation, including inverse-probability-of-censoring weighting, and its augmented version based on the semiparametric efficient influence function for the relevant marginal quantile of the outcome accounting for dependent censoring. We formally establish asymptotic coverage guarantees of the proposed methods, and demonstrate both theoretically and through empirical experiments, that the augmented approach substantially improves efficiency over all other proposed methods. Specifically, its coverage error bound is doubly robust, and therefore of second order, thus ensuring that it is asymptotically negligible relative to the coverage error of the other methods.

Proportional-integral conformal nowcasting for discrete-time state-space models

Max Sampson, Florida State University

Max Sampson (Florida State University); Kung-Sik Chan (University of Iowa)

We introduce Proportional-Integral Conformal Nowcasting (PI-CN) to construct prediction intervals for discrete-time linear state-space models. Nowcasting involves forecasting a target time-series when features arrive asynchronously, creating a “ragged edge" of data availability. To guarantee simultaneous coverage across this ragged edge, we define the conformal score as the maximum absolute standardized prediction error across all intra-period updates. The conformal multiplier is computed via a proportional-integral controller that sequentially tracks empirical coverage, ensuring valid, long-run $1-\alpha$ coverage regardless of model specification.

We target simultaneous rather than marginal coverage along the ragged edge because sequential intra-period forecasting errors in fixed-event nowcasting are structurally positively correlated, sharing the same future unobserved innovations. We establish that PI-CN asymptotically achieves nominal $1-\alpha$ simultaneous coverage under mild regularity conditions. Furthermore, we prove that if the true DGP is a consistently estimated discrete-time linear state-space process, intra-period prediction errors are strictly positively correlated. Specifically, when incoming orthogonal shocks are small relative to the prediction error, the joint coverage probability converges to the first intra-period step’s coverage. Empirical applications confirm our theoretical derivations, demonstrating highly correlated intra-period errors and valid $1-\alpha$ simultaneous long-run coverage.

ACS: an interactive framework for conformal selection

Zhimei Ren, University of Pennsylvania

Yu Gui (University of Pennsylvania), Ying Jin (University of Pennsylvania), Yash Nair (Stanford University), Zhimei Ren (University of Pennsylvania)

This paper presents adaptive conformal selection (ACS), an interactive framework for model-free selection with guaranteed error control. Building on conformal selection (Jin and Candès, 2023b), ACS generalizes the approach to support human-in-the-loop adaptive data analysis. Under the ACS framework, we can partially reuse the data to boost the selection power, make decisions on the fly while exploring the data, and incorporate new information or preferences as they arise. The key to ACS is a carefully designed principle that controls the information available for decision making, allowing the data analyst to explore the data adaptively while maintaining rigorous control of the false discovery rate (FDR). Based on the ACS framework, we provide concrete selection algorithms for various goals, including model update/selection, diversified selection, and incorporating newly available labeled data. The effectiveness of ACS is demonstrated through extensive numerical simulations and real-data applications in large language model (LLM) deployment and drug discovery.

Doubly robust conformalized survival analysis with right-censored data

Matteo Sesia, University of Southern California

Matteo Sesia; Vladimir Svetnik

We present a conformal inference method for constructing lower prediction bounds for survival times from right-censored data, extending recent approaches designed for more restrictive type-I censoring scenarios. The proposed method imputes unobserved censoring times using a machine learning model, and then analyzes the imputed data using a survival model calibrated via weighted conformal inference. This approach is theoretically supported by an asymptotic double robustness property. Empirical studies on simulated and real data demonstrate that our method leads to relatively informative predictive inferences and is especially robust in challenging settings where the survival model may be inaccurate.

Session 49: Advances in Statistical Inference for Complex Data

Location: Coney Island

Organizer: Yufeng Liu (University of Michigan)

Chair: Yufeng Liu (University of Michigan)

Presenters: Gongjun Xu; Quefeng Li; Minjie Wang; Hang Zhou

Beyond vintage rotation: bias-free sparse representation learning with oracle inference

Gongjun Xu, University of Michigan

Chengyu Cui (University of Michigan); Yunxiao Chen (London School of Economics and Political Science); Jing Ouyang (University of Hong Kong); Gongjun Xu (University of Michigan)

Learning low-dimensional latent representations is a central topic in statistics and machine learning, and rotation methods have long been used to obtain sparse and interpretable representations. Despite nearly a century of widespread use across many fields, rigorous guarantees for valid inference for the learned representation remain lacking. In this paper, we identify a surprisingly prevalent phenomenon that suggests a reason for this gap: for a broad class of vintage rotations, the resulting estimators exhibit a non-estimable bias. Because this bias is independent of the data, it fundamentally precludes the development of valid inferential procedures, including the construction of confidence intervals and hypothesis testing. To address this challenge, we propose a novel bias-free rotation method within a general representation learning framework based on latent variables. We establish an oracle inference property for the learned sparse representations: the estimators achieve the same asymptotic variance as in the ideal setting where the latent variables are observed. To bridge the gap between theory and computation, we develop an efficient computational framework and prove that its output estimators retain the same oracle property. Our results provide a rigorous inference procedure for the rotated estimators, yielding statistically valid and interpretable representation learning.

Linear hypothesis testing for high-dimensional interactive fixed effects models

Quefeng Li, University of North Carolina at Chapel Hill

Zhe Zhang (UNC-Chapel Hill); Quefeng Li (UNC-Chapel Hill)

Panel data in economics and biomedical research exhibit variation across both subjects and time, which interactive fixed effects models capture through latent factors with subject-specific loadings. Two questions remain open: how to test general linear hypotheses when the number of covariates grows exponentially with the sample size, and whether modeling the latent factors yields genuine gains for inference. For the first, we propose an estimator that penalizes all coefficients except those under test and alternates between estimating the factors and the coefficients. Leaving the tested coefficients unpenalized yields a normal limiting distribution directly, requiring neither a debiasing step nor inversion of a large matrix, and the resulting Wald test accommodates arbitrary linear restrictions. Our main contribution addresses the second question. Practitioners have long observed that interactive fixed effects models deliver sharper inference than models absorbing the factors into the error term, but no theory has explained this phenomenon. We prove that our test is asymptotically more efficient and identify the underlying mechanism: projecting the covariates onto the estimated factors incurs negligible information loss, whereas removing factor variance from the errors substantially reduces noise, so the gain is real and the cost of estimating the factors is asymptotically negligible. This analysis also clarifies that the advantage is largest when the factors account for a large share of total variation. Simulations confirm a substantial power gap that widens as factor influence increases, and in a Eurostat producer-price panel the factor-adjusted test rejects a block of production terms undetected by the factor-free test.

High-dimensional variable selection with diffusion-generated synthetic data

Minjie Wang, Binghamton University

Minjie Wang (Binghamton University); Xiaotong Shen (University of Minnesota); Wei Pan (University of Minnesota)

Recent advances in generative models and generative artificial intelligence (AI) have prompted the advent of synthetic data generation, which challenges the traditional statistical practice of focusing solely on the raw data and shifts toward a more synthetic data-centric approach. On the other hand, variable selection for high-dimensional, highly correlated data has long been a challenging problem, often yielding unstable and unreliable models. To enhance variable selection, we present a resample-aggregate framework that exploits diffusion models' ability to generate high-fidelity synthetic data. Specifically, we draw multiple pseudo-data sets from a diffusion model fitted to the original data, apply any off-the-shelf selector (e.g., lasso or SCAD), and store the resulting inclusion indicators and coefficients. Aggregating across replicas produces a stable subset of predictors with calibrated stability scores for variable selection. Theoretically, we show that the proposed method is selection consistent under mild assumptions. Extensive simulations show consistent gains over the lasso, stability selection, and knockoff baselines, especially when predictors are strongly correlated, achieving higher true-positive rates and lower false-discovery proportions. By coupling diffusion-based data augmentation with principled aggregation, our method advances variable selection methodology and broadens the toolkit for interpretable, statistically rigorous analysis in complex scientific applications.

Statistical inference for random objects

Hang Zhou, University of North Carolina at Chapel Hill

Hang Zhou (University of North Carolina at Chapel Hill); Hans-Georg

Müller (University of California, Davis)

Random objects are complex data types that take values in general metric spaces. Although such data are increasingly common in scientific research, the corresponding statistical methodology and theory remain relatively underdeveloped. A key challenge in their analysis is the absence of vector space operations, such as addition, subtraction, scalar multiplication, and inner products, that underpin classical statistical methods. In this talk, we introduce distance profiles as a useful tool for analyzing random objects. We illustrate their utility through two applications, predictive inference and independence testing, demonstrating how they enable principled inference for complex data in non-Euclidean spaces.

Session 73: AI/ML for Biomedical Research

Location: Park Slope

Organizer: Muxuan Liang (The University of Texas MD Anderson Cancer Center)

Chair: Muxuan Liang (The University of Texas MD Anderson Cancer Center)

Presenters: Sohom Bhattacharya; Tian Gu; Jin Jin; Wodan Ling

Interpolation-based learning under data heterogeneity

Sohom Bhattacharya, University of Florida

Sohom Bhattacharya (University of Florida)

Min-norm interpolators naturally arise as implicitly regularized limits of modern neural networks and other widely used algorithms. While their out-of-distribution risk has recently been analyzed in settings where test samples are unavailable during training, many practical applications allow partial access to target data. The behavior of min-norm interpolation in such settings remains poorly understood. In this talk, I will present a sharp characterization of the risk of min-norm interpolators under both covariate and model shifts. To reach these conclusions, we develop novel anisotropic local laws, representing new advances in random matrix theory for heterogeneous data problems. I will also discuss applications of our results to the challenge of combining real data with synthetic data generated by AI models.

On the connections among three transfer learning paradigms

Tian Gu, Columbia University

Tian Gu (Columbia University); Sijia Li (University of California, Los Angeles); Rui Duan (Harvard University)

We examine the solution paths of three transfer learning estimators within the context of linear models. Our analysis identifies a general solution path that characterizes how transfer learning estimators interpolate between the target and source estimators. This involves a change of basis and entry-wise weighting functions, with existing transfer learning methods identified as special cases. The proposed framework reveals connections and equivalences among transfer learning approaches, offering valuable insights for designing future estimators with improved control over the learning process. It extends beyond traditional penalized regression and gradient descent techniques and holds the potential for generalization to nonlinear cases. Extensive simulations validate the theoretically derived solution paths, and their practical utility is demonstrated by the improvement in risk prediction for end-stage renal disease in Hispanic populations.

Building AI co-scientists for genetic risk research

Jin Jin, University of Pennsylvania Perelman School of Medicine

Zhaoqian Xue (Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine); Jiyao Zhang (Applied Math and Computational Science, University of Pennsylvania Perelman School of Medicine); Bingxin Zhao (Department of Statistics and Data Science, the Wharton School of the University of Pennsylvania); Jin Jin (Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine)

Polygenic risk score (PRS) research has become increasingly challenging as the number of available GWAS datasets, prediction methods, pre-trained models, and analytical choices continues to grow. Researchers must navigate thousands of existing models, determine their suitability for a target phenotype and population, and design rigorous analyses—tasks requiring substantial domain expertise and not reliably addressed by general-purpose large language models (LLMs). We develop a domain-specialized PRS Research Agent, an agentic AI system designed to serve as an AI co-scientist for polygenic risk prediction. The system integrates LLMs with curated genetic resources, domain-specific tools, expert-defined “PRS Skills,” task-specific workflows, and verification steps. Given a target phenotype, cohort, and research goal, the agent synthesizes the PRS landscape, recommend appropriate pre-trained models or initiate new model development, and guide model evaluation and downstream analyses. A central application is evidence-based PRS model recommendation: the agent evaluates phenotype alignment, ancestry and cohort relevance, predictive performance, methodological appropriateness, and potential information leakage. Across 46 diseases in the All of Us Research Program, the agent identified more powerful PRS models than a general-purpose LLM, with an average 30.9% increase in (AUC−0.5), demonstrating the potential of domain-specialized agentic AI for accessible and reproducible genetic research.

Three-dimensional protein structure provides fine-grained classification of viral lineages

Wodan Ling, Weill Cornell Medicine

Wodan Ling (Weill Cornell Medicine)

The study of the human virome is advancing rapidly, yet current classification approaches based on nucleotide

similarity are limited by high viral mutation rates and lack functional resolution. We propose a function-based framework that leverages predicted protein three-dimensional structures from AlphaFold to cluster viral proteins and lineages at finer resolution. Compared with existing virus catalogs, our method yields more robust and biologically informative classification of viral lineages.

Session 75: Causal AI and Representation Learning

Location: Windsor Terrace

Organizer: Bryon Aragam (University of Chicago)

Chair: Bryon Aragam (University of Chicago)

Presenters: Vasilis Syrgkanis; Yixin Wang; Yue Yu

Discovering hierarchical latent capabilities of language models via causal representation learning

Vasilis Syrgkanis, Stanford University

Jikai Jin (Stanford University); Vasilis Syrgkanis (Stanford University); Sham Kakade (Harvard University); Hanlin Zhang (Harvard University)

Faithful evaluation of language model capabilities is crucial for deriving actionable insights that can inform model development. However, rigorous causal evaluations in this domain face significant methodological challenges, including complex confounding effects and prohibitive computational costs associated with extensive retraining. To tackle these challenges, we propose a causal representation learning framework wherein observed benchmark performance is modeled as a linear transformation of a few latent capability factors. Crucially, these latent factors are identified as causally interrelated after appropriately controlling for the base model as a common confounder. Applying this approach to a comprehensive dataset encompassing over 1500 models evaluated across six benchmarks from the Open LLM Leaderboard, we identify a concise three-node linear causal structure that reliably explains the observed performance variations. Further interpretation of this causal structure provides substantial scientific insights beyond simple numerical rankings: specifically, we reveal a clear causal direction starting from general problem-solving capabilities, advancing through instruction-following proficiency, and culminating in mathematical reasoning ability. Our results underscore the essential role of carefully controlling base model variations during evaluation, a step critical to accurately uncovering the underlying causal relationships among latent model capabilities.

Discrete causal representation learning

Yixin Wang, University of Michigan

Wenjin Zhang, Yixin Wang, Yuqi Gu

Causal representation learning seeks to uncover causal relationships among high-level latent variables from low-level, entangled, and noisy observations. Existing approaches often either rely on deep neural networks, which lack interpretability and formal guarantees, or impose restrictive assumptions like linearity, continuous-only observations, and strong structural priors. These limitations particularly challenge applications with a large number of discrete latent variables and mixed-type observations. To address these challenges, we propose discrete causal representation learning (DCRL), a generative framework that models a directed acyclic graph among discrete latent variables, along with a sparse bipartite graph linking latent and observed layers. This design accommodates continuous, count, and binary responses through flexible measurement models while maintaining interpretability. Under mild conditions, we prove that both the bipartite measurement graph and the latent causal graph are identifiable from the observed data distribution alone. We further propose a three-stage estimate-resample-discovery pipeline: penalized estimation of the generative model parameters, resampling of latent configurations from the fitted model, and score-based causal discovery on the resampled latents. We establish the consistency of this procedure, ensuring reliable recovery of the latent causal structure. Empirical studies on educational assessment and synthetic image data demonstrate that DCRL recovers sparse and interpretable latent causal structures.

Attention-based pretraining for unsupervised amortized causal discovery

Yue Yu, Lehigh University

Naiyu Yin; Tian Gao; Yue Yu

Directed Acyclic Graphs (DAGs) are foundational to many areas of AI research, providing a formal language for causal reasoning and model interpretability. However, learning DAGs remains challenging, due to super-exponential computational cost and diminished accuracy in small sample regimes. To address these challenges, we introduce Attention-DAG (ADAG), a novel linear transformer architecture for unsupervised amortized DAG learning. Unlike traditional unsupervised DAG learning methods which recover the graph structure from data of each individual domain, ADAG leverages the knowledge from multiple domains. It provides a nonlinear mapping from observational data of each domain to both the corresponding graph structure and the underlying parameters of Structural Equation Models (SEMs). This enables efficient zero-shot inference of causal structures in new domains with unseen SEM parameters. Remarkably, we demonstrate that for the DAG learning problem, attention acts as a fixed-point iterative scheme, so the training on multiple domains effectively discovers an efficient solver for the constraint optimization problem. Extensive evaluations on synthetic and realistic benchmarks demonstrate that ADAG significantly outperforms existing baselines in accuracy and efficiency, particularly when data is scarce.

Session 61: AI/ML and Advanced Statistical Methods in Clinical Development

Location: DUMBO

Organizer: Yue Shentu (Merck & Co.)

Chair: Yue Shentu (Merck & Co)

Presenters: Eric (Pingye) Zhang; Thomas Jemielita; Vitaly Druker

A clinical trial design benchmark for AI agents

Eric (Pingye) Zhang, Eikon Therapeutics

Eric Zhang (Eikon Therapeutics)

As AI agents become increasingly capable of performing specialized scientific tasks, robust benchmarks are needed to systematically evaluate their capabilities and limitations. Clinical trial design is a critical responsibility of statisticians in the pharmaceutical industry, yet no standardized benchmark currently exists for evaluating AI agents on this task.

We are developing a benchmark that evaluates an agent's ability to reproduce existing clinical trial designs from information contained in study protocols and Statistical Analysis Plans (SAPs). The benchmark assesses whether an agent can accurately identify and reconstruct key elements of a trial's statistical design, providing a structured and reproducible framework for evaluating AI capabilities in clinical trial design.

This benchmark establishes a foundation for systematically assessing AI agents in a clinically relevant setting and can support future applications, including evaluating the impact of domain-specific agent skills on clinical trial design tasks.

Agentic AI for external control benchmarking

Thomas Jemielita, Merck & Co., Inc.

Thomas Jemielita

To better inform oncology clinical development and identify promising therapeutics, external data is often leveraged to evaluate whether the experimental treatment arm is superior to available standard-of-care. A typical workflow is as follows: (1) Identify benchmark data, (2) Extract data of interest, (3) Compare trial data vs external with adjustment for confounding. While traditionally each step is done separately and manually, advances in AI/LLMs offer the opportunity to automate related workflows through agentic type processes. This work examines advantages and disadvantages of using an AI based workflow, such as the accuracy of identification/extraction of benchmark data which is crucial for downstream analyses such as Go/No-Go decision making in clinical trials. For example, while such workflows may enhance efficiency and reproducibility, they may also introduce challenges such as consistency of benchmark data identification and/or extraction. These challenges can further compound when assessing Go/No-Go across multiple studies or compounds, requiring effective context management for agentic systems. Overall, while challenges remain, advances in AI can potentially automate analysis workflows for external control analyses, with the end goal of improving decision making in clinical development.

Predictive biomarker graphical methods integrated with Virtual Twins and synthetic controls incorporating multi-modal data

Vitaly Druker, AstraZeneca

Gina D'Angelo; Vitaly Druker (presenting); Xiaowen Tian

Biomarkers are central across the drug development lifecycle, from early discovery to late-phase trials. We present a hybrid framework that integrates the Predictive Biomarker Graphical Approach (PRIME) with machine learning–enabled virtual twins and synthetic control arms. PRIME is a counterfactual-oriented graphical method for identifying predictive biomarkers and optimizing cutoffs to define responsive subpopulations. We extend this framework to support both two-arm and single-arm designs by leveraging external data, including clinical datasets, quantitative continuous scoring (QCS) from AI-driven IHC imaging, and proteomics, to construct comparator arms. A lung cancer case study demonstrates how PRIME and virtual twin and synthetic-arm methods can inform enrichment strategies, quantify uncertainty, and generate interpretable evidence suitable for development decision-making and regulatory interactions.

Session 58: Advances in Methodology and Theory for Network Analysis

Location: Greenpoint

Organizer: Jingming Wang (University of Virginia)

Chair: Jingming Wang (University of Virginia)

Presenters: Huimin Cheng; David Hong; Sagnik Nandy; Haixiao Wang

Network perturbation aggregation for graphon estimation

Huimin Cheng, Boston University

Da Wang (Tsinghua University); Jun S. Liu (Tsinghua University); Yongkai Chen (Harvard University); Yuyao Wang (Boston University); Huimin Cheng (Boston University)

In recent years, various methods have been proposed to estimate the edge probability under the graphon model given a single observed network. Since the presence or absence of edges in the observed network is a stochastic realization from the underlying probability structure, estimating edge probabilities based on only a single observed realization is inherently limited and leads to estimates with high variance and high mean squared error (MSE). To address this issue, we propose the Network Perturbation aggregating (Net-Paging) method. The key idea is to construct multiple perturbed networks that preserve key graphon properties, mimicking multiple

replications. We then obtain graphon estimates from each perturbed network and average these estimates to obtain the final estimate. We theoretically show that the more perturbed samples used in our algorithm, the smaller the MSE, with a linear dependency. Extensive simulation experiments and real data analysis show that Net-Paging effectively reduces the variance and MSE compared to existing methods.

Selecting the embedding dimension with signflips

David Hong, University of Delaware

David Hong (University of Delaware); Joshua Cape (University of Wisconsin–Madison)

Selecting the embedding dimension is a crucial component of spectral analysis for networks. Using too few eigenvectors risks missing out on community structure in the data, whereas using too many eigenvectors risks making the embeddings noisier. This talk presents recent work on tackling this problem in a challenging setting of large heterogeneous networks that have weakly distinguishable community structure. In this setting, the "signal" eigenvalues do not diverge, so the cut-off must be carefully calibrated to select the eigenvalues whose magnitudes rise above the appropriate "noise floor". For this task, we introduce a novel spectral method that compares the eigenvalues of the normalized adjacency matrix to those obtained after randomly signflipping its entries. The proposed method, called network signflip parallel analysis (NetFlipPA), is interpretable, simple to implement, data driven, and does not require users to carefully tune parameters. We show that NetFlipPA provably recovers the spectral noise floor, thus providing a statistically rigorous randomization-based method for selecting the embedding dimension. Compared to traditional cutoff-based methods, the data-driven threshold used in NetFlipPA is provably effective under milder assumptions on the node degree heterogeneity and the number of node communities.

PriME: privacy-aware membership profile estimation in networks

Sagnik Nandy, The Ohio State University

Abhinav Chakraborty (Columbia University); Sayak Chatterjee (University of Pennsylvania); Sagnik Nandy (Ohio State university)

In this talk, we study mixed-membership estimation in degree-corrected network models under edge-level local differential privacy. Privatizing each observed link introduces noise that not only perturbs the observed network but also substantially corrupts degree information changing the fundamental nature of the statistical problem in settings where there is substantial degree heterogeneity. We introduce PriME, a privacy-aware spectral procedure that debiases the privatized adjacency matrix and estimates mixed-membership structure. We establish high-probability error bounds together with minimax lower bounds, showing that PriME is optimal up to logarithmic factors over arbitrary edge-local private mechanisms under the stochastic block model framework. The talk highlights how privacy changes both the optimal algorithmic strategy and the fundamental limits of community membership estimation under privacy constraints in degree heterogeneous networks.

Spectral embeddings via random geometric graphs for noisy, high-dimensional, and nonlinear datasets with applications

Haixiao Wang, University of Washington

Haixiao Wang(University of Washington); Xiucai Ding(University of California Davis)

Clustering is one of the fundamental problems in statistics and machine learning. Many models, such as the Stochastic Block Model (SBM) and the Gaussian Mixture Model (GMM), have been widely used for synthetic data generation and for theoretical evaluation of algorithms. Much of the existing literature focuses on the scenario where cluster centers are linearly separable, which does not reflect the nonlinear geometry of data in many applications. We consider a multi-manifold generative model: disjoint manifolds correspond to different clusters, and their geometric centers are not necessarily linearly separable. Datasets are sampled from the multi-manifold model and then corrupted by some high-dimensional noise. We propose a two-stage algorithm, where in the first stage, we construct a Random Geometric Graph (RGG) from the data and obtain an embedding matrix from its adjacency matrix; in the second stage, we apply spectral clustering to that embedding. When the manifolds are sufficiently separated, the procedure recovers the community structure with a vanishing error rate, as we establish both theoretically and empirically. Finally, following the framework established by Ding and Ma (2023), we prove that the embeddings converge to their noiseless counterparts when the signal-to-noise ratio is sufficiently large, demonstrating the robustness of our algorithm.

Parallel Sessions 6: Monday, November 2, 2026, 3:30 pm–5:00 pm

Session 82: Artificial Intelligence in Rare Disease Therapeutics: Integrating Drug Repurposing, Hybrid Trial Design, and Real-World Evidence

Location: Salon C

Organizer: Bryan McComb (Pfizer, Inc.)

Chair: Bryan McComb (Pfizer, Inc.)

Presenters: Bryan McComb; Qing Liu; Xiaofei Wang; Ching-Ray Yu (Discussant)

AI/ML applications for drug repurposing in rare diseases

Bryan McComb, Pfizer, Inc

Mercedeh Ghadessi (Bayer U.S. LLC); Maha Karnoub (Daiichi Sankyo, Inc); Oleksandr Sverdlov (Novartis Pharmaceuticals Corporation); Gianna McMillan (PRIM&R); Bryan McComb (Pfizer, Inc); Vanja Vlajnic (Bayer U.S. LLC); Rui Tang (Astellas Pharma); Aravindhan Veerapandiyan (University of Arkansas for Medical Sciences, Arkansas Children's Hospital)

Drug repurposing offers a promising strategy for addressing the unmet needs of rare diseases, where clinical development is often hindered by limited patient populations, high costs, and sparse data. The integration of artificial intelligence (AI) and machine learning (ML) transforms this landscape by enabling the analysis of vast, heterogeneous datasets—including genomics, proteomics, and real-world evidence—to identify drug-disease relationships and predict therapeutic efficacy.

This paper evaluates AI/ML-driven drug repurposing for rare diseases, citing successful cases like Baricitinib for COVID-19 and Alpelisib for PROS. We introduce the DREAM (Drug Repurposing with Efficient AI/ML Application) framework, a systematic approach covering the entire pipeline from data acquisition and feature engineering to in silico prediction and clinical validation.

We also address critical barriers, including data scarcity, regulatory hurdles, ethical concerns, and model interpretability, while proposing strategies for mitigation. Ultimately, we emphasize that interdisciplinary collaboration among researchers, clinicians, regulators, and patient advocacy groups is essential. By harnessing AI/ML, drug repurposing can evolve into a predictive, systematic, and patient-centered discipline, providing meaningful hope to the millions affected by rare diseases.

Harnessing artificial intelligence to accelerate rare disease drug development: a strategic framework across preclinical, clinical, and regulatory domains

Qing Liu, Quantitative & Regulatory Medical Science LLC

Qing Liu

The development of therapeutics for rare diseases presents unique challenges due to limited patient populations, heterogeneous phenotypes, and sparse data. Artificial Intelligence (AI) offers transformative potential to address these obstacles by enabling data-driven decision-making, automating analytical workflows, and enhancing regulatory readiness. This talk outlines a comprehensive AI-powered framework tailored to support rare disease drug development across seven critical objectives: (1) AI-driven literature mining and knowledge graph generation to facilitate rapid disease education and biomarker discovery; (2) design of integrated preclinical and clinical strategies using predictive modeling and simulation tools; (3) protocol and Statistical Analysis Plan (SAP) development leveraging natural language generation (NLG) and adaptive trial design algorithms; (4) creation of analytical pipelines in R for real-time data summarization, visualization, and inferential statistics; (5) generation of top-line results and Clinical Study Reports (CSRs) via automated report-writing tools and reproducible analytics; (6) compilation of regulatory briefing documents with AI-assisted summarization and formatting aligned with global authority expectations; and (7) AI-augmented response systems for timely, evidence-based replies to regulatory queries post submission. This strategic integration of AI throughout the rare disease drug development lifecycle enhances scientific rigor, operational efficiency, and regulatory compliance—ultimately accelerating time to patient access for novel therapies.

Hybrid-controlled trials: real-world data, AI-enabled analysis, and lessons learned

Xiaofei Wang, Duke University

Ke Zhu (NCSU), Jiajun Liu (Duke), Shu Yang (NCSU)

Hybrid-controlled trials that augment randomized controlled trials (RCTs) with external controls from real-world data (RWD) are emerging as a key AI application for generating regulatory-grade evidence. In this session, we use a case study in stage IB non-small cell lung cancer, integrating CALGB 9633 with external controls from the National Cancer Database, to illustrate how modern machine learning and AI tools can both strengthen and threaten causal inference in this setting. First, we show that flexible ML/AI algorithms, such as ensemble methods for scoring and outcome regression, enable doubly robust estimation of average and conditional treatment effects under covariate shift between RCT and external-control populations. These approaches can substantially improve efficiency over RCT-only analyses while preserving estimand interpretability. Second, we discuss AI-assisted procedures for constructing synthetic controls and patient-level digital twins from RWD to support more precise and individualized inference. We conclude with lessons learned and regulatory insights: how to document AI-driven borrowing rules, diagnose hidden bias and data-quality issues, communicate digital-twin and hybrid-control results to clinicians and reviewers, and balance efficiency gains against the risks of over-borrowing or model misspecification in RWD-based regulatory decision-making.

Discussant: Ching-Ray Yu, Pfizer, Inc

Session 31: From Statistical Inference to Production AI: Industry Case Studies in Causality and Intelligent Systems

Location: Williamsburg

Organizer: Wanjun Liu (LinkedIn Corporation)

Chair: Wanjun Liu (LinkedIn Corporation)

Presenters: Haotian Lin; Xiufan Yu; Wanjun Liu

ReSkill: reconciling skill creation with policy optimization in agentic RL

Haotian Lin, Amazon

Zelin He (Penn State), Haotian Lin (Amazon), Boran Han (Amazon), Wei Zhu (Amazon), Haoyang Fang (Amazon), Bernie Wang (Amazon), Xuan Zhu (Amazon), Runze Li (Penn State), Matthew Reimherr (Amazon, Penn State)

Reinforcement learning lets LLM agents improve from environment rewards, but the learned policies do not systematically accumulate reusable problem-solving strategies. Modular "skills" can fill this gap, yet existing methods create skills separately from policy training, so a skill may stop helping as the policy evolves. We present ReSkill, which tests and refines skills inside the RL loop. Exploiting the group structure of GRPO, rollouts within the same group are assigned to competing skill versions, giving a controlled within-task comparison at no extra rollout cost. Which version to keep is framed as a non-stationary two-armed bandit, solved by Thompson Sampling with adaptive discounting that down-weights stale evidence as the policy changes. Across embodied reasoning, agentic search, coding, and web-shopping benchmarks, ReSkill outperforms memory- and skill-based baselines, with the largest gains on unseen tasks.

Deep causal learning for heterogeneous panel data models

Xiufan Yu, University of Notre Dame

Guanhao Zhou (University of Notre Dame); Yuefeng Han (University of Notre Dame); Xiufan Yu (University of Notre Dame)

This talk focuses on the task of estimating heterogeneous treatment effects in causal panel data models. We propose a novel Covariate-Adjusted Deep Causal Learning framework, that cohesively deals with the underlying heterogeneity and nonlinearity of both panel units and covariate effects. CoDEAL integrates nonlinear covariate effect components (parameterized by a feed-forward neural network) with nonlinear factor structures (modeled by a multi-output autoencoder) to form a heterogeneous causal panel model. The nonlinear covariate component flexibly captures complex covariate influences on outcomes, and the nonlinear factor decomposition enables CoDEAL to effectively capture both cross-sectional and temporal dependencies inherent in the data panel. This latent structural information is subsequently integrated into a customized matrix completion algorithm, thereby facilitating more accurate counterfactual imputation. Moreover, the use of a multi-output autoencoder explicitly accounts for heterogeneity across units and enhances the model interpretability of the latent factors. We establish theoretical guarantees on the convergence of the estimated counterfactuals and demonstrate the compelling performance of the proposed method using extensive simulation studies and real-world data applications.

Quantifying the effectiveness of advertising: a bootstrap proportion test for brand lift testing

Wanjun Liu, LinkedIn

Wanjun Liu (LinkedIn); Xiufan Yu (University of Notre Dame); Jialiang Mao (Uber); Xiaoxu Wu (LinkedIn); Justin Dyer (Uber)

Brand Lift test is a widely deployed statistical tool for measuring the effectiveness of online advertisements on brand perception such as ad recall, brand familiarity and favorability. By formulating the problem of interest into a two-sample test on the binomial proportions from the control group ($p0$) and the treatment group ($p1$), Brand Lift test evaluates ads impact based on the statistical significance of test results. Traditional approaches construct the test statistics based on the absolute difference between the two observed proportions, a.k.a, absolute lift. In this work, we propose a new bootstrap test based on the percentage difference between the two observed proportions, i.e., relative lift. We provide rigorous theoretical guarantees on the asymptotic validity of the proposed relative-lift-based test. Our numerical studies suggest that the relative-lift-based test requires less stringent conditions than the absolute-lift-based test for controlling the type-I error rate. Interestingly, we also prove that the relative-lift-based test is more powerful than the absolute lift-based test when the alternative is positive (i.e., $p1 - p0 > 0$), but less powerful when the alternative is negative (i.e., $p1 - p0 < 0$). The empirical performance of the proposed test is demonstrated by extensive simulation studies, an application to a publicly available A/B testing dataset from advertising, and real datasets collected from the Brand Lift Testing platform at LinkedIn.

Session 72: Methods and Theory for Machine Learning: Estimation and Prediction

Location: Fulton Ferry

Organizer: Yang Ning (Cornell University)

Chair: Yang Ning (Cornell University)

Presenters: Mengxin Yu; Xinyi Li; Yang Ning; Matey Neykov

Beyond exchangeability: predictive inference under general data symmetries

Mengxin Yu, Washington University in St. Louis

Yichen Shen (University of Chicago); Mengxin Yu (Washington University in St. Louis)

Quantifying uncertainty in predictions is a fundamental

problem in modern statistics and machine learning. Many existing methods for predictive inference, including standard conformal prediction, rely on exchangeability of the data. Yet modern datasets often exhibit richer forms of dependence and structure, arising, for example, from networks, hierarchical data, or other structured observations, for which exchangeability may no longer be appropriate.

To tackle this challenge, in the first part of the talk, I will introduce a general framework for predictive inference based on group symmetries. By exploiting distributional invariance under general group actions, this framework extends predictive inference beyond exchangeable data and provides finite-sample marginal coverage guarantees for a broad class of structured data.

In the second part of the talk, I will turn to conditional predictive inference. While marginal coverage ensures validity on average, it may mask substantial heterogeneity across observations. Building on the symmetry-based framework, I will introduce an approach to near-conditional coverage for structured, non-exchangeable data. The proposed method measures conditional miscoverage over a user-specified function class, allowing the notion of conditional validity to adapt to the problem structure. Together, these results provide a unified perspective on predictive inference beyond exchangeability, from marginal to conditional guarantees.

Linear regression using principal components from general Hilbert-space-valued covariates

Xinyi Li, Clemson University

Xinyi Li (Clemson University); Margaret Hoch (University of North Carolina at Chapel Hill); Michael R. Kosorok (University of North Carolina at Chapel Hill)

We introduce Adaptive Subspace PCA (AS-PCA), a framework for principal component analysis of random elements in a general separable Hilbert space. AS-PCA projects the covariance operator onto a data-adaptive finite-dimensional subspace prior to eigendecomposition, requiring no kernel specification and accommodating multi-dimensional functional objects including images and surfaces. Under the second-moment condition, we prove a Donsker theorem for Hilbert-space-valued empirical processes and use it to establish uniform consistency and joint Gaussian limits for the leading eigenpairs. A data-driven diagnostic verifies projection accuracy, and a consistent proportion-of-variance-explained rule selects the number of components. Building on AS-PCA, we construct Hilbert-Space Principal Component Regression (HS-PCR) for models combining Euclidean and Hilbert-space-valued covariates. The HS-PCR estimator is root-n consistent and asymptotically normal, with an explicit influence function decomposition accounting for eigenfunction estimation uncertainty. Both nonparametric and wild bootstrap procedures are shown to be asymptotically valid. Simulations with two- and three-dimensional imaging predictors confirm accurate eigenstructure recovery and nominal bootstrap coverage. HS-PCR is applied to Alzheimer's Disease Neuroimaging Initiative data in regression and precision-medicine settings.

Active subsampling for binary response models

Yang Ning, Cornell University

Jingyi Duan (Cornell); Lehao Fu (Cornell); Yang Ning (Cornell)

In the measurement-constrained problems, despite the availability of large datasets, we may be only affordable to observe the labels on a small portion of the large dataset. This poses a critical question that which data points are most beneficial to label given a budget constraint. In this paper, we focus on the estimation of the optimal individualized threshold in a measurement-constrained M-estimation framework. Our goal is to estimate a high-dimensional parameter in a linear threshold for a continuous variable X such that the discrepancy between whether X exceeds the threshold and a binary outcome is minimized. We propose a novel K-step active subsampling algorithm to estimate θ, which iteratively samples the most informative observations and solves a regularized M-estimator. We show that the two-step algorithm yields an estimator with the parametric convergence rate. Furthermore, we formulate a minimax framework for the measurement-constrained M-estimation problem and prove that our estimator is minimax rate optimal up to a logarithmic factor. Finally, we demonstrate the performance of our method in simulation studies and apply the method to analyze a large diabetes dataset.

Polynomial-time near-optimal estimation over certain type-2 convex bodies

Matey Neykov, Northwestern University

Matey Neykov

We will discuss near (minimax) optimal computationally efficient estimators for constrained estimation over certain special convex constraints. The methodology is applicable to the problem of the constrained Gaussian sequence model, as well as constrained robust mean estimation and regression. Examples of sets over which this method can be applied include all linearly transformed quadratically convex orthosymmetric sets.

Session 55: Statistical-Computational Gaps in Network and Tensor Data

Location: Sunset Park

Organizer: Keith Levin (University of Wisconsin-Madison)

Chair: Keith Levin (University of Wisconsin-Madison)

Presenters: Cheng Mao; Alessandro Rinaldo; Runshi Tang; Subhabrata Sen

Resolution of the detection threshold conjecture for random geometric graphs

Cheng Mao, Georgia Institute of Technology

Hang Du (Massachusetts Institute of Technology), Cheng Mao (Georgia Institute of Technology), Nike Sun (Massachusetts Institute of Technology), Yihong Wu (Yale University), Jiaming Xu (Duke University)

A random geometric graph (RGG) is generated by first sampling latent points $x_1, \dots, x_n$ independently and uniformly from the unit sphere in $\mathbb{R}^d$, and then connecting each pair (i, j) if $\langle x_i, x_j \rangle$ exceeds some threshold τ. We study the sharp detection threshold—the largest dimension at which the RGG can be statistically distinguished from the Erdős–Rényi graph with the same edge density p. This threshold is conjectured to be $d \asymp (nh(p))^3$, where $h(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$ is the binary entropy function. Previous works proved this conjecture for dense graphs with constant p and, up to polylogarithmic factors, very sparse graphs with $p = \Theta(1/n)$. In this paper, we prove that detection is impossible when $d \gg (nh(p))^3$ and $d \geq (1+\epsilon)n$ for any constant $\epsilon > 0$, thereby resolving the conjecture in the regime $p \gtrsim n^{-2/3}/\log n$ and improving upon the state of the art in the regime $1/n \ll p \ll n^{-2/3}/\log n$. The key to our proof is a sharp analysis of the posterior distribution of the latent points given the observed graph, obtained through an information-theoretic comparison argument combined with strong log-concavity.

Statistical and computational gaps in Gaussian graph combinatorial testing problems

Alessandro Rinaldo, The University of Texas at Austin

Ikjun Choi (UT Austin); Pratik Patil (UT Austin); Alessandro Rinaldo (UT Austin)

We investigate statistical and computational gaps in graph combinatorial testing problems in which the data take the form of an isotropic Gaussian random vector whose coordinates are indexed by the edges of complete graph. Under the null hypothesis, the overall mean is zero. Under the structured alternative hypothesis, the coordinates corresponding to a random selection from a symmetric class of edge-induced subgraphs have constant non-zero mean. We seek to identify conditions under which the testing problem has vanishing error as the size of the data and the complexity of the combinatorial class of alternatives grow. We identify statistical and computational thresholds for the signal magnitude that separate regimes for which testing succeeds from those for which it fails for various combinatorial graph alternatives, including multiple planted cliques, hyper-cliques, thick cycles and paths. These examples exhibit regimes in which exhaustive scan tests attain the statistical limit, whereas natural polynomial-time tests such as averages and coordinatewise maxima require substantially larger signal. In our analysis we rely on low degree polynomial analysis and our results hold under an appropriate low-degree polynomial conjecture. We prove a general equivalence principle for natural exponential families, by formulating natural conditions ensuring that statistical and low-degree computational thresholds from one family can be mapped to analogous ones for another family. As a by product, we show that the thresholds obtained from Gaussian combinatorial testing recover known thresholds for planted graph problems in the Erdös-Rényi model.

Detection is harder than estimation in certain regimes: inference for moment and cumulant tensors

Runshi Tang, University of Wisconsin–Madison

Runshi Tang (UW-Madison); Yuefeng Han (University of Notre Dame); Anru R. Zhang (Duke University)

We study estimation and detection of high-order cumulant tensors from n i.i.d. observations of a p-dimensional random vector, with performance measured in the tensor spectral norm. Under sub-Gaussianity, we show that the minimax rate for estimating the order-d cumulant tensor is $\sqrt{p/n} \wedge 1$. In contrast to covariance estimation, the natural plug-in estimator obtained by replacing population moments in the moment–cumulant formula with sample moments, which we call the sample cumulant tensor, is generally not rate-optimal for $d \geq 3$, and we construct an estimator that attains the minimax rate up to logarithmic factors. On the computational side, we study testing whether the d-th order cumulant tensor vanishes after whitening. We provide evidence, under the low-degree polynomial framework, that detection is computationally hard when $n \ll p^{d/2}$. At the same time, we identify a regime in which an efficiently computable estimator achieves error smaller than the separation at which low-degree polynomial tests can reliably distinguish the null from the alternative. This reveals an unusual reverse detection–estimation gap: computationally efficient detection can be harder than computationally efficient estimation. The underlying reason is that the relevant loss, tensor spectral norm, is itself NP-hard to compute, creating a new form of computational–statistical gap.

Causal inference under interference: computational barriers and algorithmic solutions

Subhabrata Sen, Harvard University

Sohom Bhattacharya (University of Florida); Subhabrata Sen (Harvard University)

We study causal effect estimation under the chain graph framework pioneered in Tchetgen Tchetgen et. al. (2021). We will discuss computational barriers in evaluating the causal effect under this framework. Subsequently, we will introduce efficient algorithms for causal effect estimation under additional structural assumptions on the underlying chain graph.

Session 12: Modern Perspectives in Bayesian Statistics

Location: Golden Boardroom

Organizer: Guanyu Hu (Michigan State University)

Chair: Guanyu Hu (Michigan State University)

Presenters: Yanxun Xu; Didong Li; Yuexi Wang; Guanyu Hu

Self-supervised representation learning for longitudinal polypharmacy patterns

Yanxun Xu, Johns Hopkins University

Yunshan Duan (Johns Hopkins University); Eero Korpela (Johns Hopkins University); Yanxun Xu (Johns Hopkins University)

Polypharmacy, commonly defined as the concurrent use of multiple medications, is increasingly prevalent in aging populations and is associated with adverse health outcomes. Motivated by longitudinal studies of aging people, we study how medication use evolves over time and characterizes multimorbidity patterns. Longitudinal medication data present substantial methodological challenges, including high dimensionality, sparsity, irregular observation times, and structured pharmacologic relationships among medications. Existing approaches are typically task-specific and lack a unified framework for learning general-purpose representations of medication trajectories. We propose LOPEL (LOngitudinal Polypharmacy Embedding Learning), a two-stage self-supervised framework for learning low-dimensional representations of longitudinal medication data. In the first stage, visit-level embeddings are learned using a Gaussian process model that incorporates pharmacologic similarity through the Anatomical Therapeutic Chemical hierarchy. In the second stage, subject-level embeddings are constructed by modeling trajectories over time and defining similarity through a Wasserstein-based representation capturing temporal dynamics and uncertainty. Simulation studies demonstrate that LOPEL accurately recovers latent structure under realistic conditions with high-dimensional sparsity and irregular sampling. In an application to aging cohorts, LOPEL identifies clinically meaningful subgroups that differ in the timing, composition, and progression of medication use, highlighting heterogeneity relevant for risk stratification and clinical management.

From human knowledge to informative priors: LLM embeddings for statistical inference

Didong Li, University of North Carolina at Chapel Hill

Hongqian Niu (University of North Carolina at Chapel Hill); Jordan Bryan (UVA); Jacob Williams (NCI); Hufeng Zhou (Harvard); Zhun Deng (UNC); Haoyu Zhang (NCI); Xihao Li (UNC); Didong Li (UNC, presenting author)

External knowledge can provide valuable prior information for statistical inference, but much of this knowledge is unstructured and therefore difficult to incorporate into conventional statistical models. Large language models (LLMs) provide a new opportunity to bridge this gap by representing unstructured knowledge through numerical embeddings that can be used to construct informative priors. In this talk, I will present two recent works illustrating this idea in genomics and genetics. First, I will discuss a method that uses gene embeddings from GPT to incorporate external biological knowledge into high-dimensional hypothesis testing, motivated by cancer genomics problems where traditional methods can be underpowered. Second, I will discuss our recent work in genetics, where we curated annotations for 9 billion genetic variants across the human genome and obtained embeddings of these variants using GPT and Qwen embedding models. These embeddings provide a representation of prior knowledge that can be incorporated into downstream statistical analyses, including genome-wide association studies and polygenic risk prediction. Together, these examples illustrate how LLMs can transform large amounts of unstructured scientific knowledge into structured prior information, providing a new approach to integrating AI with Bayesian and statistical inference.

Simulation-based inference via structured score matching

Yuexi Wang, University of Illinois Urbana-Champaign

Haoyu Jiang (University of Illinois Urbana-Champaign); Yuexi Wang (University of Illinois Urbana-Champaign); Yun Yang (University of Maryland, College Park)

Simulation-based inference (SBI) provides a powerful framework for statistical analysis when the likelihood function is intractable but simulations are available. I will present a unified framework that leverages structured score matching to enable both maximum likelihood estimation and Bayesian posterior sampling in such likelihood-free settings. The key idea is to approximate the intractable likelihood score function with neural networks that embed the statistical structures of likelihood score through architectural regularization. These structural constraints enhance estimation accuracy, scalability, and uncertainty quantification. We establish theoretical guarantees for the proposed methods and demonstrate their practical advantages on benchmark tasks and challenging high-dimensional problems, where they perform favorably against existing approaches.

How competitive is competition? A Bayesian view of parity over time

Guanyu Hu, Michigan State University

Guanyu Hu (MSU); Maoran Xu (IUB); Xinran Li (UIUC)

How competitive is a competition, really? In settings such as professional sports, mechanisms are often designed to promote parity, yet observed rankings can suggest substantial differences that reflect sampling variability as much as genuine differences in ability. This raises several fundamental questions: how should competitive balance be quantified, how does it evolve over time, and when do the data support distinct ranks rather than statistically

indistinguishable tiers? In this talk, I will present a Bayesian framework that treats competitive balance as a latent structural quantity rather than a descriptive summary of observed outcomes. Competitor abilities are constrained by a time-varying geometric structure whose scale provides an interpretable measure of competitive imbalance. Importantly, this geometry also induces data-driven rank tiers by assigning positive probability to exact equality in latent abilities. Thus, when competitors are closely matched, the model can group them into common tiers rather than forcing a strict ranking, while meaningful ability differences remain distinguishable. To capture changes over time, a hidden Markov mixture model identifies persistent competitive regimes and allows the number of such regimes to be learned from the data. Applications to professional sports leagues illustrate how the framework can reveal the rise and fall of competitive balance, identify persistent historical regimes, and quantify uncertainty in rank tiers. More broadly, the approach provides a probabilistic perspective on when observed rankings reflect meaningful differences in ability and when apparent separation is better understood as noise.

Student Paper Award Session

Time: Monday, November 2, 2026, 3:30 pm–5:00 pm

Location: Carroll Gardens

Detection is harder than estimation in certain regimes: inference for moment and cumulant tensors

Augustine Tang, Columbia University

M-estimation under two-phase multiwave sampling with applications to prediction-powered inference

Dan Kluger, MIT

Statistical guarantees for high-dimensional stochastic gradient descent

Jiaqi Li, Rice University

A statistical theory of overfitting for imbalanced classification

Jingyang Lyu, University of Wisconsin–Madison

Hierarchical clustering with confidence

Judy Wu, University of Michigan

Fast segmentation of watermarked texts from large language models through epidemic change-points framework

Soham Bonnerjee, University of Chicago

Session 10: Statistical Advances in the Analysis of Embeddings, Networks, and Graphs

Location: Coney Island

Organizer: Joshua Cape (University of Wisconsin-Madison)

Chair: Joshua Cape (University of Wisconsin-Madison)

Presenters: Xiaoyue Niu; Marianna Pensky; Jingming Wang; Siddharth Vishwanath

A dyadic latent space model for directed networks

Xiaoyue Niu, Pennsylvania State University

Xiaoran Chen (Stanford University); Hyesun Yoo (Google); Xiaoyue Niu (Pennsylvania State University); Ji Zhu (University of Michigan)

Reciprocity, the tendency of two nodes to form mutual ties, is a defining feature of directed networks, yet most models capture it only through a single global parameter, community membership, or observed covariates. In many applications the propensity to reciprocate varies across node pairs and is driven by latent node characteristics. We propose a latent space model for directed networks that models each pair of reciprocal edges jointly. The model allows the strength of reciprocity to vary flexibly across node pairs while retaining the interpretability and low-dimensional structure of latent space models, and it nests the classical p_1 model as a special case. We establish identifiability, develop a scalable projected gradient descent algorithm for maximum likelihood estimation, derive finite-sample error bounds, and extend the algorithm to link prediction. Applied to the international trade network for precision and medical instruments, the model improves link prediction and uncovers strongly heterogeneous reciprocity across country pairs.

Learning from strong layers to classify weak layers in heterogeneous multilayer networks

Marianna Pensky, University of Central Florida

Marianna Pensky (University of Central Florida) , Abdulla Toleugazin (University of Central Florida)

We study clustering of heterogeneous multilayer networks in which signal strength may vary substantially across layers. While sufficiently informative layers can be reliably clustered using estimated latent subspaces, weaker layers may not support accurate stand-alone subspace estimation. We therefore introduce a two-stage method combining clustering and classification. In the first stage, sufficiently informative active layers are identified and clustered according to their latent subspaces, and information is aggregated within the resulting groups to estimate group-specific subspaces. In the second stage, these learned subspaces are used to classify the remaining weaker layers through projection-based scores. Thus, structural information learned from stronger layers is transferred to layers with substantially weaker signals. We develop this framework for heterogeneous nonsymmetric multilayer networks under Generalized Random Dot Product Graph models. Layers share a common set of row

nodes but may differ in their column-node sets, dimensions, sparsity levels, and probability matrices. Layer groups are characterized by common left latent subspaces, while right latent spaces and other layer-specific parameters may vary. We provide non-asymptotic guarantees for perfect clustering of active layers, consistent estimation of group-specific subspaces, and correct classification of weaker layers. Numerical experiments illustrate the benefits of the proposed clustering–classification strategy.

Network goodness-of-fit for the block model family

Jingming Wang, University of Virginia

Jiashun Jin (Carnegie Mellon University); Tracy Zheng Ke (Harvard University); Jiajun Tang(Harvard University); Jingming Wang(University of Virginia)

The block model family is widely used in network modeling and includes four popular models: SBM, DCBM, MMSBM and DCMM. However, the question of which block model best fits real networks has received limited attention in the literature. In this talk, I will introduce a novel approach using cycle count statistics to address the Goodness-of-Fit for these block models. By leveraging the cycle count statistics and a network fitting scheme, we construct four GoF metrics with parameter-free limiting distributions of $N(0,1)$ under the assumed models. We apply these GoF-metrics to some frequently-used real networks for comparison. The numerical results suggest that DCMM is particularly promising for modeling undirected networks.

Multidimensional scaling from noisy data: statistical learning & uncertainty quantification

Siddharth Vishwanath, University of Virginia

Siddharth Vishwanath (University of Virginia); Ery Arias-Castro (University of California San Diego)

Multidimensional scaling (MDS) has a long history in statistics and underpins a broad class of unsupervised learning techniques from relational and network data.

The objective of MDS is to extract meaningful information from relational data (e.g., distances between sensors, correlations between brain regions, or disagreement scores between individuals) by embedding these relationships into a Euclidean space. In practice, the observed relational information is often subject to measurement errors and/or corrupted by noise. However, the resulting embeddings are typically interpreted as exploratory visualizations without accounting for these variations. This talk presents recent work developing a principled statistical framework for MDS.

We show that the classical MDS algorithm achieves minimax-optimal performance across a wide range of noise models and loss functions, and develop a framework for constructing valid confidence sets for the embedded points obtained via MDS, enabling formal uncertainty quantification for geometric structure inferred from noisy relational data.

Session 74: Enhanced Causal Inference and Clinical Trials with AI/ML

Location: Park Slope

Organizer: Muxuan Liang (The University of Texas MD Anderson Cancer Center)

Chair: Wodan Ling (Weill Cornell Medicine)

Presenters: Liangyuan Hu; Heng Zhou; Ruitao Lin

A Bayesian machine learning approach for causal inference with time-varying confounding and survival outcomes

Liangyuan Hu, Rutgers University

Keming Zhang (Brown University); Liangyuan Hu (Presenting author; Rutgers University); Yudan Zou (Rutgers University)

We propose a flexible Bayesian framework for survival analysis with longitudinal, time-varying covariates that combines a log-linear Bayesian additive regression trees (BART) component, a natural cubic spline representation of the log-baseline hazard, and subject-specific random effects. The model preserves the interpretability of a Cox-type hazard structure while leveraging BART's nonparametric flexibility to capture nonlinear effects and high-order interactions among time-varying covariates. We develop a Gibbs sampler combining blocked Metropolis–Hastings updates for the trees with random-walk Metropolis steps for the spline coefficients and random effects, run under a raw parameterization to preserve conjugate tree updates, with interpretable baseline, BART, and frailty components recovered afterward by deterministic posterior re-centering. Building on this survival model, we embed CoxBART within a Bayesian marginal structural model to estimate time-varying treatment effects, using CoxBART to flexibly estimate the time-varying treatment-assignment hazards that yield stabilized inverse-probability weights, with uncertainty propagated through a Bayesian bootstrap. Extensive simulations under proportional and non-proportional hazards, and a causal setting with hinge-type confounder effects, show that the method accurately recovers the baseline hazard, the hazard-ratio surface, and counterfactual restricted mean survival times, outperforming standard alternatives including Cox proportional hazards, kernel-based time-varying Cox, and random survival forests.

Optimizing quality tolerance limits monitoring in clinical trials through machine learning methods

Heng Zhou, Bristol-Myers Squibb

Rachael Liu(Bristol-Myers Squibb), Lei Yan(FSU), Liwen Wu(Takeda), Jianchang Lin

Traditional clinical trial monitoring relies heavily on site visits and manual review of cumulative patient data from Electronic Data Capture systems, making it time- and resource-intensive. Risk-based monitoring (RBM) and quality tolerance limit (QTL) frameworks provide a more

efficient alternative to traditional source data verification by proactively identifying systematic issues affecting patient safety and data integrity. We propose a machine learning–enabled approach for real-time, automated QTL risk assessment. Unlike traditional approaches that evaluate QTLs using single-source data and arbitrarily defined fixed thresholds, our QTL-ML framework integrates information across multiple clinical domains to predict diverse QTL risks at the program, study, site, and patient levels. The approach is assumption-free, relying on dynamically accumulating trial data rather than historical expectations to identify emerging quality risks. Embedded within ICH E6–recommended RBM principles, QTL-ML has the potential to enhance patient safety and data quality while reducing trial duration, monitoring burden, and overall trial costs.

GRACIAS: gradient-augmented calibration incorporating AI for covariate adjustment in randomized studies

Ruitao Lin, The University of Texas MD Anderson Cancer Center

Zhirui Li; Ruitao Lin; Xiudi Li; Muxuan Liang

Pretrained AI models can encode rich prognostic information, but predictive accuracy alone does not guarantee improved efficiency for treatment-effect estimation in randomized trials. We develop a general calibration framework for augmented inverse-probability weighted (AIPW) estimation in multi-arm trials and characterize conditions for simultaneous variance optimality across treatment contrasts. Under stated regularity and arm-specificity assumptions, requiring this optimality uniformly over admissible data-generating distributions implies that calibration is affine in its parameters, up to reparameterization. Motivated by this characterization, we propose GRACIAS, which combines AI-generated outcome predictions under all treatment arms with low-rank features obtained from model-parameter gradients. Separate Oaxaca–Blinder least-squares regressions within each treatment arm calibrate these features on a common basis, and the resulting outcome models enter the AIPW estimator. Under simple randomization, this construction simultaneously minimizes asymptotic variance within the chosen linear calibration space for every zero-sum treatment contrast. We further establish a local approximation result: under smoothness conditions, when the true outcome regression is represented by nearby model parameters and the relevant gradient directions are retained, the augmented calibration space contains an approximation that removes the leading first-order error. We extend the calibration framework and its optimality conditions to covariate-adaptive randomization. Simulations and analyses of 21 randomized trials demonstrate precision gains from calibration relative to directly using AI predictions, with additional improvements from gradient- and representation-augmented implementations. These results connect flexible AI prediction with classical regression adjustment and provide a principled framework for using AI-derived information to improve treatment-effect estimation.

Session 47: High-Dimensional Data in Neuroimaging and AI

Location: Windsor Terrace

Organizer: Guanqun Cao (Michigan State University)

Chair: Jiawa Zhang (Columbia University)

Presenters: Yang Long; Shan Yu; Soham Jana; Guanqun Cao

Synthetic surrogate functional regression (SSFR): a robust framework for estimating spatially varying covariate effects in surrogate-augmented neuroimaging

Yang Long, Yale University

Guannan Wang (William & Mary); Yang Long (Yale University); Shan Yu (University of Virginia); Zhiling Gu (University of South Carolina); Lily Wang (George Mason University)

Understanding how demographic, clinical, and genetic covariates relate to complex spatial imaging outcomes is a fundamental problem in brain imaging, particularly when imaging responses are high-dimensional and only partially observed. Modern generative models can synthesize missing imaging modalities from auxiliary data, but incorporating synthetic images into regression requires careful statistical treatment to ensure valid inference. We propose a two-stage Synthetic Surrogate Functional Regression (SSFR) framework for image-on-scalar regression, designed as a statistical safeguard for using synthetic neuroimages to learn spatially varying covariate effects when the primary imaging modality is incompletely observed. SSFR integrates observed and surrogate images in a unified regression framework, improving estimation efficiency while remaining robust to discrepancies between the surrogate and target regression functions. We further develop a weighted bootstrap procedure for constructing simultaneous confidence corridors (SCCs) for the estimated coefficient functions. As surrogate informativeness increases, SCC widths approach those under complete observations, improving localization of significant spatial effects. Applied to spatially normalized PET images from ADNI, SSFR yields interpretable coefficient maps and uniformly narrower SCCs than observed-only analysis, enlarging the region over which age-related metabolic decline is simultaneously significant.

Causal learning with invalid instruments for high-dimensional imaging responses

Shan Yu, University of Virginia

Yiting Wang (University of Virginia); Chunlin Li (University of Virginia); Shan Yu (University of Virginia); Lily Wang (George Mason University)

Large-scale neuroimaging studies provide high-dimensional imaging phenotypes for understanding

brain-related diseases and their causal relations with risk factors. While instrumental variable (IV) estimation is effective for uncovering causality, existing IV methods face significant limitations when applied to high-dimensional imaging data, particularly in handling large-scale data and leveraging the intrinsic spatial structure of images. To address these challenges, we propose a novel framework, High-dimensional Functional-imaging Instrumental Variable Estimation (Hi-FIVE), which enables estimation of spatially varying causal effects in the presence of many invalid IVs. Our approach uses nonparametric spatial smoothing through multivariate penalized splines to capture the spatial structure of imaging data, combined with distributed learning-based IV selection to handle potentially invalid instruments and large-scale data. We establish rigorous theoretical guarantees for the Hi-FIVE framework, including new identifiability conditions, consistent selection of the invalid instruments, and convergence rates and asymptotic normality for the effect estimator. Building on these theoretical results, we further construct data-driven simultaneous confidence regions for statistical inference. Hi-FIVE is applied to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset to illustrate its effectiveness, under the stated IV assumptions, in uncovering causal effects in large-scale observational neuroimaging studies.

Speckle noise modeling in high-dimensional image sensing

Soham Jana, University of Notre Dame

Xi Chen (Rutgers University); Soham Jana (University of Notre Dame); Christopher A. Metzler (University of Maryland); Shirin Jalali (Rutgers University); Arian Maleki (Columbia University); Hao Xing (City University of New York Graduate Center)

Speckle noise is an integral part in several modern imaging paradigms encompassing ultrasound, synthetic aperture radar, and optical coherence tomography. Deep learning-based methods have been proven fruitful for analyzing sensor data and reconstructing images when the signal resides in a low-dimensional space. However, the existing guarantees for such methods are far from being optimal. In addition, the problem becomes challenging when the number of sensors is very small compared to the image dimension, and it is often necessary to take multiple images of the scene to enable consistent reconstruction. Our recent work establishes estimation thresholds for such problems when the signal can arise from either low- or high-dimensional regimes. Real and simulated data analyses support the efficiency of our estimation techniques.

Heterogeneity-aware optimal classification for multi-source imaging data

Guanqun Cao, Michigan State University

Shuoyang Wang (U of Louisville); Guanqun Cao (MSU)

We propose a federated learning framework for training a deep neural network classifier of imaging data across multiple heterogeneous agents. By modeling imaging data as multi-dimensional functional objects, the framework enables accurate estimation of underlying signals and efficient detection of significant effects. The learner visits each agent once, updates model parameters locally, and transfers only compressed weights, thereby avoiding raw-data sharing and reducing computational overhead. To accommodate inter-cohort heterogeneity, we introduce an adaptive sequential weight-inheritance and constrained updating strategy that transfers information across source datasets while adapting the classifier to the target domain. We establish minimax-optimal excess risk bounds and identify a sharp phase-transition threshold that governs learnability under both densely and sparsely observed imaging data. Extensive simulations and real-world imaging studies demonstrate that the proposed method yields a scalable and accurate classifier for multi-source imaging data, achieves performance comparable to centralized algorithms, and does so without requiring raw-data sharing.

Session 106: From Data to Structure and Inference

Location: DUMBO

Organizer: Marianthi Markatou (SUNY Buffalo)

Chair: Zhezhen Jin (Columbia University)

Presenters: Marianthi Markatou; Raktim Mukhopadhyay; Alex Foss

Variable weighting for mixed-type data clustering

Marianthi Markatou, Department of Biostatistics, SPHHP, University at Buffalo

Marianthi Markatou (University at Buffalo), Alexander H. Foss (Sandia National Labs)

This talk discusses conceptual and technical challenges in variable weighting and variable selection in cluster analysis of mixed-type data. The specific objectives of existing weighting procedures are quite different and are often selected without a thorough discussion of competing objectives and the implications of these selection choices on the resulting clustering. We describe and investigate the different objectives of existing weighting and variable selection techniques and discuss the novel variable weighting technique MEDEA (Multivariate Eigenvalue Decomposition Error Analysis). We show that the MEDEA technique can be flexibly configured based on the choice of association metrics used. In particular, a version of MEDEA is described and tested that is robust to extreme values in the dataset to be clustered. We show that MEDEA weighting fills an important and previously unfilled niche in cluster analysis of mixed-type data, as demonstrated by superior performance in numerous experiments with both synthetic and actual datasets. We provide insight into the core assumptions underlying existing techniques and provide recommendations for the appropriate usage of these techniques.

Collecting data that impacts lives: navigating chaos, challenges and solutions in real-world data collection

Raktim Mukhopadhyay, Department of Biostatistics, School of Public Health and Health Professions, University at Buffalo

Raktim Mukhopadhyay (Department of Biostatistics, School of Public Health and Health Professions, University at Buffalo); Andrew H. Talal (Division of Gastroenterology, Hepatology and Nutrition, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo); Yihao Tan (Department of Biostatistics, School of Public Health and Health Professions, University at Buffalo); Zhezhen Jin (Department of Biostatistics, Mailman School of Public Health, Columbia University); Marianthi Markatou (Department of Biostatistics, School of Public Health and Health Professions, University at Buffalo)

Data play a foundational role in modern life and are fundamental for decision making and innovation, particularly in areas such as public health and biomedicine. However, important processes and considerations for obtaining and preparing “high quality” data for actionable insights are often overlooked and treated as “invisible labor”. This talk outlines the end-to-end lifecycle of data obtained from underserved populations within Opioid Treatment Programs (OTPs) in New York State. In particular, we discuss challenges in obtaining “high quality” data and present potential solutions with an eye towards reproducibility. Furthermore, we propose an extension to the traditional evidence pyramid incorporating the risk when AI systems are used for data-collection and evidence generation. This work is an attempt to leverage underutilized data repositories lying in often overlooked and fragmented components of the healthcare system. We also present important considerations when AI-systems are deployed in OTPs and aim to emphasize the scope of future research.

Class-incremental discriminant analysis

Alex Foss, Sandia National Laboratories

Alex Foss (Sandia National Laboratories); Carlos Llosa (Sandia National Laboratories); Tian Ma (Sandia National Laboratories)

In this talk, we discuss the challenge of designing class-incremental learning algorithms. This refers to incremental or streaming supervised learning techniques that can adaptively learn unknown classes not present in historical data. We begin with a general discussion of unique statistical challenges that arise in the streaming context, and how to balance trade-offs between classical sampling variability and estimation bias that arises due to estimator latency when data distributions evolve over time. We then consider the incremental learning problem and discuss advantages and limitations of various approaches, including a family of discriminant analysis algorithms that are immune to the pernicious problem of catastrophic forgetting. Finally, we discuss considerations in designing data analytics pipelines in which fast statistical classifiers are informed by insights from slower deep learning techniques running in parallel.

Session 103: Recent Methodological Advances in Causal Inference

Location: Greenpoint

Organizer: Caleb Miles (Columbia University)

Chair: Daniel Malinsky (Columbia University)

Presenters: Elizabeth L. Ogburn; Taehyeon Koo; Rajarshi Mukherjee; Rohit Bhattacharya

Spatial confounding adjustment via differential basis support

Elizabeth L. Ogburn, Johns Hopkins University

Anik Burman, Abhirup Datta, Elizabeth L. Ogburn

Unmeasured spatial confounding is a well-known problem in settings such as environmental epidemiology where interest is in estimating the causal effect of a spatially structured exposure. We develop a framework for causal effect estimation in spatial regression models with unmeasured spatial confounding, based on a basis expansion of the spatial component of the exposure and unmeasured confounder. Our framework generalizes previous proposals for spatial confounding in two crucial ways: we allow both the exposure and the confounder to contain spatial and non-spatial variation, and the variation that is unique to the exposure need not occur at high frequencies. We require only that some of the basis functions in the expansion are unique to the exposure. We propose a plurality-rule estimator that identifies the causal effect without prior knowledge of which basis functions are unique to the exposure. Our approach is grounded in robust instrumental variable methodology and is supported by theoretical guarantees for consistency and valid inference. Through extensive simulations and a real data application, we demonstrate that our method reliably recovers unbiased causal estimates in settings in which previous methods would fail.

Synthetic control with weight uncertainty: robust identification and statistical inference

Taehyeon Koo, Columbia University

Taehyeon Koo (Columbia University); Zijian Guo (Zhejiang University)

The synthetic control method estimates causal effects by comparing a treated unit with weighted controls matched on its pre-treatment trajectory. However, validity can be compromised when highly correlated controls leave the synthetic weights weakly determined or when treated–control relationships shift after treatment. We propose a new estimand, the weight-robust treatment effect, defined as the optimizer of a worst-case optimization problem over an uncertainty class of synthetic weights compatible with the pre-treatment fit. We establish its connection to sensitivity analysis: the uncertainty class induces an interval of plausible treatment effects, and the proposed estimand is the point in the interval closest to zero. Under the classical identification conditions, the

estimand coincides with the true treatment effect. When these conditions fail, the estimand remains point identified; if the uncertainty class contains the true post-treatment weight, it provides a conservative sign-preserving bound on the effect. The estimator of this target may have a non-normal limiting distribution, and we propose a novel perturbation-based method for constructing valid confidence intervals. Our proposal may be of independent interest for partial identification and sensitivity analysis, where estimators defined through constrained optimization can have non-standard limiting distributions

Dependence aware linear regression

Rajarshi Mukherjee, Harvard University

Alexander Mercier (Harvard University); James Robins (Harvard University); Rajarshi Mukherjee (Harvard University); Thomas Kolokotrones (Harvard University)

Classical large-sample theory for linear regression assumes independent observations, an assumption frequently violated by interdependent data from network interference or shared exposures. This chapter introduces a dependence-aware framework that explicitly models these interconnected structures through an underlying system of auxiliary variables. Under this framework, we establish general conditions for Local Asymptotic Normality in dependent settings, proving that maximum likelihood estimators remain consistent and asymptotically efficient, and that asymptotically valid confidence intervals are available. To achieve analytical tractability, we introduce techniques that cleanly separate the stochastic variation of the underlying variables from the deterministic structure that introduces dependence among observations. This structural approach yields explicit variance bounds and systematically resolves the asymptotic properties of complex models, including walk-weighted network regressors and shared-source exposures. Ultimately, this work provides a rigorous foundation for optimal asymptotic inference when data exhibit complicated structural dependence.

When the causal graph is unknown, embrace model pluralism

Rohit Bhattacharya, Williams College

Rohit Bhattacharya (Williams College); Junhui Yang (University of Massachusetts Amherst); Ted Westling (University of Massachusetts Amherst); Youjin Lee (Brown University); Ina Ocelli (Williams College)

A fundamental challenge in observational causal inference is the correct specification of a causal model. When identification assumptions are uncertain, analysts may seek to triangulate effect estimates across multiple candidate models that rely on different assumptions. Yet principled methods for quantitative triangulation remain underdeveloped. We develop two complementary frameworks to address this gap. The first concerns testing the causal null hypothesis and yields a procedure that is valid whenever at least one candidate model is correctly specified. The second focuses on effect estimation and produces a combined estimator that is consistent provided at least one model is both correct and empirically testable. Together, these frameworks formalize robustness under methodological pluralism without relying on model selection or agreement across models—an assumption often implicit in the triangulation literature. Throughout the talk, I will also weave in discussions on theoretical and practical challenges that arose (and some that still remain) in implementing such frameworks.

Parallel Sessions 7: Tuesday, November 3, 2026, 10:30 am–12:00 pm

Session 37: Valid Uncertainty Quantification in Modern Statistical Learning

Location: Salon C

Organizer: Yuan Zhang (Ohio State University)

Chair: Yuan Zhang (Ohio State University)

Presenters: Tianmin Xie; Shujie Ma; Yu Gui; Yuetian Luo

Conformal inference for open-set and imbalanced classification

Tianmin Xie, University of Southern California

Tianmin Xie (University of Southern California); Yanfei Zhou (University of Southern California); Ziyi Liang (University of California, Irvine); Stefano Favaro (University of Torino); Matteo Sesia (University of Southern California)

This paper presents a conformal prediction method for classification in highly imbalanced and open-set settings, where there are many possible classes and not all may be represented in the data. Existing approaches require a finite, known label space and typically involve random sample splitting, which works well when there is a sufficient number of observations from each class. Consequently, they have two limitations: (i) they fail to provide adequate coverage when encountering new labels at test time, and (ii) they may become overly conservative when predicting previously seen labels. To obtain valid prediction sets in the presence of unseen labels, we compute and integrate into our predictions a new family of conformal p-values that can test whether a new data point belongs to a previously unseen class. We study these p-values theoretically, establishing their optimality, and uncover an intriguing connection with the classical Good–Turing estimator for the probability of observing a new species. To make more efficient use of imbalanced data, we also develop a selective sample splitting algorithm that partitions training and calibration data based on label frequency, leading to more informative predictions. Despite breaking exchangeability, this allows maintaining finite-sample guarantees through suitable re-weighting. With both simulated and real data, we demonstrate our method leads to prediction sets with valid coverage even in challenging open-set scenarios with infinite numbers of possible labels, and produces more informative predictions under extreme class imbalance.

Learning what to share: adaptive spectral integration and clustering of multimodal data

Shujie Ma, University of California, Riverside

Shujie Ma (University of California, Riverside)

Multimodal data provide complementary views of the same subjects and can reveal population structures that are difficult to detect from any single source. Effective integration, however, requires more than pooling features because modalities may differ in signal strength, noise level, and relevance while containing both shared and modality-specific information. This talk presents MUSIC (Multimodal Spectral Integration and Clustering), an adaptive framework that extracts a low-dimensional singular subspace from each modality and decomposes it into common and modality-specific components. MUSIC iteratively estimates the common subspace and reweights modalities according to their agreement with the evolving common structure, thereby downweighting heterogeneous or uninformative sources while preserving modality-specific variation. The resulting spectral embeddings recover a coarse clustering shared across modalities together with finer, nested modality-specific refinements. Under a general low-rank model, we establish finite-sample guarantees for common and modality-specific subspace estimation and derive corresponding misclustering bounds. An application to multimodal molecular data from The Cancer Genome Atlas shows that MUSIC identifies clinically meaningful glioma subtypes and interpretable modality-specific refinements. The framework offers a principled approach to learning what information should be shared and how strongly each modality should be trusted.

Adaptive discovery of effect modification in matched observational studies

Yu Gui, Wharton School, University of Pennsylvania

Yu Gui, Dylan S Small, Zhimei Ren

Understanding effect modification – how treatment effects vary across subpopulations – is practically important in observational studies, as it helps identify which subgroups are likely to benefit from a given treatment. In this paper, we study the discovery of effect modification in matched observational studies, where each treated unit may be matched to multiple controls. We develop a finite-sample valid procedure for identifying and selecting covariate-interpretable subgroups, with exact control of the subgroup-level false discovery rate (FDR). Our method explicitly accounts for unmeasured confounding via sensitivity models, and leverages multiple matched controls to improve statistical power. We demonstrate the favorable performance of our method relative to baseline methods through extensive simulation studies and a real-world application to the economic returns to college education.

Adaptive robust confidence sets in high dimensions

Yuetian Luo, Rutgers University

Ilias Diakonikolas (University of Wisconsin–Madison); Chao Gao (University of Chicago); Daniel M. Kane (University of California, San Diego); Yuetian Luo (Rutgers University)

This talk studies adaptive confidence sets for multivariate Gaussian means under Huber contamination when the contamination proportion is unknown. We show that an information-theoretically optimal adaptive confidence set can be constructed by simultaneously quantifying uncertainty across all one-dimensional projections. The resulting confidence set achieves the optimal diameter and reveals a substantial adaptation cost compared with the minimax estimation error.

While statistically optimal, the proposed confidence set is computationally infeasible. We establish a Statistical Query (SQ) lower bound showing that any adaptive robust confidence set with shrinking diameter must either use exponentially many queries or require query tolerance corresponding to quasi-polynomial sample and computational complexity. Thus, in contrast to robust mean estimation, adaptation to the unknown contamination proportion incurs not only a statistical cost but also a computational cost.

Finally, we propose a quasi-polynomial-time adaptive confidence set based on low-degree moment matching. The procedure achieves a diameter matching the SQ lower bound up to constant factors. Together, these results characterize the statistical and computational limits of adaptive robust confidence sets under Huber contamination.

Session 91: AI Foundations: Statistical Principles for Modern Learning and Large-Scale Models

Location: Williamsburg

Organizer: Xiwei Tang (University of Texas at Dallas)

Chair: Xiwei Tang (University of Texas at Dallas)

Presenters: Xiao Wang; Rui Duan; Zhengling Qi; Emily C. Hector

Coreset-induced conditional velocity flow matching

Xiao Wang, Purdue University

Zihua She, Jianxi Su, Xiao Wang

Flow matching drives many of today's generative models. The standard recipe is simple: learn the average velocity that carries noise to data, then integrate it. But averaging discards structure. When a point could plausibly move toward several destinations, the mean velocity points somewhere between them, and the multimodal geometry of the transport is lost. This talk asks what happens if we keep that structure. Rather than regressing the conditional mean velocity, CCVFM models the full conditional law of the velocity $V = X1 - X0$. What makes this tractable is a surrogate: replace the target by an entropy-regularized coreset, a Gaussian mixture with far fewer components than

there are data points, and the induced conditional velocity law becomes available in closed form. A Sinkhorn-anchored coupling then pairs source and data velocities so that the surrogate marginal is preserved exactly and pairs concentrate on matched mixture components. The construction also closes a gap between theory and practice. Known rates for flow matching leave slack in the exponent, and the proofs rely on a diffusion-style path with the time interval partitioned into subintervals, each handled by a separately trained network. Deployed flow matching does none of this. We obtain the exact minimax exponent using a single global network on a linear path, with no time partition. Experiments on MNIST, CIFAR-10, ImageNet-32, and CelebA-HQ confirm the predicted regime.

Unsupervised aggregation of multiple learning algorithms

Rui Duan, Harvard T.H. Chan School of Public Health

Rui Duan

Across various domains, the growing advocacy for open science and open-source machine learning has made an increasing number of models publicly available. These models allow practitioners to integrate them into their own contexts, reducing the need for extensive data labeling, training, and calibration. However, selecting the best model for a specific target population remains challenging due to issues like limited transferability, data heterogeneity, and the difficulty of obtaining true labels or outcomes in real-world settings. In this paper, we propose an unsupervised model aggregation method, U-aggregation, designed to integrate multiple pre-trained models for enhanced and robust performance in new populations. Unlike existing supervised model aggregation or super learner approaches, U-aggregation assumes no observed labels or outcomes in the target population. Our method addresses limitations in existing unsupervised model aggregation techniques by accommodating more realistic settings, including heteroskedasticity at both the model and individual levels, and the presence of adversarial models. Drawing on insights from random matrix theory, U-aggregation incorporates a variance stabilization step and an iterative sparse signal recovery process. These steps improve the estimation of individuals' true underlying risks in the target population and evaluate the relative performance of candidate models. We provide a theoretical investigation and systematic numerical experiments to elucidate the properties of U-aggregation. We demonstrate its potential real-world application by using U-aggregation to enhance genetic risk prediction of complex traits, leveraging publicly available models from the PGS Catalog.

When is enough enough in self-evolving LLM systems?

Zhengling Qi, George Washington University

Zhengling Qi

Self-evolving large language model (LLM) systems repeatedly propose, evaluate, and incorporate updates to prompts, skills, or other persistent artifacts. Despite their growing effectiveness, these systems typically operate under a predetermined iteration or compute budget, without a principled criterion to determine when further evolution is no longer worthwhile. This can lead to two undesirable consequences: unnecessary computation after performance has saturated and the risk of returning late updates that overfit or exploit the evaluation signal. These issues motivate us to study two fundamental questions: when should a self-evolving system stop, and what should it output once it stops? We address the first by formulating an online sequential testing problem and constructing an anytime-valid restart detector using the per-item paired evaluation outcomes already produced by self-evolving LLM systems. We address the second by formulating a change-point estimation problem and using the estimated transition to select an earlier artifact for output. The resulting procedure is plug-and-play and requires no modification of the underlying self-evolving algorithms. Across two self-evolving frameworks, three LLM model families, and five benchmarks, our method substantially reduces computation costs while maintaining comparable unseen-test performance. For example, on SearchQA with SkillOpt and DeepSeek V4 Flash, our method stops at round 4 rather than the full budget of 40, reducing token usage by 91.6% while achieving 82.43% unseen-test accuracy versus 82.00% under the full-budget run.

Distributed, exactly valid, and efficient possibilistic uncertainty quantification

Emily C. Hector, University of Michigan

Emily C. Hector (University of Michigan); Leonardo Cella (Wake Forest University); Ryan Martin (Purdue University)

Divide-and-conquer methods use large-sample approximations when each block of data is both small enough to facilitate efficient computation and large enough to support approximately valid inferences. When the overall sample size is moderate, no suitable division of the data meets both requirements, hence the inference lacks validity guarantees. Our proposed approach, couched in the inferential model framework, is fully conditional in a Bayesian sense and provably valid in a frequentist sense. The insight is that existing divide-and-conquer approaches use a Gaussianity assumption twice: first in constructing an estimator, and second in approximating its sampling distribution. We retain the first, using a Gaussian working likelihood, but replace the second with a validification step that uses the sampling distributions of the block summaries determined by the model. This latter step, a probability-to-possibility transform, is key to the reliability guarantees enjoyed by our approach, which are uniquely general in the divide-and-conquer literature. In addition to finite-sample validity guarantees, our approach is also asymptotically efficient. Our computational strategy leverages state-of-the-art black-box likelihood emulators. We showcase our method's performance via simulations and an analysis of median PM2.5 in Maryborough during

the 2023 Australian bushfire season.

Session 40: Data Attribution in Statistical Science and AI

Location: Fulton Ferry

Organizer: Weijing Tang (Carnegie Mellon University)

Chair: Weijing Tang (Carnegie Mellon University)

Presenters: Andrew Ilyas; Arian Maleki; Yoonkyung Lee; Kiljae Lee

Data attribution in large-scale predictive models

Andrew Ilyas, Carnegie Mellon University

Andrew Ilyas (CMU); Logan Engstrom (OpenAI); Sung Min Park (Google); Kristian Georgiev (OpenAI); Guillaume Leclerc (OpenAI); Aleksander Madry (Massachusetts Institute of Technology)

Training data has a profound impact on the behavior of large-scale predictive models, but the nature of this impact is often difficult to pin down. In this talk, we present a problem statement called predictive data attribution that allows us to rigorously study the impact of training data on the behavior of large-scale models. We then introduce a few different approaches to this problem, grounded in statistical methodology, and conclude with a few open directions and recent results.

Accurate and efficient data point removal for high-dimensional settings

Arian Maleki, Columbia University

H. Zou, A. Auddy, Y. Kwon, K. Rad, A. Maleki

Consider a model trained with p parameters from n independent and identically distributed observations. To assess a data point's impact on the model, we remove it from the dataset and aim to understand the model's behavior when trained on the remaining data. This scenario is relevant in various classical and modern applications, including risk estimation, outlier detection, machine unlearning, and data valuation. Conventional approaches involve training the model on the remaining data, but these can be computationally demanding. Consequently, researchers often resort to approximate methods. This talk highlights that in high-dimensional settings, where p is either larger than n or at the same order, many approximation methods may prove ineffective. We will present and analyze an accurate approximation method tailored for high-dimensional regimes, elucidating the conditions for its accuracy. In the concluding part of the presentation, time permitting, we will briefly discuss some of the unresolved issues in this domain.

Data influence dynamics under iterative training algorithms

Yoonkyung Lee, The Ohio State University

Xuerong Wang (The Ohio State University); Yoonkyung Lee (The Ohio State University)

Influence functions are well-known tools for tracing model behavior back to the training data and provide a first-order approximation of data influence measures. However, when models are trained using iterative algorithms in a challenging optimization landscape, parameter estimates may not reach exact optima. In these situations, the assumptions of optimality and convexity in canonical influence analysis may lead to a notable discrepancy between the theoretical influence function and its numerical approximations. To overcome this limitation, we propose a new framework for influence analysis that explicitly accounts for the training trajectory and characterizes how data influence evolves during the training process. We apply the proposed framework to analyze influence functions under popular training algorithms, including SGD, SGD with momentum, and proximal gradient descent. This analysis reveals a recursive structure in data influence dynamics induced by iterative optimization. Numerical studies show that our approach consistently provides more accurate estimates of data influence than the canonical approach across various settings, and that examining data influence dynamics enables more fine-grained model diagnostics.

Generalized priority-aware Shapley value

Kiljae Lee, The Ohio State University

Kiljae Lee (The Ohio State University); Ziqi Liu (Carnegie Mellon University); Weijing Tang (Carnegie Mellon University); Yuan Zhang (The Ohio State University)

Credit allocation is a fundamental problem across modern machine learning, arising in applications ranging from feature attribution and data valuation to model evaluation. The Shapley value provides a principled and model-agnostic solution, but its classical formulation averages marginal contributions uniformly over contributor orderings, leaving no mechanism to incorporate contextual information about which contributors should precede others or be treated differently. Motivated by the random-order interpretation of the Shapley value, we incorporate such priority information by modifying the distribution over contributor orderings. We first introduce the Priority-Aware Shapley Value (PASV), which combines hard, acyclic precedence constraints represented by a general DAG with contributor-specific weights. We then generalize PASV to the Generalized Priority-Aware Shapley Value (GPASV) for settings where pairwise priorities are graded or mutually conflicting. GPASV replaces binary acyclic constraints with an arbitrary directed weighted priority graph and penalizes, rather than forbids, priority violations. Both methods admit axiomatic characterizations and scalable Monte Carlo estimation procedures, while GPASV recovers PASV in the hard-penalty limit on acyclic graphs. Finally, we develop priority-sweeping diagnostics for assessing sensitivity to priority specifications, and demonstrate the framework in data valuation, feature attribution, and LLM ensemble valuation under cyclic human-preference graphs.

Session 59: Modern Network Analysis and Applications Across Disciplines

Location: Sunset Park

Organizer: Jingming Wang (University of Virginia)

Chair: Huimin Cheng (Boston University)

Presenters: Zhiling Gu; Nicholas W. Landry; Robert Lunde; Fangting Zhou

Advancing longitudinal AD pathology with bi-directional network systems for protein deposition and brain activities

Zhiling Gu, University of South Carolina

Zhiling Gu (University of South Carolina); Simiao Gao (Augusta University); Tianxi Li (University of Minnesota); Yize Zhao (Yale University)

Alzheimer's disease progression is marked by coupled changes in regional tau pathology and brain network organization, yet most longitudinal models treat these processes separately or impose a one-way propagation mechanism. We propose LABIN, a longitudinal latent dynamical framework that jointly models tau burden and brain connectivity while allowing bidirectional dependence between them over time. The model captures stage-dependent progression, regional heterogeneity, and low-rank latent structure, providing a flexible way to study how connectivity shapes tau spread and how tau, in turn, alters network architecture. In simulations, the proposed estimator recovers key parameter blocks and latent structure under correct specification, and it improves prediction in settings where bidirectional tau-connectivity coupling is present. We then apply LABIN to longitudinal tau PET and functional connectivity data from the Anti-Amyloid Treatment in Asymptomatic Alzheimer's Disease (A4) study. The analysis reveals systematic associations between tau accumulation and network change, supporting the view that tau progression and connectivity disruption co-evolve rather than act independently. These results suggest that jointly modeling protein pathology and brain networks can provide a more informative representation of Alzheimer's disease dynamics and may improve longitudinal characterization of disease progression.

Cartographic characterization of complex contagions

Nicholas W. Landry, University of Virginia

Daniel Kaiser (University of Virginia); Katerina Tang (Cornell University); William Thompson (University of Vermont); Jean-Gabriel Young (University of Vermont); Laurent Hébert-Dufresne (University of Vermont); Nicholas W. Landry (University of Virginia)

Network scientists often use complex dynamic processes to describe network contagions, but tools for fitting contagion models typically assume simple dynamics. I'll address this gap by presenting a nonparametric method to reconstruct a network and dynamics from a series of node states, using a model that breaks the dichotomy between simple pairwise and complex neighborhood-based contagions. I'll then show that a network is more easily reconstructed when observed through the lens of complex contagions if it is dense or the dynamic saturates, and that simple contagions are better otherwise. I'll conclude by talking about the inadequacies of the term "complex contagion".

Conformal prediction for dyadic regression under complex missingness

Robert Lunde, Washington University in St. Louis

Robert Lunde (Washington University in St. Louis), Minjie Yang (Washington University in St. Louis), Elizaveta Levina (Michigan), Ji Zhu (Michigan)

Dyadic regression, which involves modeling a relational matrix given covariate information, is an important task in statistical network analysis. We consider uncertainty quantification for dyadic regression models using conformal prediction. We establish finite-sample validity of our procedures for various sampling mechanisms under a joint exchangeability assumption. Our proof uses new results related to the validity of conformal prediction beyond exchangeability, which may be of independent interest. We also show that, under certain conditions, it is possible to construct asymptotically valid prediction sets for a missing entry under a graphon assumption.

Multivariate probit models for causal discovery with binary data and latent confounders

Fangting Zhou, Florida State University

Fangting Zhou (Florida State University)

Causal discovery from multivariate binary data is challenging in the presence of latent confounding, which can induce spurious associations and obscure causal directions. Motivated by biomedical applications, we propose a causal discovery framework that leverages instrumental variables, such as genetic variants, to identify causal relationships among binary outcomes. We formulate a multivariate probit structural equation model in which binary variables arise by thresholding latent Gaussian variables, while correlated errors accommodate latent confounding. Under a majority-validity condition on candidate instruments, we establish identifiability of the causal structure and structural parameters. To address the computational challenges of multivariate probit models, we develop a scalable two-stage procedure that estimates reduced-form parameters using univariate and bivariate marginal distributions, avoiding high-dimensional Gaussian integration. The reduced-form estimates are then combined with instrumental-variable information to recover the underlying causal structure. We establish consistency of the proposed procedure, providing a computationally efficient approach to causal discovery with binary outcomes and latent confounding.

Session 84: Analyzing Complex-Structured Data with Latent Geometry and Heterogeneity

Location: Golden Boardroom

Organizer: Yinqiu He (University of Wisconsin-Madison)

Chair: Yinqiu He (University of Wisconsin-Madison)

Presenters: Dena Asta; Yun Wei; Lili Zheng; Zachary Lubberts

Quantum-inspired methods for statistical geometry and networks

Dena Asta, The Ohio State University

Dena Asta (The Ohio State University)

Manifold learning and network modeling share a fundamental objective: uncovering lower-dimensional latent structure from complex, high-dimensional observations. While classical statistical geometry relies on smooth local approximations, real-world network data is characterized by discrete interactions, sparsity, and non-Euclidean latent spaces and therefore defies traditional smooth locality. Non-commutative geometry (NCG), an extension of classical differential geometry developed to model non-smooth spaces at the quantum scale, turns out to naturally describe non-smooth network geometry. Rather than building local graph approximations, NCG suggests a global operator-theoretic method of estimating global metric distances across network nodes. In this talk, we introduce the core principles of non-commutative geometry in a statistical network setting, demonstrate how this operator-theoretic framework extends classical latent space models and spectral network inference, compare performance against existing graph-embedding and manifold learning methods, and outline open directions.

Title TBA

Yun Wei, The University of Texas at Dallas

Abstract TBA

Joint semi-symmetric tensor PCA for integrating multi-modal populations of networks

Lili Zheng, University of Illinois Urbana-Champaign

Lili Zheng (UIUC); Jiaming Liu (UChicago); Zhengwu Zhang (UNC); Genevera I. Allen (Columbia)

Multi-modal populations of networks arise in many scenarios including large-scale multi-modal neuroimaging studies that capture both functional and structural neuroimaging data for thousands of subjects. A major research question in such studies is how functional and structural brain connectivity are related and how they vary across the population. Motivated by this, we develop a novel PCA-type framework for integrating multi-modal undirected networks measured on many subjects. Specifically, we arrange these networks as semi-symmetric tensors, where each tensor slice is a symmetric matrix representing a network from an individual subject. We then propose a novel Joint, Integrative Semi-Symmetric Tensor PCA (JisstPCA) model, associated with an efficient iterative algorithm, for jointly finding low-rank representations of two or more networks across the same population of subjects. We establish one-step statistical convergence of our separate low-rank network factors as well as the shared population factors to the true factors, with finite sample statistical error bounds. Through simulation studies and a real data example for integrating multi-subject functional and structural brain connectivity, we illustrate the advantages of our method for finding joint low-rank structures in multi-modal populations of networks.

Spectral estimation in networks with vertex effects

Zachary Lubberts, University of Virginia

Lei Zhang (University of Virginia); Zachary Lubberts (University of Virginia)

Existing statistical analysis of spectral methods for random networks relies heavily on low rank and conditional independence assumptions that may not hold in real data. We formulate a “vertex effects" model where edges incident on the same vertex can be dependent on one another, and study the effect on the spectrum as this level of dependence changes. We show that the top eigenvectors of the adjacency matrix, while linearly orthogonal, exhibit nonlinear relationships that may frustrate existing clustering methodologies relying on ellipsoidal point distributions (such as the ubiquitous K-means and Gaussian Mixture Modeling), and propose alternative approaches.

Session 54: Learning from Real-World Health Data: Methods for Imbalance, Heterogeneity, Equity, and Real-World Evidence Translation

Location: Carroll Gardens

Organizer: Rui Duan (Harvard University)

Chair: Tian Gu (Columbia University)

Presenters: Anru Zhang; Rebecca Hubbard; Jiayi Tong

Making EHR ready for AI: statistics at the data frontier

Anru Zhang, Duke University

Anru Zhang (Duke University)

Electronic health records (EHR) are increasingly central to clinical research and AI, but they are not simply "big data." They are irregular, heterogeneous, imbalanced, privacy-sensitive, and sometimes internally inconsistent. Reliable statistical learning therefore often begins with making the data themselves more comparable, learnable, and reliable.

In this talk, I will highlight several examples from our group on making EHR more suitable for analysis and AI. These include aligning longitudinal patient trajectories through timeline registration, using synthetic data to address imbalance and privacy, and leveraging large language models to identify inconsistencies in clinical records. Together, these examples motivate a broader view of data

readiness as a statistical problem in its own right—and as an important foundation for reliable clinical AI.

Methodological challenges for equity and population health in risk-guided cancer screening

Rebecca Hubbard, Brown University

Rebecca Hubbard (Brown University)

Historically, cancer screening recommendations have been defined for broad population subgroups, such as age-based guidelines for breast cancer screening. More recently, "risk-guided" cancer screening has been proposed as an alternative paradigm, in which screening is tailored to an individual's predicted risk. This approach could potentially improve population health by focusing intensive screening on those most likely to benefit. However, risk-guided screening relies on models trained on electronic health record data, raising concerns about transportability and algorithmic fairness. These concerns are especially salient when models developed in data-rich settings are used to guide clinical decisions for underrepresented groups. In this talk, I will use breast cancer screening as a motivating example to illustrate the promise and pitfalls of risk-guided screening. I will discuss how distribution shifts, model misspecification, and choices made during model development and evaluation can differentially affect performance across populations. I will also highlight how seemingly well-intentioned approaches to promoting fairness in risk prediction can have unintended consequences when translated into real-world screening policies. I will conclude by outlining open methodological questions at the intersection of statistical learning and population health, and by discussing principles for developing and evaluating risk prediction models that support equitable implementation of risk-guided screening.

Unlocking multi-institutional insights into disease progression using federated learning on longitudinal electronic health records

Jiayi Tong, Johns Hopkins University

Yicheng Shen (JHU), Ji Soo Kim (JHU), Chongliang Luo (WashU), Scott L. Zeger (JHU), Robyn T. Domsic (UPMC), Ami A. Shah (JHU), Jiayi Tong (JHU)

Multisite longitudinal analyses of electronic health record (EHR) data can reveal real-world disease trajectories, but privacy constraints, high communication costs, outcome-specific missingness, and cross-site heterogeneity often prevent principled pooling and standard longitudinal modeling. In response, we propose PEAL (Privacy-preserving Efficient Aggregation for Longitudinal data), a single-round federated algorithm for multi-level linear mixed-effects models that yields estimates identical to pooled individual-level analysis. We then extend to MV-PEAL, a one-shot federated framework for multivariate mixed-effects models that reconstructs the global likelihood from pattern-specific summaries, enabling efficient estimation of fixed effects, covariance components, and cross-outcome correlations while accommodating outcome missingness in a secure manner. Extensive simulations show gold-standard-equivalent accuracy and improved efficiency over complete-case and standard imputation strategies. By applying the methods to the Johns Hopkins and University of Pittsburge systemic sclerosis cohorts, our methods recover clinically plausible single- and multi-biomarker trajectories, illustrating their utility for distributed research networks studying rare diseases and other time-evolving outcomes.

Session 22: Causal Mechanisms and Inference in Complex Data

Location: Coney Island

Organizer: Panpan Zhang (Vanderbilt University Medical Center)

Chair: Jun Yan (University of Connecticut)

Presenters: Dongliang Zhang; Michael J. Daniels; Yi Zhao; Richard Liu

Analysis of shape and texture features in longitudinal PET images via multilevel functional regression using burden-aware topological representation

Dongliang Zhang, Brown University

Dongliang Zhang (Brown University); Lorin Crawford (Microsoft); Ani Eloyan (Brown University)

Alzheimer's disease (AD) is a brain disorder marked by cerebral atrophy and abnormal deposition of amyloid plaques and tau tangles. Magnetic resonance imaging assesses cerebral atrophy, while positron emission tomography imaging measures spatial patterns of tau deposition where its estimation is key to understand AD pathology. The smooth Euler characteristic transform (SECT) has been recently proposed to uniquely represent statistical shapes as continuous curves. Yet, existing applications of SECT focus on representing shape features while texture representation is unexplored. We present a novel approach for statistical inference on both texture and shape features using a combined weighted SECT and function-on-scalar regression modeling. Specifically, we model texture features by a three-way model. With simulated studies, we propose resampling-based inference to study the effects of predictors, offering insights on the spatial relationship between brain and predictors for AD patients. We show differences in atrophy and tau patterns between cognitively normal and AD subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

A Bayesian nonparametric approach for semi-competing risks with application to cardiovascular health

Michael J. Daniels, University of Florida

Karina Gelis Cadena (University of Florida); Michael J. Daniels (University of Florida); Juned Siddique (Northwestern University)

We address causal estimation in semi-competing risks settings, where a non-terminal event may be precluded by one or more terminal events. We define a principal

stratification causal estimand for treatment effects on the non-terminal event, conditional on surviving past a specified landmark time. To estimate joint event time distributions, we use both vine copula constructions and Bayesian nonparametric enriched Dirichlet process mixtures (EDPM), enabling inference under minimal parametric assumptions. We index our causal assumptions with sensitivity parameters. We illustrate the proposed method using data from a cardiovascular health study.

Causal mediation analysis with a sparse single-index model

Yi Zhao, Indiana University School of Medicine

Yi Zhao (Indiana University School of Medicine); Siyuan Ping (Indiana University School of Medicine); Wanzhu Tu (Indiana University School of Medicine); Fan Li (Yale University)

We study causal mediation analysis in the presence of multiple mediators and pre-treatment covariates using a semiparametric single-index outcome model. The proposed model assumes that multiple mediators jointly affect the outcome through one underlying mediator index, represented as a weighted combination of individual mediators with an unknown smooth relationship to the outcome. This formulation accommodates high-dimensional mediators while preserving interpretability, because the index weights quantify the relative contributions of individual mediators to the indirect pathway. To identify the most relevant mediators, we incorporate a sparsity-inducing penalty that enables automatic mediator selection. Within the potential-outcomes framework, we define natural direct and indirect effects and present identification assumptions that extend standard mediation assumptions to the covariate-adjusted setting. We further develop two multiply robust estimators: one that retains the flexibility of unrestricted nonparametric efficiency theory, and another that exploits the single-index structure to improve efficiency and stability when the number of mediators is large. We establish consistency, asymptotic normality, and valid inference for the proposed index and causal effect estimators, as well as selection consistency for identifying active mediators under suitable tuning.

General target machine learning for modern causal mediation analysis

Richard Liu, New York University Grossman School of Medicine

Richard Liu (New York University Grossman School of Medicine); Nicholas T. Williams (University of California, Berkeley); Kara E. Rudolph (Columbia University); Iván Díaz (New York University Grossman School of Medicine)

Causal mediation analyses investigate the mechanisms through which causes exert their effects, and are therefore central to scientific progress. The literature on the non-parametric definition and identification of mediational effects in rigorous causal models has grown significantly in recent years, and there has been important progress to address challenges in the interpretation and identification of such effects. However, statistical methodology for non-parametric estimation has lagged, with few or no methods available for tackling non-parametric estimation in the presence of multiple, continuous, or high-dimensional mediators. In this paper we propose an all-purpose one-step estimation algorithm that can be coupled with machine learning in mediation studies that use any of six common mediation parameters (natural direct and indirect effects, randomized interventional effects, organic direct and indirect effects, path-specific effects based on recanting twins, path-specific effects based on separable effects, and decision theoretic effects). Our estimators are constructed by re-parameterizing the identification formulas in terms of sequential regressions, constructing a first order non-parametric von-Mises approximation of the first-order bias of a plug-in estimator, and leveraging Riesz learning for estimation of nuisance parameters. We show that the proposed one-step estimators have desirable properties, such as $\sqrt{n}$-convergence and asymptotic normality. We illustrate the properties of our methods in a simulation study and demonstrate its use on real data to estimate the extent to which pain management practices mediate the total effect of having a chronic pain disorder on opioid use disorder.

Session 88: Advances in Methods for Leveraging External Information in Precision Medicine

Location: Park Slope

Organizer: Nicholas Henderson (University of Michigan)

Chair: Nicholas Henderson (University of Michigan)

Presenters: Cathy Shyr; Peijin Wang; Nicholas Henderson; Daoyuan Lai

Human-guided in-context policy learning for rare disease diagnosis using LLMs

Cathy Shyr, Vanderbilt University Medical Center

Minh-Ha Nguyen (Vanderbilt University); Cathy Shyr (Vanderbilt University Medical Center)

Rare disease diagnosis is a small-sample generalization problem in which diagnoses are often n-of-1. Expert diagnostic reasoning is valuable but difficult to transfer and scale, while pretrained large language models (LLMs) contain broad biomedical knowledge yet remain unreliable without structured guidance. We introduce Policy Iteration with Human Feedback (PIHF), an in-context policy-learning framework that adapts a frozen LLM by iteratively revising an explicit, versioned natural-language policy and tool set rather than modifying model weights. A model critic and clinical experts review complete reasoning and tool-use trajectories, identify recurrent reasoning failures, propose targeted revisions, and admit or roll back candidate policies. We operationalize PIHF in liteOdyssey, our agentic rare disease diagnostic system. From 50 development

cases, PIHF increased disease Recall@1 from 26.5% to 59.3% across 1,243 public benchmark cases spanning 722 diseases, with comparable gains across 679 diseases unseen during development. The learned policy transferred across proprietary and open-weight LLMs. After adaptation on 50 real-world Undiagnosed Diseases Network cases, the policy also improved diagnosis in 465 independent, hold-out patients based on blinded physician review. These results show that explicit in-context policies can convert scarce expert feedback into reusable diagnostic strategies that generalize across rare diseases and LLMs without weight updates.

Bayesian adaptive design with sample size re-estimation for hybrid controlled trials

Peijin Wang, Duke University

Peijin Wang (Duke University); Hwanhee Hong (Duke University)

Hybrid controlled trials (HCTs), which augment a randomized controlled trial (RCT) with external controls (ECs), are increasingly considered in innovative drug development, particularly in rare diseases. Bayesian dynamic borrowing methods are widely used because they discount EC information according to its conflict with the trial data, improving estimation efficiency while limiting bias. Although these methods are well studied for estimation, their role in trial design, particularly sample size determination, remains underexplored. We propose a Bayesian adaptive design for HCTs that uses Bayesian power calculation to re-estimate the required trial sample size from the observed trial and EC data, accounting for heterogeneity between the trial and EC populations. Simulation studies show that the design can reduce trial sample size compared with a conventional RCT when heterogeneity is small. An application to chronic lung allograft dysfunction illustr ates planning an adaptive HCT and offers guidance on choosing design parameters to balance sample size savings against desired operating characteristics.

Ranking-based data integration for robust outcomes prediction and biomarker detection

Nicholas Henderson, University of Michigan

Nicholas Henderson

Utilizing established risk factors, prognostic models, or other external predictions often improves the construction of a newer risk model that uses novel biomarkers in a smaller, internal study. However, directly borrowing information from an established prognostic model is often unsuitable due to differences in study populations, patient outcomes measured, and other specific features of the internal study design. To better enable the use of external prognostic information, we propose an estimation approach centered around the idea that the risk rankings rather than the risk scores are often more transportable to the internal study context. To leverage external ranking information, we outline a penalized regression approach that introduces the ranking parameters associated with the regression coefficients of an internal risk model and estimates the internal risk model by penalizing a ranking-based discrepancy measure between the ranking parameters and the external model rankings. Simulation studies demonstrate that our method has competitive predictive performance and performs particularly well when the true internal and external models have high rank correlation but large discrepancies between their risk scores. We illustrate the use of this ranking-based data integration method through the development of prognostic models for advanced prostate cancer patients treated with an immune checkpoint inhibitor.

Bayesian transfer learning for enhanced estimation and inference

Daoyuan Lai, Columbia University

Daoyuan Lai (Columbia University); Oscar Hernan Madrid Padilla (University of California, Los Angeles); Xiang Li (The University of Texas MD Anderson Cancer Center); Tian Gu (Columbia University)

Transfer learning enhances model performance in a target population with limited samples by leveraging knowledge from related studies. While many works focus on improving predictive performance, challenges persist in statistical inference. Bayesian approaches naturally provide uncertainty quantification for parameter estimates; however, existing Bayesian transfer learning methods are typically limited to single-source scenarios or require individual-level data. We introduce TRansfer leArning via guideD horseshoE prioR (TRADER), a novel approach enabling multi-source transfer through pre-trained models in high-dimensional linear regression. TRADER shrinks target parameters toward an adaptively weighted average of source estimates, while remaining robust to differences in source scale and correlation. Theoretical investigation shows that TRADER achieves faster posterior contraction rates than standard continuous shrinkage priors when sources are well-aligned with the target, while preventing negative transfer from heterogeneous sources. The finite-sample marginal posterior behavior of TRADER is established. Extensive simulations show that TRADER achieves inference performance no worse than using target data alone, performs comparably to a competing frequentist method despite using only summary-level source data, and offers substantial computational advantages. Application to a high-dimensional genetic dataset further shows TRADER's effectiveness in inference under strong multicollinearity.

Session 36: Network Interference, Spillover and Temporal Effects

Location: Windsor Terrace

Organizer: Keith Levin (University of Wisconsin-Madison)

Chair: Keith Levin (University of Wisconsin-Madison)

Presenters: Emma Jingfei Zhang; Alex Hayes; Vardis Kandiros; Sharmodeep Bhattacharyya

Estimating treatment and spillover effects with the ego-cluster experimental design

Emma Jingfei Zhang, Emory University

In this talk, we introduce a new experimental design for the estimation and inference of global treatment effect and spillover effect under a model-based framework and ego-cluster randomization. Under this design, the network is partitioned into a collection of ego-clusters, each consisting of a focal unit (the ego) and its network neighbors (the alters), with randomization conducted at the cluster level. We propose model-based estimators for the global treatment effect and spillover effect and establish their consistency and asymptotic normality, with asymptotic variances determined by the ego-cluster structure. Building on these theoretical results, we introduce an ego-clustering algorithm that sequentially selects egos and assigns alters to minimize asymptotic variances.

Estimating peer effects in noisy, low-rank networks via network smoothing

Alex Hayes, Oregon State University

Alex Hayes (Oregon State University), Keith Levin (UW-Madison)

Peer effect estimation requires precise network measurement, yet most empirical networks are noisy, rendering standard estimators inconsistent. To address measurement error in networks, we propose a method to estimate peer effects in networks whose expected adjacency matrix is low-rank. Our key result shows that peer effects over a true unobserved network are asymptotically equivalent to peer effects over the expected adjacency matrix. This result reduces peer effect estimation in noisy networks to low-rank matrix estimation targeting the expected adjacency matrix. We develop our theory for weighted networks observed with additive noise, but simulations suggest approach can be applied more generally when there is a low-rank estimation method suited to a particular noise structure. We demonstrate via simulations that our approach applies to egocentric samples, aggregated relational data, and networks with missing edges, each requiring a different low-rank estimation method.

A design-based minimax theory for network experiments

Vardis Kandiros, Columbia University

Vardis Kandiros (Columbia University); Christopher Harshaw (Columbia University); Fredrik Savje (Uppsala University)

Network experiments are used throughout the social and medical sciences to investigate causal effects under the presence of interference. While a large body of work has developed improved statistical procedures, the fundamental limits of statistical estimation in these settings is less well understood. In this paper, we develop and investigate a design-based theory of minimax risk for network experiments under an arbitrary neighborhood interference model. Our notion of minimax risk describes the optimal precision among all statistical procedures for investigating a particular causal effect on the observed interference network. We show that the minimax risk is a function of the corresponding conflict graph, which captures inherent unobservability of estimand-relevant potential outcomes given the observed interference network. Our main contribution is a series of upper and lower bounds on the minimax rate in terms of local and global connectivity properties of the conflict graph. To illustrate their utility, we apply these general results to obtain minimax analyses for two commonly studied effects: the direct treatment effect and global average treatment effect.

A causal inference of interference on a continuous disease transmission network

Sharmodeep Bhattacharyya, Oregon State University

Jae Young Hwang (USDA), Dave Gent (USDA), Shirshendu Chatterjee (CUNY), Tom Marsh (Washington State), Trent VanHawkins (Oregon State)

Accounting for bias caused by interference, or spillover, is important for accurate estimation of causal effects when untreated individuals are connected to a response through a physical or social process. This is a difficult computational problem when treatments are continuous, such as with the intensity of disease management in agricultural fields. Using a well-described dataset of powdery mildew of hop (Podosphaera macularis), we estimate causal exposure-response functions that describe how growers' monthly fungicide inputs and costs change in response to the incidence of disease conditional on disease spillover from an inferred disease spreading network. Bayesian zero-inflated beta-binomial models were fit for two time transitions to estimate the probability of disease in a given yard as a nonlinear function of autoinfection due to preexisting disease, the number of fungicide applications, and disease spread from other yards as influenced by their source strength, fungicide applications, distance, and wind run and direction in the preceding month. Interference from disease in potential source yards was then estimated by summing incoming directed edge weights after covariate weighting by a generalized propensity score. Finally, we estimate the population-average overall exposure-response function marginalized over the distribution of spillover effect after passing covariate balance for both time transitions. We show that failure to account for interference can lead to substantial

misestimation and bias in exposure-response functions when field-to-field disease spread drives epidemics. In our empirical application, accounting for spillover was especially critical when inoculum was broadly distributed and dense disease transmission networks were present.

Session 107: Next-Generation Decision Intelligence: From Underwriting and Uplift Modeling to LLM-Driven Statistical Experiment Lifecycle Automation and Clinical Assessment

Location: DUMBO

Organizer: Jieying Jiao (New York Life)

Chair: Jieying Jiao (New York Life)

Presenters: Yangfan Liang; Pei-Yau Lung; Yang Liu

Standardizing experimentation workflows with modular AI copilots

Yangfan Liang, Google

Yangfan Liang (Google); Ian Lipton (Google); Kevin Quach (Google)

Day-to-day experimentation workflows often involve tedious, repetitive steps: writing data extraction SQL queries, calculating sample sizes and test duration, running standard statistical tests, checking distributions, and formatting summary reports. We built a modular LLM copilot that helps analysts automate and standardize these routine steps. In this workflow, the language model parses natural-language questions and coordinates tasks across specialized skills, while standard Python packages (such as scipy and statsmodels) compute the actual statistical numbers. The human analyst stays in the loop throughout the process to review metric definitions, inspect data distributions, check modeling assumptions (especially for observational and causal analyses), and approve final results before sharing. The extracted dataset, the analysis plan (proposed metrics, tests, and distribution checks), and the summary results are saved as files, so a completed analysis is straightforward to re-run and audit. This practical approach standardizes routine experimentation plumbing while keeping all statistical judgment and methodological decisions firmly with human analysts.

Title TBA

Pei-Yau Lung, Humana EnterpriseAI

Abstract TBA

A resampling-based method for uplift model comparison

Yang Liu, Upstart Network, Inc.

Yang Liu (Upstart Network, Inc.); Chaoyu Yuan (Upstart Network, Inc.)

Uplift models play a critical role in modern marketing applications to help understand the incremental benefits of interventions and identify optimal targeting strategies. A variety of techniques exist for building uplift models, and it is essential to understand the model differences in the context of intended applications. The uplift curve is a widely adopted tool for assessing uplift model performance on the selection universe when observations are available for the entire population. However, when it is uneconomical or infeasible to select the entire population, it becomes difficult or even impossible to estimate the uplift curve without appropriate sampling design. To the best of our knowledge, no prior work has addressed uncertainty quantification of uplift curve estimates on the entire population, using only a population subsample. We propose a two-step sampling procedure and a resampling-based approach to compare uplift models with uncertainty quantification, examine the proposed method via simulations and real data applications, and conclude with a discussion.

Session 60: Robust Causal Inference Under Real-World Complications: Extrapolation, High Dimensions, and Missing Data

Location: Greenpoint

Organizer: Siyu Heng (New York University)

Chair: Siyu Heng (New York University)

Presenters: Gary Hettinger; Hyunseung Kang; Thomas Leavitt; Zeyang Yu

Semi-parametric causal inference under required extrapolation: stationarity, positivity violations, and time series designs

Gary Hettinger, New York University Grossman School of Medicine

Gary Hettinger (New York University Grossman School of Medicine)

Semi-parametric causal inference typically relies on assumptions that connect causal estimands to directly observable data, enabling non-parametric estimation of nuisance functions and the construction of doubly robust estimators. However, many important problems such as positivity violations, transportability across populations, and controlled interrupted time series, require extrapolation beyond the observed data support, making fully non-parametric identification impossible. In these settings, part of the data-generating process must be specified parametrically, often through assumptions such as stationarity or stable trends, while other components remain non-parametrically estimable.

In this work, we formulate such problems as a semi-parametric estimation problem under required extrapolation and provide formal theory to what is done often implicitly in practice. Further, we propose a unified semi-parametric framework that separates parametric extrapolation components from nonparametric confounding mechanisms by deriving orthogonal estimating equations for target causal parameters and characterizing the resulting efficiency-robustness tradeoffs.

The framework is illustrated in settings with positivity

violations due to covariate drift and controlled interrupted time series designs. Our results clarify what robustness and efficiency are achievable under extrapolation and provide practical guidance for causal inference in modern observational studies where identification relies on structural assumptions such as stationarity.

Representation-powered sensitivity analysis with applications to outcome-wide observational epigenomic studies

Hyunseung Kang, University of Wisconsin–Madison

Kwangmoon Park (University of Pennsylvania); Hyunseung Kang (University of Wisconsin–Madison); Kirk J. Hogan (University of Wisconsin–Madison); Reid S. Alisch (University of Wisconsin–Madison); Sunduz Keles (University of Wisconsin–Madison)

When testing for treatment effects in large-scale, observational (i.e., non-randomized) genomic studies, investigators must address two important challenges: (i) bias from unmeasured confounders and (ii) multivariate outcomes that exhibit shared, biologically meaningful low-dimensional structure. A sensitivity analysis for unmeasured confounding quantifies the impact of unmeasured confounders, but it has been limited to scalar outcomes or multivariate outcomes without a low-dimensional structure. This work presents representation-powered sensitivity analysis, a novel test for both global and subgroup treatment effects under unmeasured confounding when the outcomes exhibit a low-dimensional structure. As long as the dimension reduction method used to discover the low-dimensional structure satisfies high-level rate conditions, the proposed test guarantees Type I error or false discovery rate control, and many off-the-shelf dimension reduction methods satisfy these rates. Also, the proposed test has non-trivial power to detect a wide variety of treatment effects, including dense, sparse, and small treatment effects. We conclude by investigating the effect of amyloid-β (i.e., treatment) on multiple DNA methylation regions (i.e., outcomes) from an observational epigenomic dataset. Our results discover DNA methylation regions that are significantly influenced by amyloid-β levels even in the presence of unmeasured confounders and these discoveries align with in-vitro experiments.

Randomization-based guarantees for Bayesian causal inference

Thomas Leavitt, City University of New York (CUNY)

Thomas Leavitt

A Bayesian analysis of a randomized experiment can rest on the known assignment mechanism rather than on a model of the potential outcomes. Recent work builds a posterior for the average treatment effect from a Normal working likelihood for the difference-in-means and shows that the posterior has Bayesian analogues of unbiasedness, under a flat prior, and of consistency, under any proper or flat prior that is continuous and positive at the true effect. I identify conditions on the working likelihood, the statistic, and the finite-population estimand under which these guarantees hold; the conditions on the working likelihood concern only general features of its shape, such as a unique mode and vanishing tails, and do not require Normality. Coverage, however, favors the original construction, whose Normal working likelihood yields the shortest credible intervals among working likelihoods guaranteed at least nominal coverage in large samples. I then analyze model-based Bayesian inference from the same randomization-based principles. I show that, in large samples, the posterior from a regular parametric outcome model has the same form as a randomization-based posterior from a working likelihood, so the same conditions apply. When the outcome model is misspecified, the model-based posterior still concentrates on the true effect if each arm is modeled independently by a linear exponential family, such as the Normal or Gamma. However, under standard models for skewed outcomes, such as the Weibull and lognormal, the posterior can concentrate on the wrong effect, even one of the wrong sign. Finally, I discuss what these results imply for choosing between randomization-based and model-based Bayesian inference and for using experiments to evaluate structural theories that generate distributions of potential outcomes.

Randomization inference with sample attrition

Zeyang Yu, Princeton University

Xinran Li (University of Chicago); Peizan Sheng (University of Chicago); Zeyang Yu (Princeton University)

Randomization inference is a widely-used and appealing approach for analyzing treatment effects in randomized experiments, as it is finite-sample valid and does not require any distributional assumptions. However, naive application of randomization inference may suffer from severe size distortion in the presence of sample attrition, where outcome data are missing for some units. In this paper, we propose new, computationally efficient methods for randomization inference that remain valid under a broad class of potentially informative missingness mechanisms, allowing a unit's missingness to depend on its (unobserved) potential outcomes. Specifically, we construct valid p-values for testing both sharp and bounded null hypotheses on treatment effects via a worst-case consideration of the classical Fisher randomization test. Leveraging distribution-free test statistics, these worst-case p-values admit closed-form solutions. Importantly, by incorporating both potential outcomes and potential missingness indicators into the test statistic, our methods can exploit structural assumptions such as monotone missingness, which are commonly adopted in applications due to their plausibility and ability to substantially improve inferential power. Moreover, our approach connects to a range of partial identification bounds in the literature, which in some sense suggests the sharpness of our tests. We illustrate the proposed methods through both simulation studies and an empirical application. An R package implementing the proposed methods is publicly

available.

Session 27: Advances in Statistical Learning for Real-World Evidence

Location: Navy Yard

Organizer: Rui Duan (Harvard University)

Chair: Jiwei Zhao (University of Wisconsin–Madison)

Presenters: Chixiang Chen; Sebastien Haneuse; Xu Shi

A new learning framework for integrating multiple auxiliary outcomes into primary data analysis

Chixiang Chen, University of Maryland, School of Medicine, Baltimore

Chixiang Chen (Presenter, University of Maryland); Daxuan Deng (Johns Hopkins University); Jiyuan Jiao (University of Maryland); Peisong Han (Gilead Sciences)

In addition to primary outcomes, auxiliary outcomes, including biomarkers, omics measures, imaging data, and other secondary endpoints, are increasingly available in contemporary biomedical research. Such information may arise from secondary outcomes collected in clinical trials (Source 1) or from composite measures and risk-prediction scores constructed using real-world data (Source 2). These auxiliary outcomes have substantial potential to improve estimation and inference for primary outcome analysis, yet their effective integration raises important statistical challenges. This talk will summarize key challenges in leveraging auxiliary outcomes for primary outcome analysis using both real-world data and clinical trial data, and will introduce a series of recently developed statistical learning approaches to address them. We will begin with an early-generation approach that incorporates prespecified prior information to guide auxiliary-outcome integration (Tool 1.0), followed by a more flexible and robust framework that enables information integration without requiring such prior information (Tool 2.0). Ongoing work extending these ideas toward broader learning from clinical trial data (Tool 3.0) will be presented at the end of the talk. Together, these methods provide promising tools for improving the efficiency and reliability of primary outcome analysis in the era of increasingly rich biomedical data.

Double sampling and semiparametric methods for informatively missing not at random data

Sebastien Haneuse, University of California, Irvine

Missing data arise in almost all applied settings and are ubiquitous in electronic health records (EHR). When data are missing not at random (MNAR) with respect to measured covariates, sensitivity analyses are often considered. These solutions, however, are unsatisfying in that they are not guaranteed to yield actionable conclusions. Motivated by an EHR-based study of long-term outcomes following bariatric surgery, we consider the use of double sampling as a means to mitigate MNAR outcome data when the statistical goals are estimation and inference regarding causal contrasts based on mean counterfactuals. We describe identification assumptions and derive efficient and robust estimators of the average causal treatment effect under a nonparametric model as well as under a model assuming the missing outcomes were, in fact, initially missing at random (MAR). We compare these in simulations to an approach that adaptively estimates based on evidence of violation of the MAR assumption. Finally, we show how the methods can be extended to: (i) estimation/inference regarding causal quantile treatment effects; and (ii) hypothesis testing regarding MNAR.

Addressing unmeasured confounding in real-world studies: advances in negative control methods and proximal causal inference

Xu Shi, University of Michigan

Xu Shi

Unmeasured confounding remains one of the most significant threats to the credibility of findings from observational studies. Recent developments in negative control methods, also known as proximal causal inference, offer promising strategies to strengthen causal conclusions. These approaches leverage negative controls — variables that have no direct causal relationship with either the exposure or the outcome — to detect and adjust for unmeasured confounding. In this talk, I will review the foundations of negative control methods and introduce the double negative control framework. I will then present our recent work extending this framework to settings where some candidate negative control variables may themselves be invalid. I will conclude with a discussion of open challenges and future research directions.

Session 92: Advances in Graphical Models for Causality and Missingness

Location: Brighton Beach

Organizer: Caleb Miles (Columbia University)

Chair: Oliver Hines (Columbia University)

Presenters: Trung Phung; Danielle Tsao; Razieh Nabi; Daniel Malinsky

Recursive equations for imputation of missing not at random data with sparse pattern support

Trung Phung, Johns Hopkins University

Trung Phung (Johns Hopkins University); Kyle Reese (Johns Hopkins University); Ilya Shpitser (Johns Hopkins University); Rohit Bhattacharya (Williams College)

A common approach for handling missing values in data analysis pipelines is multiple imputation via software packages such as MICE and Amelia. These packages typically assume the data are missing at random (MAR), and impose parametric or smoothing assumptions upon the imputing distributions in a way that allows imputation to proceed even if not all missingness patterns have support in the data. Such assumptions are unrealistic in practice,

and induce model misspecification bias on any analysis performed after such imputation. We provide a principled alternative. Specifically, we develop a new characterization for the full data law in graphical models of missing data. This characterization is constructive, is easily adapted for the calculation of imputation distributions for both MAR and MNAR (missing not at random) mechanisms, and is able to handle lack of support for certain patterns of missingness. We use this characterization to develop a new imputation algorithm—Multivariate Imputation via Supported Pattern Recursion (MISPR)—which uses Gibbs sampling, by analogy with the Multivariate Imputation with Chained Equations (MICE) algorithm, but which is consistent under both MAR and MNAR settings, and is able to handle missing data patterns with no support without imposing additional assumptions beyond those already imposed by the missing data model itself. In simulations, MISPR obtains comparable results to MICE when data are MAR, and superior, less biased results when data are MNAR. Our characterization and imputation algorithm based on it are a step towards making principled missing data methods more practical in applied settings, where the data are likely both MNAR and sufficiently high dimensional to yield missing data patterns with no support at available sample sizes.

Lost in aggregation: the causal interpretation of the IV estimand

Danielle Tsao, University of Washington

Danielle Tsao (University of Washington); Krikamol Muandet (CISPA Helmholtz Center for Information Security); Frederick Eberhardt (California Institute of Technology); Emilija Perković (University of Washington)

An instrumental variable (IV) based estimator of a causal effect emerges as the standard approach to mitigate the confounding bias in social science, economics, and epidemiology where randomized experiments can be too costly or even impossible to conduct. However, justifying the validity of the instrument often poses a significant challenge. In this work, we highlight a problem for IV settings that is generally neglected in justificatory arguments of IV-validity—-the presence of an “aggregate treatment variable”, where the treatment (e.g. education, GDP, caloric intake) is made up of finer-grained components that each may have a different effect on the outcome. For the linear Gaussian case we first characterize the nature of the causal effect of the aggregate treatment on the outcome and then identify the conditions under which standard IV-estimators identify the aggregate effect. The results imply major limitations on the interpretation of IV estimates based on aggregate treatments and highlight the need for a broader justificatory base of the exclusion restriction in IV settings.

Missing data DAGs: from identification to imputation

Razieh Nabi, Emory University

Razieh Nabi (Emory University); Anna Guo (Emory University)

Missing data problems can be viewed as causal inference problems, but in important ways harder ones. Recovering the complete-data law from the observed law can be interpreted as identifying a joint distribution over counterfactual variables corresponding to values that would have been observed had measurement been possible. In this talk, I will discuss recent developments in identification, estimation, and imputation under missing data DAG models. I will describe how ideas from causal identification can be adapted to this setting, as well as important ways in which the two problems differ. In particular, sequential interventions on missingness indicators can themselves induce selection bias, creating new challenges for graphical identification. I will then turn to imputation and ask a related but more targeted question: when can the distribution needed to impute missing values be recovered from the observed data? This perspective leads to graphical identification strategies tailored to observed missingness patterns, as well as ways to incorporate sensitivity assumptions when identification fails. Finally, I will discuss how these identification results can be translated into estimation and practical imputation procedures for mixed-type data under complex, nonignorable missingness.

Towards an end-to-end causal discovery and inference workflow

Daniel Malinsky, Columbia University

Ting-Hsuan Chang; Daniel Malinsky

In some recent work, we have tackled two tasks. First: combining causal discovery (graphical structure learning) with estimation of treatment effects to produce post-selection valid confidence intervals for effects in settings where the underlying causal structure is not entirely known. Second: exploiting the conditional independence structure in a graphical model for efficient estimation of bounds on a target causal effect when unmeasured confounding cannot be ruled out, involving a semiparametric sensitivity model. In this talk, we will discuss combining those two steps into an end-to-end causal discovery + inference workflow: we start with estimating an unknown graph, and then obtain efficient inference for a target effect or bounds for that effect using the graphical structure, all in a way such that the resulting intervals have desired coverage guarantees.

Parallel Sessions 8: Tuesday, November 3, 2026, 1:30 pm–3:00 pm

Session 78: AI-Augmented Design and Analysis in Clinical Trials and Scientific Research: Methods, Applications, and Responsible Implementation

Location: Salon C

Organizer: Qiqi Deng (Moderna Inc.)

Chair: Qiqi Deng (Moderna Inc.)

Presenters: Jesse Gronsbell; Xiaofei Bai; Dooti Roy

ML-powered scientific research: possibilities and pitfalls

Jesse Gronsbell, University of Toronto

Jesse Gronsbell (University of Toronto), Jianhui Gao (Harvard University)

From applications in structural biology to the analysis of electronic health record data, predictions from machine learning models increasingly complement costly gold-standard data in scientific inquiry. While ML-imputed data enables scientific studies to scale in an unprecedented manner, appropriately accounting for inaccuracies in the predictions is critical to achieving trustworthy conclusions from downstream statistical inference.

In this talk, I will explore the methodological and practical impacts of using ML-imputed data across various applications. I will introduce our recently proposed method for bias correction and draw connections with modern methods and classical statistical approaches dating back to the 1960s. I will also discuss ethical challenges of using ML-imputed data, underscoring the need for careful and thoughtful adoption of this practice in scientific research.

Prediction-based inference in randomized clinical trials

Xiaofei Bai, ModernaTX, Inc.

Xiaofei Bai (ModernaTX, Inc.), Xiaolin Chang (ModernaTX, Inc.), Qiqi Deng (ModernaTX, Inc.)

Prediction-based (PB) inference provides a principled framework for incorporating auxiliary predictions into statistical analysis while controlling bias from prediction errors and potentially improving efficiency relative to classical estimators that rely only on gold-standard outcomes. Although much of the existing work has been developed in machine learning contexts, its application in randomized clinical trials remains less explored. This paper presents research on applying PB inference to treatment effect estimation in randomized trials, where auxiliary predictions may arise from surrogate endpoints, interim outcomes, or external models. We compare PB estimators with oracle, classical, and naïve plug-in estimators, evaluating their bias and efficiency under varying predictive performance and correlation structures. The goal is to clarify when and how PB inference offers meaningful gains in randomized clinical trial settings while preserving valid inference.

Title TBA

Dooti Roy, Boehringer

Abstract TBA

Session 105: AI Safety and Evaluation

Location: Williamsburg

Organizer: Bin Yu (University of California, Berkeley) and Shihao Yang (Georgia Institute of Technology)

Chair: Shihao Yang (Georgia Institute of Technology)

Presenters: Andrew Gordon Wilson; Raaz Dwivedi; Shihao Yang

Title TBA

Andrew Gordon Wilson, New York University

Abstract TBA

Can causal inference learn from the bitter lesson? A test case in panel data

Raaz Dwivedi, Cornell Tech & Traversal

Dwaipayan Saha (Columbia University); Jacob Feitelberg (Columbia University); Kyuseong Choi (Cornell Tech); Anish Agarwal (Columbia University, Traversal); Raaz Dwivedi (Cornell Tech, Traversal)

Causal inference relies on human judgment to define questions, choose credible comparisons, and defend identifying assumptions, judgments that make analyses problem-specific. Modern AI has followed the opposite trajectory: general architectures that scale with data and compute, culminating in reusable foundation models, the pattern Sutton called the "bitter lesson."

Can causal inference borrow from this lesson? We test it on panel data, where imputing untreated counterfactuals has produced a zoo of specialized estimators: synthetic controls, matrix completion, nearest neighbors, doubly robust procedures, etc.

The test unfolds in three acts. First, TabImpute repurposes a tabular foundation model for matrix completion via entry-wise featurization. Pre-trained once on synthetic matrices, it beats established imputation baselines zero-shot, but the featurization caps its scale. Second, MatrixFormer makes the architecture matrix-native: alternating row and column attention encodes the exchangeability and factor structure of matrix completion and scales to far larger matrices. With frozen weights — never trained on real data — it matches or beats specialized baselines across causal panels, tabular imputation, LLM benchmark-score prediction, and recommender systems. Third, an auto-research agent, given only a validation metric, searches the loss and synthetic-data mixture and improves the recipe beyond our manual tuning.

Together, these results suggest that synthetic pre-training plus compute can amortize estimator design, a concrete test of the bitter lesson in causal inference.

Using agentic AI for research in data science

Shihao Yang, Georgia Institute of Technology

Shihao Yang (Georgia Institute of Technology)

Agentic AI can carry out increasingly complex data science tasks, but its use in research raises practical questions about reliability, evaluation, and the handling of sensitive data. This talk draws on our work developing and evaluating agents for scientific research. I will introduce LoopPlane, a workflow that connects research planning, experiment execution, validation, and revision while preserving plans and evidence across sessions. I will then discuss evaluations of locally deployed language models on data analysis tasks, examining how model capability, tool interfaces, and computational budgets affect performance and failure modes. Finally, I will present an AI-assisted SQL workflow for Medicaid claims analysis designed to support HIPAA-compliant research. Across these projects, I will discuss how to assess whether an agent has completed a research task correctly, what evidence should accompany its results, and where human judgment remains necessary.

Session 15: Distribution-Free Statistical Inference for AI

Location: Fulton Ferry

Organizer: Vladimir Svetnik (Merck & Co.)

Chair: Matteo Sesia (University of Southern California)

Presenters: Yaniv Romano; Adam Fisch; Anchit Jain and Kevin Zhang; Yonghoon Lee

Reliable inference under budget constraints

Yaniv Romano, Technion

Shai Feldman (Technion); Hen Davidov (Oxford); Yaniv Romano (Technion)

AI systems are moving from one-shot prediction to sequential interaction. For example, an agent may need many steps to work through a complex task. But how can we trust the output of a black-box system whose behavior unfolds over time? This question becomes especially challenging when interaction and evaluation are expensive, and therefore limited by budget constraints. In this talk, I will argue that agentic systems call for a new reliability question. Rather than asking only whether an AI system will eventually succeed or fail at a task, we ask a richer question: how long will it take until an important event occurs, if ever? For example, how many interactions are needed before a jailbreak succeeds, or before an agent completes a task? The key challenge is censoring: with a limited interaction budget, we often stop observing before the event occurs, so the data are only partially labeled. Nevertheless, building on survival analysis, I will show how to provide distribution-free predictive lower bounds for time-to-event outcomes, and how careful allocation of the interaction budget can yield sharper bounds.

Towards efficient model evaluation and routing on the frontier

Adam Fisch, Google DeepMind

Adam Fisch (Google DeepMind)

Evaluating generative AI models is a routine, but resource-intensive, process that is conducted over and over again during the course of model development. In this talk, I will first present Collaborative Evaluation (CollabEval), which builds a predictor of performance by matrix completion over historical runs of other models, and uses its noisy predictions as control variates to obtain unbiased estimates and valid confidence intervals at a fraction of the annotation budget. I will then discuss more generally how noisy performance predictions can be useful, e.g., for routing queries between models, and when it is worth paying for a more accurate estimate. We formalize this as an instance of Pandora's Box, the classical problem of optimal search with costly inspection, which gives simple rules for when reducing the noise in a prediction is worth its cost.

Distribution-free uncertainty quantification with time-to-event data

Anchit Jain and Kevin Zhang, Massachusetts Institute of Technology

Anchit Jain (Massachusetts Institute of Technology); Kevin Zhang* (Massachusetts Institute of Technology); Stephen Bates (Massachusetts Institute of Technology)*

Time-to-event data with censoring is ubiquitous in oncology. For example, we might measure progression-free survival time, but some patients are lost to follow-up before progression resumes. Moreover, with high-dimensional imaging and omics data increasingly available, it is critical to develop reliable models for time-to-event data that can incorporate such data. A general approach to this problem is given by deep Cox models, which combine deep neural networks for high-dimensional data with the Cox proportional hazards model, which gracefully handles time-to-event data with censoring. However deep Cox models are complex and the outputs may be biased or incorrect in ways that cannot be easily anticipated. As such, it is critical to augment these models with an accurate assessment of their certainty or uncertainty. We consider the problem of obtaining calibrated survival curves for time-to-event modeling problems with deep Cox models and develop a post-hoc calibration strategy that takes into account right-censoring.

Synthetic powered statistical inference

Yonghoon Lee, University of Pennsylvania

Meshi Bashari (Technion); Yonghoon Lee (Upenn); Roy Maor Lotan (Technion); Edgar Dobriban (Upenn); Yaniv Romano (Technion)

We discuss how AI-generated synthetic data can be leveraged to improve statistical inference by effectively increasing the sample size, while addressing the key challenge of unreliable AI outputs. We introduce a framework called General Synthetic-Powered Inference (GESPI), which transforms an existing inference algorithm

into one that can safely leverage synthetic data. GESPI provides a distribution-free worst-case guardrail on the quality of the resulting inference, regardless of the quality of the synthetic data. We discuss applications to hypothesis testing, predictive inference, multiple testing, and other inference problems, and present empirical results on tasks including image classification and protein structure prediction.

Session 98: Robustness Bayesian Methods with Machine Learning Applications

Location: Sunset Park

Organizer: Jami Mulgrave (NC State University)

Chair: Jami Mulgrave (NC State University)

Presenters: John Paisley; Aahlad Manas Puli; Marco Avella Medina; Alberto Gonzalez-Sanz

Gaussian process neural additive models for explainable supervised learning

John Paisley, Columbia University

John Paisley (Columbia), Wei Zhang (Columbia), Brian Barr (Capital One)

Deep neural networks have revolutionized many fields, but their black-box nature also occasionally prevents their wider adoption in fields such as healthcare and finance, where interpretable and explainable models are needed. The recent development of Neural Additive Models (NAMs) is a significant step in the direction of interpretable deep learning for tabular datasets. In this talk, we discuss a new subclass of NAMs that use a single-layer neural network construction of the Gaussian process via random Fourier features, which we call Gaussian Process Neural Additive Models (GP-NAM). GP-NAMs have the advantage of a convex objective function and number of trainable parameters that grows linearly with feature dimensionality. It suffers no loss in performance compared to deeper NAM approaches because GPs are well-suited for learning complex non-parametric univariate functions. We demonstrate the performance of GP-NAM on several tabular datasets, showing that it achieves comparable or better performance in both classification and regression tasks with a large reduction in the number of parameters.

Two tales of causal inference: from structure to scale

Aahlad Manas Puli, Center For Data Science, NYU

Aahlad Manas Puli (NYU)

Causal inference starts from assumptions about the data-generating process and turns them into estimators of causal effects. I will present two works that address different parts of this process: one by exploiting structure in a challenging causal problem, and the other by asking how causal inference can be scaled across many such problems. First, I will discuss Estimation with Functional Confounders (EFC), motivated by settings such as genome-wide association studies in which confounding is functionally specified. We introduce a new condition for nonparametric identification and develop LODE, an estimator that exploits this structure to estimate causal effects. I will then present Black Box Causal Inference (BBCI), which encodes causal assumptions through a distribution over data-generating processes and learns an inference procedure that maps observed datasets to causal estimands. This enables efficient estimation for different causal problems without requiring the same mathematical effort in deriving a new estimator for each setting.

A theoretical framework for M-posteriors

Marco Avella Medina, Columbia University

Juraj Marusic (Columbia University), Marco Avella Medina (Columbia University), Cynthia Rush (Columbia University)

We provide a theoretical framework for a wide class of generalized posteriors that can be viewed as the natural Bayesian posterior counterpart of the class of M-estimators in the frequentist world. We call the members of this class M-posteriors and show that they are asymptotically normally distributed under mild conditions on the M-estimation loss and the prior. In particular, an M-posterior contracts in probability around a normal distribution centered at an M-estimator, showing frequentist consistency and suggesting some degree of robustness depending on the reference M-estimator. We formalize the robustness properties of the M-posteriors by a new characterization of the posterior influence function and a novel definition of breakdown point adapted for posterior distributions. We illustrate the wide applicability of our theory in various popular models and discuss extensions to variational inference. We illustrate their empirical relevance of our results in some numerical examples.

Improved learning theory for kernel distribution regression with two-stage sampling

Alberto Gonzalez-Sanz, Columbia University

François Bachoc, Louis Béthune, Alberto González-Sanz, and Jean-Michel Loubes

Distribution regression, where covariates are probability distributions rather than finite-dimensional vectors, arises in many statistical and machine learning applications. In practice, these distributions are not observed directly but through finite samples, resulting in a two-stage sampling problem.

We study kernel distribution regression based on Hilbertian embeddings of probability measures and derive improved learning guarantees under two-stage sampling. Our main contribution is a near-unbiased condition for empirical Hilbertian embeddings, which enables a sharper control of the error introduced when replacing population distributions by their empirical counterparts. We establish this condition for several important classes of embeddings, including kernel mean embeddings and optimal-transport-based constructions such as sliced Wasserstein and Sinkhorn-type embeddings. This analysis

yields convergence rates that improve upon existing guarantees for kernel distribution regression. Our results provide a refined characterization of the interplay between the number of observed distributions and the number of samples drawn from each distribution, and are supported by numerical experiments illustrating the predicted statistical behavior.

Session 76: Improving Inference in Nonparametric Models

Location: Golden Boardroom

Organizer: Ted Westling (University of Massachusetts Amherst)

Chair: Ted Westling (University of Massachusetts Amherst)

Presenters: Charles Doss; Arun Kumar Kuchibhotla; Medha Agarwal; Ted Westling

Adaptive confidence bands for monotone causal continuous treatment effect curves

Charles Doss, University of Minnesota

Charles Doss (University of Minnesota); R. Oliver VandenBerg (Bucknell University)

In causal inference with observational data new inferential targets arise when treatment is continuous rather than discrete. Here we form (simultaneous) confidence bands over the continuous treatment effect curve. We form confidence bands based on assuming the curve satisfies shape constraints such as monotonicity, under the assumption of no unmeasured confounding. The assumption of monotonicity (or unimodality) is often a natural one for the treatment effect curve, at least over a reasonable domain of treatment. As is commonly the case, two nuisance functions arise, and our bands have a Neyman orthogonality property in the sense that they are asymptotically valid if the product of the error rates of the two nuisance functions is small enough. Furthermore, our bands are adaptive to a range of function smoothness values in that when the true smoothness level is unknown the bands still achieve the same (optimal) rate of convergence as when the true smoothness level is known. If time permits we will also develop analogous confidence bands under the constraint of concavity.

Statistical inference for linear programming problems: debiasing and regularity

Arun Kumar Kuchibhotla, Carnegie Mellon University

Arun Kuchibhotla (Carnegie Mellon University); Koustav Goswami (Indian Statistical Institute)

Optimal transport (OT) is an increasingly prominent tool in data analysis, with applications in causal inference, domain adaptation, and even computational biology. A chief virtue of optimal transport is that it gives rise both to a measure of similarity between two probability measures but also an optimal coupling or plan, which describes the least costly way to move mass from one measure to another. In fact, in many of the applications described above, it is the plan rather than the distance itself which is the primary object of interest. Discrete OT, i.e., transport between discrete distributions is a special case of a stochastic linear programming problem with non-stochastic objective and a stochastic linear constraint. As is well-known in the literature of mathematical programming and operations research, non-differentiability of solutions to perturbed linear programming problems implies potential non-Gaussianity of the discrete OT coupling. In a recent article, Liu, Bunea, and Niles-Weed (https://arxiv.org/pdf/2505.04312) proposed a relaxed discrete OT based on barrier function used in interior-point methods, and proved its asymptotic normality. They also propose bootstrap inference based on this penalized estimator. We show the pitfalls of this methodology and discuss similar issues with any penalized method. Finally, we propose a simple inference method for linear programming problems. The methodology also applies to other constrained problems without the need to derive/use the limiting distributions. This is based on joint work with Koustav Goswami (a masters student at Indian Statistical Institute, Kolkata.)

Sinkhorn treatment effects: a causal optimal transport measure

Medha Agarwal, University of Washington

Medha Agarwal (University of Washington); Alex Luedtke (Harvard Medical School)

We introduce the Sinkhorn treatment effect, an entropic optimal transport measure of divergence between counterfactual distributions. Unlike classical quantities such as the average treatment effect, this measure captures differences across entire distributions. We analyze this divergence as a statistical functional and show it can be written as a smooth transformation of counterfactual mean embeddings with an appropriate kernel. This characterization allows us to establish first-order pathwise differentiability in general, and second-order pathwise differentiability under the null hypothesis of equal counterfactual distributions. Leveraging this smoothness, we construct debiased estimators and use them to obtain asymptotically valid tests for distributional treatment effects with a fixed entropic regularization parameter. Because the power of the test depends on this unknown parameter, we further propose an aggregated test that combines evidence across a grid of regularization choices. Experiments on simulated and image data demonstrate the practical advantages of our estimator and testing procedure.

Inference on contrasts of monotone functions

Ted Westling, University of Massachusetts Amherst

Ted Westling (University of Massachusetts Amherst), Charlie Wolock (University of Rochester), Marco Carone (University of Washington)

In many settings, it is known that a function of interest is monotone due to probabilistic constraints or prior scientific

knowledge. General approaches to estimation of a monotone function have been proposed, the properties of these estimators have been derived, and the results have been applied to a variety of specific problems. In some cases, the contrast of two or more monotone functions is of scientific interest in addition to the functions themselves. For instance, in the context of current status censoring, investigators may be interested in the difference or ratio between the survival functions of the event of interest under two different conditions, such as treatment and control. A natural approach to estimation in such problems is to apply the contrast to monotone estimators. To obtain asymptotically valid inference based on the resulting contrast estimator, the joint asymptotic behavior of the monotone estimators is needed. We derive conditions under which two or more generalized Grenander-type estimators converge jointly in distribution, upon proper centering and scaling, to a multivariate analogue of the Chernoff distribution. We use our general results to obtain large-sample inference for contrasts of monotone functions in several examples with important real-world applications.

Session 18: Causal and Machine Learning Methods in Health Research

Location: Carroll Gardens

Organizer: Zhonghua Liu (Columbia University)

Chair: Kan Chen (Columbia University)

Presenters: Shuangge Ma; Yi Lian; Yiwang Zhou; Jason Roy

DNN-based semiparametric AFT model for integrating genomic and pathological imaging data in cancer prognosis

Shuangge Ma, Yale School of Public Health

Shuangge Ma

Modeling prognosis has critical implications in cancer research and clinical practice. Many studies have been conducted, built on genomic (and omics in general) and pathological imaging data. In recent research, a handful of studies have also integrated the two types of data, taking advantage of their complementary signals. In this study, we further advance cancer prognosis modeling by developing a semiparametric AFT model. For modeling the high-dimensional genomic variables – where strong interpretability is desired, we assume parametric effects. For modeling the high-dimensional pathological imaging features – where more flexible/effective modeling can be preferred over interpretability, we assume nonparametric effects. Different from many existing studies, such nonparametric effects are estimated using deep neural networks (DNNs). To differentiate important genomic and pathological imaging variables from noises, we impose penalization on both the parametric and nonparametric effects. In particular, for the nonparametric effects, we apply group penalization to the first-layer weights. The asymptotic selection and estimation consistency and normality properties are carefully established, which can provide a uniquely strong ground. Computation is examined. Simulation demonstrates the competitive performance of the proposed approach. In the analysis of TCGA data on lung cancer, the proposed approach leads to satisfactory and sensible findings different from the alternatives.

Bridging agentic AI and statistical reasoning for enhanced clinical trial design co-pilot

Yi Lian, University of Pennsylvania

Hank Tang (University of Pennsylvania); Peng Yang (The University of Texas MD Anderson Cancer Center); Jiayi Xin (University of Pennsylvania); Yi Lian (University of Pennsylvania); Qi Long (University of Pennsylvania)

Research on AI for clinical trials has mostly focused on evidence synthesis, clinical design components, and patient-trial matching. Until recent agentic AI systems, most tools handled one task at a time. These tools are usually trained with past trials as the ground truth, which encourages replication of past practice and discourages methodological innovation.

An end-to-end AI co-pilot for clinical trials needs AI agents that carry out statistical tasks: choosing the statistical design, collecting prior information, optimizing design parameters, calibrating operating characteristics, calculating sample size and power, and, during the trial, ongoing monitoring, interim analyses, and the final analysis. AI tools that reason from statistical principles can help trialists, including statisticians, keep up with the clinical, statistical, and regulatory landscape, reduce back-and-forth among experts during design, and ease operational work during trial conduct.

As statisticians, we can define a formal ontology of statistical design and build knowledge graphs that guide AI agents to reason from statistical principles and reduce hallucination. We can also create labeled datasets and benchmarks for training and evaluating AI tools, take part in human evaluation of AI performance, and help develop end-to-end AI support for clinical trials. In this talk, we present our research projects in this area.

Longitudinal data improve risk prediction in pediatric allogeneic hematopoietic cell transplant recipients

Yiwang Zhou, Department of Biostatistics, St. Jude Children's Research Hospital

Yiwang Zhou (Department of Biostatistics, St. Jude Children's Research Hospital); Ramakrishnan Kizhakkancherry Ganesh (Department of Bone Marrow Transplantation and Cellular Therapy, St. Jude Children's Research Hospital); Ying Zheng (Department of Biostatistics, St. Jude Children's Research Hospital); Dinesh Keerthi (Department of Bone Marrow Transplantation and Cellular Therapy, St. Jude Children's Research Hospital); Cai Li (Department of Biostatistics, St. Jude Children's Research Hospital); Yilun Sun (Department of Biostatistics, St. Jude Children's Research Hospital); Jessi Smith (Department of Bone Marrow Transplantation and Cellular Therapy, St. Jude Children's Research Hospital); Li Tang

(Department of Quantitative Health Sciences, Cleveland Clinic); Akshay Sharma (Department of Bone Marrow Transplantation and Cellular Therapy, St. Jude Children's Research Hospital)

Accurate prognostic assessment after allogeneic hematopoietic cell transplantation (allo-HCT) is essential for identifying patients at high risk of adverse outcomes, yet existing models rely on static baseline variables and fail to capture evolving clinical status before and after transplantation. We developed and validated machine learning-based risk prediction models for short- and long-term outcomes, including overall survival and acute graft-versus-host disease (GVHD), in pediatric and young adult allo-HCT recipients. Models integrated baseline characteristics, longitudinal clinical measurements, and medication administration data collected pre- and post-transplant. Longitudinal data were summarized using statistical features for survival modeling and functional principal component analysis for acute GVHD, while medication data were represented by duration and cumulative dosage. Using data from St. Jude Children's Research Hospital, models demonstrated strong internal performance and were externally validated with an independent cohort from Memorial Sloan Kettering Cancer Center. Incorporating longitudinal and medication data improved predictive accuracy for both short- and long-term outcomes based on time-dependent AUROC, with these features ranking among the most important predictors. This study highlights the value of dynamic, data-driven approaches for improving risk stratification and clinical decision-making in allo-HCT.

Bayesian structural nested failure time models for causal survival analysis with time-varying treatments

Jason Roy, Rutgers School of Public Health

Yudan Zou (Rutgers); Jason Roy (Rutgers); Dan Horton (Rutgers); Liangyuan Hu (Rutgers)

Structural nested failure time models (SNFTMs) provide a useful framework for causal survival analysis with time-varying treatments and time-dependent confounding. Existing approaches face complementary limitations. Rank-based g-estimation requires artificial recensoring and yields non-smooth objective functions. Structural nested cumulative survival time models avoid recensoring, but their unconstrained additive hazards formulation does not bound survival probabilities, creating instability under data sparsity. We develop a Bayesian likelihood formulation for SNFTMs by specifying a parametric distribution for the baseline counterfactual failure time together with models for the longitudinal covariate process. This replaces the rank-based objective with a smooth likelihood for blip parameters, enabling posterior computation via Markov chain Monte Carlo with data augmentation for censored event times. Under correct specification and regularity conditions, we establish identifiability, posterior consistency, and asymptotic normality for the causal parameters. Simulation studies evaluate finite sample performance across various scenarios and selected misspecification. In the settings considered, the proposed estimator showed stable inference and accurate recovery of counterfactual survival summaries relative to rank-based g-estimation and parametric g-formula estimators. We apply the method to CARRA Registry data to compare adalimumab withdrawal strategies in juvenile idiopathic arthritis, illustrating how the Bayesian SNFTM can estimate regime-specific flare-free survival under time-varying confounding.

Session 01: Neural Networks in Dimension Reduction and Causal Inference

Location: Coney Island

Organizer: Yanyuan Ma (Penn State University)

Chair: Yin Tang (University of Kentucky)

Presenters: Yin Tang; Shitao Fan; Wenxi Tan; Bing Li

Efficient estimation of the complier general causal effect in randomized controlled trials with one-sided noncompliance

Yin Tang, University of Kentucky

Yin Tang (University of Kentucky); Yanyuan Ma (Pennsylvania State University); Jiwei Zhao (University of Wisconsin–Madison)

A randomized controlled trial (RCT) is widely regarded as the gold standard for assessing the causal effect of a treatment or intervention, assuming perfect implementation. In practice, however, randomization can be compromised for various reasons, such as one-sided noncompliance. In this paper, we first systematically study the likelihood-based identifiability in an RCT with one-sided noncompliance. This foundational analysis naturally gives rise to the complier general causal effect (CGCE) as the primary estimand. We further develop two estimators for the CGCE: a simple estimator that requires no nonparametric procedures, and an efficient estimator that achieves the semiparametric efficiency bound. Our theoretical analysis shows that, achieving semiparametric efficiency requires only the nuisance estimators to converge in L2-norm, with no restriction on their convergence rates. This rate-free property opens the door to employing many more modern machine learning methods while still guaranteeing efficiency. Comprehensive simulation studies and a real data application are conducted to illustrate the proposed methods and to compare them with existing approaches.

Title TBA

Shitao Fan, University of Maryland

Abstract TBA

Belted engression: sufficient dimension reduction for generative distributional regression

Wenxi Tan, Pennsylvania State University

Wenxi Tan (Pennsylvania State University); Bing Li (Pennsylvania State University); Lingzhou Xue (Pennsylvania State University)

Modern conditional generative models face significant challenges when learning complex covariate dependencies. While sufficient dimension reduction (SDR) provides a principled approach to compress these dependencies, traditional SDR frameworks were not formulated for conditional generation. To bridge this gap, we propose Belted Engression, a unified and architecturally parameter-efficient framework for generative distributional regression. Our approach establishes an end-to-end compress-then-generate paradigm driven by sufficient representation learning, embedding a structural bottleneck into the generative architecture. Theoretically, we prove that the standard SDR condition is equivalent to a law-preserving generative factorization, which is achieved at the global optimum of the population Belted Engression objective. Furthermore, by uncovering a localized Bernstein-type control for the energy-score loss, we establish finite-sample convergence rates that are sharper than those of existing results. We also prove that this belted architecture is strictly smaller, operating with an asymptotically vanishing parameter count relative to the unstructured baseline. Extensive simulations and real-world applications demonstrate that Belted Engression achieves superior distributional prediction and SDR recovery with fewer trainable parameters.

Doubly robust estimation of causal effects for random object outcomes with continuous treatments

Bing Li, Pennsylvania State University

Satarupa Bhattacharjee (University of Florida), Bing Li (Pennsylvania State University), Xiao Wu (Columbia University), Lingzhou Xue (Pennsylvania State University)

Causal inference with a continuously varying treatment is usually formulated for scalar or finite-dimensional outcomes. We study a random-object outcome in a metric space whose scientifically chosen metric admits an exact isometric embedding into a separable Hilbert space and whose relevant marginal and conditional Hilbert means remain in the embedded image. The potential outcome is treated as the whole process $Y = \{Y_t : t \in \mathcal{T}\}$, while the observed response Y_T is its evaluation at the realized continuous treatment. Treatment densities are Radon–Nikodym derivatives, not point probabilities. For a fixed bandwidth, we derive the canonical gradient of the kernel-smoothed Hilbert representation and construct inverse-probability-weighted, doubly robust, and cross-fitted estimators. The augmented score evaluates both nuisance functions at the realized treatment and includes an integrated outcome-regression term; its population remainder is exactly the product of the two nuisance errors. Under explicit local positivity, moment, Lindeberg, nuisance-rate, quadrature, and outcome-reconstruction conditions, we establish pointwise Hilbert-space asymptotic normality and joint convergence on a prespecified finite dose grid. Under stronger uniform conditions, we also obtain a Hilbert-valued process limit on the bandwidth-rescaled local dose scale. We explain how the scientific meaning of the target changes with the outcome metric and evaluate the estimator in causal simulations for Wasserstein distributions, log–Euclidean positive-definite matrices, and Aitchison compositions. The method is illustrated in an analysis of county-level $PM_{2.5}$ exposure and age-at-death distributions.

Session 33: Embeddings and Dynamics in Complex Networks

Location: Park Slope

Organizer: Emma Jingfei Zhang (Emory University)

Chair: Emma Jingfei Zhang (Emory University)

Presenters: Keith Levin; Tianxi Li; Peter W. MacDonald; Jesús Arroyo

Embeddings of correlation networks

Keith Levin, University of Wisconsin–Madison

Keith Levin (University of Wisconsin–Madison)

In many applications, weighted networks are constructed based on time series data: each time series is associated to a vertex and edge weights are given by pairwise correlations. The result is a network whose edge dependency structure violates the assumptions of most common network models. Nonetheless, it is common to analyze these “correlation networks” using embedding methods derived from edge-independent network models, based on a belief that the edges are approximately independent. In this work, we put this modeling choice on firm theoretical ground. We show that when the time series are expressible in terms of a small number of Fourier basis elements (or in some other suitably-chosen basis), correlation networks correspond to latent space networks with dependent edge noise in which the vertex-level latent variables encode the basis coefficients. Further, we show that when time series are observed subject to noise, spectral embedding of the resulting noisy correlation network still recovers these true vertex-level latent representations under suitable assumptions. This characterization of embeddings as learning Fourier coefficients appears to be folklore in the signal processing community in the context of principal component analysis, but is, to the best of our knowledge, new to the statistical network analysis literature.

Nonparametric estimation of time-varying network connections by multi-stage smoothing

Tianxi Li, University of Minnesota

Jeonghwan Lee (University of Minnesota); Tianxi Li (University of Minnesota); Adam Rothman (University of Minnesota)

We consider the problem of estimating the underlying edge probabilities of a time-varying network observed at multiple time points. The probability structure is represented by a time-varying graphon that satisfies temporal Hölder smoothness and piecewise Lipschitz conditions in the latent variables. We propose a multi-stage

smoothing estimator that first applies temporal local smoothing to each edge and then performs node-domain smoothing using a data-driven neighborhood construction adapted from the method. An additional temporal smoothing step is introduced as an optional refinement when uniform accuracy over the entire time domain is required. Simulation studies demonstrate the benefits of combining temporal and node-domain smoothing under different generative models. We also apply the method to a real time-varying network dataset and show that it captures both smooth temporal evolution and structural patterns in the connectivity.

Mesoscale changepoint detection for dynamic networks

Peter W. MacDonald, University of Waterloo

Peter W. MacDonald (University of Waterloo); Shonak Duggal (University of Waterloo)

Dynamic network data arises naturally in many applications as a representation of a complex system observed over time. Changepoint detection and estimation is a well-studied problem in dynamic network analysis, but most often at the global scale, where changepoints are defined as a shift in the entire network's generative distribution. In this work, we address the more subtle problem of detecting and estimating changepoints for subsets of network edges. Conducting "mesoscale" changepoint detection on meaningful groups of edges can lead to more specific and interpretable conclusions, while still aggregating signals locally in the network. In this work, for a flexible class of independent edge network models, we develop approaches to changepoint detection which leverage information from outside the set of interest to adaptively aggregate signals, leading to sensitive detection and accurate estimation of mesoscale network changepoints.

Covariate-assisted graph matching

Jesús Arroyo, Texas A&M University

Trisha Dawn (University of California, Santa Cruz); Jesús Arroyo (Texas A&M University)

Data integration is essential across diverse domains, from historical records to biomedical research, facilitating joint statistical inference. A crucial initial step in this process involves merging multiple data sources based on matching individual records, often in the absence of unique identifiers. When the datasets are networks, this problem is typically addressed through graph matching methodologies. For such cases, auxiliary features or covariates associated with nodes or edges can be instrumental in achieving improved accuracy. However, most existing graph matching techniques do not incorporate this information, limiting their performance against non-identifiable and erroneous matches. To overcome these limitations, we propose two novel covariate-assisted seeded graph matching methods, where a partial alignment for a set of nodes, called seeds, is known. The first one solves a quadratic assignment problem (QAP) over the whole graph, while the second one only leverages the local neighborhood structure of seed nodes for computational scalability. Both methods are grounded in a conditional modeling framework, where elements of one graph's adjacency matrix are modeled using a generalized linear model (GLM), given the other graph and the available covariates. We establish theoretical guarantees for model estimation error and exact recovery of the solution of the QAP. The effectiveness of our methods is demonstrated through numerical experiments and in an application to matching the statistics academic genealogy and the collaboration networks. By leveraging additional covariates, we achieve improved alignment accuracy. Our work highlights the power of integrating covariate information in the classical graph matching setup, offering a practical and improved framework for combining network data with wide-ranging applications.

Session 45: Statistical Learning and Network Analysis for Complex Biological Data

Location: Windsor Terrace

Organizer: Quefeng Li (UNC Chapel Hill)

Chair: Quefeng Li (UNC Chapel Hill)

Presenters: Ying Guo; Rong Ma; Muxuan Liang; Mingcong Wu

A statistical AI framework for learning directed brain connectomes from neuroimaging data

Ying Guo, Emory University

Ying Guo (Emory University); Yaotian Wang (Emory University); Wenyi Wang (Emory University)

In recent years, connectome-based research has become a central focus in neuroscience, offering essential insights into brain organization and advancing predictive modeling of cognitive, behavioral, and mental health outcomes. While most existing approaches focus on undirected brain connectivity, they overlook the directionality and causal influences between brain regions. To address this limitation, we propose a statistical AI framework for learning directed brain connectomes from neuroimaging data. Our approach integrates principled statistical modeling with deep learning to infer sparse, interpretable directed connectivity graphs that characterize latent causal interactions across the brain. At the same time, the framework learns low-dimensional graph embeddings optimized for downstream prediction tasks, including demographic attributes and clinical phenotypes. The proposed method uncovers whole-brain directed connectivity patterns and reveals novel subpopulation-specific connectomic differences, highlighting its potential to advance both mechanistic understanding and predictive modeling in neuroscience.

Statistical principles of spectral data integration methods in biomedical research

Rong Ma, Harvard University

Tavor Baharav (Johns Hopkins University), Phillip Nicol (University of Chicago), Rafael Irizarry (Harvard and DFCI), Rong Ma (Harvard)

Modern data analysis increasingly requires identifying shared latent structure across multiple high-dimensional datasets. A commonly used model assumes that the data matrices are noisy observations of low-rank matrices with a shared singular subspace. In this case, two primary methods have emerged for estimating this shared structure, which vary in how they integrate information across datasets. The first approach, termed Stack-SVD, concatenates all the datasets, and then performs a singular value decomposition (SVD). The second approach, termed SVD-Stack, first performs an SVD separately for each dataset, then aggregates the top singular vectors across these datasets, and finally computes a consensus amongst them. While these methods are widely used, they have not been rigorously studied in the proportional asymptotic regime, which is of great practical relevance in today's world of increasing data size and dimensionality. This lack of theoretical understanding has led to uncertainty about which method to choose and limited the ability to fully exploit their potential. To address these challenges, we derive exact expressions for the asymptotic performance and phase transitions of these two methods and develop optimal weighting schemes to further improve both methods. Our analysis reveals that while neither method uniformly dominates the other in the unweighted case, optimally weighted Stack-SVD dominates optimally weighted SVD-Stack. We extend our analysis to accommodate multiple shared components, and provide practical algorithms for estimating optimal weights from data, offering theoretical guidance for method selection in practical data integration problems.

Multi-task learning for semiparametric models with nuisance parameter fusion

Muxuan Liang, The University of Texas MD Anderson Cancer Center

Sohom Bhattacharya (University of Florida); Yongzhuo Chen (University of Florida); Muxuan Liang (The University of Texas MD Anderson Cancer Center)

In the age of large and heterogeneous datasets, integrating information from diverse sources is essential for improving parameter estimation. Multi-task learning offers a powerful approach by enabling simultaneous learning across related tasks. In this work, we introduce a late fusion framework for multi-task learning with semiparametric models involving infinite-dimensional nuisance parameters, focusing on applications such as heterogeneous treatment effect estimation across multiple data sources. Our framework is two-step: first, initial double machine-learning estimators are obtained through individual task learning; second, these estimators are adaptively aggregated to exploit task similarities while remaining robust to task-specific differences. The framework avoids individual-level data sharing, thereby preserving privacy. Additionally, we propose a novel multi-task learning method for nuisance parameter estimation, which further improves parameter estimation when nuisance parameters are similar across tasks. We establish theoretical guarantees, demonstrating faster convergence rates than individual task learning when tasks share similar parametric components. Extensive simulations and real-data applications complement our theoretical findings and highlight the effectiveness of the framework even with moderate sample sizes.

Temporal conditional graphical modeling of disease interconnections using longitudinal clinical treatment data

Mingcong Wu, Yale University

Mingcong Wu (Yale University); Shuangge Ma (Yale University)

The fast accumulation of data in electronic health records (EHRs) and medical claims databases has provided unprecedented opportunities to study diseases not only individually but also collectively, in a longitudinal manner. This study has been motivated by the analysis of disease interconnections under a network analysis paradigm based on clinical treatment measures in the All of Us data and many similar datasets. To directly overcome limitations of the existing works, we develop a temporal conditional graphical modeling framework. Specifically, it directly models temporal dependencies among disease-specific outcomes, advancing from those primarily focused on time-varying network structures. To better distinguish the sources of interconnections, it directly accounts for relevant covariates and potential unobserved latent factors. It is also tailored to accommodate zero inflation and overdispersion in data. Theoretical properties are rigorously established, and simulation studies demonstrate its favorable performance. With the All of Us data, we analyze interconnections in the number of outpatient visits for 30 diseases using data on 176,642 samples. Demographic and genetic features are accounted for. Sensible findings are made in terms of pairwise interconnection, connectivity, module, and other properties, which differ significantly from the alternative methods. Such findings may potentially have implications for healthcare resource allocation and planning.

Session 90: Mechanism Design for Data Science and AI

Location: DUMBO

Organizer: Lihua Lei (Stanford University)

Chair: Lihua Lei (Stanford University)

Presenters: Stephen Bates; Aaron Roth; Haifeng Xu; Jason Hartline

Instance-adaptive hypothesis tests with heterogeneous agents

Stephen Bates, Massachusetts Institute of Technology

Flora Shi (Massachusetts Institute of Technology), Stephen Bates

(Massachusetts Institute of Technology), and Martin Wainwright (Massachusetts Institute of Technology)

We study hypothesis testing over a heterogeneous population of strategic agents with private information. Any single test applied uniformly across the population yields statistical error that is sub-optimal relative to the performance of an oracle given access to the private information. We show how it is possible to design menus of statistical contracts that pair type-optimal tests with payoff structures, inducing agents to self-select according to their private information. This separating menu elicits agent types and enables the principal to match the oracle performance even without a priori knowledge of the agent type. Our main result fully characterizes the collection of all separating menus that are instance-adaptive, matching oracle performance for an arbitrary population of heterogeneous agents. We identify designs where information elicitation is essentially costless, requiring negligible additional expense relative to a single-test benchmark, while improving statistical performance. Our work establishes a connection between proper scoring rules and menu design, showing how the structure of the hypothesis test constrains the elicitable information. Numerical examples illustrate the geometry of separating menus and the improvements they deliver in error trade-offs. Overall, our results connect statistical decision theory with mechanism design, demonstrating how heterogeneity and strategic participation can be harnessed to improve efficiency in hypothesis testing.

AI alignment from diverse misalignment

Aaron Roth, University of Pennsylvania

Natalie Collina (Massachusetts Institute of Technology); Surbhi Goel (Penn); Aaron Roth (Penn); Sikata Sengupta (Penn)

If aligning an AI model with the objectives of its user turns out to be hard or impossible, are we doomed as capabilities scale up? I sure hope not! In this talk we discuss one formal reason for hope: we may have multiple AI agents or providers that are each individually misaligned, but differently misaligned. We will see two mechanisms — one emerging naturally from the structure of competitive AI marketplaces, and another resulting from the pro-active design of multi-agent oversight — by which safety or alignment guarantees can naturally emerge from diverse misalignment. A key geometric structure which arises in both cases is what we call coalitional coverage: that the user's utility function lie close to the non-negative span of the AI agent utility functions. This gives a way to obtain guarantees from diversity in the direction of misalignment, which might be easier to obtain than perfect alignment. We will lead with the optimistic view. Time permitting, we'll also see a reason for pessimism.

When do prophets profit in prediction markets?

Haifeng Xu, University of Chicago and Google

Anri Gu, Nicole Kagan, Alec Sun, Jibang Wu, Haifeng Xu

Prediction markets aggregate dispersed beliefs into prices that act as probabilistic forecasts of uncertain events. Classical theory establishes a clean equivalence between forecasting accuracy and trading profit, but only for the specific automated market maker (AMM) design. However, the largest exchanges today are based on central limit order books in which informed forecasters routinely lose money while uninformed strategies can profit on simple heuristics. We resolve this discrepancy by establishing a formal equivalence between predictive accuracy and profitability. For any strictly proper scoring rule S, we exhibit a "proper" betting strategy that depends only on the forecaster's prediction p and the market price q and earns positive expected profit whenever p outperforms q under S and the market has sufficient liquidity. The proof rests on a decomposition of expected profit that strictly generalizes the classical AMM guarantee and also explains how strategies can profit without an accuracy edge.

Empirically, across thousands of forecasts from AI models, proper betting is the only strategy that reliably converts accuracy into profit, and we further identify systematic forecasting personas and show how the optimal proper strategy varies across them. A month-long live deployment achieves 80.33% return on investment with a Sharpe ratio of 3.35.

A decision theoretic framework for measuring AI reliance

Jason Hartline, Northwestern University

Ziyang Guo (Northwestern University), Yifan Wu (Microsoft Research), Jason Hartline (Northwestern University), Jessica Hullman (Northwestern University)

Humans frequently make decisions with the aid of artificially intelligent (AI) systems. A common pattern is for the AI to recommend an action to the human who retains control over the final decision. Researchers have identified ensuring that a human has appropriate reliance on an AI as a critical component of achieving complementary performance. We argue that the current definition of appropriate reliance used in such research lacks formal statistical grounding and can lead to contradictions. We propose a formal definition of reliance, based on statistical decision theory, which separates the concepts of reliance as the probability the decision-maker follows the AI's recommendation from challenges a human may face in differentiating the signals and forming accurate beliefs about the situation. Our definition gives rise to a framework that can be used to guide the design and interpretation of studies on human-AI complementarity and reliance. Using recent AI-advised decision making studies from literature, we demonstrate how our framework can be used to separate the loss due to mis-reliance from the loss due to not accurately differentiating the signals. We evaluate these losses by comparing to a baseline and a benchmark for complementary performance defined by the expected payoff achieved by a rational decision-maker facing the same decision task as the behavioral decision-makers.

Session 06: Statistical Learning in Biomedical Studies

Location: Greenpoint

Organizer: Zhezhen Jin (Columbia University)

Chair: Shanshan Ding (Department of Applied Economics and Statistics, University of Delaware)

Presenters: Yushu Shi; Lan Xue; Emily Ouyang; Yan Yu

Nutritional exposures and risk of breast, ovarian, and uterine cancer: automated meta-analyses

Yushu Shi, Weill Cornell Medicine

Margaux Delporte (University of Rhode Island), Rulla Tamimi (Weill Cornell Medicine), Saurabh Mehta (Cornell University), Eunji Choi (Weill Cornell Medicine), Yongxian Zhang (Weill Cornell Medicine), Yushu Shi (Weill Cornell Medicine)

We developed MetaFemina, an automated evidence-synthesis pipeline integrating keyword-based literature retrieval, LLM-assisted evidence extraction, and random-effects meta-analysis, with analyses of heterogeneity, publication bias, and leave-one-out sensitivity. The platform also provides sample size calculations and visual and plain-language summaries. We evaluated its performance against two recent peer-reviewed meta-analyses and compared exposure–outcome associations across cancer types. Compared with recent meta-analyses of folate and vitamin E intake and breast cancer risk, MetaFemina demonstrated higher fidelity in extracting key evidence than manual reviews, and identified additional eligible articles missed by manual screening (27 and 13, respectively). Among 446 exposures, β-carotene, calcium, lutein, soy, vitamin D, and zeaxanthin were associated with lower risks of all three cancers. Green tea was associated with lower ovarian and uterine cancer risks, total dietary fiber intake with lower breast and ovarian cancer risks, and vitamin C and whole-grain intake with lower breast and uterine cancer risks. Processed meat intake was associated with higher breast and ovarian cancer risks, whereas alcohol was associated with higher breast cancer risk and animal protein intake with higher uterine cancer risk. No exposure showed statistically significant associations in opposite directions across cancer types.

Cluster analysis of longitudinal associations between microbial abundance and environmental factors

Lan Xue, Oregon State University

Emily Palmer (Oregon State University); Caihong Qin (Indiana university); Yuan Jiang (Oregon State University); Lan Xue (Oregon State University)

Microbial communities are dynamic systems that shift in composition in response to host and environmental factors. Understanding these temporal patterns is essential for identifying taxa with shared biological roles. However, standard clustering approaches for microbiome data often neglect its longitudinal structure and fail to account for external variables that may influence microbial trajectories. We introduce a novel clustering framework named *cleverly*, standing for (Cl)ustering with (E)xternal (V)ariabl(e)s in (R) with (L)ongitudinal (Y)s, which incorporates external variables into the analysis of longitudinal count-based microbiome data. Our method accommodates the compositional and time-correlated nature of sequencing data and allows for clustering microbial taxa based either on their baseline behavior or on their responses to external perturbations. Using simulated data, we demonstrate that *cleverly* accurately recovers true groupings under a range of biologically realistic conditions, including varying correlation structures and compositional constraints. We further apply *cleverly* to longitudinal fecal microbiome data from wild African buffalo (*Syncerus caffer*), revealing distinct clusters of taxa that either maintain consistent trajectories or respond sharply to changes in resource availability. These findings underscore the importance of accounting for environmental covariates in microbiome clustering and illustrate how *cleverly* enables a deeper understanding of microbial function and host-microbiome interactions in dynamic ecosystems.

Efficient implementation of a semiparametric joint model for multivariate longitudinal biomarkers and competing risks time-to-event data

Emily Ouyang, University of California, Riverside

Shanpeng Li (Beijing Normal - Hong Kong Baptist University); Emily Ouyang (University of California, Riverside); Jin Zhou (University of California, Los Angeles); Xinping Cui (University of California, Riverside); Gang Li (University of California, Los Angeles)

Joint modeling has become increasingly popular for characterizing the association between one or more longitudinal biomarkers and competing risks time-to-event outcomes. However, semiparametric multivariate joint modeling for large-scale data encounter substantial statistical and computational challenges, primarily due to the high dimensionality of random effects and the complexity of estimating nonparametric baseline hazards. These challenges often lead to prolonged computation time and excessive memory usage, limiting the utility of joint modeling for biobank-scale datasets. In this article, we introduce an efficient implementation of a semiparametric multivariate joint model, supported by a normal approximation and customized linear scan algorithms within an expectation-maximization (EM) framework. Our method significantly reduces computation time and memory consumption, enabling the analysis of data from thousands of subjects. The scalability and estimation accuracy of our approach are demonstrated through two simulation studies. We also present an application to the Primary Biliary Cirrhosis (PBC) dataset involving five longitudinal biomarkers as an illustrative example. A user-friendly R package, FastJM, has been developed for the shared random effects joint model with efficient implementation. The package is publicly available on the Comprehensive R Archive Network:

https://CRAN.R-project.org/package=FastJM.

Knowledge integration quantile regression (KIQR) in ultra-high dimensions

Yan Yu, University of Cincinnati

Jiantong Wang (Xavier University), Heng Lian (City University of Hong Kong), Yan Yu (University of Cincinnati), Tianhai Zu (University of Texas at San Antonio), and Heping Zhang (Yale University)

We propose a Knowledge Integration Quantile Regression (KIQR) framework for simultaneous variable selection and estimation in ultra-high dimensions, motivated by an application to identifying genetic risk factors for obesity from hundreds of thousands of single nucleotide polymorphisms (SNPs) in a cohort study with a limited number of participants. KIQR integrates prior information from external large-scale studies such as the GIANT consortium and UK Biobank into a penalized quantile regression framework via a novel knowledge integration mechanism that adaptively balances prior knowledge against observed data while remaining robust to potential prior misspecification. Focusing on high conditional quantiles of body mass index (BMI) most relevant to obesity risk, KIQR addresses key limitations of standard genome-wide association studies (GWAS): the stringent significance threshold, the univariate marginal mean association, and the lack of utilization of prior knowledge. We establish oracle properties and large sample properties of the penalized estimator under the non-differentiable check loss via Huber approximation, combined with a nonconvex SCAD penalty handled through local linear approximation. Simulation studies further demonstrate its effectiveness. Applied to the Framingham Heart Study, KIQR identifies three novel SNP associations: rs3798696 in TFAP2A, rs7070523 in ITIH5, and rs178260 in AIFM3, which have not previously been reported in the GWAS literature.

Session 41: Innovative Methods in Statistics and Data Science in Aging

Location: Navy Yard

Organizer: Jaime Lynn Speiser (Wake Forest University School of Medicine)

Chair: Jaime Lynn Speiser (Wake Forest University School of Medicine)

Presenters: Lihua Li; Moumita Chakraborty; Ansu Chatterjee; Panpan Zhang

Developing a weighted causal forest framework for population-level heterogenous treatment effect in geriatric research

Lihua Li, Icahn School of Medicine at Mount Sinai

Lihua Li (Icahn School of Medicine at Mount Sinai); Chen Yang (Icahn School of Medicine at Mount Sinai); Bian Liu (Icahn School of Medicine at Mount Sinai); Madhu Mazumdar (Icahn School of Medicine at Mount Sinai); Melissa Aldridge (Icahn School of Medicine at Mount Sinai)

Heterogeneous treatment effects (HTEs) are common in geriatric care settings, where patients are a highly diverse group and often respond differently to the same intervention. Geriatric research also often utilizes population-based surveys, especially when large-scale randomized clinical trials are infeasible. However current methods for evaluating HTEs using complex survey data often suffer from several limitations such as being prone to arbitrary subgroup selection and failing to account for survey designs (strata, clusters, and survey weights). To overcome these limitations, we propose a framework that combines population-based survey data with robust tree-based machine learning (ML) causal inference methods. Specifically, we develop a weighted causal forest (wCF) approach that integrates ML with propensity scores while accounting for survey design to estimate population-level HTEs. We then evaluate the performance of the proposed wCF method for both continuous and binary outcomes, through extensive simulations. Finally, using real-world data from the Medicare Current Beneficiary Survey, we apply the wCF method to evaluate population-level heterogeneous effects of physical activity on cognitive functioning and social isolation. This study provides practical guidance for applying the wCF method in geriatric care research and advances current parametric approaches for evaluating HTEs using complex survey data.

Modeling survival in aging populations through Bayesian reconstruction of life-course weight trajectories

Moumita Chakraborty, University of Texas Medical Branch

Moumita Chakraborty (University of Texas Medical Branch)

Understanding the cumulative impact of cardiovascular risk factors on mortality, particularly in aging populations, is a crucial area of public health research. With the rise of weight-loss interventions like GLP-1 receptor agonists, studies on their potential life expectancy benefits warrant careful investigation. While bringing down body weight, these medications cannot fully mitigate the long-term effects of a history of excess BMI and other pre-existing risk factors. Most current studies evaluating the longevity benefits of these interventions rely primarily on current body mass index (BMI), which can misestimate outcomes by overlooking obesity as a cumulative life-course exposure. This calls for data-driven approaches to quantify lifetime obesity exposure, as traditional models may fail to capture the cumulative biological consequences of decades of obesity on aging and mortality risk. To address this, we developed a Bayesian framework using mortality-linked NHANES data to reconstruct latent adult BMI trajectories from sparse retrospective measurements, then jointly modeled these with all-cause mortality to assess if life-course exposure adds prognostic value. Preliminary results demonstrate that cumulative obesity burden has a stronger effect on survival than current BMI alone, with individuals having similar current status but different histories exhibiting substantial survival differences, and

projected GLP-1 longevity gains varying significantly when life-course burden is incorporated. This work reframes obesity as a latent life-course exposure, emphasizing trajectory-aware survival modeling and offering a statistically principled approach to evaluate intervention impacts by propagating uncertainty from reconstructed weight histories. This methodology addresses the central statistical challenge of estimating survival effects from partially observed life-course exposures. While obesity and GLP-1 therapies provide a motivating application, the framework is broadly applicable to other aging-related cumulative exposures, such as smoking histories and blood pressure trajectories.

Rapid change detection in neuroimage data

Ansu Chatterjee, University of Maryland, Baltimore County

Vishal Subedi (UMBC); Ansu Chatterjee (UMBC)

Early detection of neurodegenerative diseases is critical in ensuring proper treatment and care of patients, and in containment of the disease. Owing to the plasticity of the brain, neuromage data are often governed by individualized parameters that make such rapid detections challenging. We propose techniques that combine statistical and artificial intelligence methods that leverage knowledge and information from multiple individuals, known neuroscience and biological sciences, and modern computational techniques for rapidly identifying changes in neuroimages from any single individual. Thus, our framework combined the power of using a cohort, yet potentially delivers inferences applicable in a personalized healthcare setting. Our proposed framework comes with probabilistic guarantees of quality and uncertainty. Algorithmic details and numerical results will be presented, and some theoretical findings outlined.

SIGNET: a signed network spectral clustering method for proteomic module discovery in Alzheimer's disease

Panpan Zhang, Vanderbilt University Medical Center

Zongyue Teng (Vanderbilt); Hailey Adegboye (Vanderbilt); Tim Hohman (Vanderbilt); Angela Jefferson (Vanderbilt); Dandan Liu (Vanderbilt); Panpan Zhang (Vanderbilt)

Signed networks arise naturally in many biomedical applications, where both positive and negative relationships contain distinct and important biological information. However, most existing community detection methods either ignore edge signs or rely on transformations that do not adequately preserve negative correlation interpretation, potentially obscuring meaningful structure. In this paper, we propose SIGNET, a spectral clustering framework for signed, weighted networks that directly incorporates both positive and negative edge information through the signed Laplacian. We further develop silhouette score-based strategies for selecting the number of clusters, including a data-driven silhouette-gap criterion. Through extensive simulations, we demonstrate that SIGNET achieves robust and accurate clustering performance across a range of network configurations, including varying signal strengths, cluster sizes, and structural heterogeneity. Compared with existing approaches, SIGNET exhibits improved stability in cluster number selection and more reliable recovery of underlying modular structure. We further apply SIGNET to a protein-protein co-expression network from the Vanderbilt Memory and Aging Project. The identified modules are biologically coherent and capture biological processes relevant to Alzheimer's disease, highlighting the practical utility of SIGNET in high-dimensional biomedical studies. An R package, SIGNET, is publicly available to support implementation.

Session 97: Transparent Machine Learning Through Visualization and Human-Centered Design

Location: Brighton Beach

Organizer: Zachery Boner; Lesia Semenova (Duke University; Rutgers University)

Chair: Jon Donnelly (Duke University)

Presenters: George H. Chen; Juliana Freire; Cynthia Rudin

Deep kernel Aalen-Johansen estimator: an interpretable and flexible neural net framework for competing risks

George H. Chen, Carnegie Mellon University

Xiaobin Shen (CMU); George H. Chen (CMU)

In statistics, the problem of "competing risks" is about modeling which critical event will happen next and when for a data point (such as a patient). The modeler specifies what the critical events are, where classically these critical events are different causes of death. In this paper, we propose an interpretable deep competing risks model called the Deep Kernel Aalen-Johansen (DKAJ) estimator, which generalizes the classical Aalen-Johansen nonparametric estimate of cumulative incidence functions (CIFs). Each data point is represented as a weighted combination of clusters. If a data point has nonzero weight only for one cluster, then its predicted CIFs correspond to those of the classical Aalen-Johansen estimator restricted to data points from that cluster. These weights come from an automatically learned kernel function that measures how similar any two data points are. On four standard competing risks datasets, we show that DKAJ is competitive with state-of-the-art baselines while being able to provide visualizations to assist model interpretation.

Beyond chat: visual interfaces for interpretable AI-assisted data harmonization

Juliana Freire, New York University

Juliana Freire (New York University)

Data harmonization—integrating datasets that differ in schema, terminology, and representation—is a persistent bottleneck across data-driven fields, and biomedical research offers a particularly acute example. Genomic and

proteomic profiles, clinical records, and imaging data are collected by different institutions using different standards, and combining them for analysis still requires substantial manual effort, even for domain experts.

Large language models offer new opportunities for addressing this problem: they bring broad background knowledge and semantic reasoning capabilities that traditional matching algorithms lack. But harmonization remains a complex task, and human validation remains essential.

I discuss our experience developing two AI-assisted systems for biomedical data harmonization in the context of the ARPA-H Biomedical Data Fabric (BDF) program: a no-code, agentic system that lets users construct and execute harmonization workflows through a chat-based interface, and a visual analytics system that combines automated matching, interactive visualization, and LLM-generated explanations to support human validation. Contrasting these approaches reveals an important distinction. Conversational interaction works well for constructing a harmonization workflow step by step, but does not scale to validation, where users may need to assess hundreds or thousands of candidate attribute matches. At this scale, visualization can expose patterns and relationships across many matches at once, while AI can provide the context—explanations, domain knowledge, and supporting evidence—needed to understand individual suggestions and make informed decisions efficiently.

This experience also suggests a broader view of interpretability for data-centric AI: it is not only about explaining individual model outputs, but also about designing interfaces that enable users to inspect evidence, compare alternatives, recognize patterns, and validate decisions at scale. I close by discussing open questions in human-AI collaboration for data-centric tasks: when systems should act autonomously versus defer to humans, how uncertainty, rationale, and evidence should be communicated visually, and what these questions suggest more broadly for the design of interpretable and trustworthy AI-assisted data systems.

The Rashomon effect for dimension reduction for data visualization

Cynthia Rudin, Duke University

Cynthia Rudin (Duke), Yiyang Sun (Duke), Haiyang Huang (Google), Gaurav Parikh (Duke)

Dimension reduction (DR) is inherently non-unique: multiple embeddings can preserve the structure of high-dimensional data equally well while differing in layout or geometry. I will define the Rashomon set for DR—the collection of 'good' embeddings—and show how embracing this multiplicity leads to more powerful and trustworthy representations. Then, I will introduce PCA-informed alignment to steer embeddings toward principal components, making axes interpretable without distorting local neighborhoods. Second, I will introduce concept-alignment regularization that aligns an embedding dimension with external knowledge, such as class labels or user-defined concepts. Third, I will propose a method to extract common knowledge across the Rashomon set by identifying trustworthy and persistent nearest-neighbor relationships, which is used to construct refined embeddings with improved local structure while preserving global relationships. By moving beyond a single embedding and leveraging the Rashomon set, this work provides a flexible framework for building interpretable, robust, and goal-aligned visualizations.

Parallel Sessions 9: Tuesday, November 3, 2026, 3:30 pm–5:00 pm

Session 87: Human-AI Collaboration

Location: Salon C

Organizer: Lihua Lei (Stanford University)

Chair: Lihua Lei (Stanford University)

Presenters: Yifan Wu; Lezhi (Carrie) Tan; Zhun Deng; Will Dobbie

Perfectly truthful calibration errors

Yifan Wu, Microsoft Research, New England

Jason Hartline (Northwestern University); Lunjia Hu (Northeastern University); Yuxuan Lu (Peking University); Yifan Wu (Microsoft Research);

Calibration requires probabilistic predictions to be unbiased conditional on their reported probabilities. Yet evaluating calibration on a finite random sample can reward predictors for misreporting their true probabilities. A calibration measure is truthful if reporting the true probabilities minimizes its expected value. Prior work by Haghtalab et al. (2024) and Qiao and Zhao (2025) developed approximately truthful calibration measures for sequential prediction, but perfectly truthful measures remained unknown even in the simpler batch setting. This talk presents a general construction of perfectly truthful calibration measures (Hartline, Hu, and Wu, 2026). The resulting family includes a continuous calibration measure and quantile-binned ℓ_2-ECE. We also present an extension to multiclass prediction. The empirical results on multiclass prediction task suggest that truthful calibration measures are less sensitive to estimation hyperparameters (Lu, Wu, Hartline, and Hu, 2026).

When should humans step in? Optimal human dispatching in AI-assisted decisions

Lezhi (Carrie) Tan, Stanford University

Lezhi Tan (Stanford University); Naomi Sagan (Stanford University); Lihua Lei (Stanford University); Jose Blanchet (Stanford University)

AI systems increasingly assist decision making by producing cheap factor-level assessments of complex inputs, but these assessments are often biased and incomplete. We study selective human oversight: given AI

signals, which factors should be escalated to a costly human evaluator? We formulate this human–AI collaboration problem as a factor-level information-acquisition problem. Under squared loss, solving for optimal dispatching reduces to maximizing a contextual reward; under a linear model, this reward admits a closed-form decomposition into two interpretable terms: predictive relevance and residual uncertainty in the human evaluation after conditioning on the AI signal.

We instantiate the framework on peer review, decomposing papers into ten aspects and evaluating 3,408 ICLR submissions across three LLMs and multiple regression heads. Outputs based on the optimal dispatching rules match full-human-review performance using only 2–3 queried aspects out of 10. The framework remains effective when the human signal comes from a single noisy reviewer and across review years, suggesting robustness to reviewer inconsistency and and temporal distribution shift. These results position principled dispatching as a practical foundation for scalable human–AI decision systems.

Ask to learn: query-based collaborative reinforcement learning

Zhun Deng, University of North Carolina at Chapel Hill

Kaicheng Zhang (UNC), Hongqian Niu (UNC), Laixi Shi (JHU), Didong Li (UNC), Guannan Qu (CMU), Zhun Deng (UNC)

Offline reinforcement learning is fundamen- tally limited by the coverage of the avail- able data, while collecting additional tra- jectories through online exploration may be costly, unsafe, or infeasible. We study whether this limitation can be overcome without di- rectly interacting with the underlying environ- ment. Motivated by practical settings such as collaborator-mediated simulation, where a collaborator can provide information com- plementary to the offline data, we introduce query-based collaborative reinforcement learn- ing to study how to acquire this information efficiently. The learner starts with a fixed offline dataset and adaptively queries state– time pairs; the collaborator samples an ac- tion according to a fixed, possibly subopti- mal policy, and the simulator returns the corresponding reward and next state. For finite-horizon tabular problems, we derive an instance-dependent information-theoretic lower bound on the additional queries required to identify an ϵ-optimal policy and develop a local quadratic approximation. Guided by this characterization, we develop an adaptive query-allocation algorithm whose query com- plexity nearly matches the local benchmark under suitable conditions. Experiments on multiple synthetic and semi-synthetic environ- ments demonstrate the empirical efficiency of the proposed method.

Human decisions and machine predictions in multistage systems

Will Dobbie, Princeton University

Jason Baron (Duke University) ; Will Dobbie (Princeton University); Richard Lombardo (Harvard University); Ashesh Rambachan (Massachusetts Institute of Technology); Joseph Ryan (University of Michigan)

Many high-stakes decisions unfold through multistage systems, in which cases pass through successive stages of human review before a final decision is reached. We examine how machine learning can be used to improve human decision-making in multistage systems, in which an algorithm's value depends on where in the system it is deployed.

Session 100: The Role of the Rashomon Effect in Responsible AI

Location: Williamsburg

Organizer: Chudi Zhong (University of North Carolina at Chapel Hill)

Chair: Jon Donnelly (Duke University)

Presenters: Tong Wang; Flavio Calmon; Sanghamitra Dutta; Lesia Semenova

From model explanation to data misinterpretation: a cautionary analysis of post hoc explainers in empirical research

Tong Wang, Yale University

Tong Wang (Yale University), Ronilo Ragodos (University of New Hampshire), Lu Feng (Tsinghua University), Yu Jeffrey Hu (Purdue University)

Post hoc explainers such as SHAP and LIME are used widely in business research to interpret complex machine learning models. Although they were designed to explain model predictions, there has been an increasing trend in which the explanations they generate are treated as evidence about underlying data relationships. Based on a systematic review of more than four hundred studies, we document that this explanation interpretation is widespread and examine its validity. We also evaluate how well post hoc explanations reliably recover the direction and relative importance of features in true data-generating process. We introduce two metrics—direction alignment and strength alignment—and assess SHAP and LIME using simulated data with known ground truth. Although explanations often appear reasonable on average, they exhibit substantial heterogeneity in their alignment and can fail to be aligned even when their predictive accuracy is high. High predictive performance is therefore necessary but insufficient for reliable explanation. We further show that feature correlation and the Rashomon effect (where many equally accurate models rely on different feature attributions) are key drivers of misalignment. Agreement in explanations across such models provides a practical diagnostic of reliability. Overall, our findings caution against using post hoc explainers for hypothesis validation and instead position them as exploratory tools that generate, rather than confirm, substantive insights.

Circuits of LLM classifiers

Flavio Calmon, Harvard University

Sajani Vithana (Harvard University); Hadi Khalaf (Harvard University); Kunal Talreja (Georgia Tech); Haewon Jeong (University of California, Santa Barbara); Viveck R. Cadambe (Georgia Tech); Flavio Calmon (Harvard University)

Language models make it easy to build classifiers by simply prompting a model with the sample to classify. However, their predictions are neither deterministic nor uniformly reliable across samples. How can we combine predictions from different models to improve reliability? We view LLM classifiers as noisy measurement instruments of an underlying target label. Our goal is to maximize average classification accuracy at a given average query cost. To this end, we introduce circuits that organize LLM classifiers into an adaptive sequence of queries and decisions. We then characterize the optimal circuit, obtaining limits of achievable accuracy, and provide practical insights on realizing the optimal circuit. Across safety, healthcare, political science and preference benchmarks, our method improves the cost–accuracy frontier over routing and cascading baselines.

Trustworthy AI under model multiplicity: from inconsistent predictions to unreliable recourse

Sanghamitra Dutta, University of Maryland

Faisal Hamman, Pasan Dissanayake, Saumitra Mishra, Freddy Lecue, Sanghamitra Dutta

As AI percolates into high-stakes applications such as finance, hiring, and education, a critical pillar of trustworthiness is consistency under model multiplicity. How can we trust AI if equally well-performing models make conflicting predictions on the same applicant? This problem becomes more pronounced for algorithmic recourse: an institution might provide counterfactual explanations to rejected applicants on how to flip a model's decision, yet those explanations can become invalid if the model is slightly updated (e.g., changing the seed or hyperparameters). This talk addresses the challenge of fine-tuning multiplicity in Tabular LLMs. We provide a novel, retraining-free metric – that we call Stability – that quantifies the robustness of individual predictions by sampling the model's local behavior in the embedding space and comes with desirable mathematical properties. Next, we show that data points with a sufficiently high Stability value, as defined by our metric, will remain valid across a broad class of equally well-performing fine-tuned models with high probability. We also bridge this challenge to algorithmic recourse, shifting the focus from prediction stability to the robustness of actionable counterfactual explanations under model updates. Our framework leads to a unified, computationally efficient paradigm for certifying individual-level robustness under model multiplicity without expensive retraining.

Trustworthy machine learning under the Rashomon effect

Lesia Semenova, Rutgers University

Lesia Semenova (Rutgers)

In many real-world problems, the data do not single out one best model. Instead, there exists a Rashomon set of near-optimal models that achieve similar accuracy yet can behave very differently, a phenomenon Leo Breiman called the Rashomon effect. Standard practice reports a single model and discards the rest. I will argue that this multiplicity should instead be placed at the center of trustworthy machine learning. Rather than designing a new method for each trustworthiness goal, I will show how searching within the Rashomon set can reveal accurate models with desired properties such as interpretability, robustness, privacy, and fairness. Yet multiplicity cuts both ways: having many good models that behave differently can make a system more reliable, but also more vulnerable. I will discuss when multiplicity helps, when it hurts, and what this means for building trustworthy models.

Session 07: New Statistical Tools for High-Dimensional Biomedical Data Analysis

Location: Fulton Ferry

Organizer: Boxiang Wang (University of Iowa)

Chair: Boxiang Wang (University of Iowa)

Presenters: Xiaoxiao Sun; Chong Jin; Li-Xuan Qin; Siyu Heng

Interpretable scalar-on-image regression models via the generalized Dantzig selector

Xiaoxiao Sun, University of Arizona

Sijia Liao (University of Arizona); Xiaoxiao Sun (University of Arizona); Ning Hao (University of Arizona); Helen Hao Zhang (University of Arizona)

The scalar-on-image regression model examines the association between a scalar response and a bivariate function (e.g., images) through the estimation of a bivariate coefficient function. Existing approaches often impose smoothness constraints to control the bias-variance trade-off, and thus prevent overfitting. However, such assumptions can hinder interpretability, especially when only certain regions of an image influence changes in the response. In such a scenario, interpretability can be better captured by imposing sparsity assumptions on the coefficient function. To address this challenge, we propose the Generalized Dantzig Selector, a novel method that jointly enforces sparsity and smoothness on the coefficient function. The proposed approach enhances interpretability by accurately identifying regions with no contribution to the changes of response, while preserving stability in estimation. Extensive simulation studies and real data applications demonstrate that the new method is highly interpretable and achieves notable improvements over existing approaches. Moreover, we rigorously establish non-asymptotic bounds for the estimation error, providing strong theoretical guarantees for the proposed framework.

Single-cell resolution deconvolution of bulk RNA-seq

data via functional regression

Chong Jin, New Jersey Institute of Technology

Xiaotian Mu (New Jersey Institute of Technology); Chong Jin (New Jersey Institute of Technology)

The proliferation of single-cell RNA sequencing (scRNA-seq) data has spurred the development of methods to deconvolve bulk RNA-seq tissues. However, conventional approaches typically rely on discrete cell-type classifications, which overlook dynamic cell states and create a false dichotomy between changes in cell composition and cell-type-specific expression. Alternatively, strategies that map bulk data to cell subpopulations often suffer from arbitrary parameter choices and "hard" classifications. We propose BUDGIE (Bulk RNA-seq Deconvolution Using Generalized Inference and Estimation), a novel statistical framework that deconvolves bulk RNA-seq data at single-cell resolution. BUDGIE models bulk expression using a functional linear regression framework, representing the bulk sample as an abundance surface over the single-cell reference atlas. By combining a continuous representation with discrete cell-type indicators, this approach captures continuous cell-state transitions, and produces an estimated abundance for each reference cell that can also be aggregated to obtain conventional cell-type proportions. We evaluate BUDGIE in simulations across both cell and cell-type resolutions: BUDGIE achieves lower cell-level estimation error than existing methods, while benchmarks on simulated mixtures show performance competitive with established deconvolution methods. Finally, we demonstrate the utility of BUDGIE by applying it to resolve fine-scale cellular heterogeneity in bulk cancer tissue data.

AI-assisted synthesis of realistic microbiome data

Li-Xuan Qin, Memorial Sloan Kettering Cancer Center

Xinyi Wang (University of California, Davis); Yunhui Qi (Dana-Farber Cancer Institute); Teng Fei (MSKCC); Li-Xuan Qin (MSKCC)

Microbiome data present distinctive challenges for synthetic data generation because they are highly sparse, over-dispersed, and compositional. Generative models that reconstruct abundance profiles without explicitly modeling presence-absence patterns may therefore fail to preserve key characteristics required for valid downstream inference. We developed a Presence-Abundance Variational Autoencoder (PA-VAE), using a shared latent representation with separate decoder components for taxon presence and abundance, designed to preserve both taxon-specific sparsity and abundance distributions while maintaining multivariate community structure. We evaluated PA-VAE using three 16S microbiome datasets, spanning inflammatory bowel disease, pregnancy-associated vaginal microbiomes, and patients undergoing allogeneic hematopoietic cell transplantation. Evaluation included preservation of overall and taxon-specific sparsity, abundance distribution summaries, alpha and beta diversity, multivariate clustering, clinically relevant ecological patterns such as Enterococcus dominance, and concordance of differential-abundance analyses between observed and synthetic data. Across datasets, PA-VAE more faithfully reproduced the original data than a vanilla VAE and MIDASim. PA-VAE provides a flexible framework for realistic microbiome data augmentation and may support simulation, benchmarking, and prospective study-design applications in microbiome research.

Propensity score propagation: a general framework for design-based inference with unknown propensity scores

Siyu Heng, New York University

Siyu Heng (New York University); Yanxin Shen (Pennsylvania State University); Zijian Guo (Zhejiang University)

Design-based inference, also known as randomization-based or finite-population inference, provides a principled framework for trustworthy statistical inference. It attributes randomness solely to the design mechanism, such as treatment assignment, survey sampling, or missingness, without imposing super-population distributional or modeling assumptions on the outcome data. From the seminal work of Fisher and Neyman to its recent resurgence, design-based inference has played a central role in causal inference, survey sampling, and missing data analysis. However, its use in many modern applications has been limited by a fundamental obstacle: existing design-based inference methods typically assume that propensity scores (i.e., design probabilities) are known, whereas in observational studies, real-world surveys, missing data problems, and many other settings, propensity scores are usually unknown and must be estimated from data. We propose propensity score propagation, a general framework for valid design-based inference with unknown propensity scores. The framework uses a regeneration-and-union procedure to propagate uncertainty from propensity score estimation into downstream design-based inference without introducing super-population assumptions about the outcomes. It accommodates both parametric and nonparametric propensity score estimation, supports flexible machine learning estimators, integrates seamlessly with existing design-based inference methods developed for settings with known propensity scores, and applies broadly across design-based problems. Theoretical and simulation results show that the proposed framework achieves nominal coverage, even when existing approaches exhibit substantial under-coverage.

Session 02: Learning, Inference and Decision Making in Complex Structures

Location: Sunset Park

Organizer: Yanyuan Ma (Penn State University)

Chair: Tianying Wang (Colorado State University)

Presenters: Tianxi Cai; Donglin Zeng; Yuanjia Wang; Tianying Wang

Title TBA

Tianxi Cai, Harvard University

Abstract TBA

Debiased inference for high-dimensional regression models based on profile M-estimation

Donglin Zeng, University of Michigan

Yuhao Deng (Fred Hutch Cancer Center); Yi Wang (Shanghai University of International Business and Economics); Yu Gu (Hong Kong University); Yuanjia Wang (Columbia University); Donglin Zeng (University of Michigan)

Debiased inference for high-dimensional regression models has received substantial recent attention to ensure regularized estimators have valid inference. All existing methods focus on achieving Neyman orthogonality through explicitly constructing projections onto the space of nuisance parameters, which is infeasible when an explicit form of the projection is unavailable. We introduce a general debiasing framework, Debiased Profile M-Estimation (DPME), which applies to a broad class of models and does not require model-specific Neyman orthogonalization or projection derivations as in existing methods. Our approach begins by obtaining an initial estimator of the parameters by optimizing a penalized objective function. To correct for the bias introduced by penalization, we construct a one-step estimator using the Newton-Raphson update, applied to the gradient of a profile function defined as the optimal objective function with the parameter of interest held fixed. We use numerical differentiation without requiring the explicit calculation of the gradients. The resulting DPME estimator is shown to be asymptotically linear and normally distributed. Through extensive simulations, we demonstrate that the proposed method achieves better coverage rates than existing alternatives, with largely reduced computational cost. Finally, we illustrate the utility of our method with an application to estimating a treatment rule for multiple myeloma.

Deep representation learning for optimizing treatment decisions with EEG biomarkers

Yuanjia Wang, Columbia University

Bin Yang, Young-geun Kim, Yuanjia Wang

Antidepressant treatment response remains highly unpredictable, with moderate remission rates due to heterogeneous disease mechanisms. Electroencephalography (EEG) biomarkers have shown utility for guiding treatment decisions but pose challenges for direct use due to the need for manual feature extraction and lack of interpretability for their complex, high-dimensional properties. In this paper, we propose a unified deep learning framework, a MEHE (Multihead EEG-variational autoencoder for Heterogeneous treatment Effect) architecture, that incorporates an EEG-specific convolutional neural network-based encoder and extends the vanilla VAE model to estimate heterogeneous treatment effects (HTEs) and associated individualized treatment rules (ITRs). Our method jointly optimizes EEG representations and treatment effect estimation through a generative, outcome-guided latent space that achieves interpretability by learning meaningful EEG features. Our approach offers several advantages by providing 1) an end-to-end learning framework with a joint loss to eliminate the need for manual feature engineering; 2) flexibility in incorporating multi-modality data; 3) the ability to learn robust, interpretable EEG representations for accurate HTE and ITR estimation. Through simulations based on real data and application to a large randomized controlled trial, we demonstrate our method's superior performance in predicting optimal individualized treatment rules compared to existing approaches. The extracted latent encoding provides information on spatial and temporal EEG patterns that inform treatment decisions.

Data fusion for errors-in-variables

Tianying Wang, Colorado State University

Huali Zhao (Huazhong University of Science and Technology); Molei Liu (Peking University); Tianying Wang (Colorado State University)

We study errors-in-variables problems in which a target study contains only a single error-prone surrogate of an unobserved exposure, while an external study provides repeated surrogate measurements from a different population. The measurement error distribution is allowed to depend on the observed error-free variables, and the error-free variable distribution itself may differ between studies. We show that the target conditional measurement error distribution is identifiable if the conditional error distribution, given the observed error-free variables, is transportable from the external study to the target study. Building on this identification result, we develop a data-fusion estimator for a broad class of target functionals. The estimator combines conditional deconvolution, flexible nuisance estimation, and orthogonal correction that reduces first-order sensitivity to nuisance estimation. For the proposed estimator, we develop a unified spectral theory covering both diffuse-spectrum and finite atomic-spectrum target functionals, derive a general asymptotic expansion, and establish consistency and target-specific convergence-rate bounds. The resulting convergence-rate bounds depend jointly on the spectral properties of the measurement error, the latent exposure, and the target functional. Simulations demonstrate the robust performance of our method under various finite-sample settings. An application to the National

Health and Nutrition Examination Survey illustrates how accounting for population heterogeneity and error heteroscedasticity can change empirical conclusions.

Session 102: Making Sense of Multivariate Data

Location: Golden Boardroom

Organizer: Maryclare Griffin (University of Massachusetts Amherst)

Chair: Nathan Wycoff (University of Massachusetts Amherst)

Presenters: Li Ma; Anna Neufeld; Qian Zhao; Maryclare Griffin

Assessing generative models through density ratio learning

Li Ma, University of Chicago

Naoki Awaya (Waseda University); Yuliang Xu (Johns Hopkins University); Yun Wei (UT Dallas); Li Ma (University of Chicago)

Generative models have become a popular tool for simulating observations from complex, high-dimensional distributions. However, effectively assessing the quality of generative models remains a challenge. Standard approaches, such as reporting empirical distance estimates between real and generated samples, generally do not reveal the nature of the discrepancy, nor do they allow for reliable uncertainty quantification. We show that a powerful approach to understanding such discrepancies is to learn the density ratio between the generative model and the true data-generating model. We demonstrate several strategies for characterizing the discrepancy once reliable density ratio estimates and uncertainty quantification become available. Our starting point is a new loss function for learning density ratios under a two-sample design, based on the variational form of the Hellinger distance. Drawing a connection between this loss and the exponential loss commonly used in training additive tree ensembles for supervised learning, we show that the new loss can be used to train flexible function approximators such as tree ensembles and neural networks. We show that, for additive tree ensembles, tailored training algorithms that parallel boosting and BART are available, which help attain calibrated generalized Bayesian uncertainty quantification. For deep neural network–based approaches trained using standard gradient descent, we provide some theoretical guarantees in terms of consistency and convergence rates. We demonstrate how to use the inferred density ratios to carry out generative model diagnostics in the context of image data and microbiome compositions.

Inference after clustering: conditioning, data thinning, or data fission?

Anna Neufeld, Williams College

Anna Neufeld

Suppose we want to test for a difference in means across groups. When the groups are known a priori, numerous classical statistical tests are available. However, when the groups are unknown and estimated via clustering, the situation is more complex. If the same data are used for clustering and post-clustering inference, classical statistical tests fail to control the Type 1 error rate. Further complicating matters, traditional sample splitting approaches cannot be used in this setting. Motivated by the problem of testing for differential expression of genes between estimated cell types in the analysis of single-cell RNA sequencing data, we discuss the promises and limitations of three potential selective inference solutions in this setting: conditioning, data thinning, and data fission.

Beta-Trees: multivariate histograms with confidence statements

Qian Zhao, University of Massachusetts Amherst

Guenther Walther (Stanford University); Qian Zhao (University of Massachusetts Amherst)

Multivariate histograms are difficult to construct due to the curse of dimensionality. Motivated by k-d trees in computer science, we show how to construct an efficient data-adaptive partition of Euclidean space that possesses the following two properties: (1) with high confidence the distribution from which the data are generated is close to uniform on each rectangle of the partition; and (2) finite sample simultaneous confidence intervals can be provided for the probabilities of each rectangle in the partition. The method produces confidence intervals whose widths depend only on the probability content of the rectangles and not on the dimensionality of the space, thus avoiding the curse of dimensionality. Moreover, the widths essentially match the optimal widths in the univariate setting. The simultaneous validity of the confidence intervals allows us to use this construction, which we call Beta-trees, for various data-analytic purposes. We illustrate this by using Beta-trees for visualizing and for multivariate mode-hunting of the flow cytometry data.

Enough is enough: an improved stopping criterion for Markov chain Monte Carlo with many independent chains

Maryclare Griffin, University of Massachusetts Amherst

Maryclare Griffin (University of Massachusetts Amherst); Andee Kaplan (Colorado State University)

Despite the ubiquitous use of Markov Chain Monte Carlo methods for simulation from multivariate distributions, the availability of principled stopping criteria that do not make unrealistic assumptions on the initialization of a Markov Chain relative to the target multivariate distribution is limited. Furthermore, stopping criteria that are applicable for simulation from multivariate distributions mixed type are also limited. By leveraging the benefits of modern computation that allow for simultaneous simulation of many Markov chains, we introduce stopping criteria that are easy to compute, require minimal storage, and inherit desirable properties in a variety of applications. We demonstrate the effectiveness of these stopping criteria in

classification, clustering, and streaming settings.

Session 104: AI-Driven Clinical Intelligence: Integrating Generative Models, Predictive Analytics & Clinical Decision Making

Location: Carroll Gardens

Organizer: Xing Chen (Moderna)

Chair: Qiqi Deng (Moderna)

Presenters: Xing Chen (Moderna); Yilong Zhang (Meta)

Panel discussion: Xing Chen (Moderna); Yilong Zhang (Meta); Yue Shentu (Merck). Chaired by Qiqi Deng (Moderna).

From clinical knowledge to agentic analytics: leveraging AI across clinical data workflows

Xing Chen, Moderna — 30 minutes

Xing Chen (Moderna); Weijie Zhang (Moderna) ; Xiaolin Chang (Moderna) ; Mahtab Talaei (Moderna) ; You Wu (Moderna) ; Chenliang Xiao (Moderna) ; Chen Zhao (Moderna) ;

AI has significant potential to transform how clinical teams access study knowledge, generate analyses, and make data-driven decisions. However, applying AI in clinical trials requires a strong focus on accuracy, reproducibility, traceability, and appropriate governance. In this presentation, we will share our work on building a study-specific knowledge base that brings together clinical documents, definitions, and analytical context to support AI-enabled workflows. We will demonstrate how this foundation, combined with controlled and reusable analytical libraries, can support TFL generation and exploratory analyses, while highlighting key challenges and caveats. This is particularly important for TFLs, where highly regulated processes and stringent accuracy requirements make unconstrained AI generation challenging. We will also share how we evaluate AI capabilities using sample-size calculation benchmarks to better understand their accuracy, robustness, and limitations. Finally, through an oncology use case, we will demonstrate agentic workflows using Codex for ad hoc exploratory analyses and examine their efficiency and computational cost. For these clinical use cases, we partner with the Digital organization on the enabling infrastructure, data foundations, and data flywheel needed to continuously improve AI capabilities. Together, the examples illustrate a practical path from AI experimentation toward a governed, reusable, and measurable AI ecosystem for clinical trials.

YAMAA: a language-neutral YAML specification for AI-first clinical data workflows from ODM to ADaM

Yilong Zhang, Meta — 30 minutes

Yilong Zhang (Meta)

Statistical programmers transform clinical trial data into CDISC SDTM and ADaM datasets to support regulatory submissions. As AI agents become part of this process, explicit, reproducible, and reviewable specifications are essential. We introduce yamaa (`https://elong0527.github.io/yamaa`), a language-neutral YAML specification for defining clinical data pipelines that transform Operational Data Model (ODM) data into SDTM and ADaM datasets following CDISC standards. Designed for AI-first workflows, YAMAA expresses derivation logic in a structured format that supports version control, human review, and consistent implementation across programming languages. We used the framework to generate a 190-item benchmark (`https://elong0527.github.io/yamaa/benchmark/`) for evaluating agentic workflows and to assess consistency in four publicly available clinical trial submission packages (`https://rconsortium.github.io/submissions-pilot7-synthetic-data/studies/`). YAMAA provides a common specification for AI-assisted clinical data transformation, supporting reproducibility and reviewability from source data through analysis-ready datasets.

Panel discussion and audience Q&A

Chair: Qiqi Deng (Moderna)

Panelists: Xing Chen (Moderna), Yilong Zhang (Meta), and Yue Shentu (Merck).

Session 13: Navigating the AI Transformation in Biostatistics: Opportunities, Methods, and Evolving Practice

Location: Coney Island

Organizer: Wen Li (Pfizer)

Chair: Wen Li (Pfizer)

Presenters: Aiying Zhang; Maria Kudela; Junshui Ma; Margaret Gamalo (Discussant)

Beyond black-box AI: adaptive integration of scientific priors for brain disorder prediction

Aiying Zhang, University of Virginia

Song Wang, Javier Rasero, Chirag Agarwal, Aiying Zhang

As AI methods proliferate across biomedical sciences, a critical tension has emerged: deep learning models achieve strong predictive performance but operate as black boxes, offering little insight into the biological mechanisms that drive their predictions. This opacity is particularly consequential in psychiatric neuroimaging, where clinically actionable biomarkers must not only predict but also explain. We present ConceptNeuro, a concept-based diagnostic framework for resting-state fMRI data that moves beyond black-box graph neural networks (GNNs) by grounding predictions in interpretable, neurobiologically meaningful concepts. Rather than learning opaque edge representations, ConceptNeuro leverages large language models (LLMs) and expert domain knowledge to automatically generate, filter, and encode structured subgraphs — each representing a candidate functional

connectivity pattern between specific brain regions. Predictions are then made through a concept classifier, ensuring that every diagnostic decision is traceable to clinically coherent connectivity signatures. Evaluated across multiple pediatric psychiatric disorder datasets — spanning ADHD, anxiety, and autism spectrum disorder — ConceptNeuro-augmented GNNs consistently outperform their vanilla counterparts, improving both accuracy and transparency. Post-hoc concept analyses recover disorder-specific connectivity patterns that align with established neuroscience and generate testable hypotheses for future investigation. This work illustrates a broader principle for biostatisticians navigating the AI transformation: scientific priors are not a constraint on model flexibility — they are a source of both interpretability and generalization. We discuss implications for study design, model evaluation, and the evolving role of domain knowledge integration in AI-driven biomedical research.

Leveraging graph neural networks in biopharmaceutical research for smarter decisions

Maria Kudela, Pfizer

Maria Kudela (Pfizer), Yuxi Zhao (Pfizer), Margaret Gamalo (Pfizer)

The growing availability of biomedical data creates new opportunities to support decision-making across the drug development lifecycle. At the same time, the diversity of these data presents substantial analytical challenges. Information relevant to clinical development may include characteristics of the investigational treatment, target disease, trial design, pharmacokinetic and pharmacodynamic profiles, and evidence from previous studies. These sources vary considerably in structure, format, and completeness, making their integration and interpretation difficult.

Graph neural networks offer a useful framework for representing complex relationships among heterogeneous data sources. This presentation will examine the Hierarchical Interaction Network introduced by Fu et al. (2022) for clinical trial outcome prediction. The method combines multimodal data and utilizes a hierarchical interaction graph to represent relationships among datasets and generate trial-level predictions.

We will provide an overview of the architecture and discuss considerations relevant to its use in biopharmaceutical research. The presentation will also describe our exploration of selected modifications to the framework and the preparation of datasets intended to reflect practical features of drug development data. The presentation will consider both the opportunities and limitations of graph-based learning for integrating diverse sources of clinical information.

Harnessed AI agents: a framework for democratizing statistical expertise

Junshui Ma, Merck

Junshui Ma

In the AI era, we work with harnessed AI agents. This talk introduces the basic concept of a personalized harness: a dynamic system of skills, tools, subagents, memory, and controls built around general-purpose AI agents.

To demonstrate its potential, I will show how my harnessed agents completed three connected tasks within two to three hours. First, they helped me rapidly develop a deep understanding of the three most-cited papers published in Statistics in Medicine during the past two years, supported by structured reading and independent AI review. Next, we transformed the guidance from one paper into a reusable agent skill for multiple imputation of longitudinal data. Finally, the agent applied that skill to a concrete statistical problem and selected an appropriate method that recovered the full-data regression coefficient and between-subject variance.

These increasingly capable agents confront statisticians with a critical question: if knowledge of statistical methods becomes widely accessible, where will our future value lie? The three highly cited papers suggest a surprisingly consistent answer. The Our greatest value indeed lies far beyond individual methods...

Discussant: Margaret Gamalo, Pfizer, Inc

Session 80: Spatial Causal Inference

Location: Park Slope

Organizer: Ted Westling (University of Massachusetts Amherst)

Chair: Ted Westling (University of Massachusetts)

Presenters: Matias D. Cattaneo; Elena Dal Torrione; Anik Burman; Samrachana Adhikari

Estimation and inference in boundary discontinuity designs: location-based methods

Matias D. Cattaneo, Princeton University

Matias D. Cattaneo, Rocio Titiunik, Rae (Ruiqi) Yu

Boundary discontinuity designs are used to learn about causal treatment effects along a continuous assignment boundary that splits units into control and treatment groups according to a bivariate location score. We analyze location-based local polynomial treatment effect estimators that directly employ the bivariate score of each unit. We develop pointwise and uniform estimation and inference methods for the Boundary Average Treatment Effect Curve (BATEC), as well as for two aggregated causal parameters: the Weighted Boundary Average Treatment Effect (WBATE) and the Largest Boundary Average Treatment Effect (LBATE). Our results cover both sharp and fuzzy (imperfect compliance) designs. We illustrate the methods with an empirical application, and provide companion general-purpose software. The supplemental appendix includes additional substantive theoretical results, methodological details, and simulation evidence.

Estimating spatial spillover effects in regression discontinuity designs: evidence from Turkish administrative data

Elena Dal Torrione, Yale University

Elena Dal Torrione (Yale University); Tiziano Arduini (Tor Vergata University of Rome); Laura Forastiere (Yale University)

We study Regression Discontinuity Designs with spatial interference, where a unit's outcome may depend on both its own treatment and the treatments assigned to geographically neighboring units. In this setting, assignment to an "effective treatment" is determined jointly by the unit's running variable and those of their neighbors, generating multidimensional effective-treatment boundaries. Spillover effects are identified locally along these boundaries; in practice, however, few observations may lie near the boundaries, leading to limited statistical precision and effects that are local to particular configurations of neighbors' running variables. In this paper, we develop an approach that imposes additional structure on conditional outcome functions and neighbors' running variables to use observations away from the boundary, leading to both improved statistical precision and identification of spillover effects for broader subpopulations. We revisit Meyersson (2014) to estimate the spillover effects of the election of an Islamic mayor in Turkey on female high school education in neighboring municipalities.

Robust spatial confounding adjustment via basis voting

Anik Burman, Johns Hopkins University

Anik Burman (Johns Hopkins University); Elizabeth L. Ogburn (Johns Hopkins University); Abhirup Datta (Johns Hopkins University)

Estimating effects of spatially structured exposures is complicated by unmeasured spatial confounders, which undermine identifiability in spatial linear regression models unless structural assumptions are imposed. We develop a general framework for effect estimation in spatial regression models that relaxes the commonly assumed requirement that exposures contain higher-frequency variation than confounders. We propose Basis Voting, a plurality-rule estimator - novel in the spatial literature - that consistently identifies causal effects only under the assumption that, in a spatial basis expansion of the exposure and confounder, there exist several basis functions in the support of the exposure but not the confounder. This assumption generalizes existing assumptions of differential basis support used for identification of the causal effect under spatial confounding, and does not require prior knowledge of which basis functions satisfy this support condition. We design this estimator as the mode of several candidate estimators each computed based on a single working basis function. We also show that the standard projection-based candidate estimator typically used in other plurality-rule based methods is inefficient, and provide a more efficient novel candidate. Extensive simulations and a real-world application demonstrate that our approach reliably recovers unbiased causal estimates whenever exposure and confounder signals are separable on a plurality of basis functions. By not relying on higher-frequency variation, our method remains applicable to settings where exposures are smooth spatial functions, such as distance to pollution sources or major roadways, common in environmental studies.

A Bayesian approach to estimate causal peer influence effect

Samrachana Adhikari, New York University Grossman School of Medicine

Seungha Um (New York University Grossman School of Medicine), Tracy Sweet (University of Maryland) and Samrachana Adhikari (New York University Grossman School of Medicine)

Identifying and estimating causal peer influence in social network is challenging due to confounding by homophily, where people tend to connect with those who share similar characteristics. Moreover, since all the attributes driving homophily are generally not always observed, identifying and estimating causal peer influence becomes infeasible using standard causal identification assumptions. In this work, we address this challenge by leveraging latent locations inferred from the network to disentangle homophily from causal peer influence, and extend this approach to multiple networks by adopting a Bayesian hierarchical modeling framework. To accommodate nonlinear peer influence on individual behavior, we employ a Bayesian nonparametric method, Bayesian Additive Regression Trees, and propose a Bayesian framework that accounts for uncertainty in inferring latent locations.

Session 99: Digital Twins and Synthetic Data for Clinical Research

Location: Windsor Terrace

Organizer: Qilu Yu (NIH, National Center for Complementary and Integrative Health)

Chair: Qilu Yu (NIH, National Center for Complementary and Integrative Health)

Presenters: Margaret Gamalo; Daniele Bertolini; Bryan E. Shepherd; Ruthie Davi

Regulatory-grade UC digital twin with multi-endpoint causal inference and validation

Margaret Gamalo, Pfizer

Margaret Gamalo (Pfizer)

We developed a regulatory-grade digital twin framework for ulcerative colitis using harmonized data from ELEVATE UC 12, ELEVATE UC 52, and GLADIATOR. The framework combines estimand-based causal inference, observed-variable causal graphs, variable provenance controls, standardized g-computation, augmented inverse probability weighting (AIPW), and longitudinal trajectory modeling to evaluate multiple clinically relevant endpoints, including clinical remission, clinical response, endoscopic remission, Mayo measures, CRP, fecal calprotectin, and IBDQ. Internal validation incorporates overlap diagnostics, estimator triangulation, leave-one-study-out cross-study validation, and a noncompensatory credibility qualification framework. The platform generates treatment-specific counterfactual outcome distributions and longitudinal disease trajectories within a fully auditable workflow, providing a scalable foundation for regulatory evaluation, external-control generation, transportability assessment, and pediatric extrapolation of digital twin methodologies.

Building digital twins from real-world data and trial summary statistics

Daniele Bertolini, Unlearn.AI

Daniele Bertolini (Unlearn.AI); Franklin Fuller (Unlearn.AI); Samantha Liang (Unlearn.AI); Jason Christopher (Unlearn.AI); Aaron M. Smith (Unlearn.AI)

Digital twins are computational models of patients and can generate predictions of clinical outcomes under a defined treatment or standard of care. In clinical trials, they can be used to simulate cohorts, construct synthetic control comparators, and reduce the required sample size in randomized controlled studies.

Machine learning enables flexible digital twin models that can integrate heterogeneous and multimodal data sources during training. We focus on oncology trials, where two complementary data sources are often available: rich patient-level real-world data, including molecular and genomic information, and evidence from prior clinical trials that is often accessible only through published aggregate results.

We train a foundation model on patient-level real-world data and propose a framework for recalibrating it using clinical trial summary statistics. The recalibration corrects biases arising from differences between real-world and trial settings while minimally altering the model beyond what is required to reproduce the new evidence, thereby preserving information learned from the broader patient-level dataset.

This approach provides a path toward a continuously updated foundation model in which new trials can contribute information even when patient-level data are unavailable. We illustrate how digital twins generated by these models can support synthetic controls for single-arm trials and trial simulation.

Generating realistic synthetic longitudinal HIV cohort data using latent diffusion models

Bryan E. Shepherd, Vanderbilt University

Zhuohui Liang (Vanderbilt University); Chao Yan (Vanderbilt University); Bradley A. Malin (Vanderbilt University); Bryan E. Shepherd (Vanderbilt University);

Sharing international HIV cohort data is increasingly difficult due to privacy regulations and data governance constraints. Synthetic data offer a promising alternative, but generating high quality data that reflect complex real world cohorts remains challenging. HIV cohort data consist of longitudinal clinical records with varying follow up lengths, irregular visits, mixed data types, missingness, rare events, and heterogeneity across sites. Existing approaches struggle to capture these characteristics. This talk presents MeLD, a Medical Longitudinal latent Diffusion framework that embeds mixed type variables into a latent space using a variational autoencoder and models temporal dependencies with a transformer based diffusion process. Using a cohort of 50,000 people living with HIV in Latin America, we demonstrate that MeLD produces synthetic datasets that closely resemble real data while preserving privacy.

Comparison of an external control to a randomized control in pancreatic cancer; a validation exercise

Ruthie Davi, Medidata

Ruthie Davi (Medidata)

In oncology drug development, traditional randomization can face significant hurdles, particularly in rare diseases or biomarker-driven populations where patient recruitment is a barrier. External Control Arms (ECAs) may provide an alternative; however, demonstration that they provide robust, regulatory-grade evidence is needed. This presentation details a validation exercise comparing an ECA—constructed using a historical clinical trial data—against a concurrent randomized control group. We will describe the statistical methodologies used to create the external control and will conclude by examining how this validation supports the broader use of external controls to accelerate the delivery of new therapies to patients.

Session 25: Statistical Machine Learning for Complex Data

Location: DUMBO

Organizer: Tianxi Li (University of Minnesota)

Chair: Zifeng Zhang (Yale University)

Presenters: Feng Ruan; Fan Yao; Zifeng Zhang; Tianyi Chen

A theory of feature learning in kernel models

Feng Ruan, Northwestern University

Yunlu Chen, Yang Li, Keli Liu, Feng Ruan

A central phenomenon in modern machine learning is feature learning: rather than operating on a fixed representation of the data, successful models learn representations adapted to the prediction task. We study a tractable model of this phenomenon through a compositional variant of kernel ridge regression, where the kernel is applied after a learnable linear transformation. When the response depends on the input only through a low-dimensional predictive subspace, we show that optimizing the population objective automatically eliminates directions orthogonal to this subspace and, in certain regimes, exactly recovers it. Surprisingly, the same exact low-dimensional structure persists at finite sample sizes with high probability, even without explicitly penalizing the linear transformation to be low-dimensional.

The adaptation–stability tradeoff in iterative preference optimization

Fan Yao, University of North Carolina at Chapel Hill

Yurong Chen (Inria, École Normale Supérieure, PSL Research University); Yu He (Northwestern University); Michael I. Jordan (University of California, Berkeley); Fan Yao (University of North Carolina at Chapel Hill)

Iterative preference optimization trains LLMs through repeated comparisons between candidate responses, often updating both the data sampler and the reference model used to regularize each round. Using the latest model as the reference can preserve earlier changes, but it also carries feedback into subsequent updates. In this talk, I will study this adaptation–stability tradeoff through a finite-response population model. A Hodge decomposition separates the transformed preference signal into a component explained by one score per response and a cyclic component that can generate rotation as sampling adapts to the current policy. We show that reference refresh sets an exact discounted horizon for retained directional updates, while cyclic feedback limits how long ordinary updates can retain this history and remain locally stable. We then analyze sampling extrapolation and a lagged reference, two updates that use the preceding policy. When reference refresh is partial, both preserve the ordinary update's interior fixed point; they enlarge the stability region under different conditions and have distinct costs in adaptation speed. The results provide a mathematical diagnostic for the interaction between preference structure, sampling, and reference memory, and clarify the questions that arise when transferring these predictions to LLM post-training.

Dual validity of robust standard errors under complex dependence

Zifeng Zhang, Yale University

Zifeng Zhang (Yale University); Lihua Lei (Stanford University); Peng Ding (University of California, Berkeley); Wen Zhou (New York University)

Regression is a workhorse for analyzing data from randomized experiments, adaptive trials, and empirical studies, where valid inference typically relies on robust standard errors. These adjustments are usually justified as correcting dependence in unobserved errors, even though such dependence is rarely observable or reliably modeled. By contrast, the dependence structure of the regressor of interest, especially treatment assignment, is often known or engineered by design. Motivated by this asymmetry, we develop a general asymptotic framework for regression inference under unknown and potentially unstructured dependence and establish a dual validity principle: robust t-statistics remain asymptotically valid if the correlation structure encoded by the chosen standard-error adjustment is correct for either the regressor of interest or the unobserved errors, even when the other component is misspecified. This predictor-error symmetry has not, to our knowledge, been formalized in the classical robust-inference literature. For example, cluster-robust inference remains valid when either errors are clustered or assignment is clustered. The result reframes robust standard errors as design-aligned devices: when error dependence is uncertain, adjustments can be chosen to match the known assignment correlation rather than serve as a catch-all correction for unspecified error structure.

Vertex misalignment and changepoint localization in network time series

Tianyi Chen, Johns Hopkins University

Tianyi Chen (Johns Hopkins University); Mohammad Sharifi Kiasari (University of Maryland); Sijing Yu (University of Maryland); Youngser Park (Johns Hopkins University); Avanti Athreya (Johns Hopkins University); Vince Lyzinski (University of Maryland); Carey E. Priebe (Johns Hopkins University); Zachary Lubberts (University of Virginia)

Inference for time series of networks often relies on accurate vertex correspondence between network realizations at different times. In practice, however, such vertex alignments can be misspecified or unknown. We study the impact of vertex alignment on changepoint localization for dynamic networks through two illustrative models, each with a similar changepoint, with the key distinction being whether changepoint information is contained in marginal or joint distributions of the time-varying latent positions. We compare localization techniques ranging from the simple network statistic of average degree to the modern procedure of Euclidean mirrors. In one model, vertex misalignment causes little

error, and in the other, it impairs localization in ways that cannot be corrected through graph matching or optimal transport, which we show are closely related in this setting. Our results demonstrate that robust network inference necessitates reckoning with the subtle interplay of marginal and joint information in the observed network time series.

Session 05: Statistical Learning in Complex System

Location: Greenpoint

Organizer: Zhezhen Jin (Columbia University)

Chair: Yushu Shi (Weill Cornell Medicine)

Presenters: Yuchen Wu; Shanshan Ding; Jing Cao; Ying Hung

When does model collapse occur in structured interactive learning?

Yuchen Wu, Cornell University

Yuchen Wu (Cornell University); Kangjie Zhou (Columbia University); Weijie Su (University of Pennsylvania)

The proliferation of generative artificial intelligence has given rise to an interactive learning environment, where model parameters are continuously updated using not only data generated by natural processes, but also synthetic outputs produced by other models. This paradigm introduces two major challenges: (1) training data are no longer drawn exclusively from the target population, undermining a core assumption of classical statistical learning, and (2) model training processes become inherently correlated, as models interact with one another through repeated exposure to each other's synthetic outputs in a potentially complex manner. Establishing reliable statistical inference in such structured interactive learning environments therefore remains an important open problem. In particular, there is growing concern about model collapse, a phenomenon in which the performance of generative models progressively degrades as they are trained on synthetic data produced by earlier model generations.

Prior work on model collapse primarily focuses on a single model trained on its own output, failing to capture model performance in multi-model interactive settings. In this work, we fill this gap by investigating the performance of generative models in an interactive learning environment with general interaction patterns. In particular, we formalize model interactions using directed graphs and show that the occurrence of model collapse depends critically on the topology of the interaction graph. We further derive an explicit necessary and sufficient condition characterizing when model collapse occurs, and establish finite-sample results for linear regression and asymptotic guarantees for general M-estimators. We support our theoretical findings through extensive numerical experiments.

Supervised dimension reduction for high dimensional tensors

Shanshan Ding, University of Delaware

Shanshan Ding (University of Delaware), Wei Qian (University of Delaware)

A unified framework is introduced for sufficient dimension reduction (SDR) on high-dimensional and tensor-valued data. SDR is known to be a powerful tool for achieving data reduction and visualization in statistical and machine-learning problems. Robust nonparametric SDR methods are proposed for data with high-dimensional tensor-valued features under weak assumptions, and develop a new framework for high-dimensional tensor SDR problems with theoretical guarantees. Promising applications are demonstrated through simulations and real data analysis on neuroimaging data.

Interpretable sentiment analysis using the attention-based multiple instance classification model

Jing Cao, Southern Methodist University

Chenyu Yang (Amazon); Jing Cao (Southern Methodist University)

Sentiment analysis (SA) is widely used for analyzing text data to identify the underlying opinion or emotion expressed in a document. Neural network methods in natural language processing have produced state-of-the-art results in SA. However, many of those methods are 'black-box' algorithms that do not provide interpretable results. In this study, we develop an interpretable word-level context-based SA tool, which is an attention-based multiple instance classification model (AMIC). In text analysis, the attention mechanism helps models understand the context of words and sentences. AMIC uses word embedding to transform text to data, employs a multiple instance classification structure, and incorporates the attention mechanism to include context information from documents. AMIC has a transparent model structure with interpretable word-level context-based sentiment metrics. In addition, AMIC can be used to construct domain-specific sentiment dictionary without requiring prior information on seed words or a pre-trained list of sentiment words.

Optimal simulator selection

Ying Hung, Rutgers University

Li-Hsiang Lin (Georgia State University); Ying Hung (Rutgers University)

Computer simulators are widely used for the study of complex systems. In many applications, there are multiple simulators available with different scientific interpretations of the underlying mechanism, and the goal is to identify an optimal simulator based on the observed physical experiments. To achieve the goal, we propose a selection criterion based on conformal risk control that accounts for the trade-off between the accuracy and uncertainty of a simulator. Rigorous theoretical results, including unbiasedness and finite-sample risk guarantees, will be

rigorously derived. The proposed method is applied to the mathematical modeling of ecological systems.

Session 93: Causal Inference in Complex Real-World Data Applications

Location: Navy Yard

Organizer: Caleb Miles (Columbia University)

Chair: Taehyeon Koo (Columbia University)

Presenters: Luke Keele; Joseph Antonelli; Oliver J. Hines; Kara Rudolph

A weighting framework for clusters as confounders in observational studies

Luke Keele, University of Pennsylvania

Avi Feller; Luke Keele; Eli Ben-Michael

When units in observational studies are clustered in groups, such as students in schools or patients in hospitals, researchers often address confounding by adjusting for cluster-level covariates or cluster membership. In this paper, we develop a unified weighting framework that clarifies how different estimation methods control two distinct sources of imbalance: global balance (differences between treated and control units across clusters) and local balance (differences within clusters). We show that inverse propensity score weighting (IPW) with a random effects propensity score model — the current standard in the literature — targets only global balance and constant level shifts across clusters, but imposes no constraints on local balance. We then present two approaches that target both forms of balance. First, hierarchical balancing weights directly control global and local balance through a constrained optimization problem. Second, building on the recently proposed Generalized Mundlak approach, we develop a novel Mundlak balancing weights estimator that adjusts for cluster-level sufficient statistics rather than cluster indicators; this approach can accommodate small clusters where all units are treated or untreated. Critically, these approaches rest on different assumptions: hierarchical balancing weights require only that treatment is ignorable given covariates and cluster membership, while Mundlak methods additionally require an exponential family structure. We then compare these methods in a simulation study and in two applications in education and health services research that exhibit very different cluster structures.

Multivariate incremental effects for continuous treatments: studying the health effects of environmental mixtures

Joseph Antonelli, University of Florida, Department of Statistics

Zhuochao Huang (MD Anderson); Kejin Dong (University of Florida); Tuo Lin (University of Florida); Joseph Antonelli (University of Florida)

Evaluating the causal health effects of multivariate, continuous exposures, such as air pollution mixtures, is a critical public health challenge. A primary obstacle is the frequent violation of the positivity assumption, which renders the effects of standard deterministic interventions unidentified or heavily reliant on unreliable model extrapolation. In this paper, we develop a novel causal inference framework to address this challenge. We extend exponential tilting to multivariate exposures and address the critical question of how to compare different intervention directions fairly. This establishes a systematic framework for defining and evaluating various policy-relevant causal estimands, allowing researchers to address diverse scientific questions. We develop numerous methodological advancements, including efficient one-step estimation strategies, a Riemannian BFGS algorithm to solve a constrained manifold optimization problem, semiparametric efficiency bounds for causal estimands, minimax rates for estimators, and asymptotic normality results. We demonstrate our framework's utility by applying it to a nationwide environmental health dataset to identify the optimal strategy for reducing adverse health outcomes associated with a $PM_{2.5}$ chemical mixture.

The role of surrogate outcomes in combining data from multiple studies

Oliver J. Hines, Columbia University

Oliver J. Hines (Columbia University); Soohyun Kim (Weill Cornell Medicine); Amy J. Pitts (Regeneron Pharmaceuticals); Ngoc Duong (Northwestern University); Alice Medalia (Columbia University); Caleb H. Miles (Columbia University)

We consider combining data from multiple studies on Cognitive remediation therapy (CRT) for patients with schizophrenia. Our goal is to infer the effect of CRT on cognitive function, however, due to differing study protocols, cognitive function is measured using a different (sometimes overlapping) set of cognitive tests in each study. We develop a causal data fusion framework that exploits imperfect surrogate outcomes to infer the effect of a binary treatment on an outcome of primary interest. Our approach flexibly accommodates surrogate missingness patterns, does not require perfect surrogacy, and provides asymptotically normal inference that accounts for the uncertainty in regression estimators.

Using a generalization of the average treatment effect on the treated to better understand opioid prescription risks

Kara Rudolph, Columbia University

Kara Rudolph (Columbia University); Ivan Diaz (New York University); Shodai Inose (Columbia University); Herbert Susmann (New York University); Nicholas Williams (University of California, Berkeley); Katherine Hoffman (University of Washington); Allison Perry (New York University)

To reduce risks associated with using prescription opioids, the US (CDC) published opioid prescribing guidelines. However, these guidelines 1) do not consider dose strength and prescription duration as a joint exposure and 2) are

written as applying to all persons. However, perhaps the majority of opioid prescribing poses little risk and, therefore, is not in need of intervention. A large cohort of opioid-naive musculoskeletal pain patients on Medicaid is considered, and effects of modest reductions in opioid dose and duration prescribing practices (considered as a joint exposure) on risk of developing opioid use disorder are estimated, among patients with different prescribing levels. This causal effect is a generalization of the average treatment effect on the treated (ATT); it is the effect of a modified treatment policy among subsets for whom the policy would be relevant, based on their treatment status. A novel targeted minimum loss-based estimator is used with cross-fitting. It is found that universal reductions may have little effect on reducing opioid-related harms, and, plausibly, may even be counterproductive to the extent that their universal application results in uncontrolled pain for patients. In contrast, applying opioid prescribing guidelines to target patients with high-dose and/or long-duration prescriptions would be expected to yield much larger benefits.

Poster Session: Sunday, November 1, 2026, 6:30 pm–8:00 pm

Location: Salon F, G, H, and I

Held with the Welcome Reception.

Weighted conformal clustering

Anirban Nath, Columbia University

Anirban Nath; YoonHaeng Hur; Genevera Allen (Columbia University)

Clustering is a central tool for discovering latent structure in unlabeled data; yet modern clustering pipelines often end with a hard assignment of each observation to a cluster without rigorous measures of assignment uncertainty. We propose a novel weighted conformal approach for constructing valid confidence sets for cluster labels. The key difficulty is that the labels available for calibration are not observed ground-truth labels, but synthetic labels produced by a data-dependent clustering algorithm. Our method develops a conformal inference algorithm that corrects the resulting mismatch with the latent target labels through weights by formulating conformal clustering as a conditional label-distribution shift problem. We first derive an oracle procedure that attains finite-sample marginal coverage and then develop a computationally tractable and implementable version using estimated conditional label probabilities and novel augmented calibration. We show that the coverage of the estimated-weight procedure depends on the estimator, giving an explicit bound on the loss relative to the nominal level. Empirical studies demonstrate that the proposed weighted approach offers improvements over the recently proposed split conformal clustering procedure in terms of informative confidence set size, especially in nonlinear and high-dimensional clustering applications.

Confounding heterogeneity in sensitivity analyses for matched observational studies with multiple outcomes

Anthony Frazier, Colorado State University

Anthony Frazier (Colorado State University), Dr. Siyu Heng (New York University), Dr. Wen Zhou (New York University)

In matched observational studies, unobserved confounding may bias causal conclusions even after matching on observed covariates. Therefore, sensitivity analysis is often used to assess the magnitude of unobserved confounding required to alter the study's conclusions. We consider multivariate analyses that assess the causal effect of a single treatment on multiple outcomes. We identify an inherent feature of such analyses, which we call "confounding heterogeneity": different outcomes may be affected by different unmeasured confounders, and one common unmeasured confounder may affect different outcomes to varying degrees. We formulate sensitivity analysis frameworks for matched observational studies that accommodate confounding heterogeneity and show that Type-I error control of many existing sensitivity analysis methods under confounding heterogeneity is only guaranteed under fundamentally different sensitivity analysis models and modes of inference. We study asymptotic power using design sensitivity and Bahadur efficiency and examine finite-sample power through numerical studies. Our results reveal that the structure of unobserved confounding affects the results of a sensitivity analysis, and sensitivity analysis under the widely used single sensitivity parameter framework may misstate the sensitivity of a causal conclusion to unobserved confounding under confounding heterogeneity. Numerical studies and an empirical application demonstrate the importance of confounding heterogeneity in multivariate sensitivity analyses.

Privacy aware collaborative estimation and inference for GLMs

Bhaskar Ray, North Carolina State University

Bhaskar Ray (North Carolina State University), Aritra Mitra (North Carolina State University), Srijan Sengupta (North Carolina State University)

Statistical analyses in domains such as biomedical research and mobile health increasingly rely on sensitive data collected across multiple sites or devices, where privacy and regulatory constraints prohibit sharing data. Collaborative learning frameworks, notably Federated Learning (FL), have been developed for statistical estimation and inference under such constraints. However, algorithms fitting a single model globally can yield inconsistent estimates or fail to converge when the underlying model varies across sites. In this work, we propose a two-stage method for parameter estimation in a broad class of Generalized Linear Models (GLMs) when the parameter values underlying the data generating distribution vary across clusters of sites. The method uses a hypothesis-test-type classification based on one-shot local

estimates, followed by selective collaboration between similar sites. We jointly account for the statistical uncertainty and optimization error of the stochastic optimization iterates to propose precise hyperparameter tuning for computational efficiency. We use concentration inequalities to derive non-asymptotic risk bounds of the collaboration graph and the resulting estimators. We also establish their asymptotic behavior under offline learning for statistical inference with limited data. Simulation studies and real data applications support the theoretical results.

Learning cause-specific hazards with neural networks under competing risks

Chaeyeon Lee, Ewha Womans University

Chaeyeon Lee (Ewha Womans University); Sehwan Kim (Ewha Womans University); Myeonggyun Lee (NYU Langone Health); Hyungrok Do (New York University)

Despite recent advances in deep survival analysis and efforts to extend these methods to competing risks settings, direct estimation of continuous-time cause-specific instantaneous hazard functions has received less attention. We study the estimation of cause-specific instantaneous hazard functions parameterized by neural networks for time-to-event prediction under competing risks. Existing studies tend to focus on parameterizing cumulative incidence functions or survival functions, however, overlooked the importance of cause-specific instantaneous hazards, which provides valuable information that either CIF or survival function cannot deliver. Through simulation study, we investigate how different neural network parameterizations affect the accuracy of cause-specific instantaneous hazards estimation and further examine the proposed approach on real-world time-to-event prediction problems with competing risks.

Cluster LOCO: feature importance for interpreting clusters

Claire M He, Columbia University

Claire He (Columbia University); Genevera Allen (Columbia University)

Clustering is widely used for exploratory analysis and scientific discovery but its outputs can be difficult to interpret, audit, and reproduce as modern datasets become increasingly large and complex. Understanding which features drive the discovered cluster structure is therefore essential, but feature-level explanations for clustering remain limited, often tied to specific algorithms or data assumptions. We propose Cluster LOCO (Leave-One-Covariate-Out), a family of model-agnostic feature importance scores for clustering based on feature occlusion and clustering generalizability: for any chosen clustering algorithm, Cluster LOCO quantifies a feature's importance by measuring how much its removal degrades the ability to generalize the learnt cluster labels on held out samples. We first introduce Cluster LOCO-Split, which relies on data splitting, and then extend it to Cluster LOCO-MP, a minipatch ensemble-based version designed for large-scale data. We demonstrate that Cluster LOCO reliably identifies features driving cluster structure and can outperform existing clustering feature importance methods on synthetic experiments and show a useful application in cell-type discovery in single-cell transcriptomics.

Nonparametric continuous-time neural hazard estimation under heterogeneous cluster specific effects

Do Gyeong Han, Applied Statistics, Social Science, and Humanities, NYU Steinhardt

Han, Do Gyeong (New York University); Do, Hyungrok (New York University Grossman School of Medicine)

We study a deep learning-based nonparametric continuous time survival analysis model that estimates conditional instantaneous hazard functions under heterogeneous clustered data. Patients with identical covariates may have different hazards across clusters, and these differences may change over follow-up. Such heterogeneity can be modeled through a frailty distribution. However, when the interest is in the shared relationship between patient characteristics, time, and hazard, the cluster effects can be treated as nuisance. We propose a flexible nonparametric continuous time deep survival analysis method that separates the patient-level hazard of covariates and time, shared across clusters, from cluster-specific baseline hazards. Similar to the stratified Cox model, the baseline hazards are treated as nuisance without specifying their distribution. This formulation permits nonlinear covariate relationships and time-dependent hazard ratios together with time-varying differences between clusters. We investigate how accounting for cluster effects changes instantaneous hazard estimation, considering various forms of cross-cluster heterogeneity.

Forecasting with echo state networks for preliminary estimation of the chained CPI-U

Erin Boon, U.S. Bureau of Labor Statistics

Kate Eckerle (U.S. Bureau of Labor Statistics); Erin Boon (U.S. Bureau of Labor Statistics)

The Chained Consumer Price Index (C-CPI-U) is built by chaining together indices of one-month price changes. For the final C-CPI-U, individual item price indices are aggregated using the Törnqvist formula with monthly weights. These weights are based on consumer expenditure data which do not become available until four quarters after the price indices. In the interim, the U.S. Bureau of Labor Statistics (BLS) issues a preliminary estimate of the aggregate index using an econometric model. We propose an alternative method for preliminary estimation that instead retains the Törnqvist formula for aggregation and forecasts the missing expenditure data using a set of hierarchical Echo State Networks (ESNs), a class of Recurrent Neural Networks in which network-internal weights are initialized randomly and are not updated through training. ESNs are flexible, nonlinear, hidden variable models that can predict series with

complex temporal dynamics after a relatively simple training process. We develop an iterative procedure to forecast a vector of item expenditures for a given area based on its past expenditure data as well as on past and concurrent price data.

Theoretical inference of p-values methods for RVS

Giorgia Morello, Auburn University

Giorgia Morello

A central challenge in modern predictive modeling is that multiple models often achieve nearly identical predictive performance while leading to different scientific conclusions, a phenomenon known as the *Rashomon Effect*. Standard statistical inference relies on selecting a single "best" model, an approach that overlooks model uncertainty and contributes to the ongoing replication crisis. In this paper, we present a two-stage inferential framework designed to perform valid hypothesis testing directly across an entire *Rashomon Variable Set* (RVS). First, we test whether the Rashomon Variable Set contains evidence of predictive signal beyond what would be expected under random feature selection. Conditional on passing this global test, we perform variable-level inference by aggregating evidence across all admissible models in the Rashomon Variable Set and study the statistical behavior of several p-value aggregation strategies. Because the overlapping structure of the RVS induces complex dependence among tests and violates subset pivotality, we further incorporate the *Romano-Wolf* stepdown procedure. This integration provides an asymptotically valid correction that guarantees Family-Wise Error Rate (FWER) control under arbitrary dependence structures. We validate our framework through extensive simulation studies and real-world applications, offering practical guidelines for robust and reproducible variable-level inference under model ambiguity.

Regularized ensemble forecasting for learning weights from historical and current forecasts

Han Su, George Washington University

Han Su (George Washington University); Xiaojia Guo (University of Maryland, College Park); Xiaoke Zhang (George Washington University)

Combining forecasts from multiple experts often yields more accurate results than relying on a single expert. In this paper, we focus on forecasting a continuous outcome and introduce a novel regularized ensemble method that extends the traditional linear opinion pool by leveraging both current forecasts and historical performances to set the weights. In contrast to existing approaches that emphasize either current forecasts or past accuracy, our method accounts for both sources simultaneously. It learns weights by minimizing the variance of the combined forecast (or its transformed version) while incorporating a regularization term informed by historical performances. We also show that this approach has a Bayesian interpretation. Different distributional assumptions within this Bayesian framework yield different functional forms for the variance component and the regularization term, adapting the method to various scenarios. In empirical studies on Walmart sales and macroeconomic forecasting, our ensemble outperforms leading benchmark models both when experts' full forecasting histories are available and when experts enter and exit over time. Throughout, we also characterize how the optimal weights depend on current forecasts and historical performance, and use the empirical results to discuss where the framework's strengths lie and when each source of information plays a greater role.

Frailty neural networks for clustered survival data

Hangbin Lee, Chungnam National University

Hangbin Lee (Chungnam National University); Il Do Ha (Pukyong National University); Youngjo Lee (Seoul National University)

Large-scale survival data often exhibit both nonlinear effects and dependence structures induced by repeated or clustered observations. Standard frailty models provide a principled framework for modeling such dependence but typically impose restrictive linear predictor structures, while existing neural-based survival models largely focus on marginal prediction and do not explicitly account for latent cluster-specific heterogeneity. We propose FrailtyNet, a neural network extension of the Cox proportional hazards model that incorporates random frailties via hierarchical likelihood. The model combines a flexible neural representation of covariate effects with random frailties, allowing subject-specific risk prediction. Estimation is based on a joint objective that preserves the risk-set structure of the Cox model while accommodating large-scale optimization with neural networks. Through simulation studies under multiple frailty distributions, censoring levels, and nonlinear signal structures, we examine predictive performance and robustness to misspecification of the latent-effect distribution. Applications to clustered biomedical survival data further illustrate the practical advantages of explicitly modeling dependence. The proposed framework provides a bridge between classical frailty modeling and modern deep survival learning.

Improving external validation performance via mixed-effects models

Jiawa Zhang, Columbia University, Mailman School of Public Health, Department of Biostatistics

Jiawa Zhang (Department of Biostatistics, Columbia University); Myeonggyun Lee (Division of Biostatistics, Department of Population Health, NYU Grossman School of Medicine); Hyungrok Do (Division of Biostatistics, Department of Population Health, NYU Grossman School of Medicine)

The external validity of clinical and epidemiological prediction models developed on multi-institutional data is often evaluated on unobserved clusters. While linear mixed models (LMMs) are commonly used to handle clustered data, their theoretical and empirical advantage over

simpler linear models (LMs) for prediction in new clusters remains underexplored. Here we formally justify the superiority of LMMs in this setting. We derive risk bounds comparing the predictive mean squared error (MSPE) of LMMs and LMs for external cluster prediction. Our results show that the LMM fixed-effects estimator is inherently more efficient (lower variance) than the LM estimator when cluster-level heterogeneity is properly modeled. This efficiency gain ensures a lower MSPE for LMM predictors in any new cluster. Simulations further demonstrate that the LMM advantage is maximized in studies with high between-cluster variance and unbalanced cluster sizes. These findings clarify the fundamental basis for using LMMs in generalizability tasks and provide practical guidance for designing robust, cross-institutional prediction models. The approach is further illustrated using a real data application.

Multitask sufficient dimension reduction with common and heterogeneous sparsity patterns

Jiaying Weng, Bentley University

Jiaying Weng (Bentley University); Xiaoli Kong (Wayne State University)

Sufficient dimension reduction seeks low-dimensional projections that retain predictor information relevant to a response, but sparse methods usually analyze one task at a time. We study multiple nonlinear regression tasks with task-specific central subspaces and related sparse row supports. For tasks sharing active-variable locations but retaining task-specific directions, we propose common-support multitask sliced inverse regression (COMSIR), a two-stage estimator combining a convex sparse initialization and group-Lasso refinement. For partially overlapping supports, we propose heterogeneous-support multitask sliced inverse regression (HEMSIR), which uses support-similarity regularization and an alternating proximal-gradient hard-thresholding heuristic. We establish a conditional nonasymptotic estimation-error bound for COMSIR and an analogous bound for an idealized HEMSIR estimator defined by an exact global optimizer; the latter does not cover the heuristic implementation. We also develop componentwise lower-bound constructions that isolate shared and task-specific sources of estimation difficulty. Simulations under common and heterogeneous supports compare the proposed procedures with taskwise, pooled, and linear multitask baselines. In the settings examined, COMSIR generally has lower subspace-estimation loss than the taskwise and pooled alternatives, while HEMSIR more accurately recovers heterogeneous supports when supplied with the true support size. An analysis of Texas and California crime data illustrates nonlinear prediction after dimension reduction.

Transfer learning for crop yield estimation in data-limited regions

Jiaying Zhang, China Jiliang University

Jiaying Zhang (China Jiliang University)

Machine learning models often require substantial ground-truth observations, limiting their application in data-sparce regions. Here, we investigate whether transfer learning can leverage knowledge learned from data-rich regions to improve crop yield estimation in data-limited regions. We trained and validated a soybean yield estimation model using abundant fine-scale (field-level) data from the United States, and then transferred this U.S. model to Brazil, where crop phenology, climate and environmental conditions, yield respones, and the spatial scale of available yield observations differ significantly. Specifically, we fine-tuned the U.S. model using Brazilian state-level yield observations and evaluated its ability to predict soybean yield at the municipal level. This cross-scale setting allows us to assess whether information learned from a relatively coarse-scale observations can be used to estimate yield at a finer spatial scale. Compared with a conventional cross-scale model without transfer learning, cross-scale with transfer learning increases the municipal-level explained variance from an R2 of 0.29 to 0.44. When municipal-level observations are directly used for model training, the R2 further increases to 0.57. Relative to this benchmark, transfer learning increases the effectiveness of the cross-scale estimation from 50% to 78%. These results demonstrate that the pretrained model can retain useful information across substantial differences in climate, crop phenology, yield responses, and spatial scale. The results demonstrate the potential of transfer learning to reduce the need for local fine-scale observations, providing a data-efficient approach to crop yield estimation in data-limited regions.

Bayesian reconstruction of infectious disease transmission networks: a simulation benchmark of BetNet and TransPhylo

Jihyun Kim, Department of Statistics, University of Georgia

Jihyun Kim (Department of Statistics, University of Georgia); Liang Liu (Department of Statistics, University of Georgia and Institute of Bioinformatics, University of Georgia)

Reconstructing who-infected-whom networks is important for outbreak investigation, but true transmission histories are rarely observed. BetNet is a recently developed Bayesian method that directly integrates pathogen SNP differences with case-level temporal information, whereas TransPhylo infers transmission from a time-scaled pathogen phylogeny. Their relative performance across evolutionary and sampling conditions remains insufficiently characterized. We conducted a full-factorial simulation study comprising 432 parameter settings and 10 outbreak replicates per setting (4,320 simulated datasets), varying outbreak size, mutation rate, within-host diversity, genome length, and case-sampling proportion. Both methods were evaluated on matched outbreaks, sampled cases, transmission-truth definitions, and parent-level accuracy metrics. TransPhylo was supplied with a directly simulated time-scaled coalescent tree, excluding sequence-based phylogenetic reconstruction error and

providing a conservative benchmark for BetNet. Preliminary analyses showed that mutation rate was the dominant determinant of relative performance: among settings with $\mu \geq 10^{-6}$, BetNet outperformed TransPhylo in 85.8% of parameter settings, with mean overall accuracy 0.43 ± 0.12 SD versus 0.27 ± 0.03 SD. TransPhylo retained an advantage only at the lowest tested mutation rates ($\mu \leq 10^{-7}$), where simulated sequences contained little mutational information. These findings support BetNet as a competitive direct-inference approach when sufficient mutational information is available.

How well can causal prior-fitted networks generalize on out-of-distribution data?

Jingsen Zhu, Cornell University

Jingsen Zhu (Cornell University); Raphael Chan-Gon Kim (Cornell University); Khiem Pham (Cornell University); Ramin Zabih (Cornell University); Alexander Terenin (Cornell University); Michele Santacatterina (New York University)

Prior-fitted networks are a promising amortized Bayesian inference method based on in-context learning and have been applied to causal inference for treatment effect estimation tasks. Causal prior-fitted networks are pretrained on synthetic data generated by prior distributions over data-generating processes that follow a structural causal model. For prior-fitted networks to generalize at test time, the prior should have sufficient coverage across (i) various causal properties of the data-generating process, including heterogeneity, baseline-effect alignment, confounding, overlap, and others; (ii) different data-generating mechanisms outside the prior family. We examine existing methods and their evaluation protocols, and provide strong empirical evidence supporting two failure modes: first, existing priors are causally valid but encode no explicit causal properties above and, in particular, lack controllability, interpretability and coverage; second, existing benchmarks largely disregard how well a prior-fitted network can generalize with respect to these properties. We (i) assess what makes an evaluation benchmark informative on the model's generalization ability, and use this principle to design an evaluation protocol; (ii) evaluate existing causal prior-fitted networks on our proposed evaluation protocol for out-of-prior generalization ability, and identify a gap between prior-fitted networks and classical causal inference methods that previous methods did not detect.

Accelerated Bayesian functional principal components analysis for densely-observed Gaussian data

Joseph Sartini, Wake Forest University School of Medicine

Joseph Sartini (Wake Forest University School of Medicine); Xinkai Zhou (Beijing Normal-Hong Kong Baptist University); Ciprian Crainiceanu (Johns Hopkins University)

Emerging wearable health monitoring technologies, consumer and medical grade, measure signals at a very high sampling frequency. For example, patch electrocardiogram devices measure voltages within the range of 200-500Hz. Functional Principal Components Analysis (FPCA) can be used to perform efficient dimensionality reduction of these data, making analyses tractable. Traditional FPCA implementations estimate the functional principal components (FPCs) from a smoothed covariance estimate, then condition on these estimates when deriving the FPC scores. To produce valid inferences, we implemented a fully-Bayesian FPCA that jointly models the FPCs and scores. This approach combined an orthonormal spline FPC representation and parameter expansion to sample the corresponding spline weights. To adapt this method for high density Gaussian data, we demonstrate that all likelihood calculations can be moved to the spline coefficient space with appropriate normalization. This produces an implementation that scales with the dimension of the spline basis, rather than with the frequency of measurement in the original domain. The resulting approach produces identical inferences while reducing compute time by 50-200%. We demonstrate the utility of this approach by fitting FPCA and functional principal components regression using NHANES accelerometry from the 2003-2004 and 2005-2006 waves, which can be accomplished on a standard laptop.

Extreme value theory for singular subspace estimation

Junhyung Chang, University of Wisconsin–Madison

Junhyung Chang (University of Wisconsin–Madison); Joshua Cape (University of Wisconsin–Madison)

This paper studies fine-grained singular subspace estimation in the matrix denoising model where a deterministic low-rank signal matrix is additively perturbed by a stochastic matrix of Gaussian noise. We establish that the maximum Euclidean row norm (i.e., the two-to-infinity norm) of the aligned difference between the leading sample and population singular vectors approaches the Gumbel distribution in the large-matrix limit, under suitable signal-to-noise conditions and after appropriate centering and scaling. We apply our novel asymptotic distributional theory to test hypotheses of low-rank signal structure encoded in the leading singular vectors and their corresponding principal subspace. We provide de-biased estimators for the corresponding nuisance signal singular values and show that our proposed plug-in test statistic has desirable properties. Notably, compared to using the Frobenius norm subspace distance, our test statistic based on the two-to-infinity norm empirically has higher power to detect structured alternatives that differ from the null in only a few matrix entries or rows. Our main results are obtained by a novel synthesis of and technical analysis involving row-wise matrix perturbation analysis, extreme value theory, saddle point approximation methods, and random matrix theory. Our contributions complement the existing literature for matrix denoising focused on minimaxity, mean squared error analysis, unitarily invariant distances between subspaces, component-wise asymptotic distributional theory, and row-wise uniform error bounds. Numerical simulations illustrate our main results and

demonstrate the robustness properties of our testing procedure to non-Gaussian noise distributions.

Fréchet regression of multivariate distributions with nonparanormal transport

Junyoung Park, University of Michigan

Junyoung Park (University of Michigan); Irina Gaynanova (University of Michigan)

Regression with distribution-valued responses and Euclidean predictors has gained increasing scientific relevance. While methodology for univariate distributional data has advanced rapidly in recent years, multivariate distributions, which additionally encode dependence across univariate marginals, have received less attention and pose computational and statistical challenges. In this work, we address these challenges with a new regression approach for multivariate distributional responses, in which distributions are modeled within the semiparametric nonparanormal family. By incorporating the nonparanormal transport (NPT) metric – an efficient closed-form surrogate for the Wasserstein distance – into the Fréchet regression framework, our approach decomposes the problem into separate regressions of marginal distributions and their dependence structure, facilitating both efficient estimation and granular interpretation of predictor effects. We provide theoretical justification for NPT, establishing its bi-Lipschitz equivalence to the Wasserstein distance and proving that it mitigates the curse of dimensionality. We further prove uniform convergence guarantees for regression estimators, both when distributional responses are fully observed and when they are estimated from empirical samples, attaining fast convergence rates comparable to the univariate case. The utility of our method is demonstrated via simulations and an application to continuous glucose monitoring data.

Moment density estimator with scaled Hermite polynomials and its application in neural network training

Kaiming Cheng, Stony Brook University

Kaiming Cheng (Stony Brook University); Andrew P. Mullhaupt (Stony Brook University)

We address the problem of approximating an unknown probability density function from a finite set of its moments. Instead of attempting to recover the true underlying distribution, we provide a continuous density designed to closely interpolate the given moments. We use an estimator constructed from scaled Hermite polynomials as the baseline and demonstrate its equivalence to the bias-corrected Kernel Density Estimator (KDE). The bandwidth, based on minimizing the Mean Integrated Squared Error (MISE), and quadratic programming (QP) resolve the common drawbacks of existing Hermite-series-based methods and provide a density suitable for sampling, while maintaining computational efficiency and accuracy. Experimental results confirm the proposed method, consistently yields valid density functions that maintain high fidelity to the target moments with an implicit moment extension. Kolmogorov-Smirnov (K-S) test results confirm the recovered densities generate samples that are statistically indistinguishable from those drawn from a true density. We further apply this method to improve neural network training.

On testing for independence between generalized error models of several time series

Kilani Ghoudi, United Arab Emirates University

Kilani Ghoudi; Bouchra R. Nasri; Bruno N. Rémillard

We define generalized innovations associated with generalized error models having arbitrary distributions, that is, distributions that can be mixtures of continuous and discrete distributions. These models include stochastic volatility models and regime-switching models with possibly zero-inflated regimes. The main novelty of this article is to be able to test for conditional independence between several time series with arbitrary distributions, extending previous results of Duchesne, Ghoudi, and Rémillard (2012), obtained only for stochastic volatility models. We define families of empirical processes constructed from lagged generalized errors, and we show that their joint asymptotic distributions are Gaussian and independent of the estimated parameters of the individual time series. Möbius transformations of the empirical processes are used to obtain tractable covariances. Several tests statistics are then proposed, based on Cramér–von Mises type statistics and dependence measures, as well as graphical methods to visualize the dependence. In addition, numerical experiments are performed to assess the power of the proposed tests. Finally, to show the usefulness of our methodologies, examples of applications for financial data and crime data are given to cover both discrete and continuous cases.

When weighting meets neural CATE: a benchmark and a simple hybrid for stability under limited overlap

Kyungjin Sohn, University of Wisconsin–Madison

Kyungjin Sohn(University of Wisconsin–Madison); Guanhua Chen(University of Wisconsin–Madison)

Estimating conditional average treatment effects (CATE) from observational data is challenging when limited covariate overlap and complex outcome surfaces occur simultaneously. Existing approaches often emphasize either explicit covariate balancing at the design stage or flexible outcome modeling through representation learning, leaving unclear when each strategy is most useful and whether they can be combined effectively. We study this trade-off through controlled simulations and the semi-synthetic IHDP benchmark, comparing weighting-based, meta-learning, forest, and neural CATE estimators using estimation accuracy, policy value, and calibration. A pointwise error decomposition separates residual design imbalance from estimation and approximation error, clarifying when explicit balancing can

improve CATE estimation. Motivated by this decomposition, we introduce a modular hybrid that computes three-way maximum mean discrepancy (MMD) balancing weights and uses them as within-arm bootstrap sampling probabilities for neural CATE training, without modifying the network architecture or loss. Across simulations, distributional balancing is most effective when covariate imbalance is a dominant source of error, whereas flexible models are advantageous for complex response surfaces. On IHDP, MMD-weighted bootstrap aggregation reduces both mean error and dispersion for TARNet and DragonNet and improves calibration relative to their unweighted counterparts. These results suggest that explicit balancing can serve as a useful stabilization layer when limited overlap persists but flexible outcome modeling remains necessary.

A specification-grounded LLM workflow for translating business requirements into analytical data models

Maria Paula Carletto, Link School of Business

Maria Paula Carletto (Link School of Business); Francisco Isidro Massetto (Link School of Business)

Large language models (LLMs) have demonstrated strong performance in Text-to-SQL benchmarks, yet their reliability decreases in real-world enterprise environments, where analytical queries often require more than knowledge of database structure. Correct implementation may depend on domain-specific rules and business definitions that are not explicitly represented in the relational schema. We investigate an LLM-assisted workflow that combines automatically generated semantic metadata with structured analytical specifications elicited from domain experts. The semantic metadata is generated from relational schema information and representative records, while the specifications externalize business intent, including source-selection and temporal-interpretation rules. We compare SQL generation under three grounding conditions: relational schema alone; schema plus an LLM-generated semantic data dictionary; and schema plus semantic metadata and domain-expert specifications. Evaluation focuses on analytical correctness, requirement completeness, and alignment with the domain expert's intended analysis. The study examines whether semantic metadata can bridge the gap between enterprise database structures and business terminology, and whether specification grounding provides additional value for analytical query generation. The proposed workflow extends automated database-description and Text-to-SQL approaches toward undocumented enterprise environments in which both database semantics and business intent must be recovered explicitly.

Optimal tree pruning via ℓ_0 best-subset selection

Paul R Gaona Partida, University of California, Irvine

Paul R Gaona Partida (University of California, Irvine); Ana Maria Kenney (University of California, Irvine)

Decision trees create transparent rule-based structures that align with how scientists and clinicians reason about evidence, making them indispensable across precision medicine, clinical prediction, and subgroup discovery. The interpretive value of a fitted tree depends on it remaining compact, and classical cost-complexity pruning has been the standard procedure for decades. Yet cost-complexity pruning restricts attention to a predetermined nested sequence and offers no guarantee that the returned tree is the best of its size. We propose a unifying framework that reformulates pruning as a hierarchical ℓ_0-constrained variable selection problem on the internal-node basis of the maximal tree, effectively classic best subset selection. We prove that classical cost-complexity pruning is the exact solution-path algorithm of the corresponding penalized formulation on training data, and characterize when the weakest-link path coincides with the ℓ_0 optimum and when the latter strictly dominates. Building on this reformulation, we develop a mixed-integer optimization framework that returns the globally optimal subtree at any specified size and extends naturally to an augmented predictor space combining linear main effects with tree-defined partition effects. Extensive simulated experiments and several case studies demonstrate strong performance for interaction detection, subgroup identification, and prediction accuracy while maintaining interpretability, providing flexible methodology for scientific applications.

A robust transfer learning framework for high-dimensional multiclass distance weighted discrimination

Qian Tang, University of Minnesota

Qian Tang (University of Minnesota), Boxiang Wang (University of Iowa), Hui Zou (Johns Hopkins University)

Multiclass classification remains a challenging problem, particularly in high-dimensional settings where the number of features can greatly exceed the sample size. To address this challenge, we develop a transfer learning framework for multiclass distance weighted discrimination (DWD) that leverages auxiliary information from related source domains to improve target-domain classification. To mitigate negative transfer, we further propose an adaptive strategy that assesses source-target similarity and regulates the contribution of auxiliary data, thereby retaining beneficial source information while reducing the impact of irrelevant or harmful sources. Extensive simulation studies and real data applications demonstrate the effectiveness and robustness of the proposed method.

Transformers as provable approximators of sparse principal component analysis

Qinyan Liu, University of Science and Technology of China

Qinyan Liu (University of Science and Technology of China); Yan Chen (The University of Hong Kong); Canhong Wen (University of Science and Technology of China)

Transformers have recently been shown to emulate statistical algorithms in supervised and semi-supervised settings, yet their ability to perform fully unsupervised

statistical estimation remains less understood. We investigate this question through sparse principal component analysis (SPCA), a fundamental unsupervised problem in high-dimensional statistics involving structured sparsity and spectral estimation. We show that the Truncated Power Method (TPower), a classical iterative solver for SPCA, provides a constructive template for multi-layer Transformers to approximate sparse spectral updates. Under a noisy spiked covariance model, we establish approximation guarantees for Transformer-based SPCA and derive statistical error bounds by decomposing the recovery error into algorithmic approximation and statistical estimation components. Beyond the theoretical analysis, we study the empirical performance of Transformers trained with unsupervised SPCA objectives. The Transformer formulation enables amortized inference after training and efficient parallel processing of multiple inputs. Experiments on synthetic benchmarks and high-dimensional gene expression data demonstrate that the proposed approach can recover sparse principal components with performance comparable to classical SPCA baselines in the considered settings. This work provides a provable perspective on Transformers as approximators of statistical algorithms and advances the understanding of their role in unsupervised sparse spectral estimation.

Statistical inference for continuous action contextual bandits

Raphael C Kim, Cornell University

Raphael C Kim (Cornell); Michele Santacatterina (NYU); Ramin Zabih (Cornell); Rajarshi Mukherjee (Harvard); Ivan Diaz (NYU)

Statistical inference under adaptively collected data is becoming increasingly popular across e-commerce and mobile health. Many methods exist for discrete action settings, ranging from weighting to debiasing approaches. Despite the ubiquity of continuous actions in experimentation from optimal pricing to precision dosing, to the best of our knowledge, methods that support statistical inference under continuous action, adaptively collected data remain underdeveloped. In this work, we extend kernel-smoothed doubly robust estimation from i.i.d. data to adaptively collected data with continuous actions. We study a family of adaptively weighted estimators, establish their mean squared error rates, asymptotic normality, and characterize an estimation lower bound as a function of regret. We conclude with recommendations for experimental design informing efficiency and regret, and support our theory with an extensive simulation study.

A statistical approach to paired–portfolio trading

Sara Mezuri, Oakland University

Sara Mezuri (Oakland University), Ravindra Khattree (Oakland University)

In this study we introduce a statistical framework that combines canonical correlation analysis with cointegration testing to build and trade synthetic portfolio pairs. Traditional pairs trading focuses on selecting individual stocks, which in reality, it is difficult to find such pairs that remain stable long enough to trade. Our objective is to develop and validate a statistical method that builds tradeable portfolio pair that capture broader dependence structures and maintain stable long run relationships. We used daily data for ten large-cap U.S. equities from January 2021 to December 2023. We partitioned the equities into all possible two-block combinations and applied canonical correlation analysis to each partition to obtain the linear combinations with the highest correlation, which act as synthetic portfolios. We test each synthetic portfolio pair for cointegration using the Engle–Granger two-step procedure and the Johansen maximum-likelihood approach. In-sample results identify 42 canonical portfolio pairs that satisfy the cointegration criteria and follow stable AR(2) dynamics. These findings show shared stochastic trends with temporary deviations, which support pairs trading. We implement a market neutral trading strategy based on deviations from the estimated long-run equilibrium, defined following Tsay (2010, Section 8.8). Entry and exit rules depend on volatility scaled bands around the mean. We vary the band width parameter Δ to study sensitivity. Narrower bands increase trade frequency and turnover, while wider bands reduce activity but generate larger individual trades. We estimate the equilibrium relationship using ordinary least squares and orthogonal distance regression. For several pairs, orthogonal distance regression produces more symmetric and stable hedge ratios. We test the strategy out of sample from January 2024 through November 2025. Most cointegrated pairs remain profitable. Out of 42 pairs that perform well in sample, 33 pairs continue to generate positive returns, with gains as high as 62%. However, 9 pair lose stability. This result shows why continuous validation is necessary. The results show that canonical correlation analysis combined with cointegration testing provides a solid framework for pairs trading. Building portfolio pairs instead of choosing single stock pairs improves robustness. Alternative regression estimators and band choices give us direct control over risk and execution behavior.

Generalized functional linear mean residual life model with right-censored data

Seohyeon Park, New York University Grossman School of Medicine

Seohyeon Park (New York University Grossman School of Medicine); Myeonggyun Lee (New York University Grossman School of Medicine); Mengling Liu (New York University Grossman School of Medicine)

The mean residual life (MRL) function provides a clinically interpretable summary of remaining life expectancy and offers complementary insights beyond conventional hazard-based analyses. However, existing MRL regression models have mainly focused on static or scalar covariates, leaving functional predictors largely unaddressed. This gap

is increasingly relevant as wearable devices and medical imaging generate densely measured data that capture rich temporal or spatial information on disease risk and health status. To address this gap, we propose a generalized functional linear mean residual life (GFLMRL) model for right-censored time-to-event data. We approximate the functional coefficient using regression splines and develop a quasi-partial score estimation procedure to estimate the effects of both scalar and functional predictors, along with a likelihood-based criterion to jointly select the link function and the number of basis functions. We further introduce a nonparametric hypothesis testing procedure to assess whether the functional coefficient conforms to a specified parametric structure, such as linearity. We establish the asymptotic properties of the proposed method and inference procedures and evaluate the finite-sample performance through extensive simulation studies. The practical utility of the proposed method is illustrated through an application to an accelerometry dataset from the National Health and Nutrition Examination Survey (NHANES).

Learning convex functions with hyper input convex neural networks

Shayan Hundrieser, Yale University

Shayan Hundrieser (Yale University); Insung Kong (UNIST); Johannes Schmidt-Hieber (WIAS and HU Berlin)

Shape-constrained inference is a well-established field in low-dimensional nonparametric statistics, with convexity representing a key structural property for regression, density estimation, score-based sampling, and optimal transport. High-dimensional settings call for machine-learning approaches. To this end, we introduce Hyper Input Convex Neural Networks (HyCNNs), a novel neural network architecture to learn convex functions. HyCNNs are by design alway convex in the input, theoretically capable of leveraging depth, and perform reliable when trained at scale. Concretely, we prove that HyCNNs require exponentially fewer parameters than well-established input convex neural networks (ICNNs) to approximate quadratic functions up to a given precision. Throughout a series of synthetic experiments, we demonstrate that HyCNNs outperform existing ICNNs and unconstrained feedforward networks in terms of predictive performance for convex regression and interpolation tasks. We further apply HyCNNs to learn high-dimensional optimal transport maps for synthetic examples and for single-cell RNA sequencing data, where they oftentimes outperform ICNN-based neural optimal transport methods and other baselines across a wide range of settings.

The project is based on joint work with Insung Kong and Johannes Schmidt-Hieber.

Seasonal prediction of the Upper Colorado River Basin river flow based on statistical learning

Siyu Zhao, Zhejiang University

Siyu Zhao

The Colorado River is one of the largest rivers in the southwestern United States and northern Mexico, with about 90% of its streamflow coming from the Upper Colorado River Basin (UCRB). This study provides a ten-month-lead seasonal prediction of UCRB streamflow during April–July using an artificial neural network (ANN) forecast model with land-surface and atmospheric variables and sea surface temperature (SST). A Pacific SST index is developed for a negative-west–positive-east pattern over the Pacific (30°S–65°N) that correlates with UCRB streamflow. The index considers only grid points whose correlation coefficients exceed the 0.05 significance level and is calculated from the difference between the positively and negatively correlated points. It has a higher correlation with UCRB streamflow than traditional approaches that use the spatial average or leading principal components of SST. Prediction skill improves greatly when this Pacific SST index is incorporated into the ANN model. Overall, the ANN successfully predicts UCRB streamflow for 1980–2018 with a lead time of up to four months (cross-validated correlation of about 0.80, mean absolute percentage error of about 20%, and Heidke skill score above 70%) and shows some skill up to ten months ahead (correlation of about 0.55, mean absolute percentage error of about 26%, and Heidke skill score above 50%). Thus, the ANN forecast model substantially extends the lead time of existing UCRB streamflow predictions, which are made in the preceding winter, while providing comparable or greater skill. These results provide climate information for water-supply planning and decision-making related to the Colorado River, and illustrate the importance of oceanic climate signals in seasonal streamflow prediction.

Functional and classical geometric morphometrics with Bayesian aggregation for craniodental classification in shrews

Sofia R. Garcia, St. Edward's University

Sofia Garcia (St. Edward's University); Ahmad Talafha (St. Edward's University)

Geometric morphometric data derived from biological shapes are widely used for species classification, but different representation frameworks may lead to different predictive performance. In this study, we compare Functional Data Geometric Morphometrics (FDGM) and classical Geometric Morphometrics (GM) for classifying three shrew species (*Suncus murinus*, *Crocidura monticola*, and *Crocidura malayana*) from Peninsular Malaysia. We construct a workflow for extracting skull outlines from images and representing them using pseudo-landmarks across three craniodental views. Classifiers are trained using bootstrap aggregation, and predictions are combined using majority voting, out-of-bag (OOB) aggregation, and a proposed Bayesian probability aggregation approach. Model performance is evaluated using repeated Monte Carlo resampling. The results show that bootstrap aggregation consistently improves classification accuracy

over single-classifier predictions for both FDGM and GM, with the dorsal view yielding the highest accuracy overall. While GM slightly outperforms FDGM at the single-classifier level in the dorsal view, FDGM achieves comparable or higher accuracy than GM under aggregation, particularly with the proposed Bayesian scheme, which produces probability estimates that are reasonably well calibrated and competitive with standard majority voting. These findings indicate that functional representations, when combined with the proposed probabilistic aggregation approach, can be as effective as classical landmark-based approaches for morphometric classification tasks.

Certifying a sufficient Gaussian dependence regime for learned weighted FDR control

Soyeong Oh, Michigan State University

Soyeong Oh(Michigan State University)

We study whether independent reference data can certify a dependence condition sufficient for learned weighted FDR control. In a one-sided Gaussian model, we check every correlation not independently known to be nonnegative. Uniform soundness rules out high acceptance probability at boundary distributions approachable in total variation from outside the sign class. If all d checked correlations are at least $\gamma > 0$, an intersection–union certificate is uniformly complete with $N_C = O\{\gamma^{-2} \log(d/\beta)\}$ reference scores. For $0 < \gamma \leq 1/6$, this rate is necessary up to constants in the worst case over dimensions and sets of $d \geq 2$ pairs when $\beta \leq 1/16$, and also for every all-pairs problem with $m \geq 2$ when $\beta \leq 1/4$. When the certificate accepts, we use weighted BH at level q. Otherwise, we use a prespecified fallback valid under arbitrary dependence at $q^\circ = (q - \beta)/(1 - \beta)$. If the certificate and testing p-values are conditionally independent given the learning information, the fallback has conditional FDR at most q°, and the weighted branch either has conditional FDR at most q or is selected with conditional probability at most β, then routing controls FDR at q. Gaussian simulations examine certification and routing for up to $4{,}950$ checked pairs using synthetic correlations and correlations estimated from HBN.

Deep varying coefficients mixtures of regressions for binary classification

Sphiwe Bonakele Skhosana, University of Pretoria and Michigan State University

Sphiwe Bonakele Skhosana (University of Pretoria and Michigan State University); Guanqun Cao (Michigan State University); Najmeh Nakhaierad (University of Pretoria)

Semiparametric mixtures of varying coefficient models provide a flexible framework for modelling heterogeneity in data by assuming that the component regression parameters are functions of the covariates. However, these models are typically based on a continuous response variable. Moreover, in order to avoid the curse-of-dimensionality, the regression parameters are typically assumed to be functions of a single covariate. In this paper, we propose deep varying coefficient mixtures of logistic regressions for modelling heterogeneity in a discrete, in particular binary, response variable. Through a combination of deep neural networks (DNNs) with the classical Expectation-Maximization (EM) algorithm, we develop a dedicated DNN embedded EM algorithm, termed the EM-DNN algorithm, to estimate the parameters and non-parametric functions of the proposed model. The proposed procedure allows the component regression parameters of the proposed model to be functions of high-dimensional covariates. A simulation study demonstrates the practical utility of the proposed methods. Finally, the proposed methods are applied to a dataset in energy economics.

Biostatistical randomization techniques in SAS and R

Stephen Parziale, Yale University

Stephen Parziale (Yale University)

Randomized controlled trials are considered the gold standard for clinical research because they effectively minimize bias and establish causal relationships. Such techniques are highly recommended by regulatory bodies like the US Food and Drug Administration (FDA)(Higgins & Levin, 2023) and the International Council for Harmonisation (ICH, 1998). Thus, it is imperative for investigators, programmers, and biostatisticians to understand how these techniques work and how to implement them with statistical software. The purpose of this project is to demonstrate how to execute randomization techniques in both SAS and R, allowing for the generation of reliable, reproducible code for community utilization.

Evaluating alternatives to clinical measurements: a signal recovery coefficient with semiparametric inference

Suin Kim, Seoul National University

Suin Kim (Seoul National University); Haeun Moon (Seoul National University)

Costly or invasive clinical measurements motivate the search for more accessible alternatives. Although alternatives are often assessed by their association with an established measurement or its ability to predict a clinical outcome, neither criterion assesses whether an alternative can fulfill the established measurement's outcome-related role. We propose a signal recovery coefficient, defined as the proportion of variation in the outcome-related signal, $E(Y \mid X)$, that is explained by an alternative. Its estimation involves sequential regressions with estimated responses, making the required convergence rates for valid inference difficult to establish. We develop two semiparametric estimators to address this difficulty: one uses three-way cross-fitting, and the other avoids regressions on estimated responses by imposing an additional structural assumption. The second approach also provides a basis-based test of zero recovery, addressing the

degeneracy of first-order inference at zero recovery. In Alzheimer's Disease Neuroimaging Initiative data, p-tau217-based plasma biomarkers had the highest recovery for amyloid PET, tau PET, and cerebrospinal-fluid references, despite different rankings from reference and outcome prediction. In Mobilise-D data, we evaluate various wearable mobility measures according to how much they preserve the outcome-related signal captured by clinical mobility assessments.

Skewness is a friend: distribution-free a priori precision for mean estimation with skewness and higher moments

Thang Duc Nguyen, Stony Brook University

Thang Duc Nguyen (Stony Brook University); Tonghui Wang (New Mexico State University); Ce Shang (The University of Sydney); Ziyuan Wang (University of Wisconsin Oshkosh)

A priori study design is traditionally dominated by power analysis, which focuses on hypothesis testing rather than estimation accuracy. We develop a distribution-free A Priori Precision (APP) framework for population mean estimation and show that skewness and higher moments can improve sample-size determination. Under finite-variance and bounded-support assumptions, we derive explicit finite-sample APP guarantees using classical concentration inequalities. Our central theoretical result is a variance-skewness inequality for bounded distributions, showing that for a fixed support width, larger absolute skewness reduces the largest compatible variance and can therefore reduce the sample size required for a prescribed precision level. We further derive a sharper range-variance-skewness-kurtosis inequality that incorporates both asymmetry and tail behavior, yielding a stronger higher-moment-informed APP rule. We also provide an oracle finite-sample APP benchmark for population skewness under a finite sixth-moment assumption. Numerical examples and simulations illustrate the resulting improvements over conventional distribution-free benchmarks. These results show that skewness need not be merely a nuisance feature; in precision-based study design, it can be useful prior information.

Graphical modeling from marginal information via optimal transport

Ulysse Naepels, EPFL

Ulysse Naepels (EPFL); Paul Freulon (EPFL); Victor Panaretos (EPFL)

A graphical model represents multivariate interactions through conditional independence, but requires jointly observed samples to be fitted to data. When individuals cannot be tracked and only marginal observations are available, as with destructive measurements leading to cross-sectional data, the problem of graphical modeling is confounded with that of assembling a joint law. We show how simultaneous graphical modeling and joint law assembly can be realized via a multi-marginal optimal transport problem with graphical cost. Focusing on Gaussian marginals, we generate a joint law satisfying the Markov properties prescribed by a given graph. Within the class of all joint laws respecting the given graph, our construction selects one via a minimum-energy principle. In the empirical setting, where only estimates of the marginals are available, we examine the stability of the problem under perturbations of both the marginals and of the graph. The resulting error splits into an estimation term, vanishing with the sample size, and a specification term, which quantifies the cost of a misspecified graph. Finally, we illustrate these findings by numerical experiments.

A new bounded mixed regression model

Victor Eduardo Lachos Olivares, Interinstitutional Graduate Program in Statistics UFSCar-USP (PIPGEs)

V. E. Lachos-Olivares (UFSCar-USP); Jorge L. Bazan (University of São Paulo); Catalina B. García-García (University of Granada)

Regression models for bounded response variables on the unit interval are increasingly used for data such as proportions and rates. The Beta and Simplex models, parameterized in terms of the mean, are the most common, but they do not always provide a suitable fit. In this work, we consider the Vasicek distribution as a flexible alternative for modeling bounded response data. Although the Vasicek regression model has been developed only in a frequentist framework, we extend it to the Bayesian framework and further propose a mixed-effects version incorporating fixed and random effects. Parameter estimation is carried out within a Bayesian framework using the No-U-Turn Sampler algorithm implemented in `Stan` to obtain posterior samples. To evaluate the model, extensive simulation studies are conducted. First, an analysis of prior sensitivity for a key dispersion parameter is conducted, with the results highlighting the impact of different prior specifications. The performance of the Vasicek regression model and its mixed-effects extension is compared with that of the traditional Beta and Simplex models. In this study, we will examine a dataset concerning the proportion of children between the ages of 6 and 14 who are not enrolled in school in the state of Mato Grosso in Brazil. In a secondary application, we utilize Peruvian data on extreme poverty to propose mixed regression models incorporating both fixed and random effects. The findings indicate these extensions provide a competitive and complementary framework for the analysis of bounded response data.

Modeling heterogeneous mediation effects in survival analysis via an interpretable M-learner framework

Xingyu Li, University of Texas MD Anderson Cancer Center

Xingyu Li (MD Anderson Cancer Center);Qing Liu (Amgen);Xun Jiang (Astellas); Hong Amy Xia (Amgen); Brian P. Hobbs (Telperian); Peng Wei (MD Anderson Cancer Center)

Mediation analysis is a useful tool to evaluate surrogate endpoints in clinical trials. We propose a novel method, the M-survival learner, for estimating heterogeneous indirect treatment effects in the presence of censored outcomes. The proposed approach enables the identification of

interpretable patient subgroups characterized by distinct mediation pathways. To distinguish heterogeneous from homogeneous mediation effects, we introduce a new statistical criterion specifically designed for survival data. The method provides a principled framework for evaluating heterogeneity in surrogate biomarker performance across patient populations, offering evidence to support accelerated approval drug. By explicitly assessing subgroup-specific surrogate validity, the proposed approach addresses key regulatory concerns regarding the reliability of surrogate endpoints. We further establish theoretical properties of the method to justify its statistical guarantees. We apply the approach to data from a Phase III randomized clinical trial of HIV treatment, demonstrating its practical utility in real-world settings. Extensive simulation studies further evaluate and demonstrate its finite-sample performance.

Sparse-view interpretable 3D animal behavior representations for neural encoding and decoding

Xinming Dai, Columbia University

Xinming Dai (Columbia University); Qihang Jin (University of Science and Technology of China); Tianshu Tan (Harvard University); Baiyuan Chen (University of Cambridge); Hanrui Lyu (Northwestern University); Lenny Aharon (Columbia University); Kyle Daruwalla (Cold Spring Harbor Laboratory); Xun Helen Hou (Cold Spring Harbor Laboratory); Matthew R. Whiteway (Columbia University); Liam Paninski (Columbia University); Yizi Zhang (Columbia University)

A deeper understanding of brain function requires a precise, structured characterization of behavior. Yet, extracting behavioral representations from video in a form suitable for scientific analysis remains a fundamental challenge. Many prior studies represent behavior via pose estimation or nonlinear video embeddings. However, pose tracking discards rich information beyond predefined keypoints, while nonlinear video embeddings lack interpretability. We address this limitation with SABLE (**S**parse-view **A**nimal **B**ehavior **L**atent **E**mbeddings), a self-supervised framework that leverages a geometric inductive bias to learn behavior representations. By augmenting a multi-view transformer with priors from monocular depth and pose estimation, SABLE reconstructs 3D animal behavior from extremely sparse views while learning explicit 3D latent structure. Without ground-truth 3D labels, it reliably recovers 3D behavior from two-view videos, whereas state-of-the-art (SOTA) methods fail or yield degenerate solutions. Across the International Brain Lab and Cheese3D datasets, we demonstrate that SABLE learns 3D representations that match or exceed prior SOTA performance in neural encoding and decoding. Once pretrained across animals, SABLE serves as an off-the-shelf model that generalizes zero-shot to unseen animals without animal-specific calibration or retraining. Our method establishes 3D-aware video embeddings that capture complex behavior, opening new avenues for studying brain–behavior relationships.

Random-effect modeling of multiplex networks with directed and undirected layers

Xinyi (Cynthia) Li, Carnegie Mellon University

Cynthia Li (Carnegie Mellon University); Nynke M.D. Niezink (Carnegie Mellon University & Data Science Department); Anni Hong (University of Pittsburgh)

Individuals are often integrated in multiple social networks simultaneously. For example, colleagues in a workplace may be connected through professional collaboration, friendship and advice-seeking ties, and these different network layers are often interdependent. Social systems with multiple relationships are naturally represented by multiplex networks, in which some relations are directed (e.g., advice seeking), while others are inherently undirected (e.g., collaboration). The p_2 model provides a rich framework for modeling dependence in uniplex (Van Duijn et al., 2004) and multiplex networks (Hong et al., 2025). This generalized linear mixed conditionally-independent dyads model represents the probability of a tie as a function of fixed effects (e.g., covariates, density, reciprocity) and random effects that capture heterogeneity in actors' tendencies to send and receive ties. Bayesian estimation allows uncertainty in these effects to be quantified through their posterior distributions. Currently, p_2 models fail to incorporate undirected layers, constraining their applicability to real network data. In this study, we extend the multiplex p_2 model to include undirected network layers. We implement this extension in the R package multip2 and illustrate it through an application to organizational network data.

An integrated association test for multiple high-dimensional data

Xueyuan Cao, The University of Tennessee Health Science Center

Xueyuan Cao (The University of Tennessee Health Science Center); E Olusegun George (The University of Memphis)

With high-throughput technologies, investigators can measure genetic variations in multiple forms. New methods are needed to interrogate the relationship between genomic variations and an endpoint of interest. We previously developed the POST procedure to associate gene sets/pathways with a clinical variable. We use similar dimension reduction machinery on each form of omics data at a locus/gene to collectively test the association between multiform genomic data and an endpoint of interest. The probe-level signals for each omics data type at a locus/gene are first projected onto an orthogonal subspace, and the corresponding eigenvalues are then rescaled so that their sum equals 1 for each omics data type. The projected data are then subjected to a parametric association test to obtain z-statistics. The test statistic, T, is defined as a weighted sum of squares of individual z-statistics. A generalized $\chi 2$ distribution is used to approximate the p-value of T, with the correlation structure of the z-statistics approximated via bootstrap resampling. We further implement a LASSO model selection procedure to

pinpoint the features in a locus/gene. We investigated the performance in simulation studies and applied the proposed method to a gene-profiling and methylation dataset of pediatric AML from the NCI TARGET.

Improving prediction of patient-reported health outcomes among adults with diabetes across diverse populations through transfer learning

Xuzhi Wang, New York University

Xuzhi Wang (Department of Biostatistics, School of Global Public Health, New York University and Division of Biostatistics, NYU Grossman School of Medicine); Jeannette M. Beasley (Department of Nutrition and Food Studies, Steinhardt School of Culture, Education, and Human Development, New York University and Department of Medicine, Grossman School of Medicine, New York University); Ting-Fang Lee (Division of Biostatistics, NYU Grossman School of Medicine)

Population heterogeneity challenges the transportability of prediction models, and the utility of transfer learning across different populations remains insufficiently understood. We evaluated whether a regression-based transfer learning framework could improve prediction performance across heterogeneous populations. We applied the framework to two settings: between-country heterogeneity using the Gateway harmonized international aging cohorts (GATEWAY) and within-country heterogeneity using race/ethnicity groups from the Health and Retirement Study (HRS). GATEWAY included 19,879 adults with diabetes from eight cohorts, whereas HRS included 3,926 participants across four race/ethnicity groups. In both studies, vision impairment was predicted using demographic characteristics and clinical risk factors. Transfer learning models were compared with target-only and pooled models using AUC and Brier score, with bootstrap resampling to assess uncertainty. In GATEWAY, transfer learning improved performance in 75% (6/8) of cohorts, with AUC gains of approximately 0.1% to 0.6%. In HRS, all four (100%) race/ethnicity groups benefited from transfer learning, with AUC gains of approximately 0.8% to 3.1%. Overall, regression-based transfer learning produced greater and more consistent improvements under within-country than between-country heterogeneity. These findings suggest that transfer learning can improve prediction across heterogeneous populations, although benefits may vary by population setting.

Understanding US county-level partisan shifts through explainable machine learning

Yimin Zhang, Villanova University

Yimin Zhang (Villanova University); Anudeep Reddy Chityala (Villanova University)

We develop explainable machine learning approaches to explain why U.S. counties shift their partisan support. Using county-level socioeconomic and election data from 3,097 counties across three presidential election cycles (2012, 2016, and 2020), we compare and evaluate interpretable statistical models against black-box ensemble and neural network models via cross-validation. While the XGBoost model achieves the best predictive performance, its ensemble structure makes its decision-making process and predictions difficult to interpret. To uncover important predictors and their interactions, we apply Shapley-based explanation methods and related approaches, including standard SHAP, cohort SHAP, SAGE and Geo SHAP, and identify four key factors in predicting the county-level party shift: per capita healthcare spending, unemployment rate, median income, and white population share. We further derive meaningful loss-based counterfactual explanations for swing counties and find that unemployment, healthcare access, and taxable income are the most influential factors associated with shifts in partisan support in the swing counties. Together, these findings demonstrate how explainable machine learning can combine predictive performance with interpretable insights into county-level electoral change.

A consistent framework for modeling genetic interactions across orders

Yucheng Jin, Northwestern University

Yucheng Jin (Northwestern University); Enes Haxhimalia (Northwestern University); Istvan Kovacs (Northwestern University)

The effects of mutations depend on the genetic background in which they occur. Genome-wide perturbation screens measure these genetic interactions by comparing observed fitness with predictions from a baseline model. Higher-order interactions should reveal effects that cannot be explained by lower-order interactions alone, but existing measures do not reliably distinguish these contributions. Here we show that conventional approaches can allow lower-order effects to spill over into higher-order scores, with ambiguities arising even at the pairwise level - an apparent three-gene interaction can arise from two-gene effects and not an additional effect of all three genes together. This obscures the biological interpretation of genetic interactions, while the lack of a clear criterion for selecting a baseline model adds further uncertainty. We propose a mathematical framework to address these limitations at every interaction order. Using genome-wide yeast double- and triple-mutant datasets, we examine how modeling assumptions affect inferred interactions and their reproducibility. Our work provides a foundation for distinguishing genuine higher-order effects from lower-order contributions and for clarifying how combinations of mutations shape fitness at the organismal level.

Reliable decoding of sequential neural representations with a masked autoencoder

Zhishun Yang, University of California, Irvine

Yang Meng (UC Irvine); Zhishun Yang (UC Irvine); Keiland Cooper(UC Irvine); Stephan Mandt(UC Irvine); Norbert Fortin(UC Irvine); Babak Shahbaba(UC Irvine)

Decoding sequential structure from sparse, high-dimensional neural population activity requires

learning relationships among states across multiple timescales, often with incomplete observations and limited ground truth. Here we introduce NeuroMAE, a transformer-based masked autoencoder that jointly decodes neural states and reconstructs experience-associated population activity from partially observed spikes, providing measures of decoding reliability. In hippocampal recordings from a sequence memory task, NeuroMAE outperforms conventional decoding approaches, particularly as observations become increasingly incomplete, while maintaining reliable likelihood estimates. We demonstrated two applications for modeling sequential structure. To identify replay, the sequential reactivation of previously experienced events during rest, we developed Reliability-Weighted Sequence Matching (RWSM) to compare decoded state-probability traces with target sequence templates. To capture sequential information within events, our model recovers temporal information contained in sequential firing fields more accurately than conventional methods. Together, these results establish a framework for decoding and evaluating sequential structure under partial observation.

Confidence set for the number of classes in latent class model

Zhixin Lai, Carnegie Mellon University

Zhixin Lai (Carnegie Mellon University); Yinqiu He (University of Wisconsin–Madison); Weijing Tang (Carnegie Mellon University)

Latent class models (LCMs) are widely used to identify unobserved heterogeneous subgroups in multivariate categorical data. The application of LCMs requires selecting the number of latent classes. Most existing methods select the number of latent classes using information criteria, likelihood-based tests, or singular value decomposition. These methods provide only point estimates without uncertainty guarantees and can become unreliable when the separation between classes is weak or the sample size is small. In this work, we construct a confidence set for the true number of latent classes under the *repro samples* framework. The idea is to generate artificial samples that mimic the underlying data-generating process and thereby provide additional information about the parameter of interest. We establish a non-asymptotic coverage guarantee for the proposed confidence set. We compare our approach with existing methods for selecting the number of classes in LCMs under challenging settings through extensive simulation studies. We demonstrate using a real dataset that the proposed confidence set contains the number of classes supported by prior analyses.

Sparse bilinear regression for multiple connectivity matrices

Ziqian Zhang, University of Virginia

Chang Li (University of Virginia); Ziqian Zhang (University of Virginia); Nan Hu (Stony Brook University); Yue Wang (University of Colorado Anschutz Medical Campus); Aiying Zhang (University of Virginia)

Predicting phenotypic measures from functional magnetic resonance imaging (fMRI) connectivity data is an important problem in neuroimaging. Many existing approaches vectorize connectivity matrices or screen edges independently, potentially discarding the intrinsic matrix structure of the connectome. We propose a sparse bilinear regression framework that models phenotypic measures directly from matrix-valued connectivity data through quadratic forms. The framework jointly incorporates multiple connectivity matrices, allowing complementary information across connectivity sources to be integrated while reducing the number of parameters from $O(p^2)$ to $O(p)$ and providing sparse, interpretable region-level representations. Model parameters are estimated using an accelerated proximal gradient algorithm with sparsity-inducing regularization. Semi-synthetic simulations based on observed connectivity matrices demonstrate recovery of known sparse coefficient patterns. Application to real connectivity data identifies a sparse set of brain regions associated with age.

Statistical guarantees for link prediction with neural network latent representations

Zixuan Wu, University of Chicago

Zixuan Wu (University of Chicago); Olga Klopp (ESSEC Business School); Claire Donnat (University of Chicago)

We study link prediction in partially observed networks with node covariates under a latent-variable model with shared node-specific random effects. Our contributions are twofold. First, we establish constructive approximation bounds for Hölder-smooth functions using sparse deep neural networks with locally quadratic activation functions, including SiLU and GELU. These bounds explicitly relate approximation accuracy to network depth, width, sparsity, and weight magnitudes. Second, we show that empirical risk minimizers over networks with appropriately chosen architectures attain minimax-optimal rates for population prediction, up to logarithmic factors. The rates are governed by an effective sample size comparable to the minimum of the number of nodes and the expected number of observed node pairs. Matching minimax lower bounds characterize the transition from a regime limited by the availability of edge observations to one limited by the number of nodes. These results provide explicit guidance for selecting network architectures under incomplete observations.