Minghui Yu Memorial Conference

April 28, 2023

Doctoral Student Body
Department of Statistics
Columbia University

About

The 2023 Minghui Yu Memorial Conference, organized by doctoral students in the Statistics Department of Columbia University, will take place on Friday, April 28th in Uris Hall 110. Minghui Yu was a doctoral student at the statistics department, who passed away in a tragic accident in the spring of 2008. Since then, doctoral students in the Statistics Department have been organizing a conference each year to honor his memory. The conference will feature talks by doctoral students at the Statistics Department, ranging from those just beginning a research program to those who are about to defend dissertations. In addition to being an occasion to remember our friend and colleague, this event will be an opportunity to learn about exciting new research areas emerging from our department. We would like to thank the Department of Statistics for their continued support.

Contact

If you have any questions, please do not hesitate to contact Hane Lee at hl3333@columbia.edu.

Schedule

9:30 AM – 10:00 AM	Breakfast
10:00 AM – 10:10 AM	Opening Remarks
	Session 1
10:10 AM – 10:30 AM	*Bayesian Imputation of Unmeasured Metabolites*, Casey Bradshaw
10:30 AM – 10:50 AM	*Learning from Similar Linear Representations: Adaptivity, Minimaxity, and Robustness*, Ye Tian
10:50 AM – 11:10 AM	*Diagnosing Model Performance Under Distribution Shift*, Tiffany Cai
11:10 AM – 11:30 AM	Break
	Session 2
11:30 AM – 11:50 AM	*A Spectral Method for Identifiable Grade of Membership Analysis with Binary Responses*, Ling Chen
11:50 AM – 12:10 AM	*New Paradigm of Identifiability in Cognitive Diagnostic Models: Beyond Discrete Responses*, Seunghyun Lee
12:10 AM – 12:30 AM	*Multivariate Symmetry: Distribution-Free Testing via Optimal Transport*, Zhen Huang
12:30 PM – 1:40 PM	Lunch
	Session 3
1:40 PM – 2:00 PM	*The Second Moment Phenomenon for Monochromatic Subgraphs*, Jaesung Son
2:00 PM – 2:20 PM	*Theoretical Guarantees for Data Dependent Posterior Tempering*, Ruchira Ray
2:20 PM – 2:40 PM	*Unwinding Stochastic Orderflow in a Central Risk Book*, Long Zhao
2:40 PM – 3:00 PM	Break
3:00 PM – 4:00 PM	Keynote address, Jennifer Hill
4:00 PM – 4:20 PM	Break
	Session 4
4:20 PM – 4:40 PM	*Identifiability and Falsifiability: Two Challenges for Bayesian Model Expansion*, Collin Cademartori
4:40 PM – 5:00 PM	*Measuring Mass Polarization with Wasserstein Distance*, Hane Lee
5:00 PM – 5:20 PM	*MAE and its application to develop encoders for earth system dynamics*, Zhewen Hou
6:30 PM	Dinner

Keynote - Professor Jennifer Hiill

Jennifer Hill is a Professor of Applied Statistics at New York Universiy, director of the Center for Practice and Research at the Intersection of Information, Society, and Methodology (PRIISM), and co-director of and the Master's of Science Program in Applied Statistics for Social Science Research (A3SR). Professor Hill develops and evaluates methods to help answer the types of causal questions that are vital to policy research and scientific development. In particular she focuses on situations in which it is difficult or impossible to perform traditional randomized experiments, or when even seemingly pristine study designs are complicated by missing data or hierarchically structured data. Most recently Professor Hill has been pursuing two intersecting strands of research. The first focuses on Bayesian nonparametric methods that allow for flexible estimation of causal models and are less time-consuming and more precise than competing methods (e.g. propensity score approaches). The second line of work pursues strategies for exploring the impact of violations of typical causal inference assumptions such as ignorability (all confounders measured) and common support (overlap). Professor Hill has published in a variety of leading journals including Journal of the American Statistical Association, Statistical Science, American Political Science Review, American Journal of Public Health, and Developmental Psychology. Hill earned her PhD in Statistics at Harvard University in 2000 and completed a post-doctoral fellowship in Child and Family Policy at Columbia University's School of Social Work in 2002.

Student Talks

Casey Bradshaw

Bayesian Imputation of Unmeasured Metabolites

Metabolomics data is a rich source of insight into the cellular processes occurring in a biological specimen, which in turn has implications for our understanding of disease states. Modern metabolomics experiments are typically designed to measure a small subset of the metabolites present in a tissue sample, and the particular subset chosen varies across experiments. We present a method for leveraging data from multiple studies to impute metabolites which were unmeasured in some, but not all, of those studies. This method represents the metabolite abundance rankings via a Bayesian Plackett-Luce model, and performs approximate inference to generate the imputed metabolite ranks. Extending this analysis to allow imputation of metabolites from transcriptomic data is a key area of interest, as the latter is much more abundant and readily available.

Ye Tian

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

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

Tiffany Cai

Diagnosing Model Performance Under Distribution Shift

Prediction models can perform poorly when deployed to target distributions different from the training distribution. To understand these operational failure modes, we develop a method, called DIstribution Shift DEcomposition (DISDE), to attribute a drop in performance to different types of distribution shifts. Our approach decomposes the performance drop into terms for 1) an increase in harder but frequently seen examples from training, 2) changes in the relationship between features and outcomes, and 3) poor performance on examples infrequent or unseen during training. These terms are defined by fixing a distribution on X while varying the conditional distribution of $Y|X$ between training and target, or by fixing the conditional distribution of $Y|X$ while varying the distribution on X. In order to do this, we define a hypothetical distribution on X consisting of values common in both training and target, over which it is easy to compare $Y|X$ and thus predictive performance. We estimate performance on this hypothetical distribution via reweighting methods. Empirically, we show how our method can 1) inform potential modeling improvements across distribution shifts for employment prediction on tabular census data, and 2) help to explain why certain domain adaptation methods fail to improve model performance for satellite image classification.

Ling Chen

A Spectral Method for Identifiable Grade of Membership Analysis with Binary Responses

Grade of Membership (GoM) models are flexible individual-level mixture models for multivariate categorical data. GoM allows each subject to have mixed memberships in multiple extreme latent profiles. Therefore GoM models have a richer modeling capacity than the latent class model that restricts each subject to belong to a single profile. The flexibility of GoM comes at the cost of more challenging identifiability and estimation problems. We propose a novel SVD-based spectral approach to GoM analysis with multivariate binary responses. Our approach is based on the observation that the expectation of the response matrix is low-rank under a GoM model. For identifiability, we develop sufficient and almost necessary conditions for a notion of expectation identifiability. For estimation, we extract only a few leading singular vectors of the observed response matrix and exploit the simplex geometry of these vectors to estimate the mixed membership scores. Such a spectral method has a huge computational advantage over Bayesian or likelihood-based methods and is especially suitable for large-scale and high-dimensional response data. Extensive simulation studies demonstrate the superior efficiency and accuracy of our method compared to its competitor. We further illustrate our method by applying it to a personality test dataset.

Seunghyun Lee

New Paradigm of Identifiability in Cognitive Diagnostic Models: Beyond Discrete Responses

Cognitive diagnostic models (CDMs) are a popular family of discrete latent variable models that model students' mastery or deficiency of multiple fine-grained skills. CDMs have been widely used to model binary or polytomous item response data. With advances in technology and the emergence of varying test formats in modern educational assessments, new response types, including continuous responses such as response times, and count-valued responses from tests with repetitive tasks or eye-tracking sensors, have also become available. Variants of CDMs have been proposed for modeling such responses recently. However, whether these extended CDMs are identifiable is entirely unknown. We propose a very general CDM framework for modeling arbitrary responses with minimal assumptions and establish identifiability in this general setting. Surprisingly, we prove that the proposed general CDM is identifiable under conditions similar to those for traditional binary-response CDMs. Our conclusions set up a new paradigm of identifiable general-response CDMs. For estimation, we propose a universal EM algorithm to estimate the model parameters. We conduct simulation studies under various response types. The simulation results not only corroborate the identifiability theory, but also show the superior empirical performance of our algorithm. We also demonstrate the method through an application to real-world datasets containing response time and count data.

Zhen Huang

Multivariate Symmetry: Distribution-Free Testing via Optimal Transport

The sign test (Arbuthnott, 1710) and the Wilcoxon signed-rank test (Wilcoxon, 1945) are among the first examples of a nonparametric test. These procedures — based on signs, (absolute) ranks and signed-ranks — yield distribution-free tests for symmetry in one-dimension. We propose a novel and unified framework for distribution-free testing of multivariate symmetry (that includes central symmetry, sign symmetry, spherical symmetry, etc.) based on the theory of optimal transport. Our approach leads to notions of distribution-free generalized multivariate signs, ranks and signed-ranks. As a consequence, we develop analogues of the sign and Wilcoxon signed-rank tests that share many of the appealing properties of their one-dimensional counterparts. In particular, the proposed tests are exactly distribution-free in finite samples with an asymptotic normal limit, and adapt to various notions of multivariate symmetry. We study the consistency of the proposed tests and their behavior under local alternatives, and show that the proposed generalized Wilcoxon signed-rank test is particularly powerful against location shift alternatives. We show that in a large class of such models, our generalized Wilcoxon signed-rank test suffers from no loss in (asymp-

totic) efficiency, when compared to the Hotelling's T^2 test, despite being nonparametric and exactly distribution-free. A score transformed version of the generalized Wilcoxon signed-rank can even achieve locally asymptotically optimal properties.

Jaesung Son

The Second Moment Phenomenon for Monochromatic Subgraphs

What is the chance that in a room of n people, there are s people all of whom know each other? This can be formulated as the existence of a monochromatic s-clique K_s in the complete graph K_n, where every edge is colored with 2 colors corresponding to whether or not they know each other. Suppose H is a general, connected graph. We study the asymptotic distribution of $T(H, G_n)$, the number of monochromatic copies of H in a uniformly random coloring of the edges of the graph G_n with c_n colors. We show that $T(H, G_n)$ converges to Pois(λ) whenever $\mathbb{E}T(H, G_n) \to \lambda$ and Var$T(H, G_n) \to \lambda$. Furthermore, we show that $\mathbb{E}T(H, G_n) \to \lambda$ implies $T(H, G_n) \xrightarrow{D}$ Pois(λ) if and only if $H = K_3$, the triangle, or $H = K_{p,q}$, $p = 1, 2$ and $q \geq p$. As an application, we derive the limiting distribution of $T(H, G_n)$ when $G_n \sim G(n, p)$ is the Erdös-Rényi random graph. Multiple phase transitions emerge as p varies from 0 to 1, depending on the notion of balancedness of the graph.

Ruchira Ray

Theoretical guarantees for data-dependent posterior tempering

α-posteriors "robustify" standard Bayesian inference by raising the likelihood to a constant fractional power, α, effectively downweighting its influence in the posterior calculation. This procedure has been shown empirically to be robust to model misspecification in many settings. However, practical recommendations for selecting α and theoretical guarantees for these methods in the case of data-driven selection of α remain open questions. We engage with these issues by connecting α-posterior inference to ridge regression. Data-driven approaches to tuning α in this setting suggest a novel asymptotic regime where $\alpha = o_p(1)$. In this new regime, we provide sufficient conditions for (i) asymptotic Normality of the α-posterior; (ii) asymptotic Normality of its variational approximation; and (iii) consistency and asymptotic Normality of the Bayes estimator.

Long Zhao

Unwinding Stochastic Orderflow in a Central Risk Book

We study the problem of optimal execution from the perspective of a centralized trading unit, commonly referred to as the Central Risk Book (CRB), in banks and trading companies. The CRB aggregates orders from various business units within the organization in real-time and executes outstanding orders in an optimal manner to minimize transaction costs. In this study, we introduce a model that distinguishes between the in-flow and out-flow orders of the CRB, and incorporates both the price impact and the bid-ask spread paid by the out-flow orders as transaction costs. Specifically, the price impact and the in-flow orders are both modeled using the generalized Ornstein-Uhlenbeck processes, where the time-dependent parameters can be calibrated to market data. We present a tractable control problem with solutions that capture the stylized facts and highlight the trade-off between trading speed and transaction costs.

Collin Cademartori

Identifiability and Falsifiability: Two Challenges for Bayesian Model Expansion

The process of iterative model expansion often involves moving from some simple initial model to more complex, higher-dimensional models in order to obtain better fit to our observed data or to remove unrealistic assumptions. In this talk, I will argue that this model expansion process can create distinct challenges which motivate the use of more fine-grained posterior summaries in inference and model evaluation. In particular, a pair of theoretical results demonstrate how expansion can make typical model summaries and checks less informative as the model complexity grows. I will present some intuition for these results and discuss how the corresponding challenges can be mitigated by avoiding premature posterior marginalization.

Hane Lee

Measuring Mass Polarization with Wasserstein Distance

There has been long-standing disagreement over whether the American public is more ideologically polarized. While political scientists have an extensive literature on ideological polarization, studies are unsatisfactory due to lack of definition and ad-hoc operationalizations. We propose a measure of polarization to bridge this gap between conceptualization and measurement. Starting from the intuition of a maximum polarization distribution, we measure polarization with the Wasserstein distance from this maximum. We show that this measure follows traditional axioms of polarization from economics literature.

Zhewen Hou

MAE and its application to develop encoders for earth system dynamics

We develop deep learning encoders for capturing spatial temporal features that can improve predictive tasks for real-world climate data. Leveraging state of the art deep learning methods, such as masked autoencoder (MAE), we are able to get a better model. Leveraging different climate models, we are able to use transfer learning to address data sparsity challenges in using DL in climate workflow. MAE provides utility not only in predictive tasks but also provides insights on scientific connections between models and how models differ, etc.

About Minghui Yu

Minghui was born in Shandong, China in 1983. In 2002, he entered the Special Class for the Gifted Young at the University of Science and Technology of China (USTC), one of the most prestigious universities in China. Minghui possessed the rare quality of being not only smart, but also diligent, versatile, modest and easy-going. He was the type of friend who would stand by you no matter the situation. Minghui breezed through the challenging undergraduate program at USTC, ranking at the top of his class. Minghui was well liked by his fellows students having served as the class president from his sophomore year. Although under enormous academic pressure, he still found time to organize a series of student activities, such as hiking, art performances, and athletic contests for his fellow students.

After graduating summa cum laude in 2006 from USTC, Minghui entered the PhD program at the Physics Department of Columbia University. After one year, he transferred to the doctorate program in statistics. During his time at Columbia, Minghui served as the public relations head of the Columbia University's Chinese Students and Scholars Association (2007-2008), and was a member of the Columbia Chinese Basketball Association and the Columbia Graduate Student Consulting Club. His biography on the CUCSSA website mentioned his love of "movies, photography and delicacies". Minghui described himself in his blog as a boy who wants to combine art and science together.

On April 4, 2008, after attending a student- organized conference, Minghui escorted his girlfriend home on the west side of campus. On his return, he was accosted as he was crossing 122nd and Broadway and in his attempt to flee, he was struck by an automobile on Broadway. Minghui was taken to St. Luke's Hospital where he passed away a short time later.