Wharton Department of Statistics and Data Science
Lawrence D. Brown Distinguished Lectures

Xihong Lin

- Wednesday, November 4, 12:00-1:00 PM
  Location: G60 Jon M. Huntsman Hall

From Generative AI to Statistical Inference: Harnessing Synthetic Data for Validity and Power

The integration of statistics and generative AI plays a pivotal role in accelerating trustworthy, cross-domain scientific discovery. Recent advances in generative models have dramatically expanded the availability and use of synthetic data across scientific domains. While these developments create exciting opportunities to enhance data analysis, they also raise fundamental statistical challenges regarding how synthetic data can be used in a valid, reliable, and principled manner. In this talk, we first discuss the generation of high-fidelity, high-dimensional synthetic data for real-world applications using generative AI models, including transformers, diffusion models, and latent flow-matching models. We then present a principled framework for integrating observed and synthetic data in downstream analyses to ensure valid statistical inference, even when generative AI models are misspecified. Specifically, we introduce synthetic-data-assisted methods for continuous and discrete outcomes based on an augmented synthetic data likelihood that integrates observed and synthetic data. We show that these methods are robust to misspecification of black-box generative models and can improve statistical power when the generative models provide informative synthetic data. We demonstrate the utility of these methods through applications to UK Biobank data, including genome-wide association studies (GWAS) of proteomic data, whole-genome sequencing (WGS) analyses of brain imaging phenotypes, and analyses of binary mental health survey outcomes. These applications involve substantial missingness, with approximately 70–90% of the relevant data missing, highlighting the potential of synthetic-data-assisted methods to enable more powerful and reliable scientific discovery in challenging real-world settings.

- Thursday, November 5, 12:00-1:00 PM
  Location: G65 Jon M. Huntsman Hall

Large-Scale Composite-Null Testing: A Unified Inference Framework for Causal Mediation, Replication, and Pleiotropy

Large-scale causal mediation, replication, and pleiotropy analyses may appear to be unrelated problems. However, they all involve testing a large number of composite null hypotheses, a statistically challenging task arising in many scientific settings, especially in genomic and biobank studies. A composite null is defined as the union of K individual null hypotheses, and is rejected only when all individual hypotheses are false. We will first show that classical tests such as Wald and likelihood ratios fail to follow chi-square tests and are often overly conservative, leading to substantial power loss. We propose several approaches to address this problem. We show that the divide-aggregate composite-null test (DACT) by leveraging a large number of tests and Efron's empirical null framework works well when signals are weak and sparse, but can exhibit error rate inflation in the presence of dense and strong signals. To overcome this limitation, we propose a tail likelihood ratio (TLR) test that is robust and powerful across diverse settings, encompassing weak to strong signal strengths, sparse to dense signal proportions, symmetric and unsymmetric signal distributions, and correlations among the large number of tests. We show that local FDR for a large number of composite nulls can be expressed as a function of a likelihood ratio, which is approximated in the tail region to construct the TLR statistic. We derive the null distribution of the TLR statistic by leveraging information from the large number of tests, and account for correlation among tests using a principal component-based approximation. Extensive simulations demonstrate that the proposed method maintains accurate error rate control in multiple comparisons and exhibits superior power gains compared to existing methods. Applications to multiple examples further illustrate the performance and utility of the proposed framework.

- Friday, November 6, 12:00-1:00 PM
  Location: G65 Jon M. Huntsman Hall

Heterogeneity Agnostic Transfer Inference Using Summary Statistics

Knowledge transfer across data sources offers great potential for improving parameter estimation and statistical inference in a target population by leveraging information from source populations, which often have substantially larger sample sizes. However, effective knowledge transfer is challenging due to complex and pervasive population heterogeneity and limited access to individual-level data. This talk presents two transfer inference methods designed to address these challenges. We first introduce divide-and-shrink (dShrink), a transfer estimation method that obtains closed-form estimates of target population parameters using summary statistics from the target and external source populations while accounting for population heterogeneity. We show that, under arbitrary population heterogeneity, the dShrink estimator is guaranteed to perform at least as well as the target-only estimator in terms of expected quadratic error. We next consider transfer testing for large-scale inference, e.g., in genome-wide studies, in a target population by leveraging a large collection of summary statistics from both target and source populations. We propose HEART (Heterogeneity-aware Empowerment via Adaptive Robust Transfer), a method that harnesses potentially biased source-population estimators to improve statistical power in the target population. By formulating target-specific inference as a composite-null hypothesis testing problem, HEART accommodates source-specific associations and provides robustness to population heterogeneity. Extensive simulation studies and real-data analyses demonstrate the strong performance of both methods, highlighting their potential as robust and effective tools for transfer inference across heterogeneous populations.

Knowledge transfer across data sources holds great promise for improving the estimation of parameters in a target population by leveraging data from a source population, which often has a much larger sample size than the target population. However, the effectiveness of knowledge transfer is often challenged by the complex and pervasive heterogeneity between data sources and the lack of access to individual-level data. This talk focuses on transfer inference methods. We first propose the divide-and-shrink (dShrink) method, a transfer estimation method that estimates target population parameters in a closed form using summary statistics from a target population and some external source populations while accounting for population heterogeneity. The dShrink estimator is guaranteed to perform no worse than the estimator based solely on the target population in terms of expected quadratic error under arbitrary population heterogeneity. We next consider transfer testing in the setting where a large number of tests in a target population are performed by leveraging a large number of summary statistics in a source population. We propose HEART (Heterogeneity-aware Empowerment via Adaptive Robust Transfer), a method that uses potentially biased estimators in the source population to increase the power of inference in the target population. By formulating the test of the target parameter as a composite-null hypothesis testing problem, HEART accommodates source-specific associations and is robust to population heterogeneity. Extensive simulations and real data analyses demonstrate the superior performance of the proposed methods and their potentials as robust tools for transfer inference.