Matching for Causal Inference

Fang Han

Department of Statistics, University of Washington, Seattle, WA

ABSTRACT

In two landmark Econometrica papers, Abadie and Imbens established that the nearest neighbor (NN) matching estimator for the average treatment effect is asymptotically normal when using a fixed number of NNs, but it remains semiparametrically inefficient and bootstrap inconsistent. In this talk, I will demonstrate that these limitations can be overcome by simply allowing the number of NNs to grow with the sample size. Under this modification, the NN matching estimator becomes asymptotically normal, doubly robust, semiparametrically efficient, and bootstrap consistent. These results are published in Econometrica and other leading journals.

Keywords: graph-based statistics, graph central limit theorem, imputation, density ratio

Synthetic Nearest Neighbours: Extending Synthetic Controls for Matrix Completion with Missing Not at Random Data

Anish Agarwal[1], Munther Dahleh[2], Devavrat Shah[2], **Dennis Shen[3]**

[1]*Columbia University*

[2]*MIT*

[3]*USC*

ABSTRACT

We develop a causal framework for matrix completion under missing not at random (MNAR) data. Drawing on the method of synthetic controls from the econometric panel data literature, our approach relaxes two core assumptions that underlie standard MNAR matrix completion models: positivity (every entry is observed with positive probability) and independence (observations are independent across entries). Unlike traditional panel data models that often rely on rigid block-sparsity patterns, our framework accommodates flexible and heterogeneous observation structures commonly encountered in matrix completion problems. To operationalize our framework, we propose synthetic nearest neighbors (SNN), a novel algorithm that blends elements of K-nearest neighbors with synthetic controls. Under suitable assumptions on the underlying matrix and observed sparsity pattern, we prove that SNN achieves entrywise mean-squared error convergence for estimating the mean matrix, attaining a near-parametric rate. We further extend our analysis to heteroskedastic variance estimation, establishing that SNN attains entrywise mean-squared error convergence under bounded noise and asymptotic unbiasedness under general sub-Gaussian noise. Simulations studies corroborate our theoretical findings and demonstrate the robustness of SNN across a range of MNAR scenarios.

Keywords: Panel data, heteroskedastic variance estimation, causal inference

A Pleiotropy-Free Bayesian Model for Two- Sample Summary-Data Mendelian Randomization with Binary Outcomes

An-Shun Tai

Institute of Statistics and Data Science, National Tsing Hua University, Taiwan

ABSTRACT

Mendelian randomization (MR) is a powerful approach for estimating causal effects in the presence of unmeasured confounding. With the rapidly increasing sample sizes of genome-wide association studies, MR analyses based on twosample summary data across a broad range of phenotypes have become increasingly common. However, traditional summary-data MR techniques rely on the exclusion restriction, an assumption often violated due to pleiotropy. Although numerous methods have been proposed to address this issue, existing approaches still struggle to provide accurate causal effect estimates for binary outcomes. In this study, we derive the exact pleiotropy-induced bias under the counterfactual framework without requiring additional modeling assumptions. Building on this insight, we develop a pleiotropy-free Bayesian MR model for summary data with binary outcomes using a spike-and-slab prior to accommodate invalid instrumental variables. Through extensive simulation studies under diverse scenarios, we demonstrate the superior performance of our approach and offer a comprehensive comparison against current state-of-the-art methods.

Keywords: Mendelian Randomization; Causal Inference; Pleiotropy Bias; Bayesian Modeling; Two-Sample Summary Data.