Power and Sample Size Calculations for Causal Inference with Observational Data

Fan Li[1], Bo Liu[1]

Department of Statistical Science, Duke University

ABSTRACT

This paper investigates the theoretical foundation and develops analytical formulas for sample size and power calculations for causal inference with observational data. By analysing the variance of the inverse probability weighting estimator of the average treatment effect, we decompose the power calculations into three components: propensity score distribution, potential outcome distribution, and their correlation. We show that to determine the minimal sample size of an observational study, it is sufficient under mild conditions to have two parameters additional to the standard inputs in the power calculation of randomised trials, which quantify the strength of the confounder-treatment and the confounder-outcome association, respectively. For the former, we propose using the Bhattacharyya coefficient, which measures the covariate overlap and, together with the treatment proportion, leads to a uniquely identifiable and easily computable propensity score distribution. For the latter, we propose a sensitivity parameter bounded by the R-squared statistic of the regression of the outcome on covariates. Utilising the Lyapunov Central Limit Theorem on the linear combination of covariates, our procedure does not require distributional assumptions on the multivariate covariates. We develop an associated R package PSpower.

Keywords: causal inference; observational study; overlap; power; sample size

Back to Sessions List

Semiparametric Mediation Analysis Using Single-Index Models

Yen-Tsung Huang

Institute of Statistical Science, Academia Sinica

ABSTRACT

Mediation analysis is increasingly used to investigate how an exposure Z influences an outcome Y through a set of mediators $\boldsymbol{M}$. However, most existing methods either rely on parametric assumptions for the mediators and outcome, or are limited to binary exposures and a single mediator. We propose a novel algorithm that accommodates single-index models (SIMs) with non-binary exposures and adjustment for potential confounders, while avoiding parametric assumptions for both mediators and outcome. Specifically, we introduce two SIMs: one modeling the outcome conditional on the exposure, mediators, and confounders, and the other modeling the mediators conditional on the exposure and confounders. We derive a mediation estimand of the form $\mathrm{E}\big[\mathrm{E}\big[Y_j|Z_j = z_a, \boldsymbol{M}_j = \boldsymbol{M}_i\big]|Z_i = z_b\big]$, representing the expected outcome under a joint intervention setting the exposure to z_a and the mediators to their counterfactual values under z_b. The proposed estimator proceeds in four steps: (1) regress Y on Z and $\boldsymbol{M}$ using an SIM; (2) predict Y given z_a and $\boldsymbol{M}$; (3) regress the predicted Y on Z using another SIM; (4) predict the outcome at z_b. We establish the asymptotic properties of this estimator and evaluate its finite-sample performance through simulation studies. Finally, we demonstrate the practical utility of our method with two applications. The first examines the effect of socioeconomic status on body mass index mediated by DNA methylation of the *FASN* (fatty acid synthase) gene. The second investigates the effect of obesity on glycemic control through triglyceride, fasting glucose, and urine microalbumin.

Keywords: causal inference; mediation analyses; multiple mediators; single-index models

Back to Sessions List

Causal Mediation Analysis: A Summary-Data Mendelian Randomization Approach

Shu-Chin Lin[1,2], Sheng-Hsuan Lin[3], Tian Ge[4,5,6], Chia-Yen Chen[7], Yen-Feng Lin[1,8,9]

[1]Center for Neuropsychiatric Research, National Health Research Institutes, Miaoli, Taiwan

[2]Institute of Statistics and Data Science, National Taiwan University, Taipei, Taiwan

[3]Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

[4]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA

[5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[6]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

[7]Translational Medicine, Biogen, Cambridge, Massachusetts, USA

[8]Department of Public Health & Medical Humanities, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

[9]Institute of Behavioral Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan

ABSTRACT

Summary-data Mendelian randomization (MR) has become a popular tool for causal mediation analysis, offering an alternative to traditional methods. Two MR-based approaches corresponding to the difference and product methods have been implemented using inverse-variance weighted estimation (MR-IVW). However, existing methods often lack statistical efficiency, robustness, and rigorous inference procedures. In this study, we develop improved MR-based mediation frameworks using summary-level data, typically from genome-wide association studies (GWAS) data. Our contributions are threefold: we propose new variance estimators for mediation effects, establish formal inference procedures, and develop robust strategies to account for pleiotropy.

Keywords: causal inference; indirect effect; mediation analysis; mediation proportion; summary-data Mendelian randomization.

Back to Sessions List

Sobolev Gradient Ascent for Optimal Transport: Barycenter Optimization and Convergence Analysis

Kaheon Kim[1], Bohan Zhou[2], **Changbo Zhu[1]**, Xiaohui Chen[3], Arlina Shen

[1]*University of Notre Dame*

[2] University of California, Santa Barbara

[3] University of Southern California

ABSTRACT

This paper introduces a new constraint-free concave dual formulation for the Wasserstein barycenter. Tailoring the vanilla dual gradient ascent algorithm to the Sobolev geometry, we derive a scalable Sobolev gradient ascent (SGA) algorithm to compute the barycenter for input distributions supported on a regular grid. Despite the algorithmic simplicity, we provide a global convergence analysis that achieves the same rate as the classical subgradient descent methods for minimizing nonsmooth convex functions in the Euclidean space. A central feature of our SGA algorithm is that the computationally expensive c-concavity projection operator enforced on the Kantorovich dual potentials is unnecessary to guarantee convergence, leading to significant algorithmic and theoretical simplifications over all existing primal and dual methods for computing the exact barycenter. Our numerical experiments demonstrate the superior empirical performance of SGA over the existing optimal transport barycenter solvers.

Keywords: optimal transport; Wasserstein barycenter; concave dual; gradient ascent;

Back to Sessions List