Contents

Abstracts

Workshop on Causal Inference with Big Data

16–23 December 2021

1 Ding-Geng (Din) Chen

Arizona State University, USA

Big data inference and statistical meta-analysis

Abstract

Statistical meta-analysis (MA) is a common statistical approach in big data inference to combine meta-data from diverse studies to reach a more reliable and efficient conclusion. It can be performed by either synthesizing study-level summary statistics (MA-SS) or modeling individual participant-level data (MA-IPD), if available. However, it remains not fully understood whether the use of MA-IPD indeed gains additional efficiency over MA-SS. In this talk, we review the classical fixed-effects and random-effects meta-analyses, and further discuss the relative efficiency between MA-SS and MA-IPD under a general likelihood inference setting. We show theoretically that there is no gain of efficiency asymptotically by analyzing MA-IPD, provided that the random-effects follow the Gaussian distribution, and maximum likelihood estimation is used to obtain summary statistics. Our findings are further confirmed by extensive Monte-Carlo simulation studies and real data analyses.
*This talk is based on the publication: Chen, D.G, Liu, D., Min, X. and Zhang H. (2020). Relative efficiency of using summary and individual information in random-effects meta-analysis. Biometrics, 76(4): 119-1329. (https://doi.org/10.1111/biom.13238).

2 Peng Ding

University of California, Berkeley, USA

To adjust or not to adjust? Estimating the average treatment effect in randomized experiments with missing covariates

Abstract

Complete randomization allows for consistent estimation of the average treatment effect based on the difference in means of the outcomes without strong modeling assumptions on the outcome-generating process. Appropriate use of the pretreatment covariates can further improve the estimation efficiency. However, missingness in covariates is common in experiments and raises an important question: should we adjust for covariates subject to missingness, and if so, how? The unadjusted difference in means is always unbiased. The complete-covariate analysis adjusts for all completely observed covariates and improves the efficiency of the difference in means if at least one completely observed covariate is predictive of the outcome. Then what is the additional gain of adjusting for covariates subject to missingness? A key insight is that the missingness indicators act as fully observed pretreatment covariates as long as missingness is not affected by the treatment, and can thus be used in covariate adjustment to bring additional estimation efficiency. This motivates adding the missingness indicators to the regression adjustment, yielding the missingness-indicator method as a well-known but not so popular strategy in the literature of missing data. We recommend it due to its many advantages. We also propose modifications to the missingness-indicator method based on asymptotic and finite-sample considerations. To reconcile the conflicting recommendations in the missing data literature, we analyze and compare various strategies for analyzing randomized experiments with missing covariates under the design-based framework. This framework treats randomization as the basis for inference and does not impose any modeling assumptions on the outcome-generating process and missing-data mechanism.

3 Yen-Tsung Huang

Academia Sinica, Taiwan

Nonparametric causal mediation in a time-to-event setting

Abstract

A causal mediation model with multiple time-to-event mediators is exemplified by the natural course of human disease marked by sequential milestones with a time-to-event nature. For example, from hepatitis B infection to death, patients may experience intermediate events such as liver cirrhosis and liver cancer. The sequential events of hepatitis, cirrhosis, cancer, and death are susceptible to right censoring; moreover, the latter events may preclude the former events. Casting the natural course of human diseases in the framework of causal mediation modeling, we establish a model with intermediate and terminal events as the mediators and outcomes, respectively. We define the interventional analog of path-specific effects (iPSEs) as the effect of an exposure on a terminal event mediated (or not mediated) by any combination of intermediate events without parametric models. The expression of a counting process-based counterfactual hazard is derived under the sequential ignorability assumption. We employ composite nonparametric likelihood estimation to obtain maximum likelihood estimators for the counterfactual hazard and iPSEs. Our proposed estimators achieve asymptotic unbiasedness, uniform consistency, and weak convergence. Applying the proposed method, we show that hepatitis B induced mortality is mostly mediated through liver cancer and/or cirrhosis whereas hepatitis C induced mortality may be through extrahepatic diseases.

4 Zhichao Jiang

UMass Amherst, USA

Experimental evaluation of algorithm-assisted human decision making

Abstract

Despite an increasing reliance on fully-automated algorithmic decision-making in our day-to-day lives, human beings still make highly consequential decisions. As frequently seen in business, healthcare, and public policy, recommendations produced by algorithms are provided to human decision-makers to guide their decisions. While there exists a fast-growing literature evaluating the bias and fairness of such algorithmic recommendations, an overlooked question is whether they help humans make better decisions. Using the concept of principal stratification, we develop a statistical methodology for experimentally evaluating the causal impacts of algorithmic recommendations on human decisions. We propose the evaluation quantities of interest, identification assumptions, and estimation strategies. We also develop sensitivity analyses to assess the robustness of empirical findings to the potential violation of a key identification assumption. We apply the proposed methodology to preliminary data from the first-ever randomized controlled trial that evaluates the pretrial Public Safety Assessment (PSA) in the criminal justice system.

5 Hyunseung Kang

University of Wisconsin–Madison, USA

Assumption-lean analysis of cluster randomized trials in infectious diseases for intent-to-treat effects and network effects

Abstract

Cluster randomized trials (CRTs) are a popular design to study the effect of interventions in infectious disease settings. However, standard analysis of CRTs primarily relies on strong parametric methods, usually mixed-effect models to account for the clustering structure, and focuses on the overall intent-to-treat (ITT) effect to evaluate effectiveness. The article presents two assumption-lean methods to analyze two types of effects in CRTs, ITT effects and network effects among well-known compliance groups. For the ITT effects, we study the overall and the heterogeneous ITT effects among the observed covariates where we do not impose parametric models or asymptotic restrictions on cluster size. For the network effects among compliance groups, we propose a new bound-based method that uses pretreatment covariates, classification algorithms, and a linear program to obtain sharp bounds. A key feature of our method is that the bounds can become narrower as the classification algorithm improves and the method may also be useful for studies of partial identification with instrumental variables. We conclude by reanalyzing a CRT studying the effect of face masks and hand sanitizers on

transmission of 2008 interpandemic influenza in Hong Kong. This is joint work with Chan Park (UW-Madison).

6 Edward Kennedy

Carnegie Mellon University, USA
Minimax rates for heterogeneous effect estimation

Abstract

Heterogeneous effect estimation plays a crucial role in causal inference, with applications across medicine and social science. Many methods for estimating conditional average treatment effects (CATEs) have been proposed in recent years, but there are important theoretical gaps in understanding if and when such methods are optimal. This is especially true when the CATE has nontrivial structure (e.g., smoothness or sparsity). This talk surveys work across two recent papers in this context. First, we study a two-stage doubly robust CATE estimator and give a generic model-free error bound, which, despite its generality, yields sharper results than those in the current literature. The second contribution is aimed at understanding the fundamental statistical limits of CATE estimation. We resolve this open problem by deriving a minimax lower bound, with matching upper bound based on a new higher-order influence function-based estimator.

7 Jialiang Li

National University of Singapore, Singapore
Multi-threshold structural equation model

Abstract

In this paper, we consider the instrumental variable estimation for causal regression parameters with multiple unknown structural changes across subpopulations.

We propose a multiple change point detection method to determine the number of thresholds and estimate the threshold locations in the two-stage least squares procedure. After identifying the estimated threshold locations, we use the Wald method to estimate the parameters of interest, i.e., the regression coefficients of the endogenous variable. Based on some technical assumptions, we carefully establish the consistency of estimated parameters and the asymptotic normality of causal coefficients. Simulation studies are included to examine the performance of the proposed method. Finally, our method is illustrated via an application of the Philippine farm households data for which some new findings are discovered.

8 Wei-Yin Loh

University of Wisconsin-Madison, USA
Missing values, regression trees, and causal inference

Abstract

There seems to be some recent interest in applying machine learning methods, specifically regression trees and forests, to causal inference. Regression trees and forests are uniquely suited for analysis of data with missing values because they can do so without prior missing value imputation. This talk will use a handful of examples to highlight the differences between regression tree algorithms and show that missing-value imputation may be illogical or result in information loss.

9 Alex Luedkte

University of Washington, USA
Efficient estimation under data fusion

Abstract

We aim to make inferences about a smooth, finite-dimensional parameter by fusing data from multiple sources together. Previous works have studied the estimation of a variety of parameters in similar data fusion settings, including in the estimation of the average treatment effect, optimal treatment rule, and average reward, with the majority of them merging one historical data source with covariates, actions, and rewards and one data source of the same covariates. In this work, we consider the general case where one or more data sources align with each part of the distribution of the target population, for example, the conditional distribution of the reward given actions and covariates. We describe potential gains in efficiency that can arise from fusing these data sources together in a single analysis, which we characterize by a reduction in the semiparametric efficiency bound. We also provide a general means to construct estimators that achieve these bounds. In numerical experiments, we show marked improvements in efficiency from using our proposed estimators rather than their natural alternatives. Finally, we illustrate the magnitude of efficiency gains that can be realized in vaccine immunogenicity studies by fusing data from two HIV vaccine trials.

10 Caleb Miles

Columbia University, USA

Optimal tests of the composite null hypothesis arising in mediation analysis

Abstract

The indirect effect of an exposure on an outcome through an intermediate variable can be identified by a product of regression coefficients under certain causal and regression modeling assumptions. Thus, the null hypothesis of no indirect effect is a composite null hypothesis, as the null holds if either regression coefficient is zero. A consequence is that traditional hypothesis tests are severely underpowered near the origin (i.e., when both coefficients are small with respect to standard errors). We propose hypothesis tests that (i) preserve level alpha type 1 error, (ii) meaningfully improve power when both true underlying effects are small relative to sample size, and (iii) preserve power when at least one is not. One approach gives a closed-form test that is minimax optimal with respect to local power over the alternative parameter space. Another uses sparse linear programming to produce an approximately optimal test for a Bayes risk criterion. We provide an R package that implements the minimax optimal test.

11 Elizabeth Ogburn

Johns Hopkins University, USA

Disentangling confounding and nonsense associations due to dependence

Abstract

Nonsense associations can arise when an exposure and an outcome of interest exhibit similar patterns of dependence. Confounding is present when potential outcomes are not independent of treatment. This talk will describe how confusion about these two phenomena results in shortcomings in popular methods in two areas: causal inference with multiple treatments and unmeasured confounding and causal and statistical inference with social network data. For each of these areas I will demonstrate the flaws in existing methods and describe new methods that were inspired by careful consideration of dependence and confounding.

12 James Robins

Harvard University, USA

Estimation of optimal testing and treatment regimes under no direct effect (NDE) of testing

Abstract

In this talk I describe new, highly efficient estimators of optimal joint testing and treatment regimes under the no direct effect assumption that a given laboratory, diagnostic, or screening test has no effect on a patient's clinical outcomes, except through the effect of the test results on the choice of treatment. The proposed estimators attain high efficiency because they leverage this 'no direct effect of testing' (NDE) assumption. What is surprising is that, in a substantive study of HIV infected subjects, our new estimators delivered a 50-fold increase in efficiency (and, thus, a 50 fold reduction in required sample size) compared to estimators that fail to leverage the NDE assumption! In this talk I review the results of this HIV study, describe the new estimators, and provide guidance as to when such large gains in efficiency are to be expected. Areas in which our new, more efficient estimators should be particularly important is that of cost-benefit analyses wherein the costs of diagnostic tests (such as MRIs to screen for lung cancer, mammograms to screen for breast cancer, and urinary cytology to screen for bladder cancer) are weighed against the clinical value of the information supplied by the test results.

This is joint work with Lin Liu, Zach Shahn, and Andrea Rotnitzky.

13 Baoluo Sun

National University of Singapore, Singapore

On multiply robust mendelian randomization (MR^2) with many invalid genetic instruments

Abstract

Mendelian randomization (MR) is a popular instrumental variable (IV) approach, in which genetic markers are used as IVs. In order to improve efficiency, multiple markers are routinely used in MR analyses, leading to concerns about bias due to possible violation of IV exclusion restriction of no direct effect of any IV on the outcome other than through the exposure in view. To address this concern, we introduce a new class of Multiply Robust MR (MR^2) estimators that are guaranteed to remain consistent for the causal effect of interest provided that a set of at least $k^\dagger$ out of K candidate IVs are valid, for $k^\dagger \leq K$ set by the analyst ex ante, without necessarily knowing which IVs are invalid. We provide formal semiparametric theory supporting our results, and characterize the semiparametric efficiency bound for the exposure causal effect which cannot be improved upon by any regular estimator with our favorable robustness property. We conduct extensive simulation studies and apply our methods to a large-scale analysis of UK Biobank data, demonstrating the superior empirical performance of MR^2 compared to competing MR methods.

14 Zhiqiang Tan

Rutgers University, USA

Doubly robust semiparametric inference using regularized calibrated estimation with high-dimensional data

Abstract

Consider semiparametric estimation where a doubly robust estimating function for a low-dimensional parameter is available, depending on two working models. With high-dimensional data, we develop regularized calibrated estimation as a general method for estimating the parameters in the two working models, such that valid Wald confidence intervals can be obtained for the parameter of interest under suitable sparsity conditions if either of the two working models is correctly specified. We propose a computationally tractable two-step algorithm and provide rigorous theoretical analysis which justifies sufficiently fast rates of convergence for the regularized calibrated estimators in spite of sequential construction and establishes a desired asymptotic expansion for the doubly robust estimator. As concrete examples, we discuss applications to partially linear, log-linear, and logistic models and estimation of average treatment effects. Numerical studies in the former three examples demonstrate superior performance of our method, compared with debiased Lasso.

15 Eric Tchetgen Tchetgen

University of Pennsylvania, USA

Theory for identification and inference with synthetic controls: a proximal causal inference approach

Abstract

Synthetic control methods are commonly used to estimate the treatment effect on a single treated unit in panel data settings. A synthetic control (SC) is a weighted average of control units built to match the treated unit's pre-treatment outcome trajectory, with weights typically estimated by regressing pre-treatment outcomes of the treated unit to those of the control units. However, it has been established that such regression estimators can fail to be consistent because of error-in-variables problem. We introduce a proximal causal inference framework to formalize identification and inference for both the SC weights and the treatment effect on the treated. We show that control units previously perceived as unusable can be repurposed to consistently estimate the SC weights. We also propose to view the difference in the post-treatment outcomes between the treated unit and the SC as a time series, which opens the door to a rich literature on time-series analysis for treatment effect estimation. We further extend the traditional linear model to accommodate general nonlinear models allowing for binary and count outcomes which are understudied in the SC literature. We illustrate our proposed methods with simulation studies and an application to evaluation of the 1990 German Reunification.

16 Stijin Vansteelandt

Ghent University, Belgium

Assumption-lean Cox regression

Abstract

Inference for the parameters indexing Cox regression models is routinely based on the assumption that the model is correct and a priori specified. This is unsatisfactory because the chosen model is usually the result of a data-adaptive model selection process, which induces bias and excess uncertainty that is not usually acknowledged; moreover, the assumptions encoded in the resulting model rarely represent some a priori known, ground truth. Standard inferences may therefore lead to bias in effect estimates, and may moreover fail to give a pure reflection of the information that is contained in the data. Inspired by developments on assumption-free inference for so-called projection parameters, we here propose nonparametric definitions of main effect estimands which reduce to standard main effect parameters in Cox regression models when these models are correctly specified, but continue to capture the primary (conditional) association between a variable and an event time, even when these models are misspecified. We achieve an assumption-lean inference for these estimands by deriving their influence curve under the nonparametric model and invoking flexible data-adaptive algorithms.

17 Miao Wang

Peking University, China

A stableness of resistance model for nonresponse adjustment with callback data

Abstract

The survey world is rife with nonresponse and in many situations the missingness mechanism is not at random, which is a major source of bias for statistical inference. Nonetheless, the survey world is rich with paradata that track the data collection process. A traditional form of paradata is callback data that record attempts to contact. Although it has been recognized that callback data are useful for nonresponse adjust- ment, they had not been used widely in statistical analysis until recently. In particular, there have been a few attempts that use callback data to estimate response propensity scores, which rest on fully parametric models and fairly stringent assumptions. In this paper, we propose a novel stableness of resistance assumption for identifying the propensity scores and the outcome distribution of interest, without placing any parametric models. We establish the semiparametric efficiency theory, derive the efficient influence function, and pro- pose a suite of semiparametric estimation methods including doubly robust ones, which generalize existing parametric approaches. We also consider extension of this framework to causal inference for unmeasured confounding adjustment. Application to a Consumer Expenditure Survey dataset suggests a tendency to not respond among people with high housing expenditures, and reanalysis of Card (1995)'s dataset on the return to schooling shows a smaller effect of education in the overall population than in the respondents.

18 Menggang Yu

University of Wisconsin-Madison, USA

Robust sample weighting to facilitate individualized treatment rule learning for a target population

Abstract

We consider a setting when a study or source population for individualized-treatment-rule (ITR) learning can differ from the target population of interest. We assume subject covariates are available from both populations, but treatment and outcome data are only available from the source population. Existing methods use "importance" and/or "overlap" weights to adjust for the covariate differences between the two populations. We develop a general weighting framework that allow a better bias-variance trade-off than existing weights. Our method seeks covariate balance over a non-parametric function class characterized by a reproducing kernel Hilbert space. Our weights encompasse the importance weights and overlap weights as special cases. Numerical examples demonstrate that our weights can improve many ITR learning methods for the target population that rely on weighting.