丘成桐数学科学中心
YAU MATHEMATICAL SCIENCES CENTER

TSIMF
清华三亚国际数学论坛

因果推断:理论与实践

Causal Inference: Bridging Theory and Practice

January 2-6, 2025

Room A-110, TSIMF

组织者 ORGANIZERS

Fan Li(李凡 女), Duke University

Fan Yang(杨帆), Tsinghua University

Peng Ding(丁鹏), University of California, Berkeley

Zhichao Jiang(蒋智超), Sun Yat-Sen University

Contents

About

Causal Inference: Bridging Theory and Practice
因果推断：理论与实践

Date

January 2-6, 2025

Venue

Room A-110, TSIMF

Organizers

Fan Li(李凡 女), Duke University
Fan Yang(杨帆), Tsinghua University
Peng Ding(丁鹏), University of California, Berkeley
Zhichao Jiang(蒋智超), Sun Yat-Sen University

Abstract

Causality has long been central to the human philosophical debate and scientific pursuit. Among the many relevant questions on causality, statistics arguably can contribute the most to the question of measuring the effects of “causes”, or more specifically, interventions or actions. The last two decades have witnessed an explosive growth in statistical and machine learning theory and methods for causal inference. These methods have been increasingly applied to solve real world problems in many disciplines. This workshop will focus on exchanging new developments in theory and methods as well as impactful interdisciplinary applications. Main topics will include: (i) design and analysis of complex randomized experiments; (ii) natural and quasi-experimental designs, including instrumental variables; (iii) machine learning methods for causal inference; (iv) causal inference in action; (v) accessible causal inference: software and translational work.

Description of the aim

This workshop aims to bring together a group of respected and active researchers to present their ongoing work in several most important current areas of causal inference, including complex randomized experiments, natural experiments, machine learning methods, software development, and interdisciplinary applications. It will provide an opportunity for researchers at different career stages to exchange ideas with internationally leading experts and foster new collaborations in an intimate environment. In particular, the workshop will provide a platform for young researchers to showcase their achievements and network. Overall, the workshop is expected to contribute to building and strengthening the causal inference community in China.

Schedule

January 3, 2025, Friday

Time 日期	Name 报告人	Title 报告题目
7:30-8:30	Breakfast	
9:00-9:10	Opening: Fan Li(李凡 女)	
Chair: Fan Li(李凡 女)		
9:10-9:30	Fabrizia Mealli	Causal Inference when Intervention Units and Outcome Units Differ
9:30-9:35	Discussion	
9:35-9:55	Shu Yang(杨淑)	Enhancing Statistical Validity and Power in Hybrid Controlled Trials: A Randomization Inference Approach with Conformal Selective Borrowing
9:55-10:00	Discussion	
10:00-10:20	Bo Li(黎波)	Practical Performative Policy Learning with Strategic Agents
10:20-10:25	Discussion	
10:25-10:35	Break	
Chair: Bo Li(黎波)		
10:35-10:55	Guanglei Hong（洪光磊）	Two-Phase Treatment with Noncompliance: Identifying the Cumulative ATE via Multisite IV
10:55-11:00	Discussion	
11:00-11:20	Fan Yang(杨帆)	Identifiability of the instrumental variable model with the treatment and outcome missing not at random
11:20-11:25	Discussion	
11:25-11:45	Wei Li(李伟)	Discovery and inference of possibly bi-directional causal relationships with invalid instrumental variables
11:45-11:50	Discussion	
11:50-12:00	Group Photo	
12:00-13:30	Lunch	
Chair: Fan Yang(杨帆)		
14:00-14:20	Fan Li（李凡 女）	Two-stage least squares with treatment effect heterogeneity
14:20-14:25	Discussion	
14:25-14:45	Yue Liu(刘越)	Local Causal Discovery with Background Knowledge
14:45-14:50	Discussion	
14:50-15:10	Xinran Li(李欣然)	Cluster-robust inference with a single treated cluster using the t-test
15:10-15:15	Discussion	
15:15-15:25	Break	
Chair: Fan Li(李凡 男)		
15:25-15:45	Jingshu Wang(王静姝)	Causal inference and calibration for Within-Family Mendelian Randomization
15:45-15:50	Discussion	
15:50-16:10	Yuhao Wang(王禹皓)	Debiased regression adjustment in completely randomized experiments with moderately high-dimensional covariates
16:10-16:15	Discussion	
16:15-16:35	Wang Miao(苗旺)	Extreme-based causal effect learning (EXCEL) with unmeasured light-tailed confounding
16:35-16:40	Discussion	
18:00-20:00	Banquet	

January 4, 2025, Saturday

Time 日期	Name 报告人	Title 报告题目
7:30-8:30	Breakfast	
Chair: Hanzhong Liu(刘汉中)		
9:00-9:20	Rajarshi Mukherjee	Method-of-Moments Inference for GLMs and Doubly Robust Functionals under Proportional Asymptotics
9:20-9:25	Discussion	
9:25-9:45	Fan Li(李凡 男)	Causal Inference under Complex Biases
10:10-10:15	Discussion	
9:50-10:10	Lin Liu(刘林)	Covariate adjustment in RCT motivated by HOIF
10:20-10:25	Discussion	
10:15-10:25	Break	
Chair: Yunan Wu(吴宇楠)		
10:25-10:45	Huazhen Lin(林华珍)	Distribution-Type Index For Causal Inference*
10:45-10:50	Discussion	
10:50-11:10	Anpeng Wu(吴安鹏)	Causal Inference with Instrumental Variables in Complex Scenarios
11:10-11:15	Discussion	
11:15-11:35	Hanzhong Liu(刘汉中)	Estimation and inference of average treatment effects under heterogeneous additive treatment effect model
11:35-11:40	Discussion	
11:40-12:00	Xingjian Zhang(张行健)	Formulating generalized causal models from the no-signalling physical principle
12:00-12:05	Discussion	
12:05-13:30	Lunch	
Chair: Guanglei Hong(洪光磊)		
14:00-14:20	Peng Ding(丁鹏)	Factorial Difference-in-Differences
14:20-14:25	Discussion	
14:25-14:45	Jingyuan Liu(刘婧媛)	Statistical Inference for Mediation Models with High Dimensional Exposures and Mediators
14:45-14:50	Discussion	
14:50-15:10	Xiaojie Mao(毛小介)	On the Role of Surrogates in the Efficient Estimation of Treatment Effects with Limited Outcome Data
15:10-15:15	Discussion	
15:15-15:25	Break	
Chair: Huazhen Lin(林华珍)		
15:25-15:45	Yuehan Yang(杨玥含)	Design-based theory for Lasso adjustment in randomized block experiments and rerandomized experiments
15:45-15:50	Discussion	
15:50-16:10	Guangyu Tong(同广宇	Moving towards best practice when using propensity score weighting in survey observational studies
16:10-16:15	Discussion	
16:15-16:35	Yuqian Zhang(张宇谦)	Balancing Utility and Cost in Dynamic Treatment Regimes
16:35-16:40	Discussion	
16:40-17:00	Free Discussion	
17:30-19:00	Dinner	

January 5, 2025, Sunday

Time 日期	Name 报告人	Title 报告题目
7:30-8:30	Breakfast	
Chair: Peng Ding(丁鹏)		
9:00-9:20	Min Zhang(张敏)	Biostatistics, data and cohort study
9:20-9:25	Discussion	
9:25-9:45	Wei Luo(骆威)	An exhaustive selection of sufficient adjustment sets for causal inference
9:45-9:50	Discussion	
9:50-10:10	Ying Yan(严颖)	Matching-based Policy Learning (Name:Ying Yan; Affiliation: Sun Yat-sen University)
10:10-10:15	Discussion	
10:15-10:35	Yunan Wu(吴宇楠)	Model-Assisted Uniformly Honest Inference for Optimal Treatment Regimes in High Dimension
10:10-10:15	Discussion	
10:15-10:30	Free Discussion and Conclusion	
12:00-13:30	Lunch	
13:30-17:30	Free Afternoon	
17:30-19:00	Dinner(if plan to leave on Jan.6)	

January 6 - January 7, 2025

Departure

Titles and Abstracts

Opening: Fan Li(李凡 女)

Chair:Fan Li(李凡 女)

Causal Inference when Intervention Units and Outcome Units Differ

Fabrizia Mealli
European University Institute, University of Florence

We study causal inference in settings characterized by interference with a bipartite structure. There are two distinct sets of units: intervention units to which an intervention could be applied and outcome units on which the outcome of interest can be measured.
Outcome units may be affected by interventions on some, but not all, intervention units, as captured by a bipartite graph. Examples of this setting can be found across many applications, for example in analyses of the impact of pollution abatement in plants on health outcomes for individuals, or the effect of transporation network expansions on regional economic activity, as the HSR in China. We introduce and discuss a variety of causal estimands for these bipartite settings. We propose weighting estimators for these estimands from a design-based perspective, based on (partial) knowledge of the bipartite network.
We discuss how knowledge of the bipartite graph allows unbiased estimation of these causal quantities. We also derive the variance and asymptotic behavior of estimators under specific regimes.
We do not impose restrictions on the functional form of the exposure mapping and the potential outcomes, thus allowing for heterogeneity, non-linearity, non-additivity and potential interactions in treatment effects. Asymptotic results depend on the number of outcome units as well as the number of intervention units growing, and require conditions on the sparsity of the bipartite graph.

This is joint work with Georgia Papadogeorgou, Zhaoyan Song and Guido Imbens.

Enhancing Statistical Validity and Power in Hybrid Controlled Trials: A Randomization Inference Approach with Conformal Selective Borrowing

Shu Yang(杨淑)
North Carolina State University

Randomized controlled trials (RCTs) are the gold standard for causal inference on treatment effects, but they can be limited by small sample sizes due to the indications associated with rare diseases and small patient populations, where ethical concerns or patient reluctance may limit control group assignment. Hybrid controlled trials use external controls (ECs) from historical studies or large observational databases to enhance statistical efficiency. However, non-randomized ECs can introduce biases that compromise validity and inflate Type I errors for treatment discovery,

particularly in small samples. To address this, we extend the Fisher randomization test to hybrid controlled trials. Our approach involves a test statistic combining RCT and EC data and is based solely on randomization in the RCT. This method strictly controls the Type I error rate, even with biased ECs, and improves power by incorporating unbiased ECs.
To mitigate the power loss caused by biased ECs, we introduce Conformal Selective Borrowing, which uses individual conformal p-values to selectively incorporate unbiased ECs, offering the flexibility to use either computationally efficient parametric models or off-the-shelf machine learning models to construct the score function, along with model-agnostic reliability. We identify a risk-benefit trade-off in the power of FRT, associated with different selection thresholds for conformal p-values, analogous to the mean squared error trade-offs observed in the data integrative estimators. We propose a data-driven selection of the threshold value to achieve robust performance across different levels of hidden bias. The advantages of our method are demonstrated through simulations and an application to a small-sized lung cancer trial with ECs from the National Cancer Database.

Practical Performative Policy Learning with Strategic Agents

Bo Li(黎波)
Tsinghua University

This paper studies the performative policy learning problem, where agents adjust their features in response to a released policy to improve their potential outcomes, inducing an endogenous distribution shift. There has been a growing interest in training machine learning models in strategic environments, including strategic classification Hardt et al. (2016) and performative prediction Perdomo et al. (2020). However, existing approaches often rely on restrictive parametric assumptions: micro-level utility models in strategic classification and macro-level data distribution maps in performative prediction, severely limiting scalability and generalizability. We approach this problem as a complex causal inference task, relaxing parametric assumptions on both micro-level agent behavior and macro-level data distribution. Inspired by bounded rationality, we uncover a practical low-dimensional structure in distribution shifts and construct an effective mediator in the causal path from the deployed model to the shifted data. We then propose a gradient-based policy optimization algorithm with a differentiable classifier serving as a substitute for the high-dimensional distribution map. This algorithm efficiently leverages batch feedback and limited manipulation patterns, achieving significantly higher sample efficiency compared to methods reliant on bandit feedback or zero-order optimization. We also provide theoretical guarantees for algorithmic convergence. Extensive and challenging experiments on high-dimensional settings demonstrate the practical efficacy of our method.

Chair:Bo Li(黎波)

Two-Phase Treatment with Noncompliance: Identifying the Cumulative ATE via Multisite IV

Guanglei Hong(洪光磊)
University of Chicago

In a multi-phase intervention program, the cumulative average treatment effect (ATE) is often the causal estimand of key interest. Yet some members of the target population may display noncompliant behaviors in Phase II if not responding well to the Phase-I treatment. In a multisite randomized experiment, however, noncompliance tends to be constrained by the stochastic availability of slots in the alternative treatment group at a site in Phase II, which makes the notion of the "complier average treatment effect" problematic. The Phase-I treatment is expected to affect an individual's potential outcomes through several pathways and thereby violating the exclusion restriction. Extending a multisite multiple-mediator instrumental variable (MSMM-IV) strategy, we clarify conditions for identifying the cumulative ATE of a two-phase treatment by employing the random assignment of the Phase-I treatment as the IV. Our strategy requires neither the exclusion restriction nor the sequential strong ignorability assumption. We compare the performance of alternative methods for estimation and statistical inference through simulations. Reanalyzing data from the well-known Tennessee class size study in which students and teachers were assigned at random to either a small class or a regular class in kindergarten (Phase I) yet noncompliance occurred in Grade 1 (Phase II), we estimate the ATE of receiving two years of instruction in a small class as opposed to a regular class.

Identifiability of the instrumental variable model with the treatment and outcome missing not at random

Fan Yang(杨帆)
Tsinghua University

The instrumental variable model of Imbens and Angrist (1994) and Angrist et al. (1996) allow for the identification of the local average treatment effect, also known as the complier average causal effect. However, many empirical studies are challenged by the missingness in the treatment and outcome. Generally, the complier average causal effect is not identifiable without further assumptions when the treatment and outcome are missing not at random. We study its identifiability even when the treatment and outcome are missing not at random. We review the existing results and provide new findings to unify the identification analysis in the literature.

Discovery and inference of possibly bi-directional causal relationships with invalid instrumental variables

Wei Li(李伟)
Renmin University

Learning causal relationships between pairs of complex traits from observational studies is of great interest across various scientific domains. However, most existing methods assume the absence of unmeasured confounding and restrict causal relationships between two traits to be uni-directional, which may be violated in real-world systems. In this paper, we address the challenge of causal discovery and effect inference for two traits while accounting for unmeasured confounding and potential feedback loops. By leveraging possibly invalid instrumental variables, we provide identification conditions for causal parameters in a model that allows for bi-directional relationships, and we also establish identifiability of the causal direction under the introduced conditions. Then we propose a data-driven procedure to detect the causal direction and provide inference results about causal effects along the identified direction. We show that our method consistently recovers the true direction and produces valid confidence intervals for the causal effect. We conduct extensive simulation studies to show that our proposal outperforms existing methods. We finally apply our method to analyze real data sets from UK Biobank.

Chair:Fan Yang(杨帆)

Two-stage least squares with treatment effect heterogeneity

Fan Li(李凡 女)
Duke University

Treatment effect heterogeneity with respect to covariates is common in IV studies. The interacted 2SLS approach postulates a linear working model of the outcome on the treatment, covariates, and treatment-covariate interactions, and instrument it by the IV, covariates, and IV-covariate interactions. Theoretical properties of the interacted 2sls have not been fully established except when the IV is randomly assigned. We fill this gap and clarify the causal interpretation of the interacted 2sls under the local average treatment effect (LATE) framework when the IV is only valid conditional on covariates. Our contributions are twofold. First, we show that the interacted 2sls is consistent for estimating the systematic treatment effect variation among compilers if any of the following conditions holds but may be inconsistent otherwise: (i) the covariates are categorical; (ii) the IV is valid even without conditioning on the covariates; (iii) the linear outcome model underlying the interacted 2sls is correct. Second, we establish sufficient conditions for using the interacted 2sls to estimate the LATE. We show that the interacted 2SLS, with its interacted second stage, directly recovers LATE, whereas the 2SLS with an additive second stage only recovers a weighted average of the conditional LATEs. Therefore, we recommend the interacted 2SLS as the primary method to estimate LATE in presence of treatment effect heterogeneity.

This is a joint work with Anqi Zhao and Peng Ding.

Local Causal Discovery with Background Knowledge

Yue Liu(刘越)
Renmin University

Causality plays a pivotal role in various fields of study. Based on the framework of causal graphical models, previous works have proposed identifying whether a variable is a cause or non-cause of a target in every Markov equivalent graph by learning only the local structure. However, the presence of prior knowledge, often represented as a partially known causal graph, is common in many causal modeling applications. Leveraging this prior knowledge enables further identification of causal relations. In this paper, we first propose a method for learning the local structure using various types of causal background knowledge, including direct causal information, non-ancestral information and ancestral information. Then we introduce criteria for identifying causal relations based solely on the local structure in the presence of prior knowledge. We also apply our method to fair machine learning, and experiments on local structure learning, causal relation identification, and fair machine learning demonstrate that our method is both effective and efficient.

Cluster-robust inference with a single treated cluster using the t-test

Xinran Li(李欣然)
University of Chicago

We consider the situation where treatment is assigned at the cluster level and unobserved dependencies exist among units within each cluster. This situation often occurs in difference-in-means estimation, where a single treated cluster is compared to a finite number of control clusters. We assume the availability of asymptotically Gaussian cluster-level estimators, albeit with asymptotic variances that are unknown and challenging to estimate due to dependencies within clusters. Inference for treatment effects in this context is equivalent to a two-sample testing problem, where (i) one group comprises a single observation while the other includes a finite number of observations with a shared mean, and (ii) all observations follow independent Gaussian distributions with potentially heteroskedastic and unknown variances. We propose exact t-tests tailored to this problem, incorporating constraints on variance heterogeneity across groups. We illustrate the advantage of the proposed method through both simulations and empirical applications.

Chair:Fan Li(李凡 男)

Causal inference and calibration for Within-Family Mendelian Randomization

Jingshu Wang(王静姝)
University of Chicago

Understanding the causal mechanisms underlying diseases is essential for advancing clinical research. When randomized controlled trials are unfeasible, Mendelian Randomization (MR) can serve as an alternative, using genetic variants as natural "experiments" to help control for environmental confounding. However, while Mendel's law ensures random inheritance, it does not rule out the confounding parental genotype effects that may bias the MR conclusions. In this talk, I present and justify a simple linear approach to correct for parental genotype confounding by using data from trios or sibling pairs in genome-wide association studies (GWAS). In addition, to improve efficiency, I introduce a user-friendly calibration method that uses only summary statistics from both large-scale population-based GWAS and smaller family-based GWAS datasets. Our theoretical and empirical findings indicate that the calibrated estimators can achieve roughly a 50% reduction in variance compared to using trio-based GWAS alone, and a 10%–20% reduction compared to using sibling-based GWAS alone, with gains depending on the phenotypic correlation among siblings.

Debiased regression adjustment in completely randomized experiments with moderately high-dimensional covariates

Yuhao Wang(王禹皓)
Tsinghua University

Completely randomized experiment is the gold standard for causal inference. When the covariate information for each experimental candidate is available, one typical way is to include them in covariate adjustments for more accurate treatment effect estimation. In this paper, we investigate this problem under the randomization-based framework, i.e., that the covariates and potential outcomes of all experimental candidates are assumed as deterministic quantities and the randomness comes solely from the treatment assignment mechanism. Under this framework, to achieve asymptotically valid inference, existing estimators usually require either (i) that the dimension of covariates grows at a rate no faster than O(n^{3/4}) as sample size n goes to infinity; or (ii) certain sparsity constraints on the linear representations of potential outcomes constructed via possibly high-dimensional covariates. In this paper, we consider the moderately high-dimensional regime where p is allowed to be in the same order of magnitude as n. We develop a novel debiased estimator with a corresponding inference procedure and establish its asymptotic normality under mild assumptions. Our estimator is model-free and does not require any sparsity constraint on potential outcome's linear representations. We also discuss its asymptotic efficiency improvements over the unadjusted treatment effect estimator under different dimensionality constraints. Numerical analysis confirms that compared to other regression adjustment based treatment effect estimators, our debiased estimator performs well in moderately high dimensions.

Extreme-based causal effect learning (EXCEL) with unmeasured light-tailed confounding

Wang Miao(苗旺)
Peking University

Unmeasured confounding poses a significant challenge in identifying and estimating causal effects across various research domains. Existing methods to address confounding often rely on either parametric models or auxiliary variables, which strongly rest on domain knowledge and could be fairly restrictive in practice. In this paper, we propose a novel strategy for identifying causal effects in the presence of confounding under an additive structural equation with light-tailed confounding. This strategy uncovers the causal effect by exploring the relationship between the exposure and outcome at the extreme, which can bypass the need for parametric assumptions and auxiliary variables. The resulting identification is versatile, accommodating a multi-dimensional exposure, and applicable in scenarios involving unmeasured confounders, selection bias, or measurement errors. Building on this identification approach, we develop an Extreme-based Causal Effect Learning (EXCEL) method and further establish its consistency and non-asymptotic error bound. The asymptotic normality of the proposed estimator is established under the linear model. The EXCEL method is applied to causal inference problems with invalid instruments to construct a valid confidence set for the causal effect. Simulations and a real data analysis are used to illustrate the potential application of our method in causal inference.

Chair:Hanzhong Liu(刘汉中)

Method-of-Moments Inference for GLMs and Doubly Robust Functionals under Proportional Asymptotics

Rajarshi Mukherjee
Harvard University

In this paper, we consider the estimation of regression coefficients and signal-to-noise (SNR) ratio in high-dimensional Generalized Linear Models (GLMs), and explore their implications in inferring popular estimands such as average treatment effects in high-dimensional observational studies.

Under the ``proportional asymptotic'' regime and Gaussian covariates with known (population) covariance Σ, we derive Consistent and Asymptotically Normal (CAN) estimators of our targets of inference through a Method-of-Moments type of estimators that bypasses estimation of high dimensional nuisance functions and hyperparameter tuning altogether. Additionally, under non-Gaussian covariates, we demonstrate universality of our results under certain additional assumptions on the regression coefficients and Σ. We also demonstrate that knowing Σ is not essential to our proposed methodology when the sample covariance matrix estimator is invertible. Finally, we complement our theoretical results with numerical experiments and comparisons with existing literature.

How to achieve model-robust inference in stepped wedge trials with model-based methods?

Fan Li(李凡 男)
Yale University

A stepped wedge design is a unidirectional crossover design where clusters are randomized to distinct treatment sequences. While model-based analysis of stepped wedge designs is standard practice to evaluate treatment effects accounting for clustering and adjusting for covariates, their properties under misspecification have not been systematically explored. In this article, we focus on model-based methods, including linear mixed models and generalized estimating equations with an independence, simple exchangeable, or nested exchangeable working correlation structure. We study when a potentially misspecified working model can offer consistent estimation of the marginal treatment effect estimands, which are defined nonparametrically with potential outcomes and may be functions of calendar time and/or exposure time. We prove a central result that consistency for nonparametric estimands usually requires a correctly specified treatment effect structure, but generally not the remaining aspects of the working model (functional form of covariates, random effects, and error distribution), and valid inference is obtained via the sandwich variance estimator. Furthermore, an additional g-computation step is required to achieve model-robust inference under non-identity link functions or for ratio estimands. The theoretical results are illustrated via several simulation experiments and re-analysis of a completed stepped wedge cluster randomized trial.

Covariate adjustment in RCT motivated by HOIF

Lin Liu(刘林)
Shanghai Jiaotong University

Higher-Order Influence Functions (HOIF), developed in a series of papers over the past twenty years, serve as a fundamental theoretical framework for constructing rate-optimal causal-effect estimators from observational studies. However, the utility of HOIF for analyzing well-conducted randomized controlled trials (RCTs) has not been explicitly explored. Recent guidelines from the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) on covariate adjustment in RCT analysis recommend not only the simple, unadjusted difference-in-mean estimator but also an estimator adjusting for baseline covariates via a simple parametric working model, such as a linear model. In this paper, we demonstrate that a HOIF-motivated estimator for the treatment-specific mean exhibits significantly improved statistical properties compared to commonly used adjusted estimators, particularly when the number of baseline covariates p is

relatively large compared to the sample size n. We also characterize the conditions under which the HOIF-motivated estimator improves upon the unadjusted one. Furthermore, we demonstrate that a novel debiased adjusted estimator recently proposed by Lu et al. is, in fact, another HOIF-motivated estimator in disguise. Numerical and empirical studies are conducted to corroborate our theoretical findings.

Chair:Yunan Wu (吴宇楠)

Distribution-Type Index For Causal Inference*

Huazhen Lin(林华珍)
Southwestern University of Finance and Economics

The evaluation of causal effects is receiving increasing attention across various fields, primarily driven by concerns about covariate imbalances when comparing two groups. In this paper, we propose a novel metric, the Distribution-type Index (D-index), as a framework to address covariate imbalances and offer a comprehensive characterization of treatment effects. Particularly, the proposed D-index includes traditional measures like average treatment effects (ATE), Mann–Whitney statistics, and quantile treatment effects as special cases. However, unlike ATE or quantile treatment effects, which focus on population-level differences, the D-index evaluates outcome-level differences. Furthermore, to estimate the D-index in the presence of covariate imbalances, we then augment two groups to be identical in all aspects except for their treatment assignments, and estimate the D-index and its variation by a probabilistic model with varying-coefficient single-index structure. This approach combines the efficiency of the model-based approach and the robustness of the nonparametric approach and allows us to extract information as sufficient as possible. As a result, the proposed method is shown to have superior efficiency compared to existing efficient methods, and is robustness to the specification of propensity score and covariate imbalances between the two groups. We demonstrate the superior effectiveness of the proposed D-index and its derived indices through extensive numerical examples, including simulation studies and real-world data analyses.

*Joint work with Li Liu, Guizhen Li and Ling Zhou

Causal Inference with Instrumental Variables in Complex Scenarios

Anpeng Wu(吴安鹏)
Zhejiang University

In causal inference, instrumental variable (IV) regression is a classical method used to estimate causal effects in the presence of unobserved confounders. By introducing a variable (the instrument) that affects the treatment but not directly the outcome (other than through the treatment), the IV method helps identify the causal relationship. However, in complex scenarios, some studies overlook residual confounding effects from observed covariates during the two-stage regression. To address this, we propose an Instrumental Variable Regression with Confounder Balancing algorithm, to jointly correct biases from unmeasured confounders and achieve balance among observed confounders.

Estimation and inference of average treatment effects under heterogeneous additive treatment effect model

Hanzhong Liu(刘汉中)
Tsinghua University

Randomized experiments are the gold standard for estimating treatment effects, yet network interference challenges the validity of traditional estimators by violating the stable unit treatment value assumption and introducing bias. While cluster randomized experiments mitigate this bias, they encounter limitations in handling network complexity and fail to distinguish between direct and indirect effects. To address these challenges, we develop a design-based asymptotic theory for the existing Horvitz--Thompson estimators of the direct, indirect, and global average treatment effects under Bernoulli trials. We assume the heterogeneous additive treatment effect model with a hidden network that drives interference. Observing that these estimators are inconsistent in dense networks, we introduce novel eigenvector-based regression adjustment estimators to ensure consistency. We establish the asymptotic normality of the proposed estimators and provide conservative variance estimators under the design-based inference framework, offering robust conclusions independent of the underlying stochastic processes of the network and model parameters. Our method's adaptability is demonstrated across various interference structures, including partial interference and local interference in a two-sided marketplace. Numerical studies further illustrate the efficacy of the proposed estimators, offering practical insights into handling network interference.

Formulating generalized causal models from the no-signalling physical principle

Xingjian Zhang(张行健)
University of Hong Kong

Determining potential probability distributions with a given causal graph is vital for causality studies. The problem is generally challenging due to the difficulty in characterizing latent variables. Additionally, as demonstrated by Bell inequalities, the set of plausible distributions depends on the physical nature of these latent issues. Notably, quantum physics can generate correlations not implementable in classical Bayesian networks. It is thus intriguing to consider the following question: How should we define a proper causal model that accommodates the physical rules of Nature, which might even extend beyond our current physical understandings? In this talk, I shall discuss a new causal model that treats every causal structure as a low-dimensional projection from a high-dimensional Bell-test-type structure. The new model restricts distributions merely by the independence conditions in the high-dimensional structures, or equivalently the no-signalling physical principle, which is general enough to comprise all physical theories subjected to relativity principles. I shall show how this model replicates the equality constraints in the nested Markov model, including both conditional independences and Verma-type constraints. Nevertheless, I shall also discuss how these causal models differ in inequality constraints and discuss this issue from a physical viewpoint.

Chair: Guanglei Hong (洪光磊)

Factorial Difference-in-Differences

Peng Ding(丁鹏)
University of California, Berkeley

In many social science applications, researchers use the difference-in-differences (DID) estimator to establish causal relationships, exploiting cross-sectional variation in a baseline factor and temporal variation in exposure to an event that presumably may affect all units. This approach, which we term factorial DID (FDID), differs from canonical DID in that it lacks a clean control group unexposed to the event after the event occurs. In this paper, we clarify FDID as a research design in terms of its data structure, feasible estimands, and identifying assumptions that allow the DID estimator to recover these estimands. We frame FDID as a factorial design with two factors: the baseline factor, denoted by G, and the exposure level to the event, denoted by Z, and define the effect modification and causal interaction as the associative and causal effects of G on the effect of Z, respectively. We show that under the canonical no anticipation and parallel trends assumptions, the DID estimator identifies only the effect modification of G in FDID, and propose an additional factorial parallel trends assumption to identify the causal interaction. Moreover, we show that the canonical DID research design can be reframed as a special case of the FDID research design with an additional exclusion restriction assumption, thereby reconciling the two approaches. We extend this framework to allow conditionally valid parallel trends assumptions and multiple time periods, and clarify assumptions required to justify regression analysis under FDID. We illustrate these findings with empirical examples from economics and political science, and provide recommendations for improving practice and interpretation under FDID.

Statistical Inference for Mediation Models with High Dimensional Exposures and Mediators

Jingyuan Liu (刘婧媛)
Xiamen University

High-dimensional mediation analysis has gained increasing interest across various fields, particularly in genetic and medical research. Compared with existing works that mainly focus on high-dimensional mediators, this paper proposes new statistical inference methods when both exposures and mediators are high-dimensional under a partial multiple testing framework. We propose estimation procedures using a partial penalization approach, incorporating a double-layer latent factor structure, and establish theoretical properties of the estimators. F-type and Wald test procedures for the high-dimensional direct and indirect effects, respectively, are advocated based on the proposed estimators. Monte Carlo simulations are implemented to assess the finite sample performance of the new tests and to compare their efficacy against some existing methods. Finally, we apply our statistical framework to investigate direct effects of genetic variants on Alzheimer's disease (AD) and indirect effects of them mediated through changes in brain activity intensity.

On the Role of Surrogates in the Efficient Estimation of Treatment Effects with Limited Outcome Data

Xiaojie Mao (毛小介)
Tsinghua University

In many experimental and observational studies, the outcome of interest is often difficult or expensive to observe, reducing effective sample sizes for estimating average treatment effects (ATEs) even when identifiable. We study how incorporating data on units for which only surrogate outcomes not of primary interest are observed can increase the precision of ATE estimation. We refrain from imposing stringent surrogacy conditions, which permit surrogates as perfect replacements for the target outcome. Instead, we supplement the available, albeit limited, observations of the target outcome with abundant observations of surrogate outcomes, without any assumptions beyond unconfounded treatment assignment and missingness and corresponding overlap conditions. To quantify the potential gains, we derive the difference in efficiency bounds on ATE estimation with and without surrogates, both when an overwhelming or comparable number of units have missing outcomes. We develop robust ATE estimation and inference methods that realize these efficiency gains. We empirically demonstrate the gains by studying long-term-earning effects of job training.

Chair:Huazhen Lin (林华珍)

Design-based theory for Lasso adjustment in randomized block experiments and rerandomized experiments

Yuehan Yang(杨玥含)
Central University of Finance and Economics

Blocking, a special case of rerandomization, is routinely implemented in the design stage of randomized experiments to balance the baseline covariates. This study proposes a regression adjustment method based on the least absolute shrinkage and selection operator (Lasso) to efficiently estimate the average treatment effect in randomized block experiments with high-dimensional covariates. We derive the asymptotic properties of the proposed estimator and outline the conditions under which this estimator is more efficient than the unadjusted one. We provide a conservative variance estimator to facilitate valid inferences. Our framework allows one treated or control unit in some blocks and heterogeneous propensity scores across blocks, thus including paired experiments and finely stratified experiments as special cases. We further accommodate rerandomized experiments and a combination of blocking and rerandomization. Moreover, our analysis allows both the number of blocks and block sizes to tend to infinity, as well as heterogeneous treatment effects across blocks without assuming a true outcome data-generating model. Simulation studies and two real-data analyses demonstrate the advantages of the proposed method.

Moving towards best practice when using propensity score weighting in survey observational studies

Guangyu Tong (同广宇)
Yale University

Propensity score weighting is a common method for estimating treatment effects with survey data. The method is applied to minimize confounding using measured covariates that are often different between individuals in treatment and control. However, existing literature does not reach a consensus on the optimal use of survey weights for population-level inference in the propensity score weighting analysis. Under the balancing weights framework, we provided a unified solution for incorporating survey weights in both the propensity score of estimation and the outcome regression model. We derived estimators for different target populations, including the combined, treated, controlled, and overlap populations. We also proved the asymptotic normality of the survey-weighted estimator and used M-estimation to derive the sandwich variance and augmented estimators. Through an extensive series of simulation studies, we examined the performance of our derived estimators and compared the results to those of alternative methods. We further carried out a case study on the Early Childhood Longitudinal Study to illustrate the application of the different methods of propensity score analysis with complex survey data. We concluded with a thorough discussion of our findings and provided practical guidelines for the use of survey weights in propensity score weighting analysis using data from complex surveys.

Balancing Utility and Cost in Dynamic Treatment Regimes

Yuqian Zhang (张宇谦)
Renmin University

Dynamic treatment regimes (DTRs) refer to personalized, adaptive treatment strategies designed to guide the sequential allocation of treatments for individuals over time. At each stage, individual characteristics are collected to facilitate more precise adjustments to treatment plans. However, in practice, assigning treatment as well as collecting features that accurately reflect an individual's state often incurs higher costs. In this work, we propose a new strategy for estimating DTRs that accounts for the costs associated with feature collection over time. The performance of the proposed method is demonstrated through extensive numerical studies.

Chair:Peng Ding (丁鹏)

Biostatistics, data and cohort study

Min Zhang (张敏)
Tsinghua University

In this talk, I will share my 1.5 years of experience as a biostatistician in China after working in the US for 20 years. I will compare the differences between the two countries in terms of research and training in biostatistics, and data availability. Additionally, I will discuss cohort studies in China and introduce a cohort study we are in the process of developing.

An exhaustive selection of sufficient adjustment sets for causal inference

Wei Luo (骆威)
Zhejiang University

A subvector of predictor that satisfies the ignorability assumption, whose index set is called a sufficient adjustment set, is crucial for conducting reliable causal inference based on observational data. In this paper, we propose a general family of methods to detect all such sets for the first time in the literature, with no parametric assumptions on the outcome models and with flexible parametric and semiparametric assumptions on the predictor within the treatment groups; the latter induces desired sample-level accuracy. We show that the collection of sufficient adjustment sets can uniquely facilitate multiple types of studies in causal inference, including sharpening the estimation of average causal effect and recovering fundamental connections between the outcome and the treatment hidden in the dependence structure of the predictor. These findings are illustrated by simulation studies and a real data example at the end.

Matching-based Policy Learning

Ying Yan (严颖)
Sun Yat-Sen University

Treatment heterogeneity is ubiquitous in many observational studies, which motivates practitioners to search for the optimal policy that maximizes the expected outcome based on individualized characteristics. A large body of policy learning methods rely on inverse weighting, which may suffer from instability problems. To enhance the robustness of the estimated policy, we propose a matching-based estimator of the policy improvement upon a randomized baseline. After correcting the conditional bias, we learn the optimal policy by maximizing the estimate over a policy class. We derive a non-asymptotic high probability bound for the regret of the learned policy and show that the convergence rate is almost root n. The competitive finite sample performance of the proposed method is demonstrated in extensive simulation studies and a real data application.

Model-Assisted Uniformly Honest Inference for Optimal Treatment Regimes in High Dimension

Yunan Wu(吴宇楠)
Tsinghua University

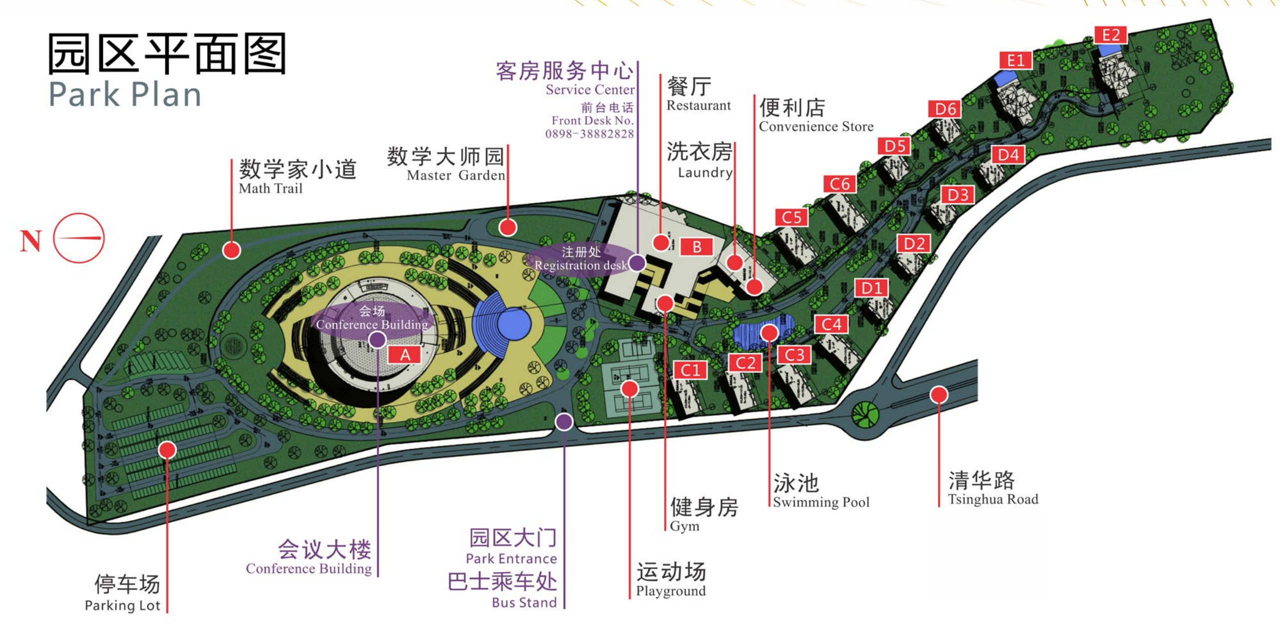
We develop new tools to quantify uncertainty in optimal decision making and to gain insight into which variables one should collect information about given the potential cost of measuring a large number of variables. We investigate simultaneous inference to determine if a group of variables is relevant for estimating an optimal decision rule in a high-dimensional semiparametric framework. The unknown link function permits flexible modeling of the interactions between the treatment and the covariates, but leads to nonconvex estimation in high dimension and imposes significant challenges for inference. We first establish that a local restricted strong convexity condition holds with high probability and that any feasible local sparse solution of the estimation problem can achieve the near-oracle estimation error bound. We further rigorously verify that a wild bootstrap procedure based on a debiased version of the local solution can provide asymptotically honest

uniform inference for the effect of a group of variables on optimal decision making. The advantage of honest inference is that it does not require the initial estimator to achieve perfect model selection and does not require the zero and nonzero effects to be well-separated. We also propose an efficient algorithm for estimation. Our simulations suggest satisfactory performance. An example from a diabetes study illustrates the real application.

The facilities of TSIMF are built on a 23-acre land surrounded by pristine environment at Phoenix Hill of Phoenix Township. The total square footage of all the facilities is over 29,000 square meter that includes state-of-the-art conference facilities (over 10,000 square meter) to hold many international workshops simultaneously, two reading rooms of library, a guest house (over 10,000 square meter) and the associated catering facilities, a large swimming pool, gym and sports court and other recreational facilities.

Management Center of Tsinghua Sanya International Forum is responsible for the construction, operation, management and service of TSIMF. The mission of TSIMF is to become a base for scientific innovations, and for nurturing of innovative human resource; through the interaction between leading mathematicians and core research groups in pure mathematics, applied mathematics, statistics, theoretical physics, applied physics, theoretical biology and other relating disciplines, TSIMF will provide a platform for exploring new directions, developing new methods, nurturing mathematical talents, and working to raise the level of mathematical research in China.

About Facilities



Registration

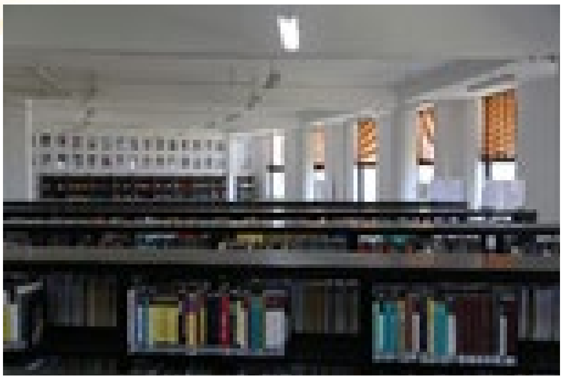
Conference booklets, room keys and name badges for all participants will be distributed at the front desk. Please take good care of your name badge. It is also your meal card and entrance ticket for all events.

Guest Room

All the rooms are equipped with: free Wi-Fi (Password:tsimf123), TV, air conditioning and other utilities.

Family rooms are also equipped with kitchen and refrigerator.

Library



Opening Hours: 09:00am-22:00pm

TSIMF library is available during the conference and can be accessed by using your room card. There is no need to sign out books but we ask that you kindly return any borrowed books to the book cart in library before your departure.

In order to give readers a better understanding of the contributions made by the Fields Medalists, the library of Tsinghua Sanya International Mathematics Forum (TSIMF) instituted the Special Collection of Fields Medalists as permanent collection of the library to serve the mathematical researchers and readers.

So far, there are 271 books from 49 authors in the Special Collection of Fields Medalists of TSIMF library. They are on display in room A220. The participants are welcome to visit.

Restaurant

All the meals are provided in the restaurant (Building B1) according to the time schedule.

Breakfast 07:30-08:45
Lunch 12:00-13:30
Dinner 17:30-19:00

Laundry

Opening Hours: 24 hours
The self-service laundry room is located in the Building(B1).

Gym

Opening Hours: 24 hours
The gym is located in the Building 1 (B1), opposite to the reception hall. The gym provides various fitness equipment, as well as pool tables, tennis tables etc.

Playground

Playground is located on the east of the central gate. There you can play basketball, tennis and badminton. Meanwhile, you can borrow table tennis, basketball, tennis balls and badminton at the reception desk.

Swimming Pool

Please enter the pool during the open hours, swimming attire and swim caps are required, if you feel unwell while swimming, please stop swimming immediately and get out of the pool. The depth of the pool is 1.2M-1.8M.
Opening Hours: 13:00-14:00 18:00-21:00

Free Shuttle Bus Service at TSIMF

We provide free shuttle bus for participants and you are always welcome to take our shuttle bus, all you need to do is wave your hands to stop the bus.

Destinations: Conference Building, Reception Room, Restaurant, Swimming Pool, Hotel etc.

Contact Information of Administration Staff

Location of Conference Affairs Office: Room 104, Building A
Tel: 0086-898-38263896
Conference Affairs: Shouxi He 何守喜
Tel:0086-186-8980-2225
Email: heshouxi@tsinghua.edu.cn

Location of Accommodation Affairs Office: Room 200, Building B1
Tel: 0086-898-38882828
Accommodation Manager: Ms. Li YE 叶莉
Tel: 0086-139-7679-8300
Email: yel@tsinghua.edu.cn

IT
Yuanhang Zhou 周远航
Tel: 0086-133-6898-0169
Email: 13368980169@163.com

*Reception duty hours: 7:00-23:00, chamber service please call: 0086-38882828 (exterior line) 80000 (internal line)
*Room maintainer night duty hours: 23:00-7:00, if you need maintenance services, please call: 0086-38263909 (exterior line) 30162 (internal line)

Director Assistant of TSIMF
Kai CUI 崔凯
Tel/Wechat: 0086- 136-1120-7077
Email :cuik@tsinghua.edu.cn

Director of TSIMF
Prof.Xuan GAO 高瑄
Tel: 0086-186-0893-0631
Email: gaoxuan@tsinghua.edu.cn

清华大学三亚国际论坛管理中心 | 三亚清华数学论坛管理中心
Tsinghua Sanya International Mathematics Forum (TSIMF)

0086-898-38882828
0086-898-38883896
FAX 0086-898-38883895
tsimf@tsinghua.edu.cn
572000
https://www.tsimf.cn
海南省三亚市天涯区清华路 100 号 , 清华三亚国际数学论坛
No.100, Tsinghua Road, Tianya District, Sanya, Hainan, P. R. China.