2023 AMERICAN CAUSAL INFERENCE CONFERENCE

ABSTRACT BOOK

ORGANIZED BY:
THE SOCIETY FOR CAUSAL INFERENCE

MAY 24-26, 2023

2023 ACIC SPONSORS

ALFRED P. SLOAN FOUNDATION

Microsoft

The University of Texas at Austin
Department of Statistics and Data Sciences

N

sas®

parexel®

Heterogeneous Treatment Effects

Weighted quantile treatment effect estimation in the presence of informatively missing data via double-sampling in EHR-based research Shuo Sun* Shuo Sun, Rajarshi Mukherjee, Sebastien Haneuse,

While electronic health records (EHR) data provide opportunities for cost-effective clinical and public health research, missing data is a widespread problem. When the data are informatively missing or missing not at random with respect to measured covariates, double-sampling is a means to augment the EHR with additional information collected on a sub-sample. Additionally, in some research areas involving low-probability, high-impact events being the outcomes of interest, the primary interest is estimating the causal effects in the tails of the outcome distribution (rather than the mean). Based on these observations, the goal of this work is to propose general distributional estimators for causal weighted quantile treatment effects (WQTE) in the presence of missing outcomes. Specifically, we propose three novel estimators of WQTEs, and corresponding identifiability conditions, each tailored to whether the missing at random assumptions holds and on whether double-sampled data is available. Asymptotic properties are derived, with finite sample performance assessed via simulation. We demonstrate the proposed methods by a EHR study of long-term outcomes following bariatric surgery.

Machine Learning and Causal Inference

Leveraging text data for causal inference using electronic health records Luke Miratrix*
Aaron Kaufman, Reagan Mozer, Leo Celi,

Text is a ubiquitous component of medical data, containing valuable information about patient characteristics and care that are often missing from structured chart data. Despite this richness, it is rarely used in clinical research, owing partly to its complexity.

Using a large database of patient records and treatment histories accompanied by extensive notes by attendant physicians and nurses, we show how text data improve research in all stages, from conception and design to analysis and interpretation, with minimal additional effort.

We focus on studies using matching for causal inference.

We augment a classic matching analysis by incorporating text in three ways: by using text to supplement a multiple imputation procedure, we improve the fidelity of imputed values to handle missing data; by incorporating text in the matching stage, we improve covariate balance; and by conditioning on text, we can estimate easily interpretable text-based heterogeneous treatment effects that may be stronger than those found across categories of structured covariates.

We introduce software to implement our procedures to ease their incorporation into existing workflows.

Using these techniques, we hope to expand the scope of secondary analysis of clinical data to domains where quantitative data is of poor quality or nonexistent but text is available, such as in developing countries.

Machine Learning and Causal Inference

Debiasing in missing data models with inaccurate estimates of outcome and missingness parameters Michael Celentano* Michael Celentano,

We consider the problem of (i) estimating linear model coefficients with data missing at random (MAR) and (ii) average treatment effect estimation with linear outcome models under strong ignorability. We study these problems in a high-dimensional regime in which the number of confounders p is proportional to the sample size n and the outcome and propensity/missingness models cannot be estimated consistently. A series of recent works (Jiang et al., 2022; Yadlowsky 2022) studied the behavior of the classical IPW, AIPW, and TMLE estimators in this regime and revealed several departures from the predictions of the classical theory, including, for example, a variance inflation of the AIPW estimator. Their analyses, however, are restricted to cases in which $n > p$, which allows for unbiased estimation of the outcome model. In this paper, we study instead the case that $n < p$ and regularization is used in estimating the outcome model and propensity/missingness models. In this case, the classical estimators of linear coefficients and average treatment effects fail to be unbiased or even consistent. We propose a debiased estimator that is provably consistent and provide confidence intervals for the estimated linear coefficients. As with classical AIPW estimator, our proposed estimator requires estimation of both the outcome and propensity/missingness models, but we combine these estimates in a non-standard way.

Heterogeneous Treatment Effects

Adaptive Experiments Toward Learning Treatment Effect Heterogeneity Waverly Wei*

Waverly Wei, Jingshen Wang,

Understanding treatment effect heterogeneity has become an increasingly popular task in various fields. For example, in e-commerce, understanding treatment effect heterogeneity helps decision-makers to design personalized advertising strategies to maximize profits. In biomedical studies, learning the impact of treatment on diverse patient subpopulations provides insights for personalized care. While much of the existing work in this research area has focused on either analyzing observational data based on untestable causal assumptions or conducting post hoc analyses of existing randomized controlled trial data, little work has gone into designing randomized experiments specifically for uncovering treatment effect heterogeneity. In this work, we develop a unified adaptive experimental design framework towards better learning treatment effect heterogeneity by efficiently identifying subpopulations with enhanced treatment effects. The adaptive nature of our framework allows practitioners to sequentially allocate experimental efforts adapting to the accrued evidence during the experiment. The resulting design framework can not only complement A/B tests in e-commerce but also unify enrichment designs and response adaptive randomization designs in clinical settings. Our theoretical investigations illustrate the trade-offs between complete randomization and our adaptive experimental algorithms. We also investigate our design in simulation studies and e-commerce data analysis.

Multilevel Causal Inference

Designing Optimal, Data-Driven Educational Policies from Multisite Randomized Trials

Youmi Suk* Youmi Suk, Chan Park,

Optimal treatment regimes (OTRs) have been popular in computer science and personalized medicine to provide data-driven, optimal recommendation rules to individuals. Unfortunately, much of the methodological work on OTRs focuses on single-site settings and there is little work on tailoring existing OTRs to educational settings where students i.e. the study unit, are nested within schools and there are hierarchical dependencies. The goal of this paper is to design OTRs from multisite randomized trials, which are frequently used in education to evaluate educational programs. We study how to modify popular OTR methods, notably Q-learning and weighting methods, in order to enhance their performances in multisite randomized trials. We consider a total of 12 modifications, 6 modifications for Q-learning and 6 modifications for weighting, based on using different multilevel models, moderators, or augmentations. Through simulation studies, we find that all Q-learning modifications enhance the performance in multisite randomized trials, but the modifications with random treatment effects show the most promise in handling cluster-level moderators. Among weighting methods, the modification that adds cluster dummies into moderator variables and augmentation terms performs the best across simulation conditions. We also demonstrate our proposals using data from a multisite randomized trial in Colombia on evaluating conditional cash transfer programs to maximize educational attainment.

Randomization-based inference

Counternull sets for randomized experiments Marie-Abele Bind* Marie-Abele Bind, Donald Rubin,

In 1990, the clinical psychologist Paul E. Meehl used the term ”counternull” to describe an unspecified nonnull alternative statistical hypothesis. Four years later, Rosenthal and Rubin introduced the less vague ”counternull value of an effect size” (1994) as ”the nonnull magnitude of effect size that is supported by exactly the same amount of evidence as is the null value of the effect size”. Continuing, evidence was explicitly illustrated using the ratio between the likelihoods of the test statistic at the maximum likelihood estimate of the estimand and at the value of that statistic at the null value of the estimand. This explication is well-defined for parametric models and in situations where asymptotic normality is used as the basis for inference, both of which are common in practice. However, in the context of randomized experiments, an arguably more natural definition of evidence is available, and this definition has the advantage of being both model-free and well-defined sub-asymptotically: a counternull value is a nonnull value that yields the same randomization-based p-value as does the null value. From this perspective, the counternull is rarely a unique scalar but rather a set of values. We illustrate this use and its potential attendant new insights using first a social science randomized experiment with 2,216 units, and then an epigenetic randomized experiment with 17 units.

Causal Inference and SUTVA/Consistencies Violations

Causal inference for environmental health data: estimating causal effects in the presence of spatial interference Nathan Wikle* Nathan Wikle, Cory Zigler,

Causal inference for environmental health data is often challenging due to the presence of spatial interference: outcomes for observational units depend on some combination of local and nonlocal treatment. One common solution includes the specification of an exposure model, in which treatment assignments are mapped to an exposure value; causal estimands of the local and spillover effects of treatment are defined through contrasts of the local treatment assignment and the exposure value. Notably, the exposure model is often defined via a network structure, which is assumed to be fixed and known a priori. However, in environmental settings, spatial interference may be dictated by complex, mechanistic processes that are both stochastic and poorly represented by a network. In this work, we develop methods for causal inference with interference when deterministic exposure models cannot be assumed or are unknown. We offer a Bayesian model for the interference structure which, when combined with a flexible nonparametric outcome model, allows us to marginalize estimates of causal effects over uncertainty in the interference structure. The interference structure is estimated from environmental data using a mechanistic model for spatial data. To illustrate our methodology, we analyze the effectiveness of air quality interventions in reducing two adverse health outcomes in Texas — asthma ED visits and Medicare all-cause mortality.

Causal Inference in Networks

Causal inference with uncertain interference structure Daniel Nevo* Daniel Nevo, Bar Weinstein,

As an alternative to the no-interference assumption, an interference structure is often represented using a network. Ubiquitously, the network structure is assumed to be known and correctly specified. However, correctly encoding the interference structure in a network can be challenging. For example, edges might be measured with error or censored, network structure can change over time, and contamination between clusters might be present. Using the exposure-function framework, we quantify the bias of commonly used estimators when the network interference structure is misspecified.

To overcome the problem of network misspecification, we propose two solutions. First, we propose a novel estimator utilizing multiple networks simultaneously, which is unbiased if one of the networks correctly represents the interference structure. As an alternative, we propose a sensitivity analysis framework that quantifies the impact of a postulated interference structure misspecification on the causal estimate as a function of parameters governing a misspecification mechanism.

We illustrate the bias arising from incorrectly specified network and study the bias-variance tradeoff entailed in our proposed misspecification-robust estimator. We demonstrate the utility of our methods in two real examples involving two different interference structures: a social network field experiment and a cluster-randomized trial.

Causal Inference in Networks

Regression Discontinuity Designs Under Interference Elena Dal Torrione* Elena Dal Torrione, Tiziano Arduini, Laura Forastiere,

Interference takes place whenever a “treatment” on one unit affects the outcome of another unit, and such a phenomenon can occur in regression discontinuity designs (RDD). For instance, in conditional cash transfer programs for education, eligible children’s schooling choices may affect the schooling choices of their ineligible peers. We propose an extension of the continuity-based framework to RDD to identify and estimate a set of causal estimands in the presence of interference. In this setting, assignment to effective treatment is determined by a unit’s score and the scores of other units—for example, her neighbors. Unlike the standard RDD, embedding the exposure mapping function as a summary of other units’ treatment may give rise to complex, multidimensional frontiers. We provide a method to characterize such frontiers for a broad class of exposure mapping functions and derive generalized continuity assumptions to identify the proposed estimands. Next, we develop three estimation methods that can handle high-dimensional—and potentially heterogeneous—score spaces, and evaluate their empirical performance in a simulation study. Finally, we apply the presented methodology to the PROGRESA/Oportunidades data to estimate the spillover effects of financial aid to families on children’s school attendance.

Causal Discovery

Causal learning with unknown interventions: algorithms and guarantees Armeen Taeb*
Armeen Taeb, Juan Gamella, Peter Buehlmann, Christina Heinze-Deml, Felix Hafenmair,

With observational data alone, causal inference is a challenging problem. The task becomes easier when having access to data collected from perturbations of the underlying system, even when the nature of these is unknown. In this talk, we will describe methods that use such perturbation data to identify plausible causal mechanisms. Specifically, in the context of Gaussian linear structural equation models, we first characterize the interventional equivalence class of DAGs. We then leverage these results to study high-dimensional consistency guarantees of a l0-penalized maximum likelihood estimator for learning said class. Since solving this estimator is generally intractable, we design a procedure called GnIES which proceeds greedily in the space of interventional equivalent models. In addition, we develop a novel procedure to generate semi-synthetic data sets with known causal ground truth but distributions closely resembling those of a real data set of choice. We leverage this procedure and evaluate the performance of GnIES on synthetic, real, and semi-synthetic data sets. Despite the strong Gaussian distributional assumption, GnIES is robust to an array of model violations and competitive in recovering the causal structure in small- to large-sample settings. We provide, in the Python packages emph{gnies} and emph{sempler}, implementations of GnIES and our semi-synthetic data generation procedure.

Instrumental Variables

Doubly Robust Proximal Causal Inference under Confounded Outcome-Dependent Sampling Kendrick Li* Kendrick Li, Xu Shi, Wang Miao, Eric Tchetgen Tchetgen,

Outcome-dependent sampling is widely used in epidemiology and econometrics to reduce time and effort when studying causal relationships between the exposure and outcome variables. In these types of studies, unmeasured confounding and selection bias are often of concern and may invalidate a causal analysis if not appropriately accounted for. In particular, a latent factor that has causal effects on the treatment, outcome, and sample selection process may cause both unmeasured confounding and selection bias, rendering standard causal parameters unidentifiable without additional assumptions. In this talk, we introduce the identification and inference of treatment effect under a logistic regression model assuming a homogeneous odds ratio effect across unmeasured strata, leveraging a pair of proxies to the source of unmeasured confounding: a negative control exposure (NCE) which is a priori known not to affect the outcome and selection, and a negative control outcome (NCO) which is a priori known not to be affected by the treatment. We introduce three estimators of the odds ratio effect, one of which is doubly robust with respect to the specification of two nuisance functions which restrict the treatment assignment mechanism and outcome distribution, respectively, such that the estimator is consistent and asymptotically normal if either model is correctly specified, without knowing which one is.

Instrumental Variables

Covariate-Assisted Nonparametric Bounds of Causal Effects with Instrumental Variables

Alexander Levis* Alexander Levis, Matteo Bonvini, Zhenghao Zeng, Luke Keele, Edward Kennedy,

When the effect of an exposure of interest is confounded by unmeasured factors, an instrumental variable (IV) can be used to identify and estimate certain causal contrasts. Identification of the marginal average treatment effect (ATE) from IVs typically relies on strong untestable structural assumptions. When one is unwilling to assert such structural assumptions, IVs can nonetheless be used to construct bounds on the ATE. Famously, linear programming techniques were employed to prove tight bounds on the ATE for a binary outcome, in a randomized trial with noncompliance and no covariate information. We demonstrate how these bounds remain useful in observational settings with baseline confounders of the IV, as well as randomized trials with measured baseline covariates. The resulting lower and upper bounds on the ATE are non-smooth functionals, and thus standard nonparametric efficiency theory is not immediately applicable. To remedy this, we propose (1) estimators of smooth approximations of these bounds, and (2) under a novel margin condition, influence function-based estimators of the ATE bounds that can attain parametric convergence rates when nuisance functions are modeled flexibly. We propose extensions to continuous outcomes, and finally, illustrate the proposed estimators in a randomized experiment studying the effects of influenza vaccination encouragement on flu-related hospital visits.

Causal Inference and SUTVA/Consistencies Violations

On the Causal Effects of Long-Term Treatments Jinglong Zhao* Jinglong Zhao, Shan Huang, Chen Wang, Yuan Yuan,

One lingering challenge of randomized controlled trials is to estimate the unobserved long-term treatment effects with limited short-term experimental data. Existing literature concerns more about the long-term effects of short-term treatments. In this paper, we focus on the long-term treatments, which would be repeatedly assigned to users once the intervention is rolled out. We propose a mathematical framework, which we refer to as a longitudinal surrogate model, to study the long-term treatment effects with historical and short-term experimental data. We show that under standard assumptions, the long-term treatment effects can be estimated by an iterative expectation expression conditional on short-term surrogates and treatment assignments. Detailed instruction on the practical estimation process and required assumptions is discussed. We verify the efficacy of our approach with empirical large-scale holdout experiments conducted on the WeChat platform. Considering the accumulated short-term effect as the benchmark, we evaluate the estimated long-term effect generated by our framework and show the validity of our approach.

Weighting

Balancing weights in factorial observational studies Ruoqi Yu* Ruoqi Yu, Peng Ding,

Factorial design is a common and easy-to-use tool to evaluate causal effects with multiple treatments. The main literature focuses on randomized experiments, but it remains challenging to draw reliable causal inferences in observational studies. In recent years, several methods have been proposed to deal with observational data, ignoring the factorial structures and treating the treatment combinations as a multi-leveled treatment. However, as the number of treatment combinations grows exponentially as the number of treatments, some treatment combinations can be rare or unobserved, raising new challenges in the definition of causal estimands and the downstream inference. To overcome the limitations, we propose a new weighting framework that (i) adjusts the confounding effects of observed covariates for all contrasts of interest, (ii) takes care of the factorial structures of any number of treatments, (iii) can be easily generalized to fractional factorial design and incomplete factorial design, (iv) is computationally efficient. We also conduct numerical studies to evaluate the performance of the newly proposed method in simulations and an empirical application.

Heterogeneous Treatment Effects

Detecting Treatment Effect Disparities at Scale Winston Chou* Winston Chou, Nathan Kallus, Danielle Rich, William Nelson,

Experimentation and causal inference increasingly drive innovation on digital platforms. Often, causal analyses focus on Average Treatment Effects (ATEs), which summarize the impact of a new product innovation across the user population. Yet, there is a growing recognition that such averages do not capture the full picture: ATEs can be dragged up or driven down by small user segments with a disproportionate reaction to the innovation, a positive ATE does not imply that a majority of members benefit from the treatment, and a null or statistically insignificant estimate of the ATE is not inconsistent with polarizing effects that lift metrics for some members and depress them for others.

In this paper, we describe a methodology, integrated into Netflix's scaled experimentation platform, for estimating the range of Conditional Average Treatment Effects (CATEs) in an experiment. The bounds of this range correspond substantively to the treatment effects on the best- and worst-affected user segments in the experiment. We term the difference between these bounds the treatment effect disparity. In surfacing this disparity, our method identifies when product innovations have distinct and even polarizing effects on users and highlights opportunities to make product wins more equally distributed.

0431

Propensity Scores

Causal Inference for Complex Continuous-time Longitudinal Studies Andrew Ying* Andrew Ying,

The existing causal inference frameworks for identifying causal effects for longitudinal studies typically assume that time advances in discrete time steps. However, medical studies nowadays with either irregular visit times or real-time monitoring have posed threats to the existing frameworks, rendering them invalid or to the very least, inefficient usage of the data. Therefore more general and advanced theory around causal inference for longitudinal data when confounders and treatments are measured continuously across time is needed. We develop a framework to identify causal effects under a user-specified treatment regime for continuous-time longitudinal studies. We provide sufficient identification assumptions including generalized consistency assumption, sequential randomization assumption, positivity assumption, and a novel “achievable” assumption designed for continuous time. Under these assumptions, we propose a g-computation process and an inverse probability weighting process, which suggest a g-computation formula and an inverse probability weighting formula for identification. For practical purposes, we also construct two classes of population estimating equations to identify these two processes, respectively, which further suggest a doubly robust formula that identifies causal effects under the user-specified treatment regime with extra robustness against process misspecification.

Difference in Differences

Double-Robust Two-Way-Fixed-Effects Regression For Panel Data Lihua Lei* Lihua Lei, Dmitry Arkhangelsky, Guido Imbens, Xiaoman Luo,

We propose a new estimator for the average causal effects of a binary treatment with panel data in settings with general treatment patterns. Our approach augments the two-way-fixed-effects specification with the unit-specific weights that arise from a model for the assignment mechanism. We show how to construct these weights in various settings, including situations where units opt into the treatment sequentially. The resulting estimator converges to an average (over units and time) treatment effect under the correct specification of the assignment model. We show that our estimator is more robust than the conventional two-way estimator: it remains consistent if either the assignment mechanism or the two-way regression model is correctly specified and performs better than the two-way-fixed-effect estimator if both are locally misspecified. This strong double robustness property quantifies the benefits from modeling the assignment process and motivates using our estimator in practice.

Machine Learning and Causal Inference

Augmented Balancing Estimators of the Average Treatment Effect on the Treated in cross-sectional and panel data Apoorva Lal* Apoorva Lal,

Recent developments in the use of machine learning methods for causal inference typically target the average treatment effect (ATE) and frequently rely on estimating a propensity score using nonparametric regression learners and inverting it to plug into the doubly-robust IPW score. In observational studies, however, the ATE is frequently difficult to target because of the failure of overlap, which is compounded by the inversion step; researchers often target the average treatment effect on the treated (ATT) in such cases. We propose a unified framework for augmented balancing estimators for the ATT in a wide variety of research designs used by applied researchers, including cross-sectional, two-period difference in differences, and longitudinal data settings. We propose set of estimators that combine doubly-robust estimators for the ATT with balancing weights that directly targets in-sample covariate balance. In simulation studies, we find that balancing weights outperform conventional estimators that involve inverting a propensity score, and conclude with empirical applications.

Continuous Interventions

Causal Inference with Continuous Multiple Time Point Interventions Michael Schomaker*
Michael Schomaker, Iván Diaz, Paolo Denti, Helen McIlleron,

Currently, there are limited options to estimate the effect of variables that are continuous and measured at multiple time points on outcomes, i.e. through the dose-response curve. However, these situations may be of relevance: in pharmacology, one may be interested in how outcomes of people living with -and treated for- HIV, such as viral failure, would vary for time-varying interventions such as different drug concentration trajectories. A challenge for doing causal inference with continuous interventions is that the positivity assumption is typically violated. To address positivity violations, we develop projection functions, which reweigh and redefine the estimand of interest based on functions of the conditional support for the respective interventions. With these functions, we obtain the desired dose-response curve in areas of enough support, and otherwise a meaningful estimand that does not require the positivity assumption. We develop g-computation type plug-in estimators for this case. Those are contrasted with using g-computation estimators in a naïve manner, i.e. applying them to continuous interventions without addressing positivity violations. The ideas are illustrated with longitudinal data from HIV+ children treated with an efavirenz-based regimen. Simulations show in which situations a naïve g-computation approach is appropriate, and in which it leads to bias and how the proposed weighted estimation approach recovers the alternative estimand of interest.

Causal Discovery

Climate Dynamics via Spatially Informed Causal Discovery J. Jake Nichol* J. Jake Nichol, Michael Weylandt, Laura P. Swiler,

Understanding the Earth’s climate is perhaps today’s largest, most complex, and most important scientific challenge, and poses many difficulties for causal inference. Accurate characterization of the underlying causal mechanisms is essential for the design and analysis of potential climate change mitigation strategies. Climate data science is challenging due to both the scale of underlying data and the difficulty in obtaining meaningful replicates and counterfactuals, even in simulation. Earth systems data typically consists of hundreds of quantities of interest, each of which is observed at over 4.5 million distinct spatial locations; by contrast, less than a thousand observations are typically available. In this ultra-high dimensional regime, popular methods for causal discovery, such as the PCMCI algorithm (Runge, et al. Science Advances, 2019), exhibit high false positive rates and are unable to separate the causal wheat from the correlative chaff. To address these challenges, we propose new approaches for causal discovery that leverage regionally homogeneous spatial dynamics to create informative pseudo-replicates, improving statistical performance and interpretability. We demonstrate the effectiveness of our approach in simulation and in an application to an important open question in atmospheric dynamics. Our approach enables causal discovery in massive spatiotemporal data and provides an important toolkit for understanding climate dynamics.

Bayesian Causal Inference

Bayesian Nonparametrics for Principal Stratification: an Application on Environmental Policies Effects on Health. Falco J. Bargagli-Stoffi* Falco J. Bargagli-Stoffi, Fabrizia Mealli, Francesca Dominici, Antonio Canale, Dafne Zorzetto,

Regulatory actions have been enacted in the United States to diminish the levels of pollutants in the air and reduce the connected environmental risks for health. Indirect accountability studies -assessing the causal effect of exposure to higher levels of air pollution- and direct accountability studies -assessing the causal impact of interventions aimed at reducing the level of air pollution- have found solid evidence of health benefits. However, the existing literature lacks robust methods that consider two crucial points in health studies: evaluate heterogeneity in the health effects of air pollution regulations across different groups of individuals, and consider the joint relations between direct and indirect effects. In this work, we develop a novel approach combining Bayesian nonparametric (BNP) methods and Principal stratification (PS) framework to deal with post-treatment variables that are potentially affected by the treatment and also affecting the response. We introduce three major innovations: (i) we rely on BNP methodologies for the imputation of missing potential outcomes for the post-treatment and outcome variables; (ii) we introduce new conditional estimands; (iii) we propose a data-driven methodology to discover causal heterogeneity. We illustrate the performance of the method through simulations. In the application we discover and estimate the heterogeneous effects of US national air quality regulations on pollution levels and health outcomes.

Machine Learning and Causal Inference

Stan + BART for causal inference: improved performance for heterogeneous effects George Perrett* Jennifer Hill, George Perrett, Vincent Dorie, Ben Goodrich,

A wide range of machine-learning-based approaches to causal inference have been developed in the past decade, increasing our ability to accurately model nonlinear and non-additive response surfaces. This has improved performance for inferential tasks such as estimating average treatment effects in situations where standard parametric models may not fit the data well. These methods have also shown promise for the related task of identifying heterogeneous treatment effects. However, the estimation of both overall and heterogeneous treatment effects can be hampered when data are structured within groups if we fail to correctly model the dependence between observations. Most machine learning methods do not readily accommodate such structure. This paper introduces a new algorithm, stan4bart, that combines the flexibility of Bayesian Additive Regression Trees (BART) for fitting nonlinear response surfaces with the computational and statistical efficiencies of using Stan for the parametric components of the model. We demonstrate how stan4bart can be used to estimate average, subgroup, and individual-level treatment effects with stronger performance than other flexible approaches that ignore the multilevel structure of the data as well as multilevel approaches that have strict parametric forms.

Machine Learning and Causal Inference

Minimax optimal counterfactual density estimation Edward Kennedy* Edward Kennedy,

Causal effects are often characterized with averages – but these can give an incomplete picture of the underlying counterfactual distribution, e.g., when treatment mostly affects spread or other more complex distributional features, beyond the mean. Therefore in this work we consider estimating the entire counterfactual density. We derive the minimax rate for counterfactual density estimation, in a nonparametric model where distributional components are Holder-smooth, and present several new estimators, giving high-level conditions under which they are minimax optimal. Importantly, our minimax results are derived via a localized version of the method of fuzzy hypotheses, combining lower bound constructions for nonparametric regression and functional estimation (thus providing connections to heterogeneous effect estimation). The minimax rate we find exhibits several interesting features, including a non-standard elbow phenomenon and an unusual interpolation between nonparametric regression and functional estimation rates. We illustrate our methods by estimating the density of CD4 count among patients with HIV, had all been treated with combination therapy versus zidovudine alone. Our results yield the practically important conclusion that combination therapy may have increased CD4 count most for high-risk patients.

Randomization Tests

Fisher Meets BART: Integrating Causal Machine Learning with Randomization Tests

JungHo Lee* JungHo Lee, Panos Toulis, David Puelz,

The Fisher randomization test provides an attractive, robust testing methodology that is finite-sample valid. However, it cannot be immediately applied for testing non-sharp, or weak, null hypotheses that are often of greater empirical interest. This paper develops an approach for testing a general class of weak null hypotheses. The key idea is to use a flexible causal machine learning model to provide plausible values of the individual treatment effects under the null, and then reject or accept these values through a standard randomization test. Hence, a contribution of this work is to integrate randomization testing and machine learning methods for causal inference. We demonstrate our methodology on weak null hypotheses involving the sample average treatment effect and treatment effect heterogeneity in simulated and real-world scenarios.

Difference in Differences

Handling Correlation in Stacked Difference-in-Differences Estimates with Application to Medical Cannabis Policy Nicholas Seewald* Nicholas Seewald, Beth McGinty, Kayla Tormohlen, Elizabeth Stuart,

Health policy researchers often have questions about the effects of state policy on individual-level outcomes collected over multiple time periods. In some cases, these studies are conducted using large-scale, individual-level administrative data such as health insurance claims. However, there are multiple open questions about the use of such individual-level data in difference-in-differences (DiD) analyses. “Stacked” DiD is one approach to estimate treatment effects when units implement the policy of interest at different times: for each unit enacting a policy, we construct a comparison group of units that never enact (or had not yet enacted) the policy, analyze each treated unit separately with its comparison group, then pool effect estimates. However, when individual-level data is available, some individuals in untreated units can contribute to comparison groups for multiple treated states, producing correlation between stacked estimates. Existing methods do not quantify or account for this sharing of controls: this leads to incorrect inference. Here, we present a framework for estimating and managing this correlation when pooling stacked effect estimates. We explore the statistical properties of this approach, examine its performance in realistic simulations and with real data, and explain its occasionally counterintuitive effects on variance estimates. This is motivated by a study investigating the effects of state medical cannabis laws on opioid prescribing for pain.

Causal Inference and SUTVA/Consistencies Violations

Fuzzy Difference-in-Differences for Spatiotemporal Data Andrej Srakar* Andrej Srakar, Marilena Vecco,

Difference-in-differences (DiD) literature is a fast growing field in econometrics. Topics such as presence of serial correlation, clustered standard errors, arbitrary covariance structures, parallel growth assumption, negative controls, matching, synthetic control, semi- and nonparametrics, nonlinear models, multiple and continuous treatments and staggered treatment adoption have been subject to recent research. In previous contributions DiD has been extended to spatial data (Delgado and Florax, 2015). We extend this in a treatment effect with network interference context by controlling for violated stable unit treatment values assumption (SUTVA), inherent for such analysis but seldom controlled so far and in a spatiotemporal autoregressive setting. We develop a time-corrected Wald ratio DiD estimator combining fuzzy DiD approach with extensions to changes-in-changes estimation (Athey and Imbens, 2006; De Chaisemartin and D'Haultfoueille, 2017) with a graphon random graph specification (Li and Wager, 2022). This allows asymptotic analysis in terms of bounds on the moments and Stein approaches including Monte Carlo simulation results. In an application, we study causal effects of the yearly Venice carnival, being able to isolate the effect respective to other competing large events in Venice in the studied period. We consider extensions using spillover double robust DiD and Bayesian approaches.

Difference in Differences

On policy evaluation with sequential exogeneity Dmitry Arkhangelsky* Dmitry Arkhangelskiy, Yahu Cong,

We develop a systematic approach to causal inference with panel data under the assumption of sequential exogeneity. First, we propose a causal model that incorporates strict and sequential exogeneity. Focusing on staggered adoption design, we show that no linear estimator can generally identify a convex combination of treatment effects. This result contrasts sharply with recent literature on strictly exogenous models and older panel data literature on sequential exogeneity. To mitigate this negative result, we propose two approaches. First, we show how to quantify the worst-case bias caused by sequential exogeneity by connecting it to the deviation from the parallel trends. Second, we show that a convex combination of treatment effects can be identified if we have access to an ex-ante control group.

Machine Learning and Causal Inference

Optimal allocation of data-collection resources in partially identified causal systems Dean Knox* Dean Knox, Guilherme Duarte,

Complications in applied research often prevent point identification of causal estimands using cheaply available data—at best, sharp bounds containing ranges of possibilities can be reported. To make progress, researchers frequently collect more information by (1) re-cleaning existing datasets, (2) gathering secondary datasets, or (3) pursuing entirely new designs. Common examples include manually correcting missingness, recontacting attrited units, validating proxies with ground-truth data, finding new instrumental variables, and conducting follow-up experiments. These auxiliary tasks are costly, forcing tradeoffs with (4) larger samples from the original approach.

We define each task's efficiency as expected information gain per unit cost. Gain is formalized as the narrowing of the confidence region on sharp bounds, capturing two kinds of benefits: point-identifying new aspects of the causal system and reducing statistical uncertainty. Leveraging recent advances in automatic bounding (Duarte et al., 2022), we prove efficiency is computable for essentially any discrete causal system, estimand, and auxiliary data task.

We propose a method for optimal adaptive allocation of data-collection resources. Users first input a causal graph, estimand, and past data. They then enumerate distributions from which future samples can be drawn, fixed and per-sample costs, and any prior beliefs. Our method automatically derives and sequentially updates the optimal data-collection strategy.

Randomized Studies

Adaptive Staggered Rollout Designs in Panel Experiments Ruoxuan Xiong* Ruoxuan Xiong, Susan Athey, Mohsen Bayati, Guido Imbens,

In this paper, we study the design and analysis of experiments conducted on a set of units over multiple time periods where the starting time of the treatment may vary by unit. The design problem involves selecting an initial treatment time for each unit in order to most precisely estimate both the instantaneous and cumulative effects of the treatment. We first consider non-adaptive experiments, where all treatment assignment decisions are made prior to the start of the experiment. For this case, we show that the optimization problem is generally NP-hard and we propose a near-optimal solution. Under this solution the fraction entering treatment each period is initially low, then high, and finally low again. Next, we study an adaptive experimental design problem, where both the decision to continue the experiment and treatment assignment decisions are updated after each period's data is collected. For the adaptive case we propose a new algorithm, the Precision-Guided Adaptive Experiment (PGAE) algorithm, that addresses the challenges at both the design stage and at the stage of estimating treatment effects, ensuring valid post-experiment inference accounting for the adaptive nature of the design. Using realistic settings, we demonstrate that our proposed solutions can reduce the opportunity cost of the experiments by over 50%, compared to static design benchmarks.

Heterogeneous Treatment Effects

A nonparametric framework for treatment effect modifier discovery in high dimensions

Philippe Boileau* Philippe Boileau, Ning Leng, Nima Hejazi, Mark van der Laan, Sandrine Dudoit,

Current approaches for uncovering treatment effect modifiers are limited to low-dimensional data or data with weakly correlated confounders, or are restricted to simple data-generating processes. We develop a general framework for defining model-agnostic treatment effect modifier variable importance parameters applicable to high-dimensional data with arbitrary correlation structure, deriving nonparametric estimators of these parameters, and establishing these estimators' asymptotic properties. We showcase this framework by deriving risk-difference- and relative-risk-based treatment effect modifier variable importance parameters for data-generating processes with continuous, binary and time-to-event outcomes with binary exposures and potentially high-dimensional confounders. One-step, estimating equation and targeted maximum likelihood estimators for each parameter are provided. Certain estimators are proven to be double-robust under non-stringent conditions. All are asymptotically linear under reasonable entropy constraints on the data-generating process and consistency-rate requirements on the nuisance parameter estimators. Numerical experiments with moderate- and high-dimensional confounders demonstrate that these estimators' asymptotic
guarantees, like false discovery rate control, are approximately achieved in realistic sample sizes for observational and randomized studies alike.

Machine Learning and Causal Inference

Policy learning with asymmetric utilities Eli Ben-Michael* Eli Ben-Michael, Kosuke Imai, Zhichao Jiang,

Data-driven decision making plays an important role even in high stakes settings like medicine and public policy. Learning optimal policies from observed data requires a careful formulation of the utility function whose expected value is maximized across a population. Although researchers typically use utilities that depend on observed outcomes alone, in many settings the decision maker's utility function is more properly characterized by the joint set of potential outcomes under all actions. For example, the Hippocratic principle to "do no harm" implies that the cost of causing death to a patient who would otherwise survive without treatment is greater than the cost of forgoing life-saving treatment. We consider optimal policy learning with asymmetric utility functions of this form. We show that asymmetric utilities lead to an unidentifiable social welfare function, and so we first partially identify it. Drawing on statistical decision theory, we then derive minimax decision rules by minimizing the maximum regret relative to alternative policies. We show that one can learn minimax decision rules from observed data by solving intermediate classification problems. We also establish that the finite sample regret of this procedure is bounded by the misclassification rate of these intermediate classifiers. We apply this conceptual framework and methodology to the decision about whether or not to use right heart catheterization for patients with possible pulmonary hypertension.

Causal Inference and Bias/Discrimination

Can a risk-based release policy reduce unnecessary pretrial detention and racial disparities in detention? Lina Montoya* Lina Montoya, Jennifer Skeem, Christopher Lowenkamp,

In the US, over 400,000 people are in pretrial detention. Some scholars argue that risk assessment instruments (RAIs) – or checklists that score a person's likelihood of reoffending or failing to appear in court – can increase the capacity to release defendants safely. Others reject these tools as racially biased, concerned they will perpetuate disparities in the criminal legal system. Using federal data from the Administrative Office of the US Courts, we assess the extent to which defendants' counterfactual outcomes would have improved had release recommendations been given using a dynamic treatment regime based on the Pretrial Risk Assessment (PTRA; a RAI), compared to the status quo (i.e., detention decisions made with no RAI support). Using Targeted Maximum Likelihood Estimation, we examine these effects overall and by racial categories. For all defendants, when examining the impacts of PTRA-based decisions, we find a substantial reduction in detention (RD: -37.47%; 95%CI: [-37.29, -37.65]) and an increase in positive outcomes (RD: 35.77%; 95%CI: [35.21%, 36.32%]), with little public safety cost (RD: 1.72%; 95%CI: [1.19%, 2.25%]). Further, under this RAI strategy, compared to White defendants, Black defendants would have had fewer detentions, more positive outcomes, and comparable risk events. These findings add to RAI research, including how these tools can offset biases embedded within current decision-making practices in the US criminal legal system.

Machine Learning and Causal Inference

Efficient estimation of modified treatment policy effects based on the generalized propensity score Nima Hejazi* Nima Hejazi, David Benkeser, Ivan Diaz, Mark van der Laan,

Continuous treatments have posed a significant challenge for causal inference, both in the formulation and identification of scientifically relevant effects and in their estimation. Traditionally, focus has been placed on techniques applicable to binary or categorical treatments with few levels, which allow for application of propensity score-based methodology with relative ease. Efforts to evaluate causal effects of continuous treatments introduced the generalized propensity score, yet estimators of this nuisance parameter often rely upon restrictive, parametric assumptions that sharply limit the robustness and efficiency of inverse probability weighted (IPW) estimators. We formulate a nonparametric generalized propensity score estimator with favorable semiparametric rate-convergence properties and use it to construct nonparametric IPW estimators of a class of causal effect estimands tailored to continuous treatments. We outline several non-restrictive selection procedures for applying a sieve estimation framework to undersmooth the generalized propensity score estimator to obtain asymptotically efficient IPW estimators. We demonstrate that these IPW estimators are capable of achieving the nonparametric efficiency bound (comparable to so-called doubly robust efficient estimators) in a setting with continuous treatments, investigate their higher-order efficiency properties, and apply them to evaluate immune correlates of protection in a vaccine efficacy trial.

Sensitivity Analysis

Long Story Short: Omitted Variable Bias in Causal Machine Learning Carlos Cinelli* Carlos Cinelli, Victor Chernozhukov, Whitney Newey, Amit Sharma, Vasilis Syrgkanis,

We derive general, yet simple, sharp bounds on the size of the omitted variable bias for a broad class of causal parameters that can be identified as linear functionals of the conditional expectation function of the outcome. Such functionals encompass many of the traditional targets of investigation in causal inference studies, such as, for example, (weighted) average of potential outcomes, average treatment effects (including subgroup effects, such as the effect on the treated), (weighted) average derivatives, and policy effects from shifts in covariate distribution — all for general, nonparametric causal models. Our construction relies on the Riesz-Frechet representation of the target functional. Specifically, we show how the bound on the bias depends only on the additional variation that the latent variables create both in the outcome and in the Riesz representer for the parameter of interest. Moreover, in many important cases (e.g, average treatment effects and avearage derivatives) the bound is shown to depend on easily interpretable quantities that measure the explanatory power of the omitted variables. Therefore, simple plausibility judgments on the maximum explanatory power of omitted variables (in explaining treatment and outcome variation) are sufficient to place overall bounds on the size of the bias. We use debiased machine learning to provide flexible and efficient statistical inference on learnable components of the bounds.

Sensitivity Analysis

Assessing Omitted Variable Bias when the Controls are Endogenous Matthew Masten*
Matthew Masten, Alexandre Poirier, Paul Diegert,

Omitted variables are one of the most important threats to the identification of causal effects. Several widely used methods, including Oster (2019) and Cinelli and Hazlett (2020), have been developed to assess the impact of omitted variables on empirical conclusions. These methods all either (1) require assuming that the omitted variables are uncorrelated with the included controls, which is often considered a strong and implausible assumption, or (2) use a method called residualization to avoid this assumption. We first formally prove that the residualization method generally leads to the wrong conclusions about robustness. We then provide a new approach to sensitivity analysis that avoids this critique, allows the omitted variables to be correlated with the included controls, and lets researchers calibrate sensitivity parameters by comparing the magnitude of selection on observables with the magnitude of selection on unobservables as in previous methods. We illustrate our results in an empirical study of the effect of historical American frontier life on modern cultural beliefs. Finally, we implement these methods in the companion Stata module regsensitivity for easy use in practice.

Case-crossover

A Formal Causal Interpretation of the Case-Crossover Design Zach Shahn* Zach Shahn, Miguel Hernan, James Robins,

The case-crossover design (Maclure, 1991) is widely used in epidemiology and other fields to study causal effects of transient treatments on acute outcomes. However, its validity and causal interpretation have only been justified under informal conditions. Here, we place the design in a formal counterfactual framework for the first time. Doing so helps to clarify its assumptions and interpretation. In particular, when the treatment effect is non-null, we identify a previously unnoticed bias arising from strong common causes of the outcome at different person-times. We analyze this bias and demonstrate its potential importance with simulations. We also use our derivation of the limit of the case-crossover estimator to analyze its sensitivity to treatment effect heterogeneity, a violation of one of the informal criteria for validity. The upshot of this work for practitioners is that, while the case-crossover design can be useful for testing the causal null hypothesis in the presence of baseline confounders, extra caution is warranted when using the case-crossover design for point estimation of causal effects.

Weighting

Approximate Balancing Weights for Clustered Observational Study Designs Luke Keele*
Luke Keele, Eli Ben-Michael,

In a clustered observational study, a treatment is assigned to groups and all units within the group are exposed to the treatment. We develop a new method for statistical adjustment in clustered observational studies using approximate balancing weights, a generalization of inverse propensity score weights that solve a convex optimization problem to find a set of weights that directly minimize a measure of covariate imbalance, subject to an additional penalty on the variance of the weights. We tailor the approximate balancing weights optimization problem to both adjustment sets by deriving an upper bound on the mean square error for each case and finding weights that minimize this upper bound, linking the level of covariate balance to a bound on the bias. We implement the procedure by specializing the bound to a random cluster-level effects model, leading to a variance penalty that incorporates the signal signal-to-noise ratio and penalizes the weight on individuals and the total weight on groups differently according to the the intra-class correlation. We also develop two extensions to the procedure for cases where overlap between treated and control clusters is poor and it is difficult to balance covariates: (i) bias-correction via an outcome model, and (ii) changing the target estimand to the maximally overlapping set. In a series of simulation studies, we inspect the performance of these estimators compared. We provide further comparisons with two empirical applications.

Machine Learning and Causal Inference

Augmented linear balancing weights as undersmoothed regressions David Bruns-Smith* Avi Feller, Betsy Ogburn, Oliver Dukes,

The augmented balancing weights framework, also known as automatic debiased machine learning (AutoDML), is a powerful recent approach for causal machine learning. In this paper, we show that, for a large class of outcome and weighting models, this approach is equivalent to a form of undersmoothed regression. In particular, when both outcome and weighting models are linear in some (possibly infinite) basis, the resulting estimator collapses to a single regularized linear outcome model, where the coefficients interpolate between the original outcome model and unpenalized OLS coefficients. When the weighting model is either ridge or lasso, the implied regularization paths are exactly analogous to ridge or lasso penalties. We then specialize these results for specific choice of outcome and weight models, showing that an earlier result for OLS is a special case and that new insights can be gleaned for (kernel) ridge regression and lasso. Specifically, when both outcome and weighting models are (kernel) ridge, the combined estimator is also a form of ridge regression; when both outcome and weighting models are lasso, the included covariates in the combined estimator are the union of the included covariates in the individual models. Finally, we explore the implications for inference and hyperparameter selection in practice.

Weighting

To Balance Covariates for Time-varying Treatment: A Unification for Methods Yige Li* Yige Li, José Zubizarreta,

Due to their complexity, ubiquity, and relevance in practice, particularly in the medical sciences, longitudinal observational studies of treatment effects are one of the active frontiers in causal inference. Here, three fundamental but seemingly unrelated methods for estimating the effects of time-varying treatments are the g-computation formula, inverse probability treatment weighting (IPTW), and augmented IPTW (AIPTW). In this paper, we offer a new interpretation and diagnostics for these methods. In particular, we present a unifying framework from the standpoint of covariate adjustment or balance. We show how these methods connect and differ in finite samples, homologating them as weighting optimization methods. From this, we propose new diagnostics for longitudinal studies of treatment effects and discuss an alternative weighting approach that directly targets them. In a simulation study, we show that the proposed approach is as efficient as the g-computation formula estimator and is robust to outcome model misspecification under regularity assumptions. In a case study, we analyze the time-varying effects of drug-taking behaviors on callous-unemotional trait changes among justice-involved male adolescents.

Propensity Scores

A Graphical Adjustment Criterion for Differential Covariate Selection for Doubly Robust Estimators Peter Steiner* Peter Steiner, Weicong Lyu,

Doubly robust (DR) estimators are widely used in causal inference because they are consistent whenever either the outcome or the treatment selection is correctly modeled (Robins et al., 1994). They rely on an implicit but in practice rarely noticed condition: the set of adjustment covariates in the outcome and selection (propensity score) model must be identical (but predictors may differ). Double robustness is not guaranteed if the two models use different covariates, even if one or both models are correctly specified on their own. Using causal graphs (Pearl, 2009) this paper explains why DR estimators are in general inconsistent when different covariate sets are used and presents a graphical DR adjustment criterion that enables researchers to check whether different covariates sets for the outcome and selection model are able to remove the entire confounding bias in DR estimation procedure. In addition to the graphical adjustment criterion, we also discuss the DR covariate selection criterion in terms of potential outcomes. Using an example and simulated data, we demonstrate the application of the DR adjustment criterion and present results for two commonly used doubly robust estimators—augmented inverse probability weighting (Scharfenstein et al., 1999) and inverse probability weighted regression estimation (Schafer & Kang, 2008; Wooldridge, 2007). We conclude with suggestions for practice.

Heterogeneous Treatment Effects

Designing Replication Studies to Evaluate Sources of Effect Heterogeneity Vivian Wong*
Vivian Wong, Peter Steiner,

Despite interest by funding agencies to promote replication studies for identifying sources of effect heterogeneity, there is not yet consensus on what replication is and how these studies should be conducted. This paper addresses these challenges by providing a formal understanding of causal replication through the potential outcomes framework. We describe 5 assumptions for two or more studies to identify the same causal estimand. The assumptions may be understood as replication design requirements and individual study design requirements to identify a causal effect. Replication failure occurs when one or more replication and/or individual design assumptions are not met. An advantage of this approach is that it is straight-forward to derive research designs for replication. Research designs for direct replication examine whether studies with the same well-defined causal estimand yield the same effect. Research designs for conceptual replications examine whether studies with potentially different causal estimands yield the same effect. Within-study comparisons, multi-site, multi-arm treatment, stepped-wedge, and switching replication designs are common research designs that have not yet been recognized as replication designs. The paper demonstrates that high-quality replication designs for identifying sources of effect heterogeneity are feasible, ethical, and desirable in field settings.

Generalizability/Transportability

Combining Randomized and Observational Data for Generalizability in Medicaid Irina Degtiar* Irina Degtiar, Sherri Rose,

While much of the causal inference literature has focused on addressing internal validity biases, both internal and external validity are necessary for unbiased estimates in a target population of interest. However, few generalizability approaches exist for estimating causal quantities in a target population that is not well-represented by a randomized study but is reflected when additionally incorporating observational data. To generalize to a target population represented by a union of these data, we propose a class of novel conditional cross-design synthesis estimators that combine randomized and observational data, while addressing their respective biases—lack of overlap and unmeasured confounding. The estimators include outcome regression, propensity weighting, and double robust approaches. All use the covariate overlap between the randomized and observational data to remove potential unmeasured confounding bias. We developed these methods to estimate the causal effect of managed care plans on health care spending among Medicaid beneficiaries in New York City, finding substantial heterogeneity in effects on spending across plans. This has major implications for our understanding of Medicaid programs, where this heterogeneity has previously been hidden. Additionally, we demonstrate that unmeasured confounding rather than lack of overlap poses a larger concern in this setting.

Machine Learning and Causal Inference

A Two-Part Machine Learning Approach to Characterizing Network Interference in A/B Testing Yuan Yuan* Yuan Yuan, Kristen Altenburger,

Randomized control trials, or “A/B tests”, have been crucial for businesses to understand the impact of new product or user experience changes. However, conventional A/B testing methods are limited in the presence of interference – a unit’s response may be affected by other units’ treatments. The current literature on addressing interference in A/B tests has two major limitations: failing to account for latent network structures of interference and relying on human experts to model interference patterns. To overcome these issues, we propose a two-part machine learning approach to automatically characterize interference based on both local network structures and the treatment assignments among users in their network neighborhood. We construct network motifs with treatment assignment information, referred to as causal network motifs, to characterize the network interference status for each unit. We then develop machine learning approaches, based on decision trees and nearest neighbors, to map each causal network motif representation to an “exposure condition.” We demonstrate the validity of our approach through two sets of experiments: a simulated experiment on Watts-Strogatz network and a 1-2 million user experiment on Instagram. Overall, our approach offers a complementary, automated method that can enhance the capabilities of human experts in addressing interference in A/B testing.

Causal Inference and SUTVA/Consistencies Violations

Understanding Shift-share Designs from the Perspective of Interference Ye Wang* Ye Wang,

Shift-share designs have been widely adopted by researchers in political economy to evaluate the impacts of random shocks on regions with varying degrees of exposure to these shocks. Examples include how import competition from China affects local labor markets or election results across the United States, and how immigrants from different origins accelerate the advance of innovations in different research fields. Nevertheless, existing approaches are built upon structural restrictions on treatment effects, including additivity and homogeneity of the shocks' impacts. These restrictions are often violated in practice, leading to biases in both estimation and inference. In this paper, I argue that a shift-share design can be understood as a bipartitie experiment under interference. From this perspective, I introduce novel estimands to capture the expected average treatment effect generated by any shock. These estimands can be nonparametrically identified and consistently estimated without aforementioned restrictions. The proposed method also allows practitioners to examine how the magnitude of the impacts varies within the sample and construct confidence intervals with the correct coverage. I test the method's performance by replicating Autor, Dorn and Hanson (2013) and detect heterogeneity of treatment effects ignored by previous studies.

Synthetic Control Method

On the Assumptions of Synthetic Control Methods Claudia Shi* Claudia Shi, Dhanya Sridhar, Vishal Misra, David Blei,

Synthetic control (SC) methods have been widely applied to estimate the causal effect of large-scale interventions, e.g. the state-wide effect of a change in policy. The idea of synthetic controls is to approximate one unit's counterfactual outcomes using a weighted combination of some other units' observed outcomes. The motivating question of this paper is: how does the SC strategy lead to valid causal inferences? We address this question by re-formulating the causal inference problem targeted by SC with a more fine-grained model, where we change the unit of the analysis from "large units" (e.g. states) to "small units" (e.g. individuals in states). Under this re-formulation, we derive sufficient conditions for the non-parametric causal identification of the causal effect. We highlight two implications of the reformulation: (1) it clarifies where "linearity" comes from, and how it falls naturally out of the more fine-grained and flexible model, and (2) it suggests new ways of using available data with SC methods for valid causal inference, in particular, new ways of selecting observations

Synthetic Control Method

Generalized Synthetic Control Method with State-Space Model Junzhe Shao* Junzhe Shao, Linda Valeri, Mingzhang Yin, Xiaoxuan Cai,

Synthetic control method (SCM) is a widely used approach to assess the treatment effect of a point-wise intervention for cross-sectional time-series data. The goal of SCM is to approximate the counterfactual outcomes of the treated unit as a combination of the control units' observed outcomes. Many studies propose a linear factor model as a parametric justification for the SCM that assumes the synthetic control weights are invariant across time. However, such an assumption does not always hold in practice. We propose a generalized SCM with time-varying weights based on state-space model (GSC-SSM), allowing for a more flexible and accurate construction of counterfactual series. GSC-SSM recovers the classic SCM when the hidden weights are specified as constant. It applies Bayesian shrinkage for a two-way sparsity of the estimated weights across both the donor pool and the time. On the basis of our method, we shed light on the role of auxiliary covariates, on nonlinear and non-Gaussian state-space model, and on the prediction interval based on time-series forecasting. We apply GSC-SSM to investigate the impact of German reunification and a mandatory certificate on COVID-19 vaccine compliance.

Synthetic Control Method

Identification and Estimation of Casual Effects with Synthetic Controls in the Presence of Interference Wang Miao* Wang Miao,

Synthetic control methods are increasingly popular for evaluating the causal effect of a treatment taking place on only one unit while no perfect control units are available.
Synthetic control methods leverages a set of imperfect control units that are not affected by the treatment to predict the potential outcome of the treated unit had the treatment did not occur, and then estimates the treatment effect.
While treatment effects on the control units may also be present in many applications,
most current synthetic control methods do not admit such interference effects.
In this paper, we propose a novel synthetic control method that allows for interference.
We establish the identification, estimation, and inference for both the treatment effect and interference effects.
Our approach requires the number of interfered units to not exceed half of the total number of units, which is plausible in certain situations including several well studied examples.
However, we do not require exact prior knowledge on the interference structure.
Robust regression is implemented to produce a consistent estimator and a reasonable inferential procedure.
We illustrate with simulations and an application to evaluating the effect of relocating the US embassy on the number of conflicts in the Middle East,
with comparisons to competing methods.

Causal Inference and Common Support Violations

Off policy evaluation without overlap through smoothness assumptions Samir Khan* Samir Khan, Johan Ugander, Martin Saveski,

Existing methods for off-policy evaluation typically require either overlap or a well-specified model. In this work, we develop a new approach to off-policy evaluation that requires neither. Instead, we assume the conditional mean of the response is Lipschitz with respect to covariates. Under this assumption, we can reweight in the overlap region, and upper and lower bound the contribution of the non-overlap region by solving LPs derived from the Lipschitz condition. This approach gives partial identification bounds for the off-policy value that generalizes the Manski bounds obtained by assumptions on the range of the response.

More specifically, we: (1) derive a closed form solution to the LP that bounds the contribution of the non-overlap region, making our method highly interpretable and transparent; (2) prove that a bootstrap confidence interval for the value of the target policy is asymptotically valid, enabling inference; (3) consider settings in which there is weak overlap, and show on real data examples that in such a setting, we can shrink our confidence intervals by treating points with weak overlap as though they have no overlap. This last perspective is analogous to sample trimming, except that we do not require a change of estimand, since our smoothness assumptions allow us to partially identify the original estimand even after trimming.

Matching

Asymptotically exact propensity and prognostic score matching, with application to recovery from pandemic-related learning loss in K-12 education Ben Hansen* Ben B. Hansen, Mark Fredrickson, Josh Wasserman,

Matches made using estimated propensity or prognostic scores are inherently inexact, but with suitable combinations of caliper restrictions make them exact in the asymptotic limit. Such matches lead to directly adjusted causal effect estimates that are consistent, assuming a parametric scoring model and the absence of hidden bias, under the weakest possible overlap condition. Asymptotic exactness is often simultaneously attainable for a propensity and a prognostic score, if not also for multiple specifications of either or both scores. This engenders a multiple robustness for matched, as opposed to inverse-probability weighted, estimation.
Following a sketch of their supporting theoretical arguments, these points will be illustrated in an observational study of an innovative Texas K-8 public education program. Research on effectivness of this program was selected for federal funding in an Institute for Education Sciences competition aimed at identifying policies to promote recovery from pandemic-related learning loss. Texas happens to have begun program rollout on the eve of the pandemic, and happens to make public sufficient school-level information to make a matched study feasible. While end analysis of the federally funded project will involve protected student data and a number of additional design and analysis elements, the simpler preliminary analysis to be described in this talk centers publicly available data, and asymptotically exact matching.

Randomized Studies

Experimenting under Stochastic Congestion Shuangning Li* Shuangning Li, Ramesh Johari, Stefan Wager, Kuang Xu,

Stochastic congestion, a phenomenon in which a system becomes overwhelmed by fluctuations in demand, occurs frequently in many industries. In this paper, we study how randomized experiments can be conducted under stochastic congestion to help gain better insights into the behavior of stochastic systems. In particular, we study two switchback experiments, where treatments are simultaneously randomized sequentially for every unit, and a local perturbation experiment, where treatments are randomized independently for each unit. We focus on a simple stochastic system that has a single queue with an outside option. Aiming to estimate the effect of a system parameter on the average arrival rate to the system, we establish central limit theorems for the proposed estimators. We establish that the estimator from the local perturbation experiment is asymptotically more accurate than the estimators from the switchback experiments.

Causal Inference and Bias/Discrimination

The Observational Target Trial: A Conceptual Model for Measuring Disparity John Jackson*

John Jackson, Yea-Jen Hsu, Raquel Greer, Romsai Boonyasai, Chanelle Howe,

We present a conceptual model for measuring disparity using an observational target trial (an inception cohort with minimal intervention). First, we discuss disparity definitions in public health and medicine and how they relate to a descriptive measure of disparity. Second, we outline the key elements of the target trial and provide inverse probability weighting and g-computation estimators to emulate it. Third, we discuss non-random selection into the eligible population and its contribution to a disparity measure from a normative and moral perspective. Fourth, for investigators who wish to do so, we extend our target trial model and its emulation to remove the contribution of non-random selection from disparity (via a stochastic intervention on all or some of the variables that establish eligibility) under various causal structures. We demonstrate our methods using electronic medical records to measure racial disparities in hypertension control in a regional health system.

Difference in Differences

Universal Difference-in-Differences Chan Park* Chan Park, Eric Tchetgen Tchetgen,

Difference-in-differences (DiD) is a popular method to evaluate causal effects of real-world policy interventions. To identify the average treatment effect on the treated, DiD relies on the parallel trends (PT) assumption, which states that the time trends for the average of treatment-free potential outcomes are parallel across the treated and control groups. A well-known limitation of the PT assumption is its lack of generalization to causal effects for discrete outcomes and to nonlinear effect measures. In this paper, we consider Universal Difference-in-Differences (UDiD) based on an alternative assumption to PT for identifying treatment effects for the treated on any scale of potential interest, and outcomes of an arbitrary nature. Specifically, we introduce the odds ratio equi-confounding (OREC) assumption, which states that the generalized odds ratios relating the treatment-free potential outcome and treatment are equivalent across time periods. Under the OREC assumption, we establish nonparametric identification for any potential treatment effect on the treated in view. Moreover, we develop a consistent, asymptotically linear, and semiparametric efficient estimator for any given treatment effect on the treated of interest which leverages recent learning theory. We illustrate UDiD with simulations and two real-world applications in labor economics and traffic safety evaluation.

Instrumental Variables

Synthetic Instrumental Variables in Diff-in-Diff designs with unmeasured confounding

Jaume Vives-i-Bastida* Jaume Vives-i-Bastida, Ahmet Gulek,

Unmeasured confounding and selection into treatment are key threats to reliable causal inference in Difference-in-Differences (DiD) designs. In practice, researchers often use instrumental variables to address endogeneity concerns, for example through shift-share instruments. However, such instruments may be correlated with unobserved confounders, exhibiting pre-trends. This paper explores the use of synthetic controls to address unmeasured confounding in IV-DiD settings. We propose a synthetic IV estimator that partials out the unmeasured confounding and derive conditions under which it is consistent when the standard two-stage least squares is not. Motivated by the finite sample properties of our estimator we then propose ensemble estimators that might address different sources of bias simultaneously. Finally, we show the relevance and pitfalls of our estimator in a simulation exercise and in an empirical application.

Machine Learning and Causal Inference

Data adaptive conditional value function estimation Ashkan Ertefaie* Ashkan Ertefaie, Mark van der Laan, Brent Johnson, Luke Duttweiler,

Flexible estimation of the mean outcome under a treatment regimen (i.e., value function) is the key step toward personalized medicine. We define our target parameter as a conditional value function given a set of baseline covariates which we refer to as a stratum based value function. We focus on semiparametric class of decision rules and propose a sieve based nonparametric regimen-response curve estimator within that class. Our work contributes in several ways. First, we propose an inverse probability weighted nonparametrically efficient estimator of the smoothed regimen-response curve function. We show that asymptotic linearity is achieved when the nuisance functions are undersmoothed sufficiently. Asymptotic and finite sample criteria for undersmoothing are proposed. Second, using Gaussian process theory, we propose simultaneous confidence intervals for the smoothed regimen-response curve function. Third, we provide consistency and convergence rate for the optimizer of the regimen-response curve estimator. The latter is important as the optimizer corresponds with the optimal dynamic treatment regimen. Some finite-sample properties are explored with simulations.

Causal Inference and Bias/Discrimination

Causal Inference with Hidden Mediators AmirEmad Ghassami* AmirEmad Ghassami, Alan Yang, Ilya Shpitser, Eric Tchetgen Tchetgen,

Proximal causal inference was recently proposed as a framework to identify causal effects from observational data in the presence of hidden confounders. In this work, we extend the proximal causal inference approach to settings where identification of causal effects hinges upon a set of mediators which are not observed, yet error prone proxies of the hidden mediators are measured. Specifically, (i) We establish causal hidden mediation analysis, which extends classical causal mediation analysis methods for identifying direct and indirect effects to a setting where the mediator of interest is hidden. (ii) We establish hidden front-door criterion, which extends the classical front-door criterion to allow for hidden mediators. (iii) We show that the identification of a certain causal effect called population intervention indirect effect remains possible with hidden mediators in settings where challenges in (i) and (ii) might co-exist. We view (i)-(iii) as important steps towards the practical application of front-door criteria and mediation analysis as mediators are almost always measured with error and thus, the most one can hope for in practice is that the measurements are at best proxies of mediating mechanisms. We propose three identification approaches for the parameters of interest in our considered models. For the estimation aspect, we propose an influence function-based estimation method and provide an analysis for the robustness the estimators.

Causal Inference and SUTVA/Consistencies Violations

Using Wearables and Apps to Characterize Your Own Recurring Average Treatment Effects

Eric J. Daza* Eric J. Daza, Logan Schneider, Igor Matias, Katarzyna Wac,

Temporally dense single-person “small data” have become widely available thanks to mobile apps (e.g., that provide patient-reported outcomes) and wearable sensors. Many caregivers and self-trackers want to use these intensive longitudinal data to help a specific person change their behavior to achieve desired health outcomes. Ideally, this involves discerning possible causes from correlations using that person’s own observational time series data. In paper one, we estimate within-individual average treatment effects of sleep duration on physical activity, and vice-versa. We introduce the model-twin randomization (MoTR; “motor”) and propensity score twin (PSTn; “piston”) methods for analyzing Fitbit sensor data. MoTR is a Monte Carlo implementation of the g-formula (i.e., standardization, back-door adjustment); PSTn implements propensity score inverse probability weighting. They estimate idiographic stable recurring effects, as done in n-of-1 trials and single case experimental designs. We characterize and apply both methods to the two authors’ own data, and compare our approaches to standard methods (with possible confounding) to show how to use causal inference to make truly personalized recommendations for health behavior change. In paper two, we apply MoTR to the three authors, thereby providing a guide for using MoTR to investigate your own recurring health conditions—and demonstrating how any suggested effects can differ greatly from those of other individuals.

Machine Learning and Causal Inference

Finite Sample Guarantees for Long Term, Dynamic, and Mediated Effects Rahul Singh* Rahul Singh,

I study a rich class of longitudinal causal parameters such as long term, dynamic, and mediated effects. The class includes heterogeneous effects that vary according to subpopulation characteristics, as well as proximal effects defined in the presence of unobserved confounding. For machine learning estimators of these parameters, I construct and justify confidence intervals. Formally, as the first main result, I prove consistency, Gaussian approximation, and semiparametric efficiency when the machine learning estimators satisfy a few simple rate conditions. To demonstrate that the rate conditions are reasonable, I verify that they hold for adversarial estimators over several machine learning function spaces. Doing so requires the second main result: a mean square rate for nested nonparametric instrumental variable regression, which appears to be new, and which is of independent interest. A key feature of these results is a multiple robustness to ill posedness for proximal causal inference in longitudinal settings.

Heterogeneous Treatment Effects

Debiasing Treatment Effect Estimation for Privacy-Protected Data: A Model Audition and Calibration Approach Ta-Wei Huang* Ta-Wei Huang, Eva Ascarza,

The growing concern for data privacy and recent regulatory changes have led organizations to implement privacy-preserving measures to protect sensitive customer information. However, there is a concern about whether these measures may hinder their ability to personalize their interventions. In this research, we examine the impact of two commonly used privacy protection methods on heterogeneous treatment effects (HTE) estimation: adding substantial noise to the data in a differentially-private way and excluding protected customer characteristics (such as gender or race) from tracking. We find that these mechanisms significantly impact the prediction accuracy of current HTE estimation methods, resulting in suboptimal targeting policies.

To overcome this problem, we propose a generic post-processing approach that combines recent advances in multi-group fairness and HTE estimation. This bias correction mechanism divides the experimental data into three folds: the first is to construct an initial HTE model, the second is to identify subgroups with large prediction errors and calibrate the model by a boosting procedure using the information on those groups, and the third is to stop the calibration procedure. Using a set of simulation analyses and real-world applications, we show that the proposed method significantly improves the accuracy of HTE estimation and provides more effective targeting policies when the data is collected under the above privacy-preserving measures.

Synthetic Control Method

Synthetic Blip Effects: Generalizing Synthetic Controls for the Dynamic Treatment Regime

Anish Agarwal* Anish Agarwal, Vasilis Syrgkanis,

We propose a generalization of the synthetic control and synthetic interventions methodology to the dynamic treatment regime. We consider the estimation of unit-specific treatment effects from panel data collected via a dynamic treatment regime and in the presence of unobserved confounding. That is, each unit receives multiple treatments sequentially, based on an adaptive policy, which depends on a latent endogenously time-varying confounding state of the treated unit. Under a low-rank latent factor model assumption and a technical overlap assumption we propose an identification strategy for any unit-specific mean outcome under any sequence of interventions. The latent factor model we propose admits linear time-varying and time-invariant dynamical systems as special cases. Our approach can be seen as an identification strategy for structural nested mean models under a low-rank latent factor assumption on the blip effects. Our method, which we term “synthetic blip effects”, is a backwards induction process, where the blip effect of a treatment at each period and for a target unit is recursively expressed as linear combinations of blip effects of a carefully chosen group of other units that received the designated treatment. Our work avoids the combinatorial explosion in the number of units that would be required by a vanilla application of prior synthetic control and synthetic intervention methods in such dynamic treatment regime settings.

Bayesian Causal Inference

Identified vaccine efficacy for binary post-infection outcomes under misclassification without monotonicity Rob Trangucci* Rob Trangucci, Yang Chen, Jon Zelner,

In order to meet regulatory approval, pharmaceutical companies often must demonstrate that new vaccines reduce the total risk of a post-infection outcome like transmission, symptomatic disease, or severe illness in randomized, placebo-controlled trials. Given that infection is necessary for a post-infection outcome, one can use principal stratification to partition the total causal effect of vaccination into two causal effects: vaccine efficacy against infection, and the principal effect of vaccine efficacy against a post-infection outcome in always-infected patients. Despite the importance of such principal effects to policymakers, these estimands are generally unidentifiable, even under strong assumptions that are rarely satisfied in real-world trials. We develop a novel method to nonparametrically point identify these principal effects while eliminating the monotonicity assumption and allowing for measurement error. Moreover, our results allow for multiple treatments, and are general enough to be applicable outside of vaccine efficacy. Our method relies on the fact that many vaccine trials are multi-center trials, and measure biologically-relevant categorical pretreatment covariates. We show our method can be applied to clinical trial settings where vaccine efficacy against infection and a post-infection outcome can be jointly inferred. This can yield new insights from existing vaccine efficacy trial data and will aid researchers in designing new multi-arm clinical trials.

Bayesian Causal Inference

Bayesian Nonparametric Estimation of Principal Causal Effects in Presence of Partial Compliance from Patients Benjamin Baer* Biraj Guha, Ashkan Ertefaie, Michael Kosorok,

In causal inference, the average causal treatment effect is typically studied conditional on given baseline covariates of patients and is termed Conditional Average Treatment Effect (CATE). In the presence of post-treatment variables like potential compliance values of patients, Principal Stratification (PS) groups them to allow for causal interpretation of the Principal Causal Effect (PCE) estimates. This estimation procedure requires the learning of the latent, partially observed stratum for each patient. The outcome model learns the relation between the outcomes and the potential compliance variables, while the strata model encodes the missing information due to the potential compliance framework. Current literature lacks the use of flexible nonparametric Bayesian regression models for the outcomes, while the strata models in several works do not enjoy a rich, yet identifiable model. Our contribution includes a modeling novelty on both these fronts. For the outcomes, we propose a random covariate Gaussian Process regression model, where the two potential outcomes are separately modeled, then connected. We discuss how to solve ensuing identifiability issues. For the strata model, we use a novel Dirichlet Process mixture of Beta distribution based Generalized Linear Models (GLM), ensuring high flexibility in learning the latent strata values. The unobserved potential compliances are discriminatively modeled in contrast to generative joint modeling in previous works.

Bayesian Causal Inference

A New Bayesian Spike and Slab Approach for Causal Effect Estimation Brandon Koch*
Brandon Koch,

Two Bayesian approaches have recently been developed for causal effect estimation that use spike and slab priors. Rather than modeling the outcome only (e.g., using a lasso to model the outcome as a function of treatment and covariates), the approaches consider models for both outcome and treatment with priors that aim to control for important confounding variables weakly related to outcome. Both approaches have been shown to estimate the treatment effect with small bias and variability compared to alternative approaches across a variety of simulations, especially in settings when the sample size is small in relation to the number of total covariates under consideration. Coverage rates of confidence intervals are also generally higher when using the spike and slab approaches compared to alternative techniques. However, one approach cannot be directly extended to non-continuous outcomes or account for treatment effect heterogeneity, and the other approach requires separately fitting heterogeneous and non-heterogeneous models and choosing one based on an information criterion. In this talk, we discuss a new Bayesian spike and slab method that addresses these limitations by introducing a model and prior that is applicable to non-continuous outcomes and allows the treatment effect to be homogeneous or heterogeneous based on the data without fitting two separate models. Simulations demonstrate improved effect estimation with the new approach over the alternative approaches.

Machine Learning and Causal Inference

Adapting Predictive Models to Distribution Shifts with Causal Structure and Rich Data

Alexander D'Amour* Alexander D'Amour, Ibrahim Alabdulmohsin, Nicole Chiou, Arthur Gretton, Sanmi Koyejo, Matt Kusner, Stephen Pfohl, Olawale Salaudeen, Jessica Schrouff, Katherine Tsai, Qingyao Sun, Sayna Ebrahimi, Kevin Murphy,

Transportability is a central challenge for applying predictive machine learning in the real world: we often need a model to make optimal predictions in populations that are distinct from its training population. This is called the domain adaptation problem. While several domain adaptation strategies currently exist (including some that mirror standard confounder adjustment), many real-world distribution shifts are too complex for these methods to handle. In this work, we describe new domain adaptation strategies that adapt to changes in (1) so-called spurious correlations, and (2) distributions of unobserved confounders. We highlight how this problem mirrors, and generalizes, causal identification. In both cases, the key idea is to train models that incorporate richer data at training time than will be available when the model is deployed; at prediction time, these submodels can be plugged into adjustment formulas that identify the optimal target predictor. Causal structure plays a key role in the derivations of these adjustment formulas. We demonstrate how these methods can be applied to modern machine learning pipelines, using examples of distribution shifts in Chest X-ray and text data.

Generalizability/Transportability

A framework for Generalization and Transportation of Causal Estimates under Covariate Shift Apoorva Lal* Apoorva Lal, Wenjing Zheng, Simon Ejdemyr,

Randomized experiments are an excellent tool for estimating internally valid causal effects with the sample at hand, but their external validity is frequently debated. While classical results on the estimation of Population Average Treatment Effects (PATE) implicitly assume random selection into experiments, this is typically far from true in many medical, social-scientific, and industry experiments. When the experimental sample is different from the target sample along observable or unobservable dimensions, experimental estimates may be of limited use for policy decisions. We begin by decomposing the extrapolation bias from estimating the Target Average Treatment Effect (TATE) using the Sample Average Treatment Effect (SATE) into covariate shift, overlap, and effect modification components, which researchers can reason about in order to diagnose the severity of extrapolation bias. Next, We cast covariate shift as a sample selection problem and propose estimators that re-weight the doubly-robust scores from experimental subjects to estimate treatment effects in the overall sample (=: generalization) or in an alternate target sample (=: transportation). We implement these estimators in the open-source R package causalTransportR and illustrate its performance in a simulation study and discuss diagnostics to evaluate its performance.

Generalizability/Transportability

A common-cause principle for eliminating selection bias in causal estimands through covariate adjustment Maya Mathur* Maya Mathur,

Average treatment effects (ATEs) may be subject to selection bias when they are estimated among only a non-representative subset of the target population. Selection bias can sometimes be eliminated by conditioning on a “sufficient adjustment set” of covariates, even for some forms of missingness not at random (MNAR). Without requiring full specification of the causal structure, we consider sufficient adjustment sets to allow nonparametric identification of conditional ATEs in the target population. Covariates in the sufficient set may be collected among only the selected sample. We establish that if a sufficient set exists, then the set consisting of common causes of the outcome and selection, excluding the exposure and its descendants, also suffices. We establish simple graphical criteria for when a sufficient set will not exist, which could help indicate whether this is plausible for a given study. Simulations considering selection due to missing data indicated that sufficiently-adjusted complete-case analysis (CCA) can considerably outperform multiple imputation under MNAR and sometimes even under missingness at random if the sample size is not large. Analogous to the common-cause principle for confounding, these sufficiency results clarify when and how selection bias can be eliminated through covariate adjustment.A common-cause principle for eliminating selection bias in causal estimands through covariate adjustment

Causal Inference and Bias/Discrimination

Causal estimands for equity Laura Hatfield* Laura Hatfield,

To estimate the impacts of programs and policies on health equity, we must first define the causal target estimand. This initial step structures everything that follows in the analysis, from defining the population and outcomes to selecting comparison groups and potential confounders. Yet many health equity evaluations fail to engage seriously with this initial process. They may simply estimate different program effects in each group (e.g., Black and white patients) and informally compare the estimated treatment effects across subgroups. These analyses may conclude that equity has been improved if the effects are more beneficial in Black patients. However, this does not correspond to a principled causal analysis. In this talk, I detail five possible target estimands, the causal assumptions that would be required to identify them, and potential estimation strategies. Among these is a proposed novel estimand that puts an adjusted measure of equity on the left-hand side (using rank-and-replace methods). I discuss the causal assumptions implied by incorporating covariates directly in the outcome measure (compared to say, regression adjustment or propensity score methods). I compare these estimands' ability to yield informative conclusions about the effects of health policies and programs on equity.

Sensitivity Analysis

Sensitivity analysis for null results: Implications for studies of racially biased policing Jake Bowers* Jake Bowers, Tom Leavitt, Luke Miratrix,

We propose a method of formal sensitivity analysis for causal inference that addresses the problem of understating rather than overstating causal effects. A null result in an observational study is no more or less likely to emerge because of hidden confounding than a strong result. We motivate this work with the problem of statistically and substantively insignificant results in the study of the causal effects of race of civilian on police use of force and show how it adds to existing critiques of null results. We build on existing criticisms of naive estimation of the effect of race on police uses of force, adding a sensitivity analysis that addresses the possibilty that a given result understates the true effect both becuase of a pattern of hidden confounding and also a pattern of post-treatment missingness like that seen in datasets use to study race and police can combine to produce a misleading null effect. And we show how our method of sensitivity analysis for null effects reveals that the null result is, in fact, sensitive to these kinds of bias.

Causal Inference and Bias/Discrimination

Detecting and Mitigating Discriminatory Bias in Treatment Assignment Policies: A Causal Algorithmic Fairness Approach with a Field Experiment Joel Persson* Joel Persson, Jurriën Bakker, Dennis Bohle, Florian von Wangenheim,

Heterogeneous treatment effect (HTE) prediction is often used to learn and optimize treatment assignment policies. Typically, this involves assigning treatment to those individuals for which the predicted HTE exceeds a specified threshold. There is discriminatory bias if the prediction error in the HTE systematically varies across protected groups (e.g., race) as, then, members of some groups are assigned incorrect treatment at a higher rate than others. We develop methods for detecting and mitigating such discriminatory bias. Our methods are based on group-wise estimation and inference of the error in the average HTE predicted by a model versus the average HTE of a consistent and unbiased estimator. Our methods are general; they make minimal assumptions on the prediction model and estimator. Here, we propose estimators based on randomized, regression discontinuity, and instrumental variable designs. Via simulations, we show that our methods are consistent and unbiased in detecting and mitigating the discriminatory bias. To test our methods in practice, we partner with a leading travel marketplace and use data from a field experiment on targeted offers exceeding 1B USD in costs. We find discriminatory bias towards people from some countries but that our methods can mitigate this. Our work contributes to previous research by developing methods for detecting and mitigating discriminatory bias in treatment assignment policies and by demonstrating their performance in practice.

general causality

In Search of the Third Number guido imbens* guido imbens,

Consider a situation where a decision maker tasks a statistician with analyzing a data set to inform a decision. The decision is whether or not to implement an intervention on all members a population of units. The decision maker has given the statistician the task to analyze the available data and report the results of his analysis to inform this decision. It is common in such settings for the statistician to report a point estimate of the average effect of the intervention and a measure of the uncertainty of that estimate, say, in the form of a standard error. However, suppose the decision maker has the sophistication to absorb more information. What else should the statistician report to the decision maker? Specifically, what should the third (and fourth and fifth) numbers be that the statistician reports to the decision maker? Compared to the apparent consensus that a point estimate and standard error are the two most relevant numbers, there is much less agreement on the additional information that would be important for the decision maker to take into account. In this paper we explore some measures that have been suggested in the literature, and discuss why there is so little agreement on what else is important to take into account for decision makers.

Heterogeneous Treatment Effects

Optimal decision rules based on algorithms and human intuition Mats Stensrud* Mats Stensrud, Aaron Sarvet,

Health care providers desire to implement decision rules that, when applied to individuals in the population of interest, yield the best possible outcomes. For example, the current focus on precision medicine reflects the search for individualized treatment decisions, adapted to a patient's characteristics. In this presentation, I will introduce superoptimal regimes, which are guaranteed to outperform conventional optimal regimes. Importantly, identification of superoptimal regimes and their values require exactly the same assumptions as identification of conventional optimal regimes in several common settings. To illustrate the utility of superoptimal regimes, we derive new identification and estimation results, including a semi-parametric efficient estimator, in a common instrumental variable setting. These results are used to study two data examples that have appeared in the optimal regimes literature, illustrating that the superoptimal regimes perform better than conventional optimal regimes.

Causal Inference and Bias/Discrimination

Parsing Taste-Based from Statistical Discrimination in Audit Experiments Viviana Rivera-Burgos* Viviana Rivera-Burgos, Thomas Leavitt,

The literature on legislative responsiveness aims to parse racial (taste-based) discrimination from statistical discrimination that is due to legislators' strategic incentives to appeal to co-partisan constituents. In this paper, we show that extant designs may be unable to do so because of a lack of symmetry in when legislators are exposed to signals of race and signals of party identification. For example, in e-mail audit studies, the putative race of the e-mail sender is signaled by the e-mail address (at which point legislators can choose whether to open the e-mail), but the party of the sender is signaled to legislators only if they open the e-mail. We derive the bias for the effect of race + party treatments that results from this lack of symmetry. We then propose two solutions: (1) We show how to implement sensitivity bounds when scholars can measure whether or not legislators open an e-mail and (2) propose a new design that uses a racially neutral e-mail address and then exposes legislators to race and party cross-cutting treatments within the body of the e-mails. We implement the former solution on an original audit experiment. Both solutions enable scholars to better discern the mechanisms behind – and hence solutions to – racial discrimination in legislators' responsiveness.

Causal Inference and SUTVA/Consistencies Violations

A Design-Based Riesz Representation Framework for Randomized Experiments Fredrik Savje* Fredrik Savje, Christopher Harshaw, Yitan Wang,

We describe a new design-based framework for drawing causal inference in randomized experiments. Causal effects in the framework are defined as linear functionals evaluated at potential outcome functions. Knowledge and assumptions about the potential outcome functions are encoded as function spaces. This makes the framework expressive, allowing experimenters to formulate and investigate a wide range of causal questions. We describe a class of estimators for estimands defined using the framework and investigate their properties. The construction of the estimators is based on the Riesz representation theorem. We provide necessary and sufficient conditions for unbiasedness and consistency. Finally, we provide conditions under which the estimators are asymptotically normal, and describe a conservative variance estimator to facilitate the construction of confidence intervals for the estimands.

Causal Inference in Networks

Bayesian inference for causal effects under interference in the presence of a partially observed diffusion process on networks Fei Fang* Fei Fang, Amir Ghasemian, Laura Forastiere, Edo Airoldi,

Behaviors are likely to spread in a connected population and the presence of a behavioral intervention may boost this spread. We consider the setting where we observe at baseline the set of treated units, and at baseline and follow-up the social network and the prevalence of behaviors. To investigate the problem, we assume a network-based diffusion models, including network susceptible-infected-susceptible (SIS) model and network susceptible-infected (SI) model, formulated as a continuous-time Markov process. We develop a Bayesian data augmentation procedure to impute over time the behavioral change as a result of diffusion from social ties or as a result of the intervention for the treated. We also extend this procedure to a setting where the network also evolves. Based on the estimated parameters, we use an imputation method to evaluate the causal effects of hypothetical treatment allocations, with different rates and network-based strategies. Under simplified network models, we also derived closed forms for the expected effect of increasing the treatment rate under different baseline behavior prevalence and network structures. We apply the proposed method to a factorial randomized experiment delivering a behavioral intervention in villages in Honduras under different treatment rates and strategies. This data allows us to compare adoption rates under a hypothetical strategy imputed in one arm with the actual adoption rates observed in the arm assigned to that strategy.

Causal Inference in Networks

Theoretical insights and new algorithms for model-assisted design of experiments under network correlated outcomes Minzhengxiong Zhang* Amir Ghasemian, Minzhengxiong Zhang, Edoardo Airoldi,

Randomized trials are methodologically justified in order to achieve an unbiased estimate in causal inference.
However, an estimator even unbiased can be inefficient if it has a large variance. The standard difference-in-means estimator in a traditional randomized design setting such as bernoulli or complete randomization may output large variance estimates and renders the inference pointless. This situation is worse on networks, where the systematic relation among the units, through the network topology of the treated and controlled units, can increase the variance of the inferences through mechanisms such as Homophily and interference.
Therefore, for a precise inference, we need to restrict the randomizations through an efficient design to reduce the variance of the estimators.
In this paper, utilizing a model-assisted design paradigm for a network setting in the presence of linear homophily, we (i) investigate a set of constraints needed for optimal randomizations that will satisfy some desired properties such as unbiasedness and minimum variance, (ii) propose mechanisms to satisfy these constraints, (iii) derive the marginal mean square error equations for the proposed mechanisms, (iv) develop efficient tools to satisfy these mechanisms, and (v) illustrate the efficiency of these suggested solutions through a set of simulations.

Sensitivity Analysis

Design sensitivity and its implications for weighted observational studies Melody Huang*
Melody Huang, Daniel Soriano, Samuel Pimentel,

Increasingly, observational studies are being used to answer causal questions in the social and biomedical sciences. Estimating causal effects in observational settings often requires an assumption that unmeasured confounding is absent. This assumption cannot usually be checked empirically, and violations are often plausible. Recent work has introduced different sensitivity analyses to assess the potential impact of an unobserved confounder on a study's results post hoc. However, sensitivity to unmeasured confounding is not typically a primary consideration in designing the treated-control comparison. We introduce a framework allowing researchers to explicitly optimize robustness to omitted variable bias at the design stage using a measure called design sensitivity. Design sensitivity, which describes the asymptotic power of a sensitivity analysis, allows researchers to transparently compare the impact of different estimation strategies on sensitivity. We show how this general framework applies to two commonly-used sensitivity models, the marginal sensitivity model and the variance-based sensitivity model. By comparing design sensitivities, we interrogate how key features of weighted designs, including estimands and model augmentation, impact robustness to unmeasured confounding, and how impacts differ for the two different sensitivity models. We illustrate the proposed framework on a study examining drivers of support in the Colombian FARC peace agreement.

Propensity Scores

Optimal Refinement of Strata to Balance Covariates Katherine Brumberg* Katherine Brumberg, Dylan Small, Paul Rosenbaum,

What is the best way to split one stratum into two if the goal is to maximally reduce the imbalance in many covariates? We formulate this problem as an integer program and show how to nearly solve it by randomized rounding of a linear program. A linear program may assign a fraction of a person to one refined stratum and the remainder to the other. Randomized rounding views fractional people as probabilities, assigning intact people to strata using biased coins. Randomized rounding of a linear program is a well-studied technique for approximating the optimal solution of classes of insoluble but amenable integer programs. When the number of people in a stratum is large relative to the number of covariates, we prove the following new results: (i) randomized rounding to split a stratum does very little randomizing, so it closely resembles the unusable linear programming solution that splits intact people, (ii) the unusable linear programming solution and the randomly rounded solution place lower and upper bounds on the unattainable integer programming solution, and because of (i), these bounds are often close, ratifying the usable randomly rounded solution. We illustrate using an observational study that balanced many covariates by forming 1008 matched pairs from 2016 patients selected from 5735 using a propensity score. Instead, we form five strata using the propensity score and refine them into ten, obtaining excellent covariate balance while retaining all 5735 patients.

Synthetic Control Method

Doubly Robust Proximal Synthetic Controls Hongxiang Qiu* Hongxiang Qiu, Xu Shi, Wang Miao, Edgar Dobriban, Eric Tchetgen Tchetgen,

To infer the treatment effect for a single treated unit using panel data, synthetic control methods search for a linear combination of control units' outcomes that mimics the treated unit's pre-treatment outcome trajectory. This linear combination is subsequently used to impute the counterfactual outcomes of the treated unit had it not been treated in the post-treatment period, and used to estimate the treatment effect. Existing synthetic control methods rely on correctly modeling certain aspects of the counterfactual outcome generating mechanism and may require near-perfect matching of the pre-treatment trajectory. Inspired by proximal causal inference, we obtain two novel nonparametric identifying formulas for the average treatment effect for the treated unit: one is based on weighting, and the other combines models for the counterfactual outcome and the weighting function. We introduce the concept of covariate shift to synthetic controls to obtain these identification results conditional on the treatment assignment. We also develop two treatment effect estimators based on these two formulas and the generalized method of moments. One new estimator is doubly robust: it is consistent and asymptotically normal if at least one of the outcome and weighting models is correctly specified. We demonstrate the performance of the methods via simulations and apply them to evaluate the effect of a tax cut in Kansas on GDP.

Federated Learning

Federated Targeted Learning Rachael Phillips* Rachael Phillips, Mark van der Laan, Maya Petersen,

In many industries, including government, health care, and social media, data reside in the form of isolated islands, with limited capacity for sharing between different organizations. Policies that prevent sensitive data from crossing established boundaries may consider individual-level data to be fundamentally different to aggregated data, so that information deemed non-identifying may be shared across institutions. Federated learning (FL) is primed for learning across many sites whose data is subject to such restrictions. It is a statistical estimation paradigm that aims to use aggregate-level information to collaboratively estimate a pooled parameter, without transferring the individual-level data to a central location. In this work, we contribute to the rapid growing field of FL by connecting it with statistical theory for semiparametric efficient estimation and causal inference. In particular, we introduce a framework for federated super learning (SL) and federated targeted minimum loss-based estimation (TMLE). The class of available federated machine learning algorithms, including federated maximum likelihood estimation for parametric models, provide a powerful library of candidates in the federated SL. We show that federated TMLE can attain similar performance as the centralized TMLE that is not subject such restrictions. Our results motivate the use of flexible federated estimators that are able to adapt to underlying similarity across sites and other factors.

Generalizability/Transportability

Targeted Optimal Treatment Regime Learning Shu Yang* Shu Yang,

Personalized decision-making, aiming to derive optimal individualized treatment rules (ITRs) based on individual characteristics, has recently attracted increasing attention in many fields, such as medicine, social services, and economics. Current literature mainly focuses on estimating ITRs from a single source population. In real-world applications, the distribution of a target population can be different from that of the source population. Therefore, ITRs learned by existing methods may not generalize well to the target population. We consider an ITR estimation problem where the source and target populations may be heterogeneous. We develop a weighting framework that tailors an ITR for a target population. Specifically, we propose a calibrated augmented inverse probability weighted estimator of the value function for the target population and estimate an optimal ITR by maximizing this estimator within a class of pre-specified ITRs. We show that the proposed calibrated estimator is consistent and asymptotically normal even with flexible semi/nonparametric models for nuisance function approximation. The framework applies to general outcomes (including censored survival outcomes) and scenarios when the target sample provides individual covariate data or only summary statistics due to privacy concerns. We demonstrate the empirical performance of the proposed method using simulation studies and real clinical data applications.

Heterogeneous Treatment Effects

Bayesian Causal Forests for Ordinal Outcomes: Effects of A Synergistic Mindset Intervention on Adolescent Self Regard Anna Morgan* Anna Morgan, Jared Murray, David Yeager,

We introduce a nonparametric Bayesian approach for estimating heterogeneous effects with ordered categorical outcomes, building upon Bayesian Causal Forests (BCF). Continuous regression methods have many notable weaknesses when used with ordinal outcome data, including sensitivity to researchers' choice of outcome scores (which may be arbitrary) and susceptibility to detecting spurious interactions due to "ceiling" and "floor" effects. Because they may give misleading reports of effect moderation, continuous methods cannot take full advantage of BCF's ability to capture treatment effect heterogeneity. Our method overcomes these concerns by modeling the ordinal outcome on the latent scale using BCF's sums-of-trees and a probit link function. Furthermore, by accurately estimating posterior category probabilities, our model allows for partial identification of parameters directly related to the joint distribution of potential outcomes, such as the probability that the treatment is strictly beneficial. We propose reasonable default priors for parameters of the treatment effect function and discuss how these priors may be informed by beliefs about the potential scale of treatment effects, illustrated by simulation studies. We demonstrate the benefits of our method with a reanalysis of an experimental study evaluating the causal effects of a synergistic mindset intervention on reported self-regard among low-SES high-school students.

Multilevel Causal Inference

Aggregate regression for policy evaluation: cheaper/faster but no less accurate/precise Dan Thal* Mariel Finucane, Dan Thal,

As the organizers of the 2022 ACIC data challenge, we generated thousands of real-world-like data sets and baked in true causal impacts unknown to participants. Participating teams then competed, using their cutting-edge methods to estimate those effects. In total, 20 teams submitted results from 58 estimators that used a range of approaches. We found several important factors driving performance that are not commonly used in business-as-usual applied policy evaluations. Among these, the most surprising to us and our policy-maker colleagues was that there was no apparent benefit to analyzing large patient-level data sets (N = 300,000 patients) instead of data sets that had been aggregated to the level at which treatment status varied (the primary care practice level, N = 500 practices). Specifically, we found that performance did not vary by patient- versus practice-level analysis among the 58 submitted estimators. And within matched pairs of benchmark estimators that we ran at both the patient and practice levels, we found higher bias, larger RMSE, and wider uncertainty intervals using the disaggregated patient-level data sets, though coverage was slightly closer to nominal. In this talk, we will present follow-up work on this intriguing finding, and share intuition for why bias and power do not suffer due to aggregation. Aggregate regressions can streamline the timelines and budgets of policy evaluations, ultimately making high-quality causal evidence more widely available.

Regression Discontinuity

Safe Policy Learning under Regression Discontinuity Designs Yi Zhang* Yi Zhang, Kosuke Imai, Eli Ben-Michael,

The regression discontinuity (RD) design is widely used for program evaluation with obser- vational data. The RD design enables the identification of the local average treatment effect (LATE) at the treatment cutoff by exploiting known deterministic treatment assignment mech- anisms. The primary focus of the existing literature has been the development of rigorous estimation methods for the LATE. In contrast, we consider policy learning under the RD de- sign. We develop a robust optimization approach to finding an optimal treatment cutoff that improves upon the existing one. Under the RD design, policy learning requires extrapolation. We address this problem by partially identifying the conditional expectation function of counterfactual outcome under a smoothness assumption commonly used for the estimation of LATE. We then minimize the worst case regret relative to the status quo policy. The resulting new treatment cutoffs have a safety guarantee, enabling policy makers to limit the probability that they yield a worse outcome than the existing cutoff. Going beyond the standard single-cutoff case, we generalize the proposed methodology to the multi-cutoff RD design by developing a doubly robust estimator. We establish the asymptotic regret bounds for the learned policy using semi-parametric efficiency theory. Finally, we apply the proposed methodology to empirical and simulated data sets.

Bayesian Causal Inference

Bayesian Safe Policy Learning with Chance Constraint Optimization: Application to Military Security Assessment in the Vietnam War Zeyang Jia* Zeyang Jia, Eli Ben-Michael, Kosuke Imai,

Algorithmic recommendations have become an integral part of our society, being utilized in high-stake decision making settings. In those applications, it is essential to control the risk before putting data-driven policies into practice. A prominent frequentist approach assumes a model class for the conditional average treatment effect (CATE) and finds an optimal policy within a pre-specified policy class by maximizing the worst-case expected utility. However, when both model class and policy class are complex, the resulting optimization and uncertainty quantification are often intractable. We propose a Bayesian safe policy learning method that controls the risk via chance constraint optimization while decoupling the estimation and optimization steps. We first estimate the CATE with a Bayesian nonparametric model, then derive a safe policy by maximizing the posterior expected utility while limiting the posterior probability that the new policy negatively affects a group of individuals on average. We also show that the chance constraint optimization can be efficiently solved as a constrained linear programming problem. Our motivating application is the military security assessment policy used during the Vietnam War. By adopting ideas in graph theory, we solve the optimization over complex decision tables, which are widely used in public policy. We find that economic and social development factors should be given greater weights.

Randomized Studies

Sequential Adaptive Designs that Learn Optimal Individualized Treatment Rules by Utilizing Surrogate Outcomes Wenxin Zhang* Wenxin Zhang, Aaron Hudson, Maya Petersen, Mark van der Laan,

Randomized trials with covariate-adjusted response-adaptive (CARA) designs can be appealing because they allow assigning less patients with inferior treatments based on prior patients' responses to treatment and current patients' covariates. To implement a CARA design, one requires the outcome to be observed shortly after treatment is administered. However, when there is a long follow-up period until the outcome of interest is observed, there may be insufficient information to learn the treatment effect on the outcome conditional on patients' covariates, therefore making it challenging to implement a CARA design. In this work, we study a setting in which multiple surrogate outcomes are observed in advance of the final outcome of interest. One can then consider implementing a CARA design by assigning treatment based on the conditional average treatment effect on the surrogate outcomes. We discuss benefits and drawbacks of using surrogate outcomes with different follow-up time in a CARA design. And we propose a target causal parameter to evaluate utility of a surrogate in this setting and estimate that under the targeting maximum likelihood estimation (TMLE) framework. We also develop a CARA design with a data-adaptive strategy for choosing and utilizing the optimal surrogate to assign treatments. We illustrate the performance of our proposed adaptive design in terms of minimizing the chance of participants receiving inferior treatments through a range of simulation studies.

Propensity Scores

Propensity-score-based predictive probability of success using surrogate endpoints Jun Lu*

Jun Lu, Sanjib Basu,

Over 50% of Phase III clinical trials failed due to a lack of significant efficacy, despite proven efficacy in the completed Phase II. This discordant efficacy can be attributed to different populations and endpoints investigated in two phases. Phase III trial failures have significant consequences for both patients and investigators, making it crucial to try to prevent them from occurring. One way to mitigate the risk of failure is to use the predictive probability of success (PPoS), which is a quantitative tool based on prior knowledge and available evidence. Properly applied, PPoS can identify failing trials early and allow for their termination. However, when calculating PPoS using accumulated data, it is important to consider heterogeneity in populations and endpoints. To address this issue, we propose a method for predicting the success of future Phase III trials based on the results of past trials. Our approach adjusts for heterogeneity in populations using the propensity scores method. PPoS can be calculated using either surrogate or both surrogate and final endpoints. Additionally, in a Bayesian framework, the propensity-score-based informative prior can be used to increase sample size with reduced bias. We have applied our method to the development of a drug for multiple sclerosis.

Machine Learning and Causal Inference

Tradeoffs in Using Surrogate Variables for Decision Making with Delayed Outcomes Steve Yadlowsky* Steve Yadlowsky, Alexander D'Amour, Avi Feller,

In many decision making problems, temporal delays in observing the outcome variable can hinder the ability of an agent to adjust their policy based on new information. To overcome this challenge, practitioners often rely on surrogate variables that they believe are related to the outcome, but are observed more quickly. However, this approach only works if the surrogates satisfy certain causal assumptions about their joint relationship with the agent's actions and the outcome. In this work, we investigate the tradeoffs in using surrogates to update policies in a multi-armed bandit feedback problem, where the goal is to minimize cumulative regret, which requires learning both quickly and accurately. We parameterize the degree to which the surrogates violate the aforementioned assumptions and the length of the temporal delay. Across this parameterization, we compare two broad strategies: updating based on possibly invalid surrogates with only a short delay, and updated based on the true outcomes observed after a long delay. We characterize the range of parameters under which using the surrogates can still be beneficial in minimizing cumulative regret. Our findings provide guidance for practitioners on when and how to use surrogate variables in decision making problems with long delays in observing the outcome variable.

Bayesian Causal Inference

A Bayesian Semiparametric Approach to Treatment Effect Variation with Noncompliance

Jared Fisher* Jared Fisher, David Puelz, Sameer Deshpande,

Estimating varying treatment effects in randomized trials with noncompliance is inherently challenging since variation comes from two separate sources: variation in the impact itself and variation in the compliance rate. In this setting, existing Frequentist and ML-based methods are quite flexible but are highly sensitive to the so-called weak instruments problem, in which the compliance rate is (locally) close to zero, and require pre-specifying subgroups of interest. Parametric Bayesian approaches, which account for noncompliance via imputation, are more robust in this case, but are much more sensitive to model specification. In this paper, we propose a Bayesian semiparametric approach that combines the best features of both approaches. Our main methodological contribution is to present a Bayesian Causal Forest model for binary response variables in scenarios with noncompliance. In this Bayesian noncompliance framework, we repeatedly impute individuals' compliance types, allowing us to flexibly estimate varying treatment effects among compliers while mitigating the weak instruments problem. We then apply the method to the detect and analyze heterogeneity in study of workplace wellness, where there are a plethora of binary outcomes of interest.

Bayesian Causal Inference

Bayesian Nonparametrics for Heterogeneity in Treatment Effect. Dafne Zorzetto* Dafne Zorzetto, Falco Joannes Bargagli-Stoffi, Antonio Canale, Francesca Dominici,

In causal inference studies, some observed characteristics play a key role in the identification of heterogeneity in the treatment effect. In this work, we propose a Bayesian nonparametric (BNP) approach that incorporates the information carried by the observed characteristics, for imputing the missing potential outcomes and data-driven discovering the heterogeneity in the causal effects. The literature for BNP framework applied to causal inference for heterogeneity in treatment effect is quite recent and has mostly focused on reworks of Bayesian Additive Regression Tree (BART) (Chipman et al., 2010), — as the works of Hill (2011) and Hahn et al. (2020) — and Dependent Dirichlet Process (DDP) (MacEachern, 2000; Quintana et al., 2020) mixture models — e.g., the works of Roy et al. (2018) and Oganisian et al. (2020). Exploiting the flexibility of the DDP, we propose a Dependent Probit Stick-breaking Process (Rodriguez and Dunson, 2011) mixture model to retrieve the conditional marginal potential outcome distributions, that allow us to: (i) estimate the individual treatment effects; (ii) identify the subgroups defined by similar conditional treatment effects, and (iii) characterize the heterogeneity in the effects in a precise and interpretable manner. We illustrate the performance of the method through simulations. We apply our method to assess the causal effect heterogeneity of long-term fine exposure to PM2.5 on mortality.

Broken Randomized Studies

Estimand strategies for safety outcomes Veronica Ballerini* Fabrizia Mealli, Alessandra Mattei, Veronica Ballerini,

Safety evaluation of new therapies is an essential aspect of clinical trials, with the primary focus of quantifying the incidence of adverse events (AEs) and comparing it to a standard treatment.
Several estimand strategies have been proposed for efficacy analysis of time-to-event outcomes in the presence of censoring and competing events, also in accordance to the ICH E9 Addendum.
Safety analysis of adverse events, instead, is often rather simplistic: AE probabilities are estimated without explicitly defining the target causal comparison and neglecting assumptions on the censoring mechanisms leading to differential follow-up times (e.g., in oncology trials, patients may discontinue the control treatment earlier than the new treatment due to Progressive Disease).

Here, we explicitly define the assumptions under which estimators typically used in the literature, such as the Exposure-Adjusted Incidence Rate, Kaplan-Meier and Aalen-Johansen estimators, have a causal interpretation. We introduce new principal stratum and hypothetical estimand strategies for safety outcomes in the presence of censoring, competing events and varying follow-up times. We also propose identifying assumptions as well as estimators under these assumptions.
Our contribution will enhance interpretation of AE risks and has the potential of changing clinical trial practice with regard to safety analysis and risk-benefit assessment.

Causal Discovery

Know Your Role: A Statistical Approach for Distinguishing Mediators, Confounders, and Colliders using Direction Dependence Analysis (DDA) Dexin Shi* Dexin Shi, Amanda Fairchild, Wolfgang Wiedermann,

In observational data, understanding the causal link when estimating the causal effect of x on y often requires researchers to identify the role of a third variable in the x-y relationship. Mediation, confounding, and colliding are three key third-variable effects that provide different theoretical and methodological implications for drawing causal inferences. However, in practice, these effects are not distinguishable using the commonly used covariance-based statistical methods (e.g., linear regression and structural equation modeling). In this study, we introduce a statistical approach for distinguishing mediators, confounders, and colliders. By using higher-moment information of variables, we propose a two-step procedure within the framework of Direction Dependence Analysis (DDA). Results from Monte Carlo simulations show that our proposed approach accurately recovers the true data-generation process of the third variable. We provide an empirical example to demonstrate the application of our proposed approach in psychological studies. Finally, we discuss the implications and future directions of our work.

Causal Discovery

TarGene: Dispensing with unnecessary assumptions in population genetics analysis Olivier Labayle* Olivier Labayle, Kelsey Tetley-Campbell,

Parametric assumptions in population genetics analysis are often made, yet a principled argument for their validity is not given. We present a unified statistical workflow, based on Targeted Learning, called TarGene, for the estimation of effect sizes, as well as two-point and higher-order epistatic interactions of genomic variants on polygenic traits, which dispenses with these unnecessary assumptions. We validate the effectiveness of our method by reproducing previously verified effect sizes on UK Biobank data, whilst also discovering non-linear effect sizes of additional allelic copies on trait or disease. We demonstrate that for the FTO variant rs1421085 effect size on weight, the addition of one copy of the C allele is associated with 0.77 kg (95% CI: 0.68 – 0.85) increase, while the addition of the second C copy non-linearly adds 1.31 kg (95% CI: 1.19 – 1.43). We further find 3 pairs of epistatic loci associated with skin colour that have been previously reported to be associated with hair colour. Finally, we illustrate how TarGene can be used to investigate higher-order interactions using 3 variants linked to the vitamin D receptor complex. TarGene thus extends the reach of current genome-wide association studies by enriching the set of parameters that can be estimated whilst data-adaptively incorporating complex non-linear relations between phenotype, genotype, and confounders, as well as accounting for strong population dependence such as island cohorts.

Causal Discovery

On Learning Time Series Summary DAGs: A Frequency Domain Approach Aramayis Dallakyan* Aramayis Dallakyan,

The fields of time series and graphical models emerged and advanced separately. Previous work on the structure learning of continuous and real-valued time series utilizes the time domain, with a focus on either structural autoregressive models or linear (non-)Gaussian Bayesian Networks. In contrast, we propose a novel frequency domain approach to identify a topological ordering and learn the structure of both real and complex-valued multivariate time series. In particular, we define a class of complex-valued Structural Causal Models (cSCM) at each frequency of the Fourier transform of the time series. Assuming that the time series is generated from the transfer function model, we show that the topological ordering and corresponding summary directed acyclic graph can be uniquely identified from cSCM. The performance of our algorithm is investigated using simulation experiments and real datasets.

Causal Discovery

Understanding impact of BRCA testing on healthcare utilization and clinical outcome

Xuyang Li* Xuyang Li, Kevin Gorman, Ilya Shpitser, Carolyn Applegate, Casey Taylor,

Clinical guidelines and previous studies suggested that the utility of BRCA1/2 testing is a determinant of patient’s prevention and treatment measurements for BRCA related cancers. As expanded offering of BRCA testing to a large population that are potentially eligible could pose a burden to clinical resources, there is a need to better understand the impact of BRCA testing on clinical decision making for cancer prevention and its effectiveness at preventing cancer. In this study we aimed to estimate the effects of BRCA testing on healthcare utilization and clinical outcomes. We leveraged the scale of the MarketScan dataset and employed causal discovery algorithms to recover the causal relationships between patient features and clinical actions, including family and personal history of breast cancer, BRCA testing, enhanced screening, surgical and chemoprevention or treatment measurements for breast cancer, and clinical diagnosis of active breast cancer and genetic susceptibility to breast cancer. The causal structure was represented as a directed acyclic graph. This enables us to estimate effects of BRCA testing on various clinical actions and breast cancer onset, while controlling for confounders and indirect effects. In addition, we will conduct counterfactual analysis by introducing simulated intervention, answering the question of what would have happened if all patients that are potentially eligible due to family history were to receive BRCA testing.

Causal Discovery

Intuition AI – A Hypotheses Iteration based Scientific Learning Framework Rabindra Chakraborty* Rabindra Chakraborty,

For rare events of high consequence that exhibit extreme behaviors in biological, chemical, and geological systems, standard machine learning fails to perform with any meaningful accuracy due to lack of data. For all these occasions, esp., when ground truth is not available instantly, industries fall back on experts' interpretations to avoid high stake consequences in an operation. Intuition Technology is a patented causation AI that builds scientific models using experts' hypotheses that are then iterated towards situational ground truth using multi-view convergence, as various diverse situational datasets are run through the model. This makes builds the fabrication of a strong model despite lack of data.

The internal state of a natural system is not observable in most cases yet is often responsible for its unexplained behavior. Intuition AI is a framework that captures different situations, time delay between a cause and an effect and the degree of the effect viz-a-vis the influencers along with the experts' explanations, thus building a situational response map useful for forewarning complex system behavior.

Intuition AI has been successfully piloted with two oil supermajors, one for interpreting real-time contamination state in fluid samples in wireline operations and the other for estimating reservoir characteristics using mud-gas logging and drilling data.

Causal Inference and Bias/Discrimination

The implementation of target trial emulation for causal inference: a scoping review Hanxiao Zuo* Hanxiao Zuo, Lin Yu, Sandra Campbell, Yan Yuan,

Background
Target trial emulation (TTE) is an increasingly-used method for causal inference by emulating randomized clinical trials using observational data. Even though many TTE research projects have been conducted, the rigor of TTE implementation has not been investigated. This review mapped the implementation of TTE, including research topics, frequently used strategies, study design, and common issues for future improvement.
Methods
We searched literature in PROSPERO, OVID Medline, Wiley Cochrane Library, SCOPUS, etc. using controlled vocabulary and keywords representing the concept of “target trial emulation”. The databases were searched from the inception to January 18, 2022, and all original research that met inclusion criteria was included for screenings and information extraction.
Results
655 papers were included for screenings, of which 96 papers were eligible for extraction. Retrieved original papers show that TTE was increasingly used from 2018, and cancer research is the top studied topic (22.9%). The main components of TTE design including time-zero, assignment procedure, and contrast strategy were not implemented well in all research. Several limitations including residual confounding, limited generalization, and lack of evaluation tools need to be improved.
Conclusion
Future improvements are needed for TTE implementation and evaluation to make it better to support the causal inference using observational data.

Causal Inference and Common Support Violations

Contingency in Causal Inference AmirEmad Ghassami* AmirEmad Ghassami, Ilya Shpitser,

Many causal inference problems involve contingent potential outcomes that are only defined if a specific state of affairs occurs. For instance, a variable such as “quality of life” is only well-defined if the individual is alive. In such settings, the contingency requirement can be represented by conditioning on a set of values of a contingency variable. In this case, seemingly natural contrasts conditioned on the contingency event may not necessarily correspond to causal effects anymore. This specifically happens when contingency variables are post-treatment. Despite prevalence of such situations, unfortunately, no general treatment of contingency in causal inference problems exists in the literature. In this work, we use the formalism of graphical causal models to propose a general methodology for characterizing which contingent contrasts correspond to causally interpretable effects and what causal effect they represent. For the case that our characterization indicates that a contingent contrast cannot be interpreted as a causal effect, we propose the use of an alternative estimand, called component-wise effect. Specifically, given the assumption that the treatment variable and certain other variables in the system have more than one component, we describe a contingent contrast, defined using interventions on treatment components, and provide graphical sufficient conditions under which our introduced contingent contrast corresponds to a causally interpretable effect.

Causal Inference and Common Support Violations

Efficient Nonparametric Causal Effect Estimation after Propensity Score Trimming with a Continuous Treatment Zach Branson* Zach Branson, Sivaraman Balakrishnan, Edward Kennedy, Larry Wasserman,

This work proposes estimators using efficient influence functions (EIFs) for average treatment effects (ATEs) after propensity score trimming in observational studies with a continuous treatment. Trimming involves estimating ATEs among subjects with propensity scores above a threshold, which addresses positivity violations that complicate estimation. Most work on trimming focuses on binary treatments, and several challenges arise with continuous treatments. First, EIFs for trimmed ATEs do not exist, due to a lack of pathwise differentiability induced by trimming and a continuous treatment. Second, if we want the trimming threshold to be estimated, uncertainty in the threshold must be accounted for. To address these challenges, we target a kernel-smoothed trimmed ATE, such that an EIF exists for an estimand close to the trimmed ATE. We allow the trimming threshold to be estimated via the quantile of the propensity score, such that confidence intervals reflect uncertainty involved in threshold estimation. Our resulting EIF-based estimators exhibit doubly-robust style guarantees, where their error can be expressed as the product of errors for the outcome and propensity score models. Thus, our estimators can exhibit parametric convergence rates even when models are estimated at slower rates via flexible machine learning. These findings are validated via simulation and an application, thereby showing how to efficiently-but-flexibly estimate a dose-response function after trimming.

Causal Inference and Missing Data

Recoverability of Causal Effects in a Longitudinal Study using missingness DAGs Michael Schomaker* Anastasiia Holovchak, Michael Schomaker, Paolo Denti, Helen McIlleron,

Missing data in multiple variables is a common issue. We investigate the applicability of the framework of graphical models for handling missing data to a complex longitudinal pharmacological study of HIV-positive children treated with an efavirenz-based regimen as part of the CHAPAS-3 trial. Specifically, we examine whether the causal effects of interest, defined through static interventions on multiple continuous variables, can be recovered (estimated consistently) from the available data only. So far, there exists no general algorithm for deciding on recoverability, and decisions have to be made on a case-by-case basis. We emphasize sensitivity of recoverability to even the smallest changes in the graph structure, and present recoverability results for three plausible missingness DAGs in the CHAPAS-3 study (directed acyclic graphs), informed by clinical knowledge. Further, we propose the concept of ”closed missingness mechanisms” and show that under these mechanisms an available case analysis is admissible for consistent estimation for any type of statistical and causal query, even if the underlying missingness mechanism is of MNAR type. Simulations demonstrate how estimation results vary depending on the modelled missingness DAG. Our analyses are possibly the first to show the applicability of missingness DAGs to complex longitudinal real-world data, while highlighting the sensitivity with respect to the assumed causal model.

Causal Inference and SUTVA/Consistencies Violations

Design of Panel Experiments with Spatial and Temporal Interference Tu Ni* Tu Ni, Iavor Bojinov, Jinglong Zhao,

Panel experiments — where we expose multiple units to some random treatments, measure their responses, and repeat the procedure for some time periods — have rapidly grown popular in marketplace companies, wishing to run randomized controlled experiments (A/B testing) in the presence of spatial interference between experimental units and temporal interference between time periods.

When running the experiments, companies group units together across zip codes, cities, or even states to form a single aggregated unit, in order to alleviate the spatial interference between units, as it ensures that each unit within the aggregated unit receives the same treatment, but it does not remove the temporal interference over time.
Unfortunately, such a drastic aggregation significantly reduces the sample size, leading to much lower power for inference. This highlights a critical trade-off when a panel experiment has interference: aggregation limits the degree of interference but reduces the volume of data.

In this work, we examine this trade-off and present a new, more powerful, randomized design for panel experiments in the presence of spatial and temporal interference.
Our proposed design has two features: the first feature is a notion of cluster-based randomization that allows us to navigate the aforementioned trade-off for aggregation transparently via the cluster size; the second feature is a notion of balanced randomization of treatment and control that incorporates an assign

Causal Inference and SUTVA/Consistencies Violations

Efficient Weighted Estimators of Interference Effects under Hypothetical Treatment Allocations in Two-Stage Randomized Experiments under Bernoulli Assignment Colleen Chan* Colleen Chan, Shinpei Nakamura Sakai, Laura Forastiere,

In many applications, the no-interference assumption in causal inference is often violated as individuals often interact with one another. Two-stage randomized experiments are incredibly useful designs for estimating causal effects of a given treatment in the presence of interference. In this design, clusters are assigned a treatment saturation level in the first stage, and each unit within a cluster is randomized to treatment or control according to the assigned saturation level in the second stage. Previous two-stage designs have been proposed under complete randomization in both stages, and simple difference-in-means estimators have been developed under the partial interference assumption. However, complete randomization in the second stage only allows the estimation of causal effects under the treatment saturations of the first stage. We propose instead a Bernoulli assignment in the second stage and weighted estimators of direct and spillover effects, combining information from all clusters. One clear advantage of using Bernoulli assignment is that it allows researchers to estimate causal effects under hypothetical treatment allocations. We derive cluster weights achieving the optimal bias-variance trade-off for our estimator. We develop simulation studies to analyze the finite sample performance of our proposed estimators. Finally, we illustrate our methodology with a data-inspired information campaign to prevent anemia in India.

Causal Inference and SUTVA/Consistencies Violations

Detecting Interference in A/B Testing with Increasing Allocation Kevin Han* Kevin Han, Shuangning Li, Jialiang Mao, Han Wu,

In the past decade, the technology industry has adopted online randomized controlled experiments (a.k.a. A/B testing) to guide product development and make business decisions. In practice, A/B tests are often implemented with increasing treatment allocation: the new treatment is gradually released to an increasing number of units through a sequence of randomized experiments. In scenarios such as experimenting in a social network setting or in a bipartite online marketplace, interference among units may exist, which can harm the validity of simple inference procedures. In this work, we introduce a widely applicable procedure to test for interference in A/B testing with increasing allocation. Our procedure can be implemented on top of an existing A/B testing platform with a separate flow and does not require a priori a specific interference mechanism. In particular, we introduce two permutation tests that are valid under different assumptions. Firstly, we introduce a general statistical test for interference requiring no additional assumption. Secondly, we introduce a testing procedure that is valid under a time fixed effect assumption. The testing procedure is of very low computational complexity, it is powerful, and it formalizes a heuristic algorithm implemented already in industry. Finally, we discuss one application at LinkedIn where a screening step is implemented to detect potential interference in all their marketplace experiments with the proposed methods in the paper.

Causal Inference and SUTVA/Consistencies Violations

A Spatial Extension of Synthetic Difference-in-Differences Renan Serenini* Renan Serenini, Frantisek Masek, Renan Serenini,

We propose a spatial extension of the Synthetic Difference-in-Differences (SDiD) estimator of Arkhangelsky et al. (2021). Our estimator handles the situation of a possible violation of the Stable Unit Treatment Value Assumption (SUTVA) when treatment may spillover to control units included in the donor pool resulting in biased and inconsistent Average Treatment Effect (ATE) estimation. We build on the approach of the Spatial Difference-in-Differences estimator of Delgado and Florax (2015) and incorporate it into SDiD. Thus, the ATE can be disentangled into direct and indirect treatment effects. We compare our approach with the SDiD estimator using an example of a violation of the SUTVA. All the features presented in Arkhangelsky et al. (2021) related to the comparison of the SDiD towards conventional Difference-in-Differences (DiD) carry forward for the direct effect. Pertaining to the indirect treatment effect, we show that our estimator may be superior to Delgado and Florax (2015) in the case when directly and indirectly treated units are similar. However, this does not hold unconditionally. We suggest a fast quantitative check to compare the synthetic control unit with the control unit using uniform weights as in Delgado and Florax (2015) to decide which of the methods better satisfies the common trend assumption for the indirectly treated units in each specific case.

Causal Inference and SUTVA/Consistencies Violations

Bias in sequential decision-making for stochastic service systems Gabriel Zayas-Caban*
Gabriel Zayas-Caban, Juan Camilo David Gomez, Amy Cochran,

Decision-makers in many service systems are often confronted with making a random number of decisions sequentially over time. How sequential decisions are made may depend on the decision-maker's perception of prior and analogous decisions, and how these prior decisions led to specific outcomes. This phenomenon, whereby prior decision-making or experiences influence current or future decision-making, has been termed sequential bias. Sequential bias violates a core assumption in causal inference that the decision for one person does not interfere with the potential outcomes of another. By drawing the connection between sequential bias in service systems to dynamic treatment regimes, and extending these latter settings to allow for a randomized number of decisions, we are to define and identify average causal effects for quantifying sequential bias. Subsequently, we propose estimators, and derive properties thereof. In a case study, we use our approach to demonstrate that the decision of a provider to route a patient one way in the Emergency Department has a measurable impact on the care of future patients.

Causal Inference Education

Does association really imply non-causation? The power of language in causal attribution.

Jennifer Hill* Jennifer Hill, George Perrett, Stacey Hancock, Le Win, Yoav Bergner,

While most statisticians and, arguably, all causal researchers have been taught to be cautious in making unwarranted causal attribution, many of us are not as careful as we might be in describing results of descriptive (i.e. non-causal) evidence. For instance, it is still common practice to use the word association to describe non-causal relationships, even if this wording is combined with causal words such as “change,” “increase/decrease,” or “gain/loss.” Are these wording choices innocuous or do they inadvertently lead less sophisticated readers to assume that these links are causal? The current study investigates the connection between the wording of study findings and causal attribution by the reader using a series of randomized experiments involving several samples of students from two large U.S. universities. It also provides evidence about the association between statistics instruction and the ability to understand appropriate causal attribution. The results suggest that specific wording choices used to describe study results noticeably impact the level of causal attribution by the reader. Moreover, the degree of causal attribution is strongly linked to the research topic. These results suggest ways to tailor wording of research findings to help decrease the probability of causal misattribution.

Causal Inference in Networks

Representational Power of Exposure Mapping Functions in Network Interference Yuchen Xiao* Yuchen Xiao, Cowrin Zigler,

We study the representational power of exposure mapping functions (EMF) used in network interference. The traditional SUTVA assumption does not hold because, under interference, the potential outcomes of a unit depend on its treatment as well as on the treatments of other units. The usual approach is to embed the neighborhood treatments into a lower dimensional representation with EMF. The most common EMFs typically assume neighborhood treatments share equal weights, limiting their ability to represent some forms of interference. For example, a person’s health status may be more heavily influenced by some contacts who are encountered more frequently than others and people with the same number of treated neighbors but different number of neighbors should be expressed differently. We use Graph Attention Networks (GAT), which employ a self-attention strategy, to compute the attention coefficients (i.e., weights) of neighboring units to each unit in a network. One of the advantages of GAT is the range of interference can be extend to k-hops away, which permits long-range causal dependence. After assigning weights to neighboring units, we compute treatment and spillover effects using a generalize propensity score approach that has been used previously with exposure mappings. We detail the GAT architecture and its relative computational advantage. Finally, we make comparisons between the results estimated with GAT and other common exposure mapping functions.

Causal Inference in Networks

Estimating network-mediated causal effects via spectral embeddings Alex Hayes* Alex Hayes, Keith Levin,

We develop a model for mediation in networks, where mediation occurs in a latent node embedding space. Under our model, nodal interventions have causal effects on nodal outcomes, and these effects can be partitioned into a direct effect independent of the network, and an indirect effect, which is induced by homophily. To estimate these network-mediated effects, we embed nodes into a low-dimensional Euclidean space via the adjacency spectral embedding. We then use the embeddings to fit two ordinary least squares models: (1) an outcome model that characterizes how nodal outcomes vary with nodal treatment, controls, and position in latent space; and (2) a mediator model that characterizes how latent positions vary with nodal treatment and controls. We prove that the estimated coefficients are asymptotically normal about the true coefficients under a sub-gamma generalization of the random dot product graph, a widely-used latent space model. Further, we show that these coefficients can be used in product-of-coefficients estimators for causal inference. Our method is easy to implement, scales to networks with millions of edges, and accommodates a wide variety of structured data.

Causal Inference in Networks

Network Synthetic Interventions: A Framework for Panel Data with Network Interference

Anish Agarwal* Anish Agarwal, Sarah Cen, Devavrat Shah, Christina Lee Yu,

We propose a generalization of the synthetic controls and synthetic interventions methodology to incorporate network interference. We consider the estimation of unit-specific treatment effects from panel data where there are spillover effects across units and in the presence of unobserved confounding. Key to our approach is a novel latent factor model that takes into account network interference and generalizes the factor models typically used in panel data settings. We propose an estimator, “network synthetic interventions”, and show that it consistently estimates the mean outcomes for a unit under an arbitrary sequence of treatments for itself and its neighborhood, given certain observation patterns hold in the data. We corroborate our theoretical findings with simulations.

Causal Inference in Networks

Simulating Potential Outcomes in the Presence of Interference Gabrielle Lemire* Gabrielle Lemire, Ashley Buchanan, Natallia Katenka, Tingfang Lee,

Simulation studies aid in understanding the behavior of statistical methods, such as measuring the finite sample performance of estimators. This is possible because we know the ground truth as we are specifying the data generating mechanism. When evaluating causal estimators, simulating potential outcomes allows the researcher to overcome both challenges due to the impossibility of perfect data collection and knowledge of the data generation mechanism, but also to have full knowledge of the counterfactual outcomes. Additional challenges arise for simulating potential outcomes when observations' outcomes are dependent due to interference (i.e., when one individual's exposure affects another's outcome). We provide 1) a paradigm that considers the assumptions about the nature of the interference structure, the identification assumptions required for causal effects, the estimators being evaluated, and the computing environment; and 2) efficient tools (i.e., R functions/package available on Git Hub) for simulating potential outcomes in the presence of interference which is designed to handle both network structures with interference and cluster structures satisfying the partial interference assumption (i.e., interference only occurs within groups). Our work aids in bridging the gap between the burgeoning literature offering new causal estimators in the presence of interference and the need for reproducible evaluation of their performance in practice.

Causal Inference in Networks

Exploiting Neighborhood Interference with Low Order Interactions Under Unit Randomized Design Mayleen Cortez-Rodriguez* Mayleen Cortez-Rodriguez, Christina Yu, Matthew Eichhorn,

Network interference, where the outcome of an individual is affected by the treatment assignment of those in their social network, is pervasive in many real-world settings. However, it poses a challenge to estimating causal effects. We consider the task of estimating the total treatment effect (TTE), or the difference between the average outcomes of the population when everyone is treated versus when no one is, under network interference. Under a Bernoulli randomized design, we utilize knowledge of the network structure to provide an unbiased estimator for the TTE when network interference effects are constrained to low order interactions among neighbors of an individual. We make no assumptions on the graph other than bounded degree, allowing for well-connected networks that may not be easily clustered. We derive a bound on the variance of our estimator and show in simulated experiments that it performs well compared with standard estimators for the TTE. We derive a minimax lower bound on the mean squared error of our estimator which suggests that the difficulty of estimation can be characterized by the degree of interactions in the potential outcomes model. We also prove that our estimator is asymptotically normal under boundedness conditions on the network degree and potential outcomes model. Central to our contribution is a new framework for balancing between model flexibility and statistical complexity as captured by this low order interactions structure.

Causal Inference in Networks

Adaptive Experiments for Exposure Mapping Estimation Ravi Sojitra* Ravi Sojitra, Ruohan Zhan,

During experimentation, treating one unit can interfere with outcomes of other units. Our goal is to estimate the proximity within which such interference can be induced in networks. For example, if some people are vaccinated against the flu, we expect unvaccinated people within the same households to benefit from reduced household level risk. Such information about treatment-control spillover is valuable for both experiment design and welfare maximization under resource constraints. First, we formalize estimands to quantify how far spillovers propagate in networks and their identifying assumptions. For instance, the (average) minimum numbers of vertices that need to be treated at varying proximities to induce a minimum spillover effect. Then, we propose algorithms for sequential experiments to simultaneously estimate these estimands and improve outcomes of experimental units during the experiment. We show how one can straightforwardly apply regret minimization and pure exploration algorithms under these assumptions. Moreover, we extend these algorithms to combine randomized treatments and observational exposures to mitigate the impact of treatment dimensionality on learning.

Continuous Interventions

Causal Inference with Continuous Multiple Time Point Interventions Michael Schomaker*

Michael Schomaker, Iván Diaz, Paolo Denti, Helen McIlleron,

Currently, there are limited options to estimate the effect of variables that are continuous and measured at multiple time points on outcomes, i.e. through the dose-response curve. However, these situations may be of relevance: in pharmacology, one may be interested in how outcomes of people living with -and treated for- HIV, such as viral failure, would vary for time-varying interventions such as different drug concentration trajectories. A challenge for doing causal inference with continuous interventions is that the positivity assumption is typically violated. To address positivity violations, we develop projection functions, which reweigh and redefine the estimand of interest based on functions of the conditional support for the respective interventions. With these functions, we obtain the desired dose-response curve in areas of enough support, and otherwise a meaningful estimand that does not require the positivity assumption. We develop g-computation type plug-in estimators for this case. Those are contrasted with using g-computation estimators in a naïve manner, i.e. applying them to continuous interventions without addressing positivity violations. The ideas are illustrated with longitudinal data from HIV+ children treated with an efavirenz-based regimen. Simulations show in which situations a naïve g-computation approach is appropriate, and in which it leads to bias and how the proposed weighted estimation approach recovers the alternative estimand of interest.

Difference in Differences

Structural Nested Mean Models Under Parallel Trends Assumptions James Robins* Zach Shahn, Oliver Dukes, David Richardson, Eric Tchetgen Tchetgen,

We link and extend two approaches to estimating time-varying treatment effects on repeated outcomes–time-varying Difference in Differences (DiD) and Structural Nested Mean Models (SNMMs). In particular, we show that SNMMs, which were previously only known to be nonparametrically identified under a no unobserved confounding assumption, are also identified under a generalized version of the parallel trends assumption typically used to justify time-varying DiD methods. Because SNMMs model a broader set of causal estimands, our results allow practitioners of existing time-varying DiD approaches to address additional substantive questions (such as characterization of time-varying effect heterogeneity, estimation of the lasting effects of a blip of treatment at a single time point, controlled direct effects, and others) under similar assumptions. Further, while some common time-varying DiD approaches are restricted to staggered adoption settings in which a binary treatment is permanently sustained once initiated, SNMMs straightforwardly accommodate potentially multivariate treatments with continuous and discrete components and arbitrary treatment patterns. We also identify optimal dynamic treatment regimes under parallel trends plus the extra assumption that unobserved confounders are not effect modifiers. Finally, using SNMMs to estimate the same effects under alternative identifying conditions can potentially enable triangulation of evidence.

Difference in Differences

Decomposing Regression Triple-Differences with Staggered Adoption Anton Strezhnev*
Anton Strezhnev,

The triple-differences (TD) design is a popular identification strategy for causal effects in settings where researchers do not believe the parallel trends assumption of conventional difference-in-differences (DD) is satisfied. TD designs augment the conventional 2×2 DD with a “placebo” stratum – observations that are nested in the same units and time periods as the DD but are known to be entirely unaffected by the treatment. However, many TD applications go beyond this simple 2x2x2 setting and use observations on many units across multiple time periods and with many “placebo” strata. A popular estimator is the “triple-differences” regression (TDR) fixed-effects estimator – an extension of the common “two-way fixed effects” estimator for DD. This paper decomposes the TDR estimator into its component two-group/two-period/two-strata triple-differences and shows that interpreting this parameter causally in settings with arbitrary staggered adoption requires strong assumptions of homogeneity in the treatment effect not only over time but also across strata. Moreover, under certain forms of treatment staggering, the regression triple-differences estimator no longer consists exclusively of 2x2x2 triple-differences but rather a mixture of triple- and double-differences, the latter of which are valid only if parallel trends holds. The decomposition illustrates the importance of being cautious when implementing triple-differences designs in settings more complex than the 2x2x2

Difference in Differences

Impacts of Gun Violence on Health: A Risk Set-Matched Difference in Differences Study

Eric Cohn* Eric Cohn, Zirui Song, Jose Zubizarreta,

Injury due to firearms is a leading cause of morbidity in the United States, though research into its health and economic effects, both on survivors and those in their communities, is limited.
In this paper, we use a large longitudinal data set of health care claims to estimate the effects of nonfatal gun violence on the health and health care spending of those injured and their family members.
We obtain causal estimates from these observational data by profile matching each exposed unit to a set of unexposed units with similar covariate values prior to the exposure, that is, by risk set matching.
A difference-in-differences identification strategy allows for the presence of unobserved, time-invariant confounding, which is likely in this application, and we perform inference using design-based (i.e., randomization-based) procedures.
Additionally, we explore the presence of treatment effect heterogeneity by both prespecified subgroups and those discovered from the data, also within the difference-in-differences and design-based frameworks.
Finally, all inferences are subject to sensitivity analyses to assess the robustness of any statistically significant findings to unobserved confounding.

Evidence Factors and Causal Inference

Statistical and Causal Robustness for Causal Null Hypothesis Tests Rohit Bhattacharya* Rohit Bhattacharya, Junhui Yang, Ted Westling,

Applications of semiparametric theory to causal inference typically focus on deriving estimators that exhibit statistical robustness under a prespecified causal model that permits identification of a desired causal parameter. However, a fundamental challenge is correct specification of such a model, which usually involves making untestable assumptions. Often, an analyst might consider multiple plausible causal models given a single observed dataset. Evidence factors have recently been proposed as an approach to combining hypothesis tests of a common causal null hypothesis under two or more candidate causal models. Under certain conditions, this yields a test that is valid if at least one of the underlying models is correct, which is a form of causal robustness. In this talk, we present a method that combines semiparametric theory with evidence factors. We propose a causal null hypothesis test based on joint asymptotic normality of k asymptotically linear semiparametric estimators, where each estimator is based on a distinct identifying functional derived from each of k candidate causal models. We show that this test provides both statistical and causal robustness in the sense that it is valid if at least one of the k proposed causal models is correct, while also allowing for slower than parametric rates of convergence in estimating nuisance parameters. We conclude by discussing relative advantages and disadvantages of the proposed method.

Generalizability/Transportability

Statistical Methods for Transporting the Effect of an Environmental Mixture Across Populations Melanie Mayer* Melanie Mayer, Adele Ribeiro, Brent Coull, Ana Navas-Acien, Elias Bareinboim, Linda Valeri,

We are constantly exposed to multiple environmental exposures which work together to cause unique outcomes. To estimate their effect, observational data from a different population from which we are interested in is often what is available. Transportability has received considerable attention, however, complexities associated with multiple, continuous exposures, as we see in environmental mixtures, require one to take extra precautions. We formalize the assumptions needed and assess methods for testable assumptions, such as overlap of exposure levels and effect modifiers across populations. We also develop a statistical approach for estimating a transported effect via a flexible, machine learning model with weighting which accounts for potential non-linear/interaction effects and skewness/correlations of exposures. We demonstrate its applicability by providing a real-world application example where we analyze the effect of exposure to multiple metals on a health outcome in a multisite cohort study. We apply the methods to explain heterogeneities across study sites and to transport effects to a target population to which we observe covariate and exposure data, but not necessarily outcome information. The developed environmental mixtures transportability framework adjusts effects based on a population’s exposure/covariate distributions, expanding the data sources available for estimating environmental mixture effects for a target population of interest.

Generalizability/Transportability

Further Results On Selected Data Models For Data Fusion: Identification And Estimation

Jaron Lee* Jaron Lee, AmirEmad Ghassami, Ilya Shpitser,

Recently, there has been significant interest in graphical models for causal data fusion, whereby graphical causal inference is conducted from a collection of interventional and observational data sources. Despite recent advances, there continues to be a significant gap in extending this framework to real-world problems due to a lack of efficient estimation methodology for graphical data fusion models.

Previously, we proposed a new graphical model called the labelled conditional acyclic directed mixed graph (L-CADMG) suited for reasoning and working with such problems, which generalized existing data fusion diagrams by explicitly introducing a selection variable that indicates the selection mechanism among the different domains, and allows for cases that selection does not happen completely at random.

Building on this work, we offer two improvements.

Our first main contribution is proposing an efficient estimation methodology:

We define a model through a parameterization property, using the Mobius function which relates a set of parameters to a probability distribution. We also define a model through a factorization property, which relates the L-CADMG and its reachable subgraphs (appropriately defined) to a collection of probability distributions.

We then prove equivalence of the two models based on factorization and parameterization by showing that these two sets of distributions are equal, a result which will lead directly to a well specified likelihood (at least for discre

Generalizability/Transportability

Multiply robust federated estimation of targeted average treatment effects Zhu Shen* Zhu Shen, Larry Han, Jose Zubizarreta,

Multi-site studies have advantages including increased generalizability, the ability to study underrepresented populations, and the opportunity to study rare exposures and outcomes. However, these studies are challenging due to the need to preserve the privacy of each individual’s data and the heterogeneity in their covariate distributions, treatment guidelines, and conditional outcome models. We propose a multiply robust federated estimator to derive valid causal inferences for a target population using multi-site observational data. Our proposed estimator is more flexible than current methods as it relaxes the requirement of homogeneous model specifications across sites; investigators from different sites can leverage site-specific knowledge such as patient preferences and treatment guidelines to propose multiple different models for the outcome and treatment. Our proposed estimator adopts a model mixing approach and is consistent for the target average treatment effect if either one of the outcome models is correctly specified or one of the propensity score models and the density ratio model are correctly specified. The estimator data-adaptively down-weights source sites that are sufficiently different from the target site to avoid the negative transfer. Numerical evaluations show that our estimator can produce comparable results to doubly robust federated estimators when models are correctly specified and offer more robustness when models are misspecified.

Generalizability/Transportability

Policy Learning under Biased Sample Selection Roshni Sahoo* Roshni Sahoo, Lihua Lei, Stefan Wager,

The empirical risk minimization approach to data-driven decision making assumes that the target distribution we want to deploy our decision rule on is the same distribution that our training set is drawn from. However, in many cases, we may be concerned that our training sample is biased, and that some groups (characterized by observable or unobservable covariates) are under- or over-represented relative to the target population, so empirical risk minimization over the training set may fail to yield rules that perform well at deployment. We propose a model of sampling bias called $Gamma$-biased sampling, where observed covariates can affect the probability of sample selection arbitrarily much but the amount of unexplained variation in the probability of sample selection is bounded by a constant factor. Applying the distributionally robust optimization framework, we propose a method for learning a decision rule that minimizes the worst-case risk incurred under the family of distributions that can generate the training distribution under $Gamma$-biased sampling. We apply a result of Rockafellar and Uryasev to show that this problem is equivalent to an augmented convex risk minimization problem. We give statistical guarantees for learning a robust model using the method of sieves and propose a deep learning algorithm whose loss function captures our robustness target. We empirically validate our proposed method in simulations and a case study on ICU length of stay prediction.

Generalizability/Transportability

Generalizing Text Experiments to Real-World Contexts Victoria Lin* Victoria Lin, Eli Ben-Michael, Louis-Philippe Morency,

As natural language processing systems are increasingly deployed in real-world settings, it is important to understand how changes in language affect how readers think, feel, and behave. For instance, we may be interested in how varying the tone of a passage of text affects a reader's response. Randomized text experiments, where texts are randomly assigned to readers, offer a way to estimate such causal effects free from confounding factors. However, the effect of a text attribute depends on its context: the many other attributes of the text (e.g., topic, formality). Therefore, an effect estimated over a randomized—and necessarily artificial—body of text may not be generalizable to real-world settings. To address this issue, we propose a stochastic intervention framework for generalizing experimental text estimates to any new corpus or distribution of texts. We introduce an empirical importance-weighting estimation approach that leverages large language models to generate robust estimates of real-world text distributions. Through several empirical studies, we demonstrate that the effect of the same text attribute on reader response can in fact vary substantially from one text setting to another.

Heterogeneous Treatment Effects

Designing randomized control trials for heterogeneous populations in social complex networks Eaman Jahani* Blas Kolic, Eaman Jahani,

Write a two-paragraph research abstract about Designing randomized control trials for heterogeneous populations in social complex networks with multiple treatments.
1) comparing different clustering methods in the context of RCTs under different exposure models, 2) modifying x-net to better deal with complex social networks (Twitter in our case), and 3) introducing multiple treatment arms for heterogeneous populations

Heterogeneous Treatment Effects

A Statistical Reinforcement Learning Approach to Personalize Renal Replacement Therapy Initiation in the ICU François Grolleau* François Grolleau, François Petit, Stéphane Gaudry, Raphaël Porcher,

Trials sequentially randomizing patients each day have never been conducted for renal replacement therapy (RRT) initiation. We used data from electronic health records and clinical trials to learn and validate optimal dynamic strategies for RRT initiation in the ICU.

We included participants from the MIMIC-III database for development and two randomized trials for evaluation. We used a doubly-robust dynamic treatment regimen estimator to learn when to start RRT after the occurrence of acute kidney injury. The decision rule to initiate RRT mimicked that of clinicians i.e., decisions are re-evaluated every day—for three days in a row—given patients' evolving characteristics. The 'crude strategy' aimed to maximize hospital-free days at day 60. The 'stringent strategy' recommended initiating RRT only when there was evidence at the 0.05 threshold that a patient would benefit from initiation. For evaluation, we estimated the causal effects of implementing our learned strategies versus following current best practices using the advantage doubly robust estimator with terminal states.

We showed that implementing our strategies could improve the average number of days ICU patients spend alive and outside the hospital. The stringent strategy entailed less frequent usage of RRT and could help save important health resources all the while reducing unnecessary treatment burdens. We developed a practical and interpretable dynamic decision support system for RRT initiation in the ICU.

Heterogeneous Treatment Effects

Image-based Treatment Effect Heterogeneity Connor Jerzak* Connor Jerzak, Fredrik Johansson, Adel Daoud,

RCTs are considered the gold standard for estimating the average treatment effect (ATE) of interventions. One important use of RCTs is to study the causes of global poverty–a subject cited in the 2019 Nobel Memorial Prize for Economics awarded for the “experimental approach to alleviating global poverty.” Because the ATE is a population summary, anti-poverty experiments often seek to unpack the treatment effect variation around the ATE by conditioning (CATE) on tabular variables such as age and ethnicity that were measured during the RCT data collection. Although such variables are key to unpacking CATE, using only such variables may fail to capture historical, geographical, or neighborhood-specific contributors to effect variation, as tabular RCT data are often only observed near the time of the experiment. In global poverty research, when the geographical location of the experiment units is approximately known, satellite imagery can provide a window into such historical and geographical factors important for understanding heterogeneity. However, there is no causal inference method that specifically enables applied researchers to analyze CATE from images. In this paper, using a deep probabilistic modeling framework, we develop such a method that estimates interpretable clusters of images with similar treatment effects distributions. Our method also identifies image segments contributing to the effect cluster prediction. We apply the method to a real anti-poverty experiment.

Heterogeneous Treatment Effects

Comparing Machine Learning Methods for Estimating Heterogeneous Treatment Effects by Combining Data from Multiple Randomized Controlled Trials Carly Brantner* Carly Brantner,

Individualized treatment decisions can improve health outcomes, but using data to make these decisions in a reliable, powerful, and generalizable way is challenging with a single dataset. Leveraging multiple randomized controlled trials allows for the combination of datasets with unconfounded treatment assignment to improve the power to estimate heterogeneous treatment effects. In this study, we discuss several non-parametric approaches for estimating heterogeneous treatment effects using data from multiple trials. We then compare different single-study methods and different ways of aggregating those to the multi-trial setting through a simulation study, with data generation scenarios that have differing levels of cross-trial heterogeneity. The simulations demonstrate that methods that directly allow for heterogeneity of the treatment effect across trials perform better than methods that do not, and that the choice of single-study method matters based on the functional form of the treatment effect. We determine which methods perform well in each setting and then apply the approaches to three randomized controlled trials comparing the effects of treatments for major depressive disorder.

Heterogeneous Treatment Effects

On counterfactual inference with unobserved confounding Abhin Shah* Abhin Shah, Raaz Dwivedi, Devavrat Shah, Gregory Wornell,

Given an observational study with n independent but heterogeneous units, our goal is to learn the counterfactual distribution for each unit using only one p-dimensional sample per unit containing covariates, interventions, and outcomes. Specifically, we allow for unobserved confounding which introduces statistical biases between interventions and outcomes as well as exacerbates the heterogeneity across units. Modeling the underlying joint distribution as an exponential family, we reduce learning the unit-level counterfactual distributions to learning n exponential family distributions with heterogeneous parameters and only one sample per distribution. We introduce a convex objective that pools all n samples to jointly learn all n parameters and provide a unit-wise mean squared error bound that scales linearly with the metric entropy of the parameter space. For example, when the parameters are s-sparse linear combination of k known vectors, the error is $O(slog k/p)$. En route, we derive sufficient conditions for compactly supported distributions to satisfy the logarithmic Sobolev inequality. As an application, our results enable consistent imputation of missing covariates when they are sparse.

Heterogeneous Treatment Effects

Causal identification with subjective outcomes Leonard Goff* Leonard Goff,

Many survey questions elicit responses on ordered scales for which the definitions of the categories are subjective, possibly varying by individual. This paper clarifies what is learned when these subjective reports are used as an outcome in regression-based causal inference. When a continuous treatment variable is statistically independent of both i) potential outcomes; and ii) heterogeneity in reporting styles, a nonparametric regression of numerical subjective reports on that variable uncovers a positively-weighted linear combination of local causal responses, among individuals who are on the margin between adjacent response categories. Though the weights do not integrate to one, the ratio of local regression derivatives with respect to two such explanatory variables identifies the relative magnitudes of convex averages of their causal effects. When results are extended to discrete regressors (e.g. a binary treatment), different weighting schemes apply to different regressors, making a comparison of their magnitudes more difficult. I obtain a partial identification result for ratios that holds when there are many categories and individual reporting functions are linear. I also provide results for identification using instrumental variables.

Heterogeneous Treatment Effects

Flexible estimation of heterogeneous area-specific causal treatment effects Katarzyna Reluga* Katarzyna Reluga, Dehan Kong, Setareh Ranjbar, Nicola Salvati, Mark van der Laan,

Small area estimation (SAE) methodology is popular for estimating parameters at the level of subpopulations with small sample sizes. It is often of interest for statistical offices to quantify the effect of an intervention at the small area level, for example, the effect of a new policy targeting unemployment across counties in a region. We develop new procedures for the estimation of heterogeneous area-specific average treatment effects at the small area level in observational studies. In particular, we use nonparametric machine learning methods which are ubiquitous in causal inference as they excel in terms of modelling flexibility and predictive abilities. We compare the empirical performance of our new estimators with the parametric alternatives.

Heterogeneous Treatment Effects

A causal mediation framework for examining treatment effect heterogeneity in longitudinal studies Hanna Kim* Hanna Kim, Jee-Seon Kim,

Heterogeneity in treatment effects have been investigated as a means to obtain contextual information on how a treatment works. In longitudinal studies where treatments are provided for multiple time periods and individuals may participate with different patterns over time, variability in such treatment participation patterns can be an important factor contributing to treatment effect variability. In this study, we propose to conceptualize the effects of participating in the national Head Start program with various patterns from age three to four on children's cognitive development as causal mediation estimands. Considering that Head Start attendance of the Head Start Impact Study (HSIS) participants was only randomized in the first year and not in the second year, causal effects such as the benefit of attending Head Start one year earlier at age three in addition to attending it at age four need to be defined as corresponding causal estimands, which is in this case the controlled direct effect of Head Start at age three given Head Start at age four. Directed acyclic graphs (DAGs) are presented to describe confounders and identification assumptions specific to each research question. Estimation methods are applied to the HSIS data to illustrate that the causal mediation framework can naturally address longitudinal treatment participation patterns as a novel source of treatment effect heterogeneity and provide substantive insights into refining well-known social interventions.

Machine Learning and Causal Inference

On the estimation of the mean number of counterfactual recurrent events before failure

Benjamin Baer* Benjamin Baer, Ashkan Ertefaie, Robert Strawderman,

On any time interval, the marginal increment in the expected number of recurrent events before failure can be decomposed as a product of two terms: the survival probability evaluated at the beginning of the time interval, multiplied by the increment in the expected number of recurrent events before failure conditional on surviving to the beginning of the time interval. For a fixed set of landmark times, the expected number of recurrent events before failure can be similarly decomposed as a sum of such terms. The resulting expression can be viewed as a generalization of the cumulative incidence function arising in semi-competing risks problems, and allows one to quantify the relative contributions of the failure time and recurrent event count distributions to the expected number of recurrent events before failure.

We define our estimand as the vector comprising each function evaluated along a sequence of landmark times. We identify the estimand in the presence of right-censoring and causal selection as an observed data functional under coarsening at random, derive the nonparametric efficiency bound, and propose a multiply-robust estimator that achieves the bound and permits nonparametric estimation of nuisance parameters. Additionally, we derive the class of influence functions when the coarsening distribution is known and frame previously published estimators as belonging to the class. Along the way, we highlight some inconsistencies in the causal survival literature.

Machine Learning and Causal Inference

Controlling Label Noise in Deep Classifiers via Causal Regularization Juan Castorena* Juan Castorena,

The explosion in data availability, computational power and scientific advances have enabled the development of deep learning (DL) methods at the forefront of every major AI advancement of the last decade. In the supervised setting, state of the art DL classifiers rely on training based on a very large dataset of clean, well annotated samples. Unfortunately, this type of effort is impractical for many real-world applications and often practitioners need to accept labeling errors as an inescapable fact. Here, we propose to address the problem of entanglement between the clean and noisy conditional distributions through a more principled approach of analysis under the lens of causality. We encode the data generative process of a variety of label noise source settings via structural causal models and analyze using the rules of the do-calculus and transportability theory the conditions for identifiability, ultimately leading to disentanglement. Empirical evidence on synthetically corrupted benchmark datasets shows that deep classifier models equipped with identifiability constraints imposed here via learning loss regularizations are more robust to out-of-distribution examples while also producing more interpretable features in contrast to their causally-unconstrained counterparts.

Machine Learning and Causal Inference

Applying causal inference in architectural engineering – thermal comfort model and control Ruiji Sun* Ruiji Sun, Stefano Schiavon, Hui Zhang, Lei Shi, Thomas Parkinson,

A rapidly growing volume of data from continuous building monitoring has been available due to their electrification and digitalization. The architectural and engineering community widely uses machine learning algorithms to transform high-dimensional data into models with high prediction accuracy. However, the main shortcoming of traditional machine learning techniques is their black-box nature which limits causal effect understanding. Understanding causality from continuous indoor environmental quality (IEQ) monitoring data would help create a sustainable, resilient, and healthy built environment. We propose a causal inference framework to study the long-term effect of IEQ on occupants' comfort, health, and performance. We first developed a collaborative causal graph across multi-disciplines to summarize exited knowledge of IEQ. We investigated one aspect of IEQ called thermal comfort, which is the subjective perception of the thermal environment. The IEQ causal graph shows the drawbacks of various classical thermal comfort models. We identify the indoor thermal environment's heterogeneous effects on long-term thermal comfort using d-separation. The structural causal model can be implemented in building mechanical systems to control the indoor environment within a comfortable temperature range.

Machine Learning and Causal Inference

A Meta-Learning Method for Estimation of Causal Excursion Effects to Assess Time-Varying Moderation Jieru Shi* Jieru Shi, Walter Dempsey,

Twin revolutions in wearable technologies and digital health interventions have significantly expanded the accessibility and uptake of mobile health (mHealth) interventions in multiple domains of health sciences. Sequentially randomized experiments called micro-randomized trials (MRTs) have grown in popularity as a means to empirically evaluate the effectiveness of these mHealth intervention components. MRTs have motivated a new class of causal estimands, termed “causal excursion effects”, which allows health scientists to assess how intervention effectiveness changes over time or is moderated by individual characteristics, context, or responses in the past. However, current data analysis methods require pre-specified features of the observed high-dimensional history to construct a working model of an important nuisance parameter. Machine learning (ML) algorithms are ideal for automatic feature construction, but their naive application to causal excursion estimation can lead to bias under model misspecification and therefore incorrect conclusions about the effectiveness of interventions. In this paper, the estimation of causal excursion effects is revisited from a meta-learner’s perspective, where ML and statistical methods such as supervised learning and regression have been explored. Asymptotic properties of the novel estimands are presented and a theoretical comparison accompanied by extensive simulation experiments demonstrates relative efficiency gains.

Machine Learning and Causal Inference

Tailored Overlap for Observational Causal Inference and Domain Adaptation David Bruns-Smith* Avi Feller, Alexander D’Amour, Steven Yadlowsky,

In observational causal inference and predictive modeling under distribution shift, assumptions about overlap between treatment/covariate groups or training/test distributions are critical for identifying causal effects or finding an optimal predictive model. Standard theory quantifies overlap in terms of bounds on the inverse propensity score, which are typically measured using chi^2 divergences. However, in modern settings with high-dimensional covariates, these standard divergence measures are often infinite. In this paper, we propose a new approach to measuring overlap that is tailored to a specific function class, which allows us to better capture the relationship between the treatment and outcome or between covariates and target variable. We show how chi^2 divergences can be generalized to this restricted function class setting, and use this to motivate more widespread use of balancing weight-based methods, which adjust the relative influence of different observations in the training data. These methods allow us to more accurately identify causal effects and optimal predictors, even in settings with high-dimensional covariates and limited overlap.

Machine Learning and Causal Inference

Modeling the Data-Generating Process is Necessary for Out-of-Distribution Generalization

Emre Kiciman* Emre Kiciman, Jivat Neet Kaur, Amit Sharma,

Recent empirical studies on domain generalization (DG) show that DG algorithms that perform well on some distribution shifts fail on others, and no
state-of-the-art DG algorithm performs consistently well on all shifts. Moreover, real-world data often has multiple distribution shifts over different attribute. In multi-attribute distribution shift datasets, we find that the accuracy of existing DG algorithms falls further. To explain this, we provide a formal characterization of generalization under multi-attribute shifts using a canonical causal graph. Based on the relationship between spurious attributes and the classification label, we obtain realizations of the canonical causal graph that characterize common distribution shifts and show that each shift entails different independence constraints over observed variables. We prove that any algorithm based on a single, fixed constraint cannot work well across all shifts, providing theoretical evidence for mixed empirical results on DG algorithms. We develop Causally Adaptive Constraint Minimization (CACM), an algorithm that uses knowledge about the data-generating process to adaptively identify and apply the correct independence constraints for regularization. Results on broad set of benchmarks show that adaptive dataset-dependent constraints lead to the highest accuracy on unseen domains whereas incorrect constraints fail, demonstrating the importance of modeling the causal relationships of the data-generating process.

Machine Learning and Causal Inference

Penalized Minimax Instrumental Variable Estimation with General Function Approximation

Masatoshi Uehara* Masatoshi Uehara, Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, Masatoshi Uehara,

We investigate instrumental variable estimation, which involves finding a solution for the equation E[Y-h(X)|Z]=0 with respect to h. It is widely utilized across a variety of fields and has seen significant progress with the use of flexible machine learning techniques. However, current approaches often have one or more of the following limitations: (1) a requirement for a unique solution, (2) an inability to accurately identify the solution with L2 rate guarantees, (3) a need for the conditional expectation operator to have smoothness in addition to the widely-accepted source condition, also known as the closedness assumption. We present the first method that avoids all of these limitations. Specifically, we propose a new minimax algorithm that has an L2 convergence guarantee under the source condition and realizability assumptions, and does not require uniqueness. This is achieved by viewing the least norm solution of E[Y-h(X)|Z]=0 as a saddle point of certain constrained minimax optimization problems, a novel perspective not taken in prior works such as Dikkala et al. 2020 and Liao et al. 2020. Finally, we demonstrate the applicability of our proposed estimator and theory to policy learning in the infinite horizon setting, a goal of great interest in fields such as mobile health and robotics.

Machine Learning and Causal Inference

Latent distribution estimation for the evaluation of the complier average causal effect

Celine Beji* Celine Beji, Raphaël Porcher,

The complier average causal effect (CACE) estimator, defined as the average of potential outcomes in the latent sub-population that complies with their assigned treatment, is more and more used in clinical trials to study the effect of a medication or an intervention rather than the effect of its prescription. Although advanced methods such as instrumental variables and G-estimation have been developed, it requires strong assumptions of exclusion restriction and principal ignorability.

We propose a new approach to CACE estimation, in the vein of principal stratification framework, that does not require these assumptions. We estimate the latent distribution of four relevant groups of individuals: compliers, never-takers, always-takers and defiers. We reframe the problem as a missing data problem and introduce a two-step procedure that estimates CACE via the latent distribution of the principal strata. We study the estimator sensitivity to assumptions using Monte Carlo simulations, for three different estimation methods using a mixture of experts, a revisited Expectation-Maximization algorithm and a neural network. We apply this approach on randomized trials which evaluate the effect of an individualized education program on risk factors reduction after an acute coronary syndrome.

Machine Learning and Causal Inference

Near-Optimal Non-Parametric Sequential Tests and Confidence Sequences with Possibly Dependent Observations Michael Lindon* Aurelien Bibaut, Nathan Kallus, Michael Lindon,

Sequential testing, always-valid p-values, and confidence sequences promise flexible statistical inference and on-the-fly decision making. However, unlike fixed-n inference based on asymptotic normality, existing sequential tests either make parametric assumptions and end up under-covering/over-rejecting when these fail or use non-parametric but conservative concentration inequalities and end up over-covering/under-rejecting. To circumvent these issues, we sidestep exact at-least-$alpha$ coverage and focus on asymptotically exact coverage and asymptotic optimality. That is, we seek sequential tests whose probability of ever rejecting a true hypothesis asymptotically approaches $alpha$ and whose expected time to reject a false hypothesis approaches a lower bound on all tests with asymptotic coverage at least $alpha$, both under an appropriate asymptotic regime. We permit observations to be both non-parametric and dependent and focus on testing whether the observations form a martingale difference sequence. We propose the universal sequential probability ratio test (uSPRT), a slight modification to the normal-mixture sequential probability ratio test, where we add a burn-in period and adjust thresholds accordingly. We show that even in this very general setting, the uSPRT is asymptotically optimal under mild generic conditions. We apply the results to stabilized estimating equations to test means, treatment effects, etc. Our results also provide corresponding guarantees

Machine Learning and Causal Inference

Conceptualizing Treatment Leakage in Text-based Causal Inference Connor Jerzak* Adel Daoud, Richard Johansson, Connor Jerzak,

Causal inference methods that control for text-based confounders are becoming increasingly important in the social sciences and other disciplines where text is readily available. However, these methods rely on a critical assumption that there is no treatment leakage: that is, the text only contains information about the confounder and no information about treatment assignment. When this assumption does not hold, these text-based adjustment methods face the problem of post-treatment (collider) bias. The assumption that there is no treatment leakage may be unrealistic in real-world situations involving text, as human language is rich and flexible. Language appearing in a public policy document or health records may refer to the future and the past simultaneously, and thereby reveal information about the treatment assignment. In this talk, we define the treatment-leakage problem, and discuss the identification as well as the estimation challenges it raises. Second, we delineate the conditions under which leakage can be addressed by removing the treatment-related signal from the text in a pre-processing step we define as text distillation. Lastly, using simulation, we show how treatment leakage introduces a bias in estimates of the average treatment effect (ATE) and how text distillation can mitigate this bias.

Machine Learning and Causal Inference

Proportional Response: Contextual Bandits for Simple and Cumulative Regret Minimization Sanath Kumar Krishnamurthy* Sanath Kumar Krishnamurthy, Ruohan Zhan, Susan Athey, Emma Brunskill,

We design adaptive experimentation (contextual bandit) algorithms to learn a targeted treatment assignment policy. Our design focuses on two important objectives: optimizing outcomes for the subjects in the experiment (cumulative regret minimization) and gathering data that will be useful for learning a policy with high expected reward (simple regret minimization). Recently, [Li et al 2022] formally showed that it is impossible for an algorithm to simultaneously achieve minimax optimal cumulative regret guarantees and instance optimal simple regret guarantees. They also proposed a novel algorithm that is the first to achieve instance-optimal simple regret guarantees, but is impractical to use due to computational considerations (their exploration policy relies on explicitly constructing a distribution over a potentially large set of policies). To address these issues, we propose a novel family of contextual bandit algorithms that are simple (easy to implement), flexible (works with any model and policy class), statistically efficient, and computationally efficient. Our family of algorithms comes with a parameter that allows for flexible trading off between guarantees on the two competing objectives, simple and cumulative regret minimization. Technically, our algorithm is enabled by the construction of a set of arms at every context that are probabilisticaly reliable over the context distribution to contain the arm recommended by the optimal policy.

Machine Learning and Causal Inference

Post-Episodic Reinforcement Learning Inference Ruohan Zhan* Ruohan Zhan, Vasilis Syrgkanis,

We consider estimation and inference with data collected from episodic reinforcement learning (RL) algorithms; i.e. adaptive experimentation algorithms that at each period (aka episode) interact multiple times in a sequential manner with a single treated unit. Our goal is to be able to evaluate counterfactual adaptive policies after data collection and to estimate structural parameters such as dynamic treatment effects, which can be used for credit assignment (e.g. what was the effect of the first period action on the final outcome). Such parameters of interest can be framed as solutions to moment equations, but not minimizers of a population loss function, leading to Z-estimation approaches in the case of static data. However, such estimators fail to be asymptotically normal in the case of adaptive data collection. We propose a re-weighted Z-estimation approach with carefully designed adaptive weights to stabilize the episode-varying estimation variance, which results from the nonstationary policy that typical episodic RL algorithms invoke. We identify proper weighting schemes to restore the consistency and asymptotic normality of the re-weighted Z-estimators for target parameters, which allows for hypothesis testing and constructing reliable confidence regions for target parameters of interest. Primary applications include dynamic treatment effect estimation and dynamic off-policy evaluation.

Matching

Covariate-adaptive randomization inference in matched designs Samuel Pimentel* Samuel Pimentel, Ruoqi Yu,

After matching treated units to controls in observational data, it is common to conduct inference by permuting treatment assignments as in a Fisher randomization test (FRT). This approach may fail to control Type I error. Firstly, it does not account for differences in treatment propensities between matched individuals, which typically do not agree exactly in practice. Secondly, it does not consider whether permuted versions of treatment would have led to the selection of the same matched-pair configuration. Recent proposals to update the FRT procedure using estimated propensity scores help address the first problem, but the second has received little attention. We show that treatment permutations incompatible with the matched pair configuration can lead to substantive Type I error violations and present new computationally efficient graph-based inference procedures for optimal pair matching with propensity scores that eliminate incompatible treatment assignments from consideration. We demonstrate their effectiveness via simulations and a re-analysis of an observational study of health outcomes.

Matching

Bias Correction for Randomization-Based Inference in Imperfectly Matched Observational Studies: Oracles, Practices, and Simulations Jianan Zhu* Jianan Zhu, Siyu Heng,

In causal inference, matching is one of the most widely used methods to mimic a randomized experiment with observational data. Ideally, treated subjects are perfectly matched with controls for the confounders, and randomization-based inference can therefore be conducted. However, imperfect matching is typically expected in practice, especially with continuous or many confounders. Previous imperfectly matched studies have routinely treated the downstream randomization-based inference as if the confounders were perfectly matched. In this work, we propose a bias correction framework for randomization-based inference with imperfectly matched datasets to further reduce confounding bias after matching. First, we derive the unbiased oracle randomization-based inference procedure with imperfect matching under oracle propensity scores. Second, based on the derived oracle inference procedure, we give some practical proposals for bias correction for imperfect matching by replacing oracle propensity scores with estimated ones. Third, we conduct simulation studies to compare the performances of various practices of randomization-based inference, including the previous practice of ignoring imperfect matching and our proposed practices of bias correction for imperfect matching. Our framework works for most existing matching designs and covers both Fisher’s sharp null and Neyman’s weak null.

Matching

Assessing Gender Disparities in Textual Response on Reddit.com Zhiyu Guo* Zhiyu Guo, Zach Branson, Reagan Mozer,

The age of social media has fostered many different online communities with distinct cultures, some with more serious gender disparity issues than others. For example, men and women may receive different reactions to posts they make in online forums, due entirely to their gender. But one can argue that the style and topics that men and women talk about on the forum might differ, thereby making it unclear if upvote differences are due to gender or other confounding factors between male and female posters. In this work, we assess whether posting as a male or female causes a change in upvotes to posts made on the /r/relationships subreddit of Reddit.com, a popular forum website. We look at all posts with self-label F or M in 2015. We combine topic modeling, sentiment analysis, and other state-of-the-art text quantification methods with both propensity score-based and cardinality matching methods to address these possible confounding effects. We compare several estimators for the average difference in upvotes between men and women: outcome regression estimators, inverse propensity score weighted estimators, and nonparametric doubly robust estimators both before and after matching. We find that matching followed by doubly robust estimation allows for a flexible analysis that also makes a fundamental causal inference assumption – positivity – more tenable. We discover that on r/relationships, women tend to get a slightly higher number of upvotes than men.

Mediation

The central role of the mediator process in mediation analysis Caleb Miles* Caleb Miles,

Traditionally, mediation analysis has involved analysis of exposure, mediator, and outcome, each observed at sequential discrete points in time. The natural direct and indirect effects are then defined based on these three time points. Identification relies on the assumption that no effect of the exposure can cause both the mediator and the outcome. However, the mediator of interest will often be a stochastic process varying from baseline to follow-up, and its value observed at an individual point in time but a coarse measurement of this process. When the intermediate variable is a mediator, I will argue that earlier instances of the intermediate variable will often be exposure-induced confounders of the mediator at its observed time. Thus, the mediated effects defined in terms of the coarsened mediator process will not be identified. Further, I will argue that the mediated effects of greatest substantive interest are those involving the full mediator process, and that the coarsened mediator process effects can have nonsensical interpretations. To make progress, one must instead rely on strong exclusion restriction assumptions or account for the full mediator process. Lastly, I will discuss an effect decomposition relating the full mediator process indirect effect to the coarsened mediator process indirect effect.

Mediation

Mediation Analysis for Case-Control Studies with Secondary and Tertiary Outcomes Zichun Cao* Zichun Cao, Honglei Chen, Chenxi Li, Yuehua Cui, Aimee D'Aloisio, Dale Sandler, Zhehui Luo,

Background: A case-control study is an epidemiological study design usually for studying the association of risk factors for a rare disease. It is not uncommon that researchers use the case-control sample for additional secondary or tertiary outcomes of interest. Statistical methods for association analysis using such secondary or tertiary outcomes have been proposed but the literature is limited for mediation analyses using such data. With the increasing amount of secondary or tertiary data, plus the fact the prospective cohort study and randomized controlled trials are expensive for mediation analysis, there is a need for using small-scale and less costly data to do preliminary mediation analysis.

Goal: To perform mediation analysis for a target population (e.g., a cohort) using case-control samples with additionally collected variables.

Methods: We proposed a new weighting method for estimating mediating effects (controlled direct effects and natural direct and indirect effects) using case-control data with a further collected (secondary) variable as the mediator and another tertiary variable collected even later as the outcome. And we also implemented the proposed estimation method in an empirical application using the NIEHS Sister Study for illustration.

Results: The proposed estimators performed very well in the Monte Carlo simulations. The implementation of the method is easy and smooth.

Missing data

Missing not at random: a cross-sectional model Anna Guo* Anna Guo, Razieh Nabi, Jiwei Zhao,

Conducting valid statistical inference is challenging in the presence of nonignorable missing mechanisms, often referred to as missing-not-at-random or MNAR for short. In this work, we consider a MNAR missingness mechanism, most appropriate in cross-sectional studies. This model is a supermodel of several popular models, including permutation model (Robins 1997), block-conditional MAR model (Zhou et al. 2010), and block-parallel model (Mohan et al. 2013), thus making it less stringent in terms of underlying statistical assumptions. The underlying complete-data distribution in our cross-sectional MNAR model is not nonparametrically identifiable from the partially observed data. We establish sufficient conditions for identification of the target law by restricting our attention to the exponential family distributions. Unlike most prior work, all variables in our model can be subject to missingness, i.e., our results do not rely on the presence of fully observed variables. Borrowing the graphical model toolkit, we propose methods for testing the independence restrictions encoded in our model. If the test result suggests further independence restrictions in the model, we show that the model is nonparametrically identifiable. We also adopt a conditional likelihood approach for independence tests via estimating pairwise odds ratios. Further, we suggest the generalized method of moments for target law estimation. Statistical properties of the estimators are established.

Missing Data and Self-Censoring

Causal Inference Under Self-Censoring Treatment and Outcome Jacob M. Chen* Jacob M. Chen, Daniel Malinsky, Rohit Bhattacharya,

“Self-censoring” is a type of missingness-not-at-random (MNAR) phenomenon that poses a particularly difficult obstacle to valid inference. If treatment and/or outcome directly determine their own missingness, this may lead to severely biased estimates for the causal effect. The possibility of unmeasured confounding in conjunction with self-censoring also complicates identifying a valid set of covariates to adjust for. Shadow variables, proposed by Miao et al. (2015), are auxiliary variables that can be used to overcome self-censoring if certain conditions are met. However, these conditions are difficult to verify: that (i) the proposed shadow variable is associated with the censored variable and (ii) it does not directly affect the missingness of the censored variable. Here, we extend prior work in covariate selection to account for both self-censoring and unmeasured confounding. We propose a two-stage test; the first stage confirms dependence between a pre-treatment variable and the missingness indicator of the outcome after conditioning on some subset of observed covariates, and the second stage confirms independence between the same variables while additionally conditioning on the treatment. We prove that if the test passes, then the pre-treatment variable is indeed a valid shadow variable while the observed covariates are a valid backdoor adjustment set. Using this information, we propose an inverse probability weight-based estimator for the causal effect of interest.

Multilevel Causal Inference

Controlling family-wise error while testing tree-structured hypotheses about treatment effects Ajinkya H. Kokandakar* Ajinkya H. Kokandakar, Sameer K. Deshpande,

We describe two applications that involve testing hypotheses about hierarchically organized treatment effects. In the first application, we wanted to determine the effect of playing organized youth sports, a treatment condition that could be defined at multiple levels of resolution (e.g. playing any sport, a collision sport, or a particular sport). The second application involved testing for significant differences between treatment effects across subgroups discovered in a data-driven manner. In both problems, hypotheses can be arranged in a binary tree such that if a parent hypothesis is false, at most one of its two children hypotheses can be true. We develop an ordered testing procedure that exploits these logical relationships to control family-wise error rate (FWER). Briefly, we keep testing along a path in the tree until we reach the first failed rejection of a null hypothesis. The key innovation in our approach is the allocation of the significance level for individual tests in a way that maintains FWER across any configuration of true hypotheses. Our procedure allows us to adaptively determine the resolutions of treatment for which there are significant effects. Similarly, we can adaptively identify subgroups exhibiting significant treatment effect differences.

Multilevel Causal Inference

Assessing Informative Cluster Size in Cluster-Randomized Trials Bryan Blette* Bryan Blette, Brennan Kahan, Michael Harhay, Fan Li,

In cluster-randomized trials, the average treatment effect among participants (p-ATE) may be different from the cluster average treatment effect (c-ATE) when informative cluster size is present, i.e., when treatment effects or participant outcomes depend on cluster size. In such scenarios, mixed-effects models and GEEs with an exchangeable correlation structure are biased for both the p-ATE and c-ATE estimands, and GEEs with an independence correlation structure or analyses of cluster-level summaries are recommended in practice. However, when the p-ATE and c-ATE are equivalent, mixed-effects models and GEEs with exchangeable correlation structure can provide unbiased estimation and notable efficiency gains over other methods. Thus, a hypothesis test of whether informative cluster size is present could be useful to formally assess the validity of this key assumption. In this study, we develop and compare model-assisted and randomization-based tests for informative cluster size in cluster-randomized trials. We construct simulation studies to examine the operating characteristics of these tests and contrast them to existing model-based tests for informative cluster size or “confounding by cluster” in the observational study setting. The proposed tests are applied to data from a recent cluster-randomized trial, and practical recommendations for using these tests are discussed.

Offline policy evaluation

Future-dependent value-based off-policy evaluation in pomdps Masatoshi Uehara* Masatoshi Uehara, Masatoshi Uehara, Andrew Bannet, Nan Jiang, Wen Sun, Nathan Kallus, Victor Chernozhukov,

We study off-policy evaluation (OPE) for partially observable MDPs (POMDPs) with general function approximation. Existing methods such as sequential importance sampling estimators and fitted-Q evaluation suffer from the curse of horizon in POMDPs. To circumvent this problem, we develop a novel model-free OPE method by introducing future-dependent value functions that take future proxies as inputs. Future-dependent value functions play similar roles as classical value functions in fully-observable MDPs. We derive a new Bellman equation for future-dependent value functions as conditional moment equations that use history proxies as instrumental variables. We further propose a minimax learning method to learn future-dependent value functions using the new Bellman equation. We obtain the PAC result, which implies our OPE estimator is consistent as long as futures and histories contain sufficient information about latent states, and the Bellman completeness. Finally, we extend our methods to learning of dynamics and establish the connection between our approach and the well-known spectral learning methods in POMDPs.

Propensity Scores

Handling covariate missingness in propensity score weighting with clustered data: A comparison of multiple imputation and complete-case analysis Xiao Liu* Xiao Liu,

Propensity scores (PS) are commonly used for causal inference of treatment effects. Two issues can complicate PS analysis: clustered data structure and missing data. Methods for handling either issue alone in PS analysis have been studied. Methods for PS analysis when both issues exist have been under-evaluated. This study considers PS weighting for average treatment effect estimation with clustered data where individual-level covariates contain missingness and (fully) unobserved cluster-level confounders exist. Simulation studies were conducted to compare different missing data methods (complete-case, single- or multi-level multiple imputation), PS models (random- or fixed-effects models), weighting (marginal, clustered), and outcome models (nonparametric, regression on treatment and random or fixed cluster intercepts; sandwich standard errors were used for methods with the nonparametric and fixed-effect outcome models). When the missingness was due to moderators (which was the unobserved cluster-level confounder in our simulation), complete-case analysis produced biased estimates. Single-level imputation performed well in some conditions but did not in a majority of conditions, particularly when dependence among variables used in the imputation was low. Overall, multilevel imputation, fixed-effect PS, clusterered weighting, and outcome regression on treatment and random cluster intercepts appeared perform well in bias, RMSE, and coverage across the simulation conditions.

Propensity Scores

When is the estimated propensity score better? High-dimensional analysis and bias correction Fangzhou Su* Fangzhou Su, Peng Ding, Wenlong Mou, Martin Wainwright,

Anecdotally, using an estimated propensity score is superior to the true propensity score in estimating the average treatment effect based on observational data. However, this claim comes with several qualifications: it holds only if propensity score model is correctly specified and the number of covariates d is small relative to the sample size n. We revisit this phenomenon by studying the inverse propensity score weighting (IPW) estimator based on a logistic model with a diverging number of covariates. We first show that the IPW estimator based on the estimated propensity score is consistent and asymptotically normal with smaller variance than the oracle IPW estimator (using the true propensity score) emph{if and only if} $n gtrsim d^2$. We then propose a debiased IPW estimator that achieves the same guarantees in the regime $n gtrsim d^{3/2}$. Our proofs rely on a novel non-asymptotic decomposition of the IPW error along with careful control of the higher order terms.

Randomized Studies

Design-Based Confidence Sequences for Anytime-valid Causal Inference Dae Woong Ham*
Dae Woong Ham, Iavor Bojinov, Michael Lindon, Martin Tingley,

Many organizations run thousands of randomized experiments, or A/B tests, to statistically quantify and detect the impact of product changes. Analysts take these results to augment decision-making around deployment and investment opportunities, making the time it takes to detect an effect a key priority. Currently, however, the analysis is only performed at the end of the study. This is undesirable because strong effects can be detected before the end of the study, which is especially relevant for risk mitigation when the treatment effect is negative. Alternatively, analysts could perform hypotheses tests more frequently and stop the experiment when the estimated causal effect is statistically significant, i.e., confidence sequences. Our paper provides valid confidence sequences from the design-based perspective, where we condition on the full set of potential outcomes and perform inference on the obtained sample. Our design-based confidence sequence accommodates a wide variety of sequential experiments in an assumption-light manner. In particular, we build confidence sequences for 1) the average treatment effect, 2) the reward mean difference in bandits, 3) the average contemporaneous treatment effect for time series/panel data settings with potential carryover effects. We further provide a variance reduction technique that incorporates modeling assumptions and covariates. We apply our proposed confidence sequences to experiments conducted by Netflix.

Randomized Studies

Experimental and Quasi-Experimental Identification of Conditional Average Treatment Effects: A Four-Arm Within-Study Comparison Bryan Keller* Bryan Keller, Vivian Wong, Sangbaek Park, Jingru Zhang, Patrick Sheehan, Peter Steiner,

In a largest-of-its-kind four-arm within-study comparison (WSC), we asked 2200 participants to request to receive either a mathematics or vocabulary training,
recorded their request, and then randomly assigned them to a session. In addition to mathematics and vocabulary outcomes, we collected over 30 baseline
covariates vetted via a pilot study such that we expect unconfoundedness to approximately hold. This design permits both experimental and quasi-experimental identification of the ATE, ATT, and ATC. For example, the ATT for the mathematics training intervention is experimentally identified by the group mean difference (T – C) for those who asked to be assigned to the mathematics group. A quasi-experiment based on self-selection may be created by comparing those who requested mathematics training and received it with those who requested vocabulary training and received it. A number of methods that condition on observed covariates to reduce bias are used to estimate ATEs with the quasi-experimental data (e.g., main effects ANCOVA, PS analysis, AIPW, BART as S-learner, causal forests, TMLE). The study was powered to test for correspondence between experimental and quasi-experimental estimates. A minimum detectable effect size corresponding with 80% bias reduction (RCT vs QED) is used as the equivalence threshold for correspondence testing. Planned analyses are described in detail in an OSF preregistration. Results will be discussed.

Randomized Studies

Berry-Esseen bounds for design-based causal inference with possibly diverging treatment levels and varying group sizes Lei Shi* Lei Shi, Peng Ding,

Neyman (1923/1990) introduced the randomization model, which contains the notation of potential outcomes to define causal effects and a framework for large-sample inference based on the design of the experiment. However, the existing theory for this framework is far from complete especially when the number of treatment levels diverges and the group sizes vary a lot across treatment levels. We provide a unified discussion of statistical inference under the randomization model with general group sizes across treatment levels. We formulate the estimator in terms of a linear permutational statistic and use results based on Stein’s method to derive various Berry–Esseen bounds on the linear and quadratic functions of the estimator. These new Berry–Esseen bounds serve as basis for design-based causal inference with possibly diverging treatment levels and diverging dimension of causal effects. We also fill an important gap by proposing novel variance estimators for experiments with possibly many treatment levels without replications. Equipped with the newly developed results, design-based causal inference in general settings becomes more convenient with stronger theoretical guarantees.

Randomized Studies

Model-robust and efficient covariate adjustment for cluster-randomized experiments

Bingkai Wang* Bingkai Wang, Chan Park, Dylan Small, Fan Li,

Cluster-randomized experiments are increasingly used to evaluate interventions in routine practice conditions, and researchers often adopt model-based methods with covariate adjustment in the statistical analyses. However, the validity of model-based covariate adjustment is unclear when the working models are misspecified, leading to ambiguity of estimands and risk of bias. In this article, we first adapt two conventional model-based methods, generalized estimating equations and linear mixed models, with weighted g-computation to achieve robust inference for cluster-average and individual-average treatment effects. Furthermore, we propose an efficient estimator for each estimand that allows for flexible covariate adjustment and additionally addresses cluster size variation dependent on treatment assignment and other cluster characteristics. Such cluster size variations often occur post-randomization and, if ignored, can lead to bias of model-based estimators. For our proposed estimator, we prove that when the nuisance functions are consistently estimated by machine learning algorithms, the estimator is consistent, asymptotically normal, and efficient. When the nuisance functions are estimated via parametric working models, the estimator is triply-robust. Simulation studies and analyses of three real-world cluster-randomized experiments demonstrate that the proposed methods are superior to existing alternatives.

Randomized Studies

Design based variance estimation for Hajek estimators of average causal effects in finely stratified, potentially clustered designs Xinhe Wang* Xinhe Wang, Ben Hansen,

Clustered randomized controlled trials are commonly used to evaluate the effectiveness of one or more treatments. Frequently, stratified or paired designs are adopted in practice. In this setting, the difference of intervention and control group means is properly construed as a Hajek estimator, at least when the means are appropriately computed with weights reflecting cluster sizes and inverse assignment probabilities. Despite the ubiquity of these designs and estimators, suitable design-based variance estimation does not appear to be available in the literature. Fogarty (2018, JRSS-B) gave such estimates for finely stratified designs, but not with clustering or variation in size; Schochet et al (2022, JASA) address clustering, but only for large-stratum designs. In this work, we derive the finite-population variance-covariance matrix and establish asymptotic normality. We propose asymptotically conservative sandwich variance estimators and covariance interval estimators. These estimators are suitable both for the large-stratum designs considered by Schochet and colleagues and in the many-small-strata scenarios considered by Pashley and Miratrix (2021, J. Educ. Behav. Stat.). In contrast to the analogous standard error for impacts estimated using stratum fixed effects, ours is consistently conservative even as the number of strata grows without bound, a theoretical finding supported by our simulation experiments.

Randomized Studies

Treatment Effect Quantiles in Stratified Randomized Experiments and Matched Observational Studies Yongchang Su* Xinran Li, Yongchang Su,

Evaluating the treatment effects has become an important topic for many applications. However, most existing literature focuses mainly on average treatment effects. When the individual effects are heavy-tailed or have outlier values, not only may the average effect not be appropriate for summarizing treatment effects, but also the conventional inference for it can be sensitive and possibly invalid due to poor large-sample approximations. In this paper, we focus on quantiles of individual treatment effects, which can be more robust in the presence of extreme individual effects. Moreover, our inference for them is purely randomization-based, avoiding any distributional assumption on the units. We first consider inference in stratified randomized experiments, extending the recent work by Caughey et al. (2021). We show that calculating valid p-values for testing null hypotheses on quantiles of individual effects is equivalent to solving multiple-choice knapsack problems, based on which we provide efficient algorithms to calculate the p-values exactly or slightly conservatively. We then extend our approach to matched observational studies and propose sensitivity analysis to investigate to what extent our inference on quantiles of individual effects is robust to unmeasured confounding. The proposed randomization inference and sensitivity analysis are simultaneously valid for all quantiles of individual effects. An R package has also been developed to implement the proposed methods.

Randomized Studies

Adjusting for Incomplete Baseline Covariates in Randomized Controlled Trials: A Cross-World Imputation Framework Yilin Song* Yilin Song, Ting Ye, James Hughes,

In randomized controlled trials, adjusting for baseline covariates is often applied to improve the precision of treatment effect estimation. However, missingness in covariates is common. Recently, Zhao & Ding (2022) studied two simple strategies, the single imputation method and the missingness indicator method (MIM), to deal with missing covariates, and showed that both methods can provide efficiency gain. To better understand and compare these two strategies, we propose and investigate a novel imputation framework termed cross-world imputation (CWI), which includes single imputation and MIM as special cases. Through the lens of CWI, we show that MIM implicitly searches for the optimal CWI values and thus achieves optimal efficiency. We also derive conditions under which the single imputation method, by searching for the optimal single imputation values, can achieve the same asymptotic efficiency as the MIM. We illustrate our findings through simulation studies and a real data analysis based on the Childhood Adenotonsillectomy Trial.

Randomized Studies

A tie-breaker design for pragmatic clinical trials Minh Nguyen* Minh Nguyen, Tim Morrison, Art Owen, Michael Baiocchi,

We study the problem of estimating the causal effect of being admitted to an intensive care unit (ICU) on health outcomes. We employ a tie-breaker design, introduced in Owen and Varian (2020), which limits randomization to a small window of covariate space. Similar to discontinuity designs, tie-breakers sort patients using a running variable. Those above an upper cutoff are given treatment and those below a lower cutoff are not. Those patients between the cutoffs are randomized at some fixed probability of receiving treatment.

We propose a novel, pragmatic study design which modifies the existing tie-breaker design in order to accommodate practical constraints faced by care providers. Using Electronic Health Record data, Nguyen et al. (2021) developed models to predict emergency room patients' risk for ICU admission – which we use as this study's running variable. The first modification is forced by the constraint of bed availability, which requires (a) forecasting bed availability and (b) a variable probability of receiving treatment. The second modification allows for physicians to override the study design assignment when necessary. This reframes the tie-breaker design into an encouragement trial. Both modifications afford instrumental variable-type analyses. Finally, we document how a careful analysis of the physicians' reasons for overriding the assignment to treatment level is valuable for improving clinical decision-making and understanding heterogeneous ICU benefits.

Regression Discontinuity

Quasi-experimental designs for learning health systems Amy Cochran* Amy Cochran, Valerie Odeh-Couvertier, Gabriel Zayas-Caban, Kenneth Nieser, Brian Patterson,

At the heart of learning health systems are risk prediction models that are validated, continually updated, and used to guide day-to-day clinical care according to patient risk profiles. While the validation and updating of risk prediction algorithms has been a major area of research, less attention has been paid to the evaluation of the consequences of using a risk prediction model in a learning health system. We developed a causal inference method for evaluating the impact of intervening on a patient within a learning system according to risk predictions. Our method builds on a regression discontinuity (RD) design to estimate (local) average treatment effects in settings when the intervention is determined according to whether a patient's predicted risk exceeds a certain value. Critically, our method allows for the specific interference that arises in a learning health system, whereby prior patients inform the care of future patients. Local average treatment effects are formally defined and identified, and estimators are proposed and analyzed. The method is tested in a simulated learning health system. Our method is a first step towards new standards for how learning health system should independently evaluate, in real-time, the use of risk predictions models in their day-to-day operations.

Sensitivity Analysis

Sensitivity Analysis for Violations of Proximal Identification Assumptions Raluca Cobzaru*
Raluca Cobzaru, Roy Welsch, Stan Finkelstein, Kenney Ng, Zach Shahn,

Causal inference from observational data often rests on the unverifiable assumption of no unmeasured confounding. Recently, Tchetgen Tchetgen and colleagues have introduced proximal inference to leverage negative control outcomes and exposures as proxies to adjust for bias from unmeasured confounding. However, some of the key assumptions that proximal inference relies on are themselves empirically untestable. Additionally, the impact of violations of proximal inference assumptions on the bias of effect estimates is not well understood. In this paper, we derive bias formulas for proximal inference estimators under a linear structural equation model data generating process. These results are a first step toward sensitivity analysis and quantitative bias analysis of proximal inference estimators. While limited to a particular family of data generating processes, our results may offer some more general insight into the behavior of proximal inference estimators.

Sensitivity Analysis

Testing unit root non-stationarity in the presence of missing data in univariate time series of mHealth studies Charlotte Fowler* Charlotte Fowler, Xiaoxuan Cai, Justin Baker, Jukka-Pekka Onnela, Linda Valeri,

The use of digital devices to collect data in mobile health (mHealth) studies introduces a novel application of time series methods, with the constraint of potential data missing at random (MAR) or missing not at random (MNAR). In causal inference for time series, testing for stationarity is an important preliminary step to inform appropriate later analyses since many causal models assume stationarity. The augmented Dickey-Fuller (ADF) test was developed to test the null hypothesis of unit root non-stationarity, under no missing data. Beyond recommendations under data missing completely at random (MCAR) for complete case analysis or last observation carry forward imputation, researchers have not extended unit root non-stationarity testing to a context with more complex missing data mechanisms. Multiple imputation with chained equations, Kalman smoothing imputation, and interpolation have also been proposed for time series data, however such methods impose constraints on the autocorrelation structure, and thus impact unit root testing. We propose maximum likelihood estimation and multiple imputation using a state space model approaches to adapt the ADF test to a context with missing data. We further develop sensitivity analysis techniques to examine the impact of MNAR data. We evaluate the performance of existing and proposed methods across different missing mechanisms in extensive simulations and in their application to a multi-year smartphone study of bipolar patients.

Simulation Methods

Simulation Design as Causal Intervention on a Data Generating Mechanism Tyrel Stokes*
Tyrel Stokes, Russell Steele, Ian Shrier,

Simulation methods are among the most ubiquitous methodological tools in statistical science. In particular, simulation is commonly used to explore properties of statistical functionals in models for which developed statistical theory is insufficient or to assess finite sample properties of theoretical results. We show that simulation experimental design can be viewed from the perspective of causal intervention on a data generating mechanism and demonstrate the use of causal tools and frameworks in this context. Most notably, our perspective is agnostic to the particular domain of the simulation experiment which increases the potential impact of our proposed approach. In this talk, we consider two illustrative examples. First, we re-examine a predictive machine learning example from a popular textbook designed to assess the relationship between mean function complexity and the mean-squared error. Second, we discuss a traditional causal inference method problem, simulating the effect of unmeasured confounding on estimation, specifically to illustrate bias amplification. In both cases, applying causal principles and using graphical models with parameters and distributions as nodes in the spirit of influence diagrams can 1) make precise the estimand the simulation targets , 2) suggest modifications to better attain the simulation goals, and 3) provide scaffolding to discuss performance criteria for a particular simulation design, such as the generalizability to other contexts.

Synthetic Control Method

On Misspecification in Synthetic Controls Claudia Shi* Claudia Shi, Achille Nazaret, David Blei,

The synthetic control (SC) method is a popular approach for estimating treatment effects from observational panel data. It rests on a crucial assumption that we can write the treated unit as a linear combination of the untreated units. This linearity assumption, however, can be unlikely to hold in practice and, when violated, the resulting SC estimates are incorrect. In this paper we examine two questions: (1) How large can the misspecification error be? (2) How can we limit it? First, we provide theoretical bounds to quantify the misspecification error. The bounds are comforting: small misspecifications induce small errors. With these bounds in hand, we then develop new SC estimators that are specially designed to minimize misspecification error. The estimators are based on additional data about each unit, which is used to produce the SC weights. (For example, if the units are countries then the additional data might be demographic information about each.) We study our estimators on synthetic data; we find they produce more accurate causal estimates than standard synthetic controls. We then re-analyze the California tobacco-program data of the original SC paper, now including additional data from the US census about per-state demographics. Our estimators show that the observations in the pre-treatment period lie within the bounds of misspecification error, and that the observations post-treatment lie outside of those bounds. This is evidence that our SC methods hav

Weighting

Physical Function Decline in Aging Cancer Survivors and Cancer-Free Controls: Accounting for Bias Due to Selective Attrition Sophia Fuller* Sophia Fuller, Sowmya Vasan, Hailey Banack, Alexandra Binder, Elizabeth Feliciano,

With improvements in screening and treatment, the number of cancer survivors is growing, increasing the need for research into long-term health and well-being after diagnosis. Investigations of how cancer and its treatment influence trajectories of aging is complicated by differential loss to follow-up and death between survivors and cancer-free individuals. Those with more severe disease, and correspondingly, more cytotoxic treatment, are also more likely to die or drop out. The confluence of cancer, stage, and treatment on censoring obscure our understanding of when and which survivors are at risk for accelerated aging. To demonstrate the value of accounting for censoring, we use longitudinal data from the Women's Health Initiative to compare trajectories of physical function between women with cancer to non-cancer controls. We estimate inverse probability of censoring weights due to loss to follow up and, separately, death, using ensemble machine learning for each wave of the study. Then we fit GEE models incorporating these weights to capture the yearly decline in physical function among cancer survivors, subset by cancer type and stage at diagnosis, and their non-cancer controls. We hypothesize that weighting to account for attrition will yield estimates of cancer survivors' yearly rates of decline that are larger in magnitude, and this decline will be greatest for women with more aggressive cancer types diagnosed at more advanced stages.

Weighting

Transfer Learning for Individualized Treatment Rules Andong Wang* Andong Wang, Johnny Rajala, Kelly Wentzlof, Miontranese Green,

Modern precision medicine aims to utilize real-world data to provide the best treatment for an individual patient. An individualized treatment rule (ITR) maps individual characteristics to a recommended treatment that maximizes the expected outcome of each patient. A problem facing modern medicine is that studies on the effect of treatment are conducted for a source population that may be different from the population of interest. Our research goal is to investigate a transfer learning algorithm to obtain targeted, optimal, and interpretable ITRs. We develop a calibrated augmented inverse probability weighting (CAIPW) estimator by maximizing the value function for the target population to estimate an optimal ITR. Additionally, we investigate transfer learning methods based on two large medical databases, eICU Collaborative Research Database (eICU-CRD) and Medical Information Mart for Intensive Care III (MIMIC-III), identifying the important covariates, treatment options, and outcomes of interest to estimate the optimal linear and tree-based ITRs for patients with sepsis. This project introduces new techniques for data merging to provide data-driven optimal ITRs, catering to each patient’s individual medical needs. These techniques extend beyond medicine, applying to a wide range of areas such as marketing, technology, social services, and education.

Weighting

Is there a survival benefit of initiating treatment earlier rather than later? Haris Fawad*
Haris Fawad,

This question arises in a variety of clinical settings, for example organ transplantations and palliative care for cancer patients. To provide an answer, we build a causal model based on counting processes and their intensities. The counting process framework is already well-suited for survival analysis. We only change the intensity of treatment to model hypothetical scenarios in which treatment occurs earlier, compared to the observational data. We highlight the structural assumptions needed for the identification of survival parameters, and present consistent estimators along with analytical expressions for their asymptotic confidence intervals.

Bayesian Causal Inference

A Bayesian approach to risk constrained iterative experiments Yufan Li* Yufan Li, Jialiang Mao, Iavor Bojinov,

Our work provides a theoretical foundation for phased releases, a widely adopted practice in the technology sector whereby a firm gradually releases a new product or update through a sequence of A/B tests. Typically, when performing a phased release, the analyst starts by releasing the new updates to a small percentage of the users (i.e., the treatment group); if the treatment is deemed not to cause harm, more users are added to the treatment group. This process continues until either the treatment is estimated as superior to the control, in which case the treatment is deployed to all users, or not, in which case all users are returned to the control. A key design question is how to determine the treatment group size that balances the risk associated with releasing unpopular products with the need to iterate quickly. To solve this problem, we propose a Bayesian approach to determine the treatment group size at each stage under a user-set risk budget that adopts to the observed data. Our method quantifies the risk in terms of the treatment effect of treated users under the Neyman-Rubin potential outcome framework. The treatment group size is determined by controlling the probability of exceeding the risk budget (risk level) below a user-set threshold (risk tolerance). Our approach involves decomposing risk tolerance to each stage by a recursive relation and deriving an upper bound of the stage-wise risk level so that the treatment group size can be analytically solved using a

Causal Discovery

Causal Discovery for Observational Categorical Data Yang Ni* Yang Ni,

Causal discovery for quantitative data has been extensively studied but less is known for categorical data. I will present novel causal models for categorical data. For ordinal categorical data, our model is based on ordinal regression whereas, for nominal categorical data, it is based on a new classifier, termed classification with optimal label permutation. Under either causal model, we establish its causal identifiability property with observation data alone. Through experiments with synthetic and real data, we demonstrate the favorable performance of the proposed causal models compared to state-of-the-art methods.

Causal Discovery

Causal inference on observational data: opportunities and challenges in earthquake engineering Henry Burton* Henry Burton,

Collecting and analyzing observational data are essential to learning and implementing lessons in earthquake engineering. Historically, the methods that have been used to analyze and draw conclusions from empirical data have been limited to traditional statistics. The models developed using these techniques are able to capture associative relationships between important variables. However, the intervention decisions geared toward seismic risk mitigation should ideally be informed by an understanding of the causal mechanisms that drive infrastructure performance and community response. This oral presentation will discuss how earthquake engineering research and practice can be transformed by the broad adoption of the language, tools, and models that have been (and continue to be) developed to draw causal conclusions from observational data. Several categories of data-driven earthquake engineering problems that can benefit from causal insights will be discussed. Two widely adopted frameworks from the broader causal inference literature will be examined and linked to hypothetical earthquake engineering problems. The presentation will conclude with a discussion of specific opportunities and challenges toward the widespread use of causal inference as a tool for knowledge discovery in earthquake engineering.

Causal Discovery

Confidence Sets for Causal Discovery Y. Samuel Wang* Y. Samuel Wang, Mladen Kolar, Mathias Drton,

Causal discovery procedures are popular methods for discovering causal structure across the physical, biological, and social sciences. However, most procedures for causal discovery only output a single estimated causal model or single equivalence class of models. In this work, we aim to quantify uncertainty in causal discovery. Specifically, we consider structural equation models where a unique graph can be identified and propose a procedure which returns a confidence set of causal orderings which are not ruled out by the data. We show that asymptotically, a true causal ordering will be contained in the returned set with some user specified probability. In addition, the confidence set can be used to form conservative sets of ancestral relationships as well as confidence intervals for causal effects which account for model uncertainty.

Causal Discovery

Differentiable Covariate Selection for Backdoor Adjustment Elijah Tamarchenko* Elijah Tamarchenko, Rohit Bhattacharya,

Covariate selection for backdoor adjustment is often made difficult due to unmeasured confounding; some adjustment sets can lead to bias due to exclusion of relevant confounders, others may be unbiased but statistically inefficient. Rotnitzky et al (2019) propose a graphical criterion for identifying the optimal adjustment set – an unbiased set with minimal asymptotic variance – in settings where the structure of the causal system is known exactly and there are no unobserved common causes. However, in most practical settings, the full causal structure is unknown and likely to exhibit unmeasured confounding. In this case, Entner et al (2013) propose a procedure for identifying an unbiased adjustment set. However, it performs an exponential number of conditional independence tests, which is infeasible in high dimensional settings, and does not consider minimizing variance. We propose a parametric continuous optimization procedure, which performs both covariate selection and effect estimation in a single step. We prove that this procedure identifies the optimal adjustment set in the absence of unmeasured confounders. We further show that under mild assumptions involving an auxiliary variable, if the continuous optimization procedure excludes this variable from the covariate selection process, then the effect estimate is provably unbiased even in settings with unmeasured confounders. Further, the procedure often leads to a practical reduction in variance as shown via simulations.

Causal Discovery

Causal Discovery and Prediction with Interventional Data James Long* James Long, Yumeng Yang, Kim-Anh Do,

The field of causal discovery has historically focused on parameter estimation, rather than the prediction performance of models. Recently, systems biology experiments have begun generating large-scale interventional data sets in which certain variables are manipulated (intervened on) and the resulting system state observed. These data sets offer the possibility to assess the prediction performance of causal discovery algorithms: parameters of a causal model are learned on a subset of the intervention data and then the model is used to predict the effects of test interventions. Here we derive some of the first analytic results connecting causal discovery models with standard regression approaches for intervention response prediction. Causal discovery models require estimation of many more parameters than standard regression approaches but can extrapolate to predict the effect of untested interventions. We study the performance of causal discovery models and regression approaches in simulations and an application to predicting the effect of drug interventions in a Melanoma cancer cell line. In the Melanoma data set, we obtain state of the art performance with regression modeling, outperforming a substantially more complex causal discovery model proposed in the computational biology literature.

Causal Inference and Bias/Discrimination

Attributable fraction and related measures: conceptual relations in the counterfactual framework Etsuji Suzuki* Etsuji Suzuki, Eiji Yamamoto,

The attributable fraction (population) has attracted much attention from a theoretical perspective and has been used extensively to assess the impact of potential health interventions. However, despite its extensive use, there is much confusion about its concept and calculation methods. In this presentation, we discuss the concepts of and calculation methods for the attributable fraction and related measures in the counterfactual framework, both with and without stratification by covariates. Generally, the attributable fraction is useful when the exposure of interest has a causal effect on the outcome. However, it is important to understand that this statement applies to the exposed group. Although the target population of the attributable fraction (population) is the total population, the causal effect should be present not in the total population but in the exposed group. As related measures, we discuss the preventable fraction and prevented fraction, which are generally useful when the exposure of interest has a preventive effect on the outcome, and we further propose a new measure called the attributed fraction. We also discuss the causal and preventive excess fractions. Finally, we discuss the relations between the aforementioned six measures and six possible patterns using a conceptual schema. It is important to have clear definitions of them in the counterfactual framework, which would improve the interpretation and use of these measures.

Causal Inference and Bias/Discrimination

A health equity perspective on data-driven treatment decisions in cardiovascular care: risk assessments versus individualized treatment rules Safiya Sirota* Safiya Sirota, Daniel Malinsky,

It is standard in clinical care to inform medical decisions based on estimated risk scores, e.g., to inform assignment of antihypertensive medications based on risk of adverse cardiac events, as is currently recommended by national ACC/AHA cardiovascular guidelines. We will investigate the consequences of this practice in cardiovascular care from the perspective of health equity and health disparities. Complex associations between racial/ethnic categories, social determinants of health, and other disease risk factors may lead to disparities in treatment allocation that are exacerbated, not mitigated, by risk-based decision-making. An alternative is to base decisions on individualized treatment rules (ITRs), which are rules sensitive to causal effect heterogeneity that optimally direct therapies to patients based on their individual characteristics. We investigate how allocations based on ITRs may mitigate disparities in treatment assignment, using both simulated data and real data from a large observational cohort study. We find that recommending treatment according to the ITR paradigm may have substantial consequences for treatment recommendations and possibly health disparities.

Causal Inference and SUTVA/Consistencies Violations

Controlling for spatial confounding and spatial interference in causal inference: Model selection advice from a computational experiment Tyler Hoffman* Tyler Hoffman, Peter Kedron,

Working with spatial data raises the possibility of encountering unique issues in causal inference—namely, spatial confounding and spatial interference. A blossoming literature on spatial causal inference is growing to address these issues, largely via regression adjustments in existing causal models. This research analyzes the usage of spatial causal models under a priori knowledge and a priori ignorance of dependence structures in a spatial dataset. We test whether spatial causal models accurately capture treatment effects in the presence of spatial confounding and interference by fitting these models on various spatial data scenarios. Based on the results of these experiments, we develop practical workflow guidelines based on the relative performance of spatial and nonspatial causal models across data scenarios. All experiments use Bayesian estimation techniques for additional uncertainty quantification. In parallel, we build a Python software package of spatial causal models and data simulators to facilitate the widespread use of these models and to enable reproduction of this work.

Causal Inference and SUTVA/Consistencies Violations

Causal Effects of Continuous Exposures in the Presence of Spatial Interference: the Effects of Air Pollution on Public Health Heejun Shin* Heejun Shin, Joseph Antonelli,

We develop new methodology to improve our understanding of the causal effects of multivariate air pollution exposures on public health. Typically, exposure to air pollution for an individual is measured at their home geographic region, though people travel to different regions with potentially different levels of air pollution. To account for this, we incorporate estimates of the mobility of individuals from cell phone mobility data to get an improved estimate of their exposure to air pollution. We treat this as an interference problem, where individuals in one geographic region can be affected by exposures in other regions due to mobility into those areas. We propose policy-relevant estimands and derive expressions showing the extent of bias one would obtain by ignoring this mobility. We additionally highlight the benefits of the proposed interference framework relative to a measurement error framework for accounting for mobility. We develop novel estimation strategies to estimate causal effects that account for this spatial spillover utilizing flexible, nonparametric Bayesian methodology. Empirically we find that this leads to substantially improved estimation of the causal effects of air pollution exposures over analyses that ignore spatial spillover caused by mobility.

Causal Inference and SUTVA/Consistencies Violations

Estimating causal effects of interventions altering social connectivity patterns under network interference Shinpei Nakamura Sakai* Shinpei Nakamura Sakai, Laura Forastiere,

Causal inference under network interference is an emerging topic as network data is ubiquitous across multiple disciplines. We say that the treatment effect ‘spills over’ to other units when one’s potential outcomes are affected by the treatment status of other units. Such a phenomenon is often due to social or physical interactions and depends on the social structure of the population. An intervention that alters social connectivity would alter the interference mechanism and consequently, the spillover effect. Current methods under interference estimate causal effects conditional on a known and fixed social connectivity graph. On the other hand, epidemic models have been used to predict the effect of hypothetical interventions altering social connectivity parameters to control the spread of infectious diseases. However, a formal and general definition of the causal effects of such interventions altering the social structure is lacking. We consider a stochastic network formation and propose causal estimands to estimate spillover effects with a modification of the network formation mechanism. These causal estimands are defined under interventions shifting the degree distribution or the network formation mechanism. We develop estimators for the counterfactual estimands under hypothetical interventions altering the network structure, and we investigate the finite sample bias and large-sample properties of these estimators.

Causal Inference and SUTVA/Consistencies Violations

Identification and estimation of interference effects with contextual multilevel models Yi Feng* Yi Feng, Peter Steiner,

Interference (spill-over) effects play an important role in the social, behavioral, and health sciences. While different types of causal interference effects can be clearly defined in terms of potential outcomes and interference graphs, their estimation remains a considerable challenge and generally employable analytic tools are still not available (Hong, 2015; Hudgens & Halloran, 2008; VanderWeele, 2015). To advance applied research on interference effects, we start by considering contextual multilevel models (MLM, mixed effects models) where the average treatment exposure of multiple subjects within groups (group-level means) is used as an additional predictor to assess contextual treatment effects (Enders, 2013). Using different data-generating interference models (compositional, direct interference, contagion; Ogburn & VanderWeele, 2014) we investigate whether contextual MLMs are able to identify average causal interference effects. When interference effects are caused by the groups' mean exposure, the MLM estimator identifies the average interference effect. However, even when spill-over effects are caused by direct interference or via contagion, which is more likely in practice than compositional effects transmitted via group means, we show that the corresponding interference effects can be recovered from MLM estimates. Theoretical identification results using causal graphs will be presented and discussed, including extensions to heterogeneous interference effects.

Causal Inference and SUTVA/Consistencies Violations

Switchback Experiments under Geometric Mixing Yuchen Hu* Yuchen Hu, Stefan Wager,

The switchback is an experimental design that measures treatment effects by repeatedly turning an intervention on and off for a whole system. Switchback experiments are a robust way to overcome cross-unit spillover effects; however, they are vulnerable to bias from temporal carryovers. In this paper, we consider properties of switchback experiments in Markovian systems that mix at a geometric rate. We find that, in this setting, standard switchback designs suffer considerably from carryover bias: Their estimation error decays as T^{-1/3} in terms of the experiment horizon T, whereas in the absence of carryovers a faster rate of T^{-1/2} would have been possible. We also show, however, that judicious use of burn-in periods can considerably improve the situation, and enables errors that decay almost as fast as T^{-1/2}. Our formal results are mirrored in an empirical evaluation.

Causal Inference Applications

Comparisons of the Treatment and Side Effects of Several Bariatric Surgery Procedures: An Observational Study via Random Forest-based and Neural Network-based Approaches

Qishuo Yin* Qishuo Yin, Qishuo Yin, Jiawei Zhang, Carlos Fernandez-Granda, Siyu Heng,

Bariatric surgery is an effective treatment for obesity, as well as diabetes, high blood pressure, sleep apnea, and high cholesterol. Over the past decades, several bariatric surgery procedures based on different techniques have been widely performed in practice. However, there is a lack of rigorous causal comparisons of their treatment and side effects. In this work, we conduct a large-scale observational study to compare the effects of several widely used bariatric surgery procedures on Bariatric surgery complication risk and weight loss using the datasets from the American College of Surgeons Metabolic and Bariatric Surgery Accreditation and Quality Improvement Program (MBSAQIP). We apply several state-of-art machine learning-based causal inference approaches such as Causal Forest, Orthogonal Random Forest, and Dragonnet to better leverage the large datasets to provide accurate and powerful causal comparisons, both at population and individual levels. On the one hand, the results from our population-level causal comparisons provide rigorous statistical evidence for the appropriateness of the current guidelines for Bariatric surgery procedures provided by the American Society for Metabolic and Bariatric Surgery (ASMBS). On the other hand, our individual-level causal comparisons offer data-driven suggestions on Bariatric surgery procedure selections for individual patients, of which the corresponding user-friendly statistical software will be provided on the website soon.

Causal Inference in Networks

Bayesian inference for the estimation of causal effects under network interference Seungha Um* Seungha Um, Samrachana Adhikari,

Causal inference in the presence of interference is challenging in observational studies on social network since the effect of treatment on a unit spills over connected units. The fact that the spillover effect is commonly confounded with latent homophily and depends on social influence, which varies by unit, makes the estimation more challenging. To disentangle the effect of the individual treatment and neighborhood treatment, we employ the estimation strategy based on generalized propensity score. Also, the homophilous attributes are estimated by relying on latent space positions and social influence is evaluated based on pairwise distances among each pair of nodes. Within Bayesian framework, the uncertainty in propensity score is quantified while avoiding model feedback and imputation of missing potential outcome is implemented. We design a simulation study to assess the performance of our proposed method and examine the spillover effect varying characteristics of network topology such as size, density and community structure. The causal effects of exposure to belief about teaching mathematics are examined on teacher advice-seeking network.

Causal Inference in Networks

Estimating Causal Spillover Effects in Smallholder Farmer Networks in Western Kenya

Medha Uppala* Medha Uppala, Bruce A Desmarais, David P Hughes,

As farmers share information on practical agricultural techniques with each other, we want to understand what demographic and social network factors lead to greater information sharing and adoption.
The goal of this research is to understand and estimate how efficiently information on disease management behaviors (DMB) spreads in smallholder farmer networks. In this research, we are concerned with DMBs with respect to plant and crop disease. More formally, the goal is to causally estimate the network spillover effects of specific DMBs in smallholder farmer networks in Western Kenya. We intend to achieve this goal through a cluster randomized trial. Methods employed include re-randomization during the design phase, randomization tests to detect interference and estimating direct and indirect effect size via exposure networks.

Causal Inference in Networks

Covariate Adjustment in Randomized Trials under General Interference Hyunseung Kang*
Ralph Trane, Hyunseung Kang,

Covariate adjustment has been a popular approach to improve precision and power in randomized trials. However, in some randomized trials say vaccine trials or evaluation of new educational programs, study units often interact and affect each other’s responses, a phenomena known as general interference. While recent work have shown how to identify and estimate treatment effects in randomized trials under general interference, to the best of our knowledge, there is little work on how to properly adjust for covariates in this setting, especially to deal with the potentially complex and unknown dependencies between study units. In this paper, we propose a class of flexible, covariate-adjusted estimators of treatment effects under general interference. Under some smoothness conditions on the response model, our estimators are consistent, asymptotically Normal, and can incorporate some, but not all, types of flexible methods from machine learning. An important corollary of our result is that ANCOVA, a popular method of covariate adjustment in randomized trials under no interference, yields a consistent, asymptotically Normal, covariate-adjusted estimator of treatment effects under general interference. Our results are demonstrated through a simulation study under some popular network models and an empirical study. We conclude by providing concrete guidelines to practitioners on how to adjust for covariates in randomized trials under general interference.

Causal Inference in Networks

Cluster Randomized Designs for One-Sided Bipartite Experiments Jennifer Brennan* Jennifer Brennan, Vahab Mirrokni, Jean Pouget-Abadie,

The conclusions of randomized controlled trials may be biased when the outcome of one unit depends on the treatment status of other units, a problem known as interference. In this work, we study interference in the setting of one-sided bipartite experiments in which the experimental units—where treatments are randomized and outcomes are measured—do not interact directly. Instead, their interactions are mediated through their connections to interference units on the other side of the graph. Examples of this type of interference are common in marketplaces and two-sided platforms. The cluster-randomized design is a popular method to mitigate interference when the graph is known, but it has not been well-studied in the one-sided bipartite experiment setting. In this work, we formalize a natural model for interference in one-sided bipartite experiments using the exposure mapping framework. We first exhibit settings under which existing cluster-randomized designs fail to properly mitigate interference under this model. We then show that minimizing the bias of the difference-in-means estimator under our model results in a balanced partitioning clustering objective with a natural interpretation. We further prove that our design is minimax optimal over the class of linear potential outcomes models with bounded interference. We conclude by providing theoretical and experimental evidence of the robustness of our design to a variety of interference graphs and potential outcomes models.

Difference in Differences

A Difference-in-Differences Framework to Estimate Causal Effects for Policy Interventions in the Presence of Heterogeneous Interference with an Application to the Philadelphia Beverage Tax Gary Hettinger* Gary Hettinger, Youjin Lee, Nandita Mitra,

Public policy interventions are often evaluated using the difference-in-differences (DiD) approach, which does not directly account for a policy affecting nearby regions, particularly when these neighboring effects vary spatially. For example, an excise tax on sweetened beverages in Philadelphia (PHL) was associated with substantial decreases in volume sales of taxed beverages in PHL as well as increases in beverage sales of nontaxed bordering counties. The latter association may be explained by cross-border shopping behaviors of PHL residents, which may vary with border proximity, transportation access, and demographics. For example, past studies have found evidence of that the tax affects Philadelphia stores differentially depending on if they border other non-taxed Pennsylvania counties, New Jersey, or are entirely surrounded by Philadelphia zip-codes. Because such effects can offset the total effect of such interventions, particularly for specific sub-populations, understanding effect dynamics is essential to holistically evaluate public policies. Further, such insights may help predict policy effects under diverse implementation strategies. To address these concerns, we extend DiD methodology to robustly identify the causal effects of policy interventions under potentially heterogeneous interference exposure. Here, we present initial work demonstrating our framework with an evaluation of the PHL Beverage Tax policy.

Difference in Differences

Controlling time-varying confounding in difference-in-differences studies using the time-varying treatments framework Leslie Myint* Leslie Myint,

I clarify how the biostatistical literature on time-varying treatments (TVT) can provide tools for dealing with time-varying confounding in difference-in-differences (DiD) studies. I use a simulation study to compare the bias and standard error of inverse probability weighting estimators from the TVT framework, a DiD framework, and hybrid approaches that combine ideas from both frameworks. I simulated longitudinal data with treatment effect heterogeneity over multiple time points using linear and logistic models. Simulation settings looked at both time-invariant confounders and time-varying confounders affected by prior treatment. Estimators that combined ideas from both frameworks had lower bias than standard TVT and DiD estimators when assumptions were unmet. The TVT framework provides estimation tools that can complement DiD tools in a wide range of applied settings. It also provides alternate estimands for consideration in policy settings.

Estimation under conditional ignorability

Regression’s weighting problem: a new analysis and simple fixes Tanvi Shinkre* Tanvi Shinkre, Chad Hazlett,

Researchers often use regression to estimate causal effects. Regressing the outcome of interest (Y) on a binary treatment indicator (D) without covariates (X) produces a coefficient equal to the difference in means and unbiased for the average treatment effect (ATE). More often, covariates are included to adjust for confounding. It is now well known that the resulting coefficient produces a weighted average of treatment effects across covariate strata, not the ATE. To address this, some researchers propose diagnostics to quantify the severity of this “weighting problem” while others suggest alternative estimation strategies. After reviewing the literature, we develop a new expression for these weights, required to recover the correct weights when D is not linear in X. We then compare a number of estimators and the assumptions under which they can recover the ATE. Central to the analysis is the recognition that multiple alternatives–including regression-imputation (g-computation) and interacting the treatment with covariates (Lin 2013)–can be motivated by assuming the treatment and non-treatment potential outcomes are separately linear in X. Further, methods sharing this justification are indeed equivalent, producing the same estimates–all of which side-step the weighting problem. Beyond providing this theoretical clarity, we recommend the simple solution of interacting D with X to solve this problem with minimal change to existing practice.

Functional Estimation and Causal Inference

Double Robust Estimation with Split Training Data: Achieving Minimax Optimality with Undersmoothed Local Averaging Linear Smoothers Alec McClean* Alec McClean, Edward Kennedy, Sivaraman Balakrishnan, Larry Wasserman,

Double robust (DR) estimators have gained popularity in causal inference due to their favorable convergence properties. However, minimax optimal estimators often rely on correcting the bias of the DR estimator through a higher-order von Mises expansion. Instead, minimax optimal DR estimators can be constructed by splitting the training data and estimating undersmoothed nuisance functions on independent samples, which we refer to as the “double robust split training” estimator (DR-ST). In this work, we estimate the expected conditional covariance and average outcome under treatment using DR-ST estimators. We derive an asymptotically linear expansion that holds under a nonparametric stability condition on the nuisance function estimators, and show when linear smoothers satisfy this condition. We examine three DR-ST estimators based on local averaging linear smoothers, assuming throughout that the nuisance functions belong to Holder smoothness classes. We demonstrate that local polynomial regression can achieve semiparametric efficiency under minimal smoothness conditions. We then propose a covariate-density-adapted local polynomial regression that is minimax optimal when the covariate density is known or well-estimated and show that a Normal limit distribution exists even in the low regularity case by deriving a slower-than-root-n central limit theorem. Finally, we present an easy-to-implement DR-ST estimator using 1-Nearest-Neighbors and illustrate its convergence properties.

Generalizability/Transportability

Harmonizing graphical methods for dynamic and static causal modelling Ian Shrier* Ian Shrier, Naftali Weinberger, Tyrel Stokes, Russell J. Steele,

Causal directed acyclic graphs (DAGs) succinctly outline assumptions about causal relationships between variables. Causal DAGs are called static models because the random variables are in a stable state. A light switch either completes or interrupts an electrical circuit. However, most biological systems operate as a dynamic model; “cancer” occurs when the development of cancer cells exceeds the immune system’s ability to kill them. Dynamic models generally describe rates of change using derivatives. “Causes” are often modelled as affecting the derivative, and causal loops are used to indicate that variables can cause each other over time. One partial harmonizing approach “unfolds” the causal loop and defines variables according to the time they are measured. We expand on this approach. First, we explicitly distinguish between the data generating process that describes the world as it is, and the data generating process as it would be in the presence of a new intervention. Second, we leverage previous work using causal DAGs with nodes representing derivatives to allow estimation for a new range of questions previously only addressed by dynamic models. Third, our modifications are more flexible than current dynamic models because they allow for the derivatives to change over time, which may occur when interventions alter the equilibrium state.

Generalizability/Transportability

Examining subgroup-specific treatment effects in multi-source data: source-specific inference and transportability to an external population Guanbo Wang* Guanbo Wang, Alex Levis, Issa Dahabreh,

One major challenge in estimating effect heterogeneity is that the sample size of the data used is typically not enough to capture how effects vary according to the effect modifiers precisely. Therefore, there is interest in synthesizing evidence across multi-source data (e.g., multi-center trials, meta-analyses of randomized trials, pooled analyses of observational cohorts) to improve the precision of estimators of heterogeneous treatment efficacy. Furthermore, when combining information from multi-source data, the samples typically do not represent a common target population of substantive interest. This raises the question of how to combine information from multi-source data in a way that is interpretable in the context of some meaningful target population of interest while using evidence across multi-source data to improve efficiency. We develop and evaluate methods for using multi-source data to estimate subgroup treatment effects in an external target population or the populations underlying the data sources. We propose a doubly robust estimator that, under mild conditions, is non-parametrically efficient and allows for nuisance functions to be estimated using machine learning methods. We illustrate the methods in meta-analyses of randomized trials for schizophrenia and bipolar disorder.

Generalizability/Transportability

An Interdisciplinary Perspective on Building Causal Knowledge using Evidential Pluralism

James Grace* James Grace,

In this talk, I present a map of the fields of science that allows us to consider some of the major differences among fields and the bases for those differences that are relevant to causal studies. In addition, I suggest a number of linguistic distinctions as aids to clarify discussion and improve communication. An important conclusion from this examination, which is in accordance with the consensus view of causal scholars/philosophers, is that the study of causality must be treated as a pluralistic enterprise. By itself, this recognition is not sufficiently helpful to point a way forward. However, a new philosophical thesis “Evidential Pluralism” (first developed for the field of medicine) provides a promising epistemological system for causal studies suitable for all scientific disciplines and a wide variety of intentions. Evidential Pluralism maintains that the ideal approach to establishing causal claims requires both empirical evidence and relevant mechanistic knowledge. This combination permits general causal claims with external validity. At present, a major challenge for this system is to decide more precisely how to use mechanistic knowledge to enhance the robustness of empirical studies, and vice versa. An important element of the solution will be to shift emphasis from simply making causal claims to building causal knowledge.

Generalizability/Transportability

Causality and robustness from heterogeneous data: going beyond mean shifts Xinwei Shen* Armeen Taeb, Xinwei Shen, Peter Buehlmann,

Despite being challenging for iid-based statistical learning, heterogeneous data provides opportunities for causal inference and for learning prediction models that generalize to unseen environments. Indeed, existing methods such as anchor regression exploit heterogeneous data arising from mean shifts to the features and the response variable of interest to learn distributionally robust predictions models. In many real-world settings, apart from mean shifts, the perturbations may also be affecting the variances of the relevant variables. Previous techniques however are not able to handle this richer perturbation class. We propose Distributionally Robust predictions via Invariant Gradients (DRIG), a method that leverages perturbations in the form of both mean and variance shifts for robust predictions. In a linear setting, we prove that DRIG produces prediction models that are robust against perturbations in strictly (and often much) more ‘directions’ than those protected by anchor regression, highlighting the additional gains from exploiting heterogeneity beyond mean shifts. Viewing causality as an extreme case of distributional robustness, we investigate the causal identifiability of DRIG under various scenarios. Moreover, we extend DRIG to the semi-supervised domain adaptation setting where a few labeled samples from the target domain are available and are used to further improve robustness. Finally, we illustrate the utility of our methods through numerical experiments.

Generalizability/Transportability

Transportability sharp bounds Guilherme Duarte* Guilherme Duarte,

Transportability is one of the biggest challenges in causal inference. Researchers are often familiar with techniques to estimate quantities such as the ATE, from experimental/observational data, but they still struggle with generalizing estimates from one environment to another. For example, they might run an experiment in Los Angeles and calculate an effect, but they still hesitate over which assumptions allow them to know what the effect would be in New York City without having to rerun the experiment there. Current papers emphasize solutions that indicate if transportation is permitted given particular structural assumptions. Nonetheless, limitations to this strategy are known as it puzzlingly says cases are not transportable when they in fact are, and even when a precise solution does not exist, it fails to provide informative bounds. Here I propose a general and complete algorithm to always provide upper and lower sharp bounds to transportable estimates. Among the inputs, one states two different environments (for instance, Los Angeles and New York City), introduces existing data to both (e.g., baseline mortality rates in each city), invariance assumptions, and an original quantity they want to transport. The algorithm, then, outputs sharp bounds for that estimate in the second environment. If both bounds converge, then one would say that the quantity is precisely transportable. To show the applicability of this method, two examples of the literature are analyzed.

Heterogeneous Treatment Effects

Causal Inference for Distributional Treatments Andrej Srakar* Andrej Srakar,

Symbolic data analysis proposes that a distribution or an interval of the individual records’ values is associated with each unit considering new variable types named symbolic variables. Wo build on contributions by Athey and Imbens (2006), Gunsilius (2020) and Pollmann (2022) to develop concept of a distributional treatment, i.e. causal variable which is of distributional nature. Different to previous authors we transform the problem in a general regression context for empirical distributional data, in particular Dias-Brito (2011) two-quantile and Irpino-Verde (2012) two-component regressions which more adequately solve the problem of negative coefficients in the regression of quantile functions. We develop explicit formulas for average treatment effect, average treatment effect for compliers and local average treatment effect in a linear regression framework. We study the performance of our approach in asymptotic and simulation context. We demonstrate that in the Dias-Brito case it transforms in a constrained OLS optimization problem with well defined optimal solutions. In Irpino-Verde case we derive an explicit form for the ATE estimator and show it is consistent and asymptotically normal. We apply this to causal relationship between health indicators and decision to retire using pooled panel dataset of Health and Retirement Study (HRS). Our analysis can be generalized to other causal approaches in econometrics and statistics.

Heterogeneous Treatment Effects

Conditionally Normalized Two-Way Fixed Effects Estimation with Sub-Gaussian Rates Adam Bouyamourn* Adam Bouyamourn,

I describe TWFE estimation as a nuisance parameter estimation problem, and describe heterogeneity in terms of a location-scale approximation that shows simply why TWFE estimation breaks down in the context of heterogeneity or imbalanced panel designs. This suggests a general approach to resolving the problem — conditional Studentization in the directions of the nuisance dimensions, which yields estimators that are ancillary for nuisance parameters, and hence consistent for the Average Treatment Effect. I show that this approach is applicable to a wide variety of designs. I explain why these estimators resolve the so-called `forbidden comparisons’ problem, and why the problem of `negative weights’ is not, in fact, a problem. I then propose and evaluate the performance of three conditionally Studentized estimators consistent for the ATE with sub-Gaussian rates: trimmed means, median of means, and an Edgeworth expansion of the sample mean. I assess their empirical performance through simulation and the reanalysis of two papers.

Heterogeneous Treatment Effects

Assessment of Heterogeneous Treatment Effect Estimation Accuracy via Matching Zijun Gao* Zijun Gao, Trevor Hastie, Robert Tibshirani,

We study the assessment of the accuracy of heterogeneous treatment effect (HTE) estimation, where the HTE is not directly observable so standard computation of prediction errors is not applicable. To tackle the difficulty, we propose an assessment approach by constructing pseudo-observations of the HTE based on matching. Our contributions are three-fold: first, we introduce a novel matching distance derived from proximity scores in random forests; second, we formulate the matching problem as an average minimum-cost flow problem and provide an efficient algorithm; third, we propose a match-then-split principle for the assessment with cross-validation. We demonstrate the efficacy of the assessment approach using simulations and a real dataset.

Heterogeneous Treatment Effects

Heterogeneous Treatment Effects and Survival Analysis with a State Indian Child Welfare Statute: Implications for Practice Claudette Grinnell-Davis* Claudette Grinnell-Davis, Richard Smith,

There is no direct way to evaluate the Indian Child Welfare Act (ICWA)'s effect on Indian children in foster care; national data (AFCARS) has no fields for Tribal citizenship (required for ICWA) or ICWA compliance. Nebraska strengthened its ICWA compliance through enhanced statutes specifying practice guidelines to facilitate safe, accelerated permanency. However, challenges in identifying eligibility and practice impact remain.

This research team used an event history AFCARS file of all race-identified Indian (AIAN) children in Nebraska whose cases opened and closed prior to statute (7/1/2015) or whose cases opened after statute and closed before 9/30/2018 (N=777). Assuming AIAN-only children are more likely to be Tribal citizens, we compared AIAN only and AIAN+ children in our ATET calculations and treatment effects survival analyses, using passage of the statute as treatment, in relationship to discharge reason (reunification, adoption, guardianship, or agency transfer) and duration of case.

After-statute cases of AIAN-only children discharged on average 132 days sooner than either before-statute or AIAN+ after-statute children. While there was no effect on reunification, treatment effects on AIAN-only adoptions demonstrated a drop (desired) while agency transfers increased (also desired). These results indicate that statute-specified practices may impact changes in case progress for the children who are intended to benefit.

Heterogeneous Treatment Effects

Distilled Heterogeneous Treatment Effect Estimation for Ad Campaigns Emre Kiciman* Jing Ma, Emre Kıcıman, Sergii Babkin,

The key of designing powerful online advertising systems is to correctly assess the impact of each ad campaign on different users' future behaviors. This task can be naturally formulated as a heterogeneous treatment effect estimation (HTE) problem which investigates the causal effects of seeing an ad campaign on users' decisions (e.g., making a purchase) in a personalized way. Nevertheless, most existing studies require sufficient observations of user behavior data related to the ad campaigns (e.g., whether a user has purchased the corresponding product). These methods, however, cannot directly estimate the effects of ad campaigns with insufficient data (e.g., new ad campaigns that have not yet been deployed). In this paper, we address the problem of HTE estimation for ad campaigns, especially those with insufficient data. To address this problem, we propose a novel two-stage distillation based framework, which extracts the HTE estimation knowledge learned from multiple HTE estimators trained on ad campaigns with sufficient user behavior data, and thus enables effective HTE estimation on other campaigns (even those with insufficient user behavior data) by effectively leveraging the features of ad campaigns. Extensive experiments on real-world data validate the effectiveness of our method in HTE estimation for ad campaigns with insufficient user behavior data, and demonstrate the improvement of our method in data utility.

Heterogeneous Treatment Effects

Resource-constrained optimal rules for HIV care retention in rural Kenya Lina Montoya*

Lina Montoya, Harriet Adhiambo, Thomas Odeny, Elvin Geng, Maya Petersen,

Missed clinic visits can compromise HIV treatment success. A recent Sequential Multiple Assignment Randomized Trial (ADAPT-R) of 1,816 HIV-positive patients in Kenya showed that, on average, conditional cash transfers (CCTs) for on-time clinic visits increased viral suppression (VS), an indicator of treatment success, compared to standard of care. We applied SuperLearning to data from ADAPT-R to estimate an optimal dynamic treatment rule for CCT use. Preliminary results suggest that this rule would assign CCTs to all persons (i.e., CCT was not harmful in the short term for any participants). In practice, however, resources may constrain the proportion of persons who can receive a CCT. One response is to selectively administer CCTs to only those persons most likely to benefit. For example, standard univariate effect modification analyses suggest that CCTs were more effective for persons living further from the clinic or self-employed. Thus, we use the approach of Luedtke and van der Laan (2016) to estimate optimal stochastic allocation rules for administering CCTs under a range of constraints on the maximum proportion of patients who can receive a CCT. We further evaluate the expected counterfactual probability of VS under each resource-constrained optimal rule and contrast it with the expected counterfactual outcome under the static rule in which everyone receives a CCT. Our work provides an applied illustration of resource-constrained optimal dynamic treatment rules.

Heterogeneous Treatment Effects

Toward Fair and Robust Estimation of Optimal Treatment Regimes Kwangho Kim* Kwangho Kim, Jose Zubizarreta,

We propose a simple and general framework for nonparametric estimation of optimal treatment regimes under fairness constraints. Under standard regularity conditions, we show that the resulting estimators possess the double robustness property. We go on to use this framework to characterize the trade-off between fairness and the maximum welfare achievable by the optimal treatment policy. We evaluate the methods in a simulation study and illustrate them in a real-world case study.

Heterogeneous Treatment Effects

Adaptive experiment design for efficient semiparametric estimation in the partially linear model Harrison Li* Harrison Li, Art Owen,

We propose an adaptive procedure to optimally choose binary treatment assignments in a multi-stage randomized experiment based on semiparametric efficient estimation in a partially linear model for heterogeneous treatment effects. The asymptotic covariance matrix of the treatment effect estimator depends on both propensity scores and nuisance functions such as conditional variances. We show that we can use data from previous stages of the experiment to consistently estimate propensity scores for future stages that optimize a scalar functional of this covariance matrix, with possible constraints on the fraction of subjects treated at each stage. Allowable functionals include standard “design criteria” from the classical theory of experiment design, such as D-optimality. With appropriate cross fitting, the data from this adaptive experiment can be used to efficiently estimate the treatment effect under weak conditions on nuisance function estimation, even if there is bounded covariate shift between stages. Here, efficiency means achieving the same first-order asymptotic behavior as a semiparametric efficient estimator that uses data collected non-adaptively according to the optimal propensity scores and knows all nuisance functions exactly. Our running example evaluates a targeting algorithm for the Reemployment Services and Eligibility Assessment (RESEA) program, a U.S. government service to help unemployment insurance recipients resume their careers.

Heterogeneous Treatment Effects

Finding Valid Adjustments under Non-ignorability with Minimal DAG Knowledge Abhin Shah* Abhin Shah, Karthikeyan Shanmugam, Kartik Ahuja,

There are two very different schools of thought for treatment effect estimation from observational data. On one hand, the Pearlian framework commonly assumes structural knowledge (provided by an expert) in the form of directed acyclic graphs and provides graphical criteria, e.g., the back-door criterion, to identify valid adjustment sets. On the other hand, the potential outcomes (PO) framework commonly assumes that all the observed features satisfy ignorability (i.e., no hidden confounding), which in general is untestable. In prior works that attempted to bridge these frameworks, there is an observational criteria to identify an anchor variable and if a subset of covariates (not involving the anchor variable) passes a suitable conditional independence criteria, then that subset is a valid back-door. Our main result strengthens these prior results by showing that under a different expert-driven structural knowledge — that one variable is a causal parent of the treatment variable — remarkably, testing for subsets (not involving the known parent variable) that are valid back-doors is equivalent to an invariance test. Importantly, we cover the non-trivial case where the entire set of observed features is not ignorable (generalizing the PO framework) without requiring the knowledge of all the parents of the treatment variable. We also leverage Invariant Risk Minimization to connect finding valid adjustments (in non-ignorable observational settings) to representation learning.

high-dimensional causal effects estimation

Causal Inference with High-dimensional Discrete Covariates Zhenghao Zeng* Zhenghao Zeng, Edward Kennedy, Sivaraman Balakrishnan,

When estimating causal effects, covariate adjustment is often applied in estimating the outcome model and propensity score. The desired properties of the estimator are typically based on fast convergence rates of these nuisance function estimators, for which additional structural assumptions (e.g. smoothness) are usually required on the nuisance functions. However, in real applications, researchers may only have access to discrete covariates (with potentially a large number of levels). In this setting, commonly used structures such as smoothness fail to hold and the behavior of the estimator has not been well-understood. In this work, we study estimation of the average causal effect in a model where the covariates required for confounding adjustment are discrete but arbitrarily high-dimensional. Specifically, we develop point estimation theory for two causal estimands: average treatment effects (ATE) and variance-weighted average treatment effects (WATE). We consider commonly used estimators and examine conditions required for consistently estimating the functionals of interests. The results are illustrated via simulation studies. Importantly, we also characterize minimax lower bounds of the target functionals.

Instrumental Variables

Relaxing IV Exclusion with Common Confounders Christian Tien* Christian Tien,

Instruments can be used to identify causal effects in the presence of unobserved confounding, under the famous relevance and exclusion assumptions. As exclusion is difficult to justify and to some degree untestable, it often invites criticism in applications. Hoping to alleviate this problem, we propose a novel identification approach, which relaxes traditional IV exclusion to exclusion conditional on some unobserved common confounders.

We assume there exist some relevant proxies for the unobserved common confounders. Unlike typical proxies, our proxies can have a direct effect on the endogenous regressor and the outcome. We provide point identification results with a linearly separable outcome model in the disturbance, and alternatively with strict monotonicity in the first stage.

Our approach is motivated by causal questions with observational data subject to unobserved common confounders in various disciplines. Using this novel method, we separate ability and selection bias in the economic returns to education problem using NLS97 data. Yet, the approach is just as relevant in health treatment evaluation with an unobserved underlying health status, or a psychological study where character traits are unobserved common confounders.

Instrumental Variables

A novel penalized inverse-variance weighted estimator for Mendelian randomization with applications to COVID-19 outcomes Zhonghua Liu* Zhonghua Liu,

Mendelian randomization utilizes genetic variants as instrumental variables (IVs) to estimate the causal effect of an exposure variable on an outcome of interest even in the presence of unmeasured confounders. However, the popular inverse-variance weighted (IVW) estimator could be biased in the presence of weak IVs, a common challenge in MR studies. In this article, we develop a novel penalized inverse-variance weighted (pIVW) estimator, which adjusts the original IVW estimator to account for the weak IV issue by using a penalization approach to prevent the denominator of the pIVW estimator from being close to zero. Moreover, we adjust the variance estimation of the pIVW estimator to account for the presence of balanced horizontal pleiotropy. We show that the recently proposed debiased IVW (dIVW) estimator is a special case of our proposed pIVW estimator. We further prove that the pIVW estimator has smaller bias and variance than the dIVW estimator under some regularity conditions. We also conduct extensive simulation studies to demonstrate the performance of the proposed pIVW estimator. Furthermore, we apply the pIVW estimator to estimate the causal effects of five obesity-related exposures on three coronavirus disease 2019 (COVID-19) outcomes. Notably, we find that hypertensive disease is associated with an increased risk of hospitalized COVID-19; and peripheral vascular disease and higher body mass index are associated with increased risks of COVID-19 outcomes.

Instrumental Variables

Nonparametric Mendelian Randomisation for Characterising Nonlinear Exposure-Outcome Relationship with Discrete Instrumental Variables Cunhao Liu* Cunhao Liu, Stephen Burgess,

Nonlinear Mendelian randomisation is an extension to standard MR to characterise the nonlinear exposure-outcome relationship using genetic variants as instrumental variables. The approach divides the population into strata with different average levels of the exposure, and estimates average causal effect in each stratum, known as the localised average causal effect (LACE). However, this method typically requires a strong constant genetic effect assumption, and is unable to capture complex nonlinearities, for example a threshold relationship. We propose a nonparametric extension to nonlinear MR with discrete instruments, which we call nonparametric MR, that does not rely on the constant genetic effect assumption and is able to flexibly characterise more complex nonlinear relationships. We argue that our nonparametric MR estimates converge to a different version of localised average causal effect, which we call quantile average causal effect (QACE), that approximates the derivative of the true exposure-outcome function. Our method works well when the instrument is weak and takes only two or three values, a situation which most of the existing nonparametric IV methods struggle to deal with. Our simulation study shows that nonparametric MR consistently outperforms nonlinear MR under various scenarios and is robust to very weak instruments. We also illustrate our method using a real data example from UK Biobank, showing a nonlinear causal relationship between BMI and pulse rate.

Longitudinal causal inference

Semi-parametric G-computation for Longitudinal Analysis of Antiretroviral Therapy Andrew Spieker* Andrew Spieker, Bryan Shepherd,

The g-formula is a longitudinal generalization of standardization designed for settings in which there is time-varying treatment subject to time-dependent confounding. While the associated g-computation algorithm is straightforward and conceptually intuitive, its dependence upon parametric models is not ideal in practical circumstances in which variables possess distributions that are difficult to specify. We discuss the utility of cumulative probability models for use in g-computation as a way to relax certain forms of distributional assumptions. Simulations show this approach to be robust and feasible to implement in the real world. We illustrate the utility of this methodology through a study of core and ancillary agents comprising longitudinal antiretroviral therapy regimens and their effects on weight gain in a large cohort of persons living with HIV. Specifically, we hypothesize that modern integrase strand transfer inhibitors and tenofovir alafenamide are associated with greater mean weight gain as compared to other core and ancillary agents.

Machine Learning and Causal Inference

Finite sample and asymptotic properties of a cross-validated risk estimator of a nonparametric regimen-response curve estimate Cuong Pham* Cuong Pham, Ashkan Ertefaie,

Dynamic treatment strategies are the decision rules that map individual characteristics to a specific type of treatment. We focus on a parametric class of rules and define a regimen-response curve function as an expected value of the counterfactual outcome under a decision rule given a set of baseline variables. The regimen-response curve is a function of the decision rule's parameters, and it is also the optimizer of the curve corresponding to an optimal dynamic treatment strategy. Existing methods impose a parametric model on the regimen-response curve function and estimate the corresponding parameters using a marginal structural model approach. The parametric model is likely to be misspecified, particularly in time-varying settings that may lead to considerably suboptimal treatment strategies. To overcome this issue, we propose a data-adaptive approach that utilizes the highly adaptive lasso model to capture the true regimen-response curve function. We also propose a nonparametric cross-validated risk estimator of the estimated regimen-response curve.

We show that:
1. Our estimated regimen-response curve converges to the true curve in loss-based-dissimilarity faster than root-n.
2. The cross-validated risk estimator is asymptotically linear.
3. The cross-validation tuning parameter selector risk is bounded in a finite sample.

We conduct extensive simulation studies to confirm our theoretical results and examine the finite sample performance of our estimator.

Machine Learning and Causal Inference

Transfer Learning With Efficient Estimators To Optimally Leverage Historical Data in Analysis of Randomized Trials Lauren Liao* Lauren Liao, Alan Hubbard, Alejandro Schuler,

Inference from randomized control trials (RCTs) yields reliable causal interpretation but is often limited by the sample size. RCTs often answer similar questions that were done in observational studies. Opposed to RCTs, previous observational studies are often abundant in data although the causal link is less well-defined. While a typical RCT is analyzed alone, estimators have been developed to incorporate historical, observational studies to increase the estimator's efficiency. Previous estimators developed to incorporate historical data in RCT analyses often have unrealistic additional assumptions that can create a biased estimate. Instead, we propose using no additional assumptions on the historical data, nor the connection between the historical and trial data, to increase efficiency in analysis of RCTs. This is performed by leveraging machine learning to estimate generalized prognostic scores from the observational studies of similar nature. These generalized prognostic scores are predictions on the current study using historical models, and they are included in the current analysis as covariates in an efficient estimator. We demonstrate the utility of this estimator leveraging historical data on a randomized blood transfusion study of trauma patients.

Machine Learning and Causal Inference

Investigating neural networks to learn an intervention response function on a continuous exposure Mauricio Tec* Mauricio Tec, Oladimeji Mudele, Kevin Josey, Falco Bargagli Stoffi, Francesca Dominici,

This work investigates neural network (NN) techniques for estimating the expected change in an outcome of interest resulting from a hypothetical change in a continuous exposure. For example, a policymaker may want to know how many hospitalizations could be avoided by an intervention reducing air pollution from its current values in each location. We formulate this goal as learning an intervention response function (IRF) using the stochastic interventions (SI) framework (Hubbard & Van Der Laan, 2005). We illustrate how IRFs differ from exposure-response functions and can better support reasoning about certain intervention policies of interest, particularly under overlap violations.

We consider three mechanisms in which NNs can improve IRF estimation:
By learning a flexible model of the conditional density ratios required by IPTW and TMLE estimators for SIs (Diaz-Munoz & Van Der Laan, 2012);
By using architectures and priors promoting smoothness in the individual potential outcome curves as a function of the exposure;
By extending recent work on doubly robust estimation via targeted regularization for NNs (Shi et al., 2019; Nie et al., 2021) to the case of SIs.

These mechanisms are evaluated using previously proposed synthetic benchmark datasets for continuous treatments. Finally, we illustrate them in a real application to estimate the impact of a reduction in airborne exposure to toxic metals (Kodros et al., 2022) on hospitalizations among Medicare enrollees in the U.S.

Machine Learning and Causal Inference

Super Efficient Estimation for a Sieve of Statistical Models Ivana Malenica* Ivana Malenica, Mark van der Laan,

There is an increasing interest in estimating causal effects in dependent settings, where dependence is prevalent across time and/or samples. In order to make progress, assumptions on the statistical model are common: usually in the form of known network or Markov order, or conditional independence. In this work, we propose a sieve-based approach for data-adaptively picking a statistical model from an initial fully nonparametric space. We are concerned with statistical inference for a pathwise differentiable target parameter based on t time-points (or n samples), where there is enough independence that the canonical gradient is asymptotically normal. One example considered is estimation of the causal effect of an intervention at time t on the proximal outcome given the past, averaged over all observed times. We consider a sequence of nested models indexed by a multivariate real-valued parameter that approximates the actual statistical model. We propose a data-adaptive selector of the index, and estimate the target parameter with the corresponding targeted minimum loss estimator (TMLE). We show that, under regularity conditions, the proposed adaptive TMLE is asymptotically normal and super-efficient. This provides an important alternative to the TMLE for the actual statistical model which might be too large to be informative, but can be captured by a smaller model in a sieve, as is often the case in structured dependent settings.

Machine Learning and Causal Inference

Inference on Strongly Identified Functionals of Weakly Identified Functions Xiaojie Mao*
Xiaojie Mao,

In a variety of applications, including nonparametric instrumental variable (NPIV) analysis, proximal causal inference under unmeasured confounding, and missing-not-at-random data with shadow variables, we are interested in inference on a continuous linear functional (e.g., average causal effects) of nuisance function (e.g., NPIV regression) defined by conditional moment restrictions. These nuisance functions are generally weakly identified, in that the conditional moment restrictions can be severely ill-posed as well as admit multiple solutions. This is sometimes resolved by imposing strong conditions that imply the function can be estimated at rates that make inference on the functional possible. In this paper, we study a novel condition for the functional to be strongly identified even when the nuisance function is not; that is, the functional is amenable to asymptotically-normal estimation at root-n rates. The condition implies the existence of debiasing nuisance functions, and we propose penalized minimax estimators for both the primary and debiasing nuisance functions. The proposed nuisance estimators can accommodate flexible function classes, and can converge to fixed limits regardless of the identifiability of the nuisances. We use the penalized nuisance estimators to form a debiased estimator for the functional of interest and prove its asymptotic normality under generic high-level conditions, which provide for asymptotically valid confidence intervals.

Machine Learning and Causal Inference

Policy learning “without” overlap: Pessimism and generalized empirical Bernstein’s inequality Zhimei Ren* Ying Jin, Zhimei Ren, Zhuoran Yang, Zhaorann Wang,

We study offline policy learning, which utilizes observations collected a priori to learn an optimal individualized decision rule that achieves the best overall outcomes for a given population. Existing policy learning methods rely on a uniform overlap assumption, i.e., the propensities of exploring all actions for all individual characteristics are lower bounded. As one has no control over the data collection process, this assumption can be unrealistic in many situations, especially when the behavior policies are allowed to evolve over time with diminishing propensities for certain actions. We propose a new algorithm that optimizes lower confidence bounds (LCBs) — instead of point estimates — of the policy values. The LCBs are constructed using the behavior policies for collecting the offline data. Without assuming any uniform overlap condition, we establish a data-dependent upper bound for the suboptimality of our algorithm, which only depends on (i) the overlap for the optimal policy, and (ii) the complexity of the policy class. As an implication, for adaptively collected data, we ensure efficient policy learning as long as the propensities for optimal actions are lower bounded over time, while those for suboptimal ones are allowed to diminish arbitrarily fast. In our analysis, we develop a new self-normalized type concentration inequality for IPW-type estimators, generalizing the empirical Bernstein’s inequality to unbounded and non-i.i.d. data.

Machine Learning and Causal Inference

Adaptive designs for best arm identification and evaluation, with application to vaccine confidence messaging Molly Offer-Westort* Molly Offer-Westort, Leah Rosenzweig,

This paper contributes to the literature on adaptive experimental designs for best arm identification. In particular, we consider a setting where the experimenter wishes not only to learn the best arm, but also to evaluate response under that arm, or treatment effects with respect to the best arm and a control condition. Adaptive designs can help experimenters more efficiently learn which treatment condition among a set of alternatives will perform best on a given response measure. However, naive frequentist estimates of response under the best arm will be upwardly biased if they are calculated on the same data used to learn which arm is best. To resolve this, we consider a two-stage design where best arm identification is learned in an adaptive first stage, and evaluated alongside a control condition in a second stage. We show improvements in simple regret, and bias and precision of treatment effect estimation as compared to alternative designs where learning and evaluation are objectives.

In our application, we use an adaptive best-arm identification algorithm to optimize informational messaging on vaccines in an online study among Facebook users in Kenya and Nigeria. We demonstrate that optimized personalized messaging improves vaccine confidence and intentions to get vaccinated over a uniform public service announcement condition or a pure control condition.

Machine Learning and Causal Inference

Cross-Validated Decision Trees with Targeted Maximum Likelihood Estimation for Nonparametric causal mixtures analysis David McCoy* David McCoy, Alan Hubbard, Alejandro Schuler, Mark van der Laan,

People often encounter multiple simultaneous exposures (e.g. several drugs or pollutants). Policymakers are interested in setting safe limits, interdictions, or recommended dosage combinations based on a combination of thresholds, one per exposure. Setting these thresholds is difficult because all relevant interactions between exposures must be accounted for. Previous statistical methods have used parametric estimators which don’t directly address the question of superadditive or subadditive effects in a mixture and rely on unrealistic assumptions. Here we present an estimator that a) automatically identifies thresholds that maximize the differential effect of self-selecting exposure within the thresholded exposure region vs. outside of it; and which b) unbiasedly and efficiently estimates the magnitude of that differential effect. This is done by combining a tree-based search algorithm with a targeted maximum likelihood estimator using cross-validation. We provide open-source software (CVtreeMLE) that implements the method.

Machine Learning and Causal Inference

Personalization to One of Many Arms Rahul Ladhania* Rahul Ladhania, Jann Spiess, Lyle Ungar,

We consider learning personalized assignments among potentially many treatment arms from a randomized controlled trial. In a theoretical model, we illustrate how a high number of treatment arms makes finding the best arm hard, while we can still achieve sizable welfare gains from personalization by direct optimization. In a practical implementation, we propose methods that optimize treatment assignment specifically in the case of many treatment arms. First, we consider a regularized forest-based assignment algorithm based on greedy recursive partitioning that includes shrinkage across treatment arms. Second, we propose a clustering scheme that combines treatment arms with consistently similar outcomes. In a simulation study, we compare the performance of these approaches to predicting arm-wise outcomes separately, and document gains of directly optimizing the treatment assignment and including regularization and clustering in the underlying model construction.

Machine Learning and Causal Inference

A Comparison of Missing Imputation Method for Covariates in Propensity Score Analysis Using Random Forests Yongseok Lee* Yongseok Lee, Walter Leite,

Propensity Score Analysis (PSA) is a prominent method to alleviate selection bias in observational studies, but missing data in covariates is prevalent and must be dealt during propensity score estimation. Through Monte Carlo simulations, this study evaluates the use of imputation methods based on multiple random forests (RF) algorithms to handle missing data in covariates: MICE-RF (CALIBER), Proximity Imputation (PI), and missForest. The results indicated that PI and missForest outperformed other methods with respect to bias of average treatment effect (ATE) regardless of sample size and missing mechanisms. A demonstration of these five methods with PSA to evaluate the effect of participation in center-based care on children’s reading ability is provided using data from the Early Childhood Longitudinal Study (ECLS-K: 2011).

Machine Learning and Causal Inference

Tree Priors for Feature Selection in Average Treatment Effect Estimation Andrew Herren*
Andrew Herren, Richard Hahn,

This paper builds on previous theoretical work justifying feature selection for average treatment effects in the context of discrete covariates. We extend to continuous covariates using the adaptive discretization provided by decision trees and propose a novel decision tree prior that incorporates the estimated propensity score. The methods are justified using finite sample comparisons to existing Bayesian tree methods for treatment effect estimation as well as the broader class of machine learning estimators.

Machine Learning and Causal Inference

Taming Individualized Treatment Effects under Interference via Representation Learning with Graph Neural Networks Mauricio Tec* Mauricio Tec, Claudio Battiloro,

Causal inference is used to estimate the impact of a treatment on a specific outcome. A widely accepted assumption is that one unit’s treatment does not impact other units’ potential outcomes, which can be violated in practical settings and lead to biased estimates. While increasing research focuses on this issue, many studies have concentrated on average direct and spillover effects. Instead, this study considers individualized treatment effects (ITE) under interference. To this end, we propose and evaluate graph neural network (GNN) methods with the minimal assumption that the potential outcomes can be written as a symmetric function of the interfering units’ treatments and covariates. We formulate our learning task using the general framework of bipartite interference, which contains standard network interference as a special case. We formalize the estimand of interest and analyze the counterfactual generalization error based on distributional shift. Our experiments with synthetic data evaluate architectures and representation learning methods previously proposed for ITE estimation (under no interference). These results provide directions for designing GNN-based estimators of ITEs under interference. In addition, we demonstrate the utility of our approach in an application to estimate the health effects of implementing an intervention to reduce emissions in selected US coal-fired power plants.

Machine Learning and Causal Inference

Methods for obtaining counterfactual predictions and quantifying associated uncertainty using observational data Karla DiazOrdaz* Karla DiazOrdaz,

Prediction models, whether statistical or AI, are often used to help decision making. However, these approaches should not be used to answer 'what if' questions. Failure to recognise when the prediction estimand is causal leads to incorrect risk predictions and suboptimal treatment or policy decisions.
Our focus is counterfactual predictions, where for each individual we predict what their outcome would be under a hypothetical policy or treatment, assuming the causal structure is known, and there are no unobserved confounders.

Targeting a causal prediction estimand brings new challenges, because we can only use the observed (factual) treated sample to develop the model, but we must make predictions for the entire population. In the presence of confounding, the distribution of the factual treated may substantially differ from the target population (i.e. covariate shift). Further, we also consider situations where relevant variables are available at the model-building stage but are not available at deployment.

We review some existing methods allowing machine learning (e.g.DR-learner) and make a simpler proposal (based on inverse weighting) under covariate shift. We also implement distribution-free prediction intervals using conformal inference. We compare the methods in a simulation study and illustrate them in a real example using electronic health records to obtain counterfactual predictions for type 2 diabetes patients under different Hba1c lowering drugs choices.

Machine Learning and Causal Inference

Doubly Robust Inference for Hazard Ratio under Covariate-Induced Dependent Left Truncation with Machine Learning Yuyao Wang* Yuyao Wang, Andrew Ying, Ronghui Xu,

In prevalent cohort studies with follow-up, the time-to-event outcome is subject to left truncation leading to selection bias. For comparing the time-to-event outcome between treatment groups, Cox proportional hazards models accounting for confounders are typically considered. While such Cox models have addressed confounding, the selection bias caused by left truncation still needs to be handled. The partial likelihood approach with risk set adjustment can properly handle left truncation under the conditional quasi-independent left truncation assumption that the truncation time and the event time are independent on the observed region given the covariates involved in the Cox model. However, this assumption can be violated when the dependence between the left truncation time and the event time is induced by other covariates. Inverse probability of truncation weighting (IPW) leveraging additional covariate information can be used in this case, but it is sensitive to misspecification of the truncation model. In this work, we propose an augmented IPW estimator that has doubly robust properties: 1) model double-robustness, that is, it is consistent and asymptotically normal (CAN) when one of the two nuisance models is correctly specified; 2) rate double-robustness, that is, it is CAN when both of the nuisance parameters are consistent and the error product rate under the two nuisance models is faster than root-n.

Matching

A Surprising Granularity when Coarsening Continuous Variables with Coarsened Exact Matching Aran Canes* Aran Canes, Jigar Shah,

Iacus, King and Porro introduced the well-known method of Coarsened Exact Matching (CEM) in 2011. As they state, “The basic idea of CEM is to coarsen each variable by recoding so that substantively indistinguishable values are grouped and assigned the same numerical value.”

We’ve used CEM in retrospective observational studies for many years. What we wanted to know is whether common continuous variables, such as age, could be coarsened into large enough categories to allow for a reasonable number of matches, and thus the evaluation of the ATE, as well as be small enough so that the limits are substantively indistinguishable.

We were able to test these assumptions by looking at the effect of coarsening continuous variables on a prior version of the outcome. By creating increasingly more granular categories until the limits were statistically insignificant predictors of the prior outcome we tested the “Substantively indistinguishable” assumption. The number of categories for certain variables was in excess of one hundred—making it impossible to perform CEM and meet all assumptions.

Given the increasing use of CEM, we believe these results are important as showing that common-sense or intuitive coarsening of continuous variables may be violating the assumptions of CEM and leading to inaccurate measures of the treatment effect.

Matching

Propensity score augmentation in matching-based estimation of causal effects Ernesto Ulloa Perez* Ernesto Ulloa Perez, Alex Luedtke, Marco Carone,

When assessing the causal effect of a binary exposure using observational data, confounder imbalance across exposure arms must be addressed. Matching methods, including propensity score-based matching, can be used to deconfound the causal relationship of interest. They have been particularly popular in practice, at least in part due to their simplicity and interpretability. However, these methods can suffer from low statistical efficiency compared to many competing methods. In this work, we propose a novel matching-based estimator of the average treatment effect based on a suitably-augmented propensity score model. Our procedure can be shown to have greater statistical efficiency than traditional matching estimators whenever prognostic variables are available, and in some cases, can nearly reach the nonparametric efficiency bound. In addition to a theoretical study, we provide numerical results to illustrate our findings. Finally, we use our novel procedure to estimate the effect of circumcision on risk of HIV-1 infection using vaccine efficacy trial data.

Mediation

Mediation analysis with mediator and outcome missing not at random Shuozhi Zuo* Shuozhi Zuo, Debashis Ghosh, Peng Ding, Fan Yang,

Mediation analysis is widely used for investigating direct and indirect causal pathways through which an effect arises. However, many mediation analysis studies are often challenged by missingness in the mediator and outcome. In general, when the mediator and outcome are missing not at random, the direct and indirect effects are not identifiable without further assumptions. In this work, we study the identifiability of the direct and indirect effects under some interpretable missing not at random mechanisms. We evaluate the performance of statistical inference under those assumptions through simulation studies and illustrate the proposed methods via the National Job Corps Study.

Mediation

Calculating Mediation Effects of High Dimensional Radiomic Data Between Exposure and Outcome Emily Mastej* Emily Mastej, Debashis Ghosh,

Radiomics involves the mathematical extraction of quantitative features from medical images. While there is a wide range of radiomic-based prediction research, there is an issue with the clinical translation of these methods as well as their explainability due to the “black box” nature of deep learning algorithms. One solution to gain understanding of the mechanistic pathway between radiomic features and outcomes is to build causal inference models, specifically mediation models. We propose a downstream radiomics analysis method that uses high dimensional mediation to explore the causal pathway between an exposure and an outcome through a radiomic mediator. This method takes an exposure, radiomic features, and an outcome and finds principal directions of mediation (PDMs) or weighted groups of radiomic features that independently mediate the indirect effect of the exposure on the outcome with the largest indirect effect being mediated by the first PDM. We applied our method to T2 MRI radiomic data obtained from 203 subjects with either a glioblastoma or a glioma. Using IDH gene mutation status as the exposure and survival outcomes as the phenotype of interest, we used our method to find groups of radiomic features that were principal directions of mediation. Original tumor shape sphericity and wavelet HHL first order median were both found to be highly involved radiomic features in the first and second PDMs.

Mediation

Regression-based mediation sample size and power determinations Yingjin Zhang* Yingjin Zhang, Chung-Chou Ho Chang,

Mediation analysis has been widely applied in many disciplines to better understand the underlying mechanism. Sufficient sample size and statistical power are essential in the study design stage in order to ensure reliable results in research. Methods to calculate sample-size and power for mediation analysis, including the Sobel test for statistical significance of the indirect effect, Monte Carlo simulations for power, and bootstrap assessment of confidence interval, have been hampered by the lack of closed forms and thus require substantial amounts of computational simulations. Therefore, these existing methods have rarely been adopted by researchers due to computational complexity, absence of software options, and limited settings on the prespecified causal pathways. In this study, we propose and derive regression-based analytic formulas for sample size and power estimations associated to the inferences on the direct effect, indirect effect, and mediation proportion under the counterfactual mediation setting with 15 different combinations of types of the main exposure, mediator, and outcome variables. Our methods focus on the cross-sectional settings with one mediator and possible multiple measured exposure–outcome and mediator–outcome confounders. Our methods rely on the estimations of covariate effects and their variance-covariance matrices in the involved regression models, which can be either provided or calculated from pilot data sets.

Mediation

Mediating pathways of neighborhood violence on adverse pregnancy outcomes in California

Caitlin Chan* Caitlin Chan, Shelley Jung, Dana Goin, Kara Rudolph, Kristen Marchi, William Dow, Paula Braveman, Mahasin Mujahid, Mark van der Laan, Jennifer Ahern,

Community violence may contribute to adverse perinatal health outcomes through exposure to violence, buffering resources, and mediating mechanisms. In particular, the contribution of mediators such as unhealthy coping behaviors and medical conditions on perinatal outcomes is not well understood, and may be an important source of disparate impacts on historically minoritized populations.

We examined mediating pathways from acute changes in neighborhood violence to adverse perinatal outcomes, restricting analyses to within-neighborhood comparisons to control for time-constant neighborhood factors.

We combined California neighborhood violence data with hospital records of singleton live births from 2007-2011. We estimated the excess risk of infant mortality, neonatal mortality, preterm birth, gestational diabetes, and preeclampsia among birthing individuals exposed to acute neighborhood violence spikes. We employed targeted maximum likelihood estimation, adjusted for individual- (age, race, parity, education, insurance, conception year and season) and neighborhood- (temperature, precipitation, unemployment) level confounders. Substance use and maternal infection during pregnancy were analyzed as mediators.

Community violence was associated with elevated risk of gestational diabetes, preeclampsia, and preterm birth, with both substance use and maternal infection mediating these effects. Further analyses will estimate the magnitude of mediation effects by racial/ethnic group.

Mediation

Exploring causal mechanisms and quantifying direct and indirect effects using a joint modeling approach for recurrent and terminal events Cheng Zheng* Cheng Zheng, Fang Niu, Lei Liu,

Recurrent events are commonly encountered in biomedical studies. In many situations, there exists a terminal event, e.g., death, which is potentially related to recurrent events. Joint models of recurrent and terminal events have been proposed to address the correlation between the recurrent event and the terminal event. However, there is a dearth of suitable methods to rigorously investigate the causal mechanisms between specific exposures, recurrent events, and terminal events. For example, it is of interest to know whether preventing the happening of certain recurrent events could lead to better overall survival and how much of the total effect of the primary exposure of interest on the terminal event is through the recurrent events. In this work, we propose a formal causal mediation analysis method to compute the natural direct and indirect effects. A novel joint modeling approach is used to take the recurrent event process as the mediator and the survival endpoint as the outcome. This new joint modeling approach allows us to relax the commonly used “sequential ignorability” assumption. Simulation studies show our new model’s good finite sample performance in estimating both model parameters and mediation effects. We apply our method to an AIDS study to evaluate how much of the comparative effectiveness of the two treatments and the effect of CD4 counts on overall survival are mediated by recurrent opportunistic infections.

Multilevel Causal Inference

Multilevel Modeling under Multisite Quasi-experimental Trials: Considerations of Coding Strategies, Unbalanced Groups, and Heterogeneous Variances Xiao Liu* Qian Zhang, Xiao Liu, Zijun Ke,

Multilevel models are often used to make causal inferences under multisite quasi-experimental trials, where participants are non-randomly assigned to treatment and control groups within each site. We are interested in inferring the average treatment effect associated with the dichotomous Level-1 treatment indicator. The dichotomous treatment indicator can be coded mainly in three ways: dummy coding, unweighted effect coding, and weighted effect coding. Among the three coding strategies, dummy coding and unweighted effect coding assume a 1:1 ratio between the treatment and control groups in the population; in contrast, weighted effect coding is recommended over dummy coding / unweighted effect coding in the situation where group proportions are unequal in the population. The goals of the study include (a) to analytically obtain the average treatment effect estimates and compare them under different coding strategies, (b) to analytically examine whether and how the inference about the average treatment effect may be impacted by heterogeneous variances of Level-1 residuals, (c) to conduct a simulation study to examine the inference accuracy about the average treatment effect using different coding strategies considering equal and unequal proportions of the two groups in the population, variability of group proportions across sites, and heterogeneous Level-1 residual variances, and (d) to illustrate the comparisons among different coding strategies via an empirical example.

Propensity Scores

Using Fractional polynomials in the Marginal Structural Model: An Application to Thoracic Aortic Endovascular Repair Hang Nguyen* Hang Nguyen, Daniel Heitjan, Haekyung Jeon-Slaughter,

Thoracic aortic aneurysm is a potentially suddenly lethal condition that may manifest no symptoms and is often detected only during medical visits for other reasons. Once identified, the surgeon can treat the aneurysm with thoracic endovascular aortic repair (TEVAR), a procedure that can be done electively. Aneurysm size is known to have a significant impact on the risk of rupture; thus, size at repair is expected to be associated with survival. We have used the marginal structural model (MSM), estimated with inverse propensity score weighting, to identify the effect of aneurysm size on post-surgery survival. Particularly, we first estimate stabilized weights as a function of aneurysm size. We then fit an MSM that is parameterized as a fractional polynomial function. This method provides a flexible parameterization for a continuous exposure that avoids some of the issues associated with spline analysis — in particular, the selection of the number and locations of knots. The use of fractional polynomials is evidently novel in applied causal inference. We illustrate our approach by applying our method to data from the Vascular Quality Initiative TEVAR registry. Our results show that the lowest hazard occurs with operations at an aneurysm size of 60mm.

Propensity Scores

Bootstrapping for Propensity Score Analysis Jason Bryer* Jason Bryer,

As the popularity of propensity score methods for estimating causal effects in observational studies increase, the choices researchers have for which methods to use has also increased. Rosenbaum (2012) suggested that there are benefits for testing the null hypothesis more than once in observational studies. With the wide availability of high power computers resampling methods such as bootstrapping (Efron, 1979) have become popular for providing more stable estimates of the sampling distribution. This paper introduces the `PSAboot` package for R that provides functions for bootstrapping propensity score methods. It deviates from traditional bootstrapping methods by allowing for different sampling specifications for treatment and control groups, mainly to ensure the ratio of treatment-to-control observations are maintained. Additionally, this framework will provide estimates using multiple methods for each bootstrap sample. Two examples are discussed: the classic National Work Demonstration and PSID (Lalonde, 1986) study and a study on tutoring effects on student grades.

Randomized Studies

The Plausibility of Experimental Findings under Selective Reporting: An Application to Randomized Experiments on Voter Turnout by Proprietary Organizations Thomas Leavitt*
Thomas Leavitt, Donald Green,

In principle, voter turnout research is well suited to drawing valid causal inferences. Most studies of voter turnout in the past decade randomly assign interventions and rely on administrative data to measure outcomes. Moreover, academic researchers who study voter turnout have been early and earnest adopters of pre-registration, which lessens the threat of publication bias. However, an increasing share of experiments in this domain is conducted by consultants or staff working within organizations that do not register and disclose all of the studies they conduct. Such selective reporting is a source of the replication crisis, which has led prominent researchers to declare that published research findings are often false. This paper takes up the issue of post-study plausibility of experimental findings under selective reporting. We propose a randomization-based, Bayesian procedure that enables the principled incorporation of assumptions about selective reporting into the prior distribution of the average effect. We pair this procedure with a Bayesian sensitivity analysis whereby researchers can assess the robustness of posterior inferences under increasingly severe degrees of selective reporting. Unlike existing methods, our procedure enables judgments about the plausibility of individual experimental findings under selective reporting, although we also show the implications of our method for meta-analyses that aggregate across multiple experiments.

Randomized Studies

Balanced and robust randomized treatment assignments: the finite selection model

Ambarish Chattopadhyay* Ambarish Chattopadhyay, Carl Morris, Jose Zubizarreta,

The Finite Selection Model (FSM) was developed in the 1970s for the design of the RAND Health Insurance Experiment (HIE), one of the largest and most comprehensive social science experiments conducted in the U.S. The idea behind the FSM is that each treatment group takes its turns selecting units in a fair and random order to optimize a common criterion. At each of its turns, a treatment group selects the available unit that maximally improves the combined quality of its resulting group of units in terms of the criterion. In the HIE and beyond, we revisit, formalize, and extend the FSM as a general tool for experimental design.

Leveraging the idea of D-optimality, we propose and analyze a new selection criterion in the FSM. The FSM using the D-optimal selection function has no tuning parameters, is affine invariant, and when appropriate retrieves several classical designs such as randomized block and matched-pair designs. For multi-arm experiments, we propose algorithms to generate a fair and random selection order of treatments. We demonstrate FSM's performance in a case study based on the HIE and in ten randomized studies from the health and social sciences. We recommend the FSM be considered in experimental design for its conceptual simplicity, balance, and robustness.

Randomized Studies

Doubly robust nearest neighbors in factor models Raaz Dwivedi* Raaz Dwivedi, Katherine Tian, Sabina Tomkins, Predrag Klasnja, Susan Murphy, Devavrat Shah,

We introduce an improved variant of nearest neighbors for counterfactual inference in panel data settings where multiple units are assigned multiple treatments over multiple time points, each sampled with constant probabilities. We call this estimator a doubly robust nearest neighbor estimator and provide a high probability non-asymptotic error bound for the mean parameter corresponding to each unit at each time. Our guarantee shows that the doubly robust estimator provides a (near-)quadratic improvement in the error compared to nearest neighbor estimators analyzed in prior work for these settings.

Randomized Studies

Covariance Adjustment with Non-Experimental Units in Randomized Studies Joshua Wasserman* Joshua Wasserman, Ben Hansen,

Researchers estimating intervention effects from randomized experiments may perform covariance adjustment to improve the precision of their estimate. Cohen and Fogarty (2022), Guo and Basse (2021), and Lin (2013) are among the proponents of the procedure, citing that when the intervention is randomly assigned, the OLS estimate of the intervention effect will be no less precise than the difference-in-means estimate. Covariance adjustment may be particularly appealing when a vast collection of auxiliary control units exists. The “rebar” method in Sales, Hansen, and Rowan (2018) leverages unmatched control units, for example. However, potential differences between the experimental and auxiliary units lead Rubin and Thomas (2000) and Ho, Imai, King, and Stuart (2007) to suggest using only matched observations to fit the covariance adjustment model. It is an open question—and indeed depends on the data at hand—whether the gain in precision outweighs the possible bias. In light of this debate, we introduce our methodology and accompanying software for covariance-adjusted intervention effect estimates through an application in the educational setting. Our approach based on stacked estimating equations allows for propagation of error from the covariance adjustment model to the intervention effect model, where the two need not be fit to the same sample. As a result, it more accurately portrays the bias-variance tradeoff in covariance adjustment than preeminent methods and software.

Randomized Studies

Model-assisted analyses of cluster-randomized experiments Fangzhou Su* Fangzhou Su, Peng Ding,

Cluster-randomized experiments are widely used due to their logistical convenience and policy relevance. To analyze them properly, we must address the fact that the treatment is assigned at the cluster level instead of the individual level. Standard analytic strategies are regressions based on individual data, cluster averages, and cluster totals, which differ when the cluster sizes vary. These methods are often motivated by models with strong and unverifiable assumptions, and the choice among them can be subjective. Without any outcome modeling assumption, we evaluate these regression estimators and the associated robust standard errors from a design-based perspective where only the treatment assignment itself is random and controlled by the experimenter. We demonstrate that regression based on cluster averages targets a weighted average treatment effect, regression based on individual data is suboptimal in terms of efficiency, and regression based on cluster totals is consistent and more efficient with a large number of clusters. We highlight the critical role of covariates in improving estimation efficiency, and illustrate the efficiency gain via both simulation studies and data analysis. Moreover, we show that the robust standard errors are convenient approximations to the true asymptotic standard errors under the design-based perspective. Our theory holds even when the outcome models are misspecified, so it is model-assisted rather than model-based. We also extend the th

Randomized Studies

Design and Analysis of Temporal Experiments in Ride-Hailing Platforms Ruoxuan Xiong*
Ruoxuan Xiong, Alex Chin, Sean Taylor,

We study the design and analysis of temporal experiments, where an intervention is repeatedly applied to the same set of experimental units over time, and units' longitudinal observations are available for the estimation of treatment effects. The motivating setting is that a ride-hailing platform tests changes to marketplace algorithms, such as pricing and matching, and estimate effects from longitudinal outcomes, such as the rate at which ride requests are completed, at the city level. The design problem involves the planning of partitioning time periods into intervals and assigning the new intervention at the interval level. We propose an autoregressive specification for the outcomes, from which treatment effects can be efficiently estimated. Based on analyzing the statistical properties of this specification, we show that the efficiency of the design depends on three factors: carryover effects from interventions at earlier times, serial correlation in outcomes, and heterogeneity and periodicity of experiment times. We further propose a meta-analysis approach that leverages massive historical data to construct beliefs about the three factors and optimally design experiments. We further conduct a careful simulation study in realistic settings to build intuition and guidance for practitioners to optimally design temporal experiments.

Sensitivity Analysis

Bounds and semiparametric inference in L^infinity and L^2 sensitivity analysis for observational studies Yao Zhang* Yao Zhang, Qingyuan Zhao,

Sensitivity analysis for the unconfoundedness assumption is a crucial component of observational studies. The marginal sensitivity model has become increasingly popular for this purpose due to its interpretability and mathematical properties. After reviewing the original marginal sensitivity model that imposes a L^infinity constraint on the maximum logit difference between the observed and full data propensity scores, we introduce a more flexible L^2 analysis framework; sensitivity value is interpreted as the “average” minimum amount of unmeasured confounding in the analysis. We derive analytic solutions to the stochastic optimization problems under the L^2 model, which can be used to bound the average treatment effect (ATE). We obtain the efficient influence functions for the optimal values and use them to develop efficient one-step estimators. We show that multiplier bootstrap can be applied to construct a simultaneous confidence band of the ATE. Our proposed methods are illustrated by simulation and real-data studies.

Sensitivity Analysis

Quantifying and interpreting reverse causation bias in epidemiological studies with sensitivity analysis Jeremy Brown* Jeremy Brown,

In epidemiological studies reverse causation occurs when we want to estimate the effect of the exposure on the outcome, but the outcome affects exposure. This is a particular concern in cross-sectional and ecological studies, where temporality of exposure and outcome is not clear, therefore, these designs are rarely used for causal inference. In study designs more commonly used for causal inference, such as cohort studies, the temporal order of exposure and outcome can typically be ascertained and therefore reverse causation prevented. However, there may still be reverse causation bias if the outcome is misclassified, for example if it was present but undiagnosed at baseline. Using directed acyclic graphs we describe the structure of this bias, which is structurally analogous to bias due to a misclassified confounder. Methods to mitigate this bias are available, but are not always feasible. Bias formulas have not been commonly used, but have been derived and under certain assumptions are equivalent to unmeasured confounding bias formulas. We describe these formulas and, as an example, apply them to the results of a published case-control study examining the effect of oral contraceptives on uterine cancer. Applying plausible bias parameters to the observed odds ratio 2.35 (95% CI 1.29-4.26) led to bias-adjusted estimates ranging from 1.96 (95% CI 1.08-3.55) to 1.24 (95% CI 0.68-2.24) indicating the potential for reverse causation amongst other biases to alter findings.

Sensitivity Analysis

Sensitivity analysis for continuous exposures and binary outcomes in matched observational studies Jeffrey Zhang* Jeffrey Zhang, Dylan Small, Siyu Heng,

Sensitivity analysis frameworks for binary treatments under the matching design have been well-studied, for both continuous and binary responses. Methods for continuous exposures, however, are not well-developed. We discuss how to perform exact randomization inference and sensitivity analysis within the Rosenbaum model with continuous exposures and binary outcomes. We also introduce inference on a generalized notion of the attributable effect. We conduct numerical studies into the design sensitivity and power of various test statistics under several data-generating mechanisms. Finally, we apply our methods to a recent study of the effect of early childhood lead exposure on juvenile delinquency.

Sequential Testing and Anytime-Valid Inference

Anytime-Valid F-Tests for Faster Sequential Experimentation Through Covariate Adjustment Michael Lindon* Michael Lindon, Dae Woong Ham, Iavor Bojinov, Martin Tingley,

Multivariate linear regression models are commonly used to perform inference about average treatment effects. The experimentation platform at Netflix relies heavily on such models. We demonstrate that performing sequential “anytime-valid” inference is no harder than classical fixed-n inference. The confidence sequences and sequential p-values we provide depend on the same set of statistics as classical confidence intervals and p-values, that is, we provide drop-in replacements which generalize guarantees to hold uniformly across time. This enables Netflix to perform sequential covariate-adjusted A/B tests, enabling peeking and optional stopping through the time-uniform nature of the guarantees, and achieving tighter confidence sequences and faster stopping times through variance reduction. Formally, we introduce sequential F-tests and confidence sequences for subsets of coefficients of a linear model. In addition to treatment effect estimation, we present applications concerning sequential tests of treatment effect heterogeneity and model selection. Our approach is based on an invariant mixture martingale, which exploits group invariance properties of the linear model to provide time-uniform Type I error coverage guarantees regardless nuisance parameters. Our test statistic is based on a group invariant Bayes factor obtained from using a right-Haar prior over nuisance parameters, which bridges frequentist and Bayesian paradigms.

Synthetic Control Method

On the Asymptotics of Synthetic Control Methods Claudia Shi* Claudia Shi, Achille Nazaret, David Blei,

Synthetic control is a method for estimating the causal effects of large-scale interventions, such as the statewide effects of policy changes.

The idea of SC is to approximate the treated unit as a weighted combination of the control units.

The SC estimators use the pre-intervention outcomes to learn the weights and use those weights to approximate the counterfactual outcomes of the treated unit.

The existing asymptotic framework suggests that as the number of time points goes to infinity, we can get an unbiased estimate of the SC.

However, this asymptotic framework may be unrealistic in practice, because we often only have data from a few time points.

In this paper, we build on the fine-grained model in Shi et al. [Shi+22] and introduce a novel asymptotic framework for synthetic control.

In the fine-grained model, the units of analysis are “small units” (e.g. individuals in states), rather than “large units” (e.g. states).

With the formulation, we show that the variance in the SC estimate could be explained by each large unit containing a finite number of small units (i.e., there are a finite number of individuals in each state).

We derive a variance quantification method and a variance-minimizing estimator. The significance of these methods is that they reduce the variance in the SC estimate by leveraging external data. We first study the properties of these methods using synthetic data, we then apply them to a real-world case study.

Synthetic Control Method

Same Root Different Leaves: Time Series and Cross-Sectional Methods in Panel Data Dennis Shen* Dennis Shen, Peng Ding, Jasjeet Sekhon, Bin Yu,

A central goal in social science is to evaluate the causal effect of a policy. One dominant approach is through panel data analysis in which the behaviors of multiple units are observed over time. The information across time and space motivates two general approaches: (i) horizontal regression (i.e., unconfoundedness), which
exploits time series patterns, and (ii) vertical regression (e.g., synthetic controls), which exploits cross-sectional patterns. Conventional wisdom states that the two approaches are fundamentally different. We establish this position to be partly false for estimation but generally true for inference. In particular, we prove that both approaches yield identical point estimates under several standard settings. For the same point estimate, however, each approach quantifies uncertainty with respect to a distinct estimand. The confidence interval developed for one estimand may have incorrect coverage for another. This emphasizes that the source of randomness that
researchers assume has direct implications for the accuracy of inference.

Weighting

Balancing Covariates via Weighted Independence Measures for Continuous Exposures Xiao Wu* Xiao Wu, Trevor Hastie,

Covariate balance plays a central role in causal inference. In this paper, we study the framework for balancing covariates in observational studies with a continuous exposure. Firstly, we overview two closely related approaches: 1) the modeling approach that maximizes the fit of a propensity model for treatment assignment, and weights by the inverse of the estimated propensity density to achieve covariate balance in large samples; 2) the balancing approach that optimizes certain measures of the covariate balance in finite samples. We propose the use of weighted independent measures to diagnose the degree of covariate balance and to achieve the uniform approximate balance for covariate functions in a reproducing-kernel Hilbert space. We provide theoretical justification that the proposed weighting estimator could achieve minimized biases under certain outcome model specifications. In simulations, the proposed methods outperform existing methods in terms of covariate balance, effective sample sizes, absolute bias, and root mean squared errors.