"Causal Inference and Machine Learning"

Virtual Satellite Invited Session of the CEN-IBS & GMDS 2020

28 September, 3-5pm (CEST, Berlin time)

The event is free - to participate please register at:

https://www.eventbrite.de/e/cen-ibsgmds-invited-session-on-causal-inference-and-machine-learning-tickets-116222778459

Registered participants will receive the link to the online event a few days before the event.

Speakers, titles, abstracts:

Karla Diaz-Ordaz, London School of Hygiene and Tropical Medicine, UK

Machine Learning estimation of Causal estimands: why and how

Machine learning (ML) methods have received a lot of attention in recent years especially in settings with a large number of variables. However, causal effect estimation often involves counterfactuals, and prediction tools from the ML literature cannot be readily used for causal inference. In the last decade, major innovations have taken place incorporating supervised ML tools into estimators for causal parameters. This holds the promise of attenuating model misspecification issues.

In this talk, I will review recent developments incorporating machine learning in the estimation of causal effects, in particular the average treatment effect (ATE), under the "no unobserved confounding" and positivity assumptions. I will explain why naïve machine learning plug-in estimators should be avoided. We will then see how machine learning estimators of the working models can be combined with doubly robust estimators to give valid inferences.

I will then propose an algorithm for estimating Structural Nested Mean Models (SNMMs), based on g-estimation with machine-learning plug-ins for the working models, in settings with time-varying binary treatment and a continuous outcome. I will present preliminary work studying its performance in a simulation.

Jonas Peters, Dpt. of Mathematical Sciences, University of Copenhagen, Denmark

The hardness of conditional independence testing

It is a common saying that testing for conditional independence, i.e., testing whether two random vectors X and Y are independent, given Z, is a hard statistical problem if Z is a continuous random variable (or vector). In this work, we prove that conditional independence is indeed a particularly

difficult hypothesis to test for and is fundamentally harder than testing for unconditional independence, for example. Solving it requires carefully chosen assumptions on the data generating process. We also provide a conditional independence test, the generalised covariance measure (GCM), that is explicit about such assumptions.

Reference: Rajen Shah and Jonas Peters: "The Hardness of Conditional Independence Testing and the Generalised Covariance Measure", Annals of Statistics (to appear).

Oliver Dukes, Dpt. of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium

Assumption-lean inference for generalised linear model parameters

Inference for the parameters indexing generalised linear models is routinely based on the assumption that the model is correct and a priori specified. This is unsatisfactory because the chosen model is usually the result of a data-adaptive model selection process, which may induce excess uncertainty that is not usually acknowledged. Moreover, the assumptions encoded in the chosen model rarely represent some a priori known, ground truth, making standard inferences prone to bias, but also failing to give a pure reflection of the information that is contained in the data. Inspired by developments on assumption-free inference for so-called projection parameters, we here propose novel nonparametric definitions of main effect estimands and effect modification estimands. These reduce to standard main effect and effect modification parameters in generalised linear models when these models are correctly specified, but have the advantage that they continue to capture respectively the primary (conditional) association between two variables, or the degree to which two variables interact (in a statistical sense) in their effect on outcome, even when these models are misspecified. We achieve an assumption-lean inference for these estimands (and thus for the underlying regression parameters) by deriving their influence curve under the nonparametric model and invoking flexible data-adaptive (e.g., machine learning) procedures.

This is joint work with Stijn Vansteelandt.

Reference: https://arxiv.org/pdf/2006.08402.pdf

Andrea Rotnitzky, Dpt. of Economics, Universidad Torcuato Di Tella, Argentina; and Harvard School of Public Health, US

Optimal adjustment sets in non-parametric graphical models

(joint work with Ezequiel Smucler and Facundo Sapienza)

We consider the selection of potential confounding variables at the stage of the design of a planned observational study. Given a tentative non-parametric graphical causal model, possibly including unobservable variables, the aim is to select the set of observable covariates that both suffices to control for confounding under the model and yields a non-parametric estimator of the causal contrast of interest with smallest variance. For studies without unobservables aimed at assessing the effect of a static point exposure we show that graphical rules recently derived for identifying optimal

covariate adjustment sets in linear causal graphical models and treatment effects estimated via ordinary least squares also apply in the non-parametric setting. Moreover, we show that, in graphs with unobservable variables, but with at least one adjustment set fully observable, there exist adjustment sets that are optimal minimal (minimum), yielding non-parametric estimators with the smallest variance among those that control for observable adjustment sets that are minimal (of minimum cardinality). In addition, although a globally optimal adjustment set among observable adjustment sets does not always exist, we provide a sufficient condition for its existence. We provide polynomial time algorithms to compute the observable globally optimal (when it exists), optimal minimal, and optimal minimum adjustment sets. For studies aimed at assessing the effects of interventions at multiple time points, static or dynamic, we derive graphical rules for comparing certain pairs of time dependent adjustment sets but we show that no global graphical rule is possible for determining optimal time dependent adjustment sets, even in graphs without unobservables. Finally, for graphs without unobservables and point interventions, we provide a sound and complete graphical criterion for determining when a non-parametric optimally adjusted estimator of the population average causal effect contrast is semiparametric efficient under the non-parametric causal graphical model.

Reference: https://arxiv.org/pdf/2004.10521.pdf