Robust causal inference under covariate shift via worst-case subpopulation treatment effects

Sookyo Jeong  
Stanford University  
SOOKYO@STANFORD.EDU

Hongseok Namkoong  
Columbia University  
HN2377@COLUMBIA.EDU

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We propose a notion of worst-case treatment effect (WTE) across all subpopulations of a given size, a conservative notion of topline treatment effect. Compared to the average treatment effect (ATE) that solely relies on the covariate distribution of collected data, WTE is robust to unanticipated covariate shifts, and ensures reliable inference uniformly over underrepresented minority groups. We develop a semiparametrically efficient estimator for the WTE, leveraging machine learning-based estimates of heterogenous treatment effects and propensity scores. By virtue of satisfying a key (Neyman) orthogonality property, our estimator enjoys central limit behavior—oracle rates with true nuisance parameters—even when estimates of nuisance parameters converge at slower-than-parameteric rates. In particular, this allows using black-box machine learning methods to construct asymptotically exact confidence intervals for the WTE. For both observational and randomized studies, we prove that our estimator achieves the optimal asymptotic variance, by establishing a semiparametric efficiency lower bound. On real datasets, we illustrate the non-robustness of ATE under even small amounts distributional shift, and demonstrate that WTE allows us to guard against brittle findings that are invalidated by unanticipated covariate shifts.

Keywords: causal inference, covariate shift, semiparametrics, distributional robustness

Summary of paper

Evaluation of platform designs, clinical treatments, and policy programs are universally based on statistical inference of average treatment effects (ATE), a de facto standard practice. However, this practice is only effective when the data-generating distribution is representative of the overall population of interest, a requirement that is frequently violated. Data is often collected from a particular set of geospatial locations, and may not represent the population of interest (Hand, 2006; Blitzer et al., 2006; Daume III and Marcu, 2006; Saenko et al., 2010; Torralba and Efros, 2011). In addition to natural covariate shifts, datasets generated from both randomized and observational studies often lack diversity, leading the ATE to ignore adverse effects on underrepresented minority groups. Although elderly patients over the age of 65 account for 61% of new cancer cases and 70% of all cancer deaths, they comprised only 25% of oncology trial participants between 1993 and 1996 (Shenoy and Harugeri, 2015). Similarly, out of 10,000+ cancer clinical trials funded by the National Cancer Institute, less than 2% focused on racial minorities, and less than 5% of participants were non-white (Chen Jr et al., 2014).

1. Extended abstract. Full version available on arXiv with the same title.

© 2020 S. Jeong & H. Namkoong.
When there is heterogeneity in the treatment effect across subpopulations, mismatch between the data-generating distribution and actual covariate distributions of interest leads to pronounced failures. This is common in high-stakes applications such as medicine, where effects of medical treatments vary over patient-specific characteristics and socioeconomic demographic variables (Imai and Ratkovic, 2013; Gijsberts et al., 2015; Basu et al., 2017; Baum et al., 2017; Duan et al., 2019; Carvalho et al., 2019; Dorie et al., 2019). Symptoms and contributing factors of cardiovascular disease, cancer, and diabetes change across different age and ethnic groups in significant ways (Leigh et al., 2016), and elderly patients have worse outcomes from surgeries and are prone to adverse effects caused by comorbidities and concomitant drugs. Even large-scale randomized trials in medicine suffer from such biases, so that ATE estimates do not reliably evaluate treatment effects on the overall population due to bias in selection into the study (Shadish et al., 2002). A prominent example is the ACCORD (ACCORD Study Group, 2010) and SPRINT (SPRINT Research Group, 2015) trials that studied effects of treatments to lower blood pressure on cardiovascular disease. Despite the large sample sizes—\(n = 4733\) for ACCORD, and \(n = 9361\) for SPRINT—the topline conclusions of the two studies had different signs, and the mechanism behind the difference could not be explained by experts even ex-post (Basu et al., 2017).

One approach is to directly estimate the conditional average treatment effect (CATE), and adaptively find potential subgroups that exhibit heterogeneity. Recently, various statistical procedures using machine learning (ML) models have been developed to estimate the CATE (Feller and Holmes, 2009; Su et al., 2009; Imai and Ratkovic, 2013; Athey and Imbens, 2016; Powers et al., 2017; Shalit et al., 2017; Nie and Wager, 2017; Wager and Athey, 2018; Künzel et al., 2019). While recent progress shows promise in fine-grained evaluation of varying treatment effects, ML models are no panacea. They are optimized for average-case performance on the collected data, and perform poorly on minority subpopulations with different demographic groupings of race, gender, and age (Amodei et al., 2016; Grother et al., 2010; Hovy and Søgaard, 2015; Blodgett et al., 2016; Sapiezynski et al., 2017; Tatman, 2017; Rajpurkar et al., 2018). For example, Buolamwini and Gebru (2018) report that commercial gender classifiers’ misclassification error on darker-skinned females can be as large as 34%, compared to around 1% error rate on lighter-skinned males. In automatic video captioning, language identification, and academic recommender systems, similar variations in performance have been observed over different demographic groupings of race, gender, and age. Statistical models have been observed to lose predictive ability on particular regions of covariates (Meinshausen and Bühlmann, 2015), and resulting estimates of CATE are often unreliable, detecting heterogeneity when there is none (Rigdon et al., 2018). Subgroups with heterogeneous treatment effects identified by CATE estimates are often underpowered, and estimates of CATE are sensitive to modeling choices, even when ATE estimates align around the true value (Carvalho et al., 2019). Benefits of modern ML models should not come at the expense of underrepresented subpopulations, and in particular, there is growing concern on fairness and ethics in the medical community (Char et al., 2018; Rajkomar et al., 2018; Goodman et al., 2018; American Medical Association, 2018).

Moreover, deploying CATE estimators can be nontrivial when personalized treatments are infeasible due to operational constraints. Societal norms (e.g. fairness concerns) bar economic policies from discriminating over demographic groups, and personalization can require prohibitive amounts of infrastructure and resources. Subgroups may exhibit strategic behavior under personalized policies, rendering previous analysis obsolete.
Motivated by these challenges, we propose the worst-case treatment effect (WTE) across all subpopulations of a given size, a conservative notion of topline treatment effect. Compared to the ATE that solely relies on the covariate distribution of collected data, WTE is robust to unanticipated covariate shifts. By ensuring treatment effects remain valid uniformly across all subgroups, WTE guarantees reliability over underrepresented groups; if patients with age $> 70$, a specific genetic marker, and cardiovascular event history represent at least $\alpha\%$ of the collected data, then our worst-case treatment effect—defined over subpopulations larger than $\alpha\%$ of collected data—guarantees reliable inference over them.

Estimation of WTE requires estimating infinite dimensional nuisance parameters: individual’s treatment effect (outcome model), and probability of receiving treatment (propensity score). We propose and analyze a procedure that can leverage machine learning (ML) estimators for estimating these high-dimensional nuisance parameters. Our approach allows flexible use of black-box prediction models, and uses them conservatively so that when it finds a nonzero treatment effect, the treatment remains effective across all subpopulations of a specified size. Building on recent advances in semiparametric inference, we show our estimator enjoys central limit behavior even when ML-based estimates of nuisance parameters converge at slower-than-parametric rates. We prove a fundamental hardness result (semiparametric efficiency lower bound) for estimating the WTE, establishing that our estimator achieves the optimal asymptotic variance in both observational and randomized studies.

On a number of real datasets, we demonstrate that while decisions based on the ATE can be unreliable under natural covariate shifts, our worst-case subpopulation treatment effect provides a robust evaluation of the causal effect of treatment. Our worst-case approach is able to identify disadvantaged subpopulations based on a priori nontrivial demographic groupings, and guarantees uniformly good performance against underrepresented subpopulations that are larger than $\alpha$. Even when estimates of CATE vary significantly across different number of observations, our estimators of the WTE yield similar conclusions, a (empirical) stability property shared with estimators of ATE. Estimates of the WTE complements usual topline estimates of ATE, and guards against brittle findings that are invalidated by unanticipated covariate shifts.

References


