Proceedings of Machine Learning Research vol 125:1–6, 2020 33rd Annual Conference on Learning Theory

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

Sookyo Jeong SOOKYO@STANFORD.EDU
Stanford University

Hongseok Namkoong HN2377@COLUMBIA.EDU
Columbia University

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

[1] 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.

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 deomgraphic 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 reliablity 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-parameteric 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 α. 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

ACCORD Study Group. Effects of intensive blood-pressure control in type 2 diabetes mellitus. *New England Journal of Medicine*, 362(17):1575–1585, 2010.

American Medical Association. AMA passes first policy recommendations on augmented intelligence., 2018. URL www.ama-assn.org/ama-passes-first-policy-recommendations-augmented-intelligence.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, and Guoliang Chen. Deep speech 2: end-to-end speech recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 173–182, 2016.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Sanjay Basu, Jeremy B Sussman, and Rod A Hayward. Detecting heterogeneous treatment effects to guide personalized blood pressure treatment: a modeling study of randomized clinical trials. *Annals of Internal Medicine*, 166(5):354–360, 2017.

Aaron Baum, Joseph Scarpa, Emilie Bruzelius, Ronald Tamler, Sanjay Basu, and James Faghmous. Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the look ahead trial. *The Lancet Diabetes & Endocrinology*, 5(10):808–815, 2017.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics, 2006.

Su L. Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of Empirical Methods for Natural Language Processing*, pages 1119–1130, 2016.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

Carlos Carvalho, Avi Feller, Jared Murray, Spencer Woody, and David Yeager. Assessing treatment effect variation in observational studies: Results from a data challenge. *arXiv:1907.07592 [stat.ME]*, 2019.

Danton S Char, Nigam H Shah, and David Magnus. Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine*, 378(11):981, 2018.

Moon S Chen Jr, Primo N Lara, Julie HT Dang, Debora A Paterniti, and Karen Kelly. Twenty years post-nih revitalization act: enhancing minority participation in clinical trials (empact): laying the groundwork for improving minority clinical trial accrual: renewing the case for enhancing minority participation in cancer clinical trials. *Cancer*, 120:1091–1096, 2014.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Tony Duan, Pranav Rajpurkar, Dillon Laird, Andrew Y Ng, and Sanjay Basu. Clinical value of predicting individual treatment effects for intensive blood pressure therapy: A machine learning experiment to estimate treatment effects from randomized trial data. *Circulation: Cardiovascular Quality and Outcomes*, 12(3):e005010, 2019.

Avi Feller and Chris C Holmes. Beyond toplines: Heterogeneous treatment effects in randomized experiments. *Unpublished manuscript, Oxford University*, 2009.

Crystel M Gijsberts, Karlijn A Groenewegen, Imo E Hoefer, Marinus JC Eijkemans, Folkert W Asselbergs, Todd J Anderson, Annie R Britton, Jacqueline M Dekker, Gunnar Engström, Greg W Evans, et al. Race/ethnic differences in the associations of the framingham risk factors with carotid imt and cardiovascular events. *PLoS One*, 10(7):e0132321, 2015.

Steven N Goodman, Sharad Goel, and Mark R Cullen. Machine learning, health disparities, and causal reasoning. *Annals of Internal Medicine*, 2018.

Patrick J Grother, George W Quinn, and P Jonathon Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST Interagency/Internal Reports (NISTIR)*, 7709, 2010.

David J Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006.

Dirk Hovy and Anders Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, volume 2, pages 483–488, 2015.

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

J Adam Leigh, Manrique Alvarez, and Carlos J Rodriguez. Ethnic minorities and coronary heart disease: an update and future directions. *Current atherosclerosis reports*, 18(2):9, 2016.

Nicolai Meinshausen and Peter Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv:1712.04912 [stat.ML]*, 2017.

Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high-dimensions. *arXiv:1707.00102 [stat.ML]*, 2017.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 2018.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2018.

Joseph Rigdon, Michael Baiocchi, and Sanjay Basu. Preventing false discovery of heterogeneous treatment effect subgroups in randomized trials. *Trials*, 19(1):382, 2018.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proceedings of the European Conference on Computer Vision*, pages 213–226. Springer, 2010.

Piotr Sapiezynski, Valentin Kassarnig, and Christo Wilson. Academic performance prediction in a gender-imbalanced environment. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, volume 1, pages 48–51, 2017.

William R Shadish, Thomas D Cook, and Donald Thomas Campbell. *Experimental and quasi-experimental designs for generalized causal inference/William R. Shedish, Thomas D. Cook, Donald T. Campbell.* Boston: Houghton Mifflin,, 2002.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085, 2017.

Premnath Shenoy and Anand Harugeri. Elderly patients' participation in clinical trials. *Perspectives in clinical research*, 6(4):184, 2015.

SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. *New England Journal of Medicine*, 373(22):2103–2116, 2015.

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158, 2009.

Rrachel Tatman. Gender and dialect bias in YouTube's automatic captions. In *First Workshop on Ethics in Natural Langauge Processing*, volume 1, pages 53–59, 2017.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE, 2011.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.