The Journal of Open Source Software

txshift: Efficient estimation of the causal effects of stochastic interventions in R

Nima S. Hejazi[1, 2] **and David Benkeser**[3]

1 Graduate Group in Biostatistics, University of California, Berkeley **2** Center for Computational Biology, University of California, Berkeley **3** Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University

DOI: 10.21105/joss.02447

Software

- Review
- Repository
- Archive

Editor: Marcos Vital

Reviewers:

- @klmedeiros
- @joethorley

Submitted: 20 May 2020
Published: 07 October 2020

Summary

Statistical causal inference has traditionally focused on effects defined by inflexible static interventions, applicable only to binary or categorical exposures. The evaluation of such interventions is often plagued by many problems, both theoretical (e.g., non-identification) and practical (e.g., positivity violations); however, stochastic interventions provide a promising solution to these fundamental issues (Díaz & van der Laan, 2018). The `txshift` R package provides researchers in (bio)statistics, epidemiology, health policy, economics, and related disciplines with access to state-of-the-art statistical methodology for evaluating the causal effects of stochastic shift interventions on *continuous-valued* exposures. `txshift` estimates the causal effects of modified treatment policies (or "feasible interventions"), which take into account the natural value of an exposure in assigning an intervention level. To accommodate use in study designs incorporating outcome-dependent two-phase sampling (e.g., case-control), the package provides two types of modern corrections, both rooted in semiparametric theory, for constructing unbiased and efficient estimates, despite the significant limitations induced by such designs. Thus, `txshift` makes possible the estimation of the causal effects of stochastic interventions in experimental and observational study settings subject to real-world design limitations that commonly arise in modern scientific practice.

Statement of Need

Researchers seeking to build upon or apply cutting-edge statistical approaches for causal inference often face significant obstacles: such methods are usually not accompanied by robust, well-tested, and well-documented software packages. Yet coding such methods from scratch is often impractical for the applied researcher, as understanding the theoretical underpinnings of these methods requires advanced training, severely complicating the assessment and testing of bespoke causal inference software. What's more, even when such software tools exist, they are usually minimal implementations, providing support only for deploying the statistical method in problem settings untouched by the complexities of real-world data. The `txshift` R package solves this problem by providing an open source tool for evaluating the causal effects of flexible, stochastic interventions, applicable to categorical or continuous-valued exposures, while providing corrections for appropriately handling data generated by commonly used but complex two-phase sampling designs.

Background

Causal inference has traditionally focused on the effects of static interventions, under which the magnitude of the exposure is set to a fixed, prespecified value for each unit. The evaluation

of such interventions faces a host of issues, among them non-identification, violations of the assumption of positivity, and inefficiency. Stochastic interventions provide a promising solution to these fundamental issues by allowing for the target parameter to be defined as the mean counterfactual outcome under a hypothetically shifted version of the observed exposure distribution (Díaz & van der Laan, 2012). Modified treatment policies, a particular class of such interventions, may be interpreted as shifting the natural exposure level at the level of a given observational unit (Díaz & van der Laan, 2018; Haneuse & Rotnitzky, 2013).

Despite the promise of such advances in causal inference, real data analyses are often further complicated by economic constraints, such as when the primary variable of interest is far more expensive to collect than auxiliary covariates. Two-phase sampling is often used to bypass these limitations – unfortunately, these sampling schemes produce side effects that require further adjustment when formal statistical inference is the principal goal of a study. Among the rich literature on two-phase designs, Rose & van der Laan (2011) stand out for providing a study of nonparametric efficiency theory under a broad class of two-phase designs. Their work provides guidance on constructing efficient estimators of causal effects under general two-phase sampling designs.

`txshift`'s Scope

Building on these prior works, Hejazi et al. (2020c) outlined a novel approach for use in such settings: augmented targeted minimum loss (TML) and one-step estimators for the causal effects of stochastic interventions, with guarantees of consistency, efficiency, and multiple robustness despite the presence of two-phase sampling. These authors further outlined a technique that summarizes the effect of shifting an exposure variable on the outcome of interest via a nonparametric working marginal structural model, analogous to a dose-response analysis. The `txshift` software package, for the `R` language and environment for statistical computing (R Core Team, 2020), implements this methodology.

`txshift` is designed to facilitate the construction of TML and one-step estimators of the causal effects of modified treatment policies that shift the observed exposure value up (or down) by an arbitrary scalar δ, which may possibly take into account the natural value of the exposure (and, in future versions, the covariates). The `R` package includes tools for deploying these efficient estimators under outcome-dependent two-phase sampling designs, with two types of corrections: (1) a reweighting procedure that introduces inverse probability of censoring weights directly into relevant loss functions, as discussed in Rose & van der Laan (2011); as well as (2) an augmented efficient influence function estimating equation, studied more thoroughly by Hejazi et al. (2020c). `txshift` integrates with the `sl3` package (Coyle, Hejazi, Malenica, & Sofrygin, 2020) to allow for ensemble machine learning to be leveraged in the estimation of nuisance parameters. What's more, the `txshift` package draws on both the `hal9001` (Coyle, Hejazi, & van der Laan, 2019; Hejazi et al., 2020b) and `haldensify` (Hejazi et al., 2020a) `R` packages to allow each of the efficient estimators to be constructed in a manner consistent with the methodological and theoretical advances of Hejazi et al. (2020c), which require fast convergence rates of nuisance parameters to their true counterparts for efficiency of the resultant estimator.

Availability

The `txshift` package has been made publicly available both via GitHub and the Comprehensive `R` Archive Network. Use of the `txshift` package has been extensively documented in the package's `README`, two vignettes, and its `pkgdown` documentation website.

Acknowledgments

Nima Hejazi's contributions to this work were supported in part by a grant from the National Institutes of Health: T32 LM012417-02.

References

Coyle, J. R., Hejazi, N. S., Malenica, I., & Sofrygin, O. (2020). *sl3: Modern pipelines for machine learning and Super Learning*. https://github.com/tlverse/sl3. doi:10.5281/zenodo.1342293

Coyle, J. R., Hejazi, N. S., & van der Laan, M. J. (2019). *hal9001: The scalable highly adaptive lasso*. https://CRAN.R-project.org/package=hal9001. doi:10.5281/zenodo.3558313

Díaz, I., & van der Laan, M. J. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics*, *68*(2), 541–549. doi:10.1111/j.1541-0420.2011.01685.x

Díaz, I., & van der Laan, M. J. (2018). Stochastic treatment regimes. In *Targeted learning in data science: Causal inference for complex longitudinal studies* (pp. 167–180). Springer Science & Business Media. doi:10.1007/978-3-319-65304-4_14

Haneuse, S., & Rotnitzky, A. (2013). Estimation of the effect of interventions that modify the received treatment. *Statistics in medicine*, *32*(30), 5260–5277. doi:10.1002/sim.5907

Hejazi, N. S., Benkeser, D. C., & van der Laan, M. J. (2020a). *haldensify: Conditional density estimation with the highly adaptive lasso*. https://CRAN.R-project.org/package=haldensify. doi:10.5281/zenodo.3698329

Hejazi, N. S., Coyle, J. R., & van der Laan, M. J. (2020b). hal9001: Scalable highly adaptive lasso regression in R. *Journal of Open Source Software*, *5*(53), 2526. doi:10.21105/joss.02526

Hejazi, N. S., van der Laan, M. J., Janes, H. E., Gilbert, P. B., & Benkeser, D. C. (2020c). Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials. *Biometrics*. doi:10.1111/biom.13375

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rose, S., & van der Laan, M. J. (2011). A targeted maximum likelihood estimator for two-stage designs. *The International Journal of Biostatistics*, *7*(1), 1–21. doi:10.2202/1557-4679.1217