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 July 2020

Summary

The `txshift` `R` package aims to provide researchers in (bio)statistics, epidemiology, health policy, econometrics, and related disciplines with access to cutting-edge statistical methodology for evaluating the causal effects of *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. What's more, the package provides independent corrections for estimating such effects under two-phase sampling (e.g., case-control) designs, allowing for the methodology to be readily applied in a diversity of real-world experimental and observational studies.

Background

Causal inference has traditionally focused on the effects of static interventions, under which the magnitude of the treatment 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 schemes are often used to bypass such limitations – unfortunately, their use produces 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 such designs. Their work can be used to construct efficient estimators of causal effects under general two-phase sampling designs.

`txshift`'s Scope

Building on these prior works, Hejazi et al. (2020b) 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 even in 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 simple 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. The `R` package includes tools for deploying these efficient estimators under two-phase sampling designs, with two types of corrections: (1) a reweighting procedure that introduces inverse probability of censoring weights directly into an appropriate loss function, as discussed in Rose & van der Laan (2011); as well as (2) a correction based on the efficient influence function, studied more thoroughly by Hejazi et al. (2020b). `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` and `haldensify` `R` packages (Coyle, Hejazi, & van der Laan, 2019; Hejazi et al., 2020a) to allow each of the estimators to be constructed in a manner consistent with the theoretical claims of Hejazi et al. (2020b). The `txshift` package has been made publicly available via GitHub and will be submitted to the Comprehensive `R` Archive Network in the near future. 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., van der Laan, M. J., Janes, H. E., Gilbert, P. B., & Benkeser, D. C. (2020b). Efficient nonparametric inference on the effects of stochastic interventions under two-

phase sampling, with applications to vaccine efficacy trials. Retrieved from http://arxiv.org/abs/2003.13771

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