The Journal of Open Source Software

haldensify: Highly adaptive lasso conditional density estimation in R

Nima S. Hejazi [1], **Mark J. van der Laan** [2,3], **and David Benkeser** [4]

1 Department of Biostatistics, T.H. Chan School of Public Health, Harvard University **2** Division of Biostatistics, School of Public Health, University of California, Berkeley **3** Department of Statistics, University of California, Berkeley **4** Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University

DOI: 10.21105/joss.04522

Software

- Review
- Repository
- Archive

Editor: Mark A. Jensen

Reviewers:

- @turgeonmaxime
- @adibender

Submitted: 18 May 2022
Published: 22 September 2022

Summary

The `haldensify` R package serves as a toolbox for nonparametric conditional density estimation based on the highly adaptive lasso, a flexible nonparametric algorithm for the estimation of functional statistical parameters (e.g., conditional mean, hazard, density). Building upon an earlier proposal (Díaz & van der Laan, 2011), `haldensify` leverages the relationship between the hazard and density functions to estimate the latter by applying pooled hazard regression to a synthetic repeated measures dataset created from the input data, relying upon the framework of cross-validated loss-based estimation to yield an optimal estimator (Dudoit & van der Laan, 2005; van der Laan et al., 2004). While conditional density estimation is a fundamental problem in statistics, arising naturally in a variety of applications (including machine learning), it plays a critical role in estimating the causal effects of continuous- or ordinal-valued treatments. In such settings this covariate-conditional treatment density has been termed the *generalized propensity score* (Hirano & Imbens, 2004; Imai & Van Dyk, 2004), and, like its analog for binary treatments (Rosenbaum & Rubin, 1983), serves as a key ingredient in developing both inverse probability weighted and doubly robust estimators of causal effects (Díaz & van der Laan, 2012, 2018; Haneuse & Rotnitzky, 2013; Hejazi et al., 2022).

Statement of Need

Conditional density estimation is an important fundamental problem in the computational sciences and statistics, having garnered (independent) attention in machine learning (Sugiyama et al., 2012; Takeuchi et al., 2009), semiparametric estimation (Cheng & Chu, 2004; Qin, 1998), and causal inference (Díaz & van der Laan, 2011; Hirano & Imbens, 2004; Zhu et al., 2015). Techniques for the nonparametric estimation of this quantity, complete with asymptotic optimality guarantees, have received comparatively limited attention. Similarly, despite the critical role of the generalized propensity score in the estimation of the causal effects of continuous treatments, this nuisance parameter is usually estimated with restrictive parametric modeling strategies, ultimately sharply limiting the quality of downstream point estimates and corresponding statistical inference (e.g., hypothesis tests, confidence intervals). Approaches for flexibly estimating the generalized propensity score have received limited attention (Díaz & van der Laan, 2011; Zhu et al., 2015), and software implementations of these techniques are, to the best of our knowledge, exceedingly rare, compared to, for example, regression algorithms for estimating conditional means. `haldensify` aims to partially fill this gap by implementing a flexible, nonparametric estimator of a conditional (or marginal) density, appropriate for estimation of the generalized propensity score and useful for the construction of inverse probability weighted or doubly robust estimators of a class of causal effect parameters tailored to continuous treatments.

Conditional Density Estimation and Modern Causal Inference

Conditional density estimation is a challenging and fundamental problem in statistical learning theory. Owing to the high frequency with which conditional density estimation arises in statistics and machine learning, a wide range of techniques have been proposed – under a correspondingly wide range of assumptions. Some techniques are based in kernel smoothing (e.g., Takeuchi et al., 2009), others in specialized neural network architectures (e.g., Neuneier et al., 1994), and others still in the direct estimation of ratios of conditional densities (e.g., Sugiyama et al., 2012). Most approaches make restrictive (parametric) assumptions about the form of the underlying density functional or fail to achieve convergence rates (of the estimator to the true, underlying conditional density) necessary for semiparametric inference. As such, analysts must often negotiate a difficult tradeoff between tractability, ease of implementation, and optimality properties of the chosen estimator. To partially resolve this open challenge, `haldensify` implements a nonparametric conditional density estimation procedure, making few assumptions regarding the underlying form of the density functional, with rate-convergence guarantees compatible with modern semiparametric inference and causal machine learning.

The algorithm implemented in `haldensify` is an improved and tailored version of the proposal of Díaz & van der Laan (2011), who formulated a nonparametric conditional density estimator based on the relationship between the density and hazard functions. This algorithm proceeds by, first, partitioning the support of the dependent variable into a user-specified number of bins and recasting the input dataset into a repeated measures structure, in which each observational unit is represented by a variable number of records (with the terminal record corresponding to the position of the bin over the discretized support into which the observed value of the dependent variable falls). Next, the hazard probability, conditional on any covariates, of the dependent variable falling in a given bin along the discretized support is estimated by applying the highly adaptive lasso (HAL) algorithm (Benkeser & van der Laan, 2016; van der Laan, 2015, 2017) (in this case, for binary regression), via the `hal9001` package (Coyle et al., 2022; Hejazi, Coyle, et al., 2020); this step is often labeled "pooled hazards" regression. Under plausible assumptions on the global variation of the target functional, HAL has been shown to converge at a suitable rate ($\approx n^{-1/3}$ per Bibaut & Laan (2019)) for standard semiparametric efficiency theory to apply to any estimators incorporating this conditional density estimator; however, in this application, the ℓ_1 (i.e., lasso) penalty of the HAL estimator is updated to utilize a loss function suitable for density estimation (Dudoit & van der Laan, 2005; van der Laan et al., 2004). In a final step, the conditional hazard estimates are rescaled to conditional density estimates by dividing the estimated hazard probabilities by the respective bin widths.

The advantages derived from the flexibility and rate-convergence properties of this algorithm are especially apparent in causal inference problems with continuous treatments. In such problems, a key nuisance parameter is the generalized propensity score (GPS), the conditional density of the treatment, given covariates. This nuisance parameter is required to be well-estimated (in a rate-convergence sense) for the construction of asymptotically efficient estimators (e.g., of treatment effects), which attain the minimal possible variance in a given regularity class. Such estimators are desirable since, theoretically speaking, they admit the tightest confidence intervals and most sensitive hypothesis tests, making inference based upon these more informative for downstream decision making. For example, the GPS is a nuisance parameter required for estimation of the counterfactual mean of a modified treatment policy (MTP) (Díaz & van der Laan, 2018; Haneuse & Rotnitzky, 2013), a type of intervention that perturbs the natural (or observed) value of the treatment. Doubly robust estimators of this counterfactual quantity are implemented in the `txshift` `R` package (Hejazi & Benkeser, 2020, 2022), which relies upon `haldensify` for estimation of the GPS and has been used in estimating counterfactual vaccine efficacy based on MTPs interpretable as corresponding to hypothetical (next-generation) vaccines that modulate the activity of target immunologic biomarkers in vaccine efficacy clinical trials (Hejazi, van der Laan, et al., 2020). Alternative, asymptotically efficient and nonparametric inverse probability weighted (IPW) estimators (Ertefaie et al., 2022) of such counterfactual mean parameters are implemented in `haldensify`'s `ipw_shift()` function, which

constructs these IPW estimators by combining haldensify's GPS estimator (implemented in the eponymous haldensify() function) with the sieve estimation framework to select an asymptotically optimal IPW estimator with respect to criteria rooted in semiparametric efficiency theory; Hejazi et al. (2022) give a formal description of these novel IPW estimators.

haldensify's Scope

The haldensify R package combines the binning and hazard estimation strategy of Díaz & van der Laan (2011) with HAL regression (Benkeser & van der Laan, 2016), resulting in a flexible, nonparametric conditional density estimator. This procedure – accessible via the eponymous haldensify() function – relies upon the hal9001 R package (Coyle et al., 2022; Hejazi, Coyle, et al., 2020) for the HAL regression step and upon the origami R package (Coyle & Hejazi, 2018) for cross-validated selection of tuning parameters (e.g., number of bins, ℓ_1 regularization) so as to empirically minimize the negative log-density loss (Dudoit & van der Laan, 2005). haldensify additionally adjusts the proposal of Díaz & van der Laan (2011) to (1) incorporate sample-level weights and (2) apply HAL regression to the repeated measures data structure in a manner tailored for density estimation on the hazard scale. The nonparametric IPW estimators of Hejazi et al. (2022) have been implemented in the ipw_shift() function.

In order to ensure a simplified and minimal API, the haldensify package exposes only a limited set of functions: (1) the haldensify() function, which facilitates the estimation of conditional (or marginal) densities as described above, and (2) the ipw_shift() function, which implements nonparametric IPW estimators of the causal effect of an additive modified treatment policy. As IPW estimators require estimation of the generalized propensity score as an intermediate (nuisance) step, this latter function internally calls the former; moreover, the ipw_shift() function and the various selectors to which it provides access (e.g., selector_gcv() for estimator selection based on "global" cross-validation) have been studied from a theoretical-methodological perspective in Hejazi et al. (2022). The haldensify() function is complemented by appropriate predict() and plot() methods, the former to allow for the estimated conditional density to be evaluated at new values of the variable of interest and its conditioning set and the latter to visualize the resultant estimators along the regularization trajectory. The ipw_shift() function is accompanied by a corresponding confint() method to easily generate confidence intervals around the IPW point estimates. The S3 classes returned by both of these functions have custom print() methods to allow for their results to be easily inspected. Several internal utility functions, including, for example, cv_haldensify(), map_hazard_to_density(), and selector_dcar(), implement core aspects of the conditional density estimation and nonparametric IPW estimation methodology.

Availability

Future software development efforts will be focused primarily along two avenues: (1) improving the computational aspects of the conditional density estimation procedure, possibly to include random sampling from the internally generated repeated measures dataset, and (2) further adjustments to the undersmoothing estimator selection strategies implemented for nonparametric IPW estimation, to be based on future methodological progress. Software maintenance efforts will focus on ensuring that the package remains compatible with future versions of the hal9001 package (Coyle et al., 2022; Hejazi, Coyle, et al., 2020). Currently, stable releases of the haldensify package are made available via the Comprehensive R Archive Network (CRAN, R Core Team, 2022) at https://CRAN.R-project.org/package=haldensify, while development efforts are carried out on the package's version-controlled repository, publicly hosted at https://github.com/nhejazi/haldensify. To date (mid-September 2022), CRAN records indicate that haldensify has been downloaded over 14,400 times.

Acknowledgments

NSH's contributions to this work were supported in part by a grant from the National Science Foundation (award number DMS 2102840).

References

Benkeser, D., & van der Laan, M. J. (2016). The highly adaptive lasso estimator. *Proceedings of the International Conference on Data Science and Advanced Analytics*, *2016*, 689. https://doi.org/10.1109/dsaa.2016.93

Bibaut, A. F., & Laan, M. J. van der. (2019). Fast rates for empirical risk minimization with cadlag losses with bounded sectional variation norm. *arXiv Preprint arXiv:1907.09244*. https://arxiv.org/abs/1907.09244

Cheng, K. F., & Chu, C.-K. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, *10*(4), 583–604. https://doi.org/10.3150/bj/1093265631

Coyle, J. R., & Hejazi, N. S. (2018). origami: A generalized framework for cross-validation in R. *The Journal of Open Source Software*, *3*(21). https://doi.org/10.21105/joss.00512

Coyle, J. R., Hejazi, N. S., Phillips, R. V., van der Laan, L. W., & van der Laan, M. J. (2022). *hal9001: The scalable highly adaptive lasso*. https://doi.org/10.5281/zenodo.3558313

Díaz, I., & van der Laan, M. J. (2011). Super learner based conditional density estimation with application to marginal structural models. *International Journal of Biostatistics*, *7*(1), 1–20. https://doi.org/10.2202/1557-4679.1356

Díaz, I., & van der Laan, M. J. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics*, *68*(2), 541–549. https://doi.org/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. https://doi.org/10.1007/978-3-319-65304-4_14

Dudoit, S., & van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, *2*(2), 131–154. https://doi.org/10.1016/j.stamet.2005.02.003

Ertefaie, A., Hejazi, N. S., & van der Laan, M. J. (2022). Nonparametric inverse probability weighted estimators based on the highly adaptive lasso. *Biometrics*, *(in press)*. https://arxiv.org/abs/2005.11303

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

Hejazi, N. S., & Benkeser, D. C. (2020). txshift: Efficient estimation of the causal effects of stochastic interventions in R. *Journal of Open Source Software*, *5*(54), 2447. https://doi.org/10.21105/joss.02447

Hejazi, N. S., & Benkeser, D. C. (2022). *txshift: Efficient estimation of the causal effects of stochastic interventions*. https://doi.org/10.5281/zenodo.4070042

Hejazi, N. S., Benkeser, D., Díaz, I., & van der Laan, M. J. (2022). *Efficient estimation of modified treatment policy effects based on the generalized propensity score*. https://arxiv.org/abs/2205.05777

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

Hejazi, N. S., van der Laan, M. J., Janes, H. E., Gilbert, P. B., & Benkeser, D. C. (2020). Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials. *Biometrics*, *77*(4), 1241–1253. https://doi.org/10.1111/biom.13375

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, *226164*, 73–84.

Imai, K., & Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, *99*(467), 854–866.

Neuneier, R., Hergert, F., Finnoff, W., & Ormoneit, D. (1994). Estimation of conditional densities: A comparison of neural network approaches. *International Conference on Artificial Neural Networks*, 689–692. https://doi.org/10.1007/978-1-4471-2097-1_162

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, *85*(3), 619–630. https://doi.org/10.1093/biomet/85.3.619

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1017/cbo9780511810725.016

Sugiyama, M., Suzuki, T., & Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Takeuchi, I., Nomura, K., & Kanamori, T. (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Computation*, *21*(2), 533–559. https://doi.org/10.1162/neco.2008.10-07-628

van der Laan, M. J. (2015). *A generally efficient targeted minimum loss based estimator* (No. 343). University of California, Berkeley. https://biostats.bepress.com/ucbbiostat/paper343/

van der Laan, M. J. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *International Journal of Biostatistics*, *13*(2). https://doi.org/10.1515/ijb-2015-0097

van der Laan, M. J., Dudoit, S., & Keles, S. (2004). Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, *3*(1), 1–23. https://doi.org/10.2202/1544-6115.1036

Zhu, Y., Coffman, D. L., & Ghosh, D. (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference*, *3*(1), 25–40. https://doi.org/10.1515/jci-2014-0022