The Journal of Open Source Software

hal9001: Scalable highly adaptive lasso regression in R

Nima S. Hejazi[1, 2, 4], **Jeremy R. Coyle**[2], **and Mark J. van der Laan**[2, 3, 4]

1 Graduate Group in Biostatistics, University of California, Berkeley **2** Division of Biostatistics, School of Public Health, University of California, Berkeley **3** Department of Statistics, University of California, Berkeley **4** Center for Computational Biology, University of California, Berkeley

DOI: 10.21105/joss.02526

Software

- Review
- Repository
- Archive

Editor: Mikkel Meyer Andersen

Reviewers:

- @daviddewhurst
- @rrrlw

Submitted: 24 June 2020
Published: 26 September 2020

Summary

The `hal9001` R package provides a computationally efficient implementation of the *highly adaptive lasso* (HAL), a flexible nonparametric regression and machine learning algorithm endowed with several theoretically convenient properties. `hal9001` pairs an implementation of this estimator with an array of practical variable selection tools and sensible defaults in order to improve the scalability of the algorithm. By building on existing R packages for lasso regression and leveraging compiled code in key internal functions, the `hal9001` R package provides a family of highly adaptive lasso estimators suitable for use in both modern large-scale data analysis and cutting-edge research efforts at the intersection of statistics and machine learning, including the emerging subfield of computational causal inference (Wong, 2020).

Background

The highly adaptive lasso (HAL) is a nonparametric regression function capable of estimating complex (e.g., possibly infinite-dimensional) functional parameters at a fast $n^{-1/3}$ rate under only relatively mild conditions (Bibaut & van der Laan, 2019; van der Laan, 2017; van der Laan & Bibaut, 2017). HAL requires that the space of the functional parameter be a subset of the set of càdlàg (right-hand continuous with left-hand limits) functions with sectional variation norm bounded by a constant. In contrast to the wealth of data adaptive regression techniques that make strong local smoothness assumptions on the true form of the target functional, HAL regression's assumption of a finite sectional variation norm constitutes only a *global* smoothness assumption, making it a powerful and versatile approach. The `hal9001` package primarily implements a zeroth-order HAL estimator, which constructs and selects by lasso penalization a linear combination of indicator basis functions, minimizing the loss-specific empirical risk under the constraint that the L_1-norm of the resultant vector of coefficients be bounded by a finite constant. Importantly, the estimator is formulated such that this finite constant is the sectional variation norm of the target functional.

Intuitively, construction of a HAL estimator proceeds in two steps. First, a design matrix composed of basis functions is generated based on the available set of covariates. The zeroth-order HAL makes use of indicator basis functions, resulting in a large, sparse matrix with binary entries; higher-order HAL estimators, which replace the use of indicator basis functions with splines, have been formulated, with implementation in a nascent stage. Representation of the target functional f in terms of indicator basis functions partitions the support of f into knot points, with such basis functions placed over subsets of sections of f. Generally, numerous basis functions are created, with an appropriate set of indicator bases then selected through lasso penalization. Thus, the second step of fitting a HAL model is performing L_1-penalized regression on the large, sparse design matrix of indicator bases. The selected HAL regression model approximates the sectional variation norm of the target functional as the absolute sum

of the estimated coefficients of indicator basis functions. The L_1 penalization parameter λ can be data adaptively chosen via a cross-validation selector (van der Laan & Dudoit, 2003; van der Vaart, Dudoit, & van der Laan, 2006); however, alternative selection criteria may be more appropriate when the estimand functional is not the target parameter but instead a nuisance function of a possibly complex parameter (e.g., van der Laan, Benkeser, & Cai, 2019; Ertefaie, Hejazi, & van der Laan, 2020). An extensive set of simulation experiments were used to assess the prediction performance of HAL regression (Benkeser & van der Laan, 2016); these studies relied upon the subsequently deprecated `halplus` `R` package.

`hal9001`'s core functionality

The `hal9001` package, for the `R` language and environment for statistical computing (R Core Team, 2020), aims to provide a scalable implementation of the HAL nonparametric regression function. To provide a single, unified interface, the principal user-facing function is `fit_hal()`, which, at minimum, requires a matrix of predictors `X` and an outcome `Y`. By default, invocation of `fit_hal()` will build a HAL model using indicator basis functions for up to a limited number of interactions of the variables in `X`, fitting the penalized regression model via the lasso procedure available in the extremely popular `glmnet` `R` package (Friedman, Hastie, & Tibshirani, 2009). As creation of the design matrix of indicator basis functions can be computationally expensive, several utility functions (e.g., `make_design_matrix()`, `make_basis_list()`, `make_copy_map()`) have been written in C++ and integrated into the package via the `Rcpp` framework (Eddelbuettel, 2013; Eddelbuettel et al., 2011). `hal9001` additionally supports the fitting of standard (Gaussian), logistic, and Cox proportional hazards models (via the `family` argument), including variations that accommodate offsets (via the `offset` argument) and partially penalized models (via the `X_unpenalized` argument).

Over several years of development and usage, it was found that the performance of HAL regression can suffer in high-dimensional settings. To alleviate these computational limitations, several screening and filtering approaches were investigated and implemented. These include screening of variables prior to creating the design matrix and filtering of indicator basis functions (via the `reduce_basis` argument) as well as early stopping when fitting the sequence of HAL models in the L_1-norm penalization parameter λ. Future software development efforts will continue to improve upon the computational aspects and performance of the HAL regression options supported by `hal9001`. Currently, stable releases of the `hal9001` package are made available on the Comprehensive `R` Archive Network at https://CRAN.R-project.org/package=hal9001, while both stable (branch `master`) and development (branch `devel`) versions of the package are hosted at https://github.com/tlverse/hal9001. Releases of the package use both GitHub and Zenodo (https://doi.org/10.5281/zenodo.3558313).

Applications

As `hal9001` is the canonical implementation of the highly adaptive lasso, the package has been relied upon in a variety of statistical applications. Speaking generally, HAL regression is often used in order to develop efficient estimation strategies in challenging estimation and inference problems; thus, we interpret *statistical applications* of HAL regression chiefly as examples of novel theoretical developments that have been thoroughly investigated in simulation experiments and with illustrative data analysis examples. In the sequel, we briefly point out a few recently successful examples:

- Ju, Benkeser, & van der Laan (2020) formulate a procedure based on HAL regression that allows the construction of asymptotically normal and efficient estimators of causal effects that are robust to the presence of instrumental variables, which can often lead

to severe issues for estimation and inference (Hernán & Robins, 2020). While a variety of procedures have been proposed to overcome the issues posed by instrumental variables, a particularly successful idea was given by Shortreed & Ertefaie (2017), who proposed standard lasso regression to select covariates for the exposure model based on an estimated outcome model. The work of Ju et al. (2020) replaces the standard lasso with HAL regression, effectively screening for *infinitesimal instrumental basis functions* rather than instrumental variables, providing much enhanced flexibility. Here, the authors demonstrate how HAL regression provides exceptionally fine-grained control over screening problematic covariates while simultaneously facilitating the construction of causal effect estimators with desirable asymptotic properties.

- Díaz & Hejazi (2020) introduce novel mediation effects based on joint stochastic interventions on exposure and mediator variables. To complement the new causal effects outlined in their work, these authors introduce efficient estimators that rely upon a fast rate of convergence of nuisance parameter estimators to their true counterparts. As the authors note, HAL is currently the only machine learning algorithm for which such a fast rate of convergence can rigorously be proven under minimal global smoothness assumptions. By relying upon HAL regression for the construction of their proposed estimators, Díaz & Hejazi (2020) advance not only the state-of-the-art in causal mediation analysis but also provide evidence, in both simulation experiments and an illustrative data analysis, of how HAL regression can be brought to bear on challenging causal inference problems to develop flexible and robust estimation strategies.
- Hejazi et al. (2020) develop novel theoretical insights for building efficient estimators of causal effects under two-phase sampling designs, relying upon the flexibility and fast convergence of HAL regression at the core of their theoretical contributions. Corrections for two-phase sampling, a family of procedures for developing efficient estimators of full-sample effects in spite of censoring introduced by the second-phase subsample, have received much attention, though developments applicable to large, unrestricted statistical models have been limited. These authors provide a formulation and theory for utilizing causal effect estimators, based on data subject to two-phase sampling, that attain asymptotic efficiency by way of the fast convergence rate of HAL regression. In effect, this works demonstrates that HAL regression has properties suitable for both flexible estimation and efficient inference in settings with complex data structures. The authors make their methodology available in the `txshift` R package (Hejazi & Benkeser, 2020a, 2020b), which relies upon `hal9001`. These authors additionally provide examples in simulation experiments and a re-analysis of a recent HIV-1 vaccine efficacy trial using their proposed statistical approach.
- Ertefaie et al. (2020) provide a considered study of the construction of inverse probability weighted (IPW) estimators that rely upon HAL regression in the estimation of the exposure mechanism. While IPW estimators classically require the use of parametric models of the exposure mechanism, these authors propose and investigate novel variants of these estimators that instead rely upon the fast convergence rate of HAL regression for the required nuisance parameter functional. In particular, Ertefaie et al. (2020) show through theoretical advances, several simulation experiments, and an illustrative data analysis of data from the well-documented NHEFS study that IPW estimators based on HAL regression can be made asymptotically linear and even efficient under an undersmoothing-based debiasing procedure. In so doing, the authors simultaneously advance the literatures on HAL regression and on IPW estimation, establishing the interface between the two as an area of viable future research. Notably, in demonstrating their proposed IPW estimators with the NHEFS data, the authors show that IPW estimators based on HAL regression can yield meaningful substantive conclusions without the typically restrictive parametric assumptions required for IPW estimation.

As further theoretical advances continue to be made with HAL regression, and the resultant statistical methodology explored, we expect both the number and variety of such examples to steadily increase.

References

Benkeser, D., & van der Laan, M. J. (2016). The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE. doi:10.1109/dsaa.2016.93

Bibaut, A. F., & van der Laan, M. J. (2019). Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm. *arXiv preprint arXiv:1907.09244*.

Díaz, I., & Hejazi, N. S. (2020). Causal mediation analysis for stochastic interventions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(3), 661–683. doi:10.1111/rssb.12362

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. Springer.

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., et al. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18.

Ertefaie, A., Hejazi, N. S., & van der Laan, M. J. (2020). Nonparametric inverse probability weighted estimators based on the highly adaptive lasso. Retrieved from http://arxiv.org/abs/2005.11303

Friedman, J., Hastie, T., & Tibshirani, R. (2009). Glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, *1*(4).

Hejazi, N. S., & Benkeser, D. C. (2020a). *txshift: Efficient estimation of the causal effects of stochastic interventions*. Retrieved from https://github.com/nhejazi/txshift

Hejazi, N. S., & Benkeser, D. C. (2020b). txshift: Efficient estimation of the causal effects of stochastic interventions in R. *under review at Journal of Open Source Software*.

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*. doi:10.1111/biom.13375

Hernán, M. A., & Robins, J. M. (2020). Causal inference: What if. *Boca Raton: Chapman & Hill/CRC*.

Ju, C., Benkeser, D., & van der Laan, M. J. (2020). Robust inference on the average treatment effect using the outcome highly adaptive lasso. *Biometrics*, *76*(1), 109–118. doi:10.1111/biom.13121

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/

Shortreed, S. M., & Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, *73*(4), 1111–1122. doi:10.1111/biom.12679

van der Laan, M. J. (2017). A generally efficient targeted minimum loss based estimator based on the Highly Adaptive Lasso. *The International Journal of Biostatistics*. doi:10.1515/ijb-2015-0097

van der Laan, M. J., Benkeser, D., & Cai, W. (2019). Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. *arXiv preprint arXiv:1908.05607*.

van der Laan, M. J., & Bibaut, A. F. (2017). Uniform consistency of the highly adaptive lasso estimator of infinite-dimensional parameters. *arXiv preprint arXiv:1709.06256*.

van der Laan, M. J., & Dudoit, S. (2003). *Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite*

sample oracle inequalities and examples (No. 130). Division of Biostatistics, University of California, Berkeley.

van der Vaart, A. W., Dudoit, S., & van der Laan, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, *24*(3), 351–371. doi:10.1524/stnd.2006.24.3.351

Wong, J. C. (2020). Computational causal inference. *arXiv preprint arXiv:2007.10979*.