The Journal of Open Source Software

hdme: High-Dimensional Regression with Measurement Error

Øystein Sørensen[1]

1 Center for Lifespan Changes in Brain and Cognition, Department of Psychology, University of Oslo

DOI: 10.21105/joss.01404

Software

- Review
- Repository
- Archive

Submitted: 23 April 2019
Published: 19 May 2019

Summary

Many problems in science involve using measured variables to explain an outcome of interest using some statistical regression model. In high-dimensional problems, characterized by having a very large number of variables, one often focuses on finding a subset of variables with good explanatory power. An example from cancer research involves finding gene expressions or other biomarkers which can explain disease progression, from a large set of candidates (Kristensen et al., 2014). Another example is customer analytics, where it may be of interest to find out which variables predict whether customers will return or not, and variables of interest include factors like previous purchasing patterns, demographics, and satisfaction measures (Baesens, 2014).

The lasso (Tibshirani, 1996) and the Dantzig selector (Candes & Tao, 2007; James & Radchenko, 2009) are popular methods for variable selection in this type of problems, combining computational speed with good statistical properties (Bühlmann & Geer, 2011). In many practical applications, the process of measuring the variables of interest is subject to measurement error (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006), but this additional source of noise is neglected by the aforementioned models. Such measurement error has been shown to lead to worse variable selection properties of the lasso (Sørensen, Frigessi, & Thoresen, 2015), typically involving an increased number of false positive selections. A corrected lasso has been proposed and analyzed by Loh & Wainwright (2012) for linear models and Sørensen et al. (2015) for generalized linear models. It has been applied by Vasquez, Hu, Roe, Halonen, & Guerra (2019) in a problem involving measurement of serum biomarkers. For the Dantzig selector, Rosenbaum & Tsybakov (2010) proposed the Matrix Uncertainty Selector (MUS) for linear models, which was extended to the generalized linear model case by Sørensen, Hellton, Frigessi, & Thoresen (2018) with an algorithm named GMUS (Generalized MUS).

`hdme` is an R (R Core Team, 2018) package containing implementations of both the corrected lasso and the MU selector for high-dimensional measurement error problems. Its main functions are `gmus()` and `corrected_lasso()`. Additional functions provide opportunities for hyperparameter tuning using cross-validation or the elbow rule (Rosenbaum & Tsybakov, 2010), and plotting tools for visualizing the model fit. The underlying numerical procedures are implemented in `C++` using the `RcppArmadillo` package (Eddelbuettel & Sanderson, 2014) and linear programming with `Rglpk` (Theussl & Hornik, 2019). `hdme` is available from the comprehensive R archive network (CRAN) at https://CRAN.R-project.org, and the latest development version is available at https://github.com/osorensen/hdme. The package vignette, which can be opened in `R` with the command `vignette("hdme")`, contains a step-by-step introduction to the models implemented in the package.

Acknowledgements

The author would like to thank Arnoldo Frigessi, Kristoffer Herland Hellton, and Magne Thoresen for helpful discussions while developing the package.

References

Baesens, B. (2014). *Analytics in a big data world: The essential guide to data science and its applications* (1st ed.). Wiley Publishing.

Bühlmann, P., & Geer, S. van de. (2011). *Statistics for high-dimensional data*. Springer series in statistics. Springer, Heidelberg. doi:10.1007/978-3-642-20192-9

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, *35*(6), 2313–2351. doi:10.1214/009053606000001523

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective, second edition*. Boca Raton, FL, USA: Chapman & Hall/CRC.

Eddelbuettel, D., & Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, *71*, 1054–1063. doi:10.1016/j.csda.2013.02.005

James, G. M., & Radchenko, P. (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika*, *96*(2), 323–337. doi:10.1093/biomet/asp013

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., & Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, *14*. doi:10.1038/nrc3721

Loh, P.-L., & Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, *40*(3), 1637–1664. doi:10.1214/12-AOS1018

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rosenbaum, M., & Tsybakov, A. B. (2010). Sparse recovery under matrix uncertainty. *The Annals of Statistics*, *38*(5), 2620–2651. doi:10.1214/10-AOS793

Sørensen, Ø., Frigessi, A., & Thoresen, M. (2015). Measurement error in lasso: Impact and likelihood bias correction. *Statistica Sinica*, *25*(2), 809–829. doi:10.5705/ss.2013.180

Sørensen, Ø., Hellton, K. H., Frigessi, A., & Thoresen, M. (2018). Covariate selection in high-dimensional generalized linear models with measurement error. *Journal of Computational and Graphical Statistics*, *27*(4), 739–749. doi:10.1080/10618600.2018.1425626

Theussl, S., & Hornik, K. (2019). *Rglpk: R/GNU linear programming kit interface*. Retrieved from https://CRAN.R-project.org/package=Rglpk

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Vasquez, M. M., Hu, C., Roe, D. J., Halonen, M., & Guerra, S. (2019). Measurement error correction in the least absolute shrinkage and selection operator model when validation data are available. *Statistical Methods in Medical Research*, *28*(3), 670–680. doi:10.1177/0962280217734241