hdme: High-Dimensional Regression with Measurement Error

Oystein Sorensen[1]

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

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

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 satisifaction measures (Baesens 2014).

The lasso (Tibshirani 1996) and the Dantzig selector (Candes and Tao 2007; James and Radchenko 2009) are popular methods for variable selection in this type of problems, combining computational speed with good statistical properties (Bühlmann and Geer 2011). In many practical applications, the process of measuring the variables of interest is subject to measurement error (Carroll et al. 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, and Thoresen 2015), typically involving an increased number of false positive selections. A corrected lasso has been proposed and analyzed by Loh and Wainwright (2012) for linear models and Sørensen, Frigessi, and Thoresen (2015) for generalized linear models. It has been applied by Vasquez et al. (2019) in a problem involving measurement of serum biomarkers. For the Dantzig selector, Rosenbaum and Tsybakov (2010) proposed the Matrix Uncertainty Selector (MUS) for linear models, which was extended to the generalized linear model case by Sørensen et al. (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 `fit_gmus()` and `fit_corrected_lasso()`. Additional functions provide opportunities for hyperparameter tuning using cross-validation or the elbow rule (Rosenbaum and Tsybakov 2010), and plotting tools for visualizing the model fit. The underlying numerical procedures are implemented in `C++` using the `RcppArmadillo` package (Eddelbuettel and Sanderson 2014) and linear programming with `Rglpk` (Theussl and 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, Bart. 2014. *Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications*. 1st ed. Wiley Publishing.

Bühlmann, Peter, and Sara van de Geer. 2011. *Statistics for High-Dimensional Data*. Springer Series in Statistics. Springer, Heidelberg. https://doi.org/10.1007/978-3-642-20192-9.

Candes, Emmanuel, and Terence Tao. 2007. "The Dantzig Selector: Statistical Estimation When P Is Much Larger Than N." *Ann. Statist.* 35 (6). The Institute of Mathematical Statistics:2313–51. https://doi.org/10.1214/009053606000001523.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Boca Raton, FL, USA: Chapman & Hall/CRC.

Eddelbuettel, Dirk, and Conrad Sanderson. 2014. "RcppArmadillo: Accelerating R with High-Performance C++ Linear Algebra." *Computational Statistics & Data Analysis* 71:1054–63. https://doi.org/10.1016/j.csda.2013.02.005.

James, Gareth M., and Peter Radchenko. 2009. "A generalized Dantzig selector with shrinkage tuning." *Biometrika* 96 (2):323–37. https://doi.org/10.1093/biomet/asp013.

Kristensen, Vessela N., Ole Christian Lingjærde, Hege G. Russnes, Hans Kristian M. Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. 2014. "Principles and Methods of Integrative Genomic Analyses in Cancer." *Nature Reviews Cancer* 14. https://doi.org/10.1038/nrc3721.

Loh, Po-Ling, and Martin J. Wainwright. 2012. "High-Dimensional Regression with Noisy and Missing Data: Provable Guarantees with Nonconvexity." *Ann. Statist.* 40 (3). The Institute of Mathematical Statistics:1637–64. https://doi.org/10.1214/12-AOS1018.

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

Rosenbaum, Mathieu, and Alexandre B. Tsybakov. 2010. "Sparse Recovery Under Matrix Uncertainty." *Ann. Statist.* 38 (5). The Institute of Mathematical Statistics:2620–51. https://doi.org/10.1214/10-AOS793.

Sørensen, Arnoldo Frigessi, and Magne Thoresen. 2015. "Measurement Error in Lasso: Impact and Likelihood Bias Correction." *Statistica Sinica* 25 (2). Institute of Statistical Science, Academia Sinica:809–29. https://doi.org/10.5705/ss.2013.180.

Sørensen, Kristoffer Herland Hellton, Arnoldo Frigessi, and Magne Thoresen. 2018. "Covariate Selection in High-Dimensional Generalized Linear Models with Measurement Error." *Journal of Computational and Graphical Statistics* 27 (4). Taylor & Francis:739–49. https://doi.org/10.1080/10618600.2018.1425626.

Theussl, Stefan, and Kurt Hornik. 2019. *Rglpk: R/GNU Linear Programming Kit Interface*. https://CRAN.R-project.org/package=Rglpk.

Tibshirani, Robert. 1996. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1):267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Vasquez, Monica M, Chengcheng Hu, Denise J Roe, Marilyn Halonen, and Stefano Guerra. 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):670–80. https://doi.org/10.1177/0962280217734241.