CODERS' CORNER

MIND & BEHAVIOUR RESEARCH GROUP

CSAE CENTRE FOR THE STUDY OF AFRICAN ECONOMIES

UNIVERSITY OF OXFORD

PENALISED LEAST SQUARES FOR HIGH-DIMENSIONAL PREDICTION AND CAUSAL INFERENCE IN R

Among economists, machine learning methods have become a popular tool for performing high-dimensional causal inference. In a high-dimensional setting, namely where the number of parameters (p) that we want to estimate is not much larger than the number of observations (n), the OLS procedure *overfits* the sample, leading to lower prediction accuracy out of sample.

Penalized least squares methods offer an intuitive solution to the problem of overfitting by introducing a constraint, known as a penalty, in the standard OLS problem. For example, for LASSO, the penalty is the l_1 norm, for Ridge, it is the l_2 norm, and for Elastic Net, it is a weighted average of the two. These penalties are given a weighting that determines how large the penalization is, and it is usually chosen by cross-validation or by ``plugging-in" a theoretically optimal value.

Here, we introduce code for two popular use cases for penalized least squares. We first implement LASSO, Ridge and Elastic Net for prediction in a high-dimensional setting. We then use LASSO to perform "double-selection"[1], which allows for model selection when we are interested in estimating causal effects.

This Kaggle Notebook implements two examples using the packages *hdm*[2][3][4]and *glmnet*[5], which are two popular R packages that offer functions for implementing penalized least squares methods. *hdm* focuses on implementing theoretically valid choices of penalization parameters while *glmnet* offers good options for implementing cross-validated choices.

The Notebook aims to illustrate:

1. The use of penalized least squares methods, namely LASSO, Ridge and Elastic Net, to predict wages from (a large number of transformations of) demographic characteristics using a subsample of never-married workers from the Current Population Survey (CPS) data from 2015.
2. The use of doubly-robust LASSO to perform model selection and test the effect of gender and interaction effects of other variables with gender on wage (heterogenous treatment effects), jointly.[6] The dataset is inbuilt into \textit{hdm}, and is from CPS 2012.

Kaggle is a platform used by data scientists and machine learning practitioners to co-create and store data sets and code snippets. It provides users an intuitive interface to manage R Markdown and Jupyter (Python-based) notebooks, which allow creators to provide "code chunks", which generate outputs within the codebook, alongside substantive commentary, offering a means of making code more accessible to less experienced users.

[1] A. Belloni, V. Chernozhukov, C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies 81(2), 608-650.
[2] A. Belloni, D. Chen, V. Chernozhukov and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80 (6), 2369-2429.
[3] A. Belloni, V. Chernozhukov and C. Hansen (2013). Inference for high-dimensional sparse econometric models. In Advances in Economics and Econometrics: 10th World Congress, Vol. 3: Econometrics, Cambirdge University Press: Cambridge, 245-295.
[4] A. Belloni, V. Chernozhukov, C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies 81(2), 608-650
[5] With thanks to the authors and maintainer, Trevor Hastie.
[6] With reference to the model of C. B. Mulligan and Y. Rubinstein (2008). Selection, investment, and women's relative wages over time. The Quarterly Journal of Economics, 1061–1110.

Users can then copy and edit these notebooks, and the associated datasets, allowing them to perform robustness checks of interest or to see if they can improve on the methodology. For less experienced users, it offers the opportunity to “learn by doing” – to alter the parameterization of functions and models to better understand exactly what the code is doing.

Further information about the hdm package can be found in the R documentation, and in this vignette by the package authors, which includes a robust discussion of the theory of penalized least squares. Further information about glmnet can be found in the R documentation and an elegant introduction can be found in this vignette.

Much of the code shared here was first drafted by Victor Chernozhukov, Christian Hansen, Martin Spindler and Jannis Kueck and is used by Victor Chernozhukov as part of his class on machine learning at MIT. I provide links to the original public code sources (those in the \textit{hdm} package vignette and that maintained on Kaggle, which I help maintain) in my Kaggle notebook and commend their commitment to making ML methods more accessible to the average user.

John Walker, Research Assistant in Development Economics in the Mind and Behaviour Research Group, CSAE
02 December 2021