Homework Challenge (4 Extra Points)

Given a linear model

$$y = \beta' x + e \tag{1}$$

, where $x = (1, x_1, \ldots, x_p)$, let β^* be the population least squares coefficients, i.e.

$$\beta^* = \arg\min_{\beta} E\left[(y - \beta' x)^2\right]$$

In high-dimensional settings ($p > N$), (1) should be estimated using regularized regression methods such as the lasso. If, in addition, we believe that β^* contains mostly zeros – if $\widetilde{p}$, the number of non-zero population coefficients, $\ll p$ – then we say there is **sparsity**. In this case, we *may* be able to improve our results by a running multi-stage procedure:

Algorithm. *Relaxed Lasso (including Post-Lasso OLS)*

Stage 1 *In the first stage, estimate (1) by the lasso*

Stage 2 *Run lasso again (or OLS) on the variables selected by the first stage. Note that if optimal shrinkage is close to* 0*, then this stage becomes an OLS.*

The idea behind the two stage procedure is that in high-dimensional sparse settings, we can have a separate variable selection – estimation procedure, in which the first stage is mainly used for variable selection, while the second stage is used for estimation.

The idea originates from Efron et al. (2004), in which the authors propose running LARS[1] in the first stage and OLS in the second stage, a procedure that they call the **LARS-OLS hybrid**. Meinshausen (2007) proposes the two-stage procedure outlined above, which the author calls **relaxed Lasso**[2,3]. Belloni and Chernozhukov (2013) pro-

[1] Least Angle Regression, which incorporates the lasso.

[2] Meinshausen (2007) specifically discusses using OLS in the second stage as a special case of relaxed lasso.

[3] Elements of Statistical Learning (P91):

pose the same two-stage procedure as Meinshausen (2007) but with OLS (rather than a lasso that incorporates the OLS) as the second stage, which they call the **post-Lasso OLS**[4]. Finally, Zhao et al. (2017) propose exactly the same post-Lasso OLS procedure, which they call "a very naive approach"[5].

In the applied economics literature, the post-Lasso OLS has gained some popularity recently. In part, this is because the second stage OLS provides statistical inference for the coefficients of interest. However, such inference is generally invalid, because the second stage model is *selected* by the first stage. Significant efforts have been made to provide theoretically sound asymptotic inference for lasso-type estimators, which has been called *post-selection inference*. See Belloni et al. (2015), Lee et al. (2016), and Taylor and Tibshirani (2017).

Challenge

Use simulation to compare the performance of the lasso (alone) and the relaxed lasso or post-Lasso OLS procedure in different settings.

- To do this, you need to: (a) design a "true" model from which you are going to simulate your data; (b) generate a *really large* test data set (say, $N = 1e7$); (c) generate R (e.g., R = 1000) training data sets; (d) train your methods on *each* training set and evaluate them on the test set; (e) compare the performance of your methods by averaging their test error over *all* R iterations, and comparing the distribution of statistics such as $\widehat{\beta}$.

> ...the lasso shrinkage causes the estimates of the non-zero coefficients to be biased towards zero and in general they are not consistent [Added Note: This means that, as the sample size grows, the coefficient estimates do not converge]. One approach for reducing this bias is to run the lasso to identify the set of non-zero coefficients, and then fit an un-restricted linear model to the selected set of features. This is not always feasible, if the selected set is large. Alternatively, one can use the lasso to select the set of non-zero predictors, and then apply the lasso again, but using only the selected predictors from the first step. This is known as the relaxed lasso (Meinshausen, 2007). The idea is to use cross-validation to estimate the initial penalty parameter for the lasso, and then again for a second penalty parameter applied to the selected set of predictors. Since the variables in the second step have less "competition" from noise variables, cross-validation will tend to pick a smaller value for [the penalty parameter], and hence their coefficients will be shrunken less than those in the initial estimate.

[4] Curiously, Belloni and Chernozhukov (2013) does not cite either Efron et al. (2004) or Meinshausen (2007).

[5] Amazingly, the authors referenced Belloni and Chernozhukov (2013), although not in their literature review, but in one of their proofs.

- For different settings, try changing the size and the sparsity of your β^*. Under what settings does either method work better? Explain your findings.

- To implement relaxed lasso, try `R` packages relaxo or relaxnet. To implement post-Lasso OLS, you can still use the relaxed lasso packages, or use the `R` package hdm, which stands for "high-dimensional metrics". Read this tutorial for an overview of the methods implemented in hdm. For post-selection inference, use the `R` package selectiveInference, which implements Lee et al. (2016) and Taylor and Tibshirani (2017), or hdm, which implements Belloni et al. (2015).

References

[1] Belloni, A. and V. Chernozhukov. 2013. "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19(2). [link]

[2] Belloni A., V. Chernozhukov, and K. Kato. 2015. "Uniform Post Selection Inference for Least Absolute Deviation Regression and Other Z-estimation Problems," *Biometrika*, 102(1). [link]

[3] Efron B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. "Least Angle Regression," *The Annals of Statistics*, 32(2). [link]

[4] Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor. 2016. "Exact Post-selection Inference with Application to the Lasso," *The Annals of Statistics*, 44(3). [link]

[5] Meinshausen, N. 2007. "Relaxed Lasso," *Computational Statistics & Data Analysis*, 52(1). [link]

[6] Taylor, J. and R. Tibshirani. 2017. "Post-selection inference for L1-penalized likelihood models," *The Canadian Journal of Statistics*, 46(1). [link]

[7] Zhao S., Shojaie A., and D. Witten. 2017. "In Defense of the Indefensible: A Very Naive Approach to High-Dimensional Inference," arXiv:1705.05543. [link]