High-Dimensional Statistics and Causal Inference Homework Challenge (2 Extra Points)

High-Dimensional Statistical Modeling

Given a data set $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ and a linear model

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

, where $x = (x_1, \ldots, x_p)$[1]. When p is large relative to N, we can estimate (1) using a shrinakge method such as ridge regression (ℓ_2−regularization) to improve its finite sample performance. If, in addition, we believe that the population β^* [2]contains mostly zeros, i.e., the number of truly non-zero coefficients is small, then we say there is **sparsity**. In this case, we can estimate (1) with the lasso (ℓ_1−regularization) in order to achieve variable selection.

But is the lasso guaranteed to select the *correct* set of non-zero coefficient variables when $N \to \infty$? This property is called the **oracle property**[3]. It turns out that the original lasso does not have the oracle property, but with appropriate modifications it can achieve the oracle. See Zou (2006).

In many cases, we can also further improve the statistical performance of the lasso using a two-stage procedure. Recall that shrinkage methods tradeoff between bias and variance. The lasso coefficient estimates are therefore biased as a price to pay for variable selection[4]. But we can reduce this bias by running OLS on the model selected by the lasso. This is called **post-lasso OLS**. See Belloni and Chernozhukov (2013).

Note, however, post-lasso OLS does not provide the correct statistical inference for coefficient estimates. This is because the OLS standard errors do not take into account model selection. Significant efforts have been made in recent years to provide theoretically sound asymptotic

[1] If (x, y) are standardized, then no intercept is needed.

[2] β^* is the *population* least squares coefficients, i.e.,

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

[3] More precisely, the oracle property requires (1) identification of the correct subset model; (2) optimal estimation rate.

[4] which reduces variance.

inference for lasso-type estimators, which has been called *post-selection inference*. See Taylor and Tibshirani (2015) for an introduction to post-selection inference. See Lee et al. (2016) and Taylor and Tibshirani (2017) for results related to the lasso. The `R` package selectiveInference implements Lee et al. (2016) and Taylor and Tibshirani (2017)[5].

Reference

- Fan, J. and R. Li. (2001). "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96(456). [paper]

- Zou, H. (2006). "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101(476). [paper]

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

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

- Taylor, J. and R. Tibshirani. (2015). "Statistical learning and selective inference," *Proceedings of the National Academy of Sciences*, 112(25). [paper]

- 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). [paper]

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

[5] Note that another way to conduct correct inference for post-selection models is to simply evaluate their performance on a test data. Doing so, however, requires one to have access to a separate test data set or split existing data into training and test sets, which results in sample loss.

High-Dimensional Causal Inference

Given a data set $\mathcal{D} = \{x_i, w_i, y_i\}_{i=1}^N$, where y is the outcome variable, w is the treatment variable, and x is the control variable. Consider a linear model

$$y = \alpha \cdot w + \beta \cdot x + e \tag{2}$$

Assume that x satisfy the *back-door criterion* such that conditional on x there is no unmeasured confounding. Then α represents the *treatment effect* of w on y.

If w is low-dimenstional and x is high-dimensional, we have a causal inference problem with *high-dimensional controls.* If we believe that the number of truly relevant control variables is small, i.e. β is *sparse*, then we can estimate (2) with the lasso by applying shrinkage *only* to β and conduct post-selection inference for α and β. However, in this case we are mainly interested in the estimation and inference of α − the treatment effect of w on y given x, while β can be treated as a **nuisance** parameter.

The Partialling Out Approach

Algorithm. *Partialling Out*

Stage 1 *Estimate the following two models by the lasso:*

$$\begin{aligned} y &= \gamma \cdot x + u \\ w &= \lambda \cdot x + v \end{aligned}$$

, from which we obtain $\widehat{u}$ *and* $\widehat{v}$.

Stage 2 *Run the following residual-on-residual regression by OLS:*

$$\widehat{u} = \alpha \cdot \widehat{v} + e$$

, from which we obtain $\widehat{\alpha}$ *and its asymptotic standard error.*

The approach has been called **partialling out** and utilizes the principal of **Neyman-orthogonality**. See Chernozhukov et al. (2017)[6]. More generally, we can use this approach in situations in which we have a linear model with *high-dimensional* regressors but are interested in obtaining valid inference only on a *low-dimensional* subset of the model parameters (e.g., the parameters associated with the treatment variables).

[6] The method introduced in Chernozhukov et al. (2017) applies to models with general functional forms (e.g., machine learning models) not limited to linear models and the lasso.

What about the selection of control variables? According to the *disjunctive cause criterion*, if there exists a set of observed variables that satisfies the back-door criterion, then we can make sure we select them by selecting all observed causes of treatment w and of outcome y. In high-dimensional sparse settings, where there are a large number of *potential* causes of w and y, but the number of *real* causes are small, Belloni et al. (2014b) propose the double-selection method.

Double-Selection of Control Variables

Algorithm. *Double Selection*

Estimate the following two models by the lasso:

$$\begin{aligned} y &= \gamma \cdot x + u \\ w &= \lambda \cdot x + v \end{aligned}$$

, where x is the set of all potential causes of x and y. Then the final selected model is

$$y = \alpha \cdot w + \beta \cdot \widetilde{x} + e$$

, where $\widetilde{x}$ is the union of the variables selected by the two lasso regressions.

In practice, one can use the partialling out approach to obtain an estimate and standard error for α and use double-selection approach to obtain the set of final control variables $\widetilde{x}$. Then estimate the following model by OLS to obtain $\widehat{\beta}$[7]:

$$y - \widehat{\alpha} \cdot w = \beta \cdot \widetilde{x} + e$$

The `R` package hdm (github repo) implements the partialling out and double-selection methods. Read this tutorial for an overview. Also see Belloni et al. (2014a) for an introduction to causal inference in high-dimensional settings.

[7] Note that this will not give us the correct standard error for $\widehat{\beta}$.

Extensions

- Instrumental Variables: Belloni et al. (2012)
- Panel Data: Belloni et al. (2016)

Reference

- Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80(6). [link]
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, 28(2). [link]
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 81(2). [link]
- Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). "Inference in high-dimensional panel models with an application to gun control," *Journal of Business & Economic Statistics*, 34(4). [link]
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). "Double/debiased/neyman machine learning of treatment effects," *American Economic Review*, 107(5). [link]