Homework Challenge (2 Extra Points)

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 x and of outcome y.

In high-dimensional sparse settings, where there are a large number of *potential* causes of x and y (relative to the number of data points), but the number of *real* causes are small, Belloni et al. (2014) propose the **post-double-selection** method:

Algorithm. *Post Double Selection*

Stage 1 *In the first stage, estimate the following two models by the lasso:*

$$y = \alpha' v + e$$
$$x = \lambda' w + \epsilon$$

, where v is the set of potential causes of x and w is the set of potential causes of y.

Stage 2 *Estimate the following model by OLS:*

$$y = \beta x + \gamma' z + \varepsilon$$

, where z is the union of the v and w selected by the first stage lasso.

Alternatively, instead of doing the post-double-selection procedure, one can just run the lasso on the union of all potential causes, i.e. estimate the following model by the lasso:

$$y = \beta x + \gamma' s + \xi$$

, where $s = v \cup w$.

Challenge

Use simulation to compare the performance of the post-double-selection procedure and running the lasso on all potential causes.

- 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 $\widehat{\beta}$ to β^*.

- To implement post-double-selection, use the `R` package hdm, which stands for "high-dimensional metrics". Read this tutorial for an overview of the methods implemented in the package.

References

[1] Belloni, A., V. Chernozhukov, and C. Hansen. 2014. "Inference on Treatment Effects after Selection amongst High-dimensional Controls," *The Review of Economic Studies*, 81(2). [link]