

## 14.387 PS NUMBER 3. DUE: LAST DAY OF THE FALL TERM.

### 1. LASSO AND POST-LASSO IN A SIMPLE MODEL

Consider a classical linear regression model

$$y_i = x_i' \beta_0 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad \text{where the noise } \epsilon_i \text{ are } i.i.d., i = 1, \dots, n,$$

$y_i$  is the response variables, and  $x_i$  are  $p$ -dimensional covariates, and  $\beta_0$  is the true value of the parameter vector  $\beta = (\beta_1, \dots, \beta_p)'$ . For convenience we normalize the covariates so that  $\mathbb{E}_n[x_{ij}^2] = 1$ , for  $j = 1, \dots, p$ , where  $\mathbb{E}_n$  stands for  $n^{-1} \sum_{i=1}^n$ . Let  $X = [x_1 \cdots x_n]'$ . The orthonormal design case occurs when  $(X'X)_{jk} = 0$  for all  $j \neq k$ ; in such a case we have that  $X'X/n = I$ , where  $I$  is a  $p \times p$  identity matrix.

Consider the LASSO estimator for  $\lambda > 0$ .

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(y_i - x_i' \beta)^2] + \frac{\lambda}{n} \|\beta\|_1,$$

and the post LASSO estimator

$$\tilde{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(y_i - x_i' \beta)^2] : \beta_j = 0 \quad \text{if } \hat{\beta}_j = 0.$$

Assuming the orthonormal design case, solve one the following questions:

- (a) Derive the explicit solution for the LASSO and Post-LASSO estimators.
- (b) Suppose  $\beta_0 = 0$ , characterize the minimal value of  $\lambda$  that correctly forces  $\hat{\beta} = 0$  in this case.

This question is going to be covered in the TA session, which provides an incentive to attend this.

### 2. POST-DOUBLE-SELECTION INFERENCE

Please do one of the following:

1. Replicate the impact of abortion on crime example from "High-Dimensional Methods and Inference on Treatment and Structural Effects in Economics," 2014, J. Economic Perspectives, by A. Belloni et.al. Stata and Matlab and data file are available from C. Hansen's webpage at Booth <http://faculty.chicagobooth.edu/christian.hansen/research/>, and Stata replication code is available via JEP at:  
[http://www.aeaweb.org/jep/data/2802/2802-0029\\_data.zip](http://www.aeaweb.org/jep/data/2802/2802-0029_data.zip)  
Also an R-code is available by request by e-mail . Write up your results as 1-2 paragraph summary of what the double-selection method is doing and how it is relevant in the empirical example.
2. Replicate the growth example from "High-Dimensional Sparse Econometrics," 2013, NBER lectures, by V. Chernozhukov and C. Hansen, [www.mit.edu/~vchern](http://www.mit.edu/~vchern). The data and Matlab and Stata code are available here:  
<http://www.mit.edu/~vchern/NBER/>.  
Also an R-code is available by request. Write up your results as 1-2 paragraph summary of what the double-selection method is doing and how it is relevant in the empirical example.
3. Try the post-double-selection method on any other interesting data-set. Here if you come up with an interesting example, there is going to be a prize in the form of the book by Aad Van der Vaart "Asymptotic Statistics", which is a great book. You can download Matlab and Stata commands from:  
<http://faculty.chicagobooth.edu/christian.hansen/research/>. Also an R-code is available by request. Write up your results as 1-2 paragraph summary of what the double-selection method is doing and how it is relevant in the empirical example.

MIT OpenCourseWare  
<http://ocw.mit.edu>

14.387 Applied Econometrics: Mostly Harmless Big Data  
Fall 2014

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.