36-710, Fall 2018
Homework 3

Due Friday Nov 2 by 5:00pm in JaeHyeok's mailbox

1. **Inference after model selection**. Suppose that we observe n independent random variables $(X_1, \ldots, X_n)$ where $X_i \sim N(\mu_i, 1)$ for all i. The means $\mu_1, \ldots, \mu_n$ are unknown but we suspect that most of them are zero and some are large in absolute value. We first perform a naive model selection procedure by computing the random set of indexes

$$\hat{I} = \{i \colon |X_i| > 1\},$$

corresponding to the variables that presumably have the largest means in absolute value. This is the model selection part. Then, for any one $i \in \hat{I}$ (assumed non-empty), we test the null hypothesis that $\mu_i = 0$ at the significance level of $\alpha = 0.05$. This the inference part. We decide to ignore the selection step, and use the test that rejects if $|X_i| > z_{\alpha/2}$, the $1 - \alpha/2$ quantile of a standard normal. What is the problem with this choice? What would you suggest to do in order to correctly take into account the selection step?

2. In earlier works on the lasso, people have used a even stronger assumptions than the restricted eigenvalue property. Here is one. Suppose that the design matrix X is such that, for some integer $k > 0$,

$$\max_{i,j} \left| \frac{X_i^\top X_j}{n} - 1(i = j) \right| \leq \frac{1}{23k} \tag{1}$$

where X_i is the ith column of X, $i = 1, \ldots, d$. Think about what that means.

   (a) Show that this condition implies that, for any subset S of $\{1, \ldots, d\}$ of cardinality no larger than $k < d$ and any $\Delta \in \mathbb{R}^d$ with $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$,

$$\|\Delta\|^2 \leq \frac{2}{n}\|X\Delta\|^2.$$

   That is, show that this condition implies the $RE(3, 1/2)$ condition given in class for all non-empty subsets S of $\{1, \ldots, d\}$ of size no larger than k. *Instead of the constant* 23 *you may take a larger one if it simplifies your calculations.*

   (b) Suppose that the entries of X are now populated by independent Rademacher variables (a Rademacher variable is one that that takes the values $+1$ and -1 with equal probability). Show that, for any $\delta \in (0, 1)$, if

$$n \geq Ck^2(\log(d) + \log(1/\delta)),$$

   for some constant $C > 0$, then X satisfies the condition (1), with probability at least $1 - \delta$. *Again, instead of* 23 *feel free to show the result for a different constant if it helps with the calculations.*

3. Exercise 7.13

4. Read tghe paper "Assumptionless consistency of the lasso", by S. Chatterjee. The paper is available at https://arxiv.org/pdf/1303.5817.pdf. Reproduce the proofs of Theorem 1 and 2. Theorem 1 in particular shows that the lasso is a good method for precision.

5. Exercise 7.17

6. **The Lasso and Fals Discoveries.** Read Sections 1-4 of the paper "False Discoveries occur Early on the Lasso Path", by Weijie Su, Malgorzata Bogda and Emmanuel J. Candés, available at https://statweb.stanford.edu/~candes/papers/LassoFDR.pdf. You are not expected to read the proofs, which are based on advanced techniques not covered in the course. Write a paragraph to summarize their findings.

7. **A sparse oracle inequality for the lasso**. Consider the following set-up, as described in class. We observe n pairs $(Y_1, x_1), \ldots, (Y_n, x_n)$, where each x_i is a fixed vector in $\mathbb{R}^d$ and

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

with $\epsilon_1, \ldots, \epsilon_n$ independent variables in $SG(\sigma^2)$. We have a dictionary $(f_1, \ldots, f_M)$ of M functions from $\mathbb{R}^d$ into $\mathbb{R}$ and would like to estimate f using a **sparse** linear combination of such functions. More precisely, for $\theta = (\theta_1, \ldots, \theta_M) \in \mathbb{R}^M$, let $f_\theta = \sum_{j=1}^M \theta_j f_j$. Then, we will estimate f with $f_{\hat{\theta}}$ where $\hat{\theta}$ is a lasso solution:

$$\hat{\theta} \in \text{argmin}_{\theta \in \mathbb{R}^M} \frac{1}{2n} \sum_{i=1}^n (Y_i - f_\theta(x_i))^2 + \lambda_n \|\theta\|_1,$$

for some $\lambda_n \geq 0$. To study the performance of this estimator, we will compared $f_{\hat{\theta}}$ to the **sparse oracle approximation** f_{θ^*}, where

$$\theta^* \in \text{argmin}_{\theta \in \mathbb{R}^M, \|\theta\|_0 \leq k} \frac{1}{2n} \sum_{i=1}^n (f(x_i) - f_\theta(x_i))^2, \tag{2}$$

and $0 < k < M$ is a fixed constant.

We will make the following assumption: let Φ be the $n \times M$ matrix with entries, $\Phi_{i,j} = f_j(x_i)$, and assume that, for some $\kappa > 0$, Φ satisfies the $RE(3, \kappa)$ condition with respect to all non-empty subsets S of $\{1, \ldots, M\}$ of cardinality no larger than k. Show that, if $\lambda_n \geq \frac{2}{n} \|\Phi^\top \epsilon\|_\infty$, where $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$, then, for any $\alpha \in (0, 1)$,

$$MSE(f_{\hat{\theta}}) \leq \inf_{\theta \in \mathbb{R}^d, \|\theta\|_0 \leq k} \left\{ \frac{1+\alpha}{1-\alpha} MSE(f_\theta) + \frac{9}{2\alpha(1-\alpha)\kappa} \|\theta\|_0 \lambda_n^2 \right\},$$

where, for $\theta \in \mathbb{R}^M$, $MSE(f_\theta) = \frac{1}{n} \sum_{i=1}^n (f_\theta(x_i) - f(x_i))^2$.

The first term on the right hand side contains the approximation error, while the term on the left the estimation error. Notice that the leading constant is not 1.

The proof is similar to the proof of the fast rate for the lasso.

Proceed in this way:

- *Consider any $\theta \in \mathbb{R}^M$ with $\|\theta\|_0 \leq k$. Derive the basic inequality*

$$\frac{1}{n} \left(\sum_{i=1}^n (f(x_i) - f_{\hat{\theta}}(x_i))^2 - \sum_{i=1}^n (f(x_i) - f_\theta(x_i))^2 \right) \leq 2\lambda_n (\|\theta\|_1 - \|\hat{\theta}\|_1) + 2 \frac{1}{n} \sum_{i=1}^n \epsilon_i (f_{\hat{\theta}}(x_i) - f_\theta(x_i))$$

- *If the term within parenthesis on the left hand side is negative, then the bound holds trivially. If not, then continue exactly as in the proof of the fast rates for the lasso (you will need the RE condition). You will then see that the right hand side of the last display is bounded by*

$$3\lambda_n \|\theta\|_0 \sqrt{\frac{\sum_{i=1}^n (f_{\hat{\theta}}(x_i) - f_\theta(x_i))^2}{\kappa}}.$$

- *To further bound the last expression, use the variational inequality, valid for all $x, y \in \mathbb{R}$,*

$$2xy = \inf_{\gamma > 0} \left(\frac{x^2}{\gamma} + y^2 \gamma \right),$$

  which implies that $2xy \leq \frac{2}{\alpha} x^2 + \frac{\alpha}{2} y^2$, for any $\alpha \in (0, 1)$.
- *Finally, break up the squares $\sum_{i=1}^n (f_{\hat{\theta}}(x_i) - f_\theta(x_i))^2$ using the inequality $(x - y)^2 \leq 2x^2 + 2y^2$.*
- *The final bound holds for any $\theta \in \mathbb{R}^M$ with $\|\theta\|_0 \leq k$, so you may take the infimum over all such vectors.*