STAT 620: Asymptotic Statistics **Spring 2022**

Lecture: Apr 28

Lecturer: Xianyang Zhang

Lasso (least absolute shrinkage and selection operator)

Consider i.i.d. samples (x_i, y_i), $i = 1, 2, \dots, n$ from the linear model

$$y_i = x_i^\top \beta_0 + \epsilon_i,$$

where $\beta_0 \in \mathbb{R}^p$ is an unknown coefficient vector, and $\{\epsilon_i\}_{i=1}^n$ are random errors with mean zero. We can more succinctly express this data model as

$$Y = X\beta_0 + \epsilon,$$

where $Y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ is the vector of responses, X is the matrix of predictor variables, with ith row $x_i^\top$, and $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$ is the vector of errors.

Regularization

Regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting. When $p \gg n$, least squares estimation is ill-posed and regularization is needed. Let's consider three canonical choices: the l_0, l_1, and l_2 norms:

$$\|\beta\|_0 = \sum_{j=1}^p \mathbf{1}\{\beta_j \neq 0\},$$
$$\|\beta\|_1 = \sum_{j=1}^p |\beta_j|,$$
$$\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2.$$

In constrained form, these norms give rise to the following problems:

$$\text{Best subset selection: } \min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 \text{ subject to } \|\beta\|_0 = \sum_{j=1}^p \mathbf{1}\{\beta_j \neq 0\} \le t,$$
$$\text{Lasso: } \min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 \text{ subject to } \|\beta\|_1 = \sum_{j=1}^p |\beta_j| \le t,$$
$$\text{Ridge regression: } \min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 \text{ subject to } \|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2 \le t.$$

In penalized form, Lasso is defined as

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\|Y - X\beta\|^2 + \lambda\|\beta\|_1.$$

Consistency of Lasso

Consider the least squares estimator in the linear model

$$\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top Y.$$

The prediction error

$$\frac{\|X(\hat{\beta}_{\text{OLS}} - \beta)\|_2^2}{n} = \frac{\epsilon^\top H \epsilon}{n}.$$

where $H = X(X^\top X)^{-1} X^\top$. When $\epsilon \sim N(0, \sigma^2 I_p)$, we have $\|X(\hat{\beta}_{\text{OLS}} - \beta)\|_2^2/\sigma^2 = \epsilon^\top H \epsilon/\sigma^2 \sim \chi_p^2$ and hence

$$E\left[\frac{\|X(\hat{\beta}_{\text{OLS}} - \beta)\|_2^2}{n}\right] = \frac{p\sigma^2}{n}.$$

Define the Lasso estimator

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n}\|Y - X\beta\|^2 + \lambda\|\beta\|_1.$$

Our goal is to show that with a proper choice for λ, one has the "oracle inequality":

$$\frac{\|X(\hat{\beta} - \beta)\|_2^2}{n} \leq C\log(p)\frac{s_0\sigma^2}{n}$$

with large probability, where s_0 is the number of nonzero components in β_0. The term $C\log(p)$ is the price we pay for not knowing the support of β_0.

Basic inequality: Note that

$$\frac{1}{n}\|Y - X\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\|Y - X\beta_0\|^2 + \lambda\|\beta_0\|_1.$$

Rearranging the terms, we have the basic inequality

$$\frac{\|X(\hat{\beta} - \beta_0)\|^2}{n} + \lambda\|\hat{\beta}\|_1 \leq \frac{2\epsilon^\top X(\hat{\beta} - \beta_0)}{n} + \lambda\|\beta_0\|_1.$$

Let $X^{(j)}$ be the jth column of X. Consider the event

$$\mathcal{T} = \{\max_{1\leq j\leq p} 2|\epsilon^\top X^{(j)}|/n \leq \lambda_0\}.$$

A useful lemma: We aim to show that

$$P(\mathcal{T}) \geq 1 - 2\exp(-t^2/2),$$

where $\lambda_0 = 2\sigma\sqrt{(t^2 + 2\log(p))/n}$. Suppose $\|X^{(j)}\|^2/n = 1$ for all $1 \leq j \leq p$. Further assume ϵ_i's are i.i.d σ^2-sub-Gaussian, and ϵ and X are independent. Then we have $\epsilon^\top X^{(j)}/\sqrt{n\sigma^2}$ is 1-sub-Gaussian. Thus

$$P(|\epsilon^\top X^{(j)}/\sqrt{n\sigma^2}| \geq u) \leq 2\exp(-u^2/2).$$

Using the union bound, we have

$$P(\mathcal{T}^c) = P(\max_{1\leq j\leq p}|\epsilon^\top X^{(j)}/\sqrt{n\sigma^2}| > \sqrt{t^2 + 2\log(p)}) \leq 2p\exp\left(-\frac{t^2 + 2\log(p)}{2}\right) = 2\exp(-t^2/2).$$

Consistency of Lasso: Set

$$\lambda = 2\lambda_0 = 4\sigma\sqrt{\frac{t^2 + 2\log(p)}{n}}.$$

On the event $\mathcal{T}$,

$$|2\epsilon^\top X(\hat{\beta} - \beta_0)/n| \leq 2\|\hat{\beta} - \beta_0\|_1 \max_{1\leq j\leq p} |\epsilon^\top X^{(j)}|/n \leq \lambda_0\|\hat{\beta} - \beta_0\|_1 \leq \lambda_0\|\hat{\beta}\|_1 + \lambda_0\|\beta_0\|_1,$$

Using the basic inequality, we obtain

$$\frac{\|X(\hat{\beta} - \beta_0)\|^2}{n} + \lambda\|\hat{\beta}\|_1 \leq \lambda_0\|\hat{\beta}\|_1 + 3\lambda_0\|\beta_0\|_1.$$

Thus with probability greater than $1 - 2\exp(-t^2/2)$, it holds that

$$\frac{2\|X(\hat{\beta} - \beta_0)\|^2}{n} \leq 3\lambda\|\beta_0\|_1 = 12\sigma\|\beta_0\|_1\sqrt{\frac{t^2 + 2\log(p)}{n}}.$$

A refined result

By the basic inequality and on the event $\mathcal{T}$, we have

$$\frac{2\|X(\hat{\beta} - \beta_0)\|^2}{n} + 2\lambda\|\hat{\beta}\|_1 \leq \lambda\|\hat{\beta} - \beta_0\|_1 + 2\lambda\|\beta_0\|_1.$$

Next we note that

$$\begin{aligned}\|\hat{\beta}\|_1 =& \|\hat{\beta}_{S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1 \\ \geq& \|\beta_{0,S_0}\|_1 - \|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1.\end{aligned}$$

where $S_0 = \{1 \leq j \leq p : \beta_j \neq 0\}$. Also

$$\|\hat{\beta} - \beta_0\|_1 = \|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1 + \|\hat{\beta}_{S_0^c}\|_1.$$

Combining the inequalities, we get

$$\frac{2\|X(\hat{\beta} - \beta_0)\|^2}{n} + \lambda\|\hat{\beta}_{S_0^c}\|_1 \leq 3\lambda\|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1. \tag{1}$$

As a consequence, we have

$$\|\hat{\beta}_{S_0^c} - \beta_{0,S_0^c}\|_1 = \|\hat{\beta}_{S_0^c}\|_1 \leq 3\|\hat{\beta}_{S_0} - \beta_{0,S_0}\|_1.$$

Compatibility condition: Let $\Sigma = X^\top X/n \in \mathbb{R}^{p\times p}$. If for some $\phi_0 > 0$, and for all β satisfying $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$, it holds that

$$\|\beta_{S_0}\|_1^2 \leq s_0(\beta^\top\Sigma\beta)/\phi_0^2.$$

Main result: Under the compatibility condition, we have

$$\|X(\hat{\beta} - \beta)\|^2/n + \lambda\|\hat{\beta} - \beta\|_1 \leq 4\lambda^2 s_0/\phi_0^2. \tag{2}$$

As a result, we have

$$\begin{aligned}&\|X(\hat{\beta} - \beta)\|^2/n \leq 4\lambda^2 s_0/\phi_0^2, \\ &\|\hat{\beta} - \beta\|_1 \leq 4\lambda s_0/\phi_0^2.\end{aligned}$$

To show (2), we know that

$$
\begin{aligned}
&2\|X(\hat{\beta}-\beta)\|^2/n+\lambda\|\hat{\beta}-\beta\|_1\\
=&2\|X(\hat{\beta}-\beta)\|^2/n+\lambda\|\hat{\beta}_{S_0}-\beta_{S_0}\|_1+\lambda\|\hat{\beta}_{S_0^c}\|_1\\
\le&4\lambda\|\hat{\beta}_{S_0}-\beta_{S_0}\|_1\\
\le&4\lambda\sqrt{s_0}\|X(\beta-\beta_0)\|_2/(\sqrt{n}\phi_0)\\
\le&\|X(\beta-\beta_0)\|_2^2/n+4\lambda^2 s_0/\phi_0^2,
\end{aligned}
$$

where the first inequality follows from the basic inequality, the second inequality is due to the compatibility condition, and the last inequality is because of $2ab \le a^2+b^2$.