36-710: Advanced Statistical Theory **Fall 2018**

Lecture 16: March 21

Lecturer: Alessandro Rinaldo *Scribes: Jinjin Tian*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Last time, we talk about the prediction bound for ordinary least square ridge regression. Let's first recall the setting of the problem.

16.1 OLS regression in high dimension

Assume that Y follows the standard linear model, such that $Y = \mathbb{X}\beta^* + \epsilon$, where $\mathbb{X}$ is some $n \times d$ fixed design matrix, and ϵ is a n dimension vector of independent $SG(\sigma^2)$ random variable. Then, we have the ordinary least square estimator

$$\widehat{\beta}_{\text{OLS}} = (\mathbb{X}^T\mathbb{X})^+\mathbb{X}^T Y, \tag{16.1}$$

where $(A)^+$ denotes the pseudo inverse of matrix A, i.e. only take inverse of the positive eigenvalue of A. Then the mean squared error for $\widehat{\beta}_{\text{OLS}}$ is

$$\text{MSE}(\widehat{\beta}_{\text{OLS}}) = \frac{1}{n}||\mathbb{X}(\widehat{\beta} - \beta^*)||^2 \lesssim \sigma^2 \frac{(r + \log 1/\delta)}{n}. \tag{16.2}$$

where $r = \text{rank}(\mathbb{X}^T\mathbb{X})$.

From last lecture, we have that

$$||\beta^* - \widehat{\beta}_{\text{OLS}}||^2 \leq \frac{\text{MSE}(\widehat{\beta}_{\text{OLS}})}{\lambda_{\min}(\frac{\mathbb{X}^T\mathbb{X}}{n})},$$

where $\lambda_{\min}(A)$ is the minimal eigenvalue of matrix A. This implies estimating β^* is much harder than $\mathbb{X}\beta^*$.

16.2 Penalized regression

One possible way to tackle the high dimensional regression is adding penalization. Say instead of minimizing the ordinary square loss, we would like to find $\widehat{\beta}$ such that

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^d}\{\frac{1}{2n}||Y - \mathbb{X}\beta||^2 + \lambda_n f(\beta)\}, \tag{16.3}$$

where λ_n is the penalty parameter which is positive, and $f(\beta)$ represents the penalty for complexity or size of β.

One classical penalty is letting $f(\beta) = ||\beta||^2$, which will result in the ridge estimator $\widehat{\beta}_{\text{ridge}} = (\mathbb{X}^T\mathbb{X} + \lambda_n I)^{-1}\mathbb{X}^T Y$. As we mentioned in the last lecture,

$$\mathbb{X}\widehat{\beta}_{\text{ridge}} = \sum_{j=1}^{r} u_j \frac{\sigma_j^2}{\lambda + \sigma_j^2}\langle u_j, Y\rangle,$$

where $u_1, \ldots, u_r$ are the orthogonal basis for the column space of $\mathbb{X}$, and $\sigma_1, \ldots, \sigma_r$ are the singular value of $\mathbb{X}$, which makes σ_j the j-th eigenvalue of $\mathbb{X}^T\mathbb{X}$. Therefore, $\mathbb{X}\widehat{\beta}_{\text{ridge}}$ is actually the projection of Y onto the column space of X based on weights $\sigma_1, \ldots, \sigma_r$. We gain an estimator with clear interpretation and not much condition on $\mathbb{X}$ simply by adding penalty term to the objective function. This intuitively tells us, penalty is good for high dimensional case !

16.2.1 Common penalized regression estimators

By setting the penalty term $f(\beta)$ in RHS of (17.3) to different norms of β, we can have the following estimators of β^* :

- Ridge estimator: $\widehat{\beta}_{\text{ridge}} = \text{argmin}_{\beta\in\mathbb{R}^d} \frac{1}{2n}||Y - \mathbb{X}\beta||^2 + \lambda||\beta||^2$
- Lasso estimator: $\widehat{\beta}_{\text{lasso}} = \text{argmin}_{\beta\in\mathbb{R}^d} \frac{1}{2n}||Y - \mathbb{X}\beta||^2 + \lambda||\beta||_1$
- Best subset selection estimator: $\widehat{\beta}_{\text{BSS}} = \text{argmin}_{\beta\in\mathbb{R}^d} \frac{1}{2n}||Y - \mathbb{X}\beta||^2 + \lambda||\beta||_0$

where $||\ ||$ denote the l_2 norm of a vector and, and $||\ ||_1$ denotes the l_1 norm of a vectoe, and $||\ ||_0$ denotes the l_0 norm of a vector, i.e. number of non-zero entries of a vector.

Specifically, if the columns of $\mathbb{X}$ are orthogonal normal, i.e. $\mathbb{X}^T\mathbb{X} = I_r$, then we will have the following simple closed form of those estimators above.

$$\widehat{\beta}_{\text{ridge}} = \frac{\mathbb{X}^T Y}{1+\lambda}, \quad \widehat{\beta}_{\text{lasso}} = t_{\text{soft}}(\mathbb{X}^T Y, \lambda/2), \quad \widehat{\beta}_{\text{BSS}} = t_{\text{hard}}(\mathbb{X}^T Y, \sqrt{\lambda}).$$

Here $t_{\text{soft}}(\cdot, \lambda)$ is an element wise soft threshold operator defined as $\forall x \in \mathbb{R}^d$, for $i = 1, \ldots, d$,

$$t_{\text{soft}}(x, \lambda)_i = \begin{cases} x_i - \lambda, \text{ if } x_i > \lambda \\ x_i + \lambda, \text{ if } x_i < -\lambda \\ 0, \text{ otherwise .} \end{cases}$$

On the other hand, $t_{\text{hard}}(\cdot, \lambda)$ is an element wise hard threshold operator defined as $\forall x \in \mathbb{R}^d$, for $i = 1, \ldots, d$, $t_{\text{hard}}(x, \lambda) = x_i \mathbb{I}\{|x_i| > \lambda\}$.

16.2.2 Normal mean problem

For random vector $Y \sim N_d(\mu, \sigma^2 \mathbb{I}_d)$, where μ is an unknown vector, and σ^2 is an known parameter, we would like to estimate the mean parameter μ. Define the mean square error of estimator $\widehat{\mu}$ for μ as $\text{MSE}(\widehat{\mu}) = \mathbb{E}||\widehat{\mu} - \mu||^2$.

Take the most familiar estimator – the maximum likelihood estimator $\widehat{\mu}_{\text{mle}} = y$ as an example, we have $\text{MSE}(\widehat{\mu}_{\text{mle}}) = d\sigma^2$; In fact, $\widehat{\mu}_{\text{mle}}$ is not a very good estimator, for example, when $d > 3$, we have the James estimator $\widehat{\mu}_{\text{JS}} = (1 - \frac{d-2}{||y||^2})y$ always dominates the $\widehat{\mu}_{\text{mle}}$, i.e. it always achieve lower MSE than $\widehat{\mu}_{\text{mle}}$.

When we have more information about μ, like the information that μ has many zeros or some other sparse structures, then we could derive other better estimators for μ. One could refer to work of Iian Johnstone at Stanford for more information.

16.2.3 Comments on Best subset selections

Same reason as μ has many zeros could result in better estimators (smaller MSE) in normal mean problem, in linear regression problem, we could also assume that some (in fact, many) coordinates of β^* are zeros (this is actually true when $d > n$), to have better estimators of β^*.

In fact, if one is interested in estimating β^* and its support $\{i : \beta_i \neq 0\}$, then ridge regression is no good, since it only makes the coordinates value of estimators as small as possible, but rarely set them to zeros, so it is hard to identify the support set of β^*. Ideally, best subset selection is more suitable here:

$$\widehat{\beta}_{\text{BSS}} = \text{argmin}\frac{1}{2n}||Y - \mathbb{X}\beta||^2 + \lambda||\beta||_0$$

And we have the following bound for the estimator $\mathbb{X}\widehat{\beta}_{\text{BSS}}$, stated as Theorem 16.1.

Theorem 16.1 *For the best subset selection estimator $\widehat{\beta}_{\text{BSS}}$, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\frac{1}{n}||\mathbb{X}(\widehat{\beta}_{\text{BSS}} - \beta^*)||^2 \lesssim ||\beta^*||_0 \frac{\sigma^2 \log(ed/\delta)}{n}$$

where e is the Euler's number.

However, the best selection question is a non-convex problem, and thus has some computational difficulty: normally we do not know the support of β^*, which means we have to try $O(C_n^k)$ candidates of $\widehat{\beta}_{\text{BSS}}$, which is really computational expensive. If we know the support of β^*, then $\widehat{\beta}_{\text{BSS}}$ is perfect since it would be easy to compute and almost as good as oracle. On the other hand, if we could put some assumptions on $\mathbb{X}$, maybe we can gain some similar results that is nearly as good as oracle and also computational feasible.

16.2.4 Comments on Lasso

Recall the definition of Lasso estimator:

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

Compare to the ridge and best subset selection, Lasso, however, could not only do model selection (i.e. setting some (depend on λ) coordinates of β as zeros), but also is a convex problem (when $n > d$) that is easy to compute. Also, Lasso will have unique solution if columns of $\mathbb{X}$ are drawn from continuous distributions.

There are other Lasso-like problems, which fundamentally is Lasso problem.

- $\min_{\beta \in \mathbb{R}^d} ||\beta||_1$, s.t. $\frac{1}{2n}||Y - \mathbb{X}\beta||^2 \leq B$
- $\min_{\beta \in \mathbb{R}^d} \frac{1}{2n}||Y - \mathbb{X}\beta||^2$, s.t. $||\beta||_1 \leq B$

The following theorem 16.2 states a bound for Lasso estimator, which is known as slow rate for Lasso.

Theorem 16.2 *If $\lambda = \lambda_n \geq ||\frac{\epsilon^T \mathbb{X}}{n}||_\infty$, then for any Lasso solution $\widehat{\beta}_{\text{lasso}}$, we have*

$$\frac{||\mathbb{X}(\widehat{\beta}_{\text{lasso}} - \beta^*)||^2}{n} \leq 4||\beta^*||_1 \lambda_n$$

Proof: We have basic inequality:

$$\frac{1}{2n}||\mathbb{X}(\widehat{\beta}_{\text{lasso}} - \beta^*)||^2 \leq \epsilon^T \frac{\mathbb{X}(\widehat{\beta}_{\text{lasso}} - \beta^*)}{n} + \lambda_n(||\beta^*||_1 - ||\widehat{\beta}_{\text{lasso}}||_1). \tag{16.4}$$

This is because $Y = \mathbb{X}\beta + \epsilon$ and

$$\frac{1}{2n}||Y - \mathbb{X}\widehat{\beta}_{\text{lasso}}||^2 + \lambda||\widehat{\beta}_{\text{lasso}}||_1 \leq \frac{1}{2n}||Y - \mathbb{X}\beta^*||^2 + \lambda||\beta^*||_1$$

Then we bound the RHS of basic inequality (16.4) with

$$\begin{aligned}
&||\frac{\epsilon^T\mathbb{X}}{n}||_\infty||\widehat{\beta}_{\text{lasso}} - \beta^*||_1 + \lambda_n(||\widehat{\beta}_{\text{lasso}}||_1 - ||\beta^*||_1) \\
&\leq \lambda_n(||\widehat{\beta}_{\text{lasso}}||_1 + ||\beta^*||_1) + \lambda_n(||\beta^*||_1 - ||\widehat{\beta}_{\text{lasso}}||_1) \\
&= 2\lambda_n||\beta^*||_1
\end{aligned}$$

This concludes the proof. ■

Theorem 16.2 is useful if we have an upper bound on λ_n, that holds with high probability. We will talk about this more carefully in the next lecture.