36-709: Advanced Statistical Theory I **Spring 2020**

Lecture 14: March 5

Lecturer: Siva Balakrishnan *Scribe: Nicholas Kissel*

14.1 Recap of LASSO Theory

Broadly, there are 4 possible goals that one may have when using the LASSO.

Prediction	**Estimation**	**Variable Selection**	**Inference**
RE + sparsity $\sim \frac{s \log p}{n}$	RE + sparsity	Mutual incoherence	Need to debias
col norm $\sim \sqrt{\frac{\log p}{n}} \|\theta^*\|_1$	$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{s \log p}{n}$	and β_{min} condition	$\approx$ RE + (approximate inv.
		$\text{supp}(\widehat{\theta}) = \text{supp}(\theta^*)$	cov. matrix)

14.2 Debiasing

If we want to do a hypothesis test or form a confidence interval for θ_j^*, we will need to debias the lasso solution. This is because $\widehat{\theta}_j$ is not centered around θ_j^*–meaning, we have a bias problem. In order to fix this, we must add a term to our estimate $\widehat{\theta}$ that depends on some new term $M \approx \left(\frac{X^T X}{n}\right)^{-1}$,

$$\begin{aligned}
\widetilde{\theta} &= \widehat{\theta} + \frac{M X^T (Y - X\widehat{\theta})}{n} \\
&= \widehat{\theta} + \frac{M X^T (X\theta^* - X\widehat{\theta})}{n} + \frac{M X^T \epsilon}{n} \\
&= \widehat{\theta} + M \frac{X^T (X^T X)}{n} (\theta^* - \widehat{\theta}) + \frac{M X^T \epsilon}{n} \\
\implies \widetilde{\theta} - \theta^* &= \left(I - M \frac{X^T X}{n}\right) (\widehat{\theta} - \theta^*) + \frac{M X^T \epsilon}{n} \\
\implies \widetilde{\theta}_j - \theta_j^* &\stackrel{d}{=} \left(\frac{M X^T \epsilon}{n}\right)_j + \underbrace{\left\| I - M \frac{X^T X}{n} \right\|_\infty}_{\leq |e|} \|\widehat{\theta} - \theta^*\|_1
\end{aligned}$$

$$\text{...or better yet } \widetilde{\theta}_j - \theta_j^* \sim N\left(0, M\frac{X^TX}{n^2}M\sigma_\epsilon^2\right)_{jj} + e$$

which allows us to form a confidence interval if the error term e is small $\implies c_j = \widetilde{\theta}_j \pm z_{\frac{\alpha}{2}}\sigma_{jj}$, which will have the correct coverage if e is small relative to σ_{jj}.

We also observe that $\widetilde{\theta}$ isn't as good as $\widehat{\theta}$ in an L_2 sense–that is, $\|\widetilde{\theta} - \theta^*\|_2 \gg \|\widehat{\theta} - \theta^*\|_2$. To answer why this happens, we can look at the expected value of its L_2 distance (ignoring e),

$$\begin{aligned}\mathbb{E}\|\widetilde{\theta} - \theta\|_2^2 &= \frac{\sigma_\epsilon^2 \text{tr}\left(M\frac{X^TX}{n}M\right)}{n} \\ &= \sigma_\epsilon^2\frac{d}{n}\end{aligned}$$

which is similar to the least squares estimator. So by debiasing the lasso in this way, we've lost all of the sparsity—that is, we did not gain from the structure that we initially assumed. By debiasing, we decreased bias but at the cost of the variance, which increased greatly.

We need to add some sort of restriction in order to get a useful debiased estimator. One option may be to impose $|e| \ll \frac{1}{\sqrt{n}}$. If we make the usual RE and sparsity assumptions, then

- $\|\widehat{\theta} - \theta\|_1 \to s\sqrt{\frac{\log p}{n}}$.

On the other hand we still need to impose conditions under which we might be able to estimate the approximate inverse covariance matrix M. Lets try to understand a simple (very special) case. Suppose that $X \sim N(0, I)$ (so we actually knew the true covariance matrix) and we chose $M = I$. Then we know from our previous lectures on estimating the covariance matrix in the ℓ_∞ norm that,

- $\left\|I - \left(\frac{X^TX}{n}\right)\right\|_\infty \ll \sqrt{\frac{s\log p}{n}}$.

This in turn suggests that as long as: $s^{3/2}\log p/\sqrt{n} \to 0$ then our error term $|e| \ll 1/\sqrt{n}$ and we can use debiasing for inference. The real story (when the covariance matrix of the design is not the identity and needs to be estimated from data) is more complicated.

The motivation for debiasing comes from the KKT conditions. Suppose we solved the lasso, then the KKT conditions give

$$\frac{-X^T(Y - X\widehat{\beta})}{n} + \lambda\widehat{Z} = 0 \implies \underbrace{\frac{-MX^T(Y - X\widehat{\beta})}{n}}_{\text{centered}} + \lambda M\widehat{Z} = 0$$

where

$$\frac{-MX^T(Y - X\widehat{\beta})}{n} = M\frac{X^TX}{n}(\beta^* - \widehat{\beta}) + \lambda M\widehat{Z} = \frac{MX^T\epsilon}{n}$$

and so the bias is going to be proportional to λ. What we can do is add $\lambda M\widehat{Z}$ to our $\widehat{\beta}$ to fix the bias problem. To be clearer,

$$\lambda M\widehat{Z} = \frac{MX^T(Y - X\widehat{\beta})}{n}$$

which is exactly what we added back to the lasso above.

14.3 Minimax Lower Bounds

Let $X_1, \ldots, X_n \sim P$ with parameter $\theta^*(P)$ and we use estimator $\widehat{\theta}(P)$ and loss $\rho(\widehat{\theta}, \theta(P))$. Here, our loss is defined as a semi-metric $\rho : \Omega \times \Omega \to [0, \infty)$. Consider the risk,

$$R(\widehat{\theta}, \theta) = \mathbb{E}\rho(\widehat{\theta}, \theta(P)).$$

We can write the minimax risk in the following ways

$$M(\rho) = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} R(\widehat{\theta}, \theta(P))$$

$$M(\Phi \circ \rho) = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}[\Phi \circ \rho(\widehat{\theta}, \theta(P))] \text{ with increasing function } \Phi \text{ on } \mathbb{R}^+$$

which can be thought of as the worst-case risk. Note that the second formulation exists so to allow for evaluating risks not defined by a norm. To do so, we define $\Phi : [0, \infty) \to [0, \infty)$. Using this more general formulation allows us to evaluate risks defined by squared norms.

14.3.1 Estimation as multiple testing

We can reduce this problem from estimation to multiple testing. Broadly, the motivation is that an estimation problem is at least as hard as a certain multiple testing problem—here, we can think of having some collection of distributions a subset of them is chosen, and we will also have a sample (from a distribution in our subet), then we must figure out which distribution our sample came from. Our goal is to construct some function $\psi(Z)$ that identifies which distribution the sample Z comes from. More explicitly, we must pick out some subset $\{\theta_1, \ldots, \theta_M\}$ that is 2δ-separated, so

$$\rho(\theta_i, \theta_j) \geq 2\delta \;\; \forall i \neq j, \text{ then } M(\Phi \circ \rho) \geq \Phi(\delta) \inf_{\psi} Q(\psi \neq j)$$

where $Q(\psi(Z) \neq j)$ is the error and $j, \psi(Z) \in \{1, \ldots, M\}$ and $X_1, \ldots, X_n \sim P_{\theta_j}$.

14.3.2 Le Cam

There 2 main ways of doing this. Option 1 is using Le Cam's method. So for some θ_1 and θ_2, we want to lower bound the testing error

$$\inf_{\psi}\left[\frac{1}{2}\mathbb{P}_{\theta_1}[\psi(Z)\neq 1]+\frac{1}{2}\mathbb{P}_{\theta_2}[\psi(Z)\neq 2]\right]$$

and we know that the optimal test to use here is the likelihood ratio test (LRT). The error of the LRT is

$$\frac{1}{2}\left[\frac{1}{2}-TV(P_{\theta_1},P_{\theta_2})\right]=Q(\psi(Z)\neq j).$$

To get a lower bound using Le Cam's lemma, we will pick θ_1 and θ_2 such that

$$TV(P_{\theta_1}^n,P_{\theta_2}^n)<1-\epsilon \implies Q(\psi\neq j)\geq\frac{\epsilon}{2} \implies M(\Phi\circ\rho)\geq\Phi(\delta)\frac{\epsilon}{2}.$$

In doing this, we want to pick our θ_1 and θ_2 to be as far away as possible without having the TV distance get too large. With this, we can get a bound on the minimax risk.

14.3.3 Fano

Option 2 is using Fano method, which allows us to get a lower bound the testing error even for a multiple testing problem. Fano's method gives,

$$\inf_{\psi}Q(\psi\neq J)\geq 1-\frac{I(Z;J)+\log 2}{\log M}$$

where $J\sim Unif[M]$ and $I(Z;J)=\mathrm{KL}(P_{ZJ}\|P_ZP_J)$. This tells us that if $I(Z;J)\ll\log M$ then $\inf Q(\psi\neq j)$ is larger than some constant (meaning the testing problem is cannot be solved). Note that it may be more helpful to think of the mutual information as

$$I(Z;J)=\mathrm{KL}(P_{ZJ}\|P_ZP_J)=\frac{1}{M}\sum_{i=1}^{M}\mathrm{KL}(P_{\theta_i}\|\widetilde{Q}) \text{ where } \widetilde{Q}=\frac{1}{M}\sum_{i=1}^{M}P_{\theta_i}.$$

Considering everything above we get the following bound,

$$M(\Phi\circ\rho)\geq\Phi(\delta)\left[1-\frac{I(Z;J)+\log 2}{\log M}\right]$$

14.3.4 Summary

In summary, we either

- pick θ_1, θ_2 with $\rho(\theta_1, \theta_2) \geq 2\delta$ and $\text{TV}(P_{\theta_1}, P_{\theta_2}) < 1$ and get $LB \geq \Phi(\rho)$.
- pick $\theta_1, \ldots, \theta_M$ with $\rho(\theta_i, \theta_j) \geq 2\delta \ \forall i \neq j$ and $I(Z; J) \leq \log M$ and get $LB \geq \Phi(\rho)$.

Aside: It is difficult to calculate $TV(P_{\theta_1}^n, P_{\theta_2}^n)$ so it will be useful to appeal to the following fact in the future $TV(P_{\theta_1}^n, P_{\theta_2}^n) \leq \sqrt{\text{KL}(P_{\theta_1}^n, P_{\theta_2}^n)} = \sqrt{n\text{KL}(P_{\theta_1}, P_{\theta_2})}$