Statistics 210B Lecture 21 Notes

Daniel Raban

April 7, 2022

1 LASSO Prediction Error Bound and High-Dimensional Principal Component Analysis

1.1 Recap: overview of results for noisy, sparse linear regression

Let's finish up our analysis of noisy, sparse linear regression. Our model is $y = X\theta^* + w \in \mathbb{R}^n$, where

$$w \in \mathbb{R}^n, \qquad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \qquad \theta^* \in \mathbb{R}^d, \qquad |S(\theta^*)| \leq s.$$

We looked at the λ formulation of the LASSO problem, where

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1.$$

We also looked at the 1-norm constrained and error-constrained formulations of the problem. We defined the $\mathbb{C}_\alpha$ cone

$$\mathbb{C}_\alpha(S) = \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\}.$$

Using this cone, we defined the restricted eigenvalue condition for efficient bounds on estimation.

Definition 1.1. $X \sim \mathrm{RE}(S, (\kappa, \alpha))$ if

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa \|\Delta\|_2^2 \qquad \forall \Delta \in \mathbb{C}_\alpha(S).$$

We proved the following result, upper bounding the estimation error.

Theorem 1.1. *Assume that* $\mathrm{RE}(s, (\kappa, 3))$. *With a proper choice of hyperparameter, we have*

$$\|\widehat{\theta} - \theta^*\|_2 \lesssim \frac{1}{\kappa}\sqrt{s} \left\| \frac{X^\top w}{n} \right\|_\infty \lesssim \sigma \sqrt{\frac{s \log d}{n}}.$$

[insert gaussian random matrix theorem

1.2 LASSO prediction error bound

Instead of bounding $\|\widehat{\theta} - \theta^*\|_2$, we would like to bound the **prediction errror** (with fixed design):

$$\frac{1}{n}\,\mathbb{E}_{\widetilde{w}}[\|\widetilde{y} - X\widehat{\theta}\|_2^2] = \frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2 + \sigma^2,$$

where $\widetilde{y} = X\theta^* + \widetilde{w}$ and $\widetilde{\ }N(0, \sigma^2 I_d)$. We can upper bound $\frac{1}{n}\|X(\widehat{\theta} - \theta^*\|_2^2 \leq \|\widehat{\theta} - \theta^*\|_2^2 \|X^\top X/n\|_{\mathrm{op}}$; however, this is not always a good bound because $\|X^\top X/n\|_{\mathrm{op}}$, which has order d/n (which blows up for $n \ll d$). Instead, we want to bound the prediction error directly

Theorem 1.2 (Prediction error bound)**.** *Let θ^* be s-sparse. Assume that the hyperparameter in the λ-formulation of the LASSO problem is $\lambda_n \geq 2\|\frac{X^\top w}{n}\|_\infty$. Then*

1. *Any optimal solution $\widehat{\theta}$ satisfies the bound*

$$\frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2 \leq 12\|\theta^*\|_1 \lambda_n.$$

2. *If X satisfies* $\mathrm{RE}(S, (\kappa, 3))$*, then*

$$\frac{1}{n}\|X(\widehat{\theta} - \theta^*)\|_2^2 \leq \frac{9}{\kappa} s \lambda_n^2.$$

Proof. As before, the proof is a basic inequality, plus some algebra. □

Remark 1.1. The first bound is $\lesssim \|\theta^*\|_1 \sqrt{\frac{\log d}{n}}$, so we get decay $O(1/\sqrt{n})$. This is called the **slow rate bound**. The second bound is $\lesssim s(\sqrt{\frac{\log d}{n}})^2$, so we get decay $O(1/n)$. This is called the **fast rate bound**. Usually, without imposing any geometric assumptions, we get a slower rate bound than we get with such assumptions.

This phenomenon occurs in many settings such as in the empirical risk minimization problem. [insert pic 1] The setting is that we have data $(z_i)_{i \in [n]} \overset{\text{iid}}{\sim} \mathbb{P}_z$ and a loss function $\ell : \Theta \times Z \to \mathbb{R}$. The **empirical risk** is

$$\widehat{R}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; X_i),$$

and the **population risk** is

$$R(\theta) = \mathbb{E}[\ell(\theta; Z_i)].$$

If we take $\widehat{\theta} = \arg\min_\theta \widehat{R}_n(\theta)$, the minimizer of the empirical risk, then our **generalization error** is

$$R(\widehat{\theta}) - R(\theta^*).$$

Without geometric assumptions, we can show a **uniform convergence bound**

$$R(\widehat{\theta}) - R(\theta^*) \leq 2 \sup_{\theta \in \Theta} |\widehat{R}_n(\theta) - R(\theta)|.$$

Suppose $\Theta = B(0, 10\|\theta^*\|)$. The upper bound of such an empirical process usually scales linearly in $\|\theta^*\|$, which does not give a very sharp prediction error bound.

Here is what we get with a geometric assumption. Assume that $\kappa\|\widehat{\theta} - \theta^*\|_2^2 \leq (R(\widehat{\theta}) - R(\theta^*))$. Here, κ is a **strong convexity parameter**. With this assumption, we can show an upper bound that is like

$$R(\widehat{\theta}) - R(\theta^*) \leq 2 \sup_{\theta \in B(\theta^*, \|\widehat{\theta} - \theta^*\|_2)} |\widehat{R}_n(\theta) - R(\theta)| \lesssim \|\widehat{\theta} - \theta^*\|_2 \sqrt{\frac{d \log d}{n}}.$$

This is nice because it scales linearly in the estimation error, which is usually smaller than $\|\theta^*\|$. We can bound $\|\widehat{\theta} - \theta^*\|_2 \lesssim \sqrt{\frac{d \log d}{n}}$. Applying the geometric assumption gives the bound

$$R(\widehat{\theta}) - R(\theta^*) \leq \frac{d \log d}{n}.$$

1.3 Principal component analysis in high dimensions

Suppose we observe covariates $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} X \in \mathbb{R}^d$ with $\mathbb{E}[X] = 0$ and $\mathrm{Cov}(X) = \Sigma \in S_+^{d \times d}$. Let the eigenvalues of Σ be $\lambda_1(\Sigma) \geq \lambda_2(\Sigma) \geq \cdots \geq \lambda_d(\Sigma) \geq 0$. We can find an orthonormal basis of eigenvectors $v_1(\Sigma), \ldots, v_d(\Sigma) \in \mathbb{R}^d$ such that $\Sigma v_i = \lambda_i v_i$ for all $i \in [d]$. If we let $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^{d \times d}$ nad $B = [v_1, \ldots, v_d] \in \mathbb{R}^{d \times d}$, then we can write $\Sigma = V \Lambda V^\top$.

The statistical interpretation of v_1 is that

$$\begin{aligned} v_1 &\in \arg\max_{\|v\|_2 = 1} \mathrm{Var}(\langle x, v \rangle) \qquad X \in \mathbb{R}^d, \mathbb{E}[X] = 0. \\ &= \arg\max_{\|v\|_2 = 1} \langle v, \mathbb{E}[XX^\top] v \rangle \\ &= \arg\max_{\|v_2\| = 1} \langle v, \Sigma v \rangle. \end{aligned}$$

More generally, if we let $V_k = [v_1, \ldots, v_k] \in \mathbb{R}^{d \times k}$, then

$$V_k \in \arg\max_{\substack{U \in R^{d \times k} \\ \text{partial orth.}}} \underbrace{\mathbb{E}[\|U^\top X\|_2^2]}_{\sum_{i=1}^k \mathrm{Var}(\langle X, u_i \rangle)}.$$

Here is our statistical question: Given samples $\{X_i\}_{i \in [n]} \overset{\text{iid}}{\sim} X \in \mathbb{R}^d$, how can we estimate the principal components? Straightforwardly, we can use the eigenvectors of the sample covariance. If we define the sample covariance matrix

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top, \qquad \mathbb{E}[\widehat{\Sigma}] = \Sigma,$$

then our estimator is

$$\widehat{\arg\max} = \arg\max_{\theta} \langle \theta \widehat{\Sigma} \theta \rangle.$$

By comparison, the ground truth is

$$\theta^* = \arg\max_{\|\theta\|_2 = 1} \langle \theta, \Sigma\theta \rangle.$$

How close is $\widehat{\theta}$ to θ^*? We want to translate the closeless of Σ and $\widehat{\Sigma}$ to closeness of θ and θ^*. To quantify this, recall Weyl's eigenvalue perturbation inequality:

Lemma 1.1 (Weyl's inequality)**.** *For any matrices* $\widehat{\Sigma}, \Sigma$*,*

$$|\lambda(\widehat{\Sigma}) - \lambda_i(\Sigma)| \leq \|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}.$$

The proof of this fact comes from the variational characterization of the eigenvalues.

For a perturbation inequality for the eigenvectors, we also need the first eigen-gap to be large.

Definition 1.2. Let $\lambda_1(\Sigma) \geq \lambda_2(\Sigma) \geq \cdots \geq \lambda_d(\Sigma)$ be the eigenvalues of Σ. Then k-th **eigen-gap** is $\nu_k = \lambda_k - \lambda_{k+1}$.

We will write $\nu = \nu_1$ to refer to the first eigen-gap. You can think of having a large eigen-gap as similar to the restricted eigenvalue condition for LASSO. The parameter ν plays a similar role to κ in LASSO, where $\mathrm{RE}(S, (\kappa, 3))$ means that $\Delta^\top \frac{X^\top X}{n} \Delta \geq \kappa \|\Delta\|_2^2$.

Example 1.1. Here is an example of instability of a matrix with a small eigengap. Suppose we have a diagonal matrix

$$Q_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1.01 \end{bmatrix}.$$

The eigenvalues are $\lambda_1(Q_0) = 1.01$ and $\lambda_2(Q_0) = 1$, so the eigengap is $\nu(Q_0) = 0.01$. In this case, $\theta^*(Q_0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Now look at the perturbation

$$Q_\varepsilon = Q_0 + \varepsilon \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 1.01 \end{bmatrix},$$

where ε is small. If $\varepsilon = 0.01$, then $\theta^*(Q_\varepsilon) \approx \begin{bmatrix} 0.53 \\ 0.85 \end{bmatrix}$, which is far from $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

1.4 General perturbation bound for eigenvectors

Theorem 1.3. *Let $\Sigma \in S_+^{d\times d}$, and let $\theta^* \in \mathbb{R}^d$ be an eigenvector for $\lambda_1(\Sigma)$. Let $\nu = \lambda_1(\Sigma) - \lambda_2(\Sigma) > 0$ be the first eigen-gap. Let the perturbation $P \in S^{d\times d}$ be such that $\|P\|_{\mathrm{op}} < \nu/2$, and let $\widehat{\Sigma} = \Sigma + P$. If $\widehat{\theta} \in \mathbb{R}^d$ is an eigenvector for $\lambda_1(\widehat{\Sigma})$, then*

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{2\|\widetilde{P}\|_2}{\nu - 2\|P\|_{\mathrm{op}}}.$$

Here

$$\widetilde{P} = U^\top P U = \begin{bmatrix} \widetilde{P}_{1,1} & \widetilde{P}^\top \\ \widetilde{P} & \widetilde{P}_{2,2} \end{bmatrix} \in \mathbb{R}^{d\times d},$$

where U is the orthogonal matrix such that $\Sigma = U\Lambda U^\top$ and the blocks of $\widetilde{P}$ have sizes

$$\begin{bmatrix} 1\times 1 & d\times(d-1) \\ (d-1)\times 1 & (d-1)\times(d-1) \end{bmatrix}.$$

If $\|P\|_{\mathrm{op}}$, then we get the bound

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{4}{\nu}\|\widetilde{P}\|_2 \leq \frac{4}{\nu}\|P\|_{\mathrm{op}}.$$

To prove this, first let $\widehat{\Delta} = \widehat{\theta} - \theta^*$, and define the quantity

$$\begin{aligned}\Psi(\widehat{\Delta}; P) &= \langle \widehat{\theta}, P\widehat{\theta}\rangle - \langle \theta^*, P\theta^*\rangle \\ &= \langle \widehat{\Delta}, P\widehat{\Delta}\rangle + 2\langle \widetilde{\Delta}, P\theta^*\rangle.\end{aligned}$$

Here is the basic inequality of PCA:

Lemma 1.2 (PCA basic inequality)**.**

$$\nu\cdot(1 - \langle \widehat{\theta}, \theta^*\rangle^2) \leq |\psi(\widehat{\Delta}; P)|.$$

The left hand side measures the distance between $\widehat{\theta}$ and θ^*. We first prove this basic inequality:

Proof. The zero order optimality condition for $\widehat{\theta}$ says that $\widehat{\theta} = \arg\max_\theta \langle \theta, \widehat{\Sigma}\theta\rangle$. Then

$$\langle \widehat{\theta}, \widehat{\Sigma}\widehat{\theta}\rangle \geq \langle \theta^*, \widehat{\Sigma}\theta^*\rangle.$$

Recall that $\widehat{\Sigma} = \Sigma + P$. We can express this inequality as

$$\langle \widehat{\theta}, \Sigma\widehat{\theta}\rangle + \langle \widehat{\theta}, P\widehat{\theta}\rangle \geq \langle \theta^*, \Sigma\theta^*\rangle + \langle \theta^*, P\theta^*\rangle.$$

Putting the like terms on each side gives

$$\langle \theta^*, \Sigma\theta^* \rangle - \langle \widehat{\theta}, \Sigma\widehat{\theta} \rangle \leq \langle \widehat{\theta}, P\widehat{\theta} \rangle - \langle \theta^*, P\theta^* \rangle.$$

The right hand side is $\psi(\widehat{\Delta}; P)$.

To figure out the left hand side, write $\widehat{\theta} = \rho\theta^* + \sqrt{1-\rho^2}z$, where $\|z\|_2 = 1$, $\langle z, \theta^* \rangle = 0$. Then $\rho = \langle \widehat{\theta}, \theta^* \rangle$. We can then expand

$$\begin{aligned}\langle \widehat{\theta}, \Sigma\widehat{\theta} \rangle &= \langle \rho\theta^* + \sqrt{1-\rho^2}z, \Sigma(\rho\theta^* + \sqrt{1-\rho^2}z) \rangle \\ &= \rho^2 \underbrace{\langle \theta^*, \Sigma\theta^* \rangle}_{=\lambda_1} + 2\rho\sqrt{1-\rho^2} \underbrace{\langle \theta^*, \Sigma z \rangle}_{=0} + (1-\rho^2) \underbrace{\langle z, \Sigma z \rangle}_{\leq 2}.\end{aligned}$$

The bound on the last term is because $\langle z, \Sigma z \rangle \leq \sup_{\|z\|_2=1, \langle z, \theta^* \rangle = 0} \langle z, \Sigma z \rangle = \lambda_2$.

$$\leq \rho^2\lambda_1 + (1-\rho^2)\lambda_2.$$

So the left hand side is

$$\begin{aligned}\langle \theta^*, \Sigma\theta^* \rangle - \langle \widehat{\theta}, \Sigma\widehat{\theta} \rangle &\geq \lambda_1 - (\rho^2\lambda_1 + (1-\rho^2)\lambda_2) \\ &= (\lambda_1 - \lambda_2)(1-\rho^2) \\ &= \nu(1-\rho^2).\end{aligned}$$

So we get

$$\nu(1 - \langle \widehat{\theta}, \theta^* \rangle^2) \leq \Psi(\widehat{\Delta}; P).$$ □

Proof. Given the basic inequality, we now upper bound

$$\Psi(\widehat{\Delta}; P) = \langle \widehat{\theta}, P\widehat{\theta} \rangle - \langle \theta^*, P\theta^* \rangle.$$

Write $\Sigma = U\Lambda U^\top$ and $P = U\widetilde{P}U^\top$. We know that $U^\top\theta^* = e_1$, the first standard basis vector, so

$$U^\top\widehat{\theta} = U^\top(\rho\theta^* + \sqrt{1-\rho^2}z) + \rho e_1 + \sqrt{1-\rho^2}\underbrace{U^\top z}_{=:\widetilde{z}},$$

where $\|\widetilde{z}\|_2 = 1$. Then

$$\begin{aligned}\Psi(\widehat{\Delta}; P) &= \langle U^\top\widehat{\theta}, \widetilde{P}U^\top\widehat{\theta} \rangle - \langle U^\top\theta^*, \widetilde{P}U^\top\theta^* \rangle \\ &= \langle \rho e_1 + \sqrt{1-\rho^2}\widetilde{z}, \widetilde{P}(\rho e_1 + \sqrt{1-\rho^2}\widetilde{z}) \rangle - \langle e_1, \widetilde{P}e_1 \rangle \\ &= \rho^2\langle e_1, \widetilde{P}e_1 \rangle + 2\rho\sqrt{1-\rho^2}\langle \widetilde{z}, \widetilde{P}e_1 \rangle + (1-\rho^2)\langle \widetilde{z}, \widetilde{P}\widetilde{z} \rangle - \langle e_1, \widetilde{P}e_1 \rangle \\ &= (1-\rho^2)\underbrace{\langle e_1, \widetilde{P}e_1 \rangle}_{\leq \|P\|_{\mathrm{op}}} + (1-\rho^2)\langle \widetilde{z}, \widetilde{P}\widetilde{z} \rangle + 2\rho\sqrt{1-\rho^2}\underbrace{\langle \widetilde{z}, \widetilde{P}e_1 \rangle}_{\leq \|P\|_2}.\end{aligned}$$

So, using the basic inequality, we get

$$\nu(1-\rho^2) \leq 2(1-\rho^2)\|P\|_{\mathrm{op}} + 2\rho\sqrt{1-\rho^2}\|\widetilde{P}\|_2.$$

We can solve this to get

$$\sqrt{1-\rho^2} \leq \frac{2\rho\|\widetilde{P}\|_2}{\nu - 2\|P\|_{\mathrm{op}}}$$

So

$$\begin{aligned}
\|\widehat{\theta} - \theta^*\|_2 &= \sqrt{2(1-\rho)} \\
&\leq \frac{\sqrt{2}\rho}{\sqrt{1+\rho}} \frac{2\|\widetilde{P}\|_2}{\nu - 2\|P\|_{\mathrm{op}}} \\
&\leq \frac{2\|\widetilde{P}\|_2}{\nu - 2\|P\|_{\mathrm{op}}}.
\end{aligned}$$

□