STAT210B Theoretical Statistics Lecture 21 - 04/06/2023

Lecture 21: High Dimensional PCA

Instructor: Song Mei *Scriber: Sharan Sahu* *Proof reader: Raghav Ramanujam*

1 LASSO Prediction Error Bound

Instead of bounding $||\hat{\theta} - \theta^*||_2^2$, we would like to bound our prediction error (with fixed design), which is given by

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

where $\tilde{y} = X\theta^* + \tilde{w}$. It is possible to bound $\frac{1}{n}||X(\hat{\theta} - \theta^*)||_2^2$ by $||\frac{1}{n}X^\top X||_{\text{op}}||\hat{\theta} - \theta^*||_2^2$, but since $||\frac{1}{n}X^\top X||_{\text{op}}$ is of order d/n (which blows up for $n << d$), this is not always desirable. Instead, we want to bound the prediction error directly

Theorem 1 (Prediction error bound)**.** *Let θ^* be s-sparse. Let $\hat{\theta}$ be the minimizer of λ-formulation with $\lambda_n \geq 2||\frac{X^T w}{n}||_\infty$. Then,*

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

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

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

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

Proof. The theorem follows from application of the basic inequality followed by algebra involving the Holder's and the triangle inequality. □

Remark 2. *Note that (1) from the theorem tells us that the bound $\lesssim C||\theta^*||_1 \cdot \sqrt{\frac{\log d}{n}} = \mathcal{O}(\frac{1}{\sqrt{n}})$. This is our slow rate bound. The second bound from the theorem gives the bound $\lesssim C||\theta^*||_1 \cdot \frac{\log d}{n} = \mathcal{O}(\frac{1}{n})$. This is our fast rate bound.*

In general, slow rates tend to be provable without local geometric assumptions whereas fast rates require such assumptions.

2 Principal Component Analysis In High Dimension

Consider $X_1, X_2, \cdots, X_n \sim_{iid} X \in \mathbb{R}^d$ such that $\mathbb{E}[X] = 0$ and $\text{Cov}(X) = \Sigma \in \mathcal{S}_+^{d\times d}$.

Definition 3 (Eigenvalue decomposition of $\Sigma \in \mathcal{S}_+^{d\times d}$)**.** *There exists*

$$\lambda_1(\Sigma) \geq \lambda_2(\Sigma) \geq \cdots \geq \lambda_d(\Sigma) \geq 0$$

and pairwise orthonormal vectors

$$v_1(\Sigma), \cdots v_n(\Sigma) \in \mathbb{R}^d$$

such that $\Sigma v_i = \lambda_i v_i, \forall i \in [d]$. *In matrix notation, we can represent this as*

$$\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_d) \in \mathbb{R}^{d \times d}$$

$$V = [v_1, \cdots, v_d] \in \mathbb{R}^{d \times d}$$

$$\Sigma = V \Lambda V^\top \iff \Sigma V = V \Lambda$$

A statistical interpretation of $v_1, \cdots, v_d$ is as follows: First, we notice that

$$\text{argmax}_{||v||_2=1} \text{Var}(\langle X, v \rangle), X \in \mathbb{R}^d, \mathbb{E}[X] = 0 \tag{2}$$

is equivalent to

$$\text{argmax}_{||v||_2=1} \mathbb{E}[\langle v, XX^\top v \rangle] \tag{3}$$

Since

$$\text{argmax}_{||v||_2=1} \text{Var}(\langle X, v \rangle) = \text{argmax}_{||v||_2=1} \mathbb{E}[\langle X, v \rangle^2] = \text{argmax}_{||v||_2=1} \mathbb{E}[v^\top X X^\top v] \tag{4}$$

Then, we have that

$$\text{argmax}_{||v||_2=1} \mathbb{E}[\langle c, XX^\top c \rangle] = \text{argmax}_{||v||_2=1} \mathbb{E}\langle v, \mathbb{E}[XX^\top] v \rangle = \text{argmax}_{||v|_2=1} \mathbb{E}\langle v, \Sigma v \rangle = v_1 \tag{5}$$

Therefore, we can interpret v_1 as the max variance direction. More generally, we have for $V_k = [v_1, ..., v_k] \in \mathbb{R}^{d \times k}$

$$V_k \in \text{argmax}_{U \in \mathbb{R}^{d \times k}} \mathbb{E}[||U^T x||_2^2] = \text{argmax}_{[u_1, ..., u_k] \in \mathbb{R}^{d \times k}} \sum_{i=1}^{k} \text{Var}(\langle u_1, X \rangle) \tag{6}$$

where $[u_1, ..., u_k]$ are orthonormal.

Now, we ask ourselves a statistical question: Given samples $\{x_i\}_{i \in [n]} \sim_{\text{iid}} X \in \mathbb{R}^d$, how to estimate the principle components? Focusing on when $k = 1$, and defining $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^T$, let

$$\text{Estimator} : \hat{\theta} \in \text{argmax}_{||\theta||_2=1} \langle \theta, \hat{\Sigma} \theta \rangle \tag{7}$$

$$\text{Ground Truth} : \theta^* \in \text{argmax}_{||\theta||_2=1} \langle \theta, \Sigma \theta \rangle \tag{8}$$

How can we bound $\hat{\theta} - \theta^*$? Recall Weyl's inequality which tells us that

$$|\lambda_i(\hat{\Sigma}) - \lambda_i(\Sigma)| \leq ||\hat{\Sigma} - \Sigma||_{\text{op}} \tag{9}$$

To bound $\hat{\theta} - \theta^*$, we also need the eigengap, which is defined as $\nu = \lambda_1(\Sigma) - \lambda_2(\Sigma)$, to be large. To see why this is important, let us take a look at an example

Example (Instability With Small Eigengap): Let

$$Q_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1.01 \end{bmatrix} \tag{10}$$

We have that

$$\lambda_1(Q_0) = 1.01, \lambda_2(Q_0) = 1, \nu(Q_0) = 0.01, v_1(Q_0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{11}$$

Now, let us define

$$Q_\epsilon = Q_0 + \epsilon \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \epsilon \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1.01 \end{bmatrix} \tag{12}$$

If we let $\epsilon = 0.01$, then $v_1(Q_\epsilon) \approx \begin{bmatrix} 0.53 \\ 0.85 \end{bmatrix}$. Therefore, if we have a small eigengap, even a small perturbation can mix our eigenspace.

3 General Perturbation Bound For Eigenvectors

We now state a theorem for the (loose) general perturbation bound for eigenvectors:

Theorem 4. *Let $\Sigma \in \mathcal{S}_+^{d\times d}$, and let $\theta^* \in \mathbb{R}^d$ be the eigenvector for $\lambda_1(\Sigma)$. Let $\nu = \lambda_1(\Sigma) - \lambda_2(\Sigma) > 0$ be the first eigengap. Let $P = \hat{\Sigma} - \Sigma \in \mathcal{S}^{d\times d}$. If $\hat{\theta} \in \mathbb{R}^d$ is the eigenvector for $\lambda_1(\hat{\Sigma})$. Then*

$$||\hat{\theta} - \theta^*||_2 \leq \frac{2\sqrt{2} \cdot ||P||_{\text{op}}}{\nu}$$

We can actually obtain a sharper bound seen below:

Theorem 5. *Under the same conditions as the Theorem above,*

$$||\hat{\theta} - \theta^*|| \leq \frac{2||\tilde{P}||_2}{\nu - 2||P||_{\text{op}}}$$

where

$$\tilde{P} = U^T P U = \begin{bmatrix} \tilde{P}_{11} & \tilde{P}^\top \\ \tilde{P} & \tilde{P}_{22} \end{bmatrix} \in \mathbb{R}^{d\times d}$$

and U is defined such that $\Sigma = U\Lambda U^\top$

Before we prove Theorem 4, we introduce the following lemma which will help us in our proof

Lemma 6 (PCA Basic Inequality)**.**

$$||\theta - \theta^*||_2^2 = 1 - \langle\hat{\theta}, \theta^*\rangle \leq 1 - \langle\hat{\theta}, \theta^*\rangle^2 \leq \frac{1}{\nu}[\langle\hat{\theta}, P\hat{\theta}\rangle - \langle\theta^*, P\theta^*\rangle]$$

Proof. First, let us note that

$$||\hat{\theta} - \theta^*|| = ||\hat{\theta}||_2^2 + ||\theta^*||_2^2 - 2\langle\hat{\theta}, \theta^*\rangle = 2(1 - \langle\hat{\theta}, \theta^*\rangle)$$

Furthermore, since $0 \leq \langle\hat{\theta}, \theta^*\rangle \leq 1$, we have that $1 - \langle\hat{\theta}, \theta^*\rangle \leq 1 - \langle\hat{\theta}, \theta^*\rangle^2$. Since $\hat{\theta} = \text{argmax}_\theta\langle\theta, \hat{\Sigma}\theta\rangle$, we have that $\langle\hat{\theta}, \hat{\Sigma}\hat{\theta}\rangle \geq \langle\theta^*, \hat{\Sigma}\theta^*\rangle$. With some manipulation, we obtain that

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

This implies that

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

Now, we decompose $\hat{\theta}$ as $\rho\theta^* + \sqrt{1-\rho^2}z$, where $\rho = \langle\hat{\theta}, \theta^*\rangle \in [0, 1]$, $||z||_2 = 1$, and z is orthogonal to θ^*. Using these definitions, we get that

$$\langle\hat{\theta}, \Sigma\hat{\theta}\rangle = \langle\rho\theta^* + \sqrt{1-\rho^2}z, \Sigma(\rho\theta^* + \sqrt{1-\rho^2}z)\rangle = \rho^2\langle\hat{\theta}, \Sigma\hat{\theta}\rangle + 2\rho\sqrt{1-\rho^2}\langle\theta^*, \Sigma z\rangle + (1-\rho^2)\langle z, \Sigma z\rangle$$

Now, notice the following

$$\langle\theta^*, \Sigma\theta^*\rangle = \lambda_1(\Sigma)$$

$$\langle z, \Sigma\theta^*\rangle = \lambda_1(\Sigma)\langle z, \theta^*\rangle = 0$$

$$\langle z, \Sigma z\rangle \leq \sup_{z:||z||_2=1, \langle z,\theta^*\rangle=0} \langle z, \Sigma z\rangle = \lambda_2(\Sigma)$$

Using these, we get that $\langle\hat{\theta}, \Sigma\hat{\theta}\rangle \leq \rho^2\lambda_1(\Sigma) + (1-\rho^2)\lambda_2(\Sigma)$. Using this, we get that

$$\langle\theta^*, \Sigma\theta^*\rangle - \langle\hat{\theta}, \Sigma\hat{\theta}\rangle = \lambda_1(\Sigma) - \rho^2\lambda_1(\Sigma) + (1-\rho^2)\lambda_2(\Sigma) = (1-\rho^2)(\lambda_1(\Sigma) - \lambda_2(\Sigma)) = (1-\rho^2)\nu$$

By combining our results, we have proved the PCA Basic Inequality □

Now that we have the PCA Basic Inequality, we can now try to bound $\Psi(P) = \langle \hat{\theta}, P\hat{\theta} \rangle - \langle \theta^*, P\theta^* \rangle$. First, we note that we can naively bound $\Psi(P) \leq 2||P||_{\text{op}}$. Then, using the PCA Basic Inequality, we see that

$$\frac{1}{2}||\hat{\theta} - \theta^*||_2^2 \leq \frac{2}{\nu}||P||_{\text{op}}$$

Thus, this implies that

$$||\hat{\theta} - \theta^*||_2 \leq 2\sqrt{\frac{||P||_{\text{op}}}{\nu}}$$

We will now show a sharper bound of $\Psi(P) \leq 4||P||_{\text{op}}\sqrt{1-\rho^2}$

Proof. Notice the following:

$$\Psi(P) \leq \langle \rho\theta^* + z\sqrt{1-\rho^2}, P(\rho\theta^* + z\sqrt{1-\rho^2}) \rangle - \langle \theta^*, P\theta^* \rangle$$

Doing some basic algebra on the right-hand side gets us

$$\Psi(P) \leq (\rho^2 - 1)\langle \theta^*, P\theta^* \rangle + 2\rho\sqrt{1-\rho^2}\langle \theta^*, Pz \rangle + (1-\rho^2)\langle z, Pz \rangle$$

The right-hand side is now bounded by $2||P||_{\text{op}}(1-\rho^2) + 2\rho||P||_{\text{op}}\sqrt{1-\rho^2}$. Thus, we get

$$\Psi(P) \leq 2||P||_{\text{op}}(1-\rho^2) + 2\rho||P||_{\text{op}}\sqrt{1-\rho^2} \leq 4||P||_{\text{op}}\sqrt{1-\rho^2}$$

□

Using this sharper bound, we can now see that

$$(1-\rho^2) \leq \frac{4}{\nu}||P||_{\text{op}}\sqrt{1-\rho^2} \tag{13}$$

Thus,

$$||\hat{\theta} - \theta^*||_2 \leq \sqrt{2}\sqrt{1-\rho^2} \leq \frac{4\sqrt{2}||P||_{\text{op}}}{\nu} \tag{14}$$

Remark 7. *It should be noted that the bound that we proved above is not the sharpest bound. In fact, the sharpest bound is* $\Psi(P) \leq 2||P||_{\text{op}}(1-\rho^2) + 2\rho||\tilde{P}||_{\text{op}}\sqrt{1-\rho^2}$

4 Consequences For A Spiked Ensemble

Let $\theta^* \in \mathbb{R}^d$, $||\theta^*||_2 = 1$. Then, the spiked covariance model is $x_i = \sqrt{\nu}\xi_i\theta^* + w_i \in \mathbb{R}^d$ with $i \in [n]$ where $\xi_i \in \mathbb{R}$, $\mathbb{E}[\xi_i] = 0$, $\mathbb{E}[\xi_i^2] = 1$ and $w_i \in \mathbb{R}^d$, $\mathbb{E}[w_i] = 0$, $\mathbb{E}[w_i w_i^\top] = I_d$. Here, $\{\xi_i\}$ and $\{w_i\}$ are mutually independent. Notice that using the properties above, one can easily note that $\mathbb{E}[x_i] = 0$ and

$$\mathbb{E}[x_i x_i^\top] = \mathbb{E}[(\sqrt{\nu}\xi_i\theta^* + w_i)(\sqrt{\nu}\xi_i\theta^* + w_i)^\top] = \nu\mathbb{E}[\xi_i^2]\theta^*\theta^{*\top} + \mathbb{E}[w_i w_i^\top] = \nu\theta^*\theta^{*\top} + I_d \tag{15}$$

We will denote this Σ. The largest eigenvalue is $\lambda_{\max}(\Sigma) = \nu + 1$. The second largest eigenvalue is $\lambda_2(\Sigma)$. Thus, $\nu = \lambda_{\max}(\Sigma) - \lambda_2(\Sigma)$ is the eigengap, and the leading eigenvector of Σ is θ^*. We can estimate θ by

$$\hat{\theta} = \arg\max_{||\theta||_2 = 1} \langle \theta, \Sigma\theta \rangle \tag{16}$$

Using Theorem 5, we have the following corollary:

Corollary 8. *Assume* $\xi_i \sim \mathrm{SG}(1)$ *and* $w_i \sim \mathrm{SG}(1)$. *If* $n > d$ *and* $\sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}} \leq \frac{1}{128}$, *then*

$$||\hat{\theta} - \theta^*||_2 \lesssim \sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}}$$

with high probability

Proof. Recall that

$$||\hat{\theta} - \theta^*|| \leq \frac{2||\tilde{P}||_2}{\nu - 2||P||_{\mathrm{op}}}$$

We need to upper bound $||\tilde{P}||_2$ and $||P||_{\mathrm{op}}$. Notice that

$$P = \hat{\Sigma} - \Sigma = \frac{1}{n}\sum_{i=1}^{n}(\sqrt{\nu}\xi_i\theta^* + w_i)(\sqrt{\nu}\xi_i\theta^* + w_i)^\top - (\nu\theta^*\theta^{*\top} + I_d)$$

Doing some algebra on the right hand side, we get

$$P = (\frac{1}{n}\sum_{i=1}^{n}\xi_i^2 - 1)\nu\theta^*\theta^{*\top} + \sqrt{\nu}(\frac{1}{n}\sum_{i=1}^{n}\xi_i w_i)\theta^{*\top} + (\frac{1}{n}\sum_{i=1}^{n}\xi_i w_i)\theta^* + (\frac{1}{n}\sum_{i=1}^{n}w_i w_i^\top - I_d)$$

Thus, we get that

$$||P||_{\mathrm{op}} \leq \nu|\frac{1}{n}\sum_{i=1}^{n}\xi_i^2 - 1| + 2\sqrt{\nu}\|\frac{1}{n}\sum_{i=1}^{n}\xi_i w_i\|_2 + \|\frac{1}{n}\sum_{i=1}^{n}w_i w_i^\top - I_d\|_{\mathrm{op}}$$

□

Notice that we can bound the first term by $\sqrt{\frac{1}{n}}$ by the sub-gaussian concentration. The second and third term can be bounded by $\sqrt{\frac{d}{n}}$ by an ε-covering argument. Thus, we get that

$$||P||_{\mathrm{op}} \lesssim \nu\sqrt{\frac{1}{n}} + (\sqrt{\nu}+1)\sqrt{\frac{d}{n}}$$

Similarly, we can bound $||\tilde{P}||_2$ as

$$||\tilde{P}||_2 \leq \sqrt{\nu}\|\frac{1}{n}\sum_{i=1}^{n}\xi_i w_i\|_2 + \|\frac{1}{n}\sum_{i=1}^{n}w_i w_i^\top - I_d\|_{\mathrm{op}}$$

Using the arguments that we used to bound $||P||_{\mathrm{op}}$, we get that

$$||\tilde{P}||_2 \lesssim (\sqrt{\nu}+1)\sqrt{\frac{d}{n}}$$

Now, if $\sqrt{\frac{d}{n}} << \frac{\nu}{\sqrt{\nu}+1}$, then $\nu - 2||P||_{\mathrm{op}} \geq \frac{\nu}{2}$. This gives us

$$||\hat{\theta} - \theta^*||_2 \lesssim \frac{4||\tilde{P}||_2}{\nu} \lesssim \sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}}$$

Let us now see an example of how to use the metric entropy bound for bounding the first term in $||\tilde{P}||_2$.

Example: Notice that

$$\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i w_i\|_2 = \sup_{||\nu||_2=1}\langle \nu, \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i w_i\rangle = \sup_{||\nu||_2=1}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle w_i, \nu\rangle \tag{17}$$

Recall the fact that if $X \sim \text{SG}(\sigma_1)$ and $Y \sim \text{SG}(\sigma_2)$, then $XY \sim \text{SE}(C\sigma_1\sigma_2, C\sigma_1\sigma_2)$ for some constant C. Notice that $\varepsilon_i \sim \text{SG}(1)$ and $\langle w_i, \nu\rangle \sim \text{SG}(1)$. Thus, $\varepsilon_i\langle w_i, \nu\rangle \sim \text{SE}(1,1)$. Using the sub-exponential tail bound, we see that

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle w_i, \nu\rangle| \geq t) \leq 2\exp\{-n\min\{t, t^2\}\} \quad \forall \nu \in \mathcal{S}^{d-1} \tag{18}$$

Now, let $\Omega_{1/4}$ be the 1/4-cover of $\mathcal{S}^{d-1}$. Thus, $|\Omega_{1/4}| \leq C^d$ for some constant C. One can show that this implies

$$\sup_{\nu\in\mathcal{S}^{d-1}}|\langle \nu, a\rangle| \leq 2\sup_{\nu\in\Omega_{1/4}}|\langle \nu, a\rangle| \tag{19}$$

Using the Union Bound,

$$\mathbb{P}(||\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i w_i||_2 \geq t) \leq \mathbb{P}(2\sup_{\nu\in\Omega_{1/4}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle w_i, \nu\rangle \geq t) \leq C^d\exp\{-n\min\{t, t^2\}\} \tag{20}$$

The examples come at the courtesy of [Wai19]

References

[Wai19] Martin J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, first ed., Cambridge University Press, New York, NY, USA, 2019.