41500: High-Dimensional Statistics **Winter 2020**

Sparse PCA

Lecturer: Chao Gao *Scribe: Xinze Li*

First let's recall the sparse PCA model, so $X_1, \cdots, X_n \sim \mathcal{N}(0, \Sigma)$ i.i.d., and suppose that $\Sigma = \lambda\theta\theta^\top + I_p$. So we could observe that the eigenvalue of Σ is

$$(1 + \lambda, 1, \cdots, 1) \tag{1}$$

The leading eigenvalue corresponds to θ. also suppose θ is normalized: $\|\theta\| = 1$ and $\theta \in \Theta(p, s)$ is sparse. We now define a new parameter space $\mathcal{V}(p, s) = \{\theta \in \Theta(p, s) : \|\theta\| = 1\}$. So now we are ready to define our estimator

$$\widehat{\theta} = \arg\max_{\theta \in \mathcal{V}(p,s)} \theta^\top \widehat{\Sigma}\theta \tag{2}$$

where $\widehat{\Sigma} = \frac{1}{n}\sum X_i X_i^\top$ is the sample covariance matrix as usual. Recall that we define matrix innner product as

$$\langle A, B\rangle = \sum A_{ij}B_{ij} = \mathrm{Tr}(A^\top B) \tag{3}$$

We say that $\widehat{\theta}$ is consistent if $\widehat{\theta}^\top\theta \to 1$, which is the same to say the angle between the two vectors goes to 0. From [Johnstone and Lu, 2009], we know that $\widehat{\theta}$ is consistent if and only if $\frac{p}{n} \to 0$. Now we are ready to do our analysis.

1 Analysis on Sparse PCA

1.1 Rough Analysis

We could write out the loss function

$$\begin{aligned} \left\|\widehat{\theta}\widehat{\theta}^\top - \theta\theta^\top\right\|_F^2 &= 2\left(1 - \left|\widehat{\theta}^\top\theta\right|^2\right) \\ &= 2\sin^2\angle\left(\widehat{\theta}, \theta\right) \end{aligned} \tag{4}$$

Also note the following relation

$$\begin{aligned} \left\langle \Sigma, \theta\theta^\top - \widehat{\theta}\widehat{\theta}^\top\right\rangle &= \left\langle \lambda\theta\theta^\top + I_p, \theta\theta^\top - \widehat{\theta}\widehat{\theta}^\top\right\rangle \\ &= \lambda\left\langle \theta\theta^\top, \theta\theta^\top - \widehat{\theta}\widehat{\theta}^\top\right\rangle + \left\langle I_p, \theta\theta^\top - \widehat{\theta}\widehat{\theta}^\top\right\rangle \\ &= \lambda\left(1 - \left|\widehat{\theta}^\top\theta\right|^2\right) \\ &= \frac{\lambda}{2}\left\|\widehat{\theta}\widehat{\theta}^\top - \theta\theta^\top\right\|_F^2 \end{aligned} \tag{5}$$

Also note that

$$\left\langle \widehat{\Sigma}, \theta\theta^\top - \widehat{\widehat{\theta}\widehat{\theta}}^\top \right\rangle \leq 0$$

due to the defintion of the estimator $\widehat{\theta}$ in 2. We now define

$$\Delta = \frac{\theta\theta^\top - \widehat{\widehat{\theta}\widehat{\theta}}^\top}{\left\| \widehat{\widehat{\theta}\widehat{\theta}}^\top - \theta\theta^\top \right\|_F} \tag{6}$$

We could observe some of Δ's property: the rank of Δ is not greater than 2 and that Δ is also sparse, with only $2s$ nonzero column and $2s$ nonzero rows. Continuing the analysis, we have that

$$\begin{aligned}
\left\langle \Sigma, \theta\theta^\top - \widehat{\widehat{\theta}\widehat{\theta}}^\top \right\rangle &= \left\langle \Sigma - \widehat{\Sigma}, \theta\theta^\top - \widehat{\widehat{\theta}\widehat{\theta}}^\top \right\rangle + \left\langle \widehat{\Sigma}, \theta\theta^\top - \widehat{\widehat{\theta}\widehat{\theta}}^\top \right\rangle \\
&\leq \left\langle \Sigma - \widehat{\Sigma}, \theta\theta^\top - \widehat{\widehat{\theta}\widehat{\theta}}^\top \right\rangle \\
&= \left\| \widehat{\widehat{\theta}\widehat{\theta}}^\top - \theta\theta^\top \right\|_F \cdot \left\langle \Sigma - \widehat{\Sigma}, \Delta \right\rangle \\
&\leq \left\| \widehat{\widehat{\theta}\widehat{\theta}}^\top - \theta\theta^\top \right\|_F \cdot \sup_{\substack{\|\Delta\|_F = 1, \Delta = \Delta^\top \\ \text{rank}(\Delta) \leq 2, |\text{rowsupp}(\Delta)| \leq 2s}} \left\langle \Sigma - \widehat{\Sigma}, \Delta \right\rangle
\end{aligned} \tag{7}$$

Now use equation 5, we have the following inequalities:

$$\left\| \widehat{\widehat{\theta}\widehat{\theta}}^\top - \theta\theta^\top \right\|_F \leq \frac{2}{\lambda} \sup_{\substack{\|\Delta\|_F = 1, \Delta = \Delta^\top \\ rk(\Delta) \leq 2, |\text{rowsupp}(\Delta)| \leq 2s}} \left\langle \Sigma - \widehat{\Sigma}, \Delta \right\rangle \tag{8}$$

Now let's analyze Δ. We know from matrix theory, for all Δ, $\|\Delta\|_F = 1$, $\Delta = \Delta^\top$, $\text{rank}(\Delta) \leq 2$, $|\text{rowsupp}(\Delta)| \leq 2s$, we could do the eigenvalue decomposition as follows

$$\Delta = d_1 v_1 v_1^\top + d_2 v_2 v_2^\top, \quad \|v_1\| = \|v_2\| = 1$$

Since the Frobenius norm of Δ is 1, we also know that $d_1^2 + d_2^2 = 1$. And a rough analysis could deduce that $v_1, v_2 \in \mathcal{V}(p, 2s)$. So now we could write that

$$\begin{aligned}
\left\langle \Sigma - \widehat{\Sigma}, \Delta \right\rangle &\leq \frac{2}{\lambda} \left\langle \Sigma - \widehat{\Sigma}, d_1 v_1 v_1^\top + d_2 v_2 v_2^\top \right\rangle \\
&\leq |d_1| \left| v_1^\top \left(\widehat{\Sigma} - \Sigma \right) v_1 \right| + |d_2| \left| v_2^\top \left(\widehat{\Sigma} - \Sigma \right) v_2 \right| \\
&\leq \left| v_1^\top \left(\widehat{\Sigma} - \Sigma \right) v_1 \right| + \left| v_2^\top \left(\widehat{\Sigma} - \Sigma \right) v_2 \right| \\
&\leq \sup_{v \in \mathcal{V}(p, 2s)} \left| v^\top \left(\widehat{\Sigma} - \Sigma \right) v \right|
\end{aligned} \tag{9}$$

And thus

$$\begin{aligned}
\left\| \widehat{\widehat{\theta}\widehat{\theta}}^\top - \theta\theta^\top \right\|_F &\leq \frac{4}{\lambda} \sup_{v \in \mathcal{V}(p, 2s)} \left| v^\top \left(\widehat{\Sigma} - \Sigma \right) v \right| \\
&\lesssim \frac{1}{\lambda} \|\Sigma\|_{op} \sqrt{\frac{s \log ep/s}{n}} \\
&= \frac{1 + \lambda}{\lambda} \sqrt{\frac{s \log ep/s}{n}}
\end{aligned} \tag{10}$$

So we've reached the final conclusion:

$$\left\|\widehat{\theta\theta}^\top - \theta\theta^\top\right\|_F^2 \lesssim \frac{\lambda^2+1}{\lambda^2}\frac{s\log ep/s}{n} \tag{11}$$

So observe that this upper bound has a natural interpretation: $\frac{s\log ep/s}{n}$ is just the error we constantly get, the denominato λ^2 is the square of the eigenvalue gap, aka the difference between the largest and the second largest eigenvalue, the numerator λ^2+1 is From the literature [Cai et al., 2013], we know that the optimal rate is actually

$$\inf_{\widehat{\theta}}\sup_{\theta\in\mathcal{V}(p,s)}\mathbf{E}\left\|\widehat{\theta\theta}^\top - \theta\theta^\top\right\|_F^2 \asymp \frac{\lambda+1}{\lambda^2}\frac{s\log ep/s}{n} \tag{12}$$

So we note that the analysis we just did is not as good as the optimal rate because λ might not be bounded. So either the estimator is not that good, or the analysis we just did is not sharp. And it turns out that our analysis is not sharp.

1.2 More Subtle Analysis

We already knew that $X_i \sim \mathcal{N}\left(0, \lambda\theta\theta^\top + I_p\right)$. So we could write $X_i = \sqrt{\lambda}w_i\theta + z_i$, where $w_i \sim \mathcal{N}(0,1)$ and $z_i \sim \mathcal{N}(0, I_p)$ and that $w \perp\!\!\!\perp z$ are independent variables. So we now could write the sample covariance matrix $\widehat{\Sigma}$ as follows

$$\begin{aligned}\frac{1}{n}\sum X_iX_i^\top = \lambda\left(\sum\frac{1}{n}w_i^2\right)\theta\theta^\top + \frac{\sqrt{\lambda}}{n}\sum w_i\theta z_i^\top \\ + \frac{\sqrt{\lambda}}{n}\sum w_iz_i\theta^\top + \frac{1}{n}\sum z_iz_i^\top\end{aligned} \tag{13}$$

Observing this equation, we note that the first part of RHS is the part that we would like to kill: it contributes the λ^2 in equation 12. In other words, that's the part that makes the bound in 1.1 not sharp. So naturally, now we plan to analyze the behavior of

$$\widetilde{\Sigma} = \lambda\left(\sum\frac{1}{n}w_i^2\right)\theta\theta^\top + I_p \tag{14}$$

And it is easy to mimic the proof we just did and have

$$\begin{aligned}\left\langle\widetilde{\Sigma}, \theta\theta^\top - \widehat{\theta\theta}^\top\right\rangle &= \frac{\lambda}{2}\left(\frac{1}{n}\sum w_i^2\right)\left\|\theta\theta^\top - \widehat{\theta\theta}^\top\right\|_F^2 \\ &\le \left\langle\widetilde{\Sigma}-\Sigma, \theta\theta^\top - \widehat{\theta\theta}^\top\right\rangle\end{aligned} \tag{15}$$

Using this relation, and mimicing the techiniques in equation 7 and 10, we could derive inequality as follows:

$$\left\|\theta\theta^\top - \widehat{\theta\theta}^\top\right\|_F \le \frac{4}{\lambda\left(\frac{1}{n}w_i^2\right)}\sup_{v\in\mathcal{V}(p,2s)}\left|v^\top\left(\widehat{\Sigma}-\widetilde{\Sigma}\right)v\right| \tag{16}$$

It is easy to prove that the term $\frac{1}{n}w_i^2$ in the denominator is bounded away from zero with high probability when n goes to infinity. Let's analyze the $\sup$ term.

$$\begin{aligned}\sup_{v\in\mathcal{V}(p,2s)}\left|v^\top\left(\widehat{\Sigma}-\widetilde{\Sigma}\right)v\right| \le & \sup_{v\in\mathcal{V}(p,2s)}\left|v^\top\left(\frac{1}{n}\sum z_iz_i^\top - I_p\right)v\right| \\ & + 2\sqrt{\lambda}\sup_{v\in\mathcal{V}(p,2s)}\left|v^\top\left(\frac{1}{n}\sum w_i\theta z_i^\top\right)v\right|\end{aligned} \tag{17}$$

From previous lectures, we know that the first term on the RHS could be bounded by

$$\sup_{v\in\mathcal{V}(p,2s)}\left|v^\top\left(\frac{1}{n}\sum z_iz_i^\top - I_p\right)v\right| \lesssim \sqrt{\frac{s\log ep/s}{n}} \tag{18}$$

As for the second term, note that we could write

$$\begin{aligned}\left|v^\top\left(\frac{1}{n}\sum w_i\theta z_i^\top\right)v\right| &= \left|\frac{1}{n}\sum w_i\left(v^\top\theta\right)\cdot\left(v^\top z_i\right)\right| \\ &\le \left|\frac{1}{n}\sum w_i\cdot\left(v^\top z_i\right)\right|\end{aligned}$$

The inequality is because θ is already normalized. Note that w_i and $v^\top z_i$ are independent one dimension normal variables, and their properties are easy to analyze. So we have the bound for the second term of the RHS of equation 17

$$2\sqrt{\lambda}\sup_{v\in\mathcal{V}(p,2s)}\left|v^\top\left(\frac{1}{n}\sum w_i\theta z_i^\top\right)v\right| \lesssim \sqrt{\lambda}\sqrt{\frac{s\log ep/s}{n}} \tag{19}$$

Plugging the estimates 18 and 19 back to 16, we finally reach the following theorem

Theorem 1.1. *Suppose* $\frac{s\log ep/s}{n} \lesssim 1$*, the nwe have the following bound with high probability*

$$\left\|\theta\theta^\top - \widehat{\theta}\widehat{\theta}^\top\right\|_F^2 \lesssim \frac{1+\lambda}{\lambda^2}\frac{s\log ep/s}{n} \tag{20}$$

From this theorem, we know that the estimator $\widehat{\theta} = \arg\min_{\theta\in\mathcal{V}(p,s)}\theta^\top\widehat{\Sigma}\theta$ achieves the minimax rate.

2 How to Compute Sparse PCA: Using Convex Relaxation

It turns out that the estimator 2 is hard to compute directly. Although note that we could use the power method iteration

$$\theta^{t+1} = \frac{\widehat{\Sigma}\theta^t}{\left\|\widehat{\Sigma}\theta^t\right\|} \tag{21}$$

to compute

$$\theta^* = \arg\min_{\|\theta\|=1} \theta^\top \widehat{\Sigma}\theta \tag{22}$$

This does not guarantee the sparsity. Our method here is first to think about

$$\theta^\top \widehat{\Sigma}\theta = \left\langle \widehat{\Sigma}, \theta\theta^\top \right\rangle \overset{F=\theta\theta^\top}{=} = \left\langle \widehat{\Sigma}, F \right\rangle$$

Note that we have to ensure that F is a rank 1 matrix (the sparsity could be controlled when adding ℓ_1 penalty). So now we consider the following two sets

$$\begin{aligned} \mathcal{P} &= \left\{ P = P^\top = P^2 : \text{rank}(P) = 1 \right\} \\ \mathcal{F} &= \text{conv}\left(\mathcal{P}\right) \end{aligned} \tag{23}$$

Actually we could explicitly write out $calF$ due to the following lemma

Lemma 2.1.

$$\mathcal{F} = \left\{ F = F^\top : 0 \preceq F \preceq I_p, \text{Tr}(F) = 1 \right\} \tag{24}$$

Proof. First, it is obvious that $\mathcal{F}$ is convex. Also need to show that $\mathcal{P} \subseteq \mathcal{F}$ which is also obviously true and that for all $F \in \mathcal{F}$, we could write F as a convex combination of matrices in $\mathcal{P}$. To prove that, we simply do eigenvalue decomposition on F and have

$$F = \sum d_j v_j v_j^\top$$

Due to the defintion of $\mathcal{F}$, we know that $0 \leq d_j \leq 1$ and that $\sum d_j = 1$, and thus F is a convex combination of rank 1 matrices. □

Note that $\mathcal{F}$ is often referred to as fantope. So now we could consider the following convex program

$$\begin{aligned} \max_{F \in \mathbb{R}^{p\times p}} \quad & \left\langle \widehat{\Sigma}, F \right\rangle - \rho \left\| F \right\|_1 \\ s.t. \quad & F \in \mathcal{F} \end{aligned} \tag{25}$$

where $\|F\|_1 = \sum_{ij} |F_{ij}|$. We also define $\|F\|_\infty = \max_{ij} |F_{ij}|$. The upper program 30 is first introduced in the paper [d'Aspremont et al., 2004], though it does not analyze its property. So as usual, we start from the basic inequality

$$\left\langle \widehat{\Sigma}, \widehat{F} \right\rangle - \rho \left\| \widehat{F} \right\|_1 \geq \left\langle \widehat{\Sigma}, F \right\rangle - \rho \left\| F \right\|_1 \tag{26}$$

Reordering the inequality gives

$$\left\langle \widehat{\Sigma}, F - \widehat{F} \right\rangle \leq \rho \left(\|F\|_1 - \left\| \widehat{F} \right\|_1 \right) \tag{27}$$

Plugging in $\widetilde{\Sigma}$

$$\left\langle \widetilde{\Sigma}, F - \widehat{F} \right\rangle \leq \rho \left(\|F\|_1 - \left\| \widehat{F} \right\|_1 \right) + \left\langle \widetilde{\Sigma} - \widehat{\Sigma}, F - \widehat{F} \right\rangle \tag{28}$$

Now let $\Delta = \widehat{F} - F$ and $S = \text{supp}(\theta)$. With slight abuse of notation, we also denote $\|\Delta_{SS}\|_1 = \sum_{(i,j)\in S\times S} |\Delta_{ij}|$ and that $\left\|\Delta_{(SS)^\complement}\right\|_1 = \sum_{(i,j)\in (S\times S)^\complement} |\Delta_{ij}|$. Now, suppose that $\left\|\widetilde{\Sigma} - \widehat{\Sigma}\right\|_\infty \leq \rho$ we have

$$\begin{aligned}
\left\langle \widetilde{\Sigma}, F - \widehat{F} \right\rangle &\leq \rho \left(\|F\|_1 - \left\|\widehat{F}\right\|_1 \right) + \left\|\widetilde{\Sigma} - \widehat{\Sigma}\right\|_\infty \cdot \left\|\widehat{F} - F\right\|_1 \\
&= \rho \|F_{SS}\|_1 - \rho \|F_{SS} + \Delta_{SS}\|_1 - \rho \left\|\Delta_{(SS)^\complement}\right\|_1 \\
&+ \left\|\widetilde{\Sigma} - \widehat{\Sigma}\right\|_\infty \cdot \left(\|\Delta_{SS}\|_1 + \left\|\Delta_{(SS)^\complement}\right\|_1 \right) \\
&\leq 2\rho \|\Delta_{SS}\|_1 \leq 2\rho\sqrt{s^2} \|\Delta\|_F
\end{aligned} \tag{29}$$

3 Analysis of the Convex Relaxation

Let's recall the convex relaxation problem we did in last lecture

$$\begin{aligned}
\max_{F\in\mathbb{R}^{p\times p}} \quad & \left\langle \widehat{\Sigma}, F \right\rangle - \rho \|F\|_1 \\
s.t. \quad & F \in \mathcal{F}
\end{aligned} \tag{30}$$

where $\|F\|_1 = \sum_{ij} |F_{ij}|$. We also define $\|F\|_\infty = \max_{ij} |F_{ij}|$. The upper program 30 is first introduced in the paper [d'Aspremont et al., 2004], though it does not analyze its property. So as usual, we start from the basic inequality

$$\left\langle \widehat{\Sigma}, \widehat{F} \right\rangle - \rho \left\|\widehat{F}\right\|_1 \geq \left\langle \widehat{\Sigma}, F \right\rangle - \rho \|F\|_1 \tag{31}$$

Reordering the inequality gives

$$\left\langle \widehat{\Sigma}, F - \widehat{F} \right\rangle \leq \rho \left(\|F\|_1 - \left\|\widehat{F}\right\|_1 \right) \tag{32}$$

Plugging in $\widetilde{\Sigma}$

$$\left\langle \widetilde{\Sigma}, F - \widehat{F} \right\rangle \leq \rho \left(\|F\|_1 - \left\|\widehat{F}\right\|_1 \right) + \left\langle \widetilde{\Sigma} - \widehat{\Sigma}, F - \widehat{F} \right\rangle \tag{33}$$

Now let $\Delta = \widehat{F} - F$ and $S = \text{supp}(\theta)$. With slight abuse of notation, we also denote $\|\Delta_{SS}\|_1 = \sum_{(i,j)\in S\times S} |\Delta_{ij}|$ and that $\left\|\Delta_{(SS)^\complement}\right\|_1 = \sum_{(i,j)\in (S\times S)^\complement} |\Delta_{ij}|$. Now, suppose that $\left\|\widetilde{\Sigma} - \widehat{\Sigma}\right\|_\infty \leq \rho$ we have

$$\begin{aligned}
\left\langle \widetilde{\Sigma}, F - \widehat{F} \right\rangle &\leq \rho \left(\|F\|_1 - \left\|\widehat{F}\right\|_1 \right) + \left\|\widetilde{\Sigma} - \widehat{\Sigma}\right\|_\infty \cdot \left\|\widehat{F} - F\right\|_1 \\
&= \rho \|F_{SS}\|_1 - \rho \|F_{SS} + \Delta_{SS}\|_1 - \rho \left\|\Delta_{(SS)^\complement}\right\|_1 \\
&+ \left\|\widetilde{\Sigma} - \widehat{\Sigma}\right\|_\infty \cdot \left(\|\Delta_{SS}\|_1 + \left\|\Delta_{(SS)^\complement}\right\|_1 \right) \\
&\leq 2\rho \|\Delta_{SS}\|_1 \leq 2\rho\sqrt{s^2} \|\Delta\|_F = 2\rho s \|\Delta\|_F
\end{aligned} \tag{34}$$

Thus, if $\left\|\widehat{\Sigma}-\widetilde{\Sigma}\right\|_{\infty} \leq \rho$, we have the following bound

$$\left\langle \widetilde{\Sigma}, F-\widehat{F} \right\rangle \leq 2\rho s \left\|\Delta\right\|_F \tag{35}$$

Also note that, from our construction of $\widetilde{\Sigma}$ in last lecture

$$\widetilde{\Sigma} = \lambda\left(\sum \frac{1}{n} w_i^2\right)\theta\theta^\top + I_p \tag{36}$$

we have

$$\begin{aligned} \left\langle \widetilde{\Sigma}, F-\widehat{F} \right\rangle &= \lambda\left(\frac{1}{n}\Sigma w_i^2\right)\left(\theta^\top F\theta - \theta^\top\widehat{F}\theta\right) + \left\langle I_p, F-\widehat{F}\right\rangle \\ &= \lambda\left(\frac{1}{n}\Sigma w_i^2\right)\left(1-\theta^\top\widehat{F}\theta\right) \end{aligned} \tag{37}$$

Now let's look at the loss function

$$\begin{aligned} \left\|\widehat{F}-F\right\|_F^2 &= \left\|\widehat{F}\right\|_F^2 + \|F\|_F^2 - 2\left\langle \widehat{F}, F\right\rangle \\ &= \left\|\widehat{F}\right\|_F^2 + 1 - 2\theta^\top\widehat{F}\theta \\ &\leq 2\left(1-\theta^\top\widehat{F}\theta\right) \end{aligned} \tag{38}$$

The inequality is because of the following evaluation

$$\left\|\widehat{F}\right\|_F^2 \leq \left\|\widehat{F}\right\|_N \cdot \left\|\widehat{F}\right\|_{op} = \mathrm{Tr}\left(\widehat{F}\right)\cdot\left\|\widehat{F}\right\|_{op} \leq 1 \tag{39}$$

by definition. Here $\|\cdot\|_N$ is the nuclear norm, i.e., the sum of singular values. And so now we have the bound of the loss function, by combining eqs. (35), (37) and (38)

$$\begin{aligned} \left\|\widehat{F}-F\right\|_F^2 &\leq \frac{2}{\lambda\left(\frac{1}{n}\sum w_i^2\right)}\left\langle \widetilde{\Sigma}, F-\widehat{F}\right\rangle \\ &\leq 2\rho s\left\|\Delta\right\|_F \cdot \frac{2}{\lambda\left(\frac{1}{n}\sum w_i^2\right)} \end{aligned} \tag{40}$$

Since $\Delta = \widehat{F} - F$, we could further simplify it as

$$\left\|\widehat{F}-F\right\|_F^2 \leq \frac{16\rho^2 s^2}{\lambda^2\left(\frac{1}{n}\sum w_i^2\right)^2} \tag{41}$$

Now let's see what ρ should we choose such that $\left\|\widehat{\Sigma}-\widetilde{\Sigma}\right\|_{\infty} \leq \rho$ holds w.h.p.. And this analysis is quite similar to the one we did in last lecture. First, we surely have

$$\begin{aligned} \widehat{\Sigma} = \frac{1}{n}\sum X_i X_i^\top &= \lambda\left(\sum\frac{1}{n}w_i^2\right)\theta\theta^\top + \frac{\sqrt{\lambda}}{n}\sum w_i\theta z_i^\top \\ &+ \frac{\sqrt{\lambda}}{n}\sum w_i z_i\theta^\top + \frac{1}{n}\sum z_i z_i^\top \end{aligned} \tag{42}$$

$$\widetilde{\Sigma} = \lambda \left(\sum \frac{1}{n} w_i^2 \right) \theta\theta^\top + I_p \tag{43}$$

And thus

$$\left\| \widehat{\Sigma} - \widetilde{\Sigma} \right\|_\infty \leq \left\| \frac{1}{n} \sum z_i z_i^\top - I_p \right\|_\infty + 2\sqrt{\lambda} \left\| \frac{1}{n} \sum w_i \theta z_i^\top \right\|_\infty \tag{44}$$

The first term is surely bounded by

$$\left\| \frac{1}{n} \sum z_i z_i^\top - I_p \right\|_\infty \lesssim \sqrt{\frac{\log p}{n}} \tag{45}$$

And for the second term in the RHS

$$\begin{aligned} \left\| \frac{1}{n} \sum w_i \theta z_i^\top \right\|_\infty &= \max_{j,k} \left| \left(\frac{1}{n} \sum w_i z_{ij} \right) \theta_k \right| \\ &\leq \max_j \left\| \frac{1}{n} \sum w_i z_{ij} \right\| \\ &\leq \sqrt{\frac{\log p}{n}} \end{aligned} \tag{46}$$

Note that θ is a normalized vector and that $w_i \perp\!\!\!\perp z_{ij} \sim \mathcal{N}(0,1)$. Combining these to estimation 44 and assuming that $\log p/n \lesssim 1$, we have the following estimate w.h.p.

$$\left\| \widehat{\Sigma} - \widetilde{\Sigma} \right\|_\infty \lesssim \left(1 + \sqrt{\lambda} \right) \sqrt{\frac{\log p}{n}} \tag{47}$$

Also as always

$$\frac{1}{n} \sum w_i^2 \geq \frac{1}{2} \tag{48}$$

holds w.h.p.. Now we could state the theorem.

Theorem 3.1. *Assume that* $\log p/p \lesssim 1$*, and choose* $\rho = C \left(1 + \sqrt{\lambda} \right) \sqrt{\frac{\log p}{n}}$*, then w.h.p.*

$$\left\| \widehat{F} - F \right\|_F^2 \leq \frac{1+\lambda}{\lambda^2} \frac{s^2 \log p}{n} \tag{49}$$

Recall that the minimax rate from last lecture is

$$\left\| \theta\theta^\top - \widehat{\theta}\widehat{\theta}^\top \right\|_F^2 \lesssim \frac{1+\lambda}{\lambda^2} \frac{s \log ep/s}{n} \tag{50}$$

So there is an extra s in the theorem. So we start to wonder is the estimator not so good or the analysis not sharp. And we could observe that the objective of the convex program is to maximize $\left\langle \widehat{\Sigma}, F \right\rangle - \rho \left\| F \right\|_1$, where $\left\| F \right\|_1 = \sum_{ij} |F_{ij}|$. So the sparsity structure is changed when we regard a

rank 1 matrix in $\mathbb{R}^{p\times p}$ as a vector of length p^2. An improvement in paper [Ma, 2013] is motivated by the power iteration we mentioned in the last lecture, and is presented as

$$\theta^t = \frac{\mathcal{T}\left(\widehat{\Sigma}\theta^{t-1}\right)}{\left\|\mathcal{T}\left(\widehat{\Sigma}\theta^{t-1}\right)\right\|} \tag{51}$$

where $\mathcal{T}$ is the thresholding operator. The paper also proved that if $|\sin\angle(\theta^0,\theta)| \le c < 1$, so the initial angle is not orthogonal to the truth, then the algorithm achieves the minimax rate. This sounds nice, but finding nontrivial angle in high-dimensional space is hard and requires exponential time if we do random sampling. However, we could smartly set the initial angle as the leading eigenvector of $\widehat{F}$ we just constructed, i.e.

$$\theta^0 = \arg\max_{\theta} \theta^\top \widehat{F}\theta \tag{52}$$

This method is doable due to the *Sin-Theta Theorem* or *Davis-Kahan Theorem* we are going to prove. Also from the paper [Berthet and Rigollet, 2013], we know that if $n \not\gtrsim s^2$, then achieving the minimax rate of PCA is a NPC problem, in other words $n \gtrsim s^2$ is necessary for a polynomial time algorithm. This result will probably be discussed next week. For now, let's prove the Sin-Theta theorem.

Proposition 3.2. *Suppose that A and B are PSD and their EVD could be written as follows*

$$A = \sum d_j u_j u_j^\top, \quad B = \sum \lambda_j v_j v_j^\top \tag{53}$$

And that $d_1 \ge d_2 \ge \cdots \ge d_p \ge 0$, $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$, then we have the following estimate

$$\left\|u_1u_1^\top - v_1v_1^\top\right\|_F^2 \le \frac{4\|A-B\|_F^2}{(d_1-d_2)^2} \tag{54}$$

Proof. First, note that

$$\begin{aligned}\left\langle A, u_1u_1^\top - v_1v_1^\top\right\rangle &= d_1\left\langle u_1u_1^\top, u_1u_1^\top - v_1v_1^\top\right\rangle + \sum_{j\ge 2} d_j\left\langle u_ju_j^\top, u_1u_1^\top - v_1v_1^\top\right\rangle \\ &\ge d_1\left\langle u_1u_1^\top, u_1u_1^\top - v_1v_1^\top\right\rangle + d_2\sum_{j\ge 2}\left\langle u_ju_j^\top, -v_1v_1^\top\right\rangle\end{aligned} \tag{55}$$

Now note that the identity matrix could be decompose as

$$I_p = \sum_{j\in[p]} u_ju_j^\top = u_1u_1^\top + \sum_{j\ne 2} u_ju_j^\top \tag{56}$$

Thus, we have

$$\begin{aligned}\left\langle A, u_1u_1^\top - v_1v_1^\top\right\rangle &\ge d_1\left\langle u_1u_1^\top, u_1u_1^\top - v_1v_1^\top\right\rangle + d_2\left\langle I_p - u_1u_1^\top, -v_1v_1^\top\right\rangle \\ &= (d_1-d_2)\left(1 - \left|u_1^\top v_1\right|^2\right) \\ &= \frac{d_1-d_2}{2}\left\|u_1u_1^\top - v_1v_1^\top\right\|_F^2\end{aligned} \tag{57}$$

Now note that since v_1 is the leading eigenvector of B, we have

$$\left\langle B, u_1 u_1^\top - v_1 v_1^\top \right\rangle = u_1^\top B u_1 - v_1^\top B v_1 \leq 0 \tag{58}$$

And thus the following inequality holds

$$\begin{aligned} \left\langle A, u_1 u_1^\top - v_1 v_1^\top \right\rangle &= \left\langle A - B, u_1 u_1^\top - v_1 v_1^\top \right\rangle + \left\langle B, u_1 u_1^\top - v_1 v_1^\top \right\rangle \\ &\leq \left\langle A - B, u_1 u_1^\top - v_1 v_1^\top \right\rangle \\ &\leq \|A - B\|_F \left\| u_1 u_1^\top - v_1 v_1^\top \right\|_F \end{aligned} \tag{59}$$

Combining eqs. (57) and (59), we reach the conclusion

$$\left\| u_1 u_1^\top - v_1 v_1^\top \right\|_F^2 \leq \frac{4 \|A - B\|_F^2}{(d_1 - d_2)^2} \tag{60}$$

□

Suppose θ^0 is the leading eigenvector of $\widehat{F}$

$$\theta^0 = \arg\max_\theta \theta^\top \widehat{F} \theta \tag{61}$$

We now could use the Sin-Theta theorem 3.2 to conclude that

$$\left\| \theta^0 \left(\theta^0\right)^\top - \theta \theta^\top \right\|_F^2 \leq 4 \left\| \widehat{F} - F \right\|_F^2 \tag{62}$$

since here for $F = \theta\theta^\top$, $d_1 = 1$ and $d_2 = 0$. Another two useful inequalities eigen-space perturbation theory are Weyl inequality and Wedin inequality.

References

[Berthet and Rigollet, 2013] Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):17801815.

[Cai et al., 2013] Cai, T. T., Ma, Z., and Wu, Y. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):30743110.

[d'Aspremont et al., 2004] d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., and Lanckriet, G. R. G. (2004). A direct formulation for sparse pca using semidefinite programming.

[Johnstone and Lu, 2009] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693. PMID: 20617121.

[Ma, 2013] Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772801.