Statistics 210B Lecture 22 Notes

Daniel Raban

April 12, 2022

1 Principle Component Analysis for Spiked and Sparse Ensembles

1.1 Recap: estimation error bound for principle component analysis

In high-dimensional principal component analysis, we observe $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} X \in \mathbb{R}^d$, where $\mathbb{E}[X] = 0$ and $\text{Cov}(X) = \Sigma \in \mathbb{R}^{n \times d}$. We have the empirical covariance matrix

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top.$$

The ground truth is

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

while our estimator is

$$\widehat{\theta} = \underset{\|\theta\|_2 = 1}{\arg\max} \langle \theta, \widehat{\Sigma} \theta \rangle.$$

We want to upper bound the estimation error $\|\widehat{\theta} - \theta^*\|_2$.

Last time, he had the following theorem:

Theorem 1.1. *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 $\|P\|_{\text{op}} < \nu/2$, and let $\widehat{\Sigma} = \Sigma + P$. If $\widehat{\theta} \in \mathbb{R}^d$ is an eigenvector for $\lambda_1(\widehat{\Sigma})$, then*

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

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

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

1.2 Consequence for a spiked ensemble

In the spiked covariance model, introduced by Jonstone in 2001, we estimate $\theta^* \in \mathbb{R}^d$ with $\|\theta^*\|_2 = 1$. We observe $x_i = \sqrt{\nu}\xi_i\theta^* + w_i$, where

$$\xi_i \in \mathbb{R}, \qquad \mathbb{E}[\xi_i] = 0, \qquad \mathbb{E}[\xi_i^2] = 1,$$
$$w_i \in \mathbb{R}^d \, \mathbb{E}[w_i] = 0, \qquad \mathbb{E}[w_i w_i^\top] = I_d.$$

The w_i and ξ_i are independent. If we calculate the covariance structure of x_i, we have

$$\begin{aligned}\mathbb{E}[x_i x_i^\top] &= \mathbb{E}(\sqrt{\nu}\xi_i\theta^* + w_i)(\sqrt{\nu}\xi_i\theta^* + w_i^\top)] \\ &= \nu\theta * (\theta^*)^\top + I_d.\end{aligned}$$

This is Σ. The largest eigenvalue is $\lambda_{\max}(\Sigma) = \nu + 1$. The second largest eigenvalue is $\lambda_2(\Sigma)$. So $\nu = \lambda_{\max}(\Sigma) - \lambda_2(\Sigma)$ is the eigengap, and the leading aigenvector of Σ is θ^*. We estimate θ by

$$\widehat{\theta} = \underset{\|\theta\|_2=1}{\arg\max} \langle \theta, \Sigma\theta \rangle.$$

Our theorem gives us the following bound on $\|\widehat{\theta} - \theta^*\|_2$.

Corollary 1.1. *Assume $\xi \sim \mathrm{sG}(1)$ 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*

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

with high probability.

If you want this to be $\leq \varepsilon$, you need $n \gtrsim d\frac{\nu+1}{\nu^2}$. For large ν, $\|\widehat{\theta} - \theta^*\|_2 \sim \frac{1}{\sqrt{\nu}}$.

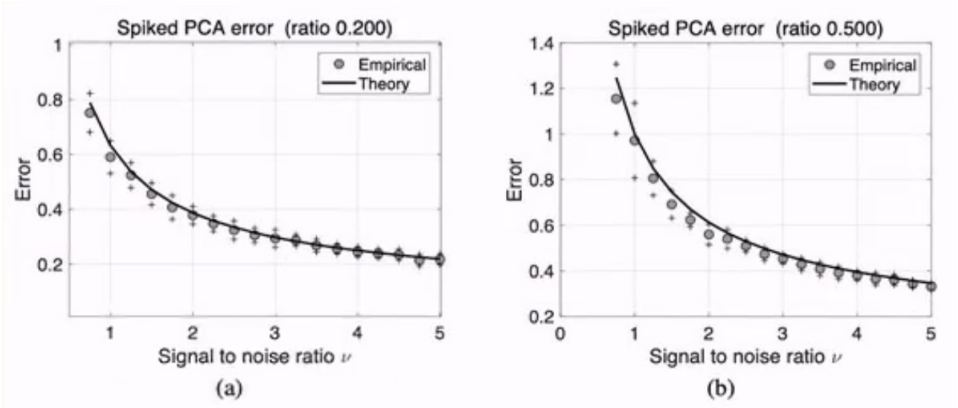

Figure 8.4 Plots of the error $\|\widehat{\theta} - \theta^*\|_2$ versus the signal-to-noise ratio, as measured by the eigengap ν. Both plots are based on a sample size $n = 500$. Dots show the average of 100 trials, along with the standard errors (crosses). The full curve shows the theoretical bound $\sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}}$. (a) Dimension $d = 100$. (b) Dimension $d = 250$.

Proof. Recall that the theorem says that $\|\widehat{\theta} - \theta\|_2 \le \frac{2\|\widetilde{P}\|_2}{\nu - 2\|P\|_{\text{op}}}$. We need to upper bound $\|\widetilde{P}\|_2$ and $\|P\|_{\text{op}}$.

$$\begin{aligned} P &= \widehat{\Sigma} - \Sigma \\ &= \frac{1}{n}\sum_{i=1}^{n}(\sqrt{\nu}\xi\theta^* + w_i)(\sqrt{\nu}\xi_i\theta^* + w_i)^\top - (\nu\theta^*(\theta^*)^\top + I_d) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\xi_i^2 - 1\right)\nu\theta^*(\theta^*)^\top + \left(\frac{1}{n}\sum_{i=1}^{n} w_i w_i^\top - I_d\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\xi_i w_i^\top\right)(\theta^*)^\top + \text{transpose}. \end{aligned}$$

So we get

$$\|P\|_{\text{op}} \le \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i^2 - 1\right|}_{a}\nu + \underbrace{\left\|\frac{1}{n}\sum_{i=1}^{n} w_i w_i^\top - I_d\right\|_{\text{op}}}_{c} + 2\sqrt{\nu}\underbrace{\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i w_i\right\|_2}_{b}.$$

We can also bound

$$\|\widetilde{P}\|_2 \le \sqrt{\nu}\underbrace{\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i w_i\right\|_2}_{b} + \underbrace{\left\|\frac{1}{n}\sum_{i=1}^{n} w_i w_i^\top - I_d\right\|_{\text{op}}}_{c},$$

so we just need to bound a, b, c.

By sub-exponential concentration, $a \lesssim \sqrt{\frac{1}{n}}$. The term c is a random matrix with mean 0, and using a metric entropy argument with matrix concentration gives $c \lesssim \sqrt{\frac{d}{n}}$. Similarly, we can show that $b \lesssim \sqrt{\frac{d}{n}}$. Given these upper bounds, we get

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

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

So if $\sqrt{\frac{d}{n}} \lesssim \frac{\nu}{\sqrt{\nu}+1}$, then $\nu - 2\|P\|_{\text{op}} \ge \frac{\nu}{2}$. This gives the bound

$$\|\widehat{\theta} - \theta^*\|_2 \lesssim \frac{2\|\widetilde{P}\|_2}{\nu/2} \lesssim \sqrt{\frac{\nu+1}{\nu^2}}\sqrt{\frac{d}{n}}.$$

Here, we give an example of how to use the metric entropy bound for the term b.

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

$$= \sup_{\|\nu\|_2=1} \frac{1}{n} \sum_{i=1}^n \underbrace{\underbrace{\varepsilon_i}_{\mathrm{sG}(1)} \underbrace{\langle w_i, \nu \rangle}_{\mathrm{sG}(1)}}_{\mathrm{sE}(1,1)}.$$

This tells us that

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

Now let $\Omega_{1/4}$ be a 1/4-cover of S^{d-1}, so $|\Omega_{1/4}| \leq C^d$ for a constant C. Show that tis implies

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

So we can use a union bound with

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

□

1.3 Sparse principle component analysis

This is an active research direction. It has been well-studied, but there are some important properties that are not well-understood. We assume that $\theta = \arg\max_{\|\theta\|_2=1} \langle \theta, \Sigma\theta \rangle$ is s-sparse, where $s \ll n \ll d$.

In the sparse spiked covariance model, $\theta^* \in \mathbb{R}^d$, $\|\theta^*\|_2 = 1$, and $|S(\theta^*)| \lesssim s$. We observe

$$x_i = \sqrt{\nu} \xi_i \theta^* + w_i, \qquad i \in [n],$$

where $\xi_i \, \mathrm{sG}(1)$ and $w_i \sim \mathrm{sG}(1)$. We have two theoretical questions:

(a) What should the sample size be to get a consistent estimator? We will see that as long as $n \gg s$, there is a consistent estimator.

(b) What is the sample size for a computationally efficient (polynomial time) consistent estimator? The best known computationally efficient estimator has $n \gg s^2$.

(c) What happens for $s \ll n \ll s^2$? This is an active research direction. It is conjectured that there exists a computational and statistical gap.

1.3.1 ℓ_1-penalized estimation

To answer part (a), we solve the estimation problem with an added ℓ_1 penalty.

- The 1-norm constrained formulation is

$$\widehat{\theta} = \arg\max_{\substack{\|\theta\|_2=1 \\ \|\theta\|_1 \le R}} \langle \theta, \widehat{\Sigma}\theta \rangle.$$

- The λ-penalized formulation is

$$\widehat{\theta} = \arg\max_{\|\theta\|_2=1} \langle \theta, \widehat{\Sigma}\theta \rangle - \lambda_n \|\theta\|_1.$$

  In this formulation, we need $\|\theta\|_1 \le (\frac{n}{\log d})^{1/4}$ for theoretical analysis.

Theorem 1.2. *Assume* $n \gtrsim s \log d . \min\{1, \frac{\nu^2}{\nu+1}\}$. *Take* $\lambda_n \asymp \sqrt{\nu+1}\sqrt{\frac{\log d}{n}}$. *Then*

$$\|\widehat{\theta} - \theta^*\|_2 \lesssim \sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{s \log d}{n}}.$$

So the required sample size is $\gtrsim s \log d$.

Proof. Here are the steps:

1. Use a basic inequality from the zero order optimality condition to derive a deterministic upper bound of $\|\widehat{\theta} - \theta^*\|_2$ by assuming a deterministic assumption on X. This is like imposing the RE condition for LASSO.
2. Prove a concentration inequality and plug in the bound. □

1.3.2 The semidefinite programing relaxation estimator

The 1-norm constrained formulation

$$\max_{\substack{\|\theta\|_2=1 \\ \|\theta\|_1 \le R}} \langle \theta, \widehat{\Sigma}\theta \rangle$$

is equivalent, by a change of variable $\Theta = \theta\theta^\top \in \mathbb{R}^{d \times d}$ to

$$\max_{\substack{\mathrm{tr}(\Theta)=1 \\ \sum_{j,k} |\Theta_{j,k}| \le R^2 \\ \mathrm{rank}(\Theta)=1}} \langle \widehat{\Sigma}, \Theta \rangle.$$

The only nonconvex constraint is the rank constraint. If we drop the rank constraint, then the optimization problem becomes convex.

Theorem 1.3 (Amini, Wainwright, 2008)**.** *If* $n \gg s^2 \log d$, *then the semidefinite programing solution has rank 1 and is consistent.*

1.3.3 The $s \ll n \ll s^2$ regime

What do we know in this regime?

Theorem 1.4 (Berthet, Rigollet, 2013)**.** *For $s \ll n \ll s^2$, sparse PCA is computationally harder or equivalent to the* ***planted clique problem*** *in the hard regime.*

It is conjectured that no polynomial time algorithm can solve this problem.

1.4 Extra topics we will not cover

This completes our discussion of the material in chapter 7 and 8 of Wainwright's book. We will not cover chapters 9, 10, or 11, which generalize the material in chapters 7 and 8. Some topics these chapters discuss are

- Logistic LASSO
- Phase retrieval (used in imaging science)
- Matrix sensing
- Matrix completion (used in recommendation systems)

Example 1.1. As an example, we will explain matrix completion. We want to estimate $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, where $\Theta^* = UV^\top$, $U \in \mathbb{R}^{d_1 \times r}$, $V \in \mathbb{R}^{d_2 \times r}$, and $r \ll \min\{d_1, d_2\}$. We can, for example, think of $\Theta_{i,j}$ as the score of user i given to movie j. Then U_i is user i's feature, and V_j is movie j's feature.

We observe $\{M_{i,j} = \Theta^*_{i,j} + \varepsilon_{i,j}\}_{(i,j)\in\Omega}$, and we want to estimate $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$. How many samples is required?

The MLE estimator is

$$\min_{\text{rank}(\Theta)\leq r} \|M_{i,j} - \Theta_{i,j}\|_2^2.$$

This rank constraint is not convex, so we can relax it to a constraint $\|\Theta\|_* \leq r$ on the nuclear norm.