

# Lecture 13: Covariance estimation and matrix completion



School of Mathematical Sciences, Xiamen University

## 1. Covariance estimation

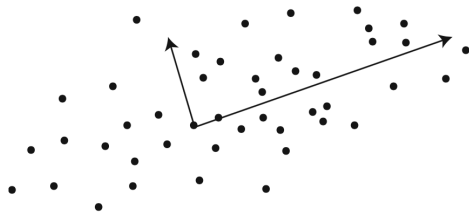
- Suppose we have a sample of data points  $X_1, \dots, X_N$  in  $\mathbb{R}^n$ . It is often reasonable to assume that these points are independently sampled from the same probability distribution (or “population”) which is unknown. We would like to learn something useful about this distribution.
- Denote by  $X$  a random vector with this (unknown) distribution. The most basic parameter of the distribution is the mean  $\mathbb{E}X$ . One can estimate  $\mathbb{E}X$  from the sample by computing the sample mean  $\sum_{i=1}^N X_i/N$ . The law of large numbers guarantees that the estimate becomes tight as the sample size  $N$  grows to infinity. In other words,

$$\frac{1}{N} \sum_{i=1}^N X_i \rightarrow \mathbb{E}X \quad \text{as } N \rightarrow \infty.$$

- The next most basic parameter of the distribution is the covariance matrix

$$\Sigma := \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^\top.$$

The eigenvectors of the covariance matrix  $\Sigma$  are called the principal components. Principal components that correspond to large eigenvalues of  $\Sigma$  are the directions in which the distribution of  $X$  is most extended, see the figure.



- This method is called Principal Component Analysis (PCA).

- One can estimate the covariance matrix  $\Sigma$  from the sample by computing the sample covariance

$$\Sigma_N := \frac{1}{N} \sum_{i=1}^N (X_i - \mathbb{E}X_i)(X_i - \mathbb{E}X_i)^\top.$$

Again, the law of large numbers guarantees that the estimate becomes tight as the sample size  $N$  grows to infinity, i.e.

$$\Sigma_N \rightarrow \Sigma \quad \text{as} \quad N \rightarrow \infty.$$

But how large should the sample size  $N$  be for covariance estimation? We are going to show that

$$N \sim n \log n$$

is enough. In other words, covariance estimation is possible with just logarithmic oversampling.

- For simplicity, we shall state the covariance estimation bound for **mean zero** distributions. (If the mean is not zero, we can estimate it from the sample and subtract. The mean can be accurately estimated from a sample of size  $N = \mathcal{O}(n)$ .)

### Theorem 1 (Covariance estimation)

Let  $X$  be a random vector in  $\mathbb{R}^n$  with covariance matrix  $\Sigma$ . Suppose that

$$\|X\|_2^2 \lesssim \mathbb{E}\|X\|_2^2 = \text{tr}\Sigma \quad \text{almost surely.}$$

Then, for every  $N \geq 1$ , we have

$$\mathbb{E} \|\Sigma_N - \Sigma\| \lesssim \|\Sigma\| \left( \sqrt{\frac{n \log n}{N}} + \frac{n \log n}{N} \right).$$

- Remark:  $N \sim \varepsilon^{-2} n \log n$  with  $\varepsilon \in (0, 1)$  guarantees a good relative error.

*Proof.* Apply matrix Bernstein's inequality (Corollary 3 of Lecture 12) for the sum of independent random matrices  $X_i X_i^\top - \Sigma$  and get

$$\begin{aligned}\mathbb{E} \|\Sigma_N - \Sigma\| &= \frac{1}{N} \mathbb{E} \left\| \sum_{i=1}^N (X_i X_i^\top - \Sigma) \right\| \\ &\lesssim \frac{1}{N} (\sigma \sqrt{\log n} + K \log n)\end{aligned}$$

where

$$\sigma^2 = \left\| \sum_{i=1}^N \mathbb{E} (X_i X_i^\top - \Sigma)^2 \right\| = N \left\| \mathbb{E} (X X^\top - \Sigma)^2 \right\|$$

and  $K$  is chosen so that

$$\|X X^\top - \Sigma\| \leq K \quad \text{almost surely.}$$

It remains to bound  $\sigma$  and  $K$ . Let us start with  $\sigma$ . We have

$$\begin{aligned}\mathbb{E}(XX^\top - \Sigma)^2 &= \mathbb{E}\|X\|_2^2 XX^\top - \Sigma^2 \\ &\lesssim \text{tr}(\Sigma) \cdot \mathbb{E}XX^\top \\ &= \text{tr}(\Sigma) \cdot \Sigma.\end{aligned}$$

Thus,  $\sigma^2 \lesssim N\text{tr}(\Sigma)\|\Sigma\|$ . Next, to bound  $K$ , we have

$$\begin{aligned}\|XX^\top - \Sigma\| &\leq \|X\|_2^2 + \|\Sigma\| \\ &\lesssim \text{tr}(\Sigma) + \|\Sigma\| \\ &\leq 2\text{tr}(\Sigma) =: K.\end{aligned}$$

Therefore,

$$\mathbb{E}\|\Sigma_N - \Sigma\| \lesssim \frac{1}{N}(\sqrt{N\text{tr}(\Sigma)\|\Sigma\|\log n} + \text{tr}(\Sigma)\log n).$$

The proof is completed by using  $\text{tr}(\Sigma) \leq n\|\Sigma\|$ . □

## 1.1 Low-dimensional distributions

- Far fewer samples are needed for covariance estimation for low-dimensional, or approximately low-dimensional, distributions. To measure approximate low-dimensionality we can use the notion of the stable rank of  $\Sigma^2$ . The stable rank of a matrix  $\mathbf{A}$  is defined as the square of the ratio of the Frobenius to operator norms:

$$r(\mathbf{A}) := \frac{\|\mathbf{A}\|_{\text{F}}^2}{\|\mathbf{A}\|^2} \leq \text{rank}(\mathbf{A}).$$

The proof of Theorem 1 yields

$$\mathbb{E} \|\Sigma_N - \Sigma\| \leq \|\Sigma\| \left( \sqrt{\frac{r \log n}{N}} + \frac{r \log n}{N} \right)$$

where  $r = r(\Sigma^{1/2}) = \text{tr}(\Sigma)/\|\Sigma\|$ . Therefore, covariance estimation is possible with  $N \sim r \log n$  samples.



## 2. Norms of random matrices

- Let  $\mathbf{A}_{i,:}$  denote the  $i$ th row of  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , we have

$$\max_i \|\mathbf{A}_{i,:}\|_2 \leq \|\mathbf{A}\| \leq \sqrt{n} \max_i \|\mathbf{A}_{i,:}\|_2.$$

For random matrices with independent entries the bound can be improved to the point where the upper and lower bounds almost match.

### Theorem 2 (Norms of random matrices without boundedness assumptions)

*Let  $\mathbf{A}$  be an  $n \times n$  symmetric random matrix whose entries on and above the diagonal are independent, mean zero random variables. Then*

$$\mathbb{E} \max_i \|\mathbf{A}_{i,:}\|_2 \leq \mathbb{E} \|\mathbf{A}\| \leq C \log n \cdot \mathbb{E} \max_i \|\mathbf{A}_{i,:}\|_2,$$

*where  $\mathbf{A}_{i,:}$  denote the rows of  $\mathbf{A}$ .*

### Lemma 3 (Symmetrization)

Let  $X_1, \dots, X_N$  be independent, mean zero random vectors in a normed space and  $\varepsilon_1, \dots, \varepsilon_N$  be independent Rademacher random variables.

Then

$$\frac{1}{2} \mathbb{E} \left\| \sum_{i=1}^N \varepsilon_i X_i \right\| \leq \mathbb{E} \left\| \sum_{i=1}^N X_i \right\| \leq 2 \mathbb{E} \left\| \sum_{i=1}^N \varepsilon_i X_i \right\|$$

*Proof.* To prove the upper bound, let  $(X'_i)$  be an independent copy of the random vectors  $(X_i)$ , i.e. just different random vectors with the same joint distribution as  $(X_i)$  and independent from  $(X_i)$ . Then

$$\begin{aligned} \mathbb{E} \left\| \sum_i X_i \right\| &= \mathbb{E} \left\| \sum_i X_i - \mathbb{E} \left( \sum_i X'_i \right) \right\| \\ &\stackrel{(\text{Jensen})}{\leq} \mathbb{E} \left\| \sum_i X_i - \sum_i X'_i \right\| = \mathbb{E} \left\| \sum_i (X_i - X'_i) \right\|. \end{aligned}$$

The distribution of the random vectors  $Y_i := X_i - X'_i$  is symmetric, which means that the distributions of  $Y_i$  and  $-Y_i$  are the same. (Why?) Thus the distribution of the random vectors  $Y_i$  and  $\varepsilon_i Y_i$  is also the same, for all we do is change the signs of these vectors at random and independently of the values of the vectors. Summarizing, we can replace  $X_i - X'_i$  in the sum above with  $\varepsilon_i(X_i - X'_i)$ . Thus

$$\begin{aligned}\mathbb{E} \left\| \sum_i X_i \right\| &\leq \mathbb{E} \left\| \sum_i \varepsilon_i (X_i - X'_i) \right\| \\ &\leq \mathbb{E} \left\| \sum_i \varepsilon_i X_i \right\| + \mathbb{E} \left\| \sum_i \varepsilon_i X'_i \right\| \\ &= 2\mathbb{E} \left\| \sum_i \varepsilon_i X_i \right\|\end{aligned}$$

This proves the upper bound in the symmetrization inequality. The lower bound can be proved by a similar argument. (Do this!) □

*Proof of Theorem 2.* The lower bound is trivial. The proof of the upper bound will be based on matrix Bernstein's inequality.

We represent  $\mathbf{A}$  as a sum of independent, mean zero, symmetric random matrices  $\mathbf{Z}_{ij}$  each of which contains a pair of symmetric entries of  $\mathbf{A}$  (or one diagonal entry):

$$\mathbf{A} = \sum_{i \leq j} \mathbf{Z}_{ij}.$$

By the symmetrization inequality (Lemma 3) for the random matrices  $\mathbf{Z}_{ij}$ , we get

$$\mathbb{E} \|\mathbf{A}\| = \mathbb{E} \left\| \sum_{i \leq j} \mathbf{Z}_{ij} \right\| \leq 2 \mathbb{E} \left\| \sum_{i \leq j} \mathbf{X}_{ij} \right\|$$

where we set  $\mathbf{X}_{ij} := \varepsilon_{ij} \mathbf{Z}_{ij}$  and  $\varepsilon_{ij}$  are independent Rademacher random variables. Now we condition on  $\mathbf{A}$ . The random variables  $\mathbf{Z}_{ij}$  become fixed values and all randomness remains in the Rademacher random variables  $\varepsilon_{ij}$ .

Note that  $\mathbf{X}_{ij}$  are (conditionally) bounded almost surely, and this is exactly what we have lacked to apply matrix Bernstein's inequality. Now we can do it. The corollary of matrix Bernstein's inequality gives

$$\mathbb{E}_\varepsilon \left\| \sum_{i \leq j} \mathbf{X}_{ij} \right\| \lesssim \sigma \sqrt{\log n} + K \log n,$$

where  $\sigma^2 = \left\| \sum_{i \leq j} \mathbb{E}_\varepsilon \mathbf{X}_{ij}^2 \right\|$  and  $K = \max_{i \leq j} \|\mathbf{X}_{ij}\|$ . A good exercise is to check that

$$\sigma \lesssim \max_i \|\mathbf{A}_{i,:}\|_2 \quad \text{and} \quad K \lesssim \max_i \|\mathbf{A}_{i,:}\|_2.$$

Then we have

$$\mathbb{E}_\varepsilon \left\| \sum_{i \leq j} \mathbf{X}_{ij} \right\| \lesssim \log n \cdot \max_i \|\mathbf{A}_{i,:}\|_2.$$

Finally, we unfix  $\mathbf{A}$  by taking expectation of both sides of this inequality with respect to  $\mathbf{A}$  and using the law of total expectation.  $\square$

- We state Theorem 2 for symmetric matrices, but it is simple to extend it to general  $m \times n$  random matrices  $\mathbf{A}$ . The bound in this case becomes

$$\mathbb{E}\|\mathbf{A}\| \leq C \log(m+n) \cdot (\mathbb{E} \max_i \|\mathbf{A}_{i,:}\|_2 + \mathbb{E} \max_j \|\mathbf{A}_{:,j}\|_2).$$

To see this, apply Theorem 2 to the  $(m+n) \times (m+n)$  symmetric random matrix

$$\begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{0} \end{bmatrix}.$$

### 3. Matrix completion

- Consider a fixed, unknown  $n \times n$  matrix  $\mathbf{X}$ . Suppose we are shown  $m$  randomly chosen entries of  $\mathbf{X}$ . Can we guess all the missing entries? This important problem is called *matrix completion*. We will analyze it using the bounds on the norms on random matrices we just obtained.

- Obviously, there is no way to guess the missing entries unless we know something extra about the matrix  $\mathbf{X}$ . So let us assume that  $\mathbf{X}$  has low rank:

$$\text{rank}(\mathbf{X}) =: r \ll n.$$

The number of degrees of freedom of an  $n \times n$  matrix with rank  $r$  is  $\mathcal{O}(rn)$ . (Why?) So we may hope that  $m \sim rn$  observed entries of  $\mathbf{X}$  will be enough to determine  $\mathbf{X}$  completely. But how?

- Here we will analyze what is probably the simplest method for matrix completion. Take the matrix  $\mathbf{Y}$  that consists of the observed entries of  $\mathbf{X}$  while all unobserved entries are set to zero. Unlike  $\mathbf{X}$ , the matrix  $\mathbf{Y}$  may not have small rank. Compute the best rank  $r$  approximation of  $\mathbf{Y}$ . The result, as we will show, will be a good approximation to  $\mathbf{X}$ .

- But before we show this, let us define sampling of entries more rigorously. Assume each entry of  $\mathbf{X}$  is shown or hidden independently of others with fixed probability  $p$ . Which entries are shown is decided by independent Bernoulli random variables

$$\delta_{ij} \sim \text{Ber}(p) \quad \text{with} \quad p := \frac{m}{n^2}$$

which are often called selectors in this context. The value of  $p$  is chosen so that among  $n^2$  entries of  $\mathbf{X}$ , the expected number of selected (known) entries is  $m$ .

- Define the  $n \times n$  matrix  $\mathbf{Y}$  with entries  $\mathbf{Y}_{ij} := \delta_{ij} \mathbf{X}_{ij}$ . We can assume that we are shown  $\mathbf{Y}$ , for it is a matrix that contains the observed entries of  $\mathbf{X}$  while all unobserved entries are replaced with zeros. The following result shows how to estimate  $\mathbf{X}$  based on  $\mathbf{Y}$ .



## Theorem 4 (Matrix completion)

Let  $\widehat{\mathbf{X}}$  be a best rank  $r$  approximation to  $p^{-1}\mathbf{Y}$ . Then

$$\mathbb{E} \frac{1}{n} \|\widehat{\mathbf{X}} - \mathbf{X}\|_{\text{F}} \leq C \log n \sqrt{\frac{rn}{m}} \|\mathbf{X}\|_{\max}.$$

Here  $\|\mathbf{X}\|_{\max} = \max_{i,j} |\mathbf{X}_{ij}|$  denotes the maximum magnitude of the entries of  $\mathbf{X}$ .

*Remark.* This theorem controls the average error per entry in the mean-squared sense. To make the error small, let us assume that we have a sample of size  $m \gg rn \log^2 n$ , which is slightly larger than the ideal size  $m \sim rn$ . This makes  $C \log n \sqrt{rn/m} = o(1)$  and forces the recovery error to be bounded by  $o(1)\|\mathbf{X}\|_{\max}$ . Summarizing, Theorem 4 says that *the expected average error per entry is much smaller than the maximal magnitude of the entries of  $\mathbf{X}$* . This is true for a sample of almost optimal size  $m$ . The smaller the rank  $r$  of the matrix  $\mathbf{X}$ , the fewer entries of  $\mathbf{X}$  we need to see in order to do matrix completion.

*Proof of Theorem 4.*

**Step 1: The error in the operator norm.** Let us first bound the recovery error in the operator norm. Decompose the error into two parts using triangle inequality:

$$\|\widehat{\mathbf{X}} - \mathbf{X}\| \leq \|\widehat{\mathbf{X}} - p^{-1}\mathbf{Y}\| + \|p^{-1}\mathbf{Y} - \mathbf{X}\|.$$

Recall that  $\widehat{\mathbf{X}}$  is a best approximation to  $p^{-1}\mathbf{Y}$ . Then the first part of the error is smaller than the second part, and we have

$$\|\widehat{\mathbf{X}} - \mathbf{X}\| \leq 2\|p^{-1}\mathbf{Y} - \mathbf{X}\| = \frac{2}{p}\|\mathbf{Y} - p\mathbf{X}\|.$$

The entries of the matrix  $\mathbf{Y} - p\mathbf{X}$ ,

$$(\mathbf{Y} - p\mathbf{X})_{ij} = (\delta_{ij} - p)\mathbf{X}_{ij},$$

are independent and mean zero random variables.

We have

$$\begin{aligned} & \mathbb{E} \|\mathbf{Y} - p\mathbf{X}\| \\ & \leq C \log n \cdot \left( \mathbb{E} \max_i \|(\mathbf{Y} - p\mathbf{X})_{i,:}\|_2 + \mathbb{E} \max_j \|(\mathbf{Y} - p\mathbf{X})_{:,j}\|_2 \right) \end{aligned}$$

All that remains is to bound the norms of the rows and columns of  $\mathbf{Y} - p\mathbf{X}$ . This is not difficult if we note that they can be expressed as sums of independent random variables:

$$\|(\mathbf{Y} - p\mathbf{X})_{i,:}\|_2^2 = \sum_{j=1}^n (\delta_{ij} - p)^2 \mathbf{X}_{ij}^2 \leq \sum_{j=1}^n (\delta_{ij} - p)^2 \cdot \|\mathbf{X}\|_{\max}^2,$$

and similarly for columns. Taking expectation and noting that

$$\mathbb{E} (\delta_{ij} - p)^2 = \text{Var} (\delta_{ij}) = p(1 - p),$$

we get

$$\mathbb{E} \|(\mathbf{Y} - p\mathbf{X})_{i,:}\|_2 \leq \left( \mathbb{E} \|(\mathbf{Y} - p\mathbf{X})_{i,:}\|_2^2 \right)^{1/2} \leq \sqrt{pn} \|\mathbf{X}\|_{\max}.$$

This is a good bound, but we need something stronger. Since the maximum appears inside the expectation, we need a uniform bound, which will say that all rows are bounded simultaneously with high probability. Such uniform bounds are usually proved by applying concentration inequalities followed by a union bound. Bernsteins inequality yields

$$\mathbb{P} \left\{ \sum_{j=1}^n (\delta_{ij} - p)^2 > t p n \right\} \leq \exp(-c t p n) \quad \text{for } t \geq 3.$$

This probability can be further bounded by  $n^{-ct}$  using the assumption that  $m = p n^2 \geq n \log n$ . A union bound over  $n$  rows leads to

$$\mathbb{P} \left\{ \max_{i \in [n]} \sum_{j=1}^n (\delta_{ij} - p)^2 > t p n \right\} \leq n \cdot n^{-ct} \quad \text{for } t \geq 3.$$

Integrating this tail, we have

$$\mathbb{E} \max_{i \in [n]} \sum_{j=1}^n (\delta_{ij} - p)^2 \lesssim pn.$$

And this yields the desired bound on the rows,

$$\mathbb{E} \max_{i \in [n]} \|(\mathbf{Y} - p\mathbf{X})_{i,:}\|_2 \lesssim \sqrt{pn} \|\mathbf{X}\|_{\max}.$$

We can do similarly for the columns. Then,

$$\mathbb{E} \|\mathbf{Y} - p\mathbf{X}\| \lesssim \log n \sqrt{pn} \|\mathbf{X}\|_{\max}.$$

Therefore, we get

$$\mathbb{E} \|\widehat{\mathbf{X}} - \mathbf{X}\| \lesssim \log n \sqrt{\frac{n}{p}} \|\mathbf{X}\|_{\max}.$$

## Step 2: Passing to Frobenius norm.

We know that  $\text{rank}(\mathbf{X}) \leq r$  by assumption and  $\text{rank}(\widehat{\mathbf{X}}) \leq r$  by construction, so  $\text{rank}(\widehat{\mathbf{X}} - \mathbf{X}) \leq 2r$ . There is a simple relationship between the operator and Frobenius norms:

$$\|\widehat{\mathbf{X}} - \mathbf{X}\|_{\text{F}} \leq \sqrt{2r} \|\widehat{\mathbf{X}} - \mathbf{X}\|.$$

Taking expectation of both sides, we get

$$\mathbb{E} \|\widehat{\mathbf{X}} - \mathbf{X}\|_{\text{F}} \leq \sqrt{2r} \mathbb{E} \|\widehat{\mathbf{X}} - \mathbf{X}\| \lesssim \log n \sqrt{\frac{rn}{p}} \|\mathbf{X}\|_{\max}.$$

Dividing both sides by  $n$ , we can rewrite this bound as

$$\mathbb{E} \frac{1}{n} \|\widehat{\mathbf{X}} - \mathbf{X}\|_{\text{F}} \lesssim \log n \sqrt{\frac{rn}{pn^2}} \|\mathbf{X}\|_{\max}.$$

The proof is completed by noting the definition of the sampling probability  $p = m/n^2$ . □