Lecture 13: Covariance estimation and matrix completion

School of Mathematical Sciences, Xiamen University

1. Covariance estimation

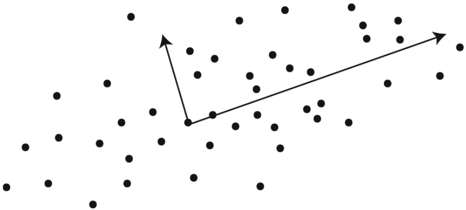
- Suppose we have a sample of data points $X_1, ..., X_N$ in $\mathbb{R}^n$. It is often reasonable to assume that these points are independently sampled from the same probability distribution (or "population") which is unknown. We would like to learn something useful about this distribution.
- Denote 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,

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

- 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 Σ are called the principal components. Principal components that correspond to large eigenvalues of Σ 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 Σ 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.

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

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 Σ. Suppose that

$$\|X\|_2^2 \lesssim \mathbb{E}\|X\|_2^2 = \mathrm{tr}\Sigma \qquad \textit{almost surely.}$$

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

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

where

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

and K is chosen so that

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

It remains to bound σ and K. Let us start with σ. We have

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

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

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

Therefore,

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

The proof is completed by using $\operatorname{tr}(\Sigma) \le n\|\Sigma\|$. $\square$

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 Σ^2. The stable rank of a matrix $\mathbf{A}$ is defined as the square of the ratio of the Frobenius to operator norms:

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

The proof of Theorem 1 yields

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

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

2. Norms of random matrices

- Let $\mathbf{A}_{i,:}$ denote the ith 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. (Better results can be found in Vershynin's HDP-book.)

Theorem 2 (Norms of random matrices without boundedness assumptions)

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

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

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

Lemma 3 (Symmetrization)

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

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

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

$$\mathbb{E}\left\|\sum_i X_i\right\| = \mathbb{E}\left\|\sum_i X_i - \mathbb{E}\left(\sum_i X_i'\right)\right\|$$
$$\overset{\text{(Jensen)}}{\leqslant} \mathbb{E}\left\|\sum_i X_i - \sum_i X_i'\right\| = \mathbb{E}\left\|\sum_i \left(X_i - X_i'\right)\right\|.$$

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

$$\boldsymbol{A} = \sum\nolimits_{i \leq j} \boldsymbol{Z}_{ij}.$$

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

$$\mathbb{E}\|\boldsymbol{A}\| = \mathbb{E}\left\|\sum\nolimits_{i \leqslant j} \boldsymbol{Z}_{ij}\right\| \leqslant 2\mathbb{E}\left\|\sum\nolimits_{i \leqslant j} \boldsymbol{X}_{ij}\right\|$$

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

Note that $\boldsymbol{X}_{ij}$ 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

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

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

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

Then we have

$$\mathbb{E}_{\varepsilon} \left\| \sum\nolimits_{i \leqslant j} \boldsymbol{X}_{ij} \right\| \lesssim \log n \cdot \max_i \|\boldsymbol{A}_{i,:}\|_2 \,.$$

Finally, we unfix $\boldsymbol{A}$ by taking expectation of both sides of this inequality with respect to $\boldsymbol{A}$ 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 $\boldsymbol{A}$. The bound in this case becomes

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

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

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

3. Matrix completion

- Consider a fixed, unknown $n \times n$ matrix $\boldsymbol{X}$. Suppose we are shown m randomly chosen entries of $\boldsymbol{X}$. 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 $\boldsymbol{X}$. So let us assume that $\boldsymbol{X}$ has low rank:

$$\operatorname{rank}(\boldsymbol{X}) =: r \ll n.$$

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

- But before we show this, let us define sampling of entries more rigorously. Assume each entry of $\boldsymbol{X}$ 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 $\boldsymbol{X}$, the expected number of selected (known) entries is m.

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

Theorem 4 (Matrix completion)

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

$$\mathbb{E}\frac{1}{n}\|\widehat{\boldsymbol{X}}-\boldsymbol{X}\|_{\mathrm{F}} \leqslant C\log n\sqrt{\frac{rn}{m}}\|\boldsymbol{X}\|_{\max}.$$

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

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 $o(1)\|\boldsymbol{X}\|_{\max}$. Summarizing, Theorem 4 says that *the expected average error per entry is much smaller than the maximal magnitude of the entries of $\boldsymbol{X}$.* This is true for a sample of almost optimal size m. The smaller the rank r of the matrix $\boldsymbol{X}$, the fewer entries of $\boldsymbol{X}$ 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:

$$\|\widehat{\boldsymbol{X}} - \boldsymbol{X}\| \le \|\widehat{\boldsymbol{X}} - p^{-1}\boldsymbol{Y}\| + \|p^{-1}\boldsymbol{Y} - \boldsymbol{X}\|.$$

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

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

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

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

are independent and mean zero random variables.

We have

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

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

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

and similarly for columns. Taking expectation and noting that

$$\mathbb{E}\,(\delta_{ij}-p)^2=\mathbb{V}\mathrm{ar}\,(\delta_{ij})=p(1-p),$$

we get

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

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

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

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

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

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,

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

We can do similarly for the columns. Then,

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

Therefore, we get

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

Step 2: Passing to Frobenius norm.

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

$$\|\widehat{\boldsymbol{X}} - \boldsymbol{X}\|_{\mathrm{F}} \leq \sqrt{2r}\|\widehat{\boldsymbol{X}} - \boldsymbol{X}\|.$$

Taking expectation of both sides, we get

$$\mathbb{E}\|\widehat{\boldsymbol{X}} - \boldsymbol{X}\|_{\mathrm{F}} \leqslant \sqrt{2r}\mathbb{E}\|\widehat{\boldsymbol{X}} - \boldsymbol{X}\| \lesssim \log n \sqrt{\frac{rn}{p}}\|\boldsymbol{X}\|_{\max}.$$

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

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

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