Matrix Concentration: Applications

SDS 391P.6, Spring 2026

Pratik Patil

1 Motivation

In the previous lecture, we developed the *matrix Laplace transform method* and its consequences for sums of independent random symmetric matrices. Those results reduced matrix concentration to a familiar-looking problem: control the matrix cumulant generating functions of the summands, and then optimize in the Laplace parameter.

Today we turn that abstract framework into concrete inequalities. Our goals are:

- to prove matrix versions of Hoeffding and Bernstein inequalities;
- to see how these matrix inequalities differ from their scalar counterparts;
- to apply them to two basic problems:
    1. a matrix Khintchine inequality for Rademacher matrix series;
    2. covariance estimation beyond the sub-Gaussian setting.

At a high level, the philosophy is the same as in the scalar case:

$$\text{CGF bound} \quad \Longrightarrow \quad \text{Chernoff/Laplace bound} \quad \Longrightarrow \quad \text{tail and expectation estimates.}$$

The main new feature here is that matrices do not commute, so additivity of scalar CGFs is replaced by the subadditivity principle from the previous lecture (proved using Lieb's concavity theorem).

There is also a useful conceptual distinction between the two inequalities we will prove:

- Matrix Hoeffding captures *purely Gaussian* behavior. It is the right tool for Rademacher matrix sums (and the matrix analog of scalar Hoeffding) and other situations where a quadratic CGF bound holds globally.
- Matrix Bernstein captures *two-regime* behavior. It is the right tool for bounded summands (and the matrix analog of scalar Bernstein): Gaussian tails at moderate scale, exponential tails at large scale.

As in the scalar case, matrix Bernstein is more flexible. In particular, it allows us to estimate sample covariance matrices under assumptions that are substantially weaker than sub-Gaussianity.

2 Recap: the matrix concentration framework

Let $X_1, \dots, X_n$ be independent random symmetric matrices in $\mathbb{S}^d$, and set

$$Y := \sum_{i=1}^{n} X_i.$$

From the previous lecture, the two structural facts we need are the following.

2.1 Matrix Laplace transform method

For any random symmetric matrix Y,

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant \inf_{\theta>0} e^{-\theta t}\, \mathbb{E}\,\mathrm{tr}\, e^{\theta Y}.$$

Likewise,

$$\mathbb{P}\{\lambda_{\min}(Y) \leqslant t\} \leqslant \inf_{\theta<0} e^{-\theta t}\, \mathbb{E}\,\mathrm{tr}\, e^{\theta Y}.$$

This is the matrix analog of Chernoff's method. The only visible difference from the scalar case is the appearance of $\mathrm{tr}\, e^{\theta Y}$ in place of $e^{\theta Y}$.

2.2 Subadditivity of matrix CGFs

If

$$\Xi_{X_i}(\theta) := \log \mathbb{E} e^{\theta X_i}$$

denotes the matrix CGF of X_i, then

$$\mathbb{E}\,\mathrm{tr}\exp\left(\theta \sum_{i=1}^{n} X_i\right) \leqslant \mathrm{tr}\exp\left(\sum_{i=1}^{n} \Xi_{X_i}(\theta)\right).$$

Thus the matrix Laplace transform method yields the tail bound

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant \inf_{\theta>0} e^{-\theta t}\ \mathrm{tr}\exp\left(\sum_{i=1}^{n} \Xi_{X_i}(\theta)\right). \tag{1}$$

This inequality is abstract but very useful. Once we have an upper bound on each matrix CGF $\Xi_{X_i}(\theta)$, the whole sum is under control.

So the main remaining technical task is to find usable Loewner-order bounds on $\log \mathbb{E} e^{\theta X}$.

3 Matrix Hoeffding inequality for Rademacher matrix series

We begin with the cleanest setting: a matrix-valued Rademacher series. This is the direct matrix analog of the scalar Rademacher series

$$\sum_{i=1}^{n} \varepsilon_i a_i,$$

which we studied earlier in the course.

Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables, and let $A_1, \dots, A_n \in \mathbb{S}^d$ be fixed symmetric matrices. Define

$$Y := \sum_{i=1}^{n} \varepsilon_i A_i.$$

We want to control $\|Y\|$.

3.1 Matrix Hoeffding CGF bound

The key observation is that the matrix exponential behaves especially well under Rademacher symmetrization.

Lemma 3.1 (Matrix Hoeffding CGF bound)**.** *Let ε be a Rademacher random variable and let $A \in \mathbb{S}^d$ be deterministic. Then for every $\theta \in \mathbb{R}$,*

$$\log \mathbb{E} e^{\theta \varepsilon A} \preceq \frac{\theta^2}{2} A^2.$$

Proof. By definition of the Rademacher distribution,

$$\mathbb{E} e^{\theta \varepsilon A} = \frac{1}{2} e^{\theta A} + \frac{1}{2} e^{-\theta A} = \cosh(\theta A),$$

where cosh is understood via functional calculus.

For scalars, one has

$$\cosh x \leqslant e^{x^2/2} \qquad \text{for all } x \in \mathbb{R}.$$

Applying this inequality to the eigenvalues of θA and using functional calculus gives

$$\cosh(\theta A) \preceq e^{\theta^2 A^2/2}.$$

Now both sides are positive definite, and the matrix logarithm is matrix monotone on the positive definite cone, so

$$\log \mathbb{E} e^{\theta \varepsilon A} = \log \cosh(\theta A) \preceq \log e^{\theta^2 A^2/2} = \frac{\theta^2}{2} A^2.$$

□

This is exactly the matrix analog of the scalar sub-Gaussian MGF bound for a Rademacher variable. The important point is that the variance proxy is not a scalar but the matrix A^2.

3.2 Matrix Hoeffding inequality

We now plug the CGF estimate into the tail bound (1).

Theorem 3.2 (Matrix Hoeffding inequality)**.** *Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables, and let $A_1, \dots, A_n \in \mathbb{S}^d$ be fixed symmetric matrices. Define*

$$Y := \sum_{i=1}^{n} \varepsilon_i A_i, \qquad \sigma^2 := \left\| \sum_{i=1}^{n} A_i^2 \right\|.$$

Then for every $t \geqslant 0$,

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant d \exp\left(-\frac{t^2}{2\sigma^2} \right).$$

Consequently,

$$\mathbb{P}\{\|Y\| \geqslant t\} \leqslant 2d \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Proof. By Lemma 3.1,

$$\Xi_{\varepsilon_i A_i}(\theta) = \log \mathbb{E} e^{\theta \varepsilon_i A_i} \preceq \frac{\theta^2}{2} A_i^2.$$

Summing over i,

$$\sum_{i=1}^n \Xi_{\varepsilon_i A_i}(\theta) \preceq \frac{\theta^2}{2} \sum_{i=1}^n A_i^2 \preceq \frac{\theta^2 \sigma^2}{2} I.$$

Using trace monotonicity for the exponential,

$$\operatorname{tr} \exp\left(\sum_{i=1}^n \Xi_{\varepsilon_i A_i}(\theta)\right) \leqslant \operatorname{tr} \exp\left(\frac{\theta^2 \sigma^2}{2} I\right) = d\, e^{\theta^2 \sigma^2/2}.$$

Now apply the tail bound (1):

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant \inf_{\theta > 0} \exp\left(-\theta t + \frac{\theta^2 \sigma^2}{2} + \log d\right).$$

Optimizing in θ at $\theta = t/\sigma^2$ yields

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant d \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Since Y and $-Y$ have the same distribution,

$$\mathbb{P}\{\lambda_{\min}(Y) \leqslant -t\} = \mathbb{P}\{\lambda_{\max}(-Y) \geqslant t\} \leqslant d \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Using

$$\|Y\| = \max\{\lambda_{\max}(Y), -\lambda_{\min}(Y)\}$$

and the union bound gives the two-sided estimate. □

A useful way to read the theorem is this: the matrix Rademacher series behaves like a scalar sub-Gaussian random variable, except for the extra factor d, or equivalently $\log d$, that appears because we are controlling a whole spectrum rather than one scalar.

4 Application: matrix Khintchine inequality

The tail bound in Theorem 3.2 immediately yields an expectation bound. This is a convenient expectation form of the matrix Khintchine inequality.

Corollary 4.1 (Matrix Khintchine inequality)**.** *Under the assumptions of Theorem 3.2,*

$$\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i A_i\right\| \leqslant C\sqrt{\log(2d)} \left\|\sum_{i=1}^n A_i^2\right\|^{1/2},$$

where $C > 0$ is an absolute constant.

Proof. Set

$$Z := \left\| \sum_{i=1}^{n} \varepsilon_i A_i \right\|.$$

Since $Z \geqslant 0$,

$$\mathbb{E}Z = \int_0^\infty \mathbb{P}\{Z \geqslant t\}\, dt.$$

By Theorem 3.2,

$$\mathbb{P}\{Z \geqslant t\} \leqslant 2d\, e^{-t^2/(2\sigma^2)}, \qquad \sigma^2 = \left\| \sum_{i=1}^{n} A_i^2 \right\|.$$

Choose

$$t_0 := \sigma\sqrt{2\log(2d)}.$$

Split the integral:

$$\mathbb{E}Z = \int_0^{t_0} \mathbb{P}\{Z \geqslant t\}\, dt + \int_{t_0}^\infty \mathbb{P}\{Z \geqslant t\}\, dt \leqslant t_0 + 2d \int_{t_0}^\infty e^{-t^2/(2\sigma^2)}\, dt.$$

Use the standard Gaussian-tail bound

$$\int_{t_0}^\infty e^{-t^2/(2\sigma^2)}\, dt \leqslant \frac{\sigma^2}{t_0}\, e^{-t_0^2/(2\sigma^2)}.$$

Since $2d\, e^{-t_0^2/(2\sigma^2)} = 1$, we obtain

$$\mathbb{E}Z \leqslant t_0 + \frac{\sigma^2}{t_0} \leqslant C\,\sigma\sqrt{\log(2d)}.$$

Substituting the definition of σ completes the proof. □

This estimate is the matrix analog of the scalar identity

$$\left(\mathbb{E}\Big| \sum_i \varepsilon_i a_i \Big|^2 \right)^{1/2} = \Big(\sum_i a_i^2 \Big)^{1/2},$$

except that the matrix case pays only a logarithmic price in the dimension.

Remark 4.2. With a little more work, one can similarly prove the L_p version

$$\left(\mathbb{E} \left\| \sum_{i=1}^{n} \varepsilon_i A_i \right\|^p \right)^{1/p} \leqslant C\,\sqrt{p + \log(2d)}\, \left\| \sum_{i=1}^{n} A_i^2 \right\|^{1/2}, \qquad p \geqslant 1.$$

We will not pursue this here.

5 Matrix Bernstein inequality

We now move to the main result of the lecture. Unlike matrix Hoeffding, which gives purely Gaussian tails, matrix Bernstein captures both moderate and large deviations.

Let $X_1, \dots, X_n$ be independent random symmetric matrices in $\mathbb{S}^d$ with

$$\mathbb{E}X_i = 0, \qquad \|X_i\| \leqslant B \quad \text{almost surely.}$$

Define

$$Y := \sum_{i=1}^n X_i, \qquad v := \left\| \sum_{i=1}^n \mathbb{E}X_i^2 \right\|.$$

The quantity v is the matrix variance proxy.

5.1 Matrix Bernstein CGF bound

The proof follows the scalar Bernstein proof very closely. The only real work is to lift the scalar exponential inequality to the matrix setting.

Lemma 5.1 (Matrix Bernstein CGF bound)**.** *Let $X \in \mathbb{S}^d$ be a random symmetric matrix with*

$$\mathbb{E}X = 0, \qquad \|X\| \leqslant B \quad \textit{almost surely.}$$

Then for every $0 \leqslant \theta < 3/B$,

$$\log \mathbb{E}e^{\theta X} \leq \frac{\theta^2/2}{1 - \theta B/3} \, \mathbb{E}X^2.$$

Proof. We start from the scalar inequality

$$e^u \leqslant 1 + u + \frac{u^2/2}{1 - |u|/3}, \qquad |u| < 3.$$

Indeed,

$$e^u = 1 + u + \sum_{k=2}^{\infty} \frac{u^k}{k!},$$

and for $k \geqslant 2$,

$$k! \geqslant 2 \cdot 3^{k-2},$$

so

$$\sum_{k=2}^{\infty} \frac{|u|^k}{k!} \leqslant \frac{u^2}{2} \sum_{k=0}^{\infty} \left(\frac{|u|}{3} \right)^k = \frac{u^2/2}{1 - |u|/3}.$$

Now let $x \in [-B, B]$ and $0 \leqslant \theta < 3/B$. Since $|\theta x| \leqslant \theta B < 3$, the previous inequality gives

$$e^{\theta x} \leqslant 1 + \theta x + \frac{\theta^2/2}{1 - \theta B/3} \, x^2.$$

By functional calculus, for a symmetric matrix X with $\|X\| \leqslant B$,

$$e^{\theta X} \leq I + \theta X + \frac{\theta^2/2}{1 - \theta B/3} \, X^2.$$

Taking expectations and using $\mathbb{E}X = 0$,

$$\mathbb{E}e^{\theta X} \preceq I + \frac{\theta^2/2}{1-\theta B/3}\,\mathbb{E}X^2.$$

Since for any symmetric matrix M,

$$I + M \preceq e^M$$

(by applying $1 + u \leqslant e^u$ to the eigenvalues of M), we obtain

$$\mathbb{E}e^{\theta X} \preceq \exp\left(\frac{\theta^2/2}{1-\theta B/3}\,\mathbb{E}X^2\right).$$

Finally, both sides are positive definite, and the matrix logarithm is matrix monotone on the positive definite cone, hence

$$\log \mathbb{E}e^{\theta X} \preceq \frac{\theta^2/2}{1-\theta B/3}\,\mathbb{E}X^2.$$

□

This is exactly the matrix counterpart of the scalar Bernstein CGF bound. The parameter B controls the radius of validity of the quadratic approximation, while $\mathbb{E}X^2$ is the variance-type term.

5.2 Matrix Bernstein tail bound

Theorem 5.2 (Matrix Bernstein inequality)**.** *Let $X_1, \dots, X_n$ be independent random symmetric matrices in $\mathbb{S}^d$ such that*

$$\mathbb{E}X_i = 0, \qquad \|X_i\| \leqslant B \quad \text{almost surely}$$

for all i. Define

$$v := \left\|\sum_{i=1}^n \mathbb{E}X_i^2\right\|.$$

Then for every $t \geqslant 0$,

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant d\exp\left(-\frac{t^2/2}{v + Bt/3}\right), \qquad Y := \sum_{i=1}^n X_i.$$

Consequently,

$$\mathbb{P}\{\|Y\| \geqslant t\} \leqslant 2d\exp\left(-\frac{t^2/2}{v + Bt/3}\right).$$

Proof. By Lemma 5.1, for every $0 \leqslant \theta < 3/B$,

$$\Xi_{X_i}(\theta) = \log \mathbb{E}e^{\theta X_i} \preceq \frac{\theta^2/2}{1-\theta B/3}\,\mathbb{E}X_i^2.$$

Summing over i,

$$\sum_{i=1}^n \Xi_{X_i}(\theta) \preceq \frac{\theta^2/2}{1-\theta B/3}\sum_{i=1}^n \mathbb{E}X_i^2 \preceq \frac{\theta^2 v/2}{1-\theta B/3}\,I.$$

Therefore,

$$\operatorname{tr}\exp\left(\sum_{i=1}^{n}\Xi_{X_i}(\theta)\right) \leqslant \operatorname{tr}\exp\left(\frac{\theta^2 v/2}{1-\theta B/3}\,I\right) = d\exp\left(\frac{\theta^2 v/2}{1-\theta B/3}\right).$$

Applying the tail bound (1),

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant \inf_{0<\theta<3/B} \exp\left(-\theta t + \frac{\theta^2 v/2}{1-\theta B/3} + \log d\right).$$

Choose

$$\theta = \frac{t}{v + Bt/3}.$$

Then $\theta < 3/B$, and a direct substitution gives

$$-\theta t + \frac{\theta^2 v/2}{1-\theta B/3} = -\frac{t^2/2}{v + Bt/3}.$$

Hence

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant t\} \leqslant d\exp\left(-\frac{t^2/2}{v + Bt/3}\right).$$

Applying the same argument to $-Y$ and using the union bound gives the two-sided estimate for $\|Y\|$. □

It is often convenient to rewrite the exponent in the more familiar two-regime form

$$\exp\left[-c\min\left(\frac{t^2}{v}, \frac{t}{B}\right)\right].$$

Thus matrix Bernstein says exactly what scalar Bernstein says: moderate deviations are Gaussian, large deviations are exponential.

5.3 Expectation form

Just as in the scalar setting, tail bounds lead to expectation bounds.

Corollary 5.3 (Expectation bound from matrix Bernstein)**.** *Under the assumptions of Theorem 5.2,*

$$\mathbb{E}\|Y\| \leqslant C\left(\sqrt{v\log(2d)} + B\log(2d)\right)$$

for an absolute constant C.

Proof. Apply the two-sided tail bound from Theorem 5.2 and integrate:

$$\mathbb{E}\|Y\| = \int_0^\infty \mathbb{P}\{\|Y\| \geqslant t\}\,dt.$$

Using the two-regime form of the Bernstein tail and splitting the integral at the natural transition scale gives the claimed estimate. □

6 Application: covariance estimation beyond sub-Gaussian data

We now return to the covariance estimation problem, but in a more general setting than before.

In earlier lectures, we used net arguments and scalar Bernstein inequalities to study covariance estimation for sub-Gaussian data. That approach exploited strong one-dimensional tail assumptions. Matrix Bernstein gives a complementary route: it works directly at the matrix level, and it requires no sub-Gaussian assumption.

6.1 Setup

Let $X \in \mathbb{R}^d$ be a centered random vector with covariance matrix

$$\Sigma := \mathbb{E}[XX^\top].$$

Let $X_1, \dots, X_N$ be i.i.d. copies of X, and define the sample covariance matrix

$$\widehat{\Sigma} := \frac{1}{N} \sum_{i=1}^{N} X_i X_i^\top.$$

Then $\mathbb{E}\widehat{\Sigma} = \Sigma$.

To obtain a matrix Bernstein bound, we need bounded summands. We therefore assume a bounded-norm condition: there exists $K \geqslant 1$ such that

$$\|X\|_2 \leqslant K \left(\mathbb{E}\|X\|_2^2\right)^{1/2} \qquad \text{almost surely.} \tag{2}$$

Since

$$\mathbb{E}\|X\|_2^2 = \operatorname{tr}(\Sigma),$$

this is equivalent to

$$\|X\|_2^2 \leqslant K^2 \operatorname{tr}(\Sigma) \qquad \text{almost surely.}$$

It is useful to introduce the *effective rank*

$$r(\Sigma) := \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|}.$$

This quantity satisfies

$$1 \leqslant r(\Sigma) \leqslant \operatorname{rank}(\Sigma) \leqslant d$$

whenever $\Sigma \neq 0$, and it measures the effective dimension of the covariance matrix.

6.2 A covariance deviation bound

Theorem 6.1 (Covariance estimation under bounded norm)**.** *Assume* (2)*, and let*

$$\widehat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} X_i X_i^\top.$$

Then there exists an absolute constant $C > 0$ such that for every $u \geqslant 0$,

$$\mathbb{P}\left\{ \|\widehat{\Sigma} - \Sigma\| \geqslant C \left(\sqrt{\frac{K^2 r(\Sigma)\,(\log(2d) + u)}{N}} + \frac{K^2 r(\Sigma)\,(\log(2d) + u)}{N} \right) \|\Sigma\| \right\} \leqslant 2e^{-u}.$$

Consequently,

$$\mathbb{E}\|\widehat{\Sigma} - \Sigma\| \leqslant C\left(\sqrt{\frac{K^2 r(\Sigma)\log(2d)}{N}} + \frac{K^2 r(\Sigma)\log(2d)}{N}\right)\|\Sigma\|.$$

Proof. Apply matrix Bernstein to the centered summands

$$Y_i := X_i X_i^\top - \Sigma.$$

Then

$$\widehat{\Sigma} - \Sigma = \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

We first bound the matrix variance proxy. Since

$$Y_i^2 = (X_i X_i^\top - \Sigma)^2,$$

we compute

$$\mathbb{E}Y_i^2 = \mathbb{E}(XX^\top - \Sigma)^2 = \mathbb{E}(XX^\top)^2 - \Sigma^2 \preceq \mathbb{E}(XX^\top)^2.$$

Now

$$(XX^\top)^2 = \|X\|_2^2\, XX^\top,$$

so by (2),

$$(XX^\top)^2 \preceq K^2 \operatorname{tr}(\Sigma)\, XX^\top.$$

Taking expectations yields

$$\mathbb{E}(XX^\top)^2 \preceq K^2 \operatorname{tr}(\Sigma)\,\Sigma.$$

Therefore

$$\|\mathbb{E}Y_i^2\| \leqslant K^2 \operatorname{tr}(\Sigma)\,\|\Sigma\| = K^2 r(\Sigma)\,\|\Sigma\|^2.$$

Summing over $i = 1, \ldots, N$,

$$v := \left\|\sum_{i=1}^{N} \mathbb{E}Y_i^2\right\| \leqslant NK^2 r(\Sigma)\,\|\Sigma\|^2.$$

Next we bound the summand norm:

$$\|Y_i\| \leqslant \|X_i X_i^\top\| + \|\Sigma\| = \|X_i\|_2^2 + \|\Sigma\|.$$

Using $\|X_i\|_2^2 \leqslant K^2 \operatorname{tr}(\Sigma) = K^2 r(\Sigma)\|\Sigma\|$ and $K \geqslant 1$,

$$\|Y_i\| \leqslant \big(K^2 r(\Sigma) + 1\big)\|\Sigma\| \leqslant 2K^2 r(\Sigma)\|\Sigma\|.$$

So we may take

$$B := 2K^2 r(\Sigma)\|\Sigma\|.$$

Now apply Theorem 5.2 to $\sum_{i=1}^N Y_i$: for every $s \geqslant 0$,

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{N} Y_i\right\| \geqslant s\right\} \leqslant 2d\exp\left(-\frac{s^2/2}{v + Bs/3}\right).$$

Set $s = Nt$. Then

$$\mathbb{P}\{\|\widehat{\Sigma} - \Sigma\| \geqslant t\} = \mathbb{P}\left\{\left\|\sum_{i=1}^{N} Y_i\right\| \geqslant Nt\right\} \leqslant 2d \exp\left(-\frac{N^2t^2/2}{v + BNt/3}\right).$$

Substituting the bounds on v and B,

$$\mathbb{P}\{\|\widehat{\Sigma} - \Sigma\| \geqslant t\} \leqslant 2d \exp\left(-\frac{Nt^2/2}{K^2 r(\Sigma)\|\Sigma\|^2 + cK^2 r(\Sigma)\|\Sigma\| t}\right)$$

for an absolute constant c.

Choosing

$$t = C\left(\sqrt{\frac{K^2 r(\Sigma)\,(\log(2d) + u)}{N}} + \frac{K^2 r(\Sigma)\,(\log(2d) + u)}{N}\right)\|\Sigma\|$$

with C large enough makes the exponent dominate $\log(2d) + u$, and therefore the right-hand side is at most $2e^{-u}$. This proves the high-probability estimate.

The expectation bound follows from Corollary 5.3 applied to $\sum_{i=1}^{N} Y_i$, followed by division by N. □

A few features of Theorem 6.1 are worth emphasizing.

First, the sample size requirement depends on the effective rank $r(\Sigma)$, not directly on the ambient dimension d, except through the mild logarithmic factor $\log d$. This matches the general principle that low-dimensional structure inside high-dimensional data should reduce sample complexity.

Second, there is no sub-Gaussian assumption. The argument only used boundedness of $\|X\|_2$, or more realistically a truncation condition that removes large outliers before applying the theorem.

Third, the proof is very different from the net-based covariance proof from earlier in the course. There we reduced the operator norm to finitely many scalar quadratic forms. Here we stay at the matrix level throughout. This is often cleaner and more flexible.

Remark 6.2 (Sample complexity)**.** If one wants

$$\|\widehat{\Sigma} - \Sigma\| \leqslant \varepsilon\|\Sigma\|$$

with high probability, Theorem 6.1 shows that it is enough to take

$$N \gtrsim K^2\, r(\Sigma)\, \frac{\log d}{\varepsilon^2}.$$

Up to the logarithmic factor, this is linear in the effective dimension.

Remark 6.3 (Rectangular matrices)**.** Everything in this lecture was formulated for symmetric matrices. Rectangular versions follow from the *Hermitian dilation* trick: for $Z \in \mathbb{R}^{d_1 \times d_2}$, define

$$\mathcal{H}(Z) := \begin{pmatrix} 0 & Z \\ Z^\top & 0 \end{pmatrix} \in \mathbb{S}^{d_1 + d_2}.$$

Then

$$\lambda_{\max}(\mathcal{H}(Z)) = \|Z\|, \qquad \mathcal{H}(Z)^2 = \begin{pmatrix} ZZ^\top & 0 \\ 0 & Z^\top Z \end{pmatrix}.$$

So matrix Bernstein and matrix Khintchine inequalities for rectangular matrices are formal consequences of the symmetric case.

7 Look ahead

In the previous lecture, we built the abstract machinery for matrix concentration: matrix Laplace transform bounds and subadditivity of matrix CGFs. In this lecture, we turned that framework into concrete inequalities.

We first proved a matrix Hoeffding inequality for Rademacher matrix series, which immediately yielded a matrix Khintchine inequality. We then proved matrix Bernstein, the main workhorse of the subject, and used it to study covariance estimation under bounded-norm assumptions, without any sub-Gaussian hypothesis.

The big picture to keep in mind is: matrix concentration works much like scalar concentration, except that noncommutativity forces us to replace exact multiplicative identities by matrix-analytic inequalities such as Lieb's concavity theorem. Once that replacement is available, many familiar scalar ideas survive with only a logarithmic dimensional penalty.

Source material

Parts of this lecture are based on references: Vershynin (2018); Tropp (2023), in addition to the author's accumulated experience working on related topics.

References

Tropp, J. A. (2023). Probability in high dimensions. Caltech CMS Lecture Notes 2021-01.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.