Covariance Estimation

SDS 391P.6, Spring 2026

Pratik Patil

1 Recall

This lecture is our first concrete use of metric entropy as a *uniformization* device: we start from concentration in a *fixed* direction, and we upgrade it to a bound that holds *simultaneously for all directions* by discretizing the sphere and taking a union bound.

The statistical target is the operator-norm error of the sample covariance: given i.i.d. samples $X_1, \dots, X_N \in \mathbb{R}^d$ with population covariance $\Sigma = \mathbb{E}[XX^\top]$, we form

$$\widehat{\Sigma} \;:=\; \frac{1}{N}\sum_{i=1}^{N} X_i X_i^\top, \qquad \mathbb{E}\widehat{\Sigma} = \Sigma,$$

and we want a high-probability bound on $\|\widehat{\Sigma} - \Sigma\|$. The key identity is the variational characterization

$$\|\widehat{\Sigma} - \Sigma\| = \sup_{u \in S^{d-1}} \left|u^\top(\widehat{\Sigma} - \Sigma)u\right|.$$

So we are staring at a supremum over the sphere, exactly the setting of the previous lecture.

Recall from the previous lecture that for $\varepsilon \in (0, 1]$, the unit sphere admits an ε-net $\mathcal{N} \subset S^{d-1}$ with

$$|\mathcal{N}| \;\leqslant\; \left(\frac{3}{\varepsilon}\right)^d, \qquad \text{so} \qquad \log|\mathcal{N}| \;\lesssim\; d\log(1/\varepsilon).$$

This is the "complexity price" we must pay when we union bound over directions.

2 The ε-net reductions

The first step is a deterministic reduction: replace a supremum over the sphere by a maximum over a finite net, at the cost of a factor depending on ε.

2.1 General matrices

Lemma 2.1 (Net reduction for operator norms)**.** *Let A be an $m \times n$ matrix and let $\mathcal{N}$ be an ε-net of S^{n-1} in $\|\cdot\|_2$ with $\varepsilon \in (0, 1)$. Then*

$$\|A\| = \sup_{\|x\|_2 = 1} \|Ax\|_2 \;\leqslant\; \frac{1}{1-\varepsilon} \sup_{x \in \mathcal{N}} \|Ax\|_2.$$

Proof. Fix $x \in S^{n-1}$ and choose $x_0 \in \mathcal{N}$ with $\|x - x_0\|_2 \leqslant \varepsilon$. Then

$$\|Ax\|_2 \leqslant \|Ax_0\|_2 + \|A(x - x_0)\|_2 \leqslant \|Ax_0\|_2 + \varepsilon\|A\|.$$

Taking sup over $x \in S^{n-1}$ gives $\|A\| \leqslant \sup_{x_0 \in \mathcal{N}} \|Ax_0\|_2 + \varepsilon\|A\|$ and rearranging yields the claim. □

2.2 Symmetric matrices

For covariance estimation we will apply the same idea to quadratic forms.

Lemma 2.2 (Net reduction for symmetric matrices)**.** *Let $A \in \mathbb{R}^{d\times d}$ be symmetric and let $\mathcal{N}$ be an ε-net of S^{d-1} with $\varepsilon \in (0, 1/2)$. Then*

$$\|A\| = \sup_{\|x\|_2=1} |x^\top Ax| \;\leqslant\; \frac{1}{1-2\varepsilon} \sup_{x\in\mathcal{N}} |x^\top Ax|.$$

Proof. Fix $x \in S^{d-1}$ and choose $x_0 \in \mathcal{N}$ with $\|x - x_0\|_2 \leqslant \varepsilon$. Expand

$$x^\top Ax - x_0^\top Ax_0 = (x - x_0)^\top Ax + x_0^\top A(x - x_0).$$

By Cauchy–Schwarz and $\|x\|_2 = \|x_0\|_2 = 1$,

$$|x^\top Ax| \leqslant |x_0^\top Ax_0| + |(x - x_0)^\top Ax| + |x_0^\top A(x - x_0)| \leqslant |x_0^\top Ax_0| + 2\|A\|\,\|x - x_0\|_2.$$

Thus $|x^\top Ax| \leqslant |x_0^\top Ax_0| + 2\varepsilon\|A\|$. Taking sup over $x \in S^{d-1}$ and rearranging gives the result. □

Lemma 2.2 is the bridge from a continuous supremum to a finite maximum. Once we are finite, we can use concentration in each direction plus a union bound.

3 Covariance estimation

We now implement the "fixed direction → net → union bound → lift back" template.

3.1 Isotropic sub-Gaussian data

We will start with a special case of isotropic covariance and then extend to general covariance.

A random vector $X \in \mathbb{R}^d$ is isotropic if

$$\mathbb{E}X = 0, \qquad \mathbb{E}[XX^\top] = I_d,$$

equivalently $\mathbb{E}\langle X, u\rangle^2 = 1$ for all $u \in S^{d-1}$.

We assume a uniform sub-Gaussian control on all one-dimensional marginals:

$$\|X\|_{\psi_2} := \sup_{u\in S^{d-1}} \|\langle X, u\rangle\|_{\psi_2} \;\leqslant\; K.$$

Let $X_1, \ldots, X_N$ be i.i.d. copies of X and set

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

Since $\mathbb{E}\widehat{\Sigma} = I_d$, our goal is a high-probability bound on $\|\widehat{\Sigma} - I_d\|$. We will proceed in the steps outlined above.

1. Concentration in one direction. Fix $u \in S^{d-1}$ and define

$$Z_i(u) := \langle X_i, u\rangle^2 - 1.$$

Then

$$u^\top(\widehat{\Sigma} - I_d)u = \frac{1}{N}\sum_{i=1}^N Z_i(u).$$

Because $\langle X_i, u\rangle$ is K-sub-Gaussian, the square is sub-exponential: there is an absolute constant C such that

$$\|Z_i(u)\|_{\psi_1} \leqslant CK^2.$$

Bernstein's inequality for sub-exponential random variables then gives constants $c, C > 0$ such that for all $t \geqslant 0$,

$$\mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^N Z_i(u)\right| \geqslant t\right\} \leqslant 2\exp\left[-cN\min\left(\frac{t^2}{K^4}, \frac{t}{K^2}\right)\right]. \tag{1}$$

The two regimes (quadratic for small t, linear for large t) are exactly the usual Bernstein behavior.

2. Union bound over a net. Let $\mathcal{N}$ be an ε-net of S^{d-1} with $\varepsilon \in (0, 1/4)$. Applying (1) to each $u \in \mathcal{N}$ and union bounding gives

$$\mathbb{P}\left\{\max_{u\in\mathcal{N}}\left|u^\top(\widehat{\Sigma} - I_d)u\right| \geqslant t\right\} \leqslant 2|\mathcal{N}|\exp\left[-cN\min\left(\frac{t^2}{K^4}, \frac{t}{K^2}\right)\right]. \tag{2}$$

This is the exact point where metric entropy enters:

$$\log|\mathcal{N}| \lesssim d\log(1/\varepsilon)$$

is the "price of uniformity."

3. Lift back from the net to the whole sphere. Apply Lemma 2.2 with $A = \widehat{\Sigma} - I_d$:

$$\|\widehat{\Sigma} - I_d\| \leqslant \frac{1}{1-2\varepsilon}\max_{u\in\mathcal{N}}|u^\top(\widehat{\Sigma} - I_d)u|.$$

Choosing $\varepsilon = 1/4$ makes $(1-2\varepsilon)^{-1} = 2$, so

$$\|\widehat{\Sigma} - I_d\| \leqslant 2\max_{u\in\mathcal{N}}|u^\top(\widehat{\Sigma} - I_d)u|. \tag{3}$$

Combining (2) and (3) yields a tail bound for $\|\widehat{\Sigma} - I_d\|$.

Theorem 3.1 (Sample covariance: isotropic sub-Gaussian case)**.** *Let $X \in \mathbb{R}^d$ be isotropic and satisfy $\|X\|_{\psi_2} \leqslant K$. Let $X_1, \dots, X_N$ be i.i.d. copies and let*

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

Then there exist absolute constants $c, C > 0$ such that for every $s \geqslant 0$,

$$\mathbb{P}\left\{\|\widehat{\Sigma} - I_d\| \geqslant CK^2\left(\sqrt{\frac{d+s}{N}} + \frac{d+s}{N}\right)\right\} \leqslant 2e^{-cs}.$$

Proof sketch. Take $\varepsilon = 1/4$, so $|\mathcal{N}| \leqslant 12^d$ (any bound of the form $(C/\varepsilon)^d$ works) and thus $\log|\mathcal{N}| \lesssim d$. Combine (2) and (3). Choose t so that the exponent

$$cN \min\left(\frac{t^2}{K^4}, \frac{t}{K^2}\right)$$

dominates $\log|\mathcal{N}| + s \lesssim d + s$. Solving yields $t \asymp K^2\sqrt{(d+s)/N}$ in the quadratic regime and $t \asymp K^2(d+s)/N$ in the linear regime, which combine (up to constants) into the displayed bound. □

Some comments are in order:

- *Sanity checks.* When $N \gg d$, the leading behavior is

  $$\|\widehat{\Sigma} - I_d\| \lesssim K^2\sqrt{\frac{d}{N}},$$

  i.e. "dimension over sample size." The second term becomes relevant only when N is not much larger than d.
- *Relative error.* Fix $\epsilon \in (0,1)$. If $N \gtrsim K^4(d+s)/\epsilon^2$, then the theorem gives $\|\widehat{\Sigma} - I_d\| \leqslant \epsilon$ with probability at least $1 - 2e^{-cs}$.
- *Expectation bound.* Integrating the tail bound (or repeating the proof with s as a random variable) yields

  $$\mathbb{E}\|\widehat{\Sigma} - I_d\| \leqslant CK^2\left(\sqrt{\frac{d}{N}} + \frac{d}{N}\right).$$

  We will often use this form when we only need a scale estimate rather than an explicit failure probability.

3.2 From isotropic to general covariance

Let $X \in \mathbb{R}^d$ have covariance $\Sigma = \mathbb{E}[XX^\top]$ (assume $\Sigma \succ 0$ for simplicity). Define the whitened vector $Y := \Sigma^{-1/2}X$, so $\mathbb{E}[YY^\top] = I_d$. If $Y_1, \ldots, Y_N$ are the whitened samples, then

$$\widehat{\Sigma}_Y = \frac{1}{N}\sum_{i=1}^N Y_i Y_i^\top = \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}.$$

Therefore

$$\widehat{\Sigma} - \Sigma = \Sigma^{1/2}(\widehat{\Sigma}_Y - I_d)\Sigma^{1/2}, \qquad \text{so} \qquad \|\widehat{\Sigma} - \Sigma\| \leqslant \|\Sigma\| \cdot \|\widehat{\Sigma}_Y - I_d\|.$$

So any isotropic bound transfers to the general case, up to scaling by $\|\Sigma\|$ (and the sub-Gaussian constant of Y, which depends on both X and Σ).

4 Perturbation theory

The main statistical reason we care about $\|\widehat{\Sigma} - \Sigma\|$ is that it controls eigenvalues and eigenvectors of the sample covariance, hence the accuracy of principal component analysis.

4.1 Eigenvalues

Lemma 4.1 (Weyl)**.** *For symmetric $d \times d$ matrices A, B,*

$$\max_{1 \leqslant i \leqslant d} |\lambda_i(A) - \lambda_i(B)| \;\leqslant\; \|A - B\|.$$

Thus, once we have an operator-norm bound on $\widehat{\Sigma} - \Sigma$, all eigenvalues are uniformly close.

4.2 Eigenvectors

Eigenvectors are only stable when there is a gap separating the relevant part of the spectrum.

Theorem 4.2 (Davis–Kahan for top-k subspaces)**.** *Let A, B be symmetric $d \times d$ matrices. Let P_A be the orthogonal projector onto* $\operatorname{span}\{v_1(A), \ldots, v_k(A)\}$*. Assume there is an eigengap at k for A,*

$$\delta \;:=\; \lambda_k(A) - \lambda_{k+1}(A) \;>\; 0,$$

and that $\|A - B\| \leqslant \delta/2$. Let P_B be the projector onto $\operatorname{span}\{v_1(B), \ldots, v_k(B)\}$*. Then*

$$\|P_A - P_B\| \;\leqslant\; \frac{2\|A - B\|}{\delta}.$$

The eigengap condition is not a technicality: when eigenvalues are nearly tied, an arbitrarily small perturbation can rotate the corresponding eigenvectors by a large angle. The stable object is the *subspace* associated with a separated spectral cluster.

To see that sample PCA $\approx$ population PCA, apply Theorem 4.2 with $A = \Sigma$ and $B = \widehat{\Sigma}$. If $\|\widehat{\Sigma} - \Sigma\| \ll \delta$, then the leading k-dimensional PCA subspace of the sample is close to the population PCA subspace.

5 Lookahead

This lecture used the most direct supremum strategy: discretize the sphere with a single net, apply a union bound, and then lift back to the continuum. It is a reliable baseline method, and it already yields the correct $N \sim d$ sample complexity for operator-norm covariance estimation under sub-Gaussian assumptions.

In the next lectures, we will refine and complement this picture. On the probability side, we will develop matrix concentration tools that can control $\left\|\sum_i Y_i\right\|$ directly without discretizing the sphere. On the geometry side, we will meet multi-scale approximations (chaining and entropy integrals) that become essential when the supremum is over a *structured* subset of directions (e.g. sparse directions, low-rank structure, restricted eigenvalues). Finally, as we transition toward random matrix theory, we will study not only operator-norm errors but also the bulk and edge behavior of the sample covariance spectrum, where new universal laws appear.

Source material

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

References

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