Lecture 8: Gaussian polynomials, and overview of robust covariance estimation

October 22, 2025

We have so far focused exclusively on the task of robust mean estimation. In this lecture, we will branch out into a somewhat more sophisticated task, namely, that of robust *covariance* estimation, specifically for Gaussian data. Here, we consider the setting where we are given ε-corrupted samples from a d-dimensional Gaussian $\mathcal{N}(\mu, \Sigma)$ with unknown covariance Σ, and our goal is to robustly recover Σ, in the appropriate sense. For now, let's also assume that we know the mean μ, and so without loss of generality let $\mu = 0$.

Remark 0.1. There is a simple trick that lets us reduce from the case of unknown μ to the case of $\mu = 0$ while at most doubling the value of ε. Do you see it?

Throughout these lectures, let $\mathbf{S}_d$ denote the set of $d \times d$ symmetric matrices.

1 Mahalanobis distance

Recall from Lecture 2 that the key criteria that will determine how well one can hope to recover Σ is the relationship between total variation distance between $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$, and the appropriate notion of distance between Σ_1 and Σ_2. On the homework, you're asked to show the following:

Theorem 1.1. *Let $\Sigma_1, \Sigma_2 \succ 0$. Then*

$$\mathrm{d_{TV}}(\mathcal{N}(0, \Sigma_1), \mathcal{N}(0, \Sigma_2)) = \Theta\left(\min\left(\left\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - I\right\|_F, 1\right)\right) . \tag{1}$$

Here $\|\cdot\|_F$ denotes the Frobenius norm of the matrix, and for $M \succeq 0$, we let $M^{-1/2}$ denote the unique positive semi-definite square root of the matrix M. We pause here to parse this esoteric-looking expression somewhat. First, observe that if $\Sigma_1 = I$, then the key part of the expression on the RHS of (1) is just $\|\Sigma_2 - I\|_F$, i.e. the Frobenius norm difference between the two covariance matrices. Indeed, when Σ_2 is close-ish to identity (i.e. if all of its eigenvalues are very close to 1), it turns out that this exactly captures the total variation distance between the two Gaussians.

However, this expression becomes very lossy when Σ_1 is ill-conditioned. This is because we should expect that for $\mathcal{N}(0, \Sigma_2)$ to be close to $\mathcal{N}(0, \Sigma_1)$, the covariances should be *multiplicatively close along every direction*—even in one dimension. Formally, this is because TV distance is *affine invariant*: if I have two random variables X, Y so that $\mathrm{d_{TV}}(X, Y) = \alpha$, and I apply any bijective linear transform $x \mapsto Ax$ for some non-degenerate matrix A, one can easily verify that $\mathrm{d_{TV}}(AX, AY) = \alpha$ as well (why?). So, in particular, if $X \sim \mathcal{N}(0, \Sigma_1)$ and $Y \sim \mathcal{N}(0, \Sigma_2)$, if I apply the map $x \mapsto \Sigma_1^{-1/2}x$ to both random variables, then we have that $\Sigma_1^{-1/2}X \sim \mathcal{N}(0, I)$ and $\Sigma_1^{-1/2}Y \sim \mathcal{N}(0, \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})$, and the TV distance between these two random variables is the same as the TV distance between X and Y, by what we asserted above, the TV distance between these two transformed random variables is exactly the Frobenius norm distance between their covariance matrices, which results in the expression above, which can be interpreted as a "preconditioned" Frobenius norm distance between the two covariance matrices.

This notion of preconditioned closeness will come up repeatedly, and is known as the *Mahalanobis distance* between the two covariance matrices [1], after the great Indian statistician P.C. Mahalanobis.

In general, given any positive semidefinite matrix Σ, we define its associated Mahalanobis norm to be $\|A\|_\Sigma = \|\Sigma^{-1/2} A \Sigma^{-1/2}\|_F$. With this, we can represent the RHS of (1) somewhat more concisely as $\min(\|\Sigma_1 - \Sigma_2\|_{\Sigma_1}, 1)$. With this alongside Theorem 1.1, we can also now state the natural notion of robustly learning a covariance matrix:

Problem 1.2. *Let $\Sigma \succ 0$ be an arbitrary positive definite matrix, and let $\varepsilon > 0$. Given an ε-corrupted set of samples from $\mathcal{N}(0, \Sigma)$, find $\widehat{\Sigma}$ so that with high probability, $\|\Sigma - \widehat{\Sigma}\|_\Sigma$ is small.*

The spoiler of the next couple of lectures will be that it turns out there is an efficient algorithm, once again based on filtering, that will be able to achieve error $O(\varepsilon \log 1/\varepsilon)$. This is what we will build up to. Moreover, on the homework, you will see that by combining this routine with a slight modification of the filtering for robust mean estimation, this will actually give an algorithm that robustly learns an arbitrary Gaussian to total variation distance $O(\varepsilon \log 1/\varepsilon)$.

2 From matrices to polynomials

For now, let's focus on the setting where the unknown covariance Σ is guaranteed to be close to the identity: i.e. let us assume that $0.5I \preceq \Sigma \preceq 2I$; that is, all of Σ's eigenvalues are between $1/2$ and 2. This is the setting where the Mahalanobis distance is essentially just the Frobenius distance up to constant factors, and so our job is more or less equivalent to learning Σ to good Frobenius error.

A first naive attempt at covariance estimation is to observe that in a formal sense, covariance estimation is just a special case of robust mean estimation! Indeed, if we define the matrix $Y_i = X_i X_i^\top$, then the whole goal of covariance estimation is to obtain an accurate estimate of Y_i given ε-corrupted samples. If we think of Y_i as a $\mathbb{R}^{d^2}$ length vector, then this is exactly a question of mean estimation in this higher dimensional setting. So what goes wrong?

The main problem is that the covariance of Y_i (which, recall, is really the covariance *of the covariance* of X_i) is kind of nasty to work with. In particular, the covariance of Y_i will depend on the unknown covariance Σ. In contrast, for mean estimation, the covariance of the data clearly does not depend on the mean.

To get around this issue, it will turn out that, in this setting, it will be convenient to interpret matrices as quadratic polynomials, in the following sense. Let S be a dataset of n points, and let $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ be the empirical second moment of this dataset. By the self-duality of Frobenius norm, we have that

$$\|\Sigma - \widehat{\Sigma}\|_F = \sup_{A \in \mathbf{S}_d : \|A\|_F = 1} \langle A, \widehat{\Sigma} - \Sigma \rangle \tag{2}$$

$$= \sup_{A \in \mathbf{S}_d : \|A\|_F = 1} \left(\frac{1}{n}\sum_{i=1}^n \langle A, X_i X_i^\top \rangle - \mathop{\mathbb{E}}_{X \sim \mathcal{N}(0,\Sigma)} \langle A, XX^\top \rangle \right) \tag{3}$$

$$= \sup_{A \in \mathbf{S}_d : \|A\|_F = 1} \left(\frac{1}{n}\sum_{i=1}^n p(X_i) - \mathop{\mathbb{E}}_{X \sim \mathcal{N}(0,\Sigma)} p(X) \right), \tag{4}$$

where $p(x) = \langle A, xx^\top \rangle = x^\top A x$ is a homogeneous, quadratic polynomial in x. In other words, the Frobenius norm difference between the Σ and $\widehat{\Sigma}$ is large if and only if there is some quadratic polynomial whose empirical expectation differs dramatically from the true expectation under $\mathcal{N}(0, \Sigma)$. This is the degree-2 analog of the same phenomena for robust mean estimation, where the empirical mean differed from the true mean if and only if there was some linear function whose expectation differed dramatically. But whereas in the linear case we didn't have to explicitly take the perspective of linear functions, and could just directly work with matrices, in this case it will be somewhat helpful to take the functional perspective.

Before we do so, it will be necessary to establish some properties of how these polynomials behave under Gaussian input. First, observe that there is are canonical mappings between $\mathbf{S}_d$, homogeneous quadratic polynomials, and vectors in $\mathbb{R}^{\binom{d+1}{2}}$:

- For any matrix $A \in \mathbf{S}_d$, we let $p_A(x) = x^\top Ax$ be its *associated polynomial*. Alternatively, for any homogeneous quadratic polynomial $p : \mathbb{R}^d \to \mathbb{R}$ with form $p(x) = \sum_{i,j} a_{ij} x_i x_j$, we let its *associated matrix* $A \in \mathbf{S}_d$ be given by $A_{ij} = \frac{1}{2}(a_{ij} + a_{ji})$, for all i, j.
- Let e_{ij} denote the $d \times d$ matrix which is 1 in the (i,j)-th position and 0 elsewhere. Note that the matrices $\{e_{ii}\}_{i=1}^d \cup \{\frac{1}{\sqrt{2}}(e_{ij} + e_{ji})\}_{i<j}$ form a orthonormal basis for $\mathbf{S}_d$. We let $L_d : \mathbf{S}_d \to \mathbb{R}^{\binom{d+1}{2}}$ be the bijective linear mapping which sends these vectors to the standard basis elements of $\mathbb{R}^{\binom{d+1}{d}}$. We call this operator the *flattening* operator, as its effect is to reshape a matrix into a corresponding vector.

It turns out that we can relate the norms of these objects in natural ways:

Lemma 2.1. *Let $A \in \mathbf{S}_d$. Then:*

$$\|A\|_F^2 = \|L_d(A)\|_2^2 = \frac{1}{2} \operatorname*{Var}_{X \sim \mathcal{N}(0,I)} p_A(X) \, .$$

Proof. That $\|A\|_F = \|L_d(A)\|_2$ follows immediately from definition. To prove the remaining equality, let $A = \sum_{i=1}^d \lambda_i v_i v_i^\top$ be the diagonalization of A. Then,

$$p_A(x) = \sum_{i=1}^d \lambda_i \langle v_i, x \rangle^2 \, .$$

In particular, since the v_i are orthogonal, if $X \sim \mathcal{N}(0, I)$, then the random variables $Y_i = \langle v_i, X \rangle$ are independent, standard normal Gaussians, and hence

$$\operatorname*{Var}_{X \sim \mathcal{N}(0,I)} p_A(X) = \sum_{i=1}^d \lambda_i^2 \operatorname*{Var}_{X \sim \mathcal{N}(0,I)} \langle v_i, X \rangle^2 \tag{5}$$

$$= \sum_{i=1}^d \lambda_i^2 \operatorname{Var} Y_i^2 = 2 \sum_{i=1}^d \lambda_i^2 = 2\|A\|_F^2 \, , \tag{6}$$

where here we are using that the variance of a chi-squared random variable is 2. □

As a consequence of this is as follows. Observe that $X \sim \mathcal{N}(0, \Sigma)$ is equivalent to saying that $X = \Sigma^{1/2} Y$ for $Y \sim \mathcal{N}(0, I)$. Therefore,

$$p_A(X) = X^\top AX = Y^\top \Sigma^{1/2} A \Sigma^{1/2} Y = p_{\Sigma^{1/2} A \Sigma^{1/2}}(Y) \, ,$$

and so

$$\operatorname*{Var}_{X \sim \mathcal{N}(0,\Sigma)} p_A(X) = 2 \cdot \|\Sigma^{1/2} Y \Sigma^{1/2}\|_F^2 \, . \tag{7}$$

3 Spectral signatures for quadratic polynomials

Now, let's see how to try to adapt the algorithmic strategy for mean estimation to this language. We will first need a natural polynomial analog of the notion of ε-goodness. Formally:

Definition 3.1. We say a set of points S is ε-good with respect to some covariance Σ if for all polynomials p_A with $\|A\|_F = 1$:

- We have that

$$\left| \frac{1}{|S|} \sum_{i \in S} p_A(X_i) - \operatorname{tr}(A\Sigma) \right| \lesssim \varepsilon \log 1/\varepsilon \quad \text{and} \quad \left| \frac{1}{|S|} \sum_{i \in S} p_A(X_i)^2 - 2\|\Sigma^{1/2} A \Sigma^{1/2}\|_F^2 \right| \lesssim \varepsilon \log^2 1/\varepsilon \, ,$$

- For all subsets $T \subset S$ with $|T| = \varepsilon n$, we have that

$$\left|\frac{1}{|T|}\sum_{i\in T} p_A(X_i)\right| \lesssim \log 1/\varepsilon \text{ and } \left|\frac{1}{|T|}\sum_{i\in T} p_A(X_i)^2\right| \lesssim \log^2 1/\varepsilon \,.$$

Note that in the second set of conditions, we're not centering around the expectations, but it doesn't matter since the bound we want is so large relative to the expectations (which are always at most constant). This is an almost direct analog the notion of goodness for mean estimation, and one can also show that this is satisfied whp by a sufficiently large set of samples from $\mathcal{N}(0, \Sigma)$:

Lemma 3.1 (see e.g. Appendix B in [2]). *Let $\varepsilon, \delta > 0$, and let S be a set of n independent draws from $\mathcal{N}(0, \Sigma)$, for some Σ satisfying $\|\Sigma\|_2 \lesssim 1$, and where*

$$n \gtrsim \frac{d^2 \operatorname{poly}\log(d/\delta)}{\varepsilon^2} \,.$$

Then with probability $1 - \delta$, S is ε-good with respect to Σ.

Let us see how this can be used to obtain a quadratic version of the spectral signatures we saw in previous lectures. Let $\widehat{\Sigma}$ denote the empirical covariance of the corrupted dataset. In fact, let's consider the setting for now where $\widehat{\Sigma} = I$, but where $\|\Sigma - I\|_F = \eta$ for some η sufficiently large.

How can this happen? By the above, this means there is some p_A with $\|A\|_F = 1$ so that

$$\frac{1}{n}\sum_{i\in S} p_A(X_i) - \mathop{\mathbb{E}}_{X\sim N(0,\Sigma)} p_A(X) = \eta \,.$$

Once again decomposing $S = S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$, ε-goodness implies that

$$\left|\frac{1}{n}\sum_{i\in S_g} p_A(X_i) - \mathop{\mathbb{E}}_{X\sim N(0,\Sigma)} p_A(X)\right| \lesssim \varepsilon \log 1/\varepsilon$$

$$\left|\frac{1}{n}\sum_{i\in S_r} p_A(X_i)\right| \lesssim \varepsilon \log 1/\varepsilon \,,$$

and so, assuming $\eta \gg \varepsilon \log 1/\varepsilon$, we obtain that

$$\frac{1}{|S_{\text{bad}}|}\sum_{i\in S_{\text{bad}}} p_A(X_i) \gtrsim \frac{\eta}{\varepsilon} \,,$$

and so by Jensen's inequality, we have that

$$\frac{1}{|S_{\text{bad}}|}\sum_{i\in S_{\text{bad}}} p_A(X_i)^2 \gtrsim \frac{\eta^2}{\varepsilon^2} \gg \log^2 \varepsilon \,.$$

We claim that this will cause p_A to have noticeably large variance. Since the empirical dataset has covariance $\widehat{\Sigma} = I$, our best guess of the true variance of any polynomial will be $2\|A\|_F^2 = 2$, so it seems reasonable to hope that

$$\frac{1}{|S|}\sum_{i\in S} p_A(X)^2 - 2$$

will be large. Indeed, by ε-goodness, we have that:

$$\begin{aligned}\frac{1}{n}\sum_{i\in S_{\text{good}}\setminus S_r} p_A^2(X) &= \frac{1}{n}\sum_{i\in S_{\text{good}}} p_A^2(X) - \frac{1}{n}\sum_{i\in S_r} p_A^2(X) \\ &= 2\|\Sigma^{1/2} A \Sigma^{1/2}\|_F^2 \pm O(\varepsilon \log 1/\varepsilon) \,.\end{aligned}$$

Therefore,

$$\frac{1}{|S|}\sum_{i\in S} p_A(X)^2 - 1 = \frac{1}{|S_{\text{bad}}|}\sum_{i\in S} p_A(X)^2 + \sum_{i\in S_{\text{good}}} p_A(X)^2 - 2$$
$$\gg \frac{\eta^2}{\varepsilon} + 2 - 2\|\Sigma^{1/2} A \Sigma^{1/2}\|_F^2 \pm O(\varepsilon \log 1/\varepsilon) \; .$$

Since $\eta \gg \varepsilon \log 1/\varepsilon$, the error term is bounded, but now there's one additional term, namely, the error because the true covariance is Σ, not I. However, it turns out that this can also be controlled, as long as Σ is not too large:

Lemma 3.2. *Suppose that $\Sigma \preceq 2I$. Then, we have that*

$$1 - \|\Sigma^{1/2} A \Sigma^{1/2}\|_F^2 \gtrsim -\|\Sigma - I\|_2 \geq -\|\Sigma - I\|_F \; .$$

Proof. The last inequality is trivial. To show the first inequality, let $\delta = \|\Sigma - I\|_2$. Then $\Sigma \preceq (1+\min(1,\delta))I$. Thus,

$$\|\Sigma^{1/2} A \Sigma^{1/2}\|_F^2 = \text{tr}\left(\Sigma^{1/2} A \Sigma A \Sigma^{1/2}\right)$$
$$\leq (1+\min(1,\delta))^2 \text{tr}(A^2) \leq 1 + 3\delta \; ,$$

from which the claim follows. □

Combining this with the previous calculation, we obtain that

$$\frac{1}{|S|}\sum_{i\in S} p_A(X)^2 - 1 \gg \frac{\eta^2}{\varepsilon} - \eta \gtrsim \frac{\eta^2}{\varepsilon} \; .$$

We have essentially proven the following "spectral signatures" lemma:

Lemma 3.3. *Let ε be sufficiently small, let $\Sigma \preceq 2I$, let S_{good} be an ε-good set of points with respect to Σ, and let $S = S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$ be an ε-corruption of S_{good} satisfying*

$$\frac{1}{n}\sum_{i\in S} X_i X_i^\top = I \; .$$

Then, we have that

$$\|\Sigma - I\|_F \lesssim \varepsilon \log 1/\varepsilon + \sqrt{\varepsilon \left(\max_{\|A\|_F = 1} \frac{1}{n}\sum_{i\in S} p_A(X_i)^2 - 2 \right)} \; .$$

4 Obtaining efficient algorithms

Given this lemma, it is natural to pursue the following algorithmic paradigm, in analogy to what we've done so far:

- Find an A with $\|A\|_F = 1$ that maximizes $\frac{1}{n}\sum_{i\in S} p_A(X_i)^2$.
- Define scores $\tau_i = p_A(X_i)^2$.
- Perform (weighted) iterative filtering with these scores.

Indeed, this more or less works, but there are several complications:

Making things isotropic Let's start with the easiest problem, which is that this lemma assumes that

$$\frac{1}{n}\sum_{i\in S} X_i X_i^\top = I \ ,$$

i.e. the empirical samples are in *isotropic position*. But this is easy enough to fix: if we let $\widehat{\Sigma} = \frac{1}{n}\sum_{i\in S} X_i X_i^\top$, we can just define $\tilde{X}_i = \widehat{\Sigma}^{-1/2} X_i$. In general, this is known as *whitening* the data. Then, these points will satisfy the desired property. Additionally, one can check that this will not alter the ε-goodness properties of S_{good} in any meaningful way.

Spectral upper bound The lemma also assumes that $\Sigma \preceq 2I$. But it turns out that this whitening procedure also guarantees this. This is because one can show that with very high probability, $(1-\varepsilon)\Sigma \preceq \widehat{\Sigma}$. Since the whitened points are of the form $\widehat{\Sigma}^{-1/2} X_i$, the new effective covariance of the good points is

$$\widehat{\Sigma}^{-1/2}\Sigma\widehat{\Sigma}^{-1/2} \preceq \frac{1}{1-\varepsilon} I \ ,$$

which is stronger than we needed.

You might worry that the whitened points are no longer independent Gaussian points—and this is true, because $\widehat{\Sigma}$ depends on the empirical datapoints. But crucially, we are only using a set of deterministic conditions on the points (namely, ε-goodness), and this will not be violated by rescaling.

Efficiently finding p_A Finally, we come to the issue that it's not *a priori* clear how to find A which maximizes this quadratic form. However, this is where we can turn back to the linear algebraic interpretation. Observe that $\langle L_d(A), L_d(X_i X_i^\top)\rangle = p_A(X_i)$, and so $\langle L_d(A), L_d(X_i X_i^\top)\rangle^2 = p_A(X_i)^2$. Thus,

$$\frac{1}{n}\sum_{i\in S} p_A(X_i^2) = \frac{1}{n}\sum_{i=S}\langle L_d(A), L_d(X_i X_i^\top)\rangle^2 = L_d(A)^\top \left(\frac{1}{n}\sum_{i\in S} L_d(X_i X_i^\top) L_d(X_i X_i^\top)^\top\right) L_d(A) \ .$$

Moreover, the set of vectors $L_d(A)$ with $\|A\|_F = 1$ is just the set of unit vectors in $R^{\binom{d+1}{2}}$! Thus, if we define $Y_i = L_d(X_i X_i^\top) \in \mathbb{R}^{\binom{d+1}{2}}$, we have that

$$\max_{\|A\|_F=1} \frac{1}{n}\sum_{i\in S} p_A(X_i^2) = \max_{\|v\|=1} v^\top \left(\frac{1}{n}\sum_{i\in S} Y_i Y_i^\top\right) v \ ,$$

and so we can find the maximizing A just be finding the top eigenvector of this $R^{\binom{d+1}{2}\times\binom{d+1}{2}}$-sized matrix, and given the eigenvector, by Lemma 2.1 we can also find the associated matrix.

References

[1] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 80:S1–S7, 1918.

[2] Jerry Li and Guanghao Ye. Robust gaussian covariance estimation in nearly-matrix multiplication time. *Advances in Neural Information Processing Systems*, 33:12649–12659, 2020.