Lecture 4: Spectral signatures and efficient certifiability

October 2, 2025

1 Introduction

Last time we saw some information-theoretic criteria which give inefficient estimators in high dimensions which achieve the correct asymptotic error rates for robust mean estimation. In the next couple of lectures, we'll develop another machinery to give matching polynomial-time estimators, in some regimes. The basic idea underlying all of these algorithmic techniques will be the following:

Corruptions to the first moment of a nice distribution caused by a small fraction of outliers necessarily result in larger than expected eigenvalues of higher moments.

We will refer to this phenomenon as a *spectral signature*, and we'll see two concrete instantiations of this in the next several classes.

The presence of such large eigenvalues is a very useful algorithmic tool, and indeed, this geometric fact is the only thing we need to get a simple, polynomial time algorithm that achieves the correct asymptotic error. This is because it does two things. First, the contrapositive of this statement gives a polynomial-time checkable certificate for whether or not the mean has been corrupted by outliers: if the top eigenvalue is small, then the mean has not been corrupted. Second, it gives a direction in which the outliers must have undue influence: namely, the associated eigenvector. Therefore, in the case when the eigenvalue is large, it also gives us a way to make progress by removing outliers.

2 Spectral signatures with bounded second moment

Let's first consider what is arguably the simplest setting in which this phenomena appears, namely, in distributions with bounded second moment. In this case, we have the following critical lemma:

Lemma 2.1 (Spectral signatures with bounded second moment)**.** *Let $\varepsilon < 1$. Let $S = S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$ be of size n. Let $\mu_g = \frac{1}{n}\sum_{i \in S_{\text{good}}} X_i$, and assume that*

$$\frac{1}{n}\sum_{i \in S_{\text{good}}} (X_i - \mu_g)(X_i - \mu_g)^\top \preceq CI\,. \tag{1}$$

If we let

$$\widehat{\mu} = \frac{1}{n}\sum_{i \in S} X_i \qquad \text{and} \qquad \widehat{\Sigma} = \frac{1}{n}\sum_{i \in S} (X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top$$

denote the empirical mean and covariance of the dataset S respectively, then we have

$$\|\mu_g - \widehat{\mu}\|_2 \le \frac{1}{1-\varepsilon}\left(\sqrt{2C\varepsilon} + \sqrt{\varepsilon\left\|\widehat{\Sigma}\right\|_2}\right)\,. \tag{2}$$

In particular, in the regime where ε is bounded away from 1 and $C = O(1)$, the RHS of (2) is at most $\lesssim \sqrt{\varepsilon} + \sqrt{\varepsilon \left\|\widehat{\Sigma}\right\|_2}$. Notice that is exactly an empirical notion of spectral signature as defined above: the largest eigenvalue of $\widehat{\Sigma}$ can be lower bounded by the mean shift caused by the outliers. Notice moreover that all quantities on the RHS of (2) are either known, or at least efficiently computable. Since μ_g is not known, this provides a very useful proxy for checking if the empirical mean has been corrupted.

In this class, we will actually prove a more general version of this lemma. Before we can do so, we need to establish a bit of terminology.

3 The set $\mathscr{W}_{S,\varepsilon}$, and spectral signatures for soft subsets

We will actually prove a slightly more general version of this lemma, as this will be very algorithmically useful for us later on. Our algorithms will work with *non-uniform* weightings of the points, which will correspond to our beliefs on how good or bad an individual point is. We will denote the set of allowable weights by Γ_S:

$$\Gamma_S = \left\{ w \in \mathbb{R}^S : \sum_{i \in S} w_i \leq 1 \text{ and } w_i \geq 0 \text{ for all } i \in S \right\} . \tag{3}$$

For any $w \in \Gamma_S$, if $\|w\|_1 = 1$, this is simply another probability distribution over the points in S, and it is easy to define the empirical mean and covariance of S with respect to these weights. For us, it will be useful to allow weights whose sum is less than one. However, we have to be slightly careful about how we define the empirical mean and covariance with respect to general weights in Γ_S. Specifically, we will set:

$$\mu(w) = \sum_{i \in S} \frac{w_i}{\|w\|_1} X_i \qquad \text{and} \qquad \Sigma(w) = \sum_{i \in S} w_i \left(X_i - \mu(w)\right)\left(X_i - \mu(w)\right)^\top . \tag{4}$$

Given two weights $w, w' \in \Gamma_S$, we will also say that $w \leq w'$ if $w_i \leq w'_i$ for all $i \in S$. If $T \subseteq S$, we also let

$$\mu(T) = \mu\left(\frac{1}{|T|}\mathbb{I}_T\right) \qquad \text{and} \qquad \Sigma(T) = \Sigma\left(\frac{1}{|T|}\mathbb{I}_T\right) \tag{5}$$

be the empirical mean and covariance of the uniform distribution over T.

3.1 The set $\mathscr{W}_{S,\varepsilon}$

Initially our prior on every point in S is that each point is equally valid. As our algorithm proceeds, in an ideal world, we would converge on the uniform distribution over the subset $S_{\text{good}} \cap S$, i.e. we uniformly value each remaining inlier, and put no mass on any outlier. Since a priori we don't know which samples are good or bad, the set of possible weights we should search over should include the set of weights which are uniform over all sets of size $|S_{\text{good}} \cap S| = (1-\varepsilon)n$.

Let's introduce some notation to capture this more formally. For any set $T \subseteq S$, let $w(T) \in \Gamma_S$ denote the set of weights which are uniform over T, i.e.

$$w(T)_i = \begin{cases} \frac{1}{n} & \text{if } i \in T; \\ 0 & \text{otherwise.} \end{cases}$$

With this notation, the above discussion indicates that our set of allowable weights should include $w(T)$ for all $T \subset S$ with $|T| = (1-\varepsilon)n$. However, this set of weights is not convex, and it will be natural in the course of developing our algorithms to consider all convex combinations of these weights. It turns out that the following convex set captures the set of $w(T)$, while maintaining a number of useful properties for us:

$$\mathscr{W}_{S,\varepsilon} = \{ w \in \Gamma_S : w \leq w(S) \qquad \text{and} \qquad \|w(S) - w\|_1 \leq \varepsilon \} . \tag{6}$$

As a special case, note that the uniform set of weights over S is in $\mathscr{W}_{S,\varepsilon}$, i.e. $w(S) \in \mathscr{W}_{S,\varepsilon}$. We also note the following fact about the width of $\mathscr{W}_{S,\varepsilon}$:

Lemma 3.1. *For all $w, w' \in \mathscr{W}_{S,\varepsilon}$, we have $\|w - w'\|_1 \le 2\varepsilon$.*

Proof. By triangle inequality, we have $\|w - w'\|_1 \le \|w - w(S)\|_1 + \|w' - w(S)\|_1 \le 2\varepsilon$. □

3.2 Spectral signatures for soft subsets

We now return to Lemma 2.1. We will show the following generalization of it, which can be expressed quite concisely with our new notation:

Lemma 3.2. *Let ε, S, μ_g be as in Lemma 2.1. Assume that (1) holds. Then, for all $w \in \mathscr{W}_{S,\varepsilon}$, we have*

$$\|\mu_g - \mu(w)\|_2 \le \frac{1}{1-\varepsilon}\left(\sqrt{2C\varepsilon} + \sqrt{\varepsilon \left\|\Sigma(w)\right\|_2}\right) . \tag{7}$$

As discussed above, this lemma is a strict generalization of Lemma 2.1 so it follows directly from this lemma. Before we prove the lemma, we first note the following elementary inequality:

Fact 3.3 (special case of Hölder's inequality)**.** *Let $a_1, \ldots, a_m, b_1, \ldots, b_m$ be arbitrary, and let $w \in \Gamma_{[m]}$. Then*

$$\left(\sum_{i=1}^m w_i a_i b_i\right)^2 \le \left(\sum_{i=1}^m w_i a_i^2\right)\left(\sum_{i=1}^m w_i b_i^2\right) .$$

Proof of Lemma 3.2. We begin with the following sequence of equalities. We adopt the convention that $w_i = 0$ for all $i \in S_{\text{good}} \setminus S$.

$$\begin{aligned}
\|w\|_1 \cdot \|\mu(w) - \mu_g\|_2^2 &= \|w\|_1 \cdot \langle \mu(w) - \mu_g, \mu(w) - \mu_g \rangle \\
&= \sum_{i \in S} w_i \cdot \langle \mu(w) - \mu_g, X_i - \mu_g \rangle \\
&= \sum_{i \in S_{\text{good}}} \frac{1}{n} \langle \mu(w) - \mu_g, X_i - \mu_g \rangle + \sum_{i \in S_{\text{bad}}} w_i \langle \mu(w) - \mu_g, X_i - \mu_g \rangle \\
&\quad + \sum_{i \in S_{\text{good}}} \left(w_i - \frac{1}{n}\right) \langle \mu(w) - \mu_g, X_i - \mu_g \rangle \\
&= \sum_{i \in S_{\text{bad}}} w_i \langle \mu(w) - \mu_g, X_i - \mu_g \rangle + \sum_{i \in S_{\text{good}}} \left(w_i - \frac{1}{n}\right) \langle \mu(w) - \mu_g, X_i - \mu_g \rangle \\
&= \sum_{i \in S_{\text{bad}}} w_i \langle \mu(w) - \mu_g, X_i - \mu(w) \rangle + \left(\sum_{i \in S_{\text{bad}}} w_i\right) \cdot \|\mu(w) - \mu_g\|_2^2 \\
&\quad + \sum_{i \in S_{\text{good}}} \left(w_i - \frac{1}{n}\right) \langle \mu(w) - \mu_g, X_i - \mu_g \rangle .
\end{aligned}$$

Rearranging yields that

$$\begin{aligned}
\left(\|w\|_1 - \left(\sum_{i \in S_{\text{bad}}} w_i\right)\right) \cdot \|\mu_g - \mu(w)\|_2^2 &= \sum_{i \in S_{\text{bad}}} w_i \langle \mu(w) - \mu_g, X_i - \mu(w) \rangle \\
&\quad + \sum_{i \in S_{\text{good}}} \left(w_i - \frac{1}{n}\right) \langle \mu(w) - \mu_g, X_i - \mu_g \rangle .
\end{aligned} \tag{8}$$

We now bound the two terms separately. We first note that

$$\left(\sum_{i\in S_{\text{bad}}} w_i \langle \mu(w)-\mu_g, X_i-\mu(w)\rangle\right)^2 \le \left(\sum_{i\in S_{\text{bad}}} w_i\right)\cdot\left(\sum_{i\in S_{\text{bad}}} w_i\cdot\langle \mu(w)-\mu_g, X_i-\mu(w)\rangle^2\right) \tag{9}$$

$$\le \varepsilon\cdot\sum_{i\in S} w_i\cdot\langle \mu(w)-\mu_g, X_i-\mu(w)\rangle^2 \tag{10}$$

$$= \varepsilon\cdot(\mu(w)-\mu_g)^\top\Sigma(w)(\mu(w)-\mu_g) \tag{11}$$

$$\le \varepsilon\|\Sigma(w)\|_2\cdot\|\mu(w)-\mu_g\|_2^2\ . \tag{12}$$

Here (9) follows from Fact 3.3 and (12) follows from the definition of spectral norm. Taking square roots yields that

$$\left|\sum_{i\in S_{\text{bad}}} w_i\langle \mu(w)-\mu_g, X_i-\mu(w)\rangle\right| \le \|\mu(w)-\mu_g\|_2\cdot\sqrt{\varepsilon\|\Sigma(w)\|_2}\ . \tag{13}$$

We will now do something similar for S_{good}. Let $\delta_i = w_i - \frac{1}{n}$. We have

$$\left(\sum_{i\in S_{\text{good}}}\delta_i\langle \mu(w)-\mu_g, X_i-\mu_g\rangle\right)^2 \le \left(\sum_{i\in S_{\text{good}}} n\delta_i^2\right)\cdot\sum_{i\in S_{\text{good}}}\frac{1}{n}\langle \mu(w)-\mu_g, X_i-\mu_g\rangle^2$$

$$\le \left(\sum_{i\in S_{\text{good}}} n\delta_i^2\right)\cdot C\cdot\|\mu(w)-\mu_g\|_2^2\ , \tag{14}$$

where (14) follows from our assumed bound on the spectral norm of the empirical covariance. At the same time, we have

$$\sum_{i\in S_{\text{good}}} n\delta_i^2 = \sum_{i\in S_r}\frac{1}{n} + \sum_{i\in S_{\text{good}}\cap S} n\delta_i^2 \tag{15}$$

$$\le \varepsilon + \sum_{i\in S_{\text{good}}\cap S}|\delta_i| \tag{16}$$

$$\le \varepsilon + \|w-w(S)\|_1 \le 2\varepsilon\ , \tag{17}$$

where (15) follows since $\delta_i = -\frac{1}{n}$ for $i\in S_r$, (16) follows since $|n\delta_i|\le 1$, and (17) follows from the definition of $\mathscr{W}_{S,\varepsilon}$.

Combining these two bounds and taking square-roots yields that

$$\left|\sum_{i\in S_{\text{good}}}\left(w_i-\frac{1}{n}\right)\langle \mu(w)-\mu_g, X_i-\mu_g\rangle\right| \le \|\mu(w)-\mu_g\|_2\cdot\sqrt{2C\varepsilon}\ . \tag{18}$$

Plugging these bounds into (8), simplifying, and using the fact that $\sum_{i\in S_{\text{bad}}} w_i \le \varepsilon$, we obtain that

$$\|\mu_g-\mu(w)\|_2 \le \frac{1}{1-\varepsilon}\left(\sqrt{2C\varepsilon}+\sqrt{\varepsilon\|\Sigma(w)\|_2}\right)\ ,$$

as claimed. □

4 Population level spectral signatures

Lemma 2.1 says that spectral signatures allow us to approximate the empirical mean of the set of uncorrupted points, if they have bounded second moment. To translate this to a population level-statement, one would

naively need (1) that the empirical mean μ_g is close to the true mean of the distribution, and (2) the empirical covariance is bounded. While (1) is not hard to prove, (2) is somewhat more troublesome. Even if the distribution itself has bounded second moments, it is not true that we can necessarily expect any finite-sample concentration for covariance of the empirical distribution (why?).

However, one can get around this problem with a bit of a "hack": one can show that given a set of m i.i.d. draws from a distribution D, there will (with high probability) exist a $(1-\varepsilon)m$-sized subset of these points so that (1) their mean is $O(\sqrt{\varepsilon})$ close to the true mean, and (2) their covariance is bounded. Then, by simply adding the remaining εm points to S_{bad} (and accordingly doubling the ε-corruption of the original problem), we can apply Lemma 2.1. More formally:

Lemma 4.1. *Let $\varepsilon \in [0, 1/2)$, and let $\delta > 0$. Let D be a distribution over $\mathbb{R}^d$ with mean μ and covariance $\Sigma \preceq I$. Let $X_1, \ldots, X_m \sim D$ be independent. Then, there exist universal constants c, c' so that with probability $1 - \delta - \exp(-\Omega(\varepsilon m))$, there exists a set $S_{\text{good}} \subseteq [m]$ so that $|S| \geq (1-\varepsilon)m$ and:*

$$\|\widehat{\mu} - \mu\|_2 \leq \frac{1}{1-\varepsilon}\left(\sqrt{\frac{2d}{m\delta}} + \sqrt{c\varepsilon}\right)$$

$$\left\|\frac{1}{|S_{\text{good}}|}\sum_{i \in S_{\text{good}}} (X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top\right\|_2 \leq \frac{1}{1-\varepsilon}\frac{d(\log d + \log 2/\delta)}{c'\varepsilon m},$$

where $\widehat{\mu} = \frac{1}{|S_{\text{good}}|}\sum_{i \in S_{\text{good}}} X_i$.

We will prove this lemma in a future class. As a corollary of this, we obtain:

Corollary 4.2. *Let $m = \Omega(d \log d/\varepsilon)$. Then, in the setting of Lemma 4.1, with probability $99/100$, we have that*

$$\|\widehat{\mu} - \mu\|_2 \lesssim \sqrt{\varepsilon}, \qquad \text{and} \qquad \left\|\frac{1}{|S_{\text{good}}|}\sum_{i \in S_{\text{good}}} (X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top\right\|_2 \lesssim 1.$$

One can show that to obtain $O(\sqrt{\varepsilon})$ error with high probability, $\Omega(d/\varepsilon)$ samples are required, even without corruption. Up to log factors, the above corollary shows that, with the same number of samples, we can begin to apply the machinery of spectral signatures to get guarantees at the population level.