Lecture 6: Stronger spectral signatures for Gaussian datasets

October 16, 2019

In the past couple of lectures, we saw how to use spectral signatures to build the filtering algorithm for robust mean estimation when you assume that your dataset has bounded second moment. In the next couple of lectures, we'll show how to get stronger guarantees when your dataset satisfies stronger regularity conditions. This happens for instance when your data is Gaussian. Before we do so, it will be helpful to understand exactly what these stronger regularity conditions should look like.

1 Regularity of Gaussian datasets

At a high level, the key fact we will use is that small subsets of points cannot cause very large variance in any one direction. This is already true for the setting we considered before: if S is a set of points with empirical mean μ_g satisfying

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_g)(X_i - \mu_g)^\top \right\|_2 \leq 1 \,,$$

then for all unit vectors $v \in \mathbb{R}^d$, we have that $\frac{1}{n} \sum_{i=1}^{n} \langle X_i - \mu_g, v \rangle^2 \leq 1$, which in turn implies for any that $T \subseteq S$ of size $|T| = \varepsilon n$, we have that $\frac{1}{|T|} \sum_{i \in T} \langle X_i - \mu_g, v \rangle^2 \leq \frac{1}{\varepsilon}$, for all unit vectors v, or in other words,

$$\left\| \frac{1}{|T|} \sum_{i \in T} (X_i - \mu_g)(X_i - \mu_g)^\top \right\|_2 \leq \frac{1}{\varepsilon} \,. \tag{1}$$

However, when the data comes from a Gaussian distribution, we will be able to show a much stronger fact. Recall that given a set of points S, for any set $T \subset S$, we let $\mu(T)$ denote the empirical mean of the points in T.

Theorem 1.1. *Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, I)$, and let $\varepsilon \in (0, 1/2)$ and $\delta \in (0, 1)$. Then, with probability $1 - \delta$, we have that for all $T \subset S$ with $|T| = \varepsilon n$,*

$$\|\mu(T) - \mu\|_2 \leq \beta_1 \,, \qquad \text{and} \qquad \left\| \frac{1}{|T|} \sum_{i \in T} (X_i - \mu)(X_i - \mu)^\top - I \right\|_2 \leq \beta_1^2 \,, \tag{2}$$

where

$$\beta_1 \lesssim \sqrt{\log 1/\varepsilon} + \sqrt{\frac{d + \log 1/\delta}{\varepsilon n}} \,.$$

Before we prove Theorem 1.1, we need the following standard tail bounds for Gaussians.

Fact 1.2 (see e.g. [1])**.** *Let $X_1, \ldots, X_m \sim \mathcal{N}(\mu, I)$. Then there exist universal constants $A, c > 0$ so that:*

$$\Pr\left[\left\| \frac{1}{m} \sum_{i=1}^{m} X_i - \mu \right\|_2 > t \right] \lesssim \exp\left(Ad - cmt^2\right)$$

$$\Pr\left[\left\| \frac{1}{m} \sum_{i=1}^{m} (X_i - \mu)(X_i - \mu)^\top - I \right\|_2 > t \right] \lesssim \exp\left(Ad - cm \min(t, t^2)\right) \,.$$

We can now prove Theorem 1.1:

Proof of Theorem 1.1. We'll just prove the concentration of the empirical mean of subsets. The corresponding proof for the covariance follows from exactly the same technique while substituting the appropriate tail bound. Note that for any fixed subset $T \subset S$ with $|T| = \varepsilon n$, Fact 1.2 immediately implies that

$$\Pr\left[\|\mu(T) - \mu\|_2 > t\right] \lesssim \exp\left(Ad - c\varepsilon n t^2\right) .$$

The difficulty with this proof is to get this bound to hold *simultaneously* for all subsets T. We will do so by union bounding over all sets T of size εn. This is a bit unusual: that as we take n larger, the number of things over which we union bound also increases. As we'll see, this is the technical reason why we will never be able to achieve error that is asymptotically better than $\sqrt{\log 1/\varepsilon}$.

Formally, there are $\binom{n}{\varepsilon n}$ subsets T of size εn, and so therefore

$$\Pr\left[\exists T : |T| = \varepsilon n \text{ and } \|\mu(T) - \mu\|_2 > t\right] \lesssim \binom{n}{\varepsilon n} \exp\left(Ad - c\varepsilon n t^2\right) .$$

We now use the estimate that $\log_2 \binom{n}{\varepsilon n} \leq nH(\varepsilon)$, where $H(y) = y \log 1/y + (1-y)\log 1/(1-y)$ is the binary entropy function. Since $H(\varepsilon) \leq 2\varepsilon \log 1/\varepsilon$ for $\varepsilon \in (0, 1/2)$, we have that

$$\Pr\left[\exists T : |T| = \varepsilon n \text{ and } \|\mu(T) - \mu\|_2 > t\right] \lesssim \exp(2\varepsilon n \log 1/\varepsilon + Ad - c\varepsilon n t^2) .$$

Therefore, by a choice of β_1, we have that $2\log 1/\varepsilon < ct^2/2$, and therefore

$$\Pr\left[\exists T : |T| = \varepsilon n \text{ and } \|\mu(T) - \mu\|_2 > \beta_1\right] \lesssim \exp\left(Ad - \frac{c}{2}\varepsilon n \beta_1^2\right) \leq \delta ,$$

as claimed. The bound for the covariance follows from the same proof, except by plugging in the corresponding tail bound for the covariance, and since $\beta_1 > 1$ and so we get subexponential-style tails (i.e. $\exp(-c\varepsilon n t)$ rather than $\exp(-c\varepsilon n t^2)$), and so we only get non-trivial estimates at β_1^2 rather than β_1. □

Motivated by Theorem 1.1, let us make the following definition.

Definition 1.1. A set of points S is ε-good with respect to μ if it satisfies the following conditions:

- we have that

$$\|\mu(S) - \mu\|_2 \lesssim \varepsilon\sqrt{\log 1/\varepsilon} , \qquad \text{and} \qquad \left\| \frac{1}{|S|} \sum_{i \in S} (X_i - \mu)(X_i - \mu)^\top - I \right\|_2 \lesssim \varepsilon \log 1/\varepsilon , \tag{3}$$

- for all $T \subset S$ with $|T| = \varepsilon n$ we have

$$\|\mu(T) - \mu\|_2 \lesssim \sqrt{\log 1/\varepsilon} , \qquad \text{and} \qquad \left\| \frac{1}{|T|} \sum_{i \in T} (X_i - \mu)(X_i - \mu)^\top - I \right\|_2 \lesssim \log 1/\varepsilon . \tag{4}$$

Notice that by Theorem 1.1 and Fact 1.2, if S is a set of $n = \Omega\left(\frac{d}{\varepsilon^2 \log 1/\varepsilon}\right)$ samples from $\mathcal{N}(\mu, I)$, then they are ε-good with respect to μ with high probability. In the definition, unlike in Theorem 1.1, we don't subtract the identity for the spectral norm bounds in (4), but since the identity only has spectral norm 1, the result still follows immediately from Theorem 1.1 by a triangle inequality. We now show that such a regularity condition yields a strong notion of spectral signatures:

Theorem 1.3. *Let $\mu \in \mathbb{R}^d$ and let $\varepsilon \in (0, 1/2)$. Let $S = S_{\text{good}} \cup S_{\text{bad}} \setminus S_r$ be an ε-corrupted set of points where S_{good} is ε-good with respect to μ. Let $w \in \mathscr{W}_{S,\varepsilon}$. Then*

$$\|\mu(w) - \mu\|_2 \lesssim \varepsilon\sqrt{\log 1/\varepsilon} + \sqrt{\varepsilon\left(\|\Sigma(w) - I\|_2 + \varepsilon \log 1/\varepsilon\right)} .$$

We note two key differences between this theorem and the theorem presented in an earlier lecture giving spectral signatures assuming bounded covariance. First, the additive term out in front is $\varepsilon\sqrt{\log 1/\varepsilon}$ rather than $\sqrt{\varepsilon}$. Second, the term involving the covariance has the identity subtracted out, whereas the previous theorem did not subtract the identity. This is crucial to get the sorts of bounds we wish to obtain in this setting, as the deviations of $\Sigma(w)$ from I will be much smaller than $\Sigma(w)$ itself. To see this, it's not hard to show that $\|\Sigma(w)\|_2 = \Omega(1)$ with high probability, for any $w \in \mathscr{W}_{S,\varepsilon}$. In contrast, Definition 1.1 allows us to hope that the deviations from the identity can have norm which is much smaller. In particular, if $\|\Sigma(w) - I\|_2 \lesssim \varepsilon \log 1/\varepsilon$, we have that $\|\mu(w) - \mu\|_2 \lesssim \varepsilon\sqrt{\log 1/\varepsilon}$. Recall that the Tukey median obtained error $O(\varepsilon)$, and this error is asymptotically optimal. Up to log factors, this allows us to certify down to the same threshold, using an efficient algorithm.

Before we prove the theorem, we note the following two facts. The first states that no set of weights over the good points that has small mass can induce a large second moment. Formally:

Lemma 1.4. *Let $\varepsilon \in (0, 1/2)$ and $\mu \in \mathbb{R}^d$. Let S_{good} be a set of points of size n which are ε-good with respect to μ. Then, for all $w \in \Gamma_n$ so that $w_i \leq \frac{1}{n}$ for all $i \in S_{\text{good}}$ and $\sum_{i \in S_{\text{good}}} w_i \leq \varepsilon$, we have*

$$\left\| \sum_{i \in S_{\text{good}}} w_i \left(X_i - \mu\right)\left(X_i - \mu\right)^\top \right\|_2 \lesssim \varepsilon \log 1/\epsilon \, .$$

Proof. Without loss of generality, we may assume that $\sum_{i \in S_{\text{good}}} w_i = \varepsilon$. Otherwise, we can form $w' \geq w$ by adding mass on coordinates which are less than $1/n$ so that $\sum_{i \in S_{\text{good}}} w'_i = \varepsilon$ and $w'_i \leq 1/n$ for all $i \in S_{\text{good}}$. Then, since

$$0 \preceq \sum_{i \in S_{\text{good}}} w_i \left(X_i - \mu\right)\left(X_i - \mu\right)^\top \preceq \sum_{i \in S_{\text{good}}} w'_i \left(X_i - \mu\right)\left(X_i - \mu\right)^\top \, ,$$

a spectral norm bound on $\Sigma(w')$ would also immediately imply the same spectral norm bound on $\Sigma(w)$. But now observe that the set of matrices

$$\left\{ \sum_{i \in S_{\text{good}}} w_i \left(X_i - \mu\right)\left(X_i - \mu\right)^\top : w_i \leq 1/n, \sum_{i \in S_{\text{good}}} w_i = 1 \right\}$$

is convex, and the vertices of this set are given by

$$\frac{1}{n} \sum_{i \in T} \left(X_i - \mu\right)\left(X_i - \mu\right)^\top \, ,$$

for $T \subset S$ with $|T| = \varepsilon n$. Therefore the desired result follows from ε-goodness and convexity. □

We note a couple of consequences of this:

Corollary 1.5. *Let $\varepsilon, w, S_{\text{good}}$ be as in Lemma 1.4. Then $\left\| \sum_{i \in S_{\text{good}}} w_i (X_i - \mu) \right\|_2 \lesssim \varepsilon\sqrt{\log 1/\varepsilon}$.*

Proof. For any unit vector v, we have

$$\left\langle v, \sum_{i \in S_{\text{good}}} w_i (X_i - \mu) \right\rangle^2 = \left(\sum_{i \in S_{\text{good}}} w_i \left\langle v, X_i - \mu \right\rangle \right)^2 \tag{5}$$

$$\leq \varepsilon \sum_{i \in S_{\text{good}}} w_i \left\langle v, X_i - \mu \right\rangle^2 \tag{6}$$

$$\lesssim \varepsilon\sqrt{\log 1/\varepsilon} \, , \tag{7}$$

where the second line follows from Hölder's inequality, and the final inequality follows from Lemma 1.4. The desired result follows by taking a supremum over all v. □

Corollary 1.6. *Let S_{good} be an ε-good set of points with respect to μ of size n. Let w be a set of weights so that $w_i \le 1/n$ for all $i \in S$ and $\|w - w(S)\|_1 \le \varepsilon$. Then there is some universal constant $c > 0$ so that*

$$\left\| \sum_{i \in S_{\text{good}}} w_i \left(X_i - \mu\right)\left(X_i - \mu\right)^\top - I \right\|_2 \le c\varepsilon \log 1/\varepsilon \,,$$

and furthermore

$$\sum_{i \in S_{\text{good}}} w_i \left(X_i - \mu(w)\right)\left(X_i - \mu(w)\right)^\top - (1 - c\varepsilon \log 1/\varepsilon) I \succeq 0 \,.$$

Proof. We first observe that

$$\sum_{i \in S_{\text{good}}} w_i \left(X_i - \mu(w)\right)\left(X_i - \mu(w)\right)^\top = \sum_{i \in S_{\text{good}}} w_i \left(X_i - \mu\right)\left(X_i - \mu\right)^\top - \|w\|_1 \cdot (\mu - \mu(w))\left(\mu - \mu(w)\right)^\top \,. \quad (8)$$

We first bound the spectral norm of the second term on the RHS:

$$\left\| \|w\|_1 \cdot (\mu - \mu(w))\left(\mu - \mu(w)\right)^\top \right\|_2 \le \|\mu - \mu(w)\|_2^2 \,. \quad (9)$$

Let $\delta_i = \frac{1}{n} - w_i$. Then, by Corollary 1.5, we have

$$\|\mu - \mu(w)\|_2 \le \|\mu - \mu(S_{\text{good}})\|_2 + \left\| \sum_{i \in S_{\text{good}}} \delta_i (X_i - \mu) \right\|_2 \lesssim \varepsilon \sqrt{\log 1/\varepsilon} \,.$$

Simultaneously, we have

$$\sum_{i \in S_{\text{good}}} w_i \left(X_i - \mu\right)\left(X_i - \mu\right)^\top = \sum_{i \in S_{\text{good}}} \frac{1}{n} \left(X_i - \mu\right)\left(X_i - \mu\right)^\top - \sum_{i \in S_{\text{good}}} \delta_i \left(X_i - \mu\right)\left(X_i - \mu\right)^\top \,. \quad (10)$$

By the ε-goodness of S, the first term can be written as $I + N_1$ where $\|N_1\|_2 \le \varepsilon$, and by Lemma 1.4, the second term has spectral norm at most $c\varepsilon \log 1/\varepsilon$. Hence overall the LHS of (8) can be written as $I + N_2$, where $\|N_2\| \le c\varepsilon \log 1/\varepsilon$. Therefore the smallest eigenvalue of the LHS of (8) is at least $1 - c\varepsilon \log 1/\varepsilon$, from which the claim follows. □

We now have the tools necessary to prove Theorem 1.3.

Proof of Theorem 1.3. As in the bounded second moment case, we begin with a sequence of equalities. Let $\Delta = \mu(w) - \mu$. Then we have

$$\begin{aligned} \|w\|_1 \cdot \|\Delta\|_2^2 &= \|w\|_1 \cdot \|\mu(w) - \mu\|_2^2 = \langle \mu(w) - \mu, \Delta \rangle \\ &= \sum_{i \in S} w_i \langle X_i - \mu, \Delta \rangle \\ &= \langle \mu(S_{\text{good}}) - \mu, \Delta \rangle + \sum_{i \in S_{\text{good}}} \left(w_i - \frac{1}{n} \right) \langle X_i - \mu, \Delta \rangle + \sum_{i \in S_{\text{bad}}} w_i \langle X_i - \mu, \Delta \rangle \,. \end{aligned} \quad (11)$$

We bound these terms separately. By Cauchy-Schwarz, we have that

$$|\langle \mu(S_{\text{good}}) - \mu, \Delta \rangle| \le \|\mu(S_{\text{good}}) - \mu\|_2 \|\Delta\|_2 \lesssim \varepsilon \sqrt{\log 1/\varepsilon} \cdot \|\Delta\|_2 \,, \quad (12)$$

by ε-goodness. We also have

$$\left(\sum_{i\in S_{\text{good}}}\left(w_i-\frac{1}{n}\right)\langle X_i-\mu,\Delta\rangle\right)^2 \le \left(\sum_{i\in S_{\text{good}}}\left(\frac{1}{n}-w_i\right)\right)\left(\sum_{i\in S_{\text{good}}}\left(\frac{1}{n}-w_i\right)\langle X_i-\mu,\Delta\rangle^2\right) \tag{13}$$

$$\le \varepsilon^2\log 1/\varepsilon\cdot\|\Delta\|_2^2\ , \tag{14}$$

where (13) follows from Hölder's inequality and (14) follows from the definition of $\mathscr{W}_{S,\varepsilon}$ and Lemma 1.4.

It remains the control the contribution from S_{bad}. We first observe that

$$\sum_{i\in S_{\text{bad}}} w_i\langle X_i-\mu,\Delta\rangle = \sum_{i\in S_{\text{bad}}} w_i\langle X_i-\mu(w),\Delta\rangle + \left(\sum_{i\in S_{\text{bad}}} w_i\right)\|\Delta\|_2^2\ .$$

As we did for the previous term, we have

$$\left(\sum_{i\in S_{\text{bad}}} w_i\langle X_i-\mu(w),\Delta\rangle\right)^2 \le \left(\sum_{i\in S_{\text{bad}}} w_i\right)\left(\sum_{i\in S_{\text{good}}} w_i\langle X_i-\mu(w),\Delta\rangle^2\right) \tag{15}$$

$$\le \varepsilon\left(\sum_{i\in S_{\text{bad}}} w_i\langle X_i-\mu(w),\Delta\rangle^2\right)\ . \tag{16}$$

We now seek to relate the term in (16) to $\|\Sigma(w)-I\|_2$. Let w' be the set of weights so that $w'_i=w_i$ for all $i\in S_{\text{good}}$, and $w'_i=0$ otherwise. Observe that $w'_i\in\mathscr{W}_{S,\varepsilon}$. By Corollary 1.6, we know that

$$\sum_{i\in S_{\text{good}}} w_i\left(X_i-\mu(w')\right)\left(X_i-\mu(w')\right)^\top - (1-c\varepsilon\log 1/\varepsilon)\|\Delta\|_2^2 \succeq 0\ ,$$

for some $c>0$. Since

$$\begin{aligned}\sum_{i\in S_{\text{good}}} w_i\left(X_i-\mu\right)\left(X_i-\mu\right)^\top &= \sum_{i\in S_{\text{good}}} w_i\left(X_i-\mu(w')\right)\left(X_i-\mu(w')\right)^\top + \|w\|_1\left(\mu(w')-\mu\right)\left(\mu(w')-\mu\right)^\top\\ &\succeq \sum_{i\in S_{\text{good}}} w_i\left(X_i-\mu(w')\right)\left(X_i-\mu(w')\right)^\top\ ,\end{aligned}$$

we also have that

$$\sum_{i\in S_{\text{good}}} w_i\left(X_i-\mu(w)\right)\left(X_i-\mu(w)\right)^\top - (1-c\varepsilon\log 1/\varepsilon)\|\Delta\|_2^2 \succeq 0\ ,$$

which in particular implies that

$$\sum_{i\in S_{\text{good}}} w_i\langle X_i-\mu,\Delta\rangle^2 - (1-c\varepsilon\log 1/\varepsilon)\|\Delta\|_2^2 \ge 0\ .$$

Therefore we have that

$$\begin{aligned}\sum_{i\in S_{\text{bad}}} w_i\langle X_i-\mu,\Delta\rangle^2 &\le \sum_{i\in S} w_i\langle X_i-\mu,\Delta\rangle^2 - \|\Delta\|_2^2 + c\varepsilon\log 1/\varepsilon\cdot\|\Delta\|_2^2\\ &\le \left(\|\Sigma(w)-I\|_2 + c\varepsilon\log 1/\varepsilon\right)\|\Delta\|_2^2\ ,\end{aligned}$$

Plugging this into (14), we obtain that

$$\left(\sum_{i\in S_{\text{bad}}} w_i\langle X_i-\mu,\Delta\rangle\right)^2 \le \varepsilon\left(\|\Sigma(w)-I\|_2 + c\varepsilon\log 1/\varepsilon\right)\|\Delta\|_2^2\ . \tag{17}$$

By simplifying (12), (14), and (17), and using that $\varepsilon<1/2$, we obtain the theorem. □

References

[1] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.