[Lecture 25]

0.1 Efficient Clustering Under Bounded Covariance

We saw that resilience is information-theoretically sufficient for agnostic clustering, but we would also like to develop efficient algorithms for clustering. This is based on work in **?** and **?**, although we will get a slightly slicker argument by using the machinery on resilience that we've developed so far.

As before, we will need a strong assumption than resilience. Specifically, we will assume that each cluster had bounded covariance and that the clusters are well-separated:

Theorem 0.1. *Suppose that the data points $x_1, \dots, x_n$ can be split into k clusters $C_1, \dots, C_k$ with sizes $\alpha_n, \cdots, \alpha_k n$ and means $\mu_1, \dots, \mu_k$, and moreover that the following covariance and separation conditions hold:*

- $\frac{1}{|C_j|} \sum_{i \in C_j} (x_i - \mu_j)(x_i - \mu_j)^\top \preceq \sigma^2 I$ *for each cluster* C_j,
- $\Delta \geq 36\sigma/\sqrt{\alpha}$, *where* $\Delta = \min_{j \neq j'} \|\mu_j - \mu_{j'}\|_2$.

Then there is a polynomial-time algorithm outputting candidate clusters $\hat{C}_1, \dots, \hat{C}_k$ and means $\hat{\mu}_1, \dots, \hat{\mu}_k$ such that:

- $|C_j \triangle \hat{C}_j| = \mathcal{O}(\sigma^2/\alpha\Delta^2)$ *(cluster recovery), and*
- $\|\mu_j - \hat{\mu}_j\|_2 = \mathcal{O}(\sigma^2/\alpha\Delta)$ *(parameter recovery).*

The basic idea behind the algorithm is to project each of the points x_i onto the span of the top k singular vectors of the data matrix $X = [x_1 \ \cdots \ x_n]$. Let P_k be the projection operator onto this space. Then since the points Px_i lie in only a k-dimensional space instead of a d-dimensional space, they are substantially easier to cluster. The algorithm itself has three core steps and an optional step:

1. Project points x_i to Px_i.
2. Form initial clusters based on the Px_i.
3. Compute the means of each of these clusters.
4. Optionally run any number of steps of k-means in the original space of x_i, initialized with the computed means from the previous step.

We will provide more formal psuedocode later [NOTE: TBD]. For now, we focus on the analysis, which has two stages: (1) showing that the initial clustering from the first two steps is "nice enough", and (2) showing that this niceness is preserved by the k-means iterations in the second two steps.

Analyzing the projection. We start by analyzing the geometry of the points $P_k x_i$. The following lemma shows that the projected clusters are still well-separated and have small covariance:

Lemma 0.2. *The projected points $P_k x_i$ satisfy the covariance and separation conditions with parameters σ and $\sqrt{\Delta^2 - 4\sigma^2/\alpha} \geq 35\sigma/\sqrt{\alpha}$:*

$$\frac{1}{|C_j|} \sum_{i \in C_j} (Px_i - P\mu_j)(Px_i - P\mu_j)^\top \preceq \sigma^2 I \text{ and } \|P\mu_j - P\mu_{j'}\|_2 \geq \sqrt{\Delta^2 - 4\sigma^2/\alpha}. \tag{1}$$

In other words, the covariance condition is preserved, and separation is only decreased slightly.

Proof. The covariance condition is preserved because the covariance matrix of the projected points for cluster j is $P_k \Sigma_j P_k$, where Σ_j is the un-projected covariance matrix. This evidently has smaller singular values than Σ_k.

The separation condition requires more detailed analysis. We start by showing that there is not much in the orthogonal component $(I - P_k)x_i$. Indeed, we have that the top singular value of $(I - P_k)x_i$ is at most σ:

$$S = \frac{1}{n}\sum_{i=1}^{n}((I - P_k)x_i)((I - P_k)x_i)^\top \preceq \sigma^2 I \tag{2}$$

This is because P_k minimizes this top singular value among all k-dimensional projection matrices, and if we take the projection Q_k onto the space spanned by the means $\mu_1, \ldots, \mu_k$, we have

$$\frac{1}{n}\sum_{i=1}^{n}((I - Q_k)x_i)((I - Q_k)x_j)^\top = \sum_{j=1}^{k}\frac{\alpha_j}{|C_j|}\sum_{i\in C_j}((I - Q_k)x_i)((I - Q_k)x_i)^\top \tag{3}$$

$$= \sum_{j=1}^{k}\frac{\alpha_j}{|C_j|}\sum_{i\in C_j}((I - Q_k)(x_i - \mu_j))((I - Q_k)(x_i - \mu_j))^\top \tag{4}$$

$$\preceq \sum_{j=1}^{k}\frac{\alpha_j}{|C_j|}\sum_{i\in C_j}(x_i - \mu_j)(x_i - \mu_j)^\top \preceq \sum_{j=1}^{k}\alpha_j\sigma^2 I = \sigma^2 I. \tag{5}$$

Given this, we know that the projections $(I - P_k)\mu_j$ must be small, since otherwise we have

$$v^\top S v = \frac{1}{n}\sum_{i=1}^{n}\langle(I - P_k)x_i, v\rangle^2 \tag{6}$$

$$\geq \frac{\alpha_j}{|C_j|}\sum_{i\in C_j}\langle(I - P_k)x_i, v\rangle^2 \tag{7}$$

$$\geq \alpha_j\Big\langle\frac{1}{|C_j|}\sum_{i\in C_j}(I - P_k)x_i, v\Big\rangle^2 \tag{8}$$

$$= \alpha_j\langle(I - P_k)\mu_j, v\rangle^2. \tag{9}$$

Consequently $\langle(I-P_k)\mu_j, v\rangle^2 \leq \sigma^2/\alpha_j$ and hence (taking v to align with $(I-P_k)\mu_j$) we have $\|(I-P_k)\mu_j\|_2 \leq \sigma/\sqrt{\alpha_j}$. In particular $\|(I - P_k)(\mu_j - \mu_{j'})\|_2 \leq 2\sigma/\sqrt{\alpha}$.

Now, by the Pythagorean theorem we have

$$\|P_k(\mu_j - \mu_{j'})\|_2^2 = \|\mu_j - \mu_{j'}\|_2^2 - \|(I - P_k)(\mu_j - \mu_{j'})\|_2^2 \geq \Delta^2 - 4\sigma^2/\alpha, \tag{10}$$

and hence the projected means are separated by at least $\sqrt{\Delta^2 - 4\sigma^2/\alpha}$, as was to be shown. □

Analyzing the initial clustering. We now analyze the initial clustering. Call a point i a *proto-center* if there are at least $\frac{\alpha}{2}n$ projected points within distance $3\sigma\sqrt{k}$ of $P_k x_i$, and call the set of these nearby points the associated *proto-cluster*.

We will show that the proto-clusters are nearly pure (have few points not from C_j) using a similar argument as when we analyzed resilient clustering. As before, call a proto-cluster *j-like* if there are at least $\frac{\alpha_j\alpha}{4}n$ points from C_j in the proto-cluster.

Lemma 0.3. *Each proto-cluster is j-like for exactly one j.*

Proof. We know that it is j-like for at least one j by the Pigeonhole principle (if not, then the proto-cluster has at most $\frac{\alpha}{4}n$ points in total, contradicting its size of at least $\frac{\alpha}{2}n$). So suppose for the sake of contradiction that it is both j-like and j'-like. By resilience, the mean of the points from C_j is at most $2\sigma/\sqrt{\alpha}$ away from $P_k\mu_j$, and similarly the mean of the points from $C_{j'}$ is at most $2\sigma/\sqrt{\alpha}$ away from $P_k\mu_{j'}$. Since the cluster has radius $3\sigma\sqrt{k} \leq 3\sigma/\sqrt{\alpha}$, this implies that $\|P_k(\mu_j - \mu_{j'})\|_2 \leq 10\sigma/\sqrt{\alpha}$, contradicting the separation condition for the projected means. Thus no proto-cluster can be j-like for multiple j, which proves the lemma. □

Now since each proto-cluster is j-like for exactly one j, at least half of the points must come from that proto-cluster.

At this point we are essentially done if all we care about is constructing an efficient algorithm for cluster recovery (but not parameter recovery), since if we just extend each proto-cluster by $\mathcal{O}(\sigma)$ we are guaranteed to contain almost all of the points in its corresponding cluster, while still containing very few points from any other cluster (assuming the data are well-separated). However, parameter recovery is a bit trickier because we need to make sure that the small number of points from other clusters don't mess up the mean of the cluster. The difficulty is that while we have control over the projected distances, and can recover the projected centers $P_k\mu_j$ well, we need to somehow get back to the original centers μ_j.

The key here is that for each proto-cluster, the Px_i are all close to each other, and the missing component $(I - P_k)x_i$ has bounded covariance. Together, these imply that the proto-cluster is *resilient*—deleting an ϵ-fraction of points can change the mean by at most $\mathcal{O}(\sigma\epsilon)$ in the P_k component, and $\mathcal{O}(\sigma\sqrt{\epsilon})$ in the $(I - P_k)$ component. In fact, we have:

Lemma 0.4. *Let B be a proto-cluster with mean ν. Then*

$$\frac{1}{|B|}\sum_{i\in B}(x_i - \nu)(x_i - \nu)^\top \preceq 11\sigma^2/\alpha. \tag{11}$$

In particular, if B is j-like then we have $\|\mu_j - \nu\|_2 \le 9\sigma/\sqrt{\alpha}$.

Proof. The covariance bound is because the covariance of the x_i are bounded in norm by at most $3\sigma\sqrt{k}$ in the P_k component and hence can contribute at most $9\sigma^2 k \le 9\sigma^2/\alpha$ to the covariance, while we get an additional $2\sigma^2/\alpha$ in an orthogonal direction because the overall second moment of the $(I - P_k)x_i$ is σ^2 and the $i \in B$ contribute to at least an $\frac{\alpha}{2}$ fraction of that.

Now, this implies that B is resilient, while we already have that C_j is resilient. Since $B \cap C_j$ contains at least half the points in both B and C_j, this gives that their means are close–within distance $2(\sqrt{11}+1)\sigma/\sqrt{\alpha} < 9\sigma/\sqrt{\alpha}$. □

Analyzing k-means. We next show that k-means iterations preserve certain important invariants. We will call the assigned means $\hat{\mu}_j$ *R-close* if $\|\hat{\mu}_j - \mu_j\|_2 \le R$ for all j, and we will call the assigned clusters $\hat{C}_j$ *ϵ-close* if $|C_j \triangle \hat{C}_j| \le \epsilon|C_j|$ for all j. We will show that if the means are R-close then the clusters are ϵ-close for some $\epsilon = f(R)$, and that the resulting new means are then $g(R)$-close. If R is small enough then we will also have $g(R) < R$ so that we obtain an invariant.

Let $\Delta_{jj'} = \|\mu_j - \mu_{j'}\|_2$, so $\Delta_{jj'} \ge \Delta$. We will show that if the $\hat{\mu}_j$ are R-close, then few points in C_j can end up in $\hat{C}_{j'}$. Indeed, if x_i ends up in $\hat{C}_{j'}$ then we must have $\|x_i - \hat{\mu}_{j'}\|_2^2 \le \|x_i - \hat{\mu}_j\|_2^2$, which after some re-arrangement yields

$$\langle x_i - \hat{\mu}_j, \hat{\mu}_{j'} - \hat{\mu}_j\rangle \ge \frac{1}{4}\langle \hat{\mu}_{j'} - \hat{\mu}_j, \hat{\mu}_{j'} - \hat{\mu}_j\rangle. \tag{12}$$

Applying the covariance bound and Chebyshev's inequality along the vector $v = \hat{\mu}_{j'} - \hat{\mu}_j$, we see that the fraction of points in C_j that end up in $\hat{C}_{j'}$ is at most $\frac{4\sigma^2}{\|\hat{\mu}_j - \hat{\mu}_{j'}\|_2^2} \le \frac{4\sigma^2}{(\Delta_{jj'}-2R)^2} \le \frac{4\sigma^2}{(\Delta-2R)^2}$. In total this means that at most $\frac{4\sigma^2 n}{(\Delta-2R)^2}$ points from other clusters end up in $\hat{C}_j$, while at most $\frac{4k\sigma^2|C_j|}{(\Delta-2R)^2}$ points from C_j end up in other clusters. Thus we have $\epsilon \le \frac{4k\sigma^2}{(\Delta-2R)^2} + \frac{4\sigma^2}{\alpha(\Delta-2R)^2} \le \frac{8\sigma^2}{\alpha(\Delta-2R)^2}$, so we can take

$$f(R) = \frac{8\sigma^2}{\alpha(\Delta - 2R)^2}. \tag{13}$$

Now suppose that $\gamma_{jj'}|C_j|$ points in C_j are assigned to $\hat{C}_{j'}$, where we must have $\gamma_{jj'} \le \frac{4\sigma^2}{(\Delta_{jj'}-2R)^2}$. By resilience, the mean of these points is within $\sigma/\sqrt{\gamma_{jj'}}$ of μ_j and hence within $\Delta_{jj'} + \sigma/\sqrt{\gamma_{jj'}}$ of $\mu_{j'}$. In total, then, these points can shift the mean $\hat{\mu}_{j'}$ by at most

$$\frac{\gamma_{jj'}\alpha_j n(\Delta_{jj'} + \sigma/\sqrt{\gamma_{jj'}})}{\frac{1}{2}\alpha n} \le \frac{2\alpha_j}{\alpha}\Big(\frac{4\sigma^2\Delta_{jj'}}{(\Delta_{jj'} - 2R)^2} + \frac{2\sigma^2}{\Delta_{jj'} - 2R}\Big) \le \frac{4\alpha_j}{\alpha}\Big(\frac{2\sigma^2\Delta}{(\Delta - 2R)^2} + \frac{\sigma^2}{\Delta - 2R}\Big). \tag{14}$$

At the same time, the $\frac{4k\sigma^2}{(\Delta-2R)^2}$ fraction of points that are missing from $C_{j'}$ can change its mean by at most $\frac{4\sigma^2\sqrt{k}}{\Delta-2R}$. Thus in total we have

$$\|\hat{\mu}_{j'} - \mu_{j'}\|_2 \leq \frac{4\sigma^2}{\Delta - 2R} \cdot \left(\sqrt{k} + \frac{1}{\alpha} + \frac{2\Delta}{\alpha(\Delta - 2R)}\right) \leq \frac{8\sigma^2(\Delta - R)}{\alpha(\Delta - 2R)^2}. \tag{15}$$

In particular we can take $g(R) = \frac{8\sigma^2(\Delta-R)}{\alpha(\Delta-2R)^2}$.

As long as $R \leq \Delta/4$ we have $g(R) \leq \frac{24\sigma^2}{\alpha\Delta}$ and $f(R) \leq \frac{32\sigma^2}{\alpha\Delta^2}$, as claimed. Since our initial R is $9\sigma/\sqrt{\alpha}$, this works as long as $\Delta \geq 36\sigma/\sqrt{\alpha}$, which completes the proof.