Robustness of SDPs for Partial Recovery of Clustering Subgaussian Mixtures

UROP+ Final Paper, Summer 2016

Siqi Chen
Mentor: Amelia Perry
Project suggested by Ankur Moitra

August 2016

Abstract

In this paper, we examine the robustness of a relax-and-round k-means clustering procedure, a method for clustering subgaussian mixtures using semidefinite programming first introduced in [MVW16]. We are interested in the robustness of the algorithm when there is an adversarial corruption of ϵN points each through distance at most R_0. We show that under such corruption this specific algorithm well-approximates the center of the subgaussians.

1 Introduction

Given N points, k-means clustering aims to partition these N observations into k clusters in which each observation belongs to the cluster with the nearest mean, or mathematically the objective is to minimize the sum of squared distances between observations and their cluster means. The term "k-means" was first used by James MacQueen [Mac67] in 1967, though the idea goes back to Hugo Steinhaus in 1957. The standard algorithm was first proposed by Stuart Lloyd [Llo82] in 1957, though it wasn't published outside of Bell Labs until 1982. This optimization problem is computationally hard to solve for an arbitrary set of input points, but may remain tractable on average in distributional models, such as gaussian or subgaussian mixtures.

So what happens when each of the unknown clusters is in some gaussian probability distribution? And what are some of the constraints on these clusters in order to have sufficient signal for successful approximation?

Dasgupta [Das99] introduced an algorithm based on random projections and showed that this algorithm well-approximates centers of Gaussians as long as the centers are well-separated. After that, performance guarantees for several algorithmic alternatives have emerged, and every existing performance guarantee either shows that the algorithm correctly clusters all points according to Gaussian mixture component, or that it well-approximates the center of each Gaussian.

Mixon, Villar and Ward [MVW16], however, looked at the problem when each of these clusters is in some unknown subgaussian probability distribution. They introduced a model-free relax-and-round algorithm for k-means clustering based on a semidefinite relaxation due to Peng and Wei [PW07]; and adapting the idea from Guédon and Vershynin [GV15] which proved the consistency of a certain semidefinite program on detecting communities in the stochastic block model from the following steps:

- the SDP gets the right answer with some reference objective (unknown to the algorithm);
- if the actual objective is close to the reference objective (in some support norm), then the solution is close to the actual solution;
- the hypothesis is satisfied: the actual objective is close to the reference objective

[MVW16] provided a performance guarantee for the algorithm to well-approximate the center of each subgaussian. Note that earlier work on SDPs for the stochastic block model addressed exact clustering of all vertices [ABH16, HWX16], and [GV15] was the first paper to prove that an SDP correctly clusters a large fraction of vertices in a setting where exact recovery is impossible. This approach of [GV15] is particularly well suited for proving partial recovery results.

Yet most real life situations may not have the exact k-means clustering set-up, that is why we are interested in the robustness of the algorithm when

there are some corruptions of points. Inspired by how Moitra, Perry and Wein [MPW16] described the robustness properties of Guédon and Vershynin analysis in the original setting of the stochastic block model, we examine the key steps of [MVW16] in Section 3, and with some characteristics of SDP as we prove in Lemma 3.3, we show that the algorithm is robust when there is an adversarial corruption of $\epsilon_0 N$ points, each through a distance at most R_0.

2 Preliminary: The Mixon-Villar-Ward Framework and Partial Recovery

The Mixon-Villar-Ward Framework develops the relax-and-round k-means clustering procedure. It provides a performance guarantee for the algorithm that well-approximates the center of each Gaussian. It consists of the following three steps:

Step 1: Approximation. In this first step, [MVW16] adapts an approach used by Guédon and Vershynin [GV15] to provide approximation guarantees for the following semidefinite program under the stochastic block model for graph clustering.

For each $t \in [k] := \{1, \cdots, k\}$, let D_t be an unknown subgaussian probability distribution over $\mathcal{R}^m$, with first moment $\gamma_t \in \mathcal{R}^m$ and second moment matrix with largest eigenvalue σ_t^2. For each t, an unknown number n_t of random points $\{x_{t,i}\}_{i\in[n_t]}$ is drawn independently from D_t. Given k and points $\{x_{t,i}\}_{i\in[n_t],t\in[k]}$, Peng and Wei [PW07] first introduced the following semidefinite program:

$$\begin{array}{ll} \text{minimize} & Tr(DX) \\ \text{subject to} & Tr(X) = k \\ & X1 = 1 \\ & X \geq 0, X \succeq 0 \end{array}$$

where D denotes the $N \times N$ matrix defined entry-wise by $D_{ij} = \|x_i - x_j\|_2^2$. Let X_D denote the minimizer of the SDP, we would hope for X_D to look like the block matrix with $1/n_i$ in the diagonal blocks and 0 in the off-diagonal blocks.

We now introduce some of the other notations we will be using in this paper:

Take $\Delta_{ab} := \|\gamma_a - \gamma_b\|_2$, and let the reference matrix R be defined as $(R_{ab})_{ij} := \xi + \Delta_{ab}^2/2 + \max\{0, \Delta_{ab}^2/2 + 2\langle r_{a,i} - r_{b,j}, \gamma_a - \gamma_b\rangle\}$, where $r_{t,i} := x_{t,i} - \gamma_t$, and $\xi > 0$ is a parameter. Denote X_R as the minimizer of the SDP when D is replaced by R. Then, as [MVW16] showed in Lemma 1:

If $1_a \in \mathbb{R}^N$ denote the indicator function for the indices i corresponding to points x_i drawn from the ath subgaussian, and $\gamma_a \neq \gamma_b$ whenever $a \neq b$, then $X_R = \sum_{t=1}^k (1/n_t) 1_t 1_t^T$ (the reference SDP recovers the truth).

Proposition 2.1. *Fix $\epsilon, \eta > 0$. There exist universal constants C, c_1, c_2, c_3 such that if $\alpha = n_{\max}/n_{\min} \lesssim k \lesssim m$ and $N > \max\{c_1 m, c_2 \log(2/\eta), \log(c_3/\eta)\}$,*

then

$$\|X_D - X_R\|_F^2 \leq \epsilon$$

with probability at least $1 - 2\eta$ *provided* $\Delta_{\min}^2 \geq \frac{C}{\epsilon} k^2 \alpha \sigma_{\max}^2$, *where* $\Delta_{\min} = \min_{a \neq b} \|\gamma_a - \gamma_b\|_2$ *is the minimal cluster center separation.*

The proof of Proposition 2.1 is shown in the proof of Theorem 2 in [MVW16].

Step 2: Denoising. Convert X_D from Step 1 into an estimate for the centers $\{\gamma_t\}_{t \in [k]}$.

Let P denote the $m \times N$ matrix whose (a,i)th column is $x_{a,i}$. Then PX_R is an $m \times N$ matrix whose (a,i)th column is $\tilde{\gamma}_a$, the centroid of the ath cluster, which converges to γ_a as $N \to \infty$.

Proposition 2.2. *Assume the points* $\{x_{a,i}\}_{i \in [n]}$ *come from* $\mathcal{N}(\gamma_a, \sigma^2 I_m)$ *in* $\mathbb{R}^m$ *for each* $a \in [k]$. *Then* $\mathbb{E}[\frac{1}{N} \sum_{a=1}^k \sum_{i=1}^n \|x_{a,i} - \gamma_a\|_2^2] = m\sigma^2$. *Let* $c_{a,i}$ *denote the* (a,i)*th column of* PX_D, *if* $k\sigma \lesssim \Delta_{\min} \leq \Delta_{\max} \lesssim K\sigma$, *then*

$$\frac{1}{N} \sum_{a=1}^k \sum_{i=1}^n \|c_{a,i} - \tilde{\gamma}_a\|_2^2 \lesssim K^2 \sigma^2$$

with high probability as $n \to \infty$.

The proof of Proposition 2.2 is shown in Corollary 3, a corollary of a more technical result (Theorem 11) in [MVW16].

Step 3: Rounding. Present a rounding scheme that provides a clustering of the original data from the denoised results from Step 2. The cost of rounding is a factor of k in the average squared deviation of the estimates.

3 Robustness of SDPs for Partial Recovery

We show that Mixon-Villar-Ward Framework performs well when one randomly moves $\epsilon_0 N$ points, each through a distance at most R_0. We will analyze the first two steps presented in Mixon-Villar-Ward:

Step 1: Approximation. Let D' be the $N \times N$ matrix defined entry-wise by $D'_{ij} = \|x'_i - x'_j\|_2^2$, where x'_i corresponds to where x_i is after we make $\epsilon_0 N$ movements, each up to distance R_0. Show that $\|X'_D - X_R\|_F^2$ is small.

Theorem 3.1. *Fix* $\epsilon, \eta > 0$. *There exist universal constants* C, c_1, c_2, c_3 *such that if* $\alpha = n_{\max}/n_{\min} \lesssim k \lesssim m$ *and* $N > \max\{c_1^2 m, c_2 \log(2/\eta), \log(c_3/\eta)\}$, *then*

$$\|X'_D - X_R\|_F^2 \leq \epsilon$$

with probability $\geq 1-2\eta$ *provided* $\Delta_{min}^2 \geq \frac{C}{\epsilon} k\alpha(\epsilon_0 R_0 + \sigma_{\max}^2 k)$*, where* $\Delta_{\min} = \min_{a\neq b} \|\gamma_a - \gamma_b\|_2$ *is the minimal cluster center separation.*

The proof of Theorem 3.1 uses the two lemmas directly derived from [MVW16].

(1) $\|X'_D - X_R\|_F^2 \leq \frac{5}{n_{min}\Delta_{min}^2} Tr(R(X'_D - X_R))$. [Lemma 8]

(2) Put $\tilde{D'} := P_{1\perp} D' P_{1\perp}$ and $\tilde{R} := P_{1\perp} R P_{1\perp}$. Then $Tr(R(X'_D - X_R)) \leq 2\mathcal{F}(\tilde{D'} - \tilde{R})$ [Lemma 9]

Here, $\mathcal{F}$ is a support norm introduced by [MVW16] as follows: Let $\mathcal{F}(M)$ denote the value of the following program: $\mathcal{F}(M) =$ maximum $|Tr(MX)|$, subject to $Tr(X) = k$, $X1 = 1$, $X \geq 0$, $X \succeq 0$ and let $\mathcal{X}_M$ denote its maximizer.

Since we aim to bound $\|X'_D - X_R\|_F^2$, we are going to show that $\mathcal{F}(\tilde{\mathcal{D}'} - \tilde{\mathcal{R}})$ does not increase by too much when we make the adversarial changes.

Lemma 3.2. *Triangular Inequality of* $\mathcal{F}$ *norm:* $\mathcal{F}(A+B) \leq \mathcal{F}(A) + \mathcal{F}(B)$

Proof. $\mathcal{F}(A+B) = \max |\langle A+B, X\rangle| = |\langle A+B, \mathcal{X}_{A+B}\rangle| \leq |\langle A, \mathcal{X}_{A+B}\rangle| + |\langle B, \mathcal{X}_{A+B}\rangle| \leq |\langle A, \mathcal{X}_A\rangle| + |\langle B, \mathcal{X}_B\rangle| = \mathcal{F}(A) + \mathcal{F}(B)$, where the first inequality follows from triangular inequality of absolute value and the second inequality follows from the definition of $\mathcal{F}$ norm. □

Lemma 3.3. *Given symmetric matrix* A *of size* $n \times n$ *where all entries except the* i*th column and* i*th row are identical.*

$$A = \begin{pmatrix} a & \cdots & a & b & a & \cdots & a \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a & \cdots & a & b & a & \cdots & a \\ b & \cdots & b & c & b & \cdots & b \\ a & \cdots & a & b & a & \cdots & a \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a & \cdots & a & b & a & \cdots & a \end{pmatrix}$$

Then there exists an $n \times n$ *matrix* X_A *such that* X_A *is a maximizer for the SDP that defines* $\mathcal{F}(A)$*, and*

$$X_A = \begin{pmatrix} x & & & y & & & \\ & \ddots & & \vdots & & & \\ & & x & y & & & \\ y & \cdots & y & z & y & \cdots & y \\ & & & y & x & & \\ & & & \vdots & & \ddots & \\ & & & y & & & x \end{pmatrix}$$

for some x, y, z with some w filling all empty entries. Here $x \leq 1$; $y \leq \frac{1}{n-1}$; $z \leq 1$; $w \leq \frac{1}{n-2}$. We call matrix X_A "in the form of xyzw".

Proof. Let X_0 be a maximizer for the SDP that defines $\mathcal{F}(A)$. Show that there exists X_A in the required form that also maximizes the SDP. The core idea of the proof is that if two entries of X_0, whose correspondents in X_A are labeled the same, (this can be separated into three cases: diagonal-diagonal, entry-entry, ith-ith) have different values, find another matrix X_1 such that it maximizes the program and the average of these two maximizers produces the same value on these two entries. Since the sum of two PSD matrices is also a PSD matrix, the average of X_0 and X_1 is a maximizer for the program with same value on those two entries. We look at entry values in the order of diagonal-diagonal, diagonal-entry, ith-ith:

Diagonal-diagonal Suppose two of the diagonal entries x_{jj} and x_{kk} are different. X_1 is generated by exchanging the two rows and then the two columns (exchanging columns before rows works in the same way). Without loss of generality, we show the case when $j = 1, k = 2, j, k \neq i$.

$$X_0 = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nn} \end{pmatrix} \Longrightarrow X_1 = \begin{pmatrix} x_{22} & x_{21} & x_{23} & \cdots & x_{2n} \\ x_{12} & x_{11} & x_{13} & \cdots & x_{1n} \\ x_{32} & x_{31} & x_{33} & \cdots & x_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n2} & x_{n1} & x_{n3} & \cdots & x_{nn} \end{pmatrix}$$

X_1 is positive semi-definite: suppose for the purpose of contradiction that there exists a vector u with entries $u_1, \cdots, u_n$ such that $u^T X_1 u < 0$, then let v be the vector defined as

$$v_t = \begin{cases} u_k & \text{if } t = j \\ u_j & \text{if } t = k \\ u_t & \text{otherwise} \end{cases}$$

Then $v^T X_0 v < 0$, a contradiction to X_0's being positive semi-definite. Hence, X_1 is positive semi-definite.

As the operation does not change the sum of any row/column/diagonal, PSD matrix X_1 satisfies all the conditions listed in the SDP that defines $\mathcal{F}$ norm. By symmetry of A and X_1, $|\langle A, X_1 \rangle| = |\langle A, X_0 \rangle|$. Therefore, X_1 also maximizes the program, and the average of X_0 and X_1 has same value $(\frac{x_{jj} + x_{kk}}{2})$ on entry jj and kk.

Entry-entry Now that entries on the diagonal except ii position have the same value, suppose entry x_{jk} and x_{pq} has different value. We can separate the program into (x_{jk}, x_{jq}) and (x_{jq}, x_{pq}) pair, where at least one of the pair has

different value. By symmetry of X_0, we can simplify the entry-entry case into the case where entries on the same column have different values, as in the case of (x_{jq}, x_{pq}). Similar to the diagonal-diagonal case, X_1 is generated by exchanging these two rows and then the two corresponding columns (exchanging columns before rows works in the same way). Without loss of generality, we show the case when X_{21} and x_{31} have different values. $(j = 2, p = 3, q = 1, j, k, q \neq i)$.

$$X_0 = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nn} \end{pmatrix} \Longrightarrow X_1 = \begin{pmatrix} x_{11} & x_{13} & x_{12} & \cdots & x_{1n} \\ x_{31} & x_{33} & x_{32} & \cdots & x_{3n} \\ x_{21} & x_{23} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n3} & x_{n2} & \cdots & x_{nn} \end{pmatrix}$$

By same reasoning as in the first case, X_1 also maximizes the program, and the average of X_0 and X_1 has same value $(\frac{x_{jq} + x_{pq}}{2})$ on entry jq and pq.

Ith-ith Suppose entry x_{ij} and x_{ki} has different value. By symmetry of X_0, $x_{ji} = x_{ij} \neq x_{ki}$. Therefore, we can simplify the ith-ith case into the case where entries on the ith column have different values, as in the case of (x_{ji}, x_{ki}). Similarly, X_1 is generated by exchanging these two rows and then the two corresponding columns. In this case, X_1 maximizes the program, and the average of X_0 and X_1 has same value $(\frac{x_{ji} + x_{ki}}{2})$ on entry ji and ki.

Therefore, given symmetric matrix A, there exists X_A in the required form. By definition of $\mathcal{F}(A)$, each row of X_A sums up to 1 and all entries are non-negative. Therefore, we have

$$\begin{cases} x + (n-2)w + y \leq 1 \\ (n-1)y + z \leq 1 \end{cases} \Longrightarrow \begin{cases} x \leq 1 \\ y \leq \dfrac{1}{n-1} \\ z \leq 1 \\ w \leq \dfrac{1}{n-2} \end{cases}$$

□

Lemma 3.4. *Of all the n points, label them 1 to n. If we move point i through a distance of at most r, then $\mathcal{F}(\tilde{D}' - \tilde{D}) \leq 14r$.*

Proof. Let u be the vector defined entry-wise by $u_j = \|D'_{ij} - D_{ij}\|_2^2$, then $|u_j| \leq r$ and $\tilde{D}' - \tilde{D} = P_{1\perp}(ue_i^T + e_i u^T)P_{1\perp}$. Find vector v such that $u = v + \lambda 1$ and the average of entries of v is 0. Then $|v_j| \leq 2r$ and $P_{1\perp}v = v$. Since $P_{1\perp}1 = 0$, we have:

$$\begin{aligned}\tilde{D}' - \tilde{D} &= P_{1\perp}(ue_i^T + e_i u^T)P_{1\perp} \\ &= P_{1\perp}((v+\lambda 1)e_i^T + e_i(v+\lambda 1)^T)P_{1\perp} \\ &= P_{1\perp}ve_i^T P_{1\perp} + P_{1\perp}e_i v^T P_{1\perp} \\ &= ve_i^T P_{1\perp} + P_{1\perp}e_i v^T \\ &= v(P_{1\perp}e_i)^T + (P_{1\perp}e_i)v^T\end{aligned}$$

Note that $(P_{1\perp}e_i)^T = (-\frac{1}{n}, \cdots, -\frac{1}{n}, \frac{n-1}{n}, -\frac{1}{n}, \cdots, -\frac{1}{n})^T$, then

$$|(\tilde{D}' - \tilde{D})_{jk}| \leq \begin{cases} \frac{4r}{n} & j,k \neq i \\ 2r & \text{otherwise} \end{cases}$$

As a result,

$$\begin{aligned}\mathcal{F}(\tilde{D}' - \tilde{D}) &= \max|\langle \tilde{D}' - \tilde{D}, X\rangle| \\ &= \max(\max\langle \tilde{D}' - \tilde{D}, X\rangle, \text{-min}\langle \tilde{D}' - \tilde{D}, X\rangle) \\ &\leq \max(\max\langle (\tilde{D}' - \tilde{D})^{inc}, X\rangle, \text{-min}\langle (\tilde{D}' - \tilde{D})^{dec}, X\rangle)\end{aligned}$$

where $(\tilde{D}' - \tilde{D})^{inc}$ (alternatively, $(\tilde{D}' - \tilde{D})^{dec}$) is entry-wise bigger (smaller) than $\tilde{D}' - \tilde{D}$,

$$(\tilde{D}' - \tilde{D})^{inc} = \begin{pmatrix} \frac{4r}{n} & \cdots & \frac{4r}{n} & 2r & \frac{4r}{n} & \cdots & \frac{4r}{n} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{4r}{n} & \cdots & \frac{4r}{n} & 2r & \frac{4r}{n} & \cdots & \frac{4r}{n} \\ 2r & \cdots & 2r & 2r & 2r & \cdots & 2r \\ \frac{4r}{n} & \cdots & \frac{4r}{n} & 2r & \frac{4r}{n} & \cdots & \frac{4r}{n} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{4r}{n} & \cdots & \frac{4r}{n} & 2r & \frac{4r}{n} & \cdots & \frac{4r}{n} \end{pmatrix}$$

and $(\tilde{D}' - \tilde{D})^{dec} = -(\tilde{D}' - \tilde{D})^{inc}$. The inequality follows as $\mathcal{X}_{\tilde{D}' - \tilde{D}}$ is entry-wise non-negative, thus $\max\langle \tilde{D}' - \tilde{D}, X\rangle = \langle \tilde{D}' - \tilde{D}, \mathcal{X}_{\tilde{D}' - \tilde{D}}\rangle \leq \langle (\tilde{D}' - \tilde{D})^{inc}, \mathcal{X}_{\tilde{D}' - \tilde{D}}\rangle \leq \langle (\tilde{D}' - \tilde{D})^{inc}, \mathcal{X}_{(\tilde{D}' - \tilde{D})^{inc}}\rangle = \max\langle (\tilde{D}' - \tilde{D})^{inc}, X\rangle$. Similarly, $\text{-min}\langle \tilde{D}' - \tilde{D}, X\rangle \leq \text{-min}\langle (\tilde{D}' - \tilde{D})^{dec}, X\rangle$.

By Lemma 3.3, there exists an $n \times n$ matrix X^{inc} in the form of $xywz$ such that

$$\begin{aligned}\mathcal{F}((\tilde{D}' - \tilde{D})^{inc}) &= \langle (\tilde{D}' - \tilde{D})^{inc}, X^{inc}\rangle \\ &= (n-1)\frac{4r}{n}x + 2(n-1)2ry + 2rz + (n-1)(n-2)\frac{4r}{n}w \\ &\leq 4r + 4r + 2r + 4r = 14r\end{aligned}$$

As a result, $\max\langle (\tilde{D}' - \tilde{D})^{inc}, X\rangle = \mathcal{F}((\tilde{D}' - \tilde{D})^{inc}) \leq 14r$. Similarly, $\min\langle (\tilde{D}' - \tilde{D})^{dec}, X\rangle \geq -14r$.

Therefore, $\mathcal{F}(\tilde{D}' - \tilde{D}) \leq 14r$. □

Proof of Theorem 3.1:

If we randomly move $\epsilon_0 N$ points, each up to distance R_0, among a total of N points,

$$\begin{aligned}\mathcal{F}(\tilde{D}' - \tilde{R}) &\leq \mathcal{F}(\tilde{D}' - \tilde{D}) + \mathcal{F}(\tilde{D} - \tilde{R}) && \text{(Lemma 3.2)}\\ &\leq \sum_{\epsilon N} \mathcal{F}(\text{each move}) + \mathcal{F}(\tilde{D} - \tilde{R}) && \text{(Lemma 3.2)}\\ &\leq 14 R_0 \epsilon N + \mathcal{F}(\tilde{D} - \tilde{R}) && \text{(Lemma 3.4)}\end{aligned}$$

Combining the proof of Theorem 2 in [MVW16], there exist constants C_1, C_2, C_3,c_1, c_2, c_3 such that with probability at least $1 - 2\eta$:

$$\begin{aligned}\|X_{D'} - X_R\|_F^2 &\leq \frac{5}{n_{\min}\Delta_{\min}^2} Tr(R(X_{D'} - X_R))\\ &\leq \frac{10}{n_{\min}\Delta_{\min}^2} \mathcal{F}(\tilde{D}' - \tilde{R})\\ &\leq \frac{140 R_0 \epsilon N}{n_{\min}\Delta_{\min}^2} + C_1 \frac{\min\{k,m\}(\sqrt{N} + c_1\sqrt{m} + \sqrt{c_2 \log(2/\eta)})^2 \sigma_{max}^2}{n_{\min}\Delta_{\min}^2}\\ &+ C_2 \frac{k n_{max} \sigma_{max}^2}{n_{\min}\Delta_{\min}^2} + C_3 \frac{\sqrt{N \log c_3/\eta}}{n_{\min}\Delta_{\min}^2}\end{aligned}$$

If additionally we require $N > \max(c_1^2 m, c_2 \log(2/\eta), \log(c_3/\eta))$, we get

$$\|X_{D'} - X_R\|_F^2 \leq Ck\alpha \frac{(\epsilon_0 R_0 + \sigma_{max}^2 \min\{k,m\})}{\Delta_{\min}^2} \leq Ck\alpha \frac{(\epsilon_0 R_0 + \sigma_{max}^2 k)}{\Delta_{\min}^2}$$

Rearranging gives the result stated in Theorem 3.1. □

Step 2: Denoising. Let P' denote the $m \times N$ matrix whose (a,i)th column is $x'_{a,i}$, where $x'_{a,i}$ corresponds to where $x_{a,i}$ is after we make $\epsilon_0 N$ movements, each up to distance R_0. Then $P'X'_D$ is an $m \times N$ matrix whose columns are estimates for the centers of clusters. We show that these estimated centers are close to the centroids of clusters.

Theorem 3.5. *Let $c'_{a,i}$, $\tilde{\gamma}_a$ respectively denote the (a,i)th column of $P'X'_D$ and PX_R, then*

$$\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n} \|c'_{a,i} - \tilde{\gamma}_a\|_2^2 \lesssim (R_0\epsilon_0 + \frac{\|\Gamma\|_{2\to 2}^2}{k}) k \frac{(\epsilon_0 R_0 + \sigma_{max}^2 k)}{\Delta_{\min}^2}$$

with high probability as $n \to \infty$. Here, Γ is the "shape matrix" whose ath column is $\tilde{\gamma}_a - \frac{1}{k}\sum_{b=1}^{k} \tilde{\gamma}_b$.

For comparison, $\mathbb{E}[\frac{1}{N}\sum_{a=1}^{k}\sum_{i=1}^{n} \|x_{a,i} - \gamma_a\|_2^2] = m\bar{\sigma}^2$, meaning the $c'_{a,i}$ serves as "denoised" versions of the $x_{a,i}$ provided that $\|\Gamma\|_{2\to 2}$ and R_0 are not too large compared to $\Delta_{\min}$, and that $\sigma_{\max}$ is not too large compared to $\bar{\sigma}$.

Proof. By triangular inequality of Frobenius norm, we have:

$$\sum_{a=1}^{k}\sum_{i=1}^{n}\|c_{a,i}-\tilde{\gamma}_a\|_2^2 = \|P'X'_D - PX_R\|_F^2$$
$$\leq (\|P'X'_D - PX'_D\|_F + \|PX'_D - PX_R\|_F)^2$$

We will introduce our own argument for bounding the first term, and for the second term we will follow the argument of Theorem 11 in [MVW16].

Using some matrix norm inequalities, we derive:

$$\|P'X'_D - PX'_D\|_F = \|X'_D(P'-P)^T\|_F \leq \|X'_D\|_{2\to 2}\|P'-P\|_F$$
$$\leq (\|X'_D - X_R\|_{2\to 2} + \|X_R\|_{2\to 2})\sqrt{R_0\epsilon_0 N}$$
$$\leq (\|X'_D - X_R\|_F + 1)\sqrt{R_0\epsilon_0 N}$$

For $\|PX'_D - PX_R\|_F$, Section 4: Denoising in [MVW16] specialized to spherical gaussians, because they used a matrix concentration result for spherical gaussians from Vershynin [Ver12]. By instead using the matrix concentration result from Vershynin's Remark 5.40, we are able to present results in higher generality than the original paper:

Assume each subgaussian has the same number n of samples, then the result from Step 1: Approximation can be simplified into:

$$\|X'_D - X_R\|_F^2 \leq Ck\frac{(\epsilon_0 R_0 + \sigma_{max}^2 k)}{\Delta_{\min}^2}$$

Following the same idea in [MVW16], without loss of generality, we assume $\sum_{a=1}^{k}\tilde{\gamma}_a = 0$. Then,

$$\|PX'_D - PX_R\|_F \leq \|P\|_{2\to 2}\|X'_D - X_R\|_F$$

Decompose $P = \Gamma \otimes 1^T + G$, where columns of G are independent random subgaussian vectors in $\mathbb{R}^m$ with second moment matrix Σ. From [MVW16],

$$\|\Gamma \otimes 1^T\|_{2\to 2}^2 = n\|\Gamma\|_{2\to 2}^2 \geq \Delta_{\min}^2/2$$

and for every $t \geq 0$, Remark 5.40 in [Ver12] gives that

$$\|G\|_{2\to 2}^2 = \|G^TG\|_{2\to 2} \leq N(\|\frac{1}{N}G^TG - \Sigma\|_{2\to 2} + \|\Sigma\|_{2\to 2})$$
$$\leq N(max(\delta, \delta^2) + \sigma_{max}^2)$$

with probability at least $1 - 2e^{-ct^2}$ for some c, and $\delta = C\sqrt{\frac{m}{N}} + \frac{t}{N}$. Therefore, when $N \gg m$, $\|G\|_{2\to 2}^2 \lesssim N\sigma_{max}^2$. Estimate $\|P\|_{2\to 2}$ from triangular inequality, assuming $\Delta_{\min}^2 \geq \frac{C}{\epsilon}k\alpha(\epsilon_0 R_0 + \sigma_{max}^2 k)$ as stated in Theorem 3.1,

$$\|P\|_{2\to 2} \lesssim \sqrt{\frac{N}{k}}\|\Gamma\|_{2\to 2} \Longrightarrow \|PX'_D - PX_R\|_F \lesssim \sqrt{\frac{N}{k}}\|\Gamma\|_{2\to 2}\|X'_D - X_R\|_F$$

Combining the two terms, we derive

$$\begin{aligned}\sum_{a=1}^{k}\sum_{i=1}^{n}\|c_{a,i}-\tilde{\gamma}_a\|_2^2 &\leq (\|P'X'_D - PX'_D\|_F + \|PX'_D - PX_R\|_F)^2 \\ &\lesssim R_0\epsilon_0 N\|X'_D - X_R\|_F^2 + \frac{N}{k}\|\Gamma\|_{2\to 2}^2\|X'_D - X_R\|_F^2 \\ &\lesssim (R_0\epsilon_0 N + \frac{N}{k}\|\Gamma\|_{2\to 2}^2)k\frac{(\epsilon_0 R_0 + \sigma_{max}^2 k)}{\Delta_{\min}^2}\end{aligned}$$

Divide both sides by N gives the result as stated in Theorem 3.5. □

4 Conclusion and Future Research

In this paper, we examined the robustness of semi-definite programs for partial recovery by showing that the estimate centers are close to the theoretical centroids under the relax-and-round k-means clustering procedure first introduced in [MVW16]. For future research, we would find a way to measure accuracy and look at how well the procedure works when we move up to $\epsilon_0 N$ points, and we could possibly apply the method on some actual data sets to see how well it recovers the centers.

Acknowledgements

I would like to thank Amelia Perry for mentoring the project and for all the help and guidance. I would like to thank Prof. Ankur Moitra for supervising this project. Finally, I would like to thank the MIT Math Department and the UROP+ program for making this project possible.

References

[ABH16] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

[Das99] S. Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644, 1999.

[GV15] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via Grothendieck's inequality. *Probability Theory and Related Fields*, pages 1–25, 2015.

[HWX16] Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016.

[Llo82] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, Mar 1982.

[Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[MPW16] Ankur Moitra, William Perry, and Alexander S Wein. How robust are reconstruction thresholds for community detection? In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 828–841. ACM, 2016.

[MVW16] Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *arXiv preprint arXiv:1602.06612*, 2016.

[PW07] Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1):186–205, 2007.

[Ver12] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing*, chapter 5, pages 210–268. Cambridge University Press, Cambridge, 2012.