

# Appendix to “Scalable Clustering: Large Scale Unsupervised Learning of Gaussian Mixture Models with Outliers” published in the Journal of Computational and Graphical Statistics

Yijia Zhou, Kyle A. Gallivan

Department of Mathematics, Florida State University  
and

Adrian Barbu

Department of Statistics, Florida State University

September 30, 2024

## 1 Preliminaries

**Lemma 1.** (From (Johnstone and Lu, 2009), (A.2)) If  $Z_1, \dots, Z_n$  are i.i.d Gaussian random variables  $Z_i \sim \mathcal{N}(0, 1)$ , then for any  $0 \leq \epsilon < 1$ ,

$$\mathbb{P} \left( \sum_{i=1}^n Z_i^2 \leq n(1 - \epsilon) \right) \leq \exp \{ -n\epsilon^2/4 \}.$$

**Lemma 2.** (From (Johnstone and Lu, 2009), (A.3)) If  $Z_1, \dots, Z_n$  are i.i.d Gaussian random variables  $Z_i \sim \mathcal{N}(0, 1)$ , then for any  $0 \leq \epsilon < \frac{1}{2}$ ,

$$\mathbb{P} \left( \sum_{i=1}^n Z_i^2 \geq n(1 + \epsilon) \right) \leq \exp \{ -3n\epsilon^2/16 \}.$$

**Corollary 1.** If  $\mathbf{x} = (X_1, \dots, X_p)$  is a multivariate Gaussian random variable  $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I_p)$ , then  $\mathbb{E}(\|\mathbf{x}\|^2) = p$  and for any  $0 \leq \epsilon < 1$ ,

$$\mathbb{P}(\|\mathbf{x}\|^2 \leq p(1 - \epsilon)) \leq \exp\{-p\epsilon^2/4\}.$$

*Proof.* Follows from Lemma 1 above taking  $Z_i = X_i, i = 1, \dots, p$ . □

**Corollary 2.** *If  $\mathbf{x} = (X_1, \dots, X_p)$  is a multivariate Gaussian random variable  $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I_p)$ , then  $\mathbb{E}(\|\mathbf{x}\|^2) = p$  and for any  $0 \leq \epsilon < \frac{1}{2}$ ,*

$$\mathbb{P}(\|\mathbf{x}\|^2 \geq p(1 + \epsilon)) \leq \exp\{-3p\epsilon^2/16\}.$$

*Proof.* Follows from Lemma 2 above taking  $Z_i = X_i, i = 1, \dots, p$ .  $\square$

Using these results, it follows that with high probability the negatives are well-separated from each other.

**Corollary 3** (Separation between negatives). *For two negatives  $\mathbf{x}_i$  and  $\mathbf{x}_k$ , with probability at least  $1 - \exp\{-p\epsilon^2/4\}$  where  $0 \leq \epsilon < 1$ , the separation satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > 2p(1 - \epsilon), \quad \mathbf{x}_i, \mathbf{x}_k \in H.$$

*Proof.* Since  $\mathbf{x}_k \sim \mathcal{N}(\mathbf{0}, I_p)$  and  $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I_p)$ , then  $\mathbf{x}_i - \mathbf{x}_k \sim \mathcal{N}(\mathbf{0}, 2I_p)$ , thus  $\mathbb{E}(\|\mathbf{x}_i - \mathbf{x}_k\|^2) = 2p$ . According to Corollary 1, it follows that

$$\mathbb{P}\left(\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2} \leq p(1 - \epsilon)\right) \leq \exp\{-p\epsilon^2/4\}.$$

Then with high probability at least  $1 - \exp\{-p\epsilon^2/4\}$ , the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > 2p(1 - \epsilon).$$

$\square$

It then follows that the positives from the same cluster are within a certain radius from each other with high probability.

**Corollary 4** (Concentration of positives in the same cluster). *For any positive cluster  $S_j$  with mean  $\boldsymbol{\mu}_j$  and covariance matrix  $\sigma_j^2 I_p$ , with probability at least  $1 - \exp\{-3p\epsilon^2/16\}$  where  $0 \leq \epsilon < \frac{1}{2}$ , the concentration is bounded as*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 < 2(1 + \epsilon)p\sigma_j^2, \quad \mathbf{x}_i, \mathbf{x}_k \in S_j.$$

*Proof.* Since  $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p)$  and  $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p)$ , then  $\mathbf{x}_i - \mathbf{x}_k \sim \mathcal{N}(\mathbf{0}, 2\sigma_j^2 I_p)$ , thus  $\mathbb{E}(\|\mathbf{x}_i - \mathbf{x}_k\|^2) = 2p\sigma_j^2$ . According to Corollary 2, it follows that

$$\mathbb{P}\left(\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2\sigma_j^2} \geq p(1 + \epsilon)\right) \leq \exp\{-3p\epsilon^2/16\}.$$

Therefore, with probability at least  $1 - \exp\{-3p\epsilon^2/16\}$ , the concentration is bounded as

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 < 2(1 + \epsilon)p\sigma_j^2.$$

$\square$

Then, it is proven that the positives are well-separated from the negatives with high probability.

**Corollary 5** (Separation between positives and negatives). *For negative  $\mathbf{x}_i$  and positive  $\mathbf{x}_k$  from cluster  $S_j$  with mean  $\boldsymbol{\mu}_j$  and covariance matrix  $\sigma_j^2 I_p$ , with probability at least  $1 - \exp\{-p\epsilon^2/4\}$  where  $0 \leq \epsilon < 1$ , the separation satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(2 + \sigma_j^2)(1 - \epsilon), \quad \mathbf{x}_i \in H, \mathbf{x}_k \in S_j.$$

*Proof.* Since  $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p)$  and  $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I_p)$ , then  $\mathbf{x}_i - \mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p + I_p)$ , thus  $\mathbf{x}_i - \mathbf{x}_k = \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_1 \sqrt{\sigma_j^2 + 1}$  with  $\boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}, I_p)$ . Since  $\boldsymbol{\mu}_j \sim \mathcal{N}(\mathbf{0}, I_p)$ , then  $\mathbf{x}_i - \mathbf{x}_k$  is a Gaussian with  $\mathbb{E}(\mathbf{x}_i - \mathbf{x}_k) = \mathbf{0}$  and

$$\mathbb{E}(\|\mathbf{x}_i - \mathbf{x}_k\|^2) = \mathbb{E} \left[ \left( \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_1 \sqrt{\sigma_j^2 + 1} \right)^T \left( \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_1 \sqrt{\sigma_j^2 + 1} \right) \right] = \mathbb{E}(\|\boldsymbol{\mu}_j\|^2) + (\sigma_j^2 + 1) \mathbb{E}(\boldsymbol{\epsilon}_1^T \boldsymbol{\epsilon}_1),$$

thus

$$\mathbb{E}(\|\mathbf{x}_i - \mathbf{x}_k\|^2) = p + (\sigma_j^2 + 1) \mathbb{E}(\|\boldsymbol{\epsilon}_1\|^2) = (2 + \sigma_j^2)p.$$

According to Corollary 1, it follows immediately that,

$$\mathbb{P} \left( \frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2 + \sigma_j^2} \leq p(1 - \epsilon) \right) \leq \exp \{-p\epsilon^2/4\}.$$

Then with probability at least  $1 - \exp\{-p\epsilon^2/4\}$ , the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(2 + \sigma_j^2)(1 - \epsilon).$$

□

Moreover, positives from different clusters are also well-separated from each other with high probability.

**Corollary 6** (Separation between positives in different clusters). *For positive  $\mathbf{x}_i$  from cluster  $S_i$  with true mean  $\boldsymbol{\mu}_i$  and covariance matrix  $\sigma_i^2 I_p$  and positive  $\mathbf{x}_k$  from another cluster  $S_j$  with true mean  $\boldsymbol{\mu}_j$  and covariance matrix  $\sigma_j^2 I_p$ , with probability at least  $1 - \exp\{-p\epsilon^2/4\}$  where  $0 \leq \epsilon < 1$ , the separation satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(2 + \sigma_i^2 + \sigma_j^2)(1 - \epsilon), \quad \mathbf{x}_i \in S_i, \mathbf{x}_k \in S_j.$$

*Proof.* Since  $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_j, \sigma_j^2 I_p)$  and  $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I_p)$ , then  $\mathbf{x}_i - \mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i, \sigma_j^2 I_p + \sigma_i^2 I_p)$ , thus  $\mathbf{x}_i - \mathbf{x}_k = \boldsymbol{\mu}_j - \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_1 \sqrt{\sigma_j^2 + \sigma_i^2}$  with  $\boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}, I_p)$ . Since  $\boldsymbol{\mu}_j \sim \mathcal{N}(\mathbf{0}, I_p)$  and  $\boldsymbol{\mu}_i \sim \mathcal{N}(\mathbf{0}, I_p)$ , then  $\boldsymbol{\mu}_j - \boldsymbol{\mu}_i \sim \mathcal{N}(\mathbf{0}, 2I_p)$ , then  $\mathbf{x}_i - \mathbf{x}_k$  is a Gaussian with  $\mathbb{E}(\mathbf{x}_i - \mathbf{x}_k) = \mathbf{0}$  and

$$\mathbb{E}(\|\mathbf{x}_i - \mathbf{x}_k\|^2) = \mathbb{E} \left[ \left( \boldsymbol{\mu}_j - \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_1 \sqrt{\sigma_j^2 + \sigma_i^2} \right)^T \left( \boldsymbol{\mu}_j - \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_1 \sqrt{\sigma_j^2 + \sigma_i^2} \right) \right],$$

thus

$$\mathbb{E}(\|\mathbf{x}_i - \mathbf{x}_k\|^2) = \mathbb{E}(\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\|^2) + (\sigma_j^2 + \sigma_i^2) \mathbb{E}(\boldsymbol{\epsilon}_1^T \boldsymbol{\epsilon}_1) = 2p + (\sigma_j^2 + \sigma_i^2) \mathbb{E}(\|\boldsymbol{\epsilon}_1\|^2) = (2 + \sigma_i^2 + \sigma_j^2)p.$$

According to Corollary 1, it follows that

$$\mathbb{P}(\|\mathbf{x}_i - \mathbf{x}_k\|^2 \leq p(2 + \sigma_i^2 + \sigma_j^2)(1 - \epsilon)) \leq \exp\{-p\epsilon^2/4\}.$$

Then with probability at least  $1 - \exp\{-p\epsilon^2/4\}$ , the separation satisfies

$$\|\mathbf{x}_i - \mathbf{x}_k\|^2 > p(2 + \sigma_i^2 + \sigma_j^2)(1 - \epsilon).$$

□

The previous corollaries will be used to prove that with high probability, all positives from each cluster are within  $2(1 + \epsilon)p\rho^2$  of each other, and  $2(1 + \epsilon)p\rho^2$  away from the other clusters and from the negatives.

## 2 Proof of Key Lemmas and Propositions

**Proposition 1.** *Given  $N$  samples from a GMM with outliers, and  $\sigma_{max} \leq \rho < \sqrt{\frac{1-\epsilon}{1+\epsilon}}$ , then with probability at least  $1 - 2N^2 \exp\{-p\epsilon^2/4\} - N^2 \exp\{-3p\epsilon^2/16\}$ , the distance between positives within a cluster satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2(1 + \epsilon)p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = l(\mathbf{x}_j) > 0,$$

*and the distance between positives from a cluster and other samples not in that cluster satisfies*

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2(1 + \epsilon)p\rho^2 \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) \neq l(\mathbf{x}_i) > 0.$$

*Proof.* From Corollary 4, with probability at least  $1 - \exp\{-3p\epsilon^2/16\}$ , the distance between two positives in the same cluster is bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2(1 + \epsilon)p\sigma_{l(\mathbf{x}_i)}^2, \quad l(\mathbf{x}_i) = l(\mathbf{x}_j) > 0.$$

Using the union bound, with probability at least  $1 - N^2 \exp\{-3p\epsilon^2/16\}$ , the distances between all positives in the same cluster are bounded as

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2(1 + \epsilon)p\sigma_{l(\mathbf{x}_i)}^2 \leq 2(1 + \epsilon)p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_i) = l(\mathbf{x}_j) > 0.$$

From Corollary 5, with probability at least  $1 - \exp\{-p\epsilon^2/4\}$ , the distance between a positive and a negative satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(2 + \sigma_{l(\mathbf{x}_i)}^2)(1 - \epsilon), \quad l(\mathbf{x}_i) > 0, l(\mathbf{x}_j) = -1.$$

Using the union bound, with probability at least  $1 - N^2 \exp\{-p\epsilon^2/4\}$ , the distance between any positive and negative satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(2 + \sigma_{l(\mathbf{x}_i)}^2)(1 - \epsilon), \quad \forall \mathbf{x}_i, \mathbf{x}_j, \text{ s.t. } l(\mathbf{x}_i) > 0, l(\mathbf{x}_j) = -1.$$

Given  $\sigma_{max} \leq \rho < \sqrt{\frac{(2+\sigma_k^2)(1-\epsilon)}{2(1+\epsilon)}}$ , since  $0 < \sigma_k < 1$ , therefore, with  $\sigma_{max} \leq \rho < \sqrt{\frac{1-\epsilon}{1+\epsilon}}$ , then with probability at least  $1 - N^2 \exp\{-p\epsilon^2/4\}$ , the distance between any positive and negative satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2(1 + \epsilon)p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j, \text{ s.t. } l(\mathbf{x}_i) > 0, l(\mathbf{x}_j) = -1.$$

From Corollary 6, with probability at least  $1 - \exp\{-p\epsilon^2/4\}$ , the distance between two positives from different clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(2 + \sigma_{l(\mathbf{x}_j)}^2 + \sigma_{l(\mathbf{x}_i)}^2)(1 - \epsilon), \quad 0 < l(\mathbf{x}_i) \neq l(\mathbf{x}_j) > 0.$$

Using the union bound, with probability at least  $1 - N^2 \exp\{-p\epsilon^2/4\}$ , the distance between any two positives from different clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > p(2 + \sigma_{l(\mathbf{x}_j)}^2 + \sigma_{l(\mathbf{x}_i)}^2)(1 - \epsilon), \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } 0 < l(\mathbf{x}_i) \neq l(\mathbf{x}_j) > 0.$$

Given  $\sigma_{max} \leq \rho < \sqrt{\frac{(2+\sigma_{l(\mathbf{x}_j)}^2+\sigma_{l(\mathbf{x}_i)}^2)(1-\epsilon)}{2(1+\epsilon)}}$ , since  $0 < \sigma_{l(\mathbf{x}_i)}, \sigma_{l(\mathbf{x}_j)} < 1$ , therefore, with  $\sigma_{max} \leq \rho < \sqrt{\frac{1-\epsilon}{1+\epsilon}}$ , then with probability at least  $1 - N^2 \exp\{-p\epsilon^2/4\}$ , the distance between any two positives from different clusters satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2(1 + \epsilon)p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } 0 < l(\mathbf{x}_i) \neq l(\mathbf{x}_j) > 0.$$

Therefore, with probability at least  $1 - 2N^2 \exp\{-p\epsilon^2/4\}$ , the distance between any positive and any sample not from that cluster satisfies

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2(1 + \epsilon)p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) \neq l(\mathbf{x}_i) > 0.$$

Therefore, with probability at least  $1 - 2N^2 \exp\{-p\epsilon^2/4\} - N^2 \exp\{-3p\epsilon^2/16\}$ , the following bounds on positives within a cluster and between clusters are satisfied

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 > 2(1 + \epsilon)p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) \neq l(\mathbf{x}_i) > 0,$$

and

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 < 2(1 + \epsilon)p\rho^2, \quad \forall \mathbf{x}_i, \mathbf{x}_j \text{ s.t. } l(\mathbf{x}_j) = l(\mathbf{x}_i) > 0.$$

□

Finally, a bound for the probability that a sample  $S$  has at least one element from each positive cluster is proven.

**Lemma 3.** *If the clusters have weights  $w_1, \dots, w_m$ , with  $\sum_{k=1}^m w_k \leq 1$ , then the probability that a sample  $S$  of size  $|S| = n$  contains at least one observation from each cluster is at least*

$$\mathbb{P}(|\{\mathbf{x} \in S, l(\mathbf{x}) = k\}| \geq 1, \forall k = \overline{1, m}) \geq 1 - \sum_{k=1}^m (1 - w_k)^n.$$

*Proof.* The probability that  $S$  contains no elements from cluster  $k$  is

$$\mathbb{P}(l(\mathbf{x}) \neq k, \forall \mathbf{x} \in S) = (1 - w_k)^n.$$

Then using the union bound, the probability that there is a  $k$  such that  $S$  does not contain any elements from cluster  $k$  is

$$\mathbb{P}(\exists k, l(\mathbf{x}) \neq k, \forall \mathbf{x} \in S) \leq \sum_{k=1}^m (1 - w_k)^n,$$

which implies the result.  $\square$

### 3 Proofs of Loss Bounds

In this section, the concentration and separation results from Appendix 1 and Appendix 2 are used to obtain bounds on the loss function values.

First, it is proven that with high probability, the loss value of a negative is  $-F$ .

**Proposition 2.** *Given  $N$  samples from a GMM with outliers, and  $\sigma_{max} \leq \rho < \sqrt{\frac{1-\epsilon}{1+\epsilon}}$ , then for a negative sample  $\mathbf{x}_j, l(\mathbf{x}_j) = -1$ , with probability at least  $1 - 2N \exp\{-p\epsilon^2/4\}$ , the loss satisfies  $L(\mathbf{x}_j; \rho) = -F = -2(1 + \epsilon)$ .*

*Proof.* From Corollary 3, for  $\mathbf{x}_i, l(\mathbf{x}_i) = -1$ , with probability at least  $1 - \exp\{-p\epsilon^2/4\}$ , the distance between a negative and  $\mathbf{x}_j$  satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 2p(1 - \epsilon), \quad l(\mathbf{x}_i) = -1, i \neq j.$$

Using the union bound, with probability at least  $1 - N \exp\{-p\epsilon^2/4\}$ , the distance between any other negative and  $\mathbf{x}_j$  satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 2p(1 - \epsilon), \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = -1, i \neq j.$$

Given  $\sigma_{max} \leq \rho < \sqrt{\frac{1-\epsilon}{1+\epsilon}}$ , then with probability at least  $1 - N \exp\{-p\epsilon^2/4\}$ , the distance between any other negative and  $\mathbf{x}_j$  satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 2p(1 - \epsilon) > 2p(1 + \epsilon)\rho^2, \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) = -1, i \neq j.$$

From Corollary 5, for  $\mathbf{x}_i, l(\mathbf{x}_i) > 0$ , with probability at least  $1 - \exp\{-p\epsilon^2/4\}$ , the distance between a positive and  $\mathbf{x}_j$  satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(2 + \sigma_{l(\mathbf{x}_i)}^2)(1 - \epsilon), \quad l(\mathbf{x}_i) > 0.$$

Using the union bound, with probability at least  $1 - N \exp\{-p\epsilon^2/4\}$ , the distance between any positive and  $\mathbf{x}_j$  satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(2 + \sigma_{l(\mathbf{x}_i)}^2)(1 - \epsilon), \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) > 0.$$

Given  $\sigma_{max} \leq \rho < \sqrt{\frac{1-\epsilon}{1+\epsilon}}$ , then with probability at least  $1 - N \exp\{-p\epsilon^2/4\}$ , the distance between any positive and  $\mathbf{x}_j$  satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > p(2 + \sigma_{l(\mathbf{x}_i)}^2)(1 - \epsilon) > 2p(1 + \epsilon)\rho^2 \quad \forall \mathbf{x}_i, l(\mathbf{x}_i) > 0.$$

Therefore, with probability at least  $1 - 2N \exp\{-p\epsilon^2/4\}$ , the distance between any other sample and  $\mathbf{x}_j$  satisfies

$$\|\mathbf{x}_j - \mathbf{x}_i\|^2 > 2(1 + \epsilon)p\rho^2, \quad \forall i \neq j.$$

Therefore, with probability at least  $1 - 2N \exp\{-p\epsilon^2/4\}$ , it follows that

$$\ell(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho) = \min\left(\frac{\|\mathbf{x}_j - \mathbf{x}_i\|^2}{p\rho^2} - F, 0\right) = 0, \forall i \neq j.$$

Therefore, with probability at least  $1 - 2N \exp\{-p\epsilon^2/4\}$ , the loss satisfies

$$L(\mathbf{x}_j; \rho) = \sum_{i=1}^N \ell(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho) = -F = -2(1 + \epsilon),$$

since  $\ell(\|\mathbf{x}_j - \mathbf{x}_j\|; \rho) = \ell(0; \rho) = -F$ . □

Next, it is proven with high probability, the loss value of a positive is less than  $-F$ .

**Proposition 3.** *Given  $N$  samples from a GMM with outliers, and  $\sigma_{max} \leq \rho < \sqrt{\frac{1-\epsilon}{1+\epsilon}}$ , then for a positive sample  $\mathbf{x}_j, l(\mathbf{x}_j) = k > 0$ , with probability  $1 - 2 \exp\{-p/128\} - \exp\{-(N - 1)w_k\}$ , the loss is bounded as  $L(\mathbf{x}_j; \rho) < -F$ .*

*Proof.* The probability that a sample of size  $N - 1$  contains no elements from cluster  $S_k$  is

$$(1 - w_k)^{N-1} \leq \exp\{-(N - 1)w_k\}.$$

Therefore, with probability at least  $1 - \exp\{-(N - 1)w_k\}$ , there is at least one more sample  $\mathbf{x}_a, a \neq j$  besides  $\mathbf{x}_j$  in cluster  $S_k$ .

From Corollary 4, with probability at least  $1 - \exp\{-3p\epsilon^2/16\}$ , the distance between  $\mathbf{x}_a$  and  $\mathbf{x}_j$  is bounded as

$$\|\mathbf{x}_j - \mathbf{x}_a\|^2 < 2(1 + \epsilon)p\sigma_k^2.$$

Given  $\sigma_{max} \leq \rho < \sqrt{\frac{1-\epsilon}{1+\epsilon}}$ , with probability at least  $1 - \exp\{-3p\epsilon^2/16\}$ , the distance between  $\mathbf{x}_a$  and  $\mathbf{x}_j$  is bounded as

$$\|\mathbf{x}_j - \mathbf{x}_a\|^2 < 2(1 + \epsilon)p\sigma_k^2 \leq 2(1 + \epsilon)p\sigma_{max}^2 \leq 2(1 + \epsilon)p\rho^2.$$

Therefore, with probability at least  $1 - \exp\{-3p\epsilon^2/16\} - \exp\{-(N - 1)w_k\}$ , the following equality holds

$$\ell(\|\mathbf{x}_j - \mathbf{x}_a\|; \rho) = \min\left(\frac{\|\mathbf{x}_j - \mathbf{x}_a\|^2}{p\rho^2} - F, 0\right) < 0.$$

Therefore, with probability at least  $1 - \exp\{-3p\epsilon^2/16\} - \exp\{-(N-1)w_k\}$ , the loss is bounded above as

$$L(\mathbf{x}_j; \rho) = \sum_{i=1}^N \ell(\|\mathbf{x}_j - \mathbf{x}_i\|; \rho) \leq \ell(\|\mathbf{x}_j - \mathbf{x}_a\|; \rho) + \ell(\|\mathbf{x}_j - \mathbf{x}_j\|; \rho) < -F.$$

□

**Proposition 4.** *Given  $N$  samples from a GMM with outliers, with  $w_i \geq a/m, i = \overline{1, m}$  for some  $a > 0$  and  $\sigma_{\max} \leq \rho < \sqrt{\frac{1-\epsilon}{1+\epsilon}}$ , randomly select a set  $S$  of  $|S| = n$  subsamples from it, then with probability at least  $1 - m \exp\{-na/m\} - m \exp\{-3p\epsilon^2/16\} - m \exp\{-a(N-1)/m\}$  for each  $k = \overline{1, m}$  there exists  $\mathbf{x}_j \in S_k = \{\mathbf{x} \in S, l(\mathbf{x}) = k\}$  such that  $L(\mathbf{x}_j, \rho) < -F$ .*

*Proof.* According to Lemma 3, the probability that a sample  $S$  of size  $n$  contains at least one observation from each cluster is

$$1 - \sum_{i=1}^m (1 - w_i)^n \geq 1 - m(1 - a/m)^n \geq 1 - m \exp\{-na/m\},$$

and without loss of generality let  $\mathbf{x}_j$  be the observation from cluster  $S_k, k = \overline{1, m}$ . Applying Proposition 3 repeatedly to these  $m$  samples and using the union bound, with probability at least  $1 - 2m \exp\{-p/128\} - \sum_{i=1}^m \exp\{-(N-1)w_i\}$ , the loss is bounded as  $L(\mathbf{x}_j, \rho) < -F$ . Since  $\forall w_i \geq a/m$ , therefore  $\sum_{i=1}^m \exp\{-(N-1)w_i\} \leq m \exp\{-a(N-1)/m\}$ . Therefore, with probability at least  $1 - m \exp\{-na/m\} - m \exp\{-3p\epsilon^2/16\} - m \exp\{-a(N-1)/m\}$ , for each  $k = \overline{1, m}$  there exists  $\mathbf{x}_j \in S_k$  such that  $L(\mathbf{x}_j, \rho) < -F$ . □

## 4 Proofs of Theorem 1 and Corollary 1 of the manuscript

In this section, the proofs of Theorem 1 and Corollary 1 of the manuscript are given.

*Proof. of Theorem 1* of the manuscript. From Proposition 2, for a negative sample  $\mathbf{x}_j$ , with probability at least  $1 - 2N \exp\{-p\epsilon^2/4\}$ ,  $L(\mathbf{x}_j, \rho) = -F = -2(1 + \epsilon)$ , then for all the negatives, with probability at least  $1 - 2N^2 \exp\{-p\epsilon^2/4\}$ , the loss satisfies  $L(\mathbf{x}_j, \rho) = -F = -2(1 + \epsilon)$ .

From Proposition 4, with probability at least  $1 - m \exp\{-na/m\} - m \exp\{-3p\epsilon^2/16\} - m \exp\{-a(N-1)/m\}$ , for each  $k = \overline{1, m}$  there is  $\mathbf{x}_j \in S_k, L(\mathbf{x}_j, \rho) < -F$ .

Combining Proposition 2 and Proposition 4, with probability at least  $1 - 2N^2 \exp\{-p\epsilon^2/4\} - m \exp\{-na/m\} - m \exp\{-3p\epsilon^2/16\} - m \exp\{-a(N-1)/m\}$ , only positives will be selected at step 8 of SCRLM.

From Proposition 1, with probability at least  $1 - 2N^2 \exp\{-p\epsilon^2/4\} - N^2 \exp\{-3p\epsilon^2/16\}$ , all positives are correctly identified in Steps 9 and 17 and removed from negatives.

So with probability at least

$$1 - 4N^2 \exp\{-p\epsilon^2/4\} - (N^2 + m) \exp\{-3p\epsilon^2/16\} - m \exp\{-na/m\} - m \exp\{-a(N-1)/m\},$$

SCRLM will have 100% accuracy. □



*Proof. of Corollary 1* of the manuscript. The condition

$$p > \frac{4}{\epsilon^2}(\log N^2 + \log \frac{16}{\delta})$$

is equivalent to

$$4N^2 \exp\{-p\epsilon^2/4\} < \frac{\delta}{4}.$$

The condition

$$p > \frac{16}{3\epsilon^2}(\log(N^2 + m) + \log \frac{4}{\delta})$$

is equivalent to

$$(N^2 + m) \exp\{-3p\epsilon^2/16\} < \frac{\delta}{4}.$$

The condition

$$n > \frac{m}{a}(\log m + \log \frac{4}{\delta})$$

is equivalent to

$$m \exp(-na/m) < \frac{\delta}{4}.$$

Finally, the condition

$$N > \frac{m}{a}(\log m + \log \frac{4}{\delta}) + 1.$$

is equivalent to:

$$m \exp\{-a(N-1)/m\} < \frac{\delta}{4}.$$

These conditions together imply that

$$1 - 4N^2 \exp\{-p\epsilon^2/4\} - (N^2 + m) \exp\{-3p\epsilon^2/16\} - m \exp\{-na/m\} - m \exp\{-a(N-1)/m\} > 1 - \delta.$$

According to Theorem 1, SCRLM has 100% accuracy with probability at least  $1 - \delta$ .  $\square$

## References

Johnstone, I. M. and Lu, A. Y. (2009), “On consistency and sparsity for principal components analysis in high dimensions,” *Journal of the American Statistical Association*, 104, 682–693.