A Spectral Algorithm for Learning Mixture Models

Santosh Vempala* Grant Wang†

Abstract

We show that a simple spectral algorithm for learning a mixture of k spherical Gaussians in $\mathbb{R}^n$ works remarkably well — it succeeds in identifying the Gaussians assuming essentially the minimum possible separation between their centers that keeps them unique (solving an open problem of [1]). The sample complexity and running time are polynomial in both n and k. The algorithm can be applied to the more general problem of learning a mixture of "weakly isotropic" distributions (e.g. a mixture of uniform distributions on cubes).

1 Introduction

Learning a mixture of distributions is a classical problem in statistics and learning theory (see [10, 14]); more recently, it has also been proposed as a model for clustering. In the basic version of the problem we are given random samples from a mixture of k distributions, $F_1, \ldots, F_k$. Each sample is drawn independently with probability w_i from the i'th distribution. The numbers $w_1, \ldots, w_k$ are called the mixing weights. The problem is to classify the random samples according to which distribution they come from (and thereby infer the mixing weights, means and other properties of the underlying distributions).

An important case of this problem is when each underlying distribution is a Gaussian. In this case, the goal is to find the mean and covariances of each Gaussian (along with the mixing weights). This problem seems to be of great practical interest and many heuristics have been used to solve it. The most famous among them is the EM algorithm [5]. Unfortunately EM is a local search heuristic that can fail.

A special case of the problem is when the Gaussians are assumed to be spherical, i.e. the variance is the same in any direction. In recent years, there has been substantial progress in developing polynomial-time algorithms for this special case, by making assumptions on the separation between the means of the Gaussians. This separation condition is crucial, so we proceed to make it explicit. Let $F_1, \ldots, F_k$ be spherical Gaussians over $\mathbb{R}^n$ with mean vectors $\mu_1, \ldots, \mu_k$ and variances $\sigma_1^2, \ldots, \sigma_k^2$. We will refer to $\sigma_i\sqrt{n}$ as the *radius* of F_i and $||\mu_i - \mu_j||$ as the *separation* between F_i and F_j. If the pairwise separation is larger than the radii, then points from different Gaussians are isolated in space and easy to classify. On the other hand, if the separation is very small, then the classification problem might not have a unique solution.

1.1 Previous work

Dasgupta [3] used random projection to learn a mixture of spherical Gaussians provided they are essentially non-overlapping, i.e. the overlap in probability mass is exponentially small in n. His algorithm is polynomial-time provided the smallest mixing weight $w_{\min}$ is $\Omega(1/k)$ and the separation is

$$||\mu_i - \mu_j|| \geq C \max\{\sigma_i, \sigma_j\} n^{\frac{1}{2}}$$

*Department of Mathematics and Laboratory for Computer Science, MIT, `vempala@math.mit.edu`

†Laboratory for Computer Science, MIT, `gjw@theory.lcs.mit.edu`. Both authors are supported in part by NSF Career award CCR-987024.

for a constant C. In other words, the separation is proportional to the larger radius (the algorithm also required all the variances to be within a bounded range). Shortly thereafter, it was shown by Dasgupta and Schulman [4] that a variant of EM works with a smaller separation (along with some technical conditions on the variances):

$$||\mu_i - \mu_j|| > C \max\{\sigma_i, \sigma_j\} n^{\frac{1}{4}} \log^{\frac{1}{4}} (n/w_{\min}) \tag{1}$$

This separation is the minimum at which random points from the same Gaussian can be distinguished from random points from two different Gaussians based on pairwise distances. So points from the Gaussian with (approximately) the smallest variance have the smallest pairwise distances. They can be identified and removed and this can be repeated on the remaining points. Arora and Kannan [1] independently proved similar results for non-spherical Gaussians. They used isoperimetric theorems to obtain distance concentration results for the non-spherical case. At this separation, their algorithm simply identifies all points at roughly the minimum distance from each other as coming from a single Gaussian, removes them and repeats on the remaining data. They also give a version that uses random projection. Learning a mixture of spherical Gaussians at a smaller separation (when distance concentration results are no longer valid) has been an open problem.

1.2 Our results

In order for the solution to the classification problem to be well-defined (i.e. unique with reasonable probability) we need a separation of at least

$$||\mu_i - \mu_j|| > C \max\{\sigma_i, \sigma_j\}.$$

At this separation the overlap in the probability mass is a constant fraction. So in particular, distance concentration results are no longer applicable.

In this paper, we show that a simple spectral algorithm can learn a mixture of k Gaussians at this minimum separation in time polynomial in $k^{O(k)}$ and n. Our main result is that with a slightly larger separation of

$$||\mu_i - \mu_j|| > C \max\{\sigma_i, \sigma_j\} \left((k \log (n/w_{\min}))^{1/4} + (\log (n/w_{\min}))^{1/2} \right) \tag{2}$$

the algorithm is polynomial in both k and n. Note that this condition is almost independent of n and is much weaker than (1) as the dimension (n) gets larger than the number of Gaussians (k).

The main step of the algorithm is to project to essentially the top k right singular vectors of the sample matrix (i.e. its k principal components). This is the rank k subspace that maximizes the squared projections of the samples. The key observation is that with high probability this subspace lies very close to the span of the mean vectors of the underlying distribution. In section 3 we first prove this for the *expected* best subspace. This result holds for any *weakly isotropic*[1] distribution.

Definition 1. *A distribution with mean μ is said to be* weakly-isotropic *if for a random sample $X \in \mathbb{R}^n$ we have*

$$\mathrm{E}[(w \cdot (X - \mu))^2] = \sigma^2 \quad \forall w \in \mathbb{R}^n, ||w|| = 1. \tag{3}$$

In other words, the variance of any 1-dimensional projection is σ^2.

In section 4 we show that with high probability the best subspace is close to the span of the means when the underlying distributions are Gaussians.

[1] The term isotropic [9] also requires that the mean is zero and the variance is 1 along any direction. Here we are allowing a radial scaling and translation for each distribution in the mixture.

On projection, the separation between the mean vectors is preserved. On the other hand, the radius of the distribution projected to any k-dimensional subspace drops by a factor of $\sqrt{\frac{n}{k}}$. Together this has the surprising effect of amplifying the ratio of the separation to the radii while reducing the dimension! After projection, we can apply distance concentration to classify points from the distributions. We prove this for Gaussians in section 5 but these results hold for any weakly isotropic distribution that has good concentration bounds on the distance between sample points.

It is worth noting that this is a problem for which direct random projection does *not* work (see Figures 1,3). Indeed on random projection to d-dimensions, the inter-center distances and the radii scale at the same rate, namely $\sqrt{\frac{n}{d}}$. So in terms of the dimension, the separation condition gets worse.

2 The Spectral Algorithm

The first step of the algorithm is based on the singular value decomposition of a matrix. Any $m \times n$ matrix A can be written as

$$A = \sum_{i=1}^{n} \lambda_i u_i v_i^T$$

where $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_n$ are the singular values of A and u_i, v_i are the left and right singular vectors corresponding to the i'th singular value λ_i. The projection to the top r right singular vectors is

$$A_r = \sum_{i=1}^{r} \lambda_i u_i v_i^T.$$

The key property of the decomposition is that the subspace spanned by the top r right singular vectors is the one that maximizes the norm of the projection of A among all r-dimensional subspaces. The algorithm below for learning mixtures of weakly isotropic distributions uses this projection.

Algorithm.

1. Compute the singular value decomposition of the sample matrix.
2. Let $r = \max\{k, C\log(n/n_{\min})\}$. Project the samples to the rank r subspace spanned by the top r right singular vectors.
3. Perform a distance-based classification in the r-dimensional space.

The classification algorithm in step (3) is spelled out in full detail in section 5 for the case of mixtures of spherical Gaussians. In this case, it suffices to set $C = 1344$ in step (2); the exact constant will vary depending on the underlying distributions. We should also note that this algorithm "learns" the mixture of distributions in the sense that it correctly classifies the samples. Learning other properties of a mixture of Gaussians can be done from a correct classification.

3 The Expected Best Subspace

In this section, we show that in *expectation*, the subspace spanned by the top k singular vectors of the sample matrix is the same subspace spanned by the mean vectors of the distributions. The results of this section hold for any mixture of *weakly isotropic* distributions.

Intuitively, this is true for a single weakly isotropic distribution. Since the distribution is spherically symmetric, it is clear that any vector that passes through the mean maximizes the sum of squared projections. Similarly, *any* k-dimensional subspace passing through the mean would be optimal. Thus, for a mixture, any subspace that passes through all the means would be the best subspace; in particular, the best k-dimensional subspace is the one spanned by the k means. In what follows, we prove this formally for mixtures of weakly isotropic distributions.

It is easy to verify that (3) is equivalent to the following:

1. For each coordinate i, $\mathrm{E}[(X_i - \mu_i)^2] = \sigma^2$.
2. Each pair of coordinates i, j are uncorrelated, i.e. $\mathrm{E}[X_i X_j] = \mathrm{E}[X_i]\mathrm{E}[X_j]$.

Suppose we sample a random point X from such a distribution with mean vector μ and variance σ^2 in every direction. Then we have the following:

Lemma 1. *For any $v \in \mathbb{R}^n$,*

$$\mathrm{E}[(X \cdot v)^2] = (\mu \cdot v)^2 + \sigma^2 ||v||^2.$$

Proof.

$$\begin{aligned}\mathrm{E}[(X \cdot v)^2] = \mathrm{E}[(\sum_{i=1}^n X_i v_i)^2] &= \mathrm{E}[\sum_{i,j=1}^n X_i X_j v_i v_j] \\ &= \sum_{i,j=1}^n \mathrm{E}[X_i X_j] v_i v_j.\end{aligned}$$

Using the assumption that $\mathrm{E}[X_i X_j] = \mathrm{E}[X_i]\mathrm{E}[X_j]$,

$$\sum_{i,j=1}^n \mathrm{E}[X_i]\mathrm{E}[X_j] v_i v_j - \sum_{i=1}^n \mathrm{E}[X_i]^2 v_i^2 + \sum_{i=1}^n \mathrm{E}[X_i^2] v_i^2$$

$$= (\mathrm{E}[X] \cdot v)^2 + \sum_{i=1}^n v_i^2 \sigma_i^2 = (\mu \cdot v)^2 + \sigma^2 ||v||^2.$$

□

Corollary 1. *For all $v \in \mathbb{R}^n$ such that $||v|| = ||\mu||$,*

$$\mathrm{E}[(X \cdot \mu)^2] \geq \mathrm{E}[(X \cdot v)^2].$$

The corollary says that the best rank 1 subspace for a distribution is the one that passes through its mean.

Proof. By the above lemma,

$$\begin{aligned}\mathrm{E}[(X \cdot \mu)^2] &= (\mu \cdot \mu)^2 + \sigma^2 ||\mu||^2. \\ \mathrm{E}[(X \cdot v)^2] &= (\mu \cdot v)^2 + \sigma^2 ||v||^2.\end{aligned}$$

So $\mathrm{E}[(X \cdot \mu)^2] - \mathrm{E}[(X \cdot v)^2] = (\mu \cdot \mu)^2 - (\mu \cdot v)^2 \geq 0$, and we have the desired result. □

Next, we consider the projection of X to a higher dimensional subspace. We write $||\mathrm{proj}_V X||^2$ to denote the squared length of the projection of X onto a subspace V. For an orthonormal basis $\{v_1 \dots v_r\}$ for V, this is just $\sum_{i=1}^r (X \cdot v_i)^2$.

Lemma 2. *Let $V \subseteq \mathbb{R}^n$ be a subspace of dimension r with orthonormal basis $\{v_1 \dots v_r\}$. Then*

$$\mathrm{E}[||\mathrm{proj}_V X||^2] = ||\mathrm{proj}_V \mathrm{E}[X]||^2 + r\sigma^2.$$

Proof.

$$\mathrm{E}[||\mathrm{proj}_V X||^2] = \mathrm{E}[\sum_{i=1}^{r} ||(X \cdot v_i)v_i||^2] = \mathrm{E}[\sum_{i=1}^{r} (X \cdot v_i)^2].$$

The equalities follow from the fact that $\{v_1 \dots v_r\}$ is an orthonormal basis. By linearity of expectation and Lemma 1, the above is

$$\sum_{i=1}^{r} \mathrm{E}[(X \cdot v_i)^2] = \sum_{i=1}^{r} (\mu \cdot v_i)^2 + \sigma^2 = ||\mathrm{proj}_V \mathrm{E}[X]||^2 + r\sigma^2.$$

□

Now consider a mixture of k distribtuions $F_1 \dots F_k$, with mean vectors μ_i and variances σ_i^2.

Let $A \in \mathbb{R}^{m \times n}$ be generated randomly from a mixture of distributions $F_1 \dots F_k$ with mixing weights $w_1 \dots w_k$. For a matrix A, and any subspace V, let $||\mathrm{proj}_V A||^2 = \sum_{i=1}^{m} ||\mathrm{proj}_V A_i||^2$.

Theorem 2. *Let $V \subseteq \mathbb{R}^n$ be a subspace of dimension r with an orthonormal basis $\{v_1 \dots v_r\}$. Then*

$$\mathrm{E}[||\mathrm{proj}_V A||^2] = ||\mathrm{proj}_V \mathrm{E}[A]||^2 + m \sum_{i=1}^{k} w_i \cdot r\sigma_i^2.$$

Proof.

$$\begin{aligned} \mathrm{E}[||\mathrm{proj}_V A||^2] &= \sum_{i=1}^{m} \mathrm{E}[||\mathrm{proj}_V A_i||^2] \\ &= \sum_{l=1}^{k} \sum_{i \in F_l} \mathrm{E}[||\mathrm{proj}_V A_i||^2]. \end{aligned}$$

The second equality follows from linearity of expectation. In expectation, there are $w_i m$ samples from each distribution. By applying Lemma 2, we have

$$\begin{aligned} \sum_{l=1}^{k} \sum_{i \in F_l} ||\mathrm{proj}_V \mathrm{E}[A_i]||^2 + r\sigma_i^2 &= m \sum_{l=1}^{k} w_i \left(||\mathrm{proj}_V \mu_i||^2 + r\sigma_i^2 \right) \\ &= ||\mathrm{proj}_V \mathrm{E}[A]||^2 + m \sum_{i=1}^{k} w_i \cdot r\sigma_i^2. \end{aligned}$$

Here and throughout this paper, by $\mathrm{E}[A]$ we mean the matrix with $w_i m$ rows that are μ_i, and not the matrix each of whose rows are $\sum_i w_i \mu_i$. □

From this it follows that the expected best subspace for a mixture of k distributions is simply the subspace spanned by the mean vectors of the distributions.

Corollary 2 (Expected Best Subspace). *Let $V \subseteq \mathbb{R}^n$ be a subspace of dimension k, and let $U = \mathrm{span}\{\mu_1 \dots \mu_k\}$. Then,*

$$\mathrm{E}[||\mathrm{proj}_U A||^2] \geq \mathrm{E}[||\mathrm{proj}_V A||^2].$$

Proof. By Theorem 2 we have,

$$\begin{aligned} \mathrm{E}[||\mathrm{proj}_U A||^2] - \mathrm{E}[||\mathrm{proj}_V A||^2] &= ||\mathrm{proj}_U \mathrm{E}[A]||^2 - ||\mathrm{proj}_V \mathrm{E}[A]||^2 \\ &= ||\mathrm{E}[A]||^2 - ||\mathrm{proj}_V \mathrm{E}[A]||^2 \geq 0. \end{aligned}$$

□

4 The Likely Best Subspace

In this section, we show that for a sufficiently large sample from a mixture of Gaussians, with high probability the subspace found by SVD is very close to the one spanned by the mean vectors.

We begin with a concentration lemma.

Lemma 3 (Concentration). *Let V be a subspace of $\mathbb{R}^n$ of dimension r, with an orthonormal basis $\{v_1 \dots v_r\}$. Let $A \in \mathbb{R}^{M\times n}$ be generated randomly from Gaussians $F_1, \dots, F_k$ with mixing weights $w_1, \dots, w_k$. Assume that A contains at least m rows from each Gaussian. Then for any $1 > \epsilon > 0$:*
1. $\Pr\left(||\text{proj}_V A||^2 > (1+\epsilon)\text{E}[||\text{proj}_V A||^2]\right) < ke^{-\frac{\epsilon^2 mr}{8}}$.
2. $\Pr\left(||\text{proj}_V A||^2 < (1-\epsilon)\text{E}[||\text{proj}_V A||^2]\right) < ke^{-\frac{\epsilon^2 mr}{8}}$.

Proof. Let $\mathcal{E}$ be the first event in consideration. We further condition $\mathcal{E}$ on the event that exactly m_i rows are generated by the ith Gaussian.

Note that $||\text{proj}_V A||^2 = \sum_{l=1}^k \sum_{i\in F_l} ||\text{proj}_V A_i||^2$. Therefore, we have the following bound on the probability of the conditioned event:

$$\Pr\left(\mathcal{E}/(m_1 \dots m_k)\right) \le k \max_l \Pr\left(\sum_{i\in F_l} ||\text{proj}_V A_i||^2 > (1+\epsilon)\text{E}[\sum_{i\in F_l} ||\text{proj}_V A_i||^2]\right).$$

Let $\mathcal{B}$ be the event that:

$$||\text{proj}_V B||^2 > (1+\epsilon)\text{E}[||\text{proj}_V B||^2]$$

where $B \in \mathbb{R}^{m_l\times n}$ is a matrix containing the rows of A that are generated by F_l, an arbitrary Gaussian. We bound the probability of $\mathcal{E}/(m_1 \dots m_k)$ by bounding the probability of $\mathcal{B}$.

Let $Y_{ij} = (B_i \cdot v_j)$. Note that $||\text{proj}_V B||^2 = \sum_{i=1}^{m_l}\sum_{j=1}^r Y_{ij}^2$. We are interested in the event $\mathcal{B} \equiv \sum_{i=1}^{m_l}\sum_{j=1}^r Y_{ij}^2 > (1+\epsilon)\text{E}[||\text{proj}_V B||^2]$. Note that Y_{ij} is a Gaussian random variable with mean $(\mu_l \cdot v_j)$ and variance σ_l^2. We can write $Y_{ij} = \sigma_l X_{ij}$, where X_{ij} is a Gaussian random variable with mean $\frac{(\mu_l\cdot v_j)}{\sigma_l}$ and variance 1. Rewriting in terms of X_{ij}, we are interested in the event

$$\mathcal{B} \equiv \sum_{i=1}^{m_l}\sum_{j=1}^{r} (\sigma_l X_{ij})^2 > (1+\epsilon)\text{E}[||\text{proj}_V B||^2].$$

Since $\text{E}[||\text{proj}_V B||^2] = \sum_{j=1}^r m_l((\mu_l\cdot v_j)^2 + \sigma_l^2)$, we have:

$$\mathcal{B} \equiv \sum_{i=1}^{m_l}\sum_{j=1}^{r} X_{ij}^2 > \frac{(1+\epsilon)\sum_{j=1}^r m_l((\mu_l\cdot v_j) + \sigma_l^2)}{\sigma_l^2}.$$

By Markov's inequality,

$$\Pr(\mathcal{B}) \le \frac{\text{E}[\exp\left(t\sum_{i=1}^{m_l}\sum_{j=1}^r X_{ij}^2\right)]}{\exp\left(\frac{t(1+\epsilon)\sum_{j=1}^r m_l((\mu_l\cdot v_j)+\sigma_l^2)}{\sigma_l^2}\right)}.$$

Note that $Z = \sum_{i=1}^{m_l}\sum_{j=1}^r X_{ij}^2$ is a chi-squared random variable with noncentrality parameter $\sum_{j=1}^r \frac{m_l(\mu_l\cdot v_j)^2}{\sigma^2}$ and $m_l r$ degrees of freedom. The moment generating function for Z is (see e.g. [7]):

$$\text{E}[e^{t\sum_{i=1}^{m_l}\sum_{j=1}^r X_{ij}^2}] = (1-2t)^{-m_l r/2}\exp\left(\sum_{j=1}^r \frac{m_l(\mu_l\cdot v_j)^2}{\sigma_l^2}\left[\frac{t}{1-2t}\right]\right).$$

So we obtain the bound on $\Pr(\mathcal{B})$:

$$\Pr(\mathcal{B}) \le (1-2t)^{-m_l r/2} \exp\left(t\left(-(1+\epsilon)m_l r - (\epsilon - 2t(\epsilon+1))\frac{\sum_{j=1}^r m_l(\mu_l \cdot v_j)^2}{(1-2t)\sigma_l^2}\right)\right).$$

Using the fact that $\frac{1}{1-2t} \le e^{2t+4t^2}$, and setting $t = \frac{\epsilon}{4}$, we have $\epsilon - 2t(\epsilon+1) > 0$, and so

$$\Pr(\mathcal{B}) \le \frac{e^{(2t+4t^2)\frac{m_l r}{2}}}{e^{t m_l r(1+\epsilon)}} \le e^{-\frac{\epsilon^2 m_l r}{8}}.$$

Therefore, we obtain the following bound on $\mathcal{E}/(m_1 \dots m_k)$

$$\Pr(\mathcal{E}/(m_1 \dots m_k)) \le k \max_l e^{-\frac{\epsilon^2 m_l r}{8}} \le k e^{-\frac{\epsilon^2 m r}{8}}.$$

Since this is true regardless of the choice of $m_1, \dots, m_k$, the probability of $\mathcal{E}$ itself is at most the above. □

We can extend this concentration lemma to show that the probability that *any* subspace of dimension r has the property that the projection of a sample matrix onto the subspace lies far from its expectation is small.

Lemma 4. *Suppose $A \in \mathbb{R}^{M \times n}$ has at least m rows from each of the k Gaussians $F_1 \dots F_k$. Then, for any $1 > \epsilon > 0$, $\frac{1}{\sqrt{n}} > \alpha > 0$, and any r such that $1 \le r \le n$, the probability that there exists a subspace W of dimension r that satisfies*

$$||\mathrm{proj}_W A||^2 \le (1-\epsilon)\mathrm{E}[||\mathrm{proj}_W A||^2] - (6r\sqrt{n}\alpha)\mathrm{E}[||A||^2]$$

is at most

$$\left(\frac{2}{\alpha}\right)^{rn} k e^{\frac{-\epsilon^2 m r}{8}}.$$

Proof. Let $\mathcal{W}$ be the set of all r-dimensional subspaces. Let S be a finite set of r dimensional subspaces, with $|S| = N$, such that for any $W \in \mathcal{W}$ with orthonormal basis $\{w_1 \dots w_r\}$, there exists a $W^* \in S$ with orthonormal basis $\{w_1^* \dots w_r^*\}$, such that for all i and j, $|w_{ij} - w_{ij}^*| < \alpha$, component-wise. Then for any $W \in \mathcal{W}$, and any $a \in \mathbb{R}^n$,

$$\begin{aligned}
||\mathrm{proj}_{W^*} a||^2 &= \sum_{i=1}^r (a \cdot w_i^*)^2 \\
&= \sum_{i=1}^r (a \cdot (w_i^* - w_i + w_i))^2 \\
&\le \sum_{i=1}^r (a \cdot w_i)^2 + (a \cdot (w_i^* - w_i))^2 + 2(a \cdot (w_i^* - w_i))(a \cdot w_i) \\
&= ||\mathrm{proj}_W a||^2 + \sum_{i=1}^r \left(\sum_{j=1}^n |a_j||w_{ij}^* - w_{ij}|\right)^2 + \\
&\quad 2\sum_{i=1}^r \left(\sum_{j=1}^n |a_j||w_{ij}^* - w_{ij}|\right)\left(\sum_{j=1}^n |a_j||w_{ij}|\right) \\
&\le ||\mathrm{proj}_W a||^2 + rn\alpha^2||a||^2 + 2r\sqrt{n}\alpha||a||^2 \\
&\le ||\mathrm{proj}_W a||^2 + 3rn\alpha||a||^2
\end{aligned}$$

The last line follows from $\alpha < \frac{1}{\sqrt{n}}$. The above also gives us a bound on the projection of matrices:

$$||\text{proj}_W A||^2 \geq ||\text{proj}_{W^*} A||^2 - 3r\sqrt{n}\alpha||A||^2.$$

The same sequence of inequalities starting with $\text{E}[||\text{proj}_W A||^2]$ can be used to show:

$$\text{E}[||\text{proj}_W A||^2] \leq \text{E}[||\text{proj}_{W^*} A||^2] + 3r\sqrt{n}\alpha\text{E}[||A||^2].$$

Combining these two we get

$$||\text{proj}_W A||^2 - (1-\epsilon)\text{E}[||\text{proj}_W A||^2] \geq ||\text{proj}_{W^*} A||^2 -$$

$$(1-\epsilon)\text{E}[||\text{proj}_{W^*} A||^2] - (3r\sqrt{n}\alpha)(||A||^2 + (1-\epsilon)\text{E}[||A||^2]).$$

Using Lemma 3 and the union bound, we get that

$$\Pr\left(\exists W, ||\text{proj}_W A||^2 \leq (1-\epsilon)\text{E}[||\text{proj}_W A||^2] - (6r\sqrt{n}\alpha)\text{E}[||A||^2]\right)$$

is at most $(N+1)ke^{\frac{-\delta^2 mr}{8}}$. Here we have used that the probability that $||A||^2 \geq (1+\epsilon)\text{E}[||A||^2]$ is much lower than the probability above. A simple upper bound on $N+1$ is $(\frac{2}{\alpha})^{rn}$, the number of grid points in the cube $[-1,1]^{nr}$ with grid size α. The lemma follows. □

We now proceed to prove the main theorem. The intuition is as follows: from Lemma 3 above, we know that the norm of a sample matrix projected to a particular subspace stays close to its expectation. From Theorem 2, we know that in expectation, the subspace that maximizes the norm of the projection of the sample matrix is exactly the span of the mean vectors. If a subspace is "far" from the span of the mean vectors, then the expected norm of the projection of the sample matrix is much smaller than the expected norm of the projection onto the mean vectors. By considering a net of these subspaces that are "far" away, in addition to the concentration lemma, we show that it is unlikely that the subspace spanned by the top r singular vectors is "far" away.

Theorem 3. *Let the rows of $A \in \mathbb{R}^{m\times n}$ be picked according to a mixture of Gaussians $F_1, \ldots, F_k$ with mixing weights $w_1, \ldots, w_k$, means $\mu_1 \ldots \mu_k$ and variances $\sigma_1^2 \ldots \sigma_k^2$. Let $r = \max\{k, 96\ln\left(\frac{4m}{\delta}\right)\}$, and let $V \subseteq \mathbb{R}^n$ be the r-dimensional subspace spanned by the top r right singular vectors, and let U be a r dimensional subspace that contains the mean vectors $\mu_1 \ldots \mu_k$. Then for any $\frac{1}{2} > \epsilon > 0$, with*

$$m > \frac{5000}{\epsilon^2 w_{\min}}\left(n\left(\ln\left(\frac{n}{\epsilon}\right) + \ln\left(\max_i \frac{||\mu_i||^2}{\sigma_i^2}\right)\right) + \frac{1}{n-r}\ln\left(\frac{k}{\delta}\right)\right)$$

we have with probability at least $1-\delta$,

$$||\text{proj}_U \text{E}[A]||^2 - ||\text{proj}_V \text{E}[A]||^2 \leq \epsilon m(n-r)\sum_{i=1}^k w_i\sigma_i^2.$$

Proof. We lower bound the probability of the desired event by upper bounding the probability of the opposite event,

$$\mathcal{E} \equiv ||\text{proj}_U \text{E}[A]||^2 - ||\text{proj}_V \text{E}[A]||^2 > \epsilon m(n-r)\sum_{i=1}^k w_i\sigma_i^2.$$

In particular, we consider a weaker event in terms of the orthogonal subspaces. This will allow us to bound the probability of $\mathcal{E}$ in terms of the concentration lemmas we have proven where the deviation depends on the *variances* of the distributions, instead of the means. First, note that:

$$||\text{proj}_U \text{E}[A]||^2 - ||\text{proj}_V \text{E}[A]||^2 = ||\text{proj}_{\overline{V}} \text{E}[A]||^2 - ||\text{proj}_{\overline{U}} \text{E}[A]||^2$$

where $\overline{U}$ is the orthogonal subspace. This holds because, for any subspace V we have:

$$||A||^2 = ||\text{proj}_V A||^2 + ||\text{proj}_{\overline{V}} A||^2.$$

Therefore, if V maximizes $||\text{proj}_V A||^2$ for all r dimensional subspaces, then $\overline{V}$ minimizes $||\text{proj}_{\overline{V}} A||^2$ for all $n-r$ dimensional subspaces. We weaken $\mathcal{E}$ by considering the probability that *some* $n-q$ dimensional subspace $\overline{W}$ has the property that $||\text{proj}_{\overline{W}} A||^2 \leq ||\text{proj}_{\overline{U}} A||^2$ (note that $\overline{V}$ would achieve the minimum such value) *and*

$$||\text{proj}_{\overline{W}} \mathrm{E}[A]||^2 - ||\text{proj}_{\overline{U}} \mathrm{E}[A]||^2 > \epsilon m(n-r) \sum_{i=1}^{k} w_i \sigma_i^2. \tag{4}$$

We can relate (4) to the concentration lemmas. By Theorem 2,

$$||\text{proj}_{\overline{W}} \mathrm{E}[A]||^2 - ||\text{proj}_{\overline{U}} \mathrm{E}[A]||^2 = \mathrm{E}[||\text{proj}_{\overline{W}} A||^2] - \mathrm{E}[||\text{proj}_{\overline{U}} A||^2]$$

Now, $||\text{proj}_{\overline{U}} \mathrm{E}[A]||^2 = 0$, since $\overline{U}$ and U are orthogonal subspaces. Also,

$$\begin{aligned} \mathrm{E}[||\text{proj}_{\overline{U}} A||^2] &= ||\text{proj}_{\overline{U}} \mathrm{E}[A]||^2 + m(n-r) \sum_{i=1}^{k} w_i \sigma_i^2 \\ &= m(n-r) \sum_{i=1}^{k} w_i \sigma_i^2. \end{aligned}$$

So, $\Pr(\mathcal{E})$ is at most the probability that there exists $\overline{W}$ such that $||\text{proj}_{\overline{W}} A||^2 \leq ||\text{proj}_{\overline{U}} A||^2$ *and* the following event $\mathcal{E}_1$ happens

$$\mathrm{E}[||\text{proj}_{\overline{W}} A||^2] > (1+\epsilon)\mathrm{E}[||\text{proj}_{\overline{U}} A||^2]$$

Note that we can rewrite $||\text{proj}_{\overline{W}} A||^2 \leq ||\text{proj}_{\overline{U}} A||^2$ as:

$$\mathrm{E}[||\text{proj}_{\overline{W}} A||^2] - ||\text{proj}_{\overline{W}} A||^2 + ||\text{proj}_{\overline{U}} A||^2 - \mathrm{E}[||\text{proj}_{\overline{U}} A||^2] \geq \mathrm{E}[||\text{proj}_{\overline{W}} A||^2] - \mathrm{E}[||\text{proj}_{\overline{U}} A||^2]$$

The required probability is thus at most the sum of the probabilities of the following two events (one of the following must occur for $\mathcal{E}$ to occur):

$$\begin{aligned} \mathcal{A} &\equiv \exists \overline{W} : \mathcal{E}_1 \text{ and} \\ &\quad ||\text{proj}_{\overline{U}} A||^2 - \mathrm{E}[||\text{proj}_{\overline{U}} A||^2] \geq \frac{1}{2}\left(\mathrm{E}[||\text{proj}_{\overline{W}} A||^2] - \mathrm{E}[||\text{proj}_{\overline{U}} A||^2]\right) \\ \mathcal{B} &\equiv \exists \overline{W} : \mathcal{E}_1 \text{ and} \\ &\quad \mathrm{E}[||\text{proj}_{\overline{W}} A||^2] - ||\text{proj}_{\overline{W}} A||^2 \geq \frac{1}{2}\left(\mathrm{E}[||\text{proj}_{\overline{W}} A||^2] - \mathrm{E}[||\text{proj}_{\overline{U}} A||^2]\right) \end{aligned}$$

The probability of $\mathcal{A}$ is at most the probability that there exists $\overline{W}$ such that $||\text{proj}_{\overline{U}} A||^2 - \mathrm{E}[||\text{proj}_{\overline{U}} A||^2] \geq \frac{1}{2}\left((1+\epsilon)\mathrm{E}[||\text{proj}_{\overline{U}} A||^2] - \mathrm{E}[||\text{proj}_{\overline{U}} A||^2]\right)$, which is at most

$$\Pr(\mathcal{A}) \leq \Pr\left(||\text{proj}_{\overline{U}} A||^2 \geq (1+\frac{\epsilon}{2})\mathrm{E}[||\text{proj}_{\overline{U}} A||^2]\right) \leq e^{\frac{-\epsilon^2 m \cdot w_{\min}(n-r)}{32}}$$

The last inequality follows from Lemma 3, and the fact that each distribution generates at least $\frac{w_{\min} m}{2}$ rows (the probability this does not happen is much smaller than δ). Now the probability of $\mathcal{B}$ is at most

$$\begin{aligned} \Pr(\mathcal{B}) &\leq \Pr\left(\exists \overline{W} : \mathcal{E}_1 \text{ and } ||\text{proj}_{\overline{W}} A||^2 \leq \mathrm{E}[||\text{proj}_{\overline{W}} A||^2] - \frac{1}{2}\left(\left(1 - \frac{1}{1+\epsilon}\right)\mathrm{E}[||\text{proj}_{\overline{W}} A||^2]\right)\right) \\ &\leq \Pr\left(\exists \overline{W} : \mathcal{E}_1 \text{ and } ||\text{proj}_{\overline{W}} A||^2 \leq \left(1 - \frac{\epsilon}{4}\right)\mathrm{E}[||\text{proj}_{\overline{W}} A||^2]\right) \\ &\leq \Pr\left(\exists \overline{W} : ||\text{proj}_{\overline{W}} A||^2 \leq (1 - \frac{\epsilon}{8})\mathrm{E}[||\text{proj}_{\overline{W}} A||^2] - \frac{\epsilon}{8}\mathrm{E}[||\text{proj}_{\overline{U}} A||^2]\right) \end{aligned}$$

Here we used the fact that $\epsilon < \frac{1}{2}$. By applying Lemma 4 with

$$\alpha = \frac{\epsilon \mathrm{E}[||\mathrm{proj}_{\overline{U}} A||^2]}{48\sqrt{n}(n-r)\mathrm{E}[||A||^2]}$$

and

$$\frac{512}{\epsilon^2}\left(n \ln\left(\frac{2}{\alpha}\right) + \frac{1}{n-r}\ln\left(\frac{k}{\delta}\right)\right)$$

samples from each Gaussian, we have that the probability of $\mathcal{B}$, and therefore $\mathcal{E}$ is at most δ, which is the desired result. □

As a result, we have that with enough samples, the distance between the original mean vectors and the projected mean vectors is not large.

Corollary 3 (Likely Best Subspace). *Let $\mu_1, \dots, \mu_k$ be the means of the k Gaussians in the mixture and w_{min} the smallest mixing weight. Let $\mu'_1, \dots, \mu'_k$ be their projections onto the subspace spanned by the top r right singular vectors of the sample matrix A. With a sample of size of $m = O^*(\frac{n}{\epsilon^2 w_{min}})$ we have with high probability*

$$\sum_{i=1}^{k} w_i(||\mu_i||^2 - ||\mu'_i||^2) \leq \epsilon(n-r)\sum_{i=1}^{k} w_i\sigma_i^2.$$

Proof. Let V be the optimal rank r subspace. Let m be as large as required in Theorem 3. By the theorem, we have that with probability at least $1-\delta$,

$$\mathrm{E}[||\mathrm{proj}_U A||^2] - \mathrm{E}[||\mathrm{proj}_V A||^2] \leq \epsilon m(n-r)\sum_{i=1}^{k} w_i\sigma_i^2$$

which is equivalent to

$$m\sum_{i}^{k} w_i(||\mu_i||^2 - ||\mu'_i||^2) \leq \epsilon m(n-r)\sum_{i=1}^{k} w_i\sigma_i^2.$$

□

4.1 Random projection vs. spectral projection: examples

As mentioned in the introduction, random projection does not preserve the distance between the means. The following figures illustrate the difference between random projection and the SVD-based projection for a mixture of Gaussians.

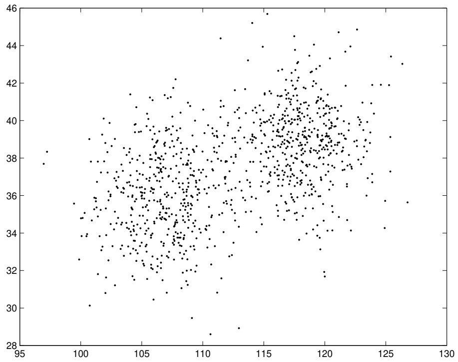

Figure 1: RP1

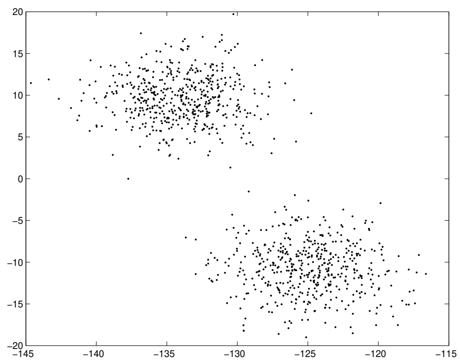

Figure 2: SVD1

Figure 1 and figure 3 are 2-dimensional random projections of samples from two different 49-dimensional mixtures (one with $k = 2$, the other with $k = 4$). Figure 2 and figure 4 are the projections to the best rank 2 subspaces of the same data sets.

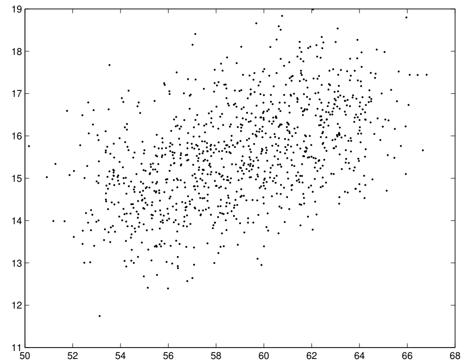

Figure 3: RP2

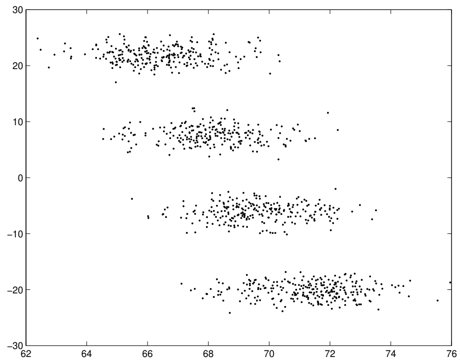

Figure 4: SVD2

5 After Projecting to the Best Subspace

To see the main idea, first project the sample matrix to the top r right singular vectors. Let $D = \max_{i,j} \frac{\sigma_i^2}{w_j \sigma_j^2}$. If D is small, we can apply Corollary 3 with $\epsilon < \frac{\hat{\epsilon}^2}{D(n-r)}$ so that for every i,

$$||\mu_i - \mu_i'||^2 = ||\mu_i||^2 - ||\mu_i'||^2 < \hat{\epsilon}^2 \sigma_i^2. \tag{5}$$

Therefore, for any pair i, j after projection,

$$||\mu_i' - \mu_j'|| \quad > \quad ||\mu_i - \mu_j|| - \hat{\epsilon}(\sigma_i + \sigma_j).$$

Now the projection of a Gaussian distribution onto any subspace remains a Gaussian distribution and so the radius is now $\sigma_i \sqrt{r}$. With such a large radius, and a separation as in (2), we can use distance concentration to correctly classify a random subsample of size $poly(n)/w_{\min}$. The details will be clear in what follows, but the idea is that points from any distribution with approximately the smallest radius will be separated from the rest of the sample. We can identify such a distribution and repeat the procedure.

Now let us consider the general case (when D is large or unknown) for a mixtures of spherical Gaussians. We first start with a distance concentration lemma.

Lemma 5. *Let $X \in F_s$, and $Y \in F_t$, where F_s, F_t are r dimensional Gaussians with means μ_s', μ_t', and variances σ_s^2, σ_t^2. Then for $\alpha > 0$, the probability that*

$$\left| ||X - Y||^2 - \mathrm{E}[||X - Y||^2] \right| \geq \alpha \left((\sigma_s^2 + \sigma_t^2)\sqrt{r} + 2||\mu_s' - \mu_t'||\sqrt{\sigma_s^2 + \sigma_t^2} \right)$$

is at most $4e^{-\alpha^2/8}$.

Proof. We consider the probability that

$$||X - Y||^2 - \mathrm{E}[||X - Y||^2] \geq \alpha((\sigma_s^2 + \sigma_t^2)\sqrt{r} + 2||\mu_s' - \mu_t'||\sqrt{\sigma_s^2 + \sigma_t^2}).$$

The other part is analogous. We write $||X - Y||^2 = \sum_{i=1}^r (\sqrt{\sigma_s^2 + \sigma_t^2} Z_i + (\mu_{si} - \mu_{ti}))^2$, where the Z_i are $N(0,1)$ random variables. Therefore, we can rewrite the above event as:

$$\sum_{i=1}^r (\sqrt{\sigma_s^2 + \sigma_t^2} Z_i + (\mu_{si}' - \mu_{ti}'))^2 \geq (\sigma_s^2 + \sigma_t^2) r + ||\mu_s' - \mu_t'||^2 + \alpha \left((\sigma_s^2 + \sigma_t^2)\sqrt{r} + 2||\mu_s' - \mu_t'||\sqrt{\sigma_s^2 + \sigma_t^2} \right).$$

The probability of this occurring is at most the sum of the probabilities of the following events

$$\mathcal{A} \equiv \sum_{i=1}^{r} (\sigma_s^2 + \sigma_t^2) Z_i^2 \geq (\sigma_s^2 + \sigma_t^2)(r + \alpha\sqrt{r})$$

$$\mathcal{B} \equiv \sum_{i=1}^{r} 2(\mu'_{si} - \mu'_{ti})\sqrt{\sigma_s^2 + \sigma_t^2} Z_i \geq \alpha \left(2||\mu'_s - \mu'_t||\sqrt{\sigma_s^2 + \sigma_t^2} \right).$$

Simplifying, we get:

$$\Pr(\mathcal{A}) \leq \Pr\left(\sum_{i=1}^{r} Z_i^2 \geq r + \alpha\sqrt{r} \right) \leq e^{-\alpha^2/8}$$

$$\Pr(\mathcal{B}) \leq \Pr\left(\sum_{i=1}^{r} (\mu_{si} - \mu_{ti}) Z_i \geq \alpha ||\mu'_s - \mu'_t|| \right) \leq e^{-\alpha^2/8}.$$

The above inequalities hold by applying Markov's inequality and moment generating functions as in the proof of Lemma 3. □

With this distance concentration lemma in hand, we can learn the mixture of Gaussians by the following implementation of step (3) of the algorithm. In the description of the algorithm below, S is the set of m projected points.

Algorithm.

(a) Let $R = \max_{x \in S} \min_{y \in S} ||x - y||$.

(b) Discard all points from S whose closest point lies at squared distance at most $3\hat{\epsilon}R^2$ to form a new set S'.

(c) Let x, w be the two closest points in the remaining set, and let H be the set of all points at squared distance at most $l = ||x - w||^2(1 + 8\sqrt{\frac{6 \ln \frac{4m}{\delta}}{r}})$ from x.

(d) Report H as a Gaussian, and remove the points in H from S'. Repeat step (c) till there are no points left.

(e) Output all Gaussians learned in step (d) with variance greater than $3\epsilon R^2/r$.

(f) Remove the points from step (e) from the original sample S, and repeat the entire algorithm (including the SVD calculation and projection) on the rest.

We show that each Gaussian output in step (e) of the algorithm above is correct with high probability, i.e. it is exactly the subsample of points from a single Gaussian. Further, at least one Gaussian is output during every iteration. The general idea is that on projection, the center of a Gaussian with a large variance does not move much relative to its variance. By first removing points from Gaussians with small variance from the sample set, we can classify points from the largest Gaussians by distance concentration.

Theorem 4. *With a sample of size*

$$m = \Omega\left(\frac{n^3}{w_{\min}^2} \left(\ln \frac{n}{w_{\min}} + \ln \left(\max_{i=1}^{k} \frac{|\mu_i|^2}{\sigma_i^2} \right) + \ln \frac{n}{\delta} \right) \right)$$

and initial separation

$$||\mu_i - \mu_j|| \geq 14 \max\{\sigma_i, \sigma_j\} \left(r \ln \left(\frac{4m}{\delta} \right) \right)^{1/4}$$

the algorithm correctly classifies all Gaussians with probability at least $1 - \delta$.

Proof. First, let us apply Corollary 3 with $\epsilon = \frac{w_{\min}\hat{\epsilon}}{n-r}$. Let σ^2 be the largest variance. It follows that,

$$\sum_{i=1}^{k} w_i(||\mu_i||^2 - ||\mu_i'||^2) \leq w_{\min}\hat{\epsilon} \sum_{i=1}^{k} w_i \sigma_i^2 \leq w_{\min}\hat{\epsilon}\sigma^2.$$

So in particular, for all i,

$$||\mu_i - \mu_i'||^2 = ||\mu_i||^2 - ||\mu_i'||^2 \leq \hat{\epsilon}\sigma^2.$$

This implies that for any Gaussian F_i with variance larger than $\hat{\epsilon}\sigma^2$, we have that for all j:

$$||\mu_i' - \mu_j'|| \geq 14\sigma_i \left(r \ln \left(\frac{4m}{\delta} \right) \right)^{1/4} - 2\sqrt{\hat{\epsilon}}\sigma_i \geq 12\sigma_i \left(r \ln \left(\frac{4m}{\delta} \right) \right)^{1/4}$$

provided that $\hat{\epsilon} \leq 1$. By applying Lemma 5 with this separation between mean vectors and $\alpha = \sqrt{24 \ln \left(\frac{4m}{\delta} \right)}$, we obtain the following with probability at least $1 - \frac{\delta}{m}$:

- For any Gaussian F_i and any two points x, y drawn from it,

$$2\sigma_i^2 r - 4\sigma_i^2 \sqrt{6r \ln \left(\frac{4m}{\delta} \right)} \leq ||x - y||^2 \leq 2\sigma_i^2 r + 4\sigma_i^2 \sqrt{6r \ln \left(\frac{4m}{\delta} \right)}. \tag{6}$$

  It will also be helpful to have the following bounds on $||x - y||^2$ in terms of only σ_i^2:

$$\sigma_i^2 r \leq ||x - y||^2 \leq 3\sigma_i^2 r \tag{7}$$

  This follows by upper bounding the deviation term $4\sigma_i^2 \sqrt{6r \ln \left(\frac{4m}{\delta} \right)}$ by $\sigma_i^2 r$. This holds by applying $r \geq 96 \ln \left(\frac{4m}{\delta} \right)$.

- For any Gaussians $F_i \neq F_j$, with $\sigma_i^2 \geq \sigma_j^2$ and $\sigma_i^2 > \hat{\epsilon}\sigma^2$ and any two points x from F_i and y from F_j,

$$||x - y||^2 \geq (\sigma_i^2 + \sigma_j^2) r + 38\sigma_i^2 \sqrt{6r \ln \left(\frac{4m}{\delta} \right)}. \tag{8}$$

  This follows from lemma 5, which states that:

$$||x - y||^2 \geq (\sigma_i^2 + \sigma_j^2) r + ||\mu_i' - \mu_j'||^2 - \alpha \left((\sigma_i^2 + \sigma_j^2)\sqrt{r} + 2||\mu_i' - \mu_j'|| \sqrt{\sigma_i^2 + \sigma_j^2} \right)$$

  With $\alpha = \sqrt{24 \ln \left(\frac{m}{\delta} \right)}$, $||\mu_i' - \mu_j'|| \geq 12\sigma_i \left(r \ln \left(\frac{4m}{\delta} \right) \right)^{1/4}$ and $r \geq 96 \ln \left(\frac{4m}{\delta} \right)$, we obtain the above bound. We can also obtain a lower bound only in terms of σ_i^2 as we did above for two points from the same Gaussian by the same upper bound on the deviation term:

$$||x - y||^2 \geq 12\sigma_i^2 r. \tag{9}$$

Using these bounds, we can classify the smallest Gaussian using (6), since inter-Gaussian distances are smaller than intra-Gaussian distances. However, this only holds for Gaussians with variance larger than $\hat{\epsilon}\sigma^2$. We first show that the first step in the algorithm removes all such small Gaussians, and then show that any Gaussians classified in step (e) are complete Gaussians. The next few observations prove the correctness of the algorithm, conditioned on Lemma 5 to obtain (6), (7), (8) and (9).

1. Let $y \in S'$ be from some Gaussian F_i. Then $\sigma_i^2 \geq \hat{\epsilon}\sigma^2$.
2. Let $x \in F_j, H$ be the point and set used in any iteration of step (c). Then $H = S' \cap F_j$, i.e. it is the points in S' from F_j.
3. For any Gaussian F_i with $\sigma_i^2 > 3\hat{\epsilon}R^2/r$, we have $F_i \subseteq S'$, i.e. any Gaussian with sufficiently large variance will be contained in S'.

We proceed to prove 1-3.

1. Let x be any point from F, the Gaussian with largest variance, and let w be any other point. By (7) and (9), $||x-w||^2 \geq \sigma^2 r$, so $R^2 \geq \sigma^2 r$.

Now suppose by way of contradiction that $\sigma_i^2 < \hat{\epsilon}\sigma^2$. We will show that if this is the case, y would have removed from S, contradicting its membership in S'. Let z be another point in S from F_i. Then by (7) we have that:

$$\begin{aligned} ||y-z||^2 &\leq 3\sigma_i^2 r \\ &< 3\hat{\epsilon}\sigma^2 r \leq 3\hat{\epsilon}R^2 \end{aligned}$$

Since y's closest point lies at a distance at most $3\hat{\epsilon}R^2$, this contradicts $y \in S'$.
2. First, we show that x, w in step (c) of the algorithm belong to the same Gaussian. If not, suppose without loss of generality that $w \in F_i$, and that $\sigma_i \leq \sigma_j$. But then by (6) and (8), there exists a point z from F_j such that:

$$||x-z||^2 \leq 2\sigma_i^2 r + 4\sigma_i^2\sqrt{6r\ln\left(\frac{4m}{\delta}\right)}.$$

However,

$$||x-w||^2 \geq 2\sigma_i^2 r + 38\sigma_i^2\sqrt{6r\ln\left(\frac{4m}{\delta}\right)}.$$

This contradicts the fact that x, w are the two closest points in S'.

With the bounds on $||x-w||^2$ for $x, w \in F_j$ from (6), we obtain the following bounds on l, our estimate for the furthest point from x that is still in F_j:

$$2\sigma_j^2 r + 4\sigma_j^2\sqrt{6r\ln\left(\frac{4m}{\delta}\right)} \leq l \leq 2\sigma_j^2 r + 28\sigma_j^2\sqrt{6r\ln\left(\frac{4m}{\delta}\right)}.$$

The lower bound on l ensures that every point in F_j and S' is included in H.

The upper bound on l ensures that any point $z \in F_i \neq F_j$ will not be included in H. If $\sigma_i \geq \sigma_j$, this follows from the fact that the upper bound on l is less than the lower bound on $||x-z||^2$ in (8). Now suppose $\sigma_i \leq \sigma_j$. Since x, w are the closest points in S', it must be the case that:

$$||x-w||^2 \leq 2\sigma_i^2 r + 4\sigma_i^2\sqrt{6r\ln\left(\frac{4m}{\delta}\right)}$$

since otherwise, two points from F_i would be the two closest points in S'. Since $||x-w||^2 \geq 2\sigma_j^2 r - 4\sigma_j^2\sqrt{6r\ln\left(\frac{4m}{\delta}\right)}$, we have that

$$\sigma_i^2 r \geq \sigma_j^2 r - 4\sigma_j^2\sqrt{6r\ln\left(\frac{4m}{\delta}\right)}.$$

Applying this to (8), we have that:

$$||x-z||^2 \geq 2\sigma_j^2 r + 34\sigma_j^2\sqrt{6r\ln\left(\frac{4m}{\delta}\right)}.$$

As this is larger than the upper bound on l, we have that $F_j \cap S' = H$.
3. Let $x \in F_i$, with $\sigma_i^2 > 3\hat{\epsilon}R^2/r$. We want to show that $x \in S'$, so we need to show that step (b) never removes x from S. This holds if:

$$\forall z, ||x - z||^2 > 3\hat{\epsilon}R^2$$

which is true by (7) and (9).

We have just shown that any Gaussian with variance at least $3\hat{\epsilon}R^2/r$ will be correctly classified. By setting $\hat{\epsilon} < \frac{1}{9}$, at least the largest Gaussian is classified in step (e), since $R^2 \leq 3\sigma^2 r$ by (7). So we remove at least one Gaussian from the sample during each iteration.

It remains to bound the success probability of the algorithm. For this we need

- Corollary 3 holds up to k times,
- Steps (a)-(e) are successful up to k times

The probability of the first event is at least $(1 - \frac{\delta}{4k})^k \geq 1 - \delta/2$. The probability of the latter event is just the probability that distance concentration holds, which is $(1 - \frac{\delta}{m}) \geq (1 - \frac{\delta}{4k})$. Therefore, we have the probability that the algorithm succeeds is: $(1 - \frac{\delta}{4k})^k(1 - \delta/2) \geq 1 - \delta$. □

Note that if we assume only the minimum possible separation,

$$||\mu_i - \mu_j|| > C \max\{\sigma_i, \sigma_j\},$$

then after projection to the top k right singular vectors the separation is

$$||\mu_i' - \mu_j'|| > (C - 2\hat{\epsilon}) \max\{\sigma_i, \sigma_j\}$$

which is the radius divided by $\sqrt{k}$ (note that we started with a separation of radius divided by $\sqrt{n}$). Here we could use an exponential in k algorithm using $O(\frac{k}{w_{\min}})$ samples from the Gaussian to obtain a maximum-likelihood estimation. First, project the samples to the k right singular vectors of the sample matrix. Then, consider each of the $O\left(\frac{k}{w_{\min}}\right)^k$ partitions of the points into k clusters. For each of these clusters, we can compute the mean and variance, as well as the mixing weight of the cluster. Since the points were generated from a spherical Gaussian, and we know the density function F for a spherical Gaussian with a given mean and variance, we can compute the likelihood of the partition. Let x be any point in the sample, and let $l(x)$ denote the cluster that contains it. Then the likelihood of the sample is:

$$\prod_{x \in S} F_{l(x)}(x).$$

By examining each of the partitions, we can determine the partition that has the maximum-likelihood, obtaining estimates of the means, variances, and mixing weights of the mixture.

6 Remarks

The algorithm and its guarantees can be extended to mixtures of weakly-isotropic distributions, provided they have two types of concentration bounds. For a guarantee as in Theorem 3, we require an analog of Lemma 3 to hold, and for a guarantee as in Theorem 4, we need a lemma similar to Lemma 5. A particular class of distributions that possess good concentration bounds is the class of logconcave distributions. A distribution is said to be *logconcave* if its density function f is logconcave, i.e. for any $x, y \in \mathbb{R}^n$, and any $0 \leq \alpha \leq 1$, $f(\alpha x + (1 - \alpha)y) \geq f(x)^\alpha f(y)^{1-\alpha}$. Examples of log-concave distributions include the special case of the uniform distribution on a weakly isotropic convex body, e.g. cubes, balls, etc. Although the weakly isotropic property might sound restrictive, it is worth noting that any single log-concave distribution can be made weakly isotropic by a linear transformation (see, e.g. [11]).

We remark that the SVD is tolerant to noise whose 2-norm is bounded [13, 12, 2]. Thus even after corruption, the SVD of the sample matrix will recover a subspace that is close to one spanning the mean vectors of the underlying distributions. In this low-dimensional space, one could exhaustively examine subsets of the data to learn the mixture (the ignored portion of the data corresponds to the noise).

We also note that the classical algorithm for SVD takes $O(mn^2)$ time for an $m \times n$ matrix. It is worth investigating whether the faster randomized methods of [8, 6] can be used in this setting.

The spectral approach does not seem to directly apply to distributions that are not weakly isotropic, e.g. non-spherical Gaussians. It is an open question as to whether it can be generalized/adapted.

References

[1] S. Arora and R. Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the 33rd ACM STOC*, 2001.

[2] Y. Azar, A. Fiat, A. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *Proceedings of the 33rd ACM STOC*, 2001.

[3] S. DasGupta. Learning mixtures of gaussians. In *Proceedings of the 40th IEEE FOCS*, 1999.

[4] S. DasGupta and L. Schulman. A two-round variant of em for gaussian mixtures. In *Uncertainty in Artificial Intelligence*, 2000.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. In *J. Royal Statistics Soc. Ser. B*, volume 39, pages 1–38, 1977.

[6] P. Drineas, R. Kannan, A. Frieze, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *Proceedings of the 10th SIAM SODA*, 1999.

[7] M. L. Eaton. *Multivariate Statistics.* Wiley, New York, 1983.

[8] A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. In *Proceedings of the 39th IEEE FOCS*, 1998.

[9] R. Kannan, L. Lovász, and M. Simonovits. Isoperimetric problems for convex bodies and a localization lemma. In *Discrete Computational Geometry*, volume 13, pages 541–559, 1995.

[10] B. Lindsay. Mixture models: theory, geometry and applications. In *American Statistical Association*, 1995.

[11] L. Lovász and S. Vempala. Logconcave functions: Geometry and efficient sampling algorithms. In *Proceedings of the 44th IEEE FOCS (to appear)*, 2003.

[12] C. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent semantic indexing: a probabilistic analysis. In *Journal of Computer and System Sciences*, volume 61, pages 217–235, 2000.

[13] G. Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. In *SIAM review*, volume 15(4), pages 727–64, 1973.

[14] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical analysis of finite mixture distributions.* Wiley, 1985.