Electronic Colloquium on Computational Complexity, Report No. 67 (2004)

# The Spectral Method for Mixture Models

Ravindran Kannan
Yale University
kannan@cs.yale.edu

Hadi Salmasian
Yale University
hadi.salmasian@yale.edu

Santosh Vempala
MIT
vempala@math.mit.edu

### Abstract

We present an algorithm for learning a mixture of distributions. The algorithm is based on spectral projection and is efficient when the components of the mixture are logconcave distributions in $\Re^n$ and their means are well-separated. The separation required grows only with $k$, the number of components and $\log n$. This improves substantially on previous results, which either focus on the special case of spherical Gaussians or Gaussians with a separation that has a much larger dependence on $n$.

## 1 Introduction

Mixture models are widely used for statistical estimation, unsupervised concept learning, text and image classification etc. [5, 10]. A finite mixture model for an unknown distribution is a weighted combination of a finite number of distributions of a known type. The problem of learning or estimating a mixture model is formulated as follows. We assume that we get samples from a distribution $F$ on $\Re^n$ which is a mixture (convex combination) of unknown distributions $F_1, F_2, \ldots, F_k$, with (unknown) mixing weights $w_1, w_2, \ldots, w_k > 0$ i.e., $F = \sum_{i=1}^k w_i F_i$ and $\sum_{i=1}^k w_i$. The goal is to (a) classify the sample points according to the underlying distributions and (b) estimate essential parameters of the components, such as the mean and covariance matrix of each component. This problem has been widely studied, particularly for the special case when each $F_i$ is a Gaussian.

One algorithm that is often used is the EM (Expectation-Maximization) algorithm. It is quite general, but does not have guarantees on efficiency

ISSN 1433-8092

and could even converge to an incorrect or suboptimal classification. A second known technique, called "projection pursuit" in statistics, projects the sample points to a random low-dimensional subspace and then tries to find the right classification by exploiting the low dimensionality and exhaustively examining all possible classifications. The trouble is that two different densities may overlap after projection — the means of the projected densities may coincide (or get closer), making it hard to separate the samples.

## 1.1 Recent theoretical work

There has been progress in recent years in finding algorithms with rigorous theoretical guarantees [2, 1, 3, 11], mostly for the important special case of learning mixtures of Gaussians. These algorithms assume a separation between the means of each pair of component distributions which depends on the variances of the two distributions and also on $n$ and $k$. For a component $F_i$ of the mixture let $\mu_i$ denotes its mean and $\sigma_i$ denote the maximum standard deviation along any direction in $\Re^n$. In order for the classification problem to have a well-defined (unique) solution with high probability, any two components $i, j$ must be separated by $\sigma_i + \sigma_j$ times a logarithmic factor; if the separation is smaller than this, then the distributions overlap significantly and some of the samples have a good chance of coming from more than one component. Dasgupta [2] showed that if each mixing weight is $\Omega(1/k)$ and the variances are within a bounded range, then a separation of (the $\Omega^*$ notation suppresses logarithmic terms and error parameters)

$$|\mu_i - \mu_j| = (\sigma_i + \sigma_j)\Omega^*(n^{1/2})$$

is enough to efficiently learn the mixture.

Shortly thereafter, this result was improved by Dasgupta and Schulman [3] and Arora and Kannan [1] who reduced the separation required to

$$|\mu_i - \mu_j| = (\sigma_i + \sigma_j)\Omega^*(n^{1/4}).$$

In [3], the algorithm used is a variant of EM (and requires some technical assumptions on the variances), while the result of [1] works for general Gaussians using distance-based classification. The idea is that at this separation, it is possible to examine just the pairwise distances of the sample point and infer the right classification with high probability.

The dependence on $n$ is critical; typically $n$ represents the number of attributes and is much larger than $k$, the size of the model. Further, the underlying method used in these papers, namely, distance-based classification, inherently needs such a large separation that grows with $n$ [1].

In [11], a spectral algorithm was used for the special case of spherical Gaussians and the separation required was reduced to

$$|\mu_i - \mu_j| = (\sigma_i + \sigma_j)\Omega^*(k^{1/4}).$$

Since $k$ is usually a constant and much less than $n$, this is a substantial improvement for the spherical case. The algorithm uses a projection of the sample to the subspace spanned by the top $k$ singular vectors of the distribution (i.e., the singular vectors of a matrix, each of whose rows is one of the iid samples drawn according to the mixture), also called the SVD subspace. The idea there was that for spherical Gaussians, that the SVD subspace of the distribution contains the means of the $k$ components. Hence, after projection to this subspace the separation between the means is preserved. On the other hand each component is still a Gaussian and the dimension is only $k$, and so the separation required is only a function of $k$. Further, even for a sample, the SVD subspace is "close" to the means and this is used in the algorithm.

## 1.2 New results

Given the success of the spectral method for spherical Gaussians, a natural question is whether it can be used for more general distributions, in particular for non-spherical Gaussians. At first sight, the method does not seem to be applicable. The property that the SVD subspace of the distribution contains the means is clearly false for non-spherical Gaussians, e.g. see Figure 1. In fact, the SVD subspace can be orthogonal to the one spanned by the means and so using spectral methods might seem hopeless.

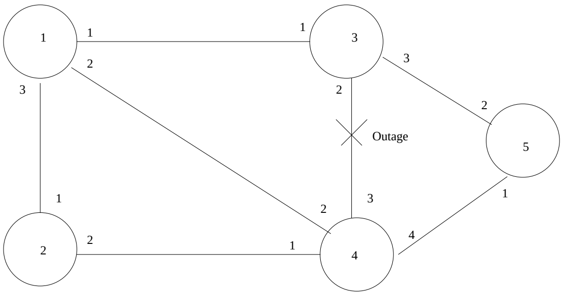

Figure 1: The SVD subspace $W$, the plane that minimizes the average squared distance, might miss the means of the components entirely.

In this paper, we show that this example is misleading and while it does not contain the means, the SVD subspace is always *close* (in an average sense) to the means of the distributions (Theorem 1). As a result, upon projection to this subspace, the inter-mean distances are approximately preserved "on average". Furthermore, this property is true for a mixture of *arbitrary* distributions.

It is then a reasonable idea to project the sample to the SVD subspace to reduce the dimensionality. To identify individual components in this subspace, we need them to remain nonoverlapping. If the mixture is arbitrary, then even though the means are separated on average, the samples could intermingle. To overcome this, we assume that the component distributions are logconcave. (A function $f : \Re^n \to \Re_+$ is logconcave if its logarithm is concave). Logconcave distributions are a powerful generalization of Gaussians, with the following properties: (a) the projection of a logconcave distribution remains logconcave (b) the distance of a random point from the mean has an exponential tail. Besides Gaussians, many other common probability measures, like the exponential family and the uniform measure over a convex set are logconcave. So for example, the mixture could consist of a Gaussian and a cube.

In Section 3, we give an iterative spectral algorithm that identifies one component of the mixture in each iteration. We assume that each mixing weight is at least $\varepsilon$ and the pairwise separation satisfies

$$|\mu_i - \mu_j| = (\sigma_i + \sigma_j)\Omega^*(k^{\frac{3}{2}}/\varepsilon^2).$$

More precisely, our algorithm only requires a lower bound $\varepsilon$, a probability of error $\delta$ an upper bound $k$, and a sample set from an $n$-dimensional mixture distribution of size $\Omega(\frac{n}{\varepsilon}\log^5(nk/\delta))$ which satisfies the given separation, and it classifies all but a fixed number with probability at least $1 - \delta$ (Theorem 3). It is easy to see that it requires time polynomial in $n, \varepsilon, \log(\frac{1}{\delta})$. The means and covariance matrices of the components can be estimated using $O(\frac{n}{\varepsilon}\log^5(nk/\delta))$ samples (Theorem 2). For the special case of Gaussians, $O(n\log^3(n/\delta)/\varepsilon)$ samples suffice. Table 1 presents a comparison of algorithms for learning mixtures (logarithmic terms are suppressed).

### 1.3 Notation

A mixture $F$ has $k$ components $F_1, \dots, F_k$. We denote their mixing weights by $w_1, \dots, w_k$ and their means by $\mu_1, \dots \mu_k$. The maximum variance of $F_i$ in any direction is denoted by $\sigma_i^2$. For any subspace $W$, we denote the maximum variance of $F_i$ along any direction in $W$ by $\sigma_{i,W}^2$.

Let $S$ be a set of iid samples $S$ from $F$. One can think of $S$ as being picked as follows: first $i$ is picked from $\{1, 2, \dots, k\}$ with probability $w_i$ (unknown to the algorithm); then a sample is picked from $F_i$. We can partition $S$ as

| Authors | Separation | Assumptions | Method |
|---|---|---|---|
| Dasgupta [2] | $n^{1/2}$ | Gaussians, bounded variances and $w_i = \Omega(1/k)$ | Random projection |
| Dasgupta-Schulman [3] | $n^{1/4}$ | Spherical Gaussians | EM+distances |
| Arora-Kannan [1] | $n^{1/4}$ | Gaussians | Distances |
| Vempala-Wang [11] | $k^{1/4}$ | Spherical Gaussians | Spectral projection |
| This paper | $\frac{k^{\frac{3}{2}}}{\varepsilon^2}$ | Logconcave distributions, $w_i \geq \varepsilon$ | Spectral projection |

Table 1: Comparison

$S = S_1 \cup S_2 \cup \ldots \cup S_k$ where each $S_i$ is from $F_i$ (note: this partition of $S$ is unknown to the algorithm). For each $i$, we denote by $\mu_i^S$ the sample mean, i.e.,

$$\mu_i^S = \frac{1}{|S_i|} \sum_{x \in S_i} x.$$

For a subspace $V$ and a vector $x$, we write $d(x, V)$ for the orthogonal distance of $x$ from $V$.

For any set of points $S$, we can form a matrix $A$ whose rows are the points in $S$. The subspace spanned by the top $k$ right singular vectors of $A$ will be called the *SVD subspace* of $S$. For any subspace $W$, we denote the maximum variance of a set of sample points in $S = S_1 \cup \ldots \cup S_k$ which belong to $S_i$ along any direction in $W$ by $\hat{\sigma}^2_{i,W}(S)$.

## 2 The SVD subspace

In this section, we prove an important property of spectral projection. The theorem says that the SVD subspace of a sample is close to the means of the samples from each component of the mixture, where "close" is in terms of the sample variances. Note that the theorem holds for *any* mixture. In the analysis of our algorithm, we will apply it only to mixtures of logconcave distributions.

**Theorem 1** *Let $S = S_1 \cup S_2 \ldots \cup S_k$ be a sample from a mixture $F$ with $k$ components such that $S_i$ is from the $i$th component $F_i$ and let $W$ be the SVD subspace of $S$. For each $i$, let $\mu_i^S$ be the mean of $S_i$ and $\hat{\sigma}^2_{i,W}(S)$ be the* *maximum variance of $S_i$ along any direction in $W$. Then,*

$$\sum_{i=1}^{k} |S_i| d(\mu_i^S, W)^2 \leq k \sum_{i=1}^{k} |S_i| \hat{\sigma}_{i,W}^2(S).$$

**Proof.** Let $\tilde{W}$ be the span of $\mu_1^S, \mu_2^S, \ldots, \mu_k^S$. For $x \in \Re^n$, write $\tilde{x}$ for the projection of $x$ onto $\tilde{W}$ and $\hat{x}$ for the projection of $x$ onto $W$.

We have (since $\mu_i^S$ is the average of $x \in S_i$ and $\mu_i^S \in \tilde{W}$),

$$\begin{aligned} \sum_{x \in S} |\tilde{x}|^2 &= \sum_{i=1}^{k} \sum_{x \in S_i} |\tilde{x} - \mu_i^S|^2 + \sum_{i=1}^{k} |S_i||\mu_i^S|^2 \\ &\geq \sum_{i=1}^{k} |S_i||\mu_i^S|^2 \\ &= \sum_{i=1}^{k} |S_i||\hat{\mu}_i^S|^2 + \sum_{i=1}^{k} |S_i| d(\mu_i^S, W)^2. \end{aligned} \tag{1}$$

On the other hand,

$$\begin{aligned} \sum_{x \in S} |\hat{x}|^2 &= \sum_{i=1}^{k} \sum_{x \in S_i} |\hat{x} - \hat{\mu}_i^S|^2 + \sum_{i=1}^{k} |S_i||\hat{\mu}_i^S|^2 \\ &\leq k \sum_{i=1}^{k} |S_i| \hat{\sigma}_{i,W}^2(S) + \sum_{i=1}^{k} |S_i||\hat{\mu}_i^S|^2. \end{aligned} \tag{2}$$

It is well-known that the SVD subspace maximizes the sum of squared projections among all subspaces of rank at most $k$ (alternatively, it minimizes the sum of squared distances to the subspace; see e.g. [4]). From this, we get

$$\sum_{x \in S} |\hat{x}|^2 \geq \sum_{x \in S} |\tilde{x}|^2.$$

Using this, the RHS of (2) is at least the RHS of (1) and the theorem follows. □

The same proof also yields an inequality for the entire distribution:

$$\sum_{i=1}^{k} w_i d(\mu_i, W)^2 \leq k \sum_{i=1}^{k} w_i \sigma_i^2.$$

Here $W$ is the SVD subspace of the entire distribution (subspace spanned by the top $k$ principal components of the distribution).

## 3 An iterative spectral algorithm

In this section, we describe the algorithm. It follows the method suggested by Theorem 1, namely, to project on the SVD subspace and to try to identify components in that subspace. However, since pairwise distances are only preserved in an average sense, it is possible that some means are very close to each other in the projected subspace and we cannot separate the corresponding samples. To get around this, we will show that all "large" components remain well-separated from the rest and there is at least one large component. We identify this component, filter it from the sample and repeat. For technical reasons (see below), the samples used to compute the SVD are discarded. The input to the algorithm below is a set of $N$ iid samples and a parameter $N_0 < N$.

**Algorithm.**

```
Repeat while there are samples left:

  1. For a subset $S$ of size $N_0$, find the $k$-dimensional SVD subspace $W$.

  2. Discard $S$ and project the rest, $T$, to the subspace $W$.

  3. For each projected point $p$:
     --- Find the closest $\varepsilon N/2$ points.  Let this set be $T(p)$ with mean $\mu(p)$.
     --- Form the matrix $A(p)$ whose rows are $x$-$\mu(p)$ for each $x$ in $T(p)$.
     Compute the largest singular value $\sigma(p)$ of $A(p)$ (Note:  this is the
     maximum standard deviation of $T(p)$ over all directions in $W$).

  4. Find a point $p_0$ for which $\sigma(p_0)$ is maximum.  Let $T_0$ be the set of all
     points of $T$ whose projection to $W$ is within distance $\frac{\sqrt{k}\log N}{\varepsilon}\sigma(p)$ of $p_0$.

  5. Label $T_0$ as one component; estimate its mean and covariance matrix.

  6. Delete $T_0$ from $T$.
```

In step 3 of the algorithm, for any point $p$, the top singular value $\sigma(p)$ of $A(p)$ can also be expressed as follows:

$$\sigma(p)^2 = \max_{v \in W, |v|=1} \frac{1}{|T_p|} \sum_{q \in T_p} |q \cdot v|^2 - \left( \frac{1}{|T_p|} \sum_{q \in T_p} q \cdot v \right)^2 .$$

This value is an estimate of the maximum variance of the entire subsample of the component to which $p$ belongs.

There is a technical issue concerning independence. If we use the entire sample to compute the SVD subspace $W$, then the sample is not independent from $W$. So we use a subset $S$ to compute the SVD subspace in each iteration and discard it. The rest of the sample, i.e., the part not used for SVD computation is classifed correctly with high probability. The size of the subset $S$ in each iteration is $N_0$.

We can state guarantees for the algorithm in two different ways. The first is a guarantee that the estimated means and covariances are approximations of the means and covariances of individual components. Recall that the covariance matrix of a distribution $G$ with mean $\mu$ is $\mathsf{E}_G((x-\mu)(x-\mu)^T)$, the matrix whose $ij$'th is the covariance of the $i$th and $j$th coordinates of a random point $x$ drawn from $G$.

**Theorem 2** *For $0 < \eta < 1$, let*

$$N_0 = C\frac{n}{\varepsilon\eta^2}\log^5\left(\frac{n}{\eta\delta}\right).$$

*Suppose that have have $2kN_0$ iid samples from a mixture $F$ of $k$ logconcave distributions in $\Re^n$, with mixing weights at least $\varepsilon$ and the means separated as*

$$\forall\, i,j \quad |\mu_i - \mu_j| \geq 1024(\sigma_i + \sigma_j)\left(\frac{k^{\frac{3}{2}}}{\varepsilon^2}\right)\log 2kN_0.$$

*Then the iterative spectral algorithm with this setting of $N_0$ finds approximations $\mu_1', \ldots, \mu_k'$ to the means and $A_1, \ldots, A_k$ to the covariance matrices such that with probability at least $1-\delta$, for $1 \leq i \leq k$,*

$$|\mu_i - \mu_i'| \leq \eta\sigma_i \quad \text{and} \quad \left\|A_i^{-1}\mathsf{E}_{F_i}\left((x-\mu_i)(x-\mu_i)^T\right) - I\right\|_F \leq \eta$$

*where $||\cdot||_F$ is the Frobenius norm, the square root of the sum of the squares of all the entries.*

A second guarantee is when we have $N$ samples and the separation grows with $\log N$. In this case, we can classify all but $kN_0$ samples.

**Theorem 3** *Suppose we have $N$ iid samples from a mixture $F$ of $k$ logconcave distributions with mixing weights at least $\varepsilon$ and the means of the components separated as*

$$\forall\, i,j \quad |\mu_i - \mu_j| \geq 1024(\sigma_i + \sigma_j)\left(\frac{k^{\frac{3}{2}}}{\varepsilon^2}\right)\log N.$$

*Then, for any $0 < \delta < 1$, with*

$$N_0 = C\frac{n}{\varepsilon}\log^5\frac{n}{\delta},$$

*the iterative spectral algorithm correctly classifies $N - 2kN_0$ samples with probability at least $1-\delta$ (a subset of $2kN_0$ samples are used by the algorithm and discarded).*

We will prove Theorem 3 in the next section. The proof of Theorem 2 is very similar.

# 4 Analysis

## 4.1 Preliminaries

We begin with some properties of logconcave distributions, paraphrased from [6, 7]. The proof of the first uses a theorem from [9].

**Lemma 1** *Let $0 < \eta < 1$ and $y_1, \ldots, y_m$ be iid samples from a logconcave distribution $G$ in $\Re^n$ whose mean is the origin. There is an absolute constant $C$ such that for*

$$m > C\frac{n}{\eta^2}\log^5\left(\frac{n}{\eta\delta}\right)$$

*with probability at least $1-\delta$, for any vector $v \in R^n$,*

$$(1-\eta)\mathsf{E}_G((v^Ty)^2) \le \frac{1}{m}\sum_{i=1}^{n}(v^Ty_i)^2 \le (1+\eta)\mathsf{E}_G((v^Ty)^2).$$

**Lemma 2** *Let $F$ be any logconcave distribution in $\Re^n$ with mean $\mu$ and second moment $\mathsf{E}_F(|X-\mu|^2) = R^2$. Then, for any $t > 1$,*

$$\Pr(|X-\mu| > tR) < e^{-t}.$$

**Lemma 3** *Let $f : \Re \to \Re_+$ be a logconcave density function with variance $\sigma^2$. Then*

$$\max_{\Re} f(x) \le \frac{1}{\sigma}.$$

### 4.2 Sample properties

Assume that $N > 2kN_0$. If $T$ is the subset of samples that are not used for SVD computation, then there is a partition $T$ as $T = T_1 \cup T_2 \cup \ldots \cup T_k$ where $T_i$ is the set of samples from $F_i$.

**Lemma 4** *With probability at least* $1 - \delta/4k$*, for every* $i \in \{1, 2, \ldots, k\}$*,*

*a.* $w_i|T| + \frac{\varepsilon}{4}|T| \leq |T_i| \leq w_i|T| + \frac{\varepsilon}{4}|T|$.

*b.* $|\mu_i - \mu_i^T| \leq \frac{\sigma_i}{4}$.

*c. For any subspace* $W$*,* $\frac{7}{8}\sigma_{i,W} \leq \hat{\sigma}_{i,W}^2(T) \leq \frac{8}{7}\sigma_{i,W}^2$.

**Proof.**

a. Follows easily from a Chernoff bound [8].

b. For any fixed $|T_i|$, the random variable $\mu_i^T = \frac{1}{|T_i|}\sum_{x \in T_i} x$ is a convolution of logconcave distributions and hence is also logconcave. Its variance is $n\sigma_i^2/|T_i|$. We apply Lemma 2 to this distribution to get the bound.

c. Follows immediately from Lemma 1.

□

In our proof, we would like to apply Theorem 1. However, the theorem holds for the sample $S$ that is used to compute the SVD subspace. The next theorem derives a similar bound for an independent sample $T$ that is not used in the SVD computation.

**Lemma 5** *Suppose* $T = T_1 \cup \ldots \cup T_k$ *is the set of sample points not used for the SVD computation in the algorithm. Then we have*

$$\sum_{i=1}^{k} |T_i| d(\mu_i^S, W)^2 \leq 2k \sum_{i=1}^{k} |T_i| \hat{\sigma}_i^2 \tag{3}$$

*where* $\hat{\sigma}_i^2 = \hat{\sigma}_{i,W}^2(T)$ *is the maximum variance of* $T_i$ *along any direction in* $W$*.*

**Proof.** First, we apply Theorem 1 to $S$. Then, using Lemma 4(a), we can related $|T_i|$ to $|S_i|$ and we have

$$\sum_{i=1}^{k} |T_i| d(\mu_i^S, W) \leq \frac{3}{2} k \sum_{i=1}^{k} |T_i| \hat{\sigma}_{i,W}^2(S).$$

Next, Lemma 1 implies that

$$\hat{\sigma}_{i,W}^2(S) \leq \frac{7}{6}\sigma_{i,W}^2.$$

Finally, we use the lower bound in Lemma 4(c) to get the desired inequality. □

### 4.3 Proof of Theorem 3

We will prove the following claim: With probability at least $1 - (\delta/2k)$, the algorithm identifies one component exactly in any one iteration. We will prove the claim for the first iteration and it will follow inductively for all subsequent iterations.

Let $T = T_1 \cup T_2 \cup \ldots \cup T_k$ be the partition of the current sample $T$ according to the components $F_i$. For each $i$, recall that $\mu_i^T$ is the sample mean and define $\hat{\mu}_i^T$ to be the projection of $\mu_i^T$ to the subspace spanned by $W$. Similarly, we have $\mu_i^S$ and $\hat{\mu}_i^T$. For convenience, we write $\hat{\sigma}_{i,W}(T)^2$ as $\hat{\sigma}_i^2$. Let

$$\alpha = 1024\frac{k^{\frac{3}{2}}}{\varepsilon^2}\log N \quad \text{and} \quad \beta = \frac{\varepsilon^3}{8096k}.$$

We say that a component $F_r$ is *large* if the following condition holds:

$$|T_r|\hat{\sigma}_r^2 \geq \beta \max_i |T_i|\hat{\sigma}_i^2. \tag{4}$$

The proof is based on the next two lemmas.

**Lemma 6** *For any large component $F_r$, for every $i \neq r$,*

$$|\hat{\mu}_i^T - \hat{\mu}_r^T| > \frac{\alpha}{8}(\sigma_i + \sigma_r).$$

**Proof.** Let $d_r = d(\mu_r^S, W)$. For any large component $r$ satisfying (4), by (3),

$$|T_r|d_r^2 \leq 2k\sum_i |T_i|\hat{\sigma}_i^2 \leq \frac{2k^2}{\beta}|T_r|\hat{\sigma}_r^2. \tag{5}$$

Thus,

$$d_r^2 \leq \frac{2k^2}{\beta}\hat{\sigma}_r^2 \leq \frac{\alpha^2}{16}\hat{\sigma}_r^2.$$

Next, let

$$R = \{i \neq r \,:\, |\hat{\mu}_i^S - \hat{\mu}_r^S| \leq \frac{\alpha}{4}(\sigma_i + \sigma_r)\}.$$

Then, by the assumed separation, for each $i \in R$, we must have (using Lemma 4b)

$$d_i = d(\mu_i^S, W) \geq |\mu_i - \mu_r| - |\mu_i - \mu_i^S| - |\mu_r - \mu_r^S| - d_r - |\hat{\mu}_i^S - \hat{\mu}_r^S| \geq \frac{\alpha}{3}\sigma_r \geq \frac{\alpha}{4}\hat{\sigma}_r.$$

Therefore, using (5),

$$\begin{aligned} \frac{2k^2}{\beta}|T_r|\hat{\sigma}_r^2 \geq 2k\sum_{i=1}^{k}|T_i|\hat{\sigma}_i^2 &\geq \sum_{i=1}^{k}|T_i|d_i^2 \\ &\geq \sum_{i\in R}|T_i|d_i^2 \geq \sum_{i\in R}|T_i|\frac{\alpha^2}{16}\hat{\sigma}_r^2. \end{aligned}$$

As a result,

$$\sum_{i\in R}|T_i| \leq \frac{32k^2}{\alpha^2\beta}|T_r| < \frac{\varepsilon}{2}|T|.$$

However, since each $|T_i| \geq \frac{\varepsilon}{2}|T|$ (by Lemma 4(a)), this implies that $R$ is empty.

To complete the lemma, we note that by Lemma 4(b), for any $j$,

$$|\hat{\mu}_j^T - \hat{\mu}_j^S| \leq |\mu_j^T - \mu_j^S| \leq |\mu_j^T - \mu_j| + |\mu_j - \mu_j^S| \leq 2\sigma_j,$$

and then use triangle inequality. □

**Lemma 7** *Let $p \in T_i$. With probability at least $1 - \delta/4k$,*

$$\sigma(p)^2 \leq 16k\hat{\sigma}_i^2.$$

*Further, if $i$ is a large component, then*

$$\sigma(p)^2 \geq \frac{w_i^2}{512}\hat{\sigma}_i^2.$$

**Proof.** By Lemma 2, within a radius of $\sqrt{2k}\sigma_{i,W}$ of any point $p$ from $T_i$, there will be at least $\varepsilon N/2$ points from the same component. Even if some points from other components are within this distance of $p$, they cannot increase $\sigma(p)$ beyond this value. To complete the proof, we use Lemma 4(c).

For the second iequality, note that by Lemma 6 the set of samples used to compute $\sigma(p)$ are all from $T_i$. If $v$ is the direction in $W$ for which the distribution $F_i$ has maximum variance, then Lemma 3 implies that for

$$H = \{x \in \Re^n : \mu^{T(p)} \cdot v - \frac{\varepsilon}{8}\sigma_{i,W} \leq v.x \leq \mu^{T(p)} \cdot v + \frac{\varepsilon}{8}\sigma_{i,W}\}$$

we have $F_i(H) \leq \frac{1}{\sigma_{i,W}} \times 2\frac{\varepsilon\sigma_{i,W}}{8} = \frac{\varepsilon}{4}$.

Now we apply VC-dimension techniques (see [12]). Suppose $|T_i| > \frac{1}{\varepsilon}$. This is guaranteed by Lemma 4a. Since the VC-dimension of intervals on a line is 2, with probability $1 - \frac{\delta}{4}$ the following statement is true:

- For any interval $I$ along the direction $v$, if $H_I = \{x \in \Re^n : x \cdot v \in I\}$, then
$$|\frac{|T_i \cap H_I|}{|T_i|} - F_i(H_I)| \leq \frac{\varepsilon}{8}.$$

Therefore $|T(p) \cap H| \leq \frac{3\varepsilon}{8}|T_i| \leq \frac{3}{4} \times \frac{\varepsilon|T_i|}{2} \leq \frac{3|T(p)|}{4}$. This means that at least $\frac{|T(p)|}{4}$ samples in $T(p)$ are out of the strip $H$, i.e. they are at least as far as $\frac{\varepsilon}{8}\sigma_{i,W}$ apart from $\mu^{T(p)}$ in the direction of $v$. Hence, using Lemma 4c

$$\begin{aligned}\sigma(p)^2 &\geq \frac{1}{|T(p)|}\sum_{x \in T(p)} (x \cdot v - \mu^{T(p)} \cdot v)^2 \\ &\geq \frac{1}{|T(p)|} \times \frac{|T(p)|}{4} \times (\frac{\varepsilon}{8}\sigma_{i,W})^2 \\ &\geq \frac{\varepsilon^2}{256}\sigma_{i,W}^2 \geq \frac{\varepsilon^2\hat{\sigma}_i^2}{512}\end{aligned}$$

which completes the proof.

□

We continue with the proof of Theorem 3. By Lemma 2, and the first part of Lemma 7, if the point $p_0$ in Step 4 that maximizes $\sigma(p)$ is from a component $r$ satisfying (4), then the set of samples identified is entirely from $T_r$. Next we will show that the point $p_0$ in step (4) must indeed be from a large component. Let $r$ be the component for which $|T_r|\hat{\sigma}_r^2$ is maximum. Take any $p \in T_i$ for an $i$ which is not large, i.e.,

$$|T_i|\hat{\sigma}_i^2 < \beta|T_r|\hat{\sigma}_r^2. \tag{6}$$

Therefore,

$$\hat{\sigma}_i^2 \leq \beta\frac{|T_r|}{|T_i|}\hat{\sigma}_r^2 \leq \frac{\beta}{\varepsilon}\hat{\sigma}_r^2.$$

By Lemma 7,

$$\sigma(p)^2 < 16k\hat{\sigma}_i^2 \leq \frac{16k\beta}{\varepsilon}\hat{\sigma}_r^2 = \frac{\varepsilon^2}{512}\hat{\sigma}_r^2.$$

On the other hand, for any point $q \in T_r$,

$$\sigma(q)^2 \geq \frac{\varepsilon^2}{512}\hat{\sigma}_r^2 > \sigma(p)^2.$$

Hence the point $p_0$ chosen in step 4 will be from a large component.

Now by Lemma 4, the number of samples we have in $T_0$ is enough to estimate the mean and covariance matrix. Finally, using these estimates, by Lemma 2, the set $T_0$ contains all the sample points from a single component with high probability.

## 5 Conclusion

Spectral projection, or principal component analysis, is fairly easy to implement and commonly used in practice for many applications. Most guarantees for spectral methods assume that the data is generated from some restricted model, such as a random model. Our algorithm is also for "random" data, but the distributions considered are more general. Spectral projection seem to be best suited for such models (unlike say random projection, which has guarantees for arbitrary input data) and our result can be viewed as further evidence of this.

## References

[1] S. Arora, R. Kannan. Learning mixtures of arbitrary Gaussians. *Proc. 33st ACM STOC*, 2001.

[2] S. DasGupta: Learning mixtures of Gaussians. *Proc. of FOCS*, 1999.

[3] S. DasGupta, L. Schulman: A two-round variant of EM for Gaussian mixtures. *Uncertainty in Artificial Intelligence*, 2000.

[4] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 1989.

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

[6] L. Lovász and S. Vempala: Logconcave functions: Geometry and Efficient Sampling Algorithms, *Proc. of FOCS*, 2003.

[7] L. Lovász and S. Vempala: The Geometry of Logconcave Functions and an $O^*(n^3)$ sampling algorithm, Microsoft Research Tech. Report MSR-TR-2003-04.

[8] R. Motwani and P. Raghavan: Randomized Algorithms, Cambridge University Press, 1995.

[9] M. Rudelson: Random vectors in the isotropic position, *J. Funct. Anal.* **164** (1999), 60–72.

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

[11] S. Vempala and G. Wang: A spectral algorithm for learning mixtures of distributions, *Proc. of FOCS*, 2002.

[12]