SIAM J. COMPUT.
Vol. 38, No. 3, pp. 1141–1156
© 2008 Society for Industrial and Applied Mathematics

THE SPECTRAL METHOD FOR GENERAL MIXTURE MODELS*

RAVINDRAN KANNAN†, HADI SALMASIAN‡, AND SANTOSH VEMPALA§

Abstract. We present an algorithm for learning a mixture of distributions based on spectral projection. We prove a general property of spectral projection for arbitrary mixtures and show that the resulting algorithm is efficient when the components of the mixture are logconcave distributions in $\Re^n$ whose means are separated. The separation required grows with k, the number of components, and $\log n$. This is the first result demonstrating the benefit of spectral projection for *general* Gaussians and widens the scope of this method. It improves substantially on previous results, which focus either on the special case of spherical Gaussians or require a separation that has a considerably larger dependence on n.

Key words. logconcave distributions, mixture models, principal component analysis, singular value decomposition

AMS subject classifications. 62H30, 68Q32, 68T05

DOI. 10.1137/S0097539704445925

1. Introduction.

Mixture models are widely used for statistical estimation, unsupervised concept learning, and text and image classification [12, 17]. Roughly speaking, a finite mixture model is a weighted combination of a finite number of distributions of a known type. More precisely, 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., $F = \sum_{i=1}^k w_i F_i$, where $\sum_{i=1}^k w_i = 1$. 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 expectation-maximization (EM) algorithm [5, 21]. It is quite general but does not have guarantees on efficiency and could even converge to an incorrect or suboptimal classification. A second known technique, from 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.

In this paper, we investigate a learning method which is based on principal component analysis (PCA). The idea of this method is known as spectral projection, i.e., representation of data in the subspace spanned by its top k principal components. We present our results following a discussion of the relevant literature.

*Received by the editors September 6, 2004; accepted for publication (in revised form) March 23, 2008; published electronically July 2, 2008. This research was supported by NSF awards ITR-0312354 and ITR-0312339 and the Sloan Foundation.

http://www.siam.org/journals/sicomp/38-3/44592.html

†Department of Computer Science, Yale University, New Haven, CT 06511 (kannan@cs.yale.edu).

‡Department of Mathematics and Statistics, University of Windsor, Windsor, ON N9B 3P4, Canada (hs79@uwindsor.ca).

§College of Computing, Georgia Tech, Atlanta, GA 30332, and Department of Mathematics, MIT, Cambridge, MA 02139 (vempala@cc.gatech.edu).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

1.1. Recent theoretical work. There has been progress in recent years in finding algorithms with rigorous theoretical guarantees [3, 2, 4, 19], 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 μ_i denote its mean and σ_i denote the maximum standard deviation along any direction in $\Re^n$ (see section 1.3). 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 may "overlap" significantly; namely, some of the samples have a good chance of coming from more than one component. Dasgupta [3] showed that if each mixing weight is $\Omega(1/k)$ and the variances are within a bounded range, then a separation of (the Ω^* 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 [4] and Arora and Kannan [2], who reduced the separation required to

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

In [4], the algorithm used is a variant of EM (and requires some technical assumptions on the variances), while the result of [2] 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 points 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 [2].

In [19], 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 in several applications k is a constant which is much less than n, this result is a substantial improvement for the spherical case. The algorithm uses a projection of the sample onto 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 independent and identically distributed (i.i.d) samples drawn according to the mixture), also called the SVD subspace. The idea there is that the SVD subspace of a mixture of spherical Gaussians *contains* the means of the k components. Hence, after projection onto 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, the SVD subspace computed from a random sample 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 nonspherical Gaussians. At first sight, the method does not seem to

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

be applicable. The property that the SVD subspace of the distribution contains the means is clearly false for nonspherical 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.

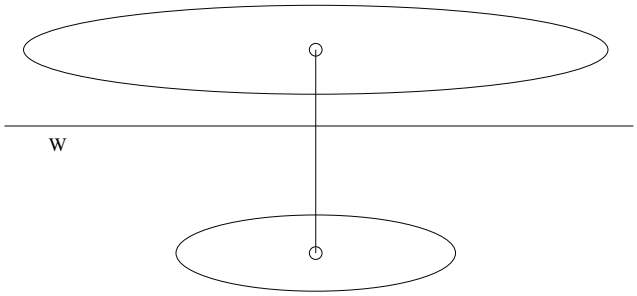


FIG. 1. *The SVD subspace W; the plane that minimizes the average squared distance and is represented by the horizontal line might miss the means of the components.*

The key insight of this paper is that while according to this (misleading) example the SVD subspace may not contain the means, it is always *close* (in an average sense) to the means of the distributions (Theorem 1). As a result, upon projection onto this subspace, the intermean distances are approximately preserved "on average." Moreover, 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; i.e., for any two points $x, y \in \Re^n$ and any $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \geq f(x)^{\lambda} f(y)^{1-\lambda}.$$

These functions have many useful properties; e.g., the product, minimum, and convolution of two logconcave functions are also logconcave [6, 11, 15]. Logconcave densities are a powerful generalization of Gaussians. Besides Gaussians, many other common probability measures, such as exponential families and the uniform measure over a convex set, are logconcave. So, for example, one component of the mixture could be a Gaussian while another is the uniform distribution over a cube. The following properties make these distributions suitable for our purpose: (a) the projection of a logconcave distribution remains logconcave; (b) the distance of a random point from the mean has an exponentially decreasing distribution.

In section 3, we give an iterative spectral algorithm that identifies one component of the mixture in each iteration. It should be emphasized that there are many possible alternatives for identifying the components *after* projection (e.g., the EM algorithm), and we expect they will also benefit from the enhancement provided by projection. For the postprojection algorithm presented here, we assume that each mixing weight is at least ε and the pairwise separation satisfies

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

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

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

TABLE 1
Comparison.

Authors	Separation	Assumptions	Method
Dasgupta [3]	$n^{\frac{1}{2}}$	Gaussians, bounded variances, and $w_i = \Omega(1/k)$	random projection
Dasgupta and Schulman [4]	$n^{\frac{1}{4}}$	spherical Gaussians	EM + distances
Arora and Kannan [2]	$n^{\frac{1}{4}}$	Gaussians	distances
Vempala and Wang [19]	$k^{\frac{1}{4}}$	spherical Gaussians	spectral projection
This paper	$\frac{k^{\frac{3}{2}}}{\varepsilon^2}$	logconcave distributions	spectral projection

1.3. Notation. This section introduces the notation used throughout the paper. For readers' convenience, in Table 2 we provide a quick reference to some of our notation. A mixture F has k components $F_1, \ldots, F_k$. We denote their mixing weights by $w_1, \ldots, w_k$ and their means by $\mu_1, \ldots \mu_k$. The maximum variance of F_i in any direction is denoted by σ_i^2. For any subspace W, we denote the maximum variance of F_i along any direction in W by $\sigma_{i,W}^2$. Namely, we set

$$\sigma_i^2 = \max_{v \in \Re^n, ||v||=1} \left\{ \int_{\Re^n} |(x-\mu_i)\cdot v|^2 F_i(x)dx \right\}$$

and

$$\sigma_{i,W}^2 = \max_{v \in W, ||v||=1} \left\{ \int_{\Re^n} |(x-\mu_i)\cdot v|^2 F_i(x)dx \right\}.$$

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

$$S = S_1 \cup S_2 \cup \cdots \cup S_k,$$

where each S_i is from F_i (note: this partition of S is unknown to the algorithm). For each i, we denote by μ_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.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

TABLE 2
Notation.

Symbol	Appears for the first time in:
$d(x, W)$	section 1.3
μ_i, μ_i^S	section 1.3
$\hat{\mu}_i, \hat{\mu}_i^S$	section 4.3
$\sigma_i^2, \hat{\sigma}_i^2$	section 1.3
$\hat{\sigma}_{i,W}^2, \hat{\sigma}_{i,W}^2(S)$	section 1.3
$\sigma(p)^2$	section 3
$T(p)$	ALGORITHM
A_i, A_i', A_i''	section 3

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. The SVD subspace can be found efficiently (i.e., in polynomial time). For more on properties and applications of SVD subspaces, see [8].

For any subspace W, we denote the maximum variance of a set of sample points in $S = S_1 \cup \cdots \cup S_k$ which belong to S_i along any direction in W by $\hat{\sigma}_{i,W}^2(S)$. Namely, we set

$$\hat{\sigma}_{i,W}^2(S) = \max_{v \in \Re^n, ||v||=1} \left\{ \frac{1}{|S_i|} \sum_{x \in S_i} |(x - \mu_i^S) \cdot v|^2 \right\}.$$

If there is no ambiguity about the set S or the space W, we use a simplified notation such as $\hat{\sigma}_{i,W}^2$ or $\hat{\sigma}_i^2$.

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, but in the analysis of our algorithm, we will apply it only to mixtures of logconcave distributions. We prove that

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

Here W is the SVD subspace of the entire distribution (subspace spanned by the top k principal components of the distribution). In Theorem 1 we state and prove a variation of this inequality for samples, but it can also be proved in a similar fashion.

THEOREM 1. *Let $S = S_1 \cup S_2 \cup \cdots \cup S_k$ be a sample from a mixture F with k components such that S_i is from the ith component F_i, and let W be the SVD subspace of S (see section 1.3). For each i, let μ_i^S be the mean of S_i and $\hat{\sigma}_{i,W}^2(S)$ be the maximum variance of S_i along any direction in W. Then*

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

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Proof. We begin by stating a general lemma which follows immediately from the Pythagorean theorem.

LEMMA 2. *Given $p, p_1, \ldots, p_K \in \Re^n$, if $\mu_P = \frac{1}{K}\sum_{i=1}^K p_i$, then*

$$\sum_{i=1}^K |p_i - p|^2 = K|p - \mu_P|^2 + \sum_{i=1}^K |p_i - \mu_P|^2.$$

Lemma 2 is used in the proof of the following theorem.

Let M be the span of $\mu_1^S, \mu_2^S, \ldots, \mu_k^S$. For $x \in \Re^n$, write $\pi_M(x)$ for the projection of x onto M and $\pi_W(x)$ for the projection of x onto W.

Using Lemma 2 and the facts that μ_i^S is the average of $x \in S_i$ and $\mu_i^S \in M$, we write

$$\begin{aligned}\sum_{x\in S} |\pi_M(x)|^2 &= \sum_{i=1}^k \sum_{x\in S_i} |\pi_M(x) - \mu_i^S|^2 + \sum_{i=1}^k |S_i||\mu_i^S|^2 \\ &\geq \sum_{i=1}^k |S_i||\mu_i^S|^2 \\ &= \sum_{i=1}^k |S_i||\pi_W(\mu_i^S)|^2 + \sum_{i=1}^k |S_i| d(\mu_i^S, W)^2. \qquad (2)\end{aligned}$$

Let $\vec{e}_1, \ldots, \vec{e}_k$ be an orthonormal basis for W. Using Lemma 2 and the fact that the variance of S_i along any direction in the k-dimensional subspace W is at most $\hat{\sigma}_{i,W}^2(S)$, we have

$$\begin{aligned}\sum_{x\in S} |\pi_W(x)|^2 &= \sum_{i=1}^k \sum_{x\in S_i} |\pi_W(x - \mu_i^S)|^2 + \sum_{i=1}^k |S_i||\pi_W(\mu_i^S)|^2 \\ &\leq \sum_{i=1}^k \sum_{j=1}^k \sum_{x\in S_i} |\pi_W(x - \mu_i^S)\cdot \vec{e}_j|^2 + \sum_{i=1}^k |S_i||\pi_W(\mu_i^S)|^2 \\ &\leq k\sum_{i=1}^k |S_i|\hat{\sigma}_{i,W}^2(S) + \sum_{i=1}^k |S_i||\pi_W(\mu_i^S)|^2. \qquad (3)\end{aligned}$$

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., [8]). From this, we get

$$\sum_{x\in S} |\pi_W(x)|^2 \geq \sum_{x\in S} |\pi_M(x)|^2.$$

Using this, the right-hand side (RHS) of (3) is at least the RHS of (2), and the theorem follows. ∎

Although we will apply Theorem 1 only for logconcave component distributions, it suggests a benefit for spectral projection more generally. Inequality (1) puts a lower bound on the average squared distance between component means after projection; if the means are well separated to begin with, they continue to be, in an average sense. On the other hand, the distance of a point from the mean of its distribution can shrink only upon projection, thus magnifying the ratio of intercomponent distance to intracomponent distance. This aspect is studied further along with empirical results in [20].

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

3. An iterative spectral algorithm. In this section, we describe the algorithm. It follows the method suggested by Theorem 1, namely, to project onto the SVD subspace and to try to identify components in that subspace. However, since pairwise distances are preserved only 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 i.i.d. samples, a weight lower bound $0 < \varepsilon < 1$, a maximum probability of error δ, a lower bound k for the number of distributions, 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 $\frac{2^8\sqrt{k}\log(\frac{Nk}{\delta})}{\varepsilon}\sigma(p)$ 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 classified correctly with high probability. The size of the subset S in each iteration is N_0, where

$$N_0 = 100 \frac{n}{\varepsilon^2} \log^5 \frac{nk}{\delta}.$$

Let $0 < \varepsilon, \delta < 1$. Our main result is the following guarantee for the proposed algorithm.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

THEOREM 3. *Suppose we have N i.i.d. samples from a mixture F of k logconcave distributions with mixing weights at least ε and the means of the components separated as follows:*

$$\forall\, i,j \quad |\mu_i - \mu_j| \geq 2^{11}(\sigma_i + \sigma_j)\left(\frac{k^{\frac{3}{2}}}{\varepsilon^2}\right)\log^2\left(\frac{Nk}{\delta}\right).$$

There exists an absolute constant C such that if $N > C\frac{kN_0}{\varepsilon}$, then the iterative spectral algorithm correctly classifies $N - kN_0$ samples with probability at least $1-\delta$ (a subset of kN_0 samples is used by the algorithm and discarded).

Remark. A crude upper bound of $C < 10^6$ can be obtained quite easily by tracing through the proof of Theorem 3. More careful estimates should yield a much smaller upper bound.

We will prove Theorem 3 in the next section. The following corollary of Theorem 3 provides a second guarantee for the algorithm in terms of estimating the means and covariances of the distributions.

COROLLARY 4. *Let $0 < \eta < 1$, and suppose we are given N i.i.d. samples from a mixture F of k logconcave distributions in $\Re^n$, with mixing weights at least ε and the means separated as stated in Theorem 3. For any $1 \leq i \leq k$, let μ_i be the mean and $A_i = \mathsf{E}_{F_i}\left((x-\mu_i)(x-\mu_i)^T\right)$ be the covariance matrix of F_i. If $N = \Omega^*(\frac{nk^2}{\delta^2\eta^2\varepsilon^3})$, then using the iterative spectral algorithm we can find approximations $\mu_1', \ldots, \mu_k'$ to the means and $A_1', \ldots, A_k'$ to the covariance matrices of the components $F_1, \ldots, F_k$ of the sample such that with probability at least $1-\delta$, for $1 \leq i \leq k$,*

$$|\mu_i - \mu_i'| \leq \eta\sigma_i \quad \text{and} \quad ||A_i - A_i'|| \leq \eta\sigma_i^2,$$

where $||\cdot||$ is the spectral norm of the matrix.

The proof of Corollary 4 is given in section 4.4.

4. Analysis.

4.1. Preliminaries.

We begin with some properties of logconcave distributions, paraphrased from [13]. The proof of the first uses a theorem from [16] (see also Proposition 2.11 of [7]).

LEMMA 5. *Let $0 < \eta < 1$ and $y_1, \ldots, y_m$ be i.i.d. 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 $v \in R^n$,

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

The next lemma is an adaptation of Lemma 5.7 of [13].

LEMMA 6. *Let F be any logconcave distribution in $\Re^n$ with mean μ and second moment $\mathsf{E}_F(|X-\mu|^2) = R^2$. There is an absolute constant c such that, for any $t > 1$,*

$$\Pr(|X-\mu| > tR) < e^{1-t}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Lemma 5.5 of [13] implies the following lemma.

LEMMA 7. *Let $f : \Re \to \Re_+$ be a logconcave density function with variance σ^2. Then*

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

4.2. Sample properties. Assume that $m > N_0$ and T is a set of m elements consisting of i.i.d. samples generated by the mixture F. Then there is a partition T as $T = T_1 \cup T_2 \cup \cdots \cup T_k$, where T_i is the set of samples from F_i.

LEMMA 8. *With probability at least $1 - \delta/4k$, for every $i \in \{1, 2, \ldots, k\}$ we have the following:*

(a) $w_i|T| - \frac{\varepsilon}{4}|T| \le |T_i| \le w_i|T| + \frac{\varepsilon}{4}|T|$.

(b) $|\mu_i - \mu_i^T| \le \frac{\sigma_i}{4}$.

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

Proof.

(a) A point $x \in T$ belongs to T_i with probability w_i. If $T = \{x_1, \ldots, x_m\}$, define a random variable $Y_j, 1 \le j \le m$, as follows:

$$Y_j = \begin{cases} 1 & \text{if } x_j \in T_i, \\ 0 & \text{otherwise.} \end{cases}$$

The Y_j's are Bernoulli variables and $\mathrm{E}(Y_j) = w_i$. Therefore by Chernoff's bound (see Theorem 4.1 of [14]) we have

$$\Pr\left(\frac{Y_1 + \cdots + Y_m}{m} - w_i > \frac{\varepsilon}{4}\right) < \left(\frac{e^{\frac{\varepsilon}{4w_i}}}{(1 + \frac{\varepsilon}{4w_i})^{(1+\frac{\varepsilon}{4w_i})}}\right)^{mw_i}. \tag{4}$$

Since $w_i \ge \varepsilon$, $\varepsilon = 4tw_i$ for some $t \le \frac{1}{4}$. The RHS of (4) equals $(\frac{e^t}{(1+t)^{(1+t)}})^{mw_i}$, and one can check that for $t \le \frac{1}{4}$ we have $\frac{e^t}{(1+t)^{(1+t)}} < e^{-\frac{t^2}{4}}$. Similarly we have

$$\Pr\left(\frac{Y_1 + \cdots + Y_m}{m} - w_i < -\frac{\varepsilon}{4}\right) < (e^{-\frac{t^2}{4}})^{mw_i}.$$

It follows that Lemma 8(a) fails for distribution F_i with probability at most $2e^{-\frac{t^2}{4}mw_i}$, which is equal to $2e^{-\frac{m\varepsilon^2}{64w_i}}$. Now Lemma 8(a) follows immediately from $m \ge N_0$.

(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 bounded from above by $n\sigma_i^2/|T_i|$. We apply Lemma 6 to the random variable μ_i^T. Now Lemma 8(b) follows from Lemma 8(a).

(c) The proof follows immediately from Lemma 5. ∎

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 lemma derives a similar bound for an independent sample T that is not used in the SVD computation.

LEMMA 9. *Let W be the SVD subspace obtained in the algorithm, and suppose $T = T_1 \cup \cdots \cup T_k$ is the set of sample points not used for the SVD computation in the algorithm. Then we have*

$$\sum_{i=1}^{k} |T_i| d(\mu_i^S, W)^2 \le 2k \sum_{i=1}^{k} |T_i| \hat{\sigma}_i^2, \tag{5}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

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 8(a), we can relate $|T_i|$ to $|S_i|$, and we have

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

Next, Lemma 5 implies that

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

Finally, we use the lower bound in Lemma 8(c) to get the desired inequality. ∎

4.3. Proof of Theorem 3. We will prove the following claim: With probability at least $1 - \frac{\delta}{2k}$, 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. From now on, we assume that Lemma 8 holds. Note that this happens with probability at least $1 - \frac{\delta}{4k}$.

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

$$\alpha = 2^{11} \frac{k^{\frac{3}{2}}}{\varepsilon^2} \log^2\left(\frac{Nk}{\delta}\right) \quad \text{and} \quad \beta = \frac{\varepsilon^3}{2^{14} k \log^2(\frac{Nk}{\delta})}.$$

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

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

Recall that $|T| > N_0$. The proof of Theorem 3 is based on the next two lemmas.

LEMMA 10. *For any large component* F_r*, for every* $i \ne r$*,*

$$|\hat{\mu}_i^T - \hat{\mu}_r^T| > \frac{\alpha}{5}(\sigma_i + \sigma_r).$$

Proof. For any $1 \le j \le r$, let $d_j = d(\mu_j^S, W)$. By (5),

$$|T_r| d_r^2 \le 2k \sum_i |T_i| \hat{\sigma}_i^2 \le \frac{2k^2}{\beta} |T_r| \hat{\sigma}_r^2. \tag{7}$$

Thus,

$$d_r^2 \le \frac{2k^2}{\beta} \hat{\sigma}_r^2 \le \frac{\alpha^2}{16} \hat{\sigma}_r^2. \tag{8}$$

Next, let

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

Then we have

$$\begin{aligned} d_i = d(\mu_i^S, W) &= |\mu_i^S - \hat{\mu}_i^S| \\ &= |(\mu_i^S - \mu_i) + (\mu_i - \mu_r) + (\mu_r - \mu_r^S) + (\mu_r^S - \hat{\mu}_r^S) + (\hat{\mu}_r^S - \hat{\mu}_i^S)| \end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

and, by the triangle inequality,

$$\begin{aligned} d_i &\geq |\mu_i - \mu_r| - |(\mu_i^S - \mu_i) + (\mu_r - \mu_r^S) + (\mu_r^S - \hat{\mu}_r^S) + (\hat{\mu}_r^S - \hat{\mu}_i^S)| \\ &\geq |\mu_i - \mu_r| - |\mu_i - \mu_i^S| - |\mu_r - \mu_r^S| - d_r - |\hat{\mu}_i^S - \hat{\mu}_r^S|. \end{aligned}$$

Next, we use the inequality above to show that

$$d_i \geq \frac{\alpha}{3}\sigma_r. \tag{9}$$

To this end, we need inequalities given in (10), (11), and (12). By the separation assumed in Theorem 3, we have

$$|\mu_i - \mu_r| \geq \alpha(\sigma_i + \sigma_r). \tag{10}$$

From Lemma 8(b) we have

$$|\mu_i - \mu_i^S| \leq \frac{\sigma_i}{4} \qquad \text{and} \qquad |\mu_r - \mu_r^S| \leq \frac{\sigma_r}{4}. \tag{11}$$

Moreover, from (8) and Lemma 8(c) it follows that

$$d_r \leq \frac{\alpha}{4}\hat{\sigma}_r \leq \frac{2\alpha}{7}\sigma_{r,W} \leq \frac{2\alpha}{7}\sigma_r. \tag{12}$$

From these facts (9) follows immediately.

Next, from Lemma 8(c) and (9) it follows that

$$d_i \geq \frac{\alpha}{4}\hat{\sigma}_r.$$

Therefore, using (7) and Lemma 9,

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

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 8(a)), this implies that R is empty. Consequently, for any $i \neq r$ we have

$$|\hat{\mu}_i^S - \hat{\mu}_r^S| > \frac{\alpha}{4}(\sigma_i + \sigma_r). \tag{13}$$

To complete the proof of the lemma, we note that by Lemma 8(b), for any $1 \leq j \leq r$,

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

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Therefore, by (13) and the triangle inequality,

$$\begin{aligned}|\hat{\mu}_i^T - \hat{\mu}_r^T| &= |(\hat{\mu}_i^T - \hat{\mu}_i^S) + (\hat{\mu}_i^S - \hat{\mu}_r^S) + (\hat{\mu}_r^S - \hat{\mu}_r^T)| \\ &\geq |\hat{\mu}_i^S - \hat{\mu}_r^S| - |\hat{\mu}_r^S - \hat{\mu}_r^T| - |\hat{\mu}_i^S - \hat{\mu}_i^T| \\ &\geq \frac{\alpha}{4}(\sigma_i + \sigma_r) - \frac{\sigma_i}{2} - \frac{\sigma_r}{2} \\ &> \frac{\alpha}{5}(\sigma_i + \sigma_r). \qquad \square\end{aligned}$$

LEMMA 11. *Let $p \in T_i$. With probability at least $1 - \frac{\delta}{4k}$,*

$$\sigma(p)^2 \leq 16k\hat{\sigma}_i^2 \log^2\left(\frac{Nk}{\delta}\right).$$

Further, if i is a large component, then

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

Proof. Consider the samples in T_i after projection onto W. These points are i.i.d. samples from a logconcave distribution with maximum variance at most $k\sigma_i^2$. Fix a (projected) sample point p. By Lemma 6,

$$\Pr\left(|p - \hat{\mu}_i| > \frac{3}{2}\sqrt{k}\sigma_{i,W}\log\left(\frac{Nk}{\delta}\right)\right) < \frac{\delta}{8kN}. \tag{14}$$

Therefore, with probability at least $1 - \frac{\delta}{8k}$, every projected sample point from any F_i lies within distance at most $\frac{3}{2}\sqrt{k}\sigma_{i,W}\log(\frac{Nk}{\delta})$ of $\hat{\mu}_i$. Consequently, with probability at least $1 - \frac{\delta}{8k}$, any pair of projected samples from T_i are within $3\sqrt{k}\sigma_{i,W}\log(\frac{Nk}{\delta})$ of each other. Since by Lemma 8(a) $|T_i| > \frac{N\varepsilon}{2}$, it follows that elements of $T(p)$ lie within a ball of radius at most $3\sqrt{k}\sigma_{i,W}\log(\frac{Nk}{\delta})$ centered at p. From the definition of $\sigma(p)$, Lemma 2, and Lemma 8(c) it follows that

$$\sigma(p)^2 \leq \frac{1}{|T(p)|}\sum_{x \in T(p)}|x - p|^2 \leq 9k\sigma_{i,W}^2\log^2\left(\frac{Nk}{\delta}\right) \leq 16k\hat{\sigma}_i^2\log^2\left(\frac{Nk}{\delta}\right).$$

For the second inequality, note that by Lemma 10 the set $T(p)$ 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 7 implies that for

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

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 a result which follows from VC-dimension techniques (see [10]). To make the paper self-contained, we will give a proof of this result in section 4.5. For any interval I along the direction v, let $H_I = \{x \in \Re^n : x \cdot v \in I\}$.

LEMMA 12. *Suppose $T = T_1 \cup T_2 \cup \cdots \cup T_k$ is a set of i.i.d. samples generated from the mixture F such that $|T| > N_0$. Fix $i \in \{1, \ldots, r\}$, and let I be any interval along the direction of v such that $F_i(H_I) \leq \frac{\varepsilon}{4}$. With probability at least $1 - \frac{\delta}{8k}$, we have*

$$\left|\frac{|T_i \cap H_I|}{|T_i|} - F_i(H_I)\right| \leq \frac{\varepsilon}{8}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

We now complete the proof of Lemma 11. From Lemma 12 we have

$$|T(p) \cap H| \leq \frac{3\varepsilon}{8}|T_i| \leq \frac{3}{4} \times \frac{\varepsilon|T_i|}{2} \leq \frac{3|T(p)|}{4}.$$

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 8(c),

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

which completes the proof. ∎

We continue with the proof of Theorem 3. First, note that from Lemma 8(a) it follows that in each iteration the size of the set of samples is at least $2N_0$. The algorithm uses N_0 samples for SVD computation, and the rest are classified. Since there are at least N_0 samples left, all of the necessary lemmas will be valid.

Suppose the point p_0 which maximizes $\sigma(p)$ in step 4 of the algorithm is actually from a large component F_r (i.e., one which satisfies (6)). By Lemma 6 and the first part of Lemma 11, we have

$$\frac{\alpha}{8}\sigma_r > 2^8\frac{\sqrt{k}}{\varepsilon}\sigma(p)\log\left(\frac{Nk}{\delta}\right) > 3\sqrt{k}\sigma_{r,W}\log\left(\frac{Nk}{\delta}\right). \tag{15}$$

It was shown that from (14) it follows that with probability $1-\frac{\delta}{8k}$, any pair of projected samples onto W from F_i are within $3\sqrt{k}\sigma_{i,W}\log(\frac{Nk}{\delta})$ of each other. This, together with (15), means that the output of an iteration contains T_r entirely. Moreover, from Lemma 10 and the above consequence of (14) it follows that the distance between any projected samples of F_r and F_i, $i \neq r$, is at least $\frac{\alpha}{8}\sigma_r$. This, together with (15), implies that the output of the iteration does not contain any samples from any F_i, $i \neq r$. Therefore the output of the iteration exactly classifies samples from one distribution.

Next, we will show that the point p_0 in step 4 of the algorithm 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.,

$$|T_i|\hat{\sigma}_i^2 < \beta|T_r|\hat{\sigma}_r^2. \tag{16}$$

Therefore,

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

By Lemma 11,

$$\sigma(p)^2 < 16k\hat{\sigma}_i^2\log^2\left(\frac{Nk}{\delta}\right) \leq \frac{32k\beta}{\varepsilon}\hat{\sigma}_r^2\log^2\left(\frac{Nk}{\delta}\right) = \frac{\varepsilon^2}{512}\hat{\sigma}_r^2.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

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 of the algorithm will be from a large component. This completes the proof of Theorem 3.

4.4. Proof of Corollary 4. Assume $N > \frac{C'k^2}{\varepsilon^3\eta^2\delta^2}\log^5(\frac{nk}{\delta\eta})$, where C' is a large enough constant. We use our spectral algorithm to obtain a classification of $N - kN_0$ samples such as $T = T_1 \cup \cdots \cup T_k$. By Theorem 3, with probability at least $1-\delta$ the algorithm classifies the samples correctly. For any $1 \leq i \leq k$, set

$$\mu_i' = \mu_i^T = \frac{1}{|T_i|}\sum_{x\in T_i} x \quad \text{and} \quad A_i' = \frac{1}{|T_i|}\sum_{x\in T_i}(x-\mu_i')(x-\mu_i')^T.$$

From Lemma 8(a) it follows that with probability at least $1-\delta$, for every $1 \leq i \leq k$ we have $|T_i| > \frac{4n}{\eta^2}\log^2(\frac{2k}{\delta})$. Now the samples in T_i are i.i.d. and all generated from F_i. With a method analogous to the proof of Lemma 8(b) given below, we can prove that for any $1 \leq i \leq k$, with probability at least $1 - \frac{\delta}{2k}$ we have $|\mu_i - \mu_i'| < \frac{\eta}{2}\sigma_i$.

Finally, let G_i be the shifted logconcave distribution given by

$$G_i(x) = F_i(x + \mu_i).$$

Clearly the mean of G_i is located at the origin. Set

$$A_i'' = \frac{1}{|T_i|}\sum_{x\in T_i}(x-\mu_i)(x-\mu_i)^T.$$

Since $|T_i| = \Omega\Big(\frac{n}{(\frac{\delta^2\eta^2}{4k^2})}\log^5(\frac{n}{(\frac{\delta^2\eta^2}{4k^2})})\Big)$, Lemma 5 implies that for any $1 \leq i \leq k$, with probability at least $1 - \frac{\delta}{2k}$ for any vector $v \in \Re^n$ we have

$$\begin{aligned}|v^T A_i'' v| &= \left|\left(\frac{1}{|T_i|}\sum_{y\in T_i-\mu_i}(v^T y)^2\right) - \mathsf{E}_{G_i}((v^T y)^2)\right| \\ &\leq \frac{\eta}{2}\mathsf{E}_{G_i}((v^T y)^2) = \frac{\eta}{2}v^T A_i v,\end{aligned}$$

which implies that $||\sum_{x\in T_i}(x-\mu_i)(x-\mu_i)^T|| \leq \frac{\eta}{2}||A_i|| \leq \frac{\eta}{2}\sigma_i^2$. Next, a simple calculation shows that

$$A_i' = A_i'' - (\mu_i - \mu_i')(\mu_i - \mu_i')^T,$$

which, together with $|\mu_i - \mu_i'| < \frac{\eta}{2}\sigma_i$, implies that $||A_i' - A_i|| < \eta\sigma_i^2$.

4.5. Proof of Lemma 12. In this section we prove Lemma 12. Our notation will be the same as in the statement of the lemma. Let $\mathcal{L}_v$ be the line in the direction of v. Let $t = \lceil\frac{30}{\varepsilon}\rceil$. Choose a sequence of closed intervals $I_1, \ldots, I_t$ on $\mathcal{L}_v$ such that the following hold:

- For any $1 \leq j \leq t$, we have $F(H_{I_j}) = \frac{1}{t}$.
- $\mathcal{L}_v = I_1 \cup \cdots \cup I_t$.
- The interiors of $I_1, \ldots, I_t$ are disjoint.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

For any $1 \le j \le t$, let r_j be the number of samples in T_i which belong to H_{I_j}. Using Chernoff's inequality one can see that

$$\Pr\left(\left|\frac{r_j}{|T_i|} - \frac{1}{t}\right| > \frac{1}{100t}\right) < 2e^{-\frac{|T_i|}{10^4 t}}.$$

From Lemma 8(a) it follows that with probability at least $1 - \frac{\delta}{8k}$, for any j we have $|\frac{r_j}{|T_i|} - \frac{1}{t}| < \frac{1}{100t}$. Next, we show that when this holds, the inequality of the statement of the lemma holds as well.

Let I be an interval such that $F_i(H_I) = l \le \frac{\varepsilon}{4}$. Suppose a, b are chosen such that $I_{a+1} \cup \cdots \cup I_{b-1} \subset I \subset I_a \cup \cdots \cup I_b$. Then $(b-a+1) > lt$ and $b-a-1 < lt$. Therefore

$$\begin{aligned}\frac{|T_i \cap H_I|}{|T_i|} &< \frac{r_a + \cdots + r_b}{|T_i|} < (b-a+1)\left(\frac{1}{t} + \frac{1}{100t}\right) < (lt+2)\left(\frac{101}{100t}\right) \\ &= l + \frac{l}{100} + \frac{101}{50t} < l + \frac{\varepsilon}{8}\end{aligned}$$

and, similarly,

$$\begin{aligned}\frac{|T_i \cap H_I|}{|T_i|} &> \frac{r_{a+1} + \cdots + r_{b-1}}{|T_i|} > (b-a-1)\left(\frac{1}{t} - \frac{1}{100t}\right) > (lt-2)\left(\frac{99}{100t}\right) \\ &= l - \frac{1}{100t} - \frac{99}{50t} > l - \frac{\varepsilon}{8}.\end{aligned}$$

5. Concluding remarks. In this paper we showed that assuming a mild separation for the means a PCA-based learning algorithm provably classifies samples from a mixture of logconcave distributions.

From the example in Figure 1, it is not hard to see that spectral projection requires a separation between means that grows with the largest variance of individual components. Following the preliminary version of our results [18, 9], Achlioptas and McSherry [1] have improved the polynomial dependence on k and ε using a more sophisticated algorithm *after* projection. It remains an open problem to learn (nonspherical) Gaussians at smaller separation. The near disjointness of Gaussian components (e.g., total variation distance nearly 1 for two Gaussians), and thus their learnability, is implied by the assumption that the distance between two means is of the order of the variance *in the direction of* the line joining the means. Spectral projection fails at such a small separation, and a different technique will have to be used.

On the other hand, the main technique used in this paper is fairly easy to implement and commonly used in practice for many applications. In [20], Vempala and Wang present empirical evidence and propose an explanation for why the method is effective for real data where the assumption that the components are Gaussian (or logconcave) might not be valid.

Finally, 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 well suited for such models, and our result can be viewed as further evidence of this.

Acknowledgment. We thank the referees for patiently reading the original version of the paper and making several useful suggestions.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

REFERENCES

[1] D. Achlioptas and F. McSherry, *On spectral learning of mixtures of distributions*, in Learning Theory, Springer, Berlin, 2005, pp. 458–469.
[2] S. Arora and R. Kannan, *Learning mixtures of arbitrary Gaussians*, Ann. Appl. Probab., 15 (2005), pp. 69–92.
[3] S. Dasgupta, *Learning mixtures of Gaussians*, in Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science, 1999, pp. 634–644.
[4] S. Dasgupta and L. Schulman, *A two-round variant of EM for Gaussian mixtures*, in Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, 2000, pp. 152–159.
[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm. With discussion*, J. Roy. Statist. Soc. Ser. B, 39 (1977), pp. 1–38.
[6] A. Dinghas, *Über eine Klasse superadditiver Mengenfunktionale von Brunn–Minkowski–Lusternik-schem Typus*, Math. Z., 68 (1957), pp. 111–125.
[7] A. A. Giannopoulos and V. D. Milman, *Concentration property on probability spaces*, Adv. Math., 156 (2000), pp. 77–106.
[8] G. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.
[9] R. Kannan, H. Salmasian, and S. Vempala, *The spectral method for general mixture models*, in Learning Theory, Lecture Notes in Comput. Sci. 3559, Springer, Berlin, 2005, pp. 444–457.
[10] M. Kearns and U. Vazirani, *An Introduction to Computational Learning Theory*, MIT Press, Cambridge, MA, 1994.
[11] L. Leindler, *On a certain converse of Hölder's Inequality* II, Acta Sci. Math. (Szeged), 33 (1972), pp. 217–223.
[12] B. Lindsay, *Mixture Models: Theory, Geometry and Applications*, American Statistical Association, Alexandria, VA, 1995.
[13] L. Lovász and S. Vempala, *The geometry of logconcave functions and sampling algorithms*, Random Structures Algorithms, 30 (2007), pp. 307–358.
[14] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, Cambridge, UK, 1995.
[15] A. Prékopa, *Logarithmic concave measures and functions*, Acta Sci. Math. (Szeged), 34 (1973), pp. 335–343.
[16] M. Rudelson, *Random vectors in the isotropic position*, J. Funct. Anal., 164 (1999), pp. 60–72.
[17] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley, Chichester, UK, 1985.
[18] S. Vempala, *On the Spectral Method for Mixture Models*, IMA Workshop on Data Analysis and Optimization, 2003; http://www.ima.umn.edu/talks/workshops/5-6-9.2003/vempala/vempala.html.
[19] S. Vempala and G. Wang, *A spectral algorithm for learning mixtures of distributions*, J. Comput. System Sci., 68 (2004), pp. 841–860.
[20] S. Vempala and G. Wang, *The benefit of spectral projection for document clustering*, in Proceedings of the 3rd Annual Workshop on Clustering High Dimensional Data and its Applications, SIAM International Conference on Data Mining, 2005.
[21] C. F. J. Wu, *On the convergence properties of the EM algorithm*, Ann. Statist., 11 (1983), pp. 95–103.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.