Learning mixtures of arbitrary gaussians

[Extended Abstract] *

Sanjeev Arora†
Dept of Computer Science
Princeton University

Ravi Kannan‡
Dept of Computer Science
Yale University

ABSTRACT

Mixtures of gaussian (or normal) distributions arise in a variety of application areas. Many techniques have been proposed for the task of finding the component gaussians given samples from the mixture, such as the *EM algorithm*, a local-search heuristic from Dempster, Laird and Rubin (1977). However, such heuristics are known to require time exponential in the dimension (i.e., number of variables) in the worst case, even when the number of components is 2.

This paper presents the first algorithm that provably learns the component gaussians in time that is polynomial in the dimension. The gaussians may have arbitrary shape provided they satisfy a "nondegeneracy" condition, which requires their high-probability regions to be not "too close" together.

Categories and Subject Descriptors

G.4 [**Mathematical Software**]; I.5.3 [**Pattern Recognition**]: Clustering; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Mixtures of gaussians

Keywords

Mixture distributions, gaussians, learning, clustering

1. INTRODUCTION

Finite mixture models are ubiquitous in a host of areas that use statistical techniques, including AI, computer vision, medical imaging, psychology, geology etc. (see [13, 18]). A mixture of distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots$ with mixing weights $w_1, w_2, w_3, \ldots$ (where $\sum_i w_1 = 1$) is the distribution in which a sample is produced by first picking a component distribution —the ith one is picked with probability w_i— and then producing a sample from that distribution. In many applications the component distributions are multivariate gaussians.

Given samples from the mixture distribution, how can one learn (i.e., reconstruct) the component distributions and their mixing weights? The most popular method is probably the EM algorithm of Dempster, Laird and Rubin (1977). EM is a local search heuristic that tries to converge to a *maximum likelihood* description of the data by trying to cluster samplepoints according to which gaussian they came from. Though moderately successful in practice, it often fails to converge or gets stuck in local optima. Much research has gone into fixing these problems, but has not yet resulted in an algorithm that provably runs in polynomial time.

Recently, Dasgupta (1999) took an important step towards the design of such an algorithm by showing how a mixture of "spherelike" gaussians could be learned in polynomial time provided they have approximately the same "radius." (Roughly speaking, "spherelike" gaussians are those in which most of the probability mass is concentrated in a thin spherical shell; see Section 2.1.) Dasgupta's algorithm uses a statistical technique called *projection pursuit* (Huber [9]): first he projects data into a random subspace of logarithmic dimension and then performs simple clustering to identify the gaussian clusters (which have become spheres after projection). Recently, independently of our work, Dasgupta and Schulman [5] generalized and strengthened this result, so that the spherelike gaussians can have different radii provided the spheres are nonoverlapping.

The current paper substantially generalizes such results: we allow spheres to overlap, and also let the gaussians have arbitrary shape. Results of Dasgupta (1999) and Dasgupta and Schulman can be derived as simple subcases (see the easy analysis of Section 3.2) from our results. We also formalize the problem of learning a mixture of gaussians more generally than Dasgupta did (see Section 2.2) and use techniques developed in context of approximation algorithms to design algorithms for one of the problems (see Section 5).

The paper presents a variety of polynomial-time algorithms. Each algorithm has its own strong points, and we do not yet have a single algorithm that combines these strong

*A full version of this paper is available from `www.cs.princeton.edu/ arora/publist.htm`

†Supported by an NSF Career Award, Alfred Sloan Fellowship, and a David and Lucille Packard Fellowship.`arora@cs.princeton.edu`

‡Supported in part by NSF Grant CCR-9820850. `kannan-ravindran@cs.yale.edu`

STOC'01 July 6-8, 2001, Hersonissos, Greece

points. In general, if the gaussians in the mixture are restricted to be spherelike (possibly of different radii), the algorithms are quite efficient and work correctly and in polynomial time even if the component gaussians are quite close together. If the gaussians could have arbitrary shape, then our algorithm requires a somewhat more stringent pairwise separation. If the separation condition is violated, our algorithm may detect "inbetween" gaussians which do not really exist in the mixture. The running time is polynomial in n, the number of dimensions. The running time is polynomial also on k, the number of components, except for one of the algorithms this dependence is quasipolynomial (it is polynomial only if $k < 2^{\log^{1/4} n}$). We do not consider this a serious constraint, since even the case $k = 2$ was open thus far.

We also present a combinatorial algorithm for (approximately) maximum likelihood fit of a mixture of identical spherical gaussians to to any (possibly unstructured) set of data points. The exact problem appears to be NP-hard.

Practical matters. Our algorithm's running time is measured asymptotically, assuming the number of dimensions is "large." Our algorithms for mixtures of spherical gaussians in Section 3.2 and Section 5 may even be practical when the number of dimensions is small, say < 20. The other algorithms are less practical for so few dimensions.

2. DEFINITIONS AND OVERVIEW

The univariate distribution $N(\mu, \sigma)$ on $\Re$ has the density function $f(x) = (\sqrt{2\pi}\sigma)^{-1}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$. It satisfies $E[x^2] = \sigma^2$. The analogous distribution in $\Re^n$ is the *axis-aligned* gaussian $N(\overline{\mu}, \overline{\sigma})$ where $\overline{\mu}, \overline{\sigma} \in \Re^n$ and the density function is the product distribution of $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$, $\ldots, N(\mu_n, \sigma_n)$. A random sample $(x_1, x_2, \ldots, x_n)$ satisfies $E[\sum_i x_i^2] = \sum_i \sigma_i^2$. (Similarly, $E[\sum_i x_i^2/\sigma_i^2] = n$.)

A general gaussians in $\Re^n$ is obtained from an axis-aligned gaussian by applying an arbitrary rotation. Specifically, its probability density function has the form

$$F_{Q,p}(x) = (2\pi)^{-n/2} \prod_i \sqrt{\lambda_i(Q)} \exp\left(-(x-p)^T Q (x-p)/2\right). \tag{1}$$

where Q is an $n \times n$ positive definite matrix with eigenvalues $\lambda_1(Q), \ldots, \lambda_n(Q) > 0$, and $p \in \Re^n$ is the *center*. Since Q can be rewritten as $R^{-1} diag(\lambda_i(Q)) R$ where R is a rotation, the quantities $1/\lambda_i(Q)$ play the same role as the variances σ_i^2 in the axis-aligned case. For this reason the quantity $(x-p)^T Q(x-p)$ is a norm, called *Mahalonobis distance with respect to* Q. From our earlier discussion, $E[(x-p)^T(x-p)]$ (the "radius") is $\sum_i \lambda_i^{-1}(Q)$ and $E[(x-p)^T Q(x-p)] = \int_x F_{Q,p}(x-p)^T Q(x-p) = n$.

For any finite sample of points in $\mathbf{R}^n$ we can try to fit a gaussian by computing their *variance-covariance matrix*. Let $x_1, x_2, \ldots x_N$ be N points in $\mathbf{R}^n$. Let X be the $n \times N$ matrix whose columns are the vectors $x_1 - q, x_2 - q, \ldots x_N - q$, where $q = \frac{1}{N}(x_1 + x_2 + \ldots x_N)$ is the sample mean. Then the variance-covariance matrix[1] $A = \frac{1}{N} X X^T$; note that it is positive definite by definition. The *maximum likelihood gaussian fit* for these points is $F_{A,q}$.

This fit may, of course, be poor for an arbitrary point set. But if the points are sampled from gaussian $F_{G,p}$, then $F_{A,p}$ converges rapidly to $F_{G,p}$: with $\Omega(n \log n/c_\epsilon)$ samples (where c_ϵ is a constant depending upon ϵ), $F_{A,q}$ is whp a $(1+\epsilon)$-fit to $F_{G,p}$ in every direction [2, 17]. (The proof is elementary for $\Omega(n^2/\epsilon^2)$ samples.)

2.1 Distance Concentration: Spherical and Spherelike gaussians

In an axis-aligned gaussian with variances $\sigma_1^2, \ldots, \sigma_n^2$, the quantity $\sum_i x_i^2/\sigma_i^2$ is the sum of n iid variables from $N(0,1)$ so this sum is tightly concentrated about its mean n. In a *spherical* gaussian, all σ_i's are the same, so even $\sum_i x_i^2$ is tightly concentrated. (These observations go back to Borel.) More generally, $E[\sum_i x_i^2] = \sum_i \sigma_i^2$. Since $\sum_i x_i^2$ is also the square of the distance from $(x_1, x_2, \ldots, x_n)$ to the origin, we call $(\sum_i \sigma_i^2)^{-1/2}$ the *radius* of the gaussian. (Often we denote it by R.) If the σ_i's are not "too different", then moment-generating methods (see Lemma 13 in the Appendix) show that that almost all of the probability mass is concentrated in a thin spherical shell about this radius; such gaussians may be thought of as *spherelike.* The sharpness of this concentration depends roughly on the quantity $\sigma_{\max}/R$, which we call the *eccentricity* [2]. Roughly speaking if $R/\sigma_{\max} = \Omega(\log n)$ then $1 - 1/\mathrm{poly}(n)$ of the probability mass lies inside a shell of thickness $o(R)$. (Consequently, a sample of size poly(n) will whp lie inside this spherical shell.)

As a point of comparison, note that if all σ_i's are equal —as in a spherical gaussian—then $R/\sigma_{\max} = \sqrt{n}$. Thus "smeared" gaussians, which have $R/\sigma_{\max} = O(\log n)$ are quite atypical. Our algorithm will deal with both spherelike and smeared gaussians.

For $\gamma < 1$, the γ-radius of a gaussian is the radius of the smallest sphere (concentric with the gaussian) such that at most γ of the probability mass lies outside. Section 3 will use $\gamma = 1/2$, and we call the 1/2-radius as *median* radius. The median radius R' is related to radius defined above, for example $R' \leq \sqrt{2}R$. Other relationships between the two follow from Lemma 13.

2.2 Likelihoods, classification, etc.

Now we formalize the learning problem. Consider a mixture of gaussians $(w_1, F_1, w_2, F_2, \ldots, w_m, F_m)$ in $\Re^n$, where the w_i's are the mixing weights. With any point $x \in \Re^n$, one can associate m numbers $(F_i(x))_{i=1,..,m}$ corresponding to the probabilities assigned to it by the various gaussians according to formula (1). For any sample $S \subseteq \Re^n$ this imposes a natural partition into m blocks: each point $x \in S$ is labelled with an integer $l(x) \in \{1,,,m\}$ indicating the distribution that assigns the highest probability to x. (Ties are broken arbitrarily.) The *likelihood* of the sample is

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

It is customary to work with the logarithm of this likelihood.

Thus one may mean several things when talking about "learning mixtures of gaussians" [16].

[1] The name "variance-covariance" refers to the fact that the i,jth term in A is is $\frac{1}{N}\sum_{k \leq N}(x_k - q)_i (x_k - q)_j$, the covariance of the ith and jth coordinates. The diagonal terms represent variances.

[2] Dasgupta defines eccentricity to be $\sigma_{\max}/\sigma_{min}$. However, our definition has a closer connection to the concentration properties of the gaussian.

1. In *Max Likelihood estimation*, we are given an arbitrary sample $S \subseteq \Re^n$, a number k, and desire the gaussian mixture with k components that maximizes the likelihood of S.

2. In the *classification problem*, we are given a sample generated from an unknown mixture of k gaussians. The goal is to find the "correct" labeling of the sampled points. (Note that this labeling would immediately yield estimates of the unknown gaussians and their mixing weights.) Viewing the unknown mixture as a "source" we may view this as the "source learning" problem.

3. In the *approximation* version of Problem 1, one desires mixtures whose likelihood is only α*-approximate* —by which we mean that the log-likelihood is within a factor α of the optimum.

4. In the *approximation* version of Problem 2, one desires labelings whose log-likelihood is within a factor α of the optimum.

5. Versions of all the above problems in which labels for some of the samplepoints may be provided as part of the input. This may happen for example if the sample is obtained from a census and the labellings of most samplepoints are removed for privacy reasons [16].

Problem 2 is a subcase of Problem 1 and Problem 4 is a subcase of Problem 3.

2.2.1 Our results

Our results may be summarized as follows.

A. PTAS for Problem 4 when the gaussians are spherelike; this appears in Section 3.2. Dasgupta (1999) dealt with the subcase of this problem when all gaussians have the same radius. However, even in this subcase our algorithm gives much better approximation of the loglikelihood. The loglikelihood with respect to the estimated gaussians is within a factor $(1 + 1/n^{1/4-\epsilon})$ of the optimum. (The corresponding factor for Dasgupta's algorithm would be something like $(1 + 1/\log n)$.)

B. An exact solution for Problem 2 when the gaussians are reasonably well-separated (but could be of arbitrary shape).

C. PTAS for Problem 3 when the gaussians are spheres of equal radii. This "PTAS" actually produces a solution whose loglikelihood is within a constant *additive* factor of the optimum.

The algorithm of Part C is combinatorial and is outlined in Section 5. We note in this context that obtaining maximum likelihood estimates is at least as hard as the clustering problem k-median (sum-of-squares version with steiner nodes), which may be NP-hard. (We haven't been able to track down an exact reference.) Indeed, our algorithm in Part C is obtained by reducing to the k-median algorithm of [3] (recent more efficient k-median algorithms would also work). We feel that this way of viewing the learning problem as an approximation problem may be fruitful in other contexts.

Although a clustering is also sought in Problems 2 and 4, we feel that these problems are simpler than general clustering because the clusters are guaranteed to have a special structure. Our algorithms in parts A, B exploit the properties of these clusters (i.e., of gaussians). For example, the algorithm in Section 3 uses distance-based clustering; this algorithm would fail badly if the sample deviated even a little from being a mixture of gaussians.

The algorithm in Section 3 also requires a known lowerbound on the smallest mixing weight $w_{\min}$. In Section 4 we describe a more complicated algorithm that does not have this restriction, but its running time is quasipolynomial in k. Both algorithms have an underlying "bootstrapping" idea, whereby we learn a little about the gaussians from a coarse examination of the data and then bootstrap from that information to find a better clustering. Both sections omit many details —including an analysis in terms of loglikelihood— for lack of space.

We note that Dasgupta (personal communication) has suggested a variant of the classification problem in which the sample comes from a "noisy" gaussian. Roughly speaking, the samples come from a mixture of sources, where each source is within distance ϵ of a gaussian. This problem may be seen as somewhere in between Problems 1 and 2. The algorithm of Section 4 can be adapted to this model when the "noise" level is small.

Note that our algorithm in Section 4 also generalizes to mixtures of other logconcave distributions, provided the components are "well-separated". The same may also be true for the algorithm in Section 3 but we have not yet proved this. Details will appear in the final version.

3. DISTANCE BASED CLUSTERING

In this section, we give an algorithm for Problem 2 —find the correct labelling of all samples— of Section 2.2. The gaussians in the mixture have to satisfy a certain separation condition. This separation condition is sharp enough that projection pursuit as in [4] may not work: random projection into $O(\log n)$ dimensions would merge the component gaussians with high probability and thus destroy the structure we are trying to learn. Consequently, the algorithm works in $\Re^n$. It uses "distance-based" clustering: samples within a certain distance r of a particular sample x are all put into one cluster. The choice of x, r is of course the crucial new element we provide.

The idea of distance-based clustering is not new, and appears in many heuristics, including [4, 5]. However, such heuristics typically founder when the gaussians are of different sizes and overlap with each other. In such cases, a samplepoint from gaussian F_i might be closer to some samplepoints of another gaussian F_j than to all the samplepoints of F_i. One crucial insight in our algorithm is that this is unlikely to happen if we look at the smallest gaussian in the mixture; hence we should use clustering to first identify this gaussian, and then iterate to find the remaining gaussians.

Suppose $F_1, F_2, \ldots F_k$ are the gaussians in the mixture with centers $p_1, p_2, \ldots p_k$ respectively in $\Re^n$ and suppose the maximum variance of F_i in any direction is $\sigma_{i,max}$ and the median radius of F_i is R_i (i.e., the probability mass in a ball of radius R_i is $F_i(B(p_i, R_i)) = 1/2$). In this section we will often refer to R_i as just radius. Suppose also that the mixing weights are $w_1, w_2, \ldots w_k$ respectively and the algorithm is given a w_{min} such that each $w_i \geq w_{min}$.

The algorithm requires a sample S of i.i.d. samples from the mixture distribution, where $|S| = 10^7 n^2 k^2 \log(kn^2)/w_{min}^4$. The running time is quadratic in $|S|$ although this may be improveable.

Note that S can be partitioned into $S = S_1 \cup S_2 \cup \ldots S_k$ where S_i is the set of samples picked according to F_i. The algorithm seeks to determine this partition. Note that this problem may be ill-posed in general, in the sense that if a pair of gaussians are very close together, then the maximum-likelihood labeling (as defined in Section 2.2) is incorrect whp (the reader may wish to verify this using two spherical gaussians of the same radius). However, if the mixture is t-separated for some $t \in O(\log(|S|/\delta))$ (where $\delta > 0$ is the failure probability of the algorithm), then whp the maximum likelihood solution correctly labels all samplepoints with the gaussian they came from.

Definition 1. *For any $t > 0$, we say that the mixture is* t-separated *if every gaussian pair F_i, F_j satisfy:*

$$|p_i - p_j|^2 \geq (R_i^2 - R_j^2)^- + 500t(R_i + R_j)(\sigma_{i,max} + \sigma_{j,max}) + 100t^2(\sigma_{i,max}^2 + \sigma_{j,max}^2). \tag{2}$$

where $(R_i^2 - R_j^2)^- = R_i^2 - R_j^2$ if $R_i^2 \leq R_j^2$ and 0 otherwise.

Note that if the radii are unequal then $(R_i^2 - R_j^2)^-$ is negative, which allows one gaussian to be possibly inside the other with even the same center.

To gain some intuition on our separation requirement, consider when all gaussians are spherical, implying $\sigma_{i,max} \approx R_i/\sqrt{n}$ (remember that we are talking about the median radius). Then the separation required is:

$$|p_i - p_j|^2 \geq (R_i^2 - R_j^2)^- + c\log(|S|/\delta)\frac{(R_i + R_j)^2}{\sqrt{n}}. \tag{3}$$

Thus if $R_i \approx R_j$, then the distance between the centers is merely $O(\log(|S|/\delta)) \cdot \sigma_{\max}$. It can be shown that this separation condition is essentially necessary (within a factor of $\sqrt{\log n}$ in the second term) for the solution to the Classification Problem to be unique with high probability (a proof will appear in the final version). Thus the separation condition is essentially optimum for spherical gaussians. The condition is not optimum for general gaussians, and designing a learning algorithm with a tighter condition is open; an idea different from distance-based clustering is needed (see Section 6).

Now we describe why distance-based clustering works when this separation condition holds. We will show below that for every pair of samples in S_i, the distance between the samples lies in an interval I_i of the real line and for every $i \neq j$, the distance between a point in S_i and one in S_j lies in an interval $I_{i,j}$, so that at least one of I_i or I_j is disjoint from $I_{i,j}$. Furthermore, there is a "guaranteed gap" between I_{ij} and this interval, allowing them to be distinguished whp. All this will follow from "distance concentration" results stating that the random variable which is the distance between a sample in S_i and one in S_j is tightly concentrated about its median value.

Some distance concentration results —at least for spherical or spherelike gaussians— were known prior to our work, showing a sharp concentration around the mean/median. In Section 4 we will also use the traditional moment-generating function approach to prove such a concentration about the median/mean value — see Lemma 13.

However, for the current algorithm we also need concentration around values that are quite far from the mean/median. For example, we will need the following result: If we have a ball containing at least 3/4 of the probability mass according to a Gaussian, then for any positive λ, increasing the radius of the ball by an additive $\lambda\sigma_{max}$ captures all but $\frac{3e^{-\lambda}}{4}$ of the mass. Such concentration results (around values other than the median/mean) are not in general provable by the traditional moment-generating function approach. We introduce a new approach in this context: Isoperimetric inequalities (see Theorem 1). These do not always prove the tightest possible concentration results (the concentration about mean is stronger in Lemma 13, which uses traditional techniques), but are more general. They also seem to apply to a more general class of distributions than just Gaussians (we do not use this generality in this draft but plan to in the final version).

Now we give an overview of the algorithm. Let F_i be the gaussian of smallest radius. Suppose we try to identify S_i by growing a ball from any $x \in S_i$. Since we have proved the existence of the guaranteed gap, it may appear that we are essentially done. Indeed, this is the case for spherical Gaussians, as explained in Section 3.2. Distances of points in S_i to x are provably tightly concentrated so that the first observed gap (in distances to x) signals the end of S_i. We may peel off all samples up to this distance, fit a gaussian to them, and continue to recover other gaussians.

But for general Gaussians, there is a "smear" possible in the pairwise distances in S_i. So the first observed gap in distances to some sample x does not necessarily signify the end of S_i. However, even here, we can say something. Suppose we are growing a ball centered at a point x. If we observe a gap of $\Omega(\sigma_{i,max})$, (namely, an annulus of thickness $\Omega(\sigma_{i,max})$ devoid of samples), then indeed by distance concentration results, one may show that the probability mass left outside is small. In other words, incrementing the radius by $\Omega(\sigma_{max})$ after seeing the first gap captures all of the S_i whp. But this still leaves one problem - we do not know σ_{max}. We tackle this by bootstrapping. We show that: (i) if we have any fraction f of the samples in S_i, then we may estimate $\sigma_{i,max}$ to a factor of $O(1/f^2)$ with high probability. We use this to get a rough estimate β of $\sigma_{i,max}$. Using β, we increment the radius in steps which are guaranteed to be less than $\sigma_{i,max}$ (which ensures that we do not step over the "gap" into another S_j) until we observe a gap; by then, we have provably picked up *most* of S_i. Now we use this to better estimate $\sigma_{i,max}$ (a kind of bootstrapping again) and then incrementing the radius again by $\Omega(\sigma_{i,max})$, we capture all of S_i. (The guaranteed gap ensures that we have not picked up any of the other S_j.)

3.1 Concentration results using isoperimetric inequalities

Suppose we have some probability density F in $\Re^n$ and a point x in space. For proving distance concentration results, we would like to measure the rate of growth / decline of $F(B(x,r))$ as a function of r. This will be provided by the Isoperimetric inequality (see Corollary 2).

Theorem 1. *[11] Let $F(x) = e^{-x^T A^{-1} x} g(x)$ be a function defined on $\mathbf{R}^n$, where A is a positive definite matrix* whose largest eigenvalue is σ_{max}^2 and $g(x)$ is any positive real valued log-concave function. Then for any measurable set K in $\mathbf{R}^n$, whose surface ∂K is also measurable, we have

$$\int_{\partial K} F(x)dx \geq \frac{2}{\sqrt{\pi}} \frac{1}{\sigma_{max}} \min\left(\int_K F(x)dx, \int_{\mathbf{R}^n \setminus K} F(x)dx\right).$$

COROLLARY 2. *We borrow notation from Theorem 1 and assume that $F(\Re^n) = 1$.*

(i) If a ball $B(x,r)$ has $F(B(x,r)) \leq 1/2$, then

$$\frac{d(F(B(x,r)))}{dr} \geq \frac{2F(B(x,r))}{\sqrt{\pi}\sigma_{max}}.$$

(ii) If a ball $B(x,r)$ has $F(B(x,r)) \geq 1/2$, then

$$\frac{d(F(B(x,r)))}{dr} \geq \frac{2(1 - F(B(x,r)))}{\sqrt{\pi}\sigma_{max}}.$$

Remark: Noting that

$$\frac{d\ln(F(B(x,r)))}{dr} = \frac{1}{F(B(x,r))} \frac{dF(B(x,r))}{dr},$$

the corollary says that $\ln(F(B(x,r))$ grows at a rate of $\Omega(1/\sigma_{max})$ until $F(B(x,r))$ is 1/2, and then $\ln(1-F(B(x,r)))$ declines at a rate of at a rate of $\Omega(1/\sigma_{max})$. Intuitively, it is easy to see that this would lead to distance concentration results since once we increase (decrease) r by $O(\sigma_{max})$ from its median value, the mass outside $B(x,r)$ (inside $B(x,r)$) is small. The first lemma below (Lemma 3) is derived exactly on these lines; the subsequent three Lemmas 4,5,6 discuss the distances between different samples from the same and from different Gaussians.

LEMMA 3. *Suppose F is a general Gaussian in $\mathbf{R}^n$ with maximum variance in any direction σ, median radius R and center p. Then for any $t > 0$, we have*

$$F(\{x : R - t\sigma \leq |x-p| \leq R + t\sigma\}) \geq 1 - e^{-t}.$$

LEMMA 4. *Let F, p, R, σ be as in Lemma 3 and suppose z is any point in space. Let $t \geq 1$. If x is picked according to F, we have that with probability at least $1 - 2e^{-t}$,*

$$\begin{aligned}(R + t\sigma)^2 + |z-p|^2 + 2\sqrt{t}|z-p|\sigma \qquad &\geq |x-z|^2 \\ \geq ((R-t\sigma)^+)^2 + |z-p|^2 - 2\sqrt{t}|z-p|\sigma &\end{aligned} \tag{4}$$

where $(R - t\sigma)^+$ is $R - t\sigma$ if this quantity is positive and 0 otherwise.

PROOF. We have

$$\begin{aligned}|x-z|^2 &= ((x-p) + (p-z)) \cdot ((x-p) + (p-z)) \\ &= |x-p|^2 + |p-z|^2 + 2(x-p)\cdot(p-z).\end{aligned} \tag{5}$$

Now $2(x-p)\cdot(p-z)$ is a normal random variable with mean 0 and variance at most $4|p-z|^2\sigma^2$, so the probability that $|2(x-p)\cdot(p-z)|$ is greater than $2\sqrt{t}|z-p|\sigma$ is at most e^{-t}. From lemma (3), we have that $R - t\sigma \leq |x-p| \leq R + t\sigma$ with probability at least $1-e^{-t}$. Combining these two facts, the current lemma follows. $\square$

LEMMA 5. *Suppose F, p, R, σ as in Lemma 3. Suppose x, y are indendent samples each picked according to F. Then for any $t \geq 1$, with probability at least $1 - 3e^{-t}$, we have*

$$2R^2 - 8t\sigma R \leq |x-y|^2 \leq 2(R + 2t\sigma)^2.$$

LEMMA 6. *Suppose F_i, F_j are two Gaussians in $\mathbf{R}^n$ with centers p_i, p_j, maximum variances $\sigma_{i,max}, \sigma_{j,max}$ and median radii R_i, R_j respectively. Let $t \geq 1$. Suppose F_i, F_j are $t-$ separated. If x is a random sample picked according to F_i and y is picked independently according to F_j, then with probability at least $1 - 6e^{-t}$, we have*

$$\begin{aligned}|x-y|^2 \geq 2\min(R_i^2, R_j^2) &+ 60t(\sigma_{i,max} + \sigma_{j,max})(R_i + R_j) \\ &+ 30t^2(\sigma_{i,max}^2 + \sigma_{j,max}^2)\end{aligned} \tag{6}$$

3.2 Warmup: case of Spherical Gaussians

As a consequence of our concentration results we present a trivial algorithm for the case when all the F_i are spherical. (It appears that even this simple case was open; the algorithm of Dasgupta [4] applies only when the gaussians have the same radius.) In this case, $\sigma_{i,max} \approx R_i/\sqrt{n}$, where the error is small enough that our calculations below are valid. Choosing $t = \Omega(\log(|S|/\delta))$ as before, it is easy to see that from the distance concentration results that whp,

$$|x-y|^2 \in 2R_i^2(1 \pm \frac{3t}{\sqrt{n}}) \qquad \forall x, y \in S_i \qquad \forall i \tag{7}$$

and by appropriately choosing the constant in the definition of t-separation we can also ensure that whp there is a positive constant $c' > 6$ such that

$$\begin{aligned}|x-y|^2 \geq 2\min(R_i^2, R_j^2) &+ \frac{c't(R_i+R_j)^2}{\sqrt{n}} \\ \forall x \in S_i \qquad \forall y &\in S_j, \quad \forall i \neq j.\end{aligned} \tag{8}$$

Now we describe a trivial algorithm to label the samples correctly. Let $x_1, x_2, \ldots, x_M$ be the samples. A *similarity graph* for the sample is an undirected graph on $|S|$ nodes where $\{i,j\}$ is an edge iff for each sample point x_l (where $l \neq i, j$) the distances $|x_i - x_l|, |x_j - x_l|$ are within a multiplicative factor $(1 + 6t/\sqrt{n})$ of each other. Consider the samples from the smallest gaussian. From (7) and (8) it follows that there is an edge between each pair, and no edge to a sample from another gaussian. In other words we have:

Claim 1: *Whp (over the choice of the samples), nodes corresponging to the smallest gaussian form an isolated clique.*

In fact, a little thought shows that

Claim 2: *Every isolated clique corresponds to samples from a single gaussian.*

Thus one can look for isolated cliques in the similarity graph, and try to fit a spherical gaussian to each. The one corresponding to smallest sphere is identified, and removed from the sample. Then we iterate to find the remaining $k-1$ gaussians.

The above algorithm generalizes to spherelike gaussians, thus also providing a simple way to derive the main result of [4]. We project to a random subspace of dimension $O(\log n)$: with high probability, spherelike gaussians become 2-circuloids and are still reasonably well-separated (see Section 4). It is easy to see that all the ideas of this section apply to 2-circuloids.

3.3 The general case

Now we consider the case when the gaussians may not be spherical or even spherelike. Let $\delta > 0$ be the probability of failure allowed. In what follows, we choose

$$t = \frac{100\log|S|}{\delta}.$$

The Algorithm
Initialization : $T \leftarrow S$.

1. Let $\alpha > 0$ be the smallest value such that a ball $B(x, \alpha)$ of radius α centered at some point in T has at least $3w_{min}|S|/4$ points from T. [This will identify a Gaussian F_i with approximately the least radius.]

2. Find the maximum variance of the set $Q = B(x, \alpha) \cap T$ in any direction. I.e., find
$$\beta = \max_{w:|w|=1} \frac{1}{|Q|} \sum_{y \in Q} (w \cdot y - w \cdot (\frac{1}{|Q|} \sum_{z \in Q} z))^2.$$
[This β is our first estimate of σ_{max}. Note that computing β is an eigenvalue computation, and an approximate eigenvalue suffices.]

3. Let $\nu = \sqrt{\frac{w_{min}\beta}{8}}$. [We will later show that $\nu \leq \sigma_{max}$; so increasing the radius in steps of ν ensures that we do not miss the "gap" between the S_i that x belongs to and the others.] Find least positive integer s such that [we will later prove that such a s exists].
$$B(x, \alpha + s\nu) \cap T = B(x, \alpha + (s-1)\nu) \cap T.$$

4. Let $Q' = B(x, \alpha + s\nu) \cap T$. As in Step 3, find the maximum variance β' of the set Q' in any direction. [We will prove that this β' gives a better estimate of σ_{max} than β.]

5. Remove $B(x, \alpha + s\nu + 3\sqrt{\beta'}(\log|S| - \log\delta + 1)) \cap T$ from T. [We will show that the set so removed is precisely one of the S_i. Ball $B(x, \alpha + s\nu)$ will be shown to contain all but $w_{min}/(10w_i)$ of the mass of the Gaussian F_i we are dealing with; the bigger radius of $B(x, \alpha + s\nu + 3\sqrt{\beta'}(\log|S| - \log\delta + 1))$ will be shown to include all but $\delta/(10|S|^2)$ of the mass of F_i. This will follow using Isoperimetry. Then we may argue that whp all of S_i is now inside this ball. An easier argument shows that none of the other S_j interesect this ball.]

6. Repeat with until T is empty.

Now we sketch a proof why this works as claimed; proofs of the Lemmas will appear in the final paper. Let $\delta > 0$ be the probability of failure allowed. We choose
$$t = \frac{100 \log |S|}{\delta}.$$

We will now show using the distance concentration results that several desirable events — described below in (9),(10),(11), (12), (13), (14) and (15) happen, each with probability at least $1 - \frac{\delta}{10}$. **We will assume from now on that the conditions hold after allowing for the failure probability of at most** $7\delta/10$. The bottomline is that the sample is very likely to represent the mixture accurately: the component gaussians are represented essentially in proportion to their mixing weights; the number of samples in every sphere and halfspace is about right; and so forth.

First, since $|S_i|$ can be viewed as the sum of $|S|$ Bernoulli independent 0-1 random variables, where each is 1 with probability w_i, we have (using standard results - eg. Hoeffding) that with probability at least $1 - \delta/10$,
$$|S_i| \geq (0.9)w_i|S| \forall i. \tag{9}$$

For each $i, 1 \leq i \leq k$ and each $x \in S_i$, let $\eta(x)$ be the least positive real number such that
$$F_i(B(x, \eta(x))) \geq 1 - \frac{\delta}{10|S|^2}.$$
Then we have with probability at least $1 - \frac{\delta}{10}$,
$$\forall i, 1 \leq i \leq k, \forall x \in S_i, \qquad S_i \subseteq B(x, \eta(x)) = \emptyset. \tag{10}$$
We have from Lemma 5 that with probability at least $1 - \delta/10$, the following is true for each $i, 1 \leq i \leq k, \forall x, y \in S_i$:
$$2R_i^2 - 8t\sigma_{i,max}R_i \quad \leq |x-y|^2 \quad \leq 2(R_i + 2t\sigma_{i,max})^2 \tag{11}$$
Further, from Lemma 6, we have that with probability at least $1 - \frac{\delta}{10}$,
$$\forall i, j \leq k, i \neq j, \forall x \in S_i, \forall y \in S_j,$$
$$|x-y|^2 \geq 2\min(R_i^2, R_j^2) + 60t(R_i + R_j)(\sigma_{i,max} + \sigma_{j,max}) + 30t^2(\sigma_{i,max}^2 + \sigma_{j,max}^2). \tag{12}$$

A (spherical) annulus A is a set of the form $B(x, R_1) \setminus B(x, R_2)$. From a VC dimension argument —Theorem 16— it follows that with probability at least $1 - \delta/4$, **every** annulus in space gets about the right number of points. More precisely, using (9), with probability at least $1 - \delta/10$, the following holds for a small enough $w_0 \leq w_{\min}$:
$$\forall i, \qquad \forall \text{ annulii } A \text{ anywhere in space}$$
$$\left| \frac{|S_i \cap A|}{|S_i|} - F_i(A) \right| \leq \frac{w_0^2}{80w_i} \tag{13}$$

Also the VC dimension of half spaces is $O(n)$ and it follows from VC dimension arguments that with probability at least $1 - \delta/10$, we have
$$\forall i \qquad \forall \text{ half spaces } H$$
$$||S_i \cap H| - |S_i|F_i(H)| \leq |S_i|/80 \tag{14}$$

From Lemma 11, it follows that with probability at least $1 - \delta/10$, we have
$$\forall \text{ unit length vectors } w, \qquad \forall i,$$
$$\frac{1}{|S_i|} \sum_{x \in S_i} (w \cdot (x - p_i))^2 \leq 2\sigma_{i,max}^2. \tag{15}$$

LEMMA 7. *Each execution of steps 1-5 removes precisely one of the S_i*

The lemma will be proved by induction on the number of executions of the loop. Suppose we have finished $l-1$ executions and are starting on the l th .

Let P be the set of j such that S_j has not yet been removed. (By the inductive assumption at the start of the loop, T is the union of S_j, $j \in P$.)

LEMMA 8. *Suppose $x \in S$ is the center of the ball $B(x, \alpha)$ found in the l th execution of step (1) of the algorithm and suppose x belongs to S_i (i unknown to us). Then, $B(x, \alpha) \cap S \subseteq S_i$ and furthermore, for each samplepoint y of another gaussian F_j we have*
$$|x-y|^2 \geq 2R_i^2 + 50t(\sigma_{i,max} + \sigma_{j,max})(R_i + R_j) + 20t^2(\sigma_{i,max}^2 + \sigma_{j,max}^2) \tag{16}$$

LEMMA 9. *The β, Q computed in step 2 of the algorithm satisfy*

$$\frac{2|S_i|}{|Q|}\sigma_{i,max}^2 \geq \beta \geq \frac{|Q|^2}{4|S_i|^2}\sigma_{i,max}^2.$$

Now we can show that β is a rough approximation to $\sigma_{i,max}$ (recall that w_0 is somewhat smaller than $w_{\min}$).

COROLLARY 10. *The β computed in step 2) of the algorithm satisfies :*

$$\frac{4}{w_0}\sigma_{i,max}^2 \geq \beta \geq \frac{1}{8}w_0^2\sigma_{i,max}^2.$$

Now since we are increasing the radius in steps of ν which is at most $\sigma_{i,max}$ (Corollary 10), and Lemma 8 implies that samples from other gaussians are "far" from x compared to the samples of F_i, it follows that there is an annulus around x which contains no samples. Thus we have proved the existence of the desired s in step 3), and also that

$$B(x, \alpha + s\nu) \cap S \subseteq S_i.$$

The trouble of course is that such a gap may exist even inside of S_i, so $B(x, \alpha + s\nu)$ may not contain all of S_i. To complete the induction we have to argue that the final two steps will succeed in identifying *every* point of S_i. For $\gamma > 0$, let

$$g(\gamma) = F_i(B(x, \gamma)).$$

From $B(x, \alpha + s\nu) \cap T = B(x, \alpha + (s-1)\nu) \cap T$, (see step 3 of the algorithm), we get using (13) that

$$g(\alpha + s\nu) - g(\alpha + (s-1)\nu) \leq \frac{w_0^2}{80w_i}.$$

So there exists a $\gamma' \in [(s-1)\nu + \alpha \ \ s\nu + \alpha]$ with

$$\left(\frac{dg(\gamma)}{d\gamma}\right)_{\gamma=\gamma'} \leq \frac{w_0^2}{80w_i\nu} \leq \frac{w_0}{20w_i\sigma_{i,max}}.$$

Now, we also have $g(\alpha+s\nu) \geq w_0/(2w_i)$. Thus Isoperimetry (Theorem 1) implies that

$$g(\alpha + s\nu) \geq 1 - \frac{w_0}{10w_i}.$$

This implies from (13) that $|Q'| \geq 0.8|S_i|$ (note that Q' is found in step 4)). Thus from Lemma 9, (noting that the proof of works for any subset of S_i), we get that β' is a fairly good approximation to $\sigma_{i,max}$:

$$1.6\sigma_{i,max}^2 \geq \beta' \geq 0.16\sigma_{i,max}^2. \tag{17}$$

From the definition of s in step (3) of the algorithm, it follows that there is some $y \in S_i$ with $|x - y| \geq \alpha + (s-1)\nu$. So, from (11), we have $\alpha + (s-1)\nu \leq \sqrt{2}(R_i + 2t\sigma_{i,max})$. So, we have

$$\alpha + s\nu + 3\sqrt{\beta'}(\log\frac{|S|}{\delta} + 1) \leq \sqrt{2}(R_i + 2t\sigma_{i,max}) + \sigma_{i,max}$$
$$+ 4\sigma_{i,max}(\log\frac{|S|}{\delta} + 1). \tag{18}$$

Thus from (16), no point of $S_j,\ j \in P \setminus \{i\}$ is contained in $B(x, \alpha + s\nu + 3\sqrt{\beta'}(\log\frac{|S|}{\delta} + 1))$. So the set removed from T in step 5) is a subset of S_i.

Finally, using $g(\alpha + s\nu) \geq 9/10$, and Isoperimetry (Theorem 1), we see that

$$g(\alpha + s\nu + 3\sqrt{\beta'}(\log\frac{|S|}{\delta} + 1)) \geq 1 - \frac{\delta}{10|S|^2},$$

whence by (10), all of S_i is in $B(x, \alpha+s\nu+3\sqrt{\beta'}(\log\frac{|S|}{\delta}+1))$. This completes the inductive proof of correctness.

Now we prove a lemma that was used above when we estimated $\sigma_{\max}$.

LEMMA 11. *Suppose F is a (general) Gaussian in $\Re^n$ If L is a set of independent identically distriubuted samples, each distributed according to F, then with probability at least $1 - \frac{\delta}{10k}$, we have (with $\epsilon = 10n(\sqrt{\log n} + \sqrt{\log(1/\delta)})/\sqrt{|L|}$), every vector w satisfies*

$$E_F((w.(x - E_F(x))^2)(1-\epsilon) \leq E_S((w.(x - E_F(x))^2)$$
$$\leq E_F((w.(x - E_F(x))^2)(1+\epsilon). \tag{19}$$

PROOF. Also wlg assume that $E_F(x)$ is the origin. Suppose Q is the square root of the inverse of the variance-covariance matrix of F. We wish to prove for all vectors w :

$$E_F((w.x)^2)(1-\epsilon) \leq E_S((w.x)^2) \leq E_F((w.x)^2)(1+\epsilon).$$

Putting $Q^{-1}w = u$ (noting that Q is nonsingular and symmetric), this is equivalent to saying for all vectors u,

$$E_F((u.(Qx))^2)(1-\epsilon) \leq E_S((u.(Qx))^2) \leq E_F((u.(Qx))^2)(1+\epsilon).$$

But Qx is a random sample drawn according to the standard normal, so it suffices to prove the statement for the standard normal.

To prove it for the standard normal, we proceed as follows. First for each i, we have (noting that $E_F(x_i^4) = O(1)$),

$$\text{Prob}\left(|E_S(|x_i|^2) - 1| \leq s\right) \geq 1 - e^{-|L|s^2/4}.$$

Now consider a pair $i, j \in \{1, 2, \ldots n\}$, where $i \neq j$. The random variable x_ix_j has mean 0 and variance $O(1)$. $E_S(x_ix_j)$ being the average of N i.i.d. samples (each not bounded, but we may use the normal property to argue concentration - to be elaborated in final version) concentrated about its mean :

$$\text{Prob}\left(E_S|x_ix_j| \leq s\right) \geq 1 - e^{-|L|s^2/100}.$$

Putting $s = 10\frac{\sqrt{\log n}}{\sqrt{|L|}}$, we see that all these $O(n^2)$ upper bounds hold simultaneously with probability at least $1 - \delta/n^8$.

Thus we have that the "moment" of inertia matrix M of S whose i,j entry is $E_S(x_ix_j)$ has entries between $1 - \frac{1}{2}\epsilon$ and $1+\frac{1}{2}\epsilon$ on its diagonal and the sum of the absolute values of the entries in each row is at most $\epsilon/2$. Thus by standard Linear Algebra (basically arguments based on the largest absolute value entry of any eigenvector), we have that the eigenvalues of M are between $1-\epsilon$ and $1+\epsilon$ proving what we want. $\square$

3.4 More general distributions

We believe (but have not yet proved) that the algorithm above can be extended to many other log-concave densities besides the general Gaussian. The basic ingredient that we use throughout is that in every direction, the Gaussian declines at a certain rate (i.e. along any direction, after a distance of $\lambda\sigma_{max}$ (where λ is any positive real), it has fallen by an exponential factor in λ). Under this condition, the Isoperimetric inequality and distance concentration results should all extend to give results similar to the Gaussian case. This will be explored in the final version.

4. CLASSIFICATION WITH NEAR-OPTIMUM SAMPLE COMPLEXITY

The algorithm of the previous section has high sample complexity, specifically, it grows as $O(1/w_{\min}^4)$. Now we give a different polynomial-time algorithm that has a sample complexity $O(n\text{poly}(k)\log n/w_{\min})$. If k is at most $2^{\log^{0.1} n}$ or so —which we need also to ensure that the running time is $\text{poly}(n)$ — then this sample complexity is $O(n^{1+o(1)}/w_{\min})$, which is close to the information-theoretic lowerbound of $\Omega(n \log n/w_{\min})$. (This lowerbound follows from the need to get $\Omega(n \log n)$ samples from the gaussian with lowest mixing weight.)

The algorithm has another important advantage: it needs no lowerbound on $w_{\min}$. It peels off gaussians in roughly decreasing order by mixing weight. For any desired $\delta > 0$, this process can be stopped once some $1-\delta$ fraction of the samples have been "described." Then the sample complexity is only $O(n\text{poly}(k)\log n/\delta)$. The running time is roughly quadratic in the sample complexity.

4.1 The main ideas

Like Dasgupta's algorithm, our algorithm first projects datapoints into a randomly-chosen subspace of dimension $m = \log^2 k$. Projections of gaussians are gaussians, so our learning problem is unchanged except for a vast reduction in the number of dimensions. We describe how to learn in m dimensions, and indicate later how to "lift" to $\Re^n$.

How can one learn the gaussian mixture in $\Re^m$? We make two observations. First, a gaussian in $\Re^m$ can be learned from $O(m \log m)$ samples. Second, at least one gaussian has mixing weight $\geq 1/k$, so a sample from the mixture of size $\Omega(km \log m)$ already contains $\Omega(m \log m)$ samples from that gaussian.

Thus we temporarily restrict attention to a subsample of size $K = \text{poly}(km)$, and enumerate *all* possible subsets of size $M = O(m \log m)$, and for each of them compute the maximum likelihood fit gaussian. Since $\binom{K}{M} = k^{O(m \log m)} = k^{\log^2 k \log\log k} = n^{o(1)}$, all of this can be done in polynomial time.

Now we have $n^{o(1)}$ candidate gaussians in $\Re^m$, at least one of which is (whp) the projection of a gaussian in the mixture. To identify this gaussian, we test these candidates: the one whose density distribution closely matches the distribution found among the sampled points and which explains the largest fraction of the data is retained at the end. This test takes around $\text{poly}(k)m^{O(m)}$ time per candidate. Having identified one gaussian, we can recurse with k reduced to $k-1$ and find the remaining gaussians.

In order for the algorithm to work correctly, we need a guarantee that gaussians remain distinct under projection, which requires them to be somewhat separated to begin with.

As before, the *radius* R of the gaussian is $\sqrt{\sum_i \sigma_i^2}$. For $\gamma < 1$, the γ-radius of a gaussian is the radius of the smallest sphere (concentric with the gaussian) such that at most γ of the probability mass lies outside. Below, a gaussian is *spherelike* if $R/\sigma_{\max} > (\log k)^3 \log 1/\gamma$, where $1/\gamma$ is the sample complexity of the algorithm, $O(\text{poly}(k)n \log n/w_{\min})$. Under random projection into a space of dimension $O(\log 1/\gamma)$, such a gaussian turns into a 2-circuloid [3] with probability at least $1 - 1/k^2$. (See Theorem 15.) Note that for a spherelike gaussian, the γ-radius is essentially the same as its radius (see Lemma 13).

DEFINITION 2. *Let $c, \gamma > 0$. A mixture of gaussians is said to be* (c, γ)-separated *if the following is true for each pair. If their γ-radii are r_1, r_2 respectively then they satisfy: (a) If both are sphere-like then either $r_1/r_2 \geq 1 + c$ or the intercenter distance is at least $c \cdot \max\{r_1, r_2\}$. (b) If the first is spherelike and the second is not, then the intercenter distance is at least $c \cdot r_1 + r_2$. (c) If neither is spherelike then the the intercenter distance is at least $r_1 + r_2$.*

Thus the definition allows gaussians to overlap (or to lie inside one another provided the larger one is spherelike) but in a way that their high probability regions remain distinct. The algorithm will require the mixture to be (c, γ)-separated, where $c > 0$ is any contant independent of n.

We will use the following facts about random projections. Let the mixture be projected to a random subspace of dimension m between $O(\log n/c^2)$ and $\Omega(\log k/c^2)$. Under such projection, a spherelike gaussian becomes whp a 2-circuloid. Furthermore, the well-known lemma of Johnson and Lindenstrauss [10] implies that for any set of $2^{\Omega(m)}$ points, the ratio of their distance changes by at most a factor $1 + c/2$ under projection. This has two implications. First, all intercenter distances scale down in proportion, since there are only $\binom{k}{2}$ of them. Second, if we consider the δ-radius of the gaussians, where $\delta \geq 2^{-\Omega(m)}$ then those radii also scale down in proportion. In other words the following is true.

Let $F_1, F_2, \ldots, F_k$ be the original gaussians and $E_1, E_2, \ldots, E_k$ be their projections. Let $w_1, w_2, \ldots, w_k$ be their mixing weights.

LEMMA 12. *Let m be between $O((\log 1/\gamma)/c^2)$ and $\Omega(\log k/c^2)$. If $F_1, F_2, \ldots, F_k$ are (c, γ)-separated then $E_1, E_2, \ldots, E_k$ are $(c/2, 2^{-\Omega(m)})$-separated with probability at least $1 - 1/k^2$.*

Below, we first describe how to learn the projection in $O(\log^2 k)$ dimensions. This algorithm shows how to obtain a "pure" sample of size $\text{poly}(\log n)$ from a single component gaussian, whereupon one can "lift" the reconstruction to $\Re^n$ (See Section *4.2.1*.)

4.2 Some of the details

For ease of exposition we assume in this section that $w_{\min} \geq 1/k^{10}$, and show how to reconstruct the mixture after it has been projected to a random subspace of dimension $m = \Omega(\log^2 k)$. As mentioned, the algorithm uses a random subsample of size k^{20} and enumerates all subsets of size $O(m \log m)$ to compute $k^{O(m \log m)}$ candidate gaussians. Each candidate is tested using the FILTER procedure below.

Below, δ is an error parameter obtained by the Lemma on VC-dimension arguments Theorem 16. It will be at most $1/k^{10}$

[3] A t-circuloid, where $t \geq 1$, is a gaussian with $\sigma_{\max}/\sigma_{\min} \leq t$. It behaves very much like a spherical gaussian.

FILTER
Given: Candidate gaussian G, sample S_1 of size k^{20}, error parameter δ.
Let G have radius R and $\mathrm{Ann}(G)$ be the annulus of G in Mahalonobis norm that contains all but $1/k^{30}$ of the probability mass. Associate with G a set $S_G \subseteq S_1$ that initially is the set of samplepoints lying in $\mathrm{Ann}(G)$.

1. If G is a 2-circuloid, discard every pair of points in S_G which are separated by distance at most $\sqrt{2}R - 6\sqrt{20 \log k}R/\sqrt{m}$.
2. If G is not a 2-circuloid, replace G by the gaussian fitted to S_G using the variance-covariance matrix. Enumerate all spheres intersecting $\mathrm{Ann}(G)$ that have radius at most mR and whose radius and center involve numbers that are integer multiples of R/k^6. (At most $(mk^6)^{m+1}$ such spheres.) A light sphere is one that has probability mass (under G) at least $1/k^5$ but that less than $\gamma|S_1|$ samples. If such a sphere exists, make S_G empty.

At the end the algorithm picks the candidate that has the largest S_G (i.e., explains the most samplepoints) and takes a random sample of k points from S_G. The claim is that these k points form a pure sample from some gaussian in the mixture.

The proof of this claim goes as follows. First, the standard Hoeffding argument shows that if a gaussian F_i has mixing weight $w_i \geq 1/2k$ then it contributes about w_i fraction of both S and S_1. Thus the exhaustive enumeration of candidates will "discover" this gaussian; in other words, one of the candidate gaussians, say G will be very close to the projection of F_i in $\Re^m$.

Furthermore, one can show using standard VC dimension arguments —see Theorem 16 and the similar arguments in Section 3— that the samples of F_i are quite representative of F_i, in the following sense: for every sphere in $\Re^m$, the number of points of F_i in it is proportional (upto some error term) to the probability mass of this sphere with respect to F_i.

Now we argue that such a candidate G survives the FILTER procedure and S_G at the end consists of almost all samplepoints of F_i with possibly some small amount of "impurities," i.e., samples from other gaussians. (We will deal with the effect of these impurities later on.)

Let us first argue this for F_i that is not spherelike. Definition 2 says that every other gaussian has negligible probability mass inside the γ-ball of F_i. The same continues to be true whp after projection in $\Re^m$. Thus by the VC dimension argument, every other gaussian has a negligible number of samples in this γ ball (this negligible number corresponds to the error parameter in the VC dimension argument).

Now assume F_i is spherelike. Then G is a 2-circuloid. One can show using Definition 2 and Corollary 14 that every other gaussian that "pollutes" $\mathrm{Ann}(G)$ by contributing a significant number of samples to it is somewhat smaller in radius. So the distance concentration results imply that Step 1 of FILTER would remove all these polluting points. (Note that a priori it seems possible that filter may remove pairs of points (x, y) where x comes from F_i and y from another gaussian, but distance concentration can be used to prove that this is unlikely.)

This brings us to the difficult part of the proof. We have to show that "fake" candidates —those composed of points from two or more "true" gaussians— end up with small S_G's at the end of FILTER and hence are not picked. This has two cases: either G is a 2-circuloid or it is not. The first case is proved easily using concentration results and results about when two spheres can have significant overlap. The second case uses the Leindler-Prekopa inequality, which says that the the probability mass functional for logconcave distributions is logconcave: if A, B are two convex bodies, then the probability mass of an inbetween body $\lambda A + (1 - \lambda)B$ is the geometric mean of the probability masses of A and B (Fact 18). In our application, the gaussian distribution is the "fake" candidate G, and A, B are balls containing "true" gaussians that contribute points to $\mathrm{Ann}(G)$. We show that the separation condition (Definition 2) implies that some inbetween ball $\lambda A + (1 - \lambda)B$ has very few samplepoints. However, Leindler-Prekopa provides a good lowerbound its probability mass under G. Thus this inbetween ball has fewer samplepoints than one would expect it to, and this ball is light (in the sense defined in the FILTER procedure). Thus this ball is a "proof" that G is a fake gaussian, and procedure correctly rejects it.

Finally, note that the analysis of FILTER does *not* show that the winning candidate G (especially if it happens to be a 2-circuloid) has a sample set S_G composed *only* of points from a single gaussian. Instead, it shows that S_G is almost pure: it has size at least $|S_1|/k$, and consists of almost all the points of some E_i with perhaps δ fraction of the points of E_i getting replaced by points of other E_p's. This issue of impurities is addressed by our final step which takes a random subsample of size k from S_G. Whp this will be a pure sample from E_i, as can be seen by realizing that a sample of k points from S_G is different from a sample from S'_G ($=$ all points of E_i) with probability at most $\delta k \leq 1/k^9$.

4.2.1 *Lifting to $\Re^n$*

Now we address the classification problem in $\Re^n$.

We have already indicated how to produce a pure sample of size k from a single gaussian, thus getting a very good approximation to this gaussian in $\Re^m$. Now, we notice that in fact if $k > \log^3 n$ (if not, just use $O(\log^3 n)$ instead of k in the description above) this many samples are enough to reconstruct the gaussian even in a random subspace of dimension $O(\log n/c^2)$. We proceed to do this reconstruction. Let H denote this gaussian. Note that unlike in the previous section, H is guranteed to be not a "fake." (This is what all our work in $\Re^m$ has achieved for us, and also the reason why we describe our approach as bootstrapping.) Suppose H is the projection of F_i. If H is not a 2-circuloid then F_i is not spherelike.

Now we work with the entire set of $S = O(\mathrm{poly}(k) n \log n / w_{\min})$ samples. Let $\mathrm{Ann}_0(H)$ denote the annulus of H in Mahalonobis norm that contains all but $1/|S|k$ of the probability mass. Consider the subset of S that lies in $\mathrm{Ann}_0(H)$. There are two cases. If H is not a 2-circuloid, then (c, γ)-separation implies that all these points come from F_i. Thus we can reconstruct F_i.

If H is a 2-circuloid on the other hand, then it is possible that $\mathrm{Ann}_0(H)$ is "polluted" by points from other gaussians, but one can argue as using the distance concentration results and Corollary 14 that the pairwise distance among such points is substantially less than $\sqrt{2}R$ (where R is the radius of H). Then we can easily identify those points and get with high probability an essentially pure sample from F_i.

5. MAX-LIKELIHOOD ESTIMATION

We consider max-likelihood fit of a mixture of k spherical gaussians of equal radius to (possibly) unstructured data. Recall the density function of a spherical gaussin of variance σ (and radius $\sigma\sqrt{n}$) is $(2\pi\sigma^n)^{-1}\exp(-|x-p|^2/\sigma^2)$.

Let $x_1, x_2, \ldots, x_M \in \Re^n$ be the points. Let $p_1, p_2, \ldots, p_k$ denote the centers of the gaussians in the max-likelihood solution. For each datapoint x_j let $p_{c(j)}$ denote the closest center. Then the mixing weights of the optimum mixture $w_1, w_2, \ldots, w_k$ are determined by considering, for each i, the fraction of points whose closest center is p_i.

The loglikelihood expression is obtained by adding terms for the individual points to obtain the following.

$$\text{Constant} + Mn \log \sigma \sum_i w_i + \sum_j \frac{\left|x_j - p_{c(j)}\right|^2}{2\sigma^2}.$$

The optimum value $\hat{\sigma}$ is obtained by differentiation (and noting $\sum_i w_i = 1$)

$$\hat{\sigma}^2 = \frac{1}{Mn} \sum_j \left|x_j - p_{c(j)}\right|^2, \tag{20}$$

which simplifies the likelihood expression to

$$\text{Constant} + Mn \log \hat{\sigma} + \frac{Mn}{2}.$$

Thus the goal is to minimize $\hat{\sigma}$, which from (20) involves minimizing the familiar objective function from the sum-of-square version of the k-median problem. Note however that we do not know the centers of the gaussians. However, picking the center to be one of the sample points changes the objective of the k-median problem by at most a factor 4. Hence results of Charikar et al. [3] provide an $O(1)$ approximation to $\hat{\sigma}^2$, and hence an approximation to $\log(\hat{\sigma})$ that is correct within an $O(1)$ *addtive* factor. More efficient algorithms for k-median are now known, so there is some hope that the observations of this section may lead to some practical learning algorithms.

6. CONCLUSIONS

We have presented several algorithms. Combining them into a single algorithm as well as extending to more families of logconcave distributions seems feasible and will be attempted for the final version.

Several open problems remain. The first concerns gaussians with significant overlap. For example, consider mixtures of spherical gaussians with pairwise intercenter distance only $O(\max\{\sigma_1, \sigma_2\})$. In this case, a constant fraction of their probability masses overlap, and the solution to the classification problem is not unique.Can the gaussians still be learnt? The second problem concerns general gaussians whose probability masses do not overlap much but which appear to coalesce under random projection. For example, consider a pair of concentric gaussians that have the same axis orientation. (Of course, these axes are unknown and are not the same as the coordinate axes.) In $n-2$ axis directions their variance is σ^2, and in other remaining two directions their variances are $1, \sigma$ and $\sigma, 1$ respectively. If $\sigma = \Omega(\log n)$ the difference in the last two coordinates is enough to differentiate their samples with probability $1 - 1/\text{poly}(n)$. But after projection to a $O(\log n)$ dimensional subspace, this difference disappears. Hence neither distance-based clustering nor projection-based clustering seems able to distinguish their samples.

The third open problem concerns max-likelihood estimation, which seems to involve combinatorial optimization with very bizarre objective criteria once we allow nonspherical gaussians.

We suspect all the above open problems may prove difficult.

Acknowledgements

We thank Sanjoy Dasgupta for many helpful discussions, including drawing our attention to [16]. We also thank Laci Lovasz and David McAllester for useful suggestions.

7. REFERENCES

[1] N. Alon, J. Spencer and P. Erdös. The probabilistic method. Wiley Interscience, 1992.

[2] J. Bourgain. Random points in isotropic convex sets. Convex geometric analysis (Berkeley, CA, 1996), 53–58, Math. Sci. Res. Inst. Publ., 34, Cambridge Univ. Press, Cambridge, 1999.

[3] M. Charikar, S. Guha, E. Tardos, and D. Shmoys. A constant-factor approximation algorithm for the k-median problem. *Proc. 31st ACM STOC*, 1999.

[4] S. DasGupta. Learning mixtures of gaussians. *Proc. IEEE Foundations of Computer Science*, 1999.

[5] S. Dasgupta and L. Schulman. *Personal communication*, March 2000.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistics Soc. Ser. B*,**39**:1-38, 1977.

[7] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. *ACM Conference on Computational Learning Theory*, 1999.

[8] W. J. Hoeffding. Probability inequalities for sums of bounded random variables. *J. American Statistical Assoc*, **58**(301):13-30, March 1963.

[9] P. J. Huber. Projection pursuit. *Annals of Statistics*, **13**(2):435-475, June 1985.

[10] W. B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. *Contemp. Math.* **26**:189-206, 1984.

[11] R. Kannan, G. Li, *Sampling according to the multivariate normal density*, in the Proceedings of the 37 th Annual Symposium on the Foundations of Computer Science, IEEE (1996) pp 204-213.

[12] L. Leindler. On a certain converse of Hölder's inequality II. stochastic programming. *Acta Sci. Math. Szeged* **33**:217-223(1972).

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

[14] A. Prékopa. Logarithmic concave measures with applications to stochastic programming. *Acta Sci. Math. Szeged* **32**:301-316 (1971).

[15] A. Prékopa. On logarithmic concave measures and functions. *Acta Sci. Math. Szeged* **34**:335-343 (1973).

[16] R. A. Redner, H. F. Walker. Mixture densities, maximum likelihood and the EM Algorithm. *SIAM Review*, **26**(2):195-239, 1984.

[17] M. Rudelson. Random vectors in the isotropic position. *J. Func. Anal.* **164**:60-72, 1999.
[18] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*, Wiley, 1985.

APPENDIX

Now we state several technical facts used in the paper.

Concentration results for sum of gaussian variables. We state a result regarding concentration (around the mean) of sums of gaussian variables. This is stronger than the results in Section 3 and are useful in Section 4.

LEMMA 13. *Let $y = (y_1, y_2, \ldots y_n)$ be a vector where each y_i is distributed as $N(0,1)$. Let $\sigma_1, \sigma_2, \ldots \sigma_n$ be any positive reals, with σ_{max} their maximum. Consider the random variable $Y = \sum_{i=1}^n \sigma_i^2 y_i^2$ with R^2 denoting $E(Y) = \sum_{i=1}^n \sigma_i^2$. Then for any $t > 0$, we have*

$$\begin{aligned} Prob\left(Y \geq R^2 + t\right) &\leq \exp\left(-\frac{t^2}{16\sum_{i=1}^n \sigma_i^4}\right) + \exp\left(-\frac{t}{8\sigma_{max}^2}\right) \\ Prob\left(Y \leq R^2 - t\right) &\leq \exp\left(-\frac{t^2}{4\sum_{i=1}^n \sigma_i^4}\right). \end{aligned} \quad (21)$$

Let $s = \sqrt{\sum_i \sigma_i^4}$. In a sample of size M, the probability is at least $1 - \epsilon$ that for each Y satisfies

$$R^2 - 2s\sqrt{\log M + \log\frac{1}{\epsilon}} \leq Y \leq R^2 + 8(\log M + \log\frac{1}{2\epsilon})\sigma_{max}^2 \quad (22)$$

$$+ 4s\sqrt{\log M + \log\frac{1}{2\epsilon}} \quad (23)$$

PROOF. (idea) Note that (23) follows directly from (21) and (21).

The rest of the proof is Chernoff-like in that it uses exponential generating function of Y.

□

COROLLARY 14. *Let G_1, G_2 be spherical gaussians of radii r_1, r_2 and centers p_1, p_2. Let $|p_1 - p_2|_2 = c \cdot r_1$ for $c \geq 0$. Let $Vol_{G_1}(A_2) \geq \gamma$, where A_2 is the spherical shell consisting of points whose distance-squared to the center of G_2 is between $r_2^2(1-\Delta)$ and $r_2^2(1+\Delta)$. Then*

$$r_2^2 \in (1+c^2)r_1^2 \pm \left(\Delta r_2^2 + \frac{r_1^2 \log 1/\gamma}{\sqrt{n}}(1+3c)\right).$$

Remark: This corollary says that either G_2 and G_1 have essentially the same radius and center, or else there is a close connection between the ratio of their radii and their intercenter distance.

Eccentricity changes of a projected gaussian

Dasgupta [4] shows that the parameter $\sigma_{\max}/\sigma_{\min}$ of a gaussia tends to reduce under random projection. Here we have a stronger version of his lemma, whose notable feature (required in our algorithm in Section 4) is that the statement makes no mention of n, the dimension we start with. Recall that a t-circuloid is a gaussian with $\sigma_{\max}/\sigma_{\min} \leq t$.

THEOREM 15. *There is a m_0 such that the following is true for all $m > m_0$ and all $\delta > 0$. If an ellipsoid in $\Re^n$ with $\sigma_{\max}^2 \leq R^2/128(m\log^2 m + \log(1/2\delta))$ is projected to a random subspace of dimension m, then with probability at least $1-\delta$, the projected Gaussian is a 2-circuloid of radius $R\frac{\sqrt{m}}{\sqrt{n}}$.*

PROOF. A careful version of Dasgupta's proof, using the above concentration results. □

VC Dimension and Sampling Errors

When we draw random samples from some distribution (in our case, this distribution is a mixture of gaussians) then the fraction that lie in any ellipsoid is approximately the same as the ellipsoid's probability mass. The same is also true for any solid that is a boolean combination of $O(1)$ ellipsoids. This is a simple consequence of the fact that the VC-dimension of ellipsoids is $O(n \log n)$. For an introduction to VC-dimension and its use in the proof of the next theorem, see Alon and Spencer [1].

We will use the following notation. For a distribution $\mathcal{D}$, the probability mass of a region R is denoted $\mathrm{Vol}_{\mathcal{D}}(R)$. If the distribution is a uniform distribution on a finite set S, $\mathrm{Vol}_S(R)$ is short-hand for $|R \cap S| / |S|$.

THEOREM 16. *Let h be a positive integer and $\epsilon > 0$. Then there is a constant c such that for all n the following is true. Let $\mathcal{D}$ be any distribution in $\Re^n$ and let S be a sample of size at least*

$$\frac{20n\log nh}{\epsilon^2}\log^2(nh/\epsilon)$$

from D. Then with probability at least $1 - 1/n$ the following is true: for every set of h ellipsoids $E_1, E_2, \ldots, E_h$,

$$|Vol_{\mathcal{D}}(\cap_i E_i) - Vol_S(\cap_i E_i)| \leq \epsilon. \quad (24)$$

A simple corollary of this theorem is that an expression similar to (24) is also true for any other boolean combination of two ellipsoids instead of $\cap$.

Log-concave distributions: Brunn-Minkowski theorems

The *Minkowski sum* of two convex bodies A, B is defined as

$$A + B = \{a + b : a \in A, b \in B\}.$$

FACT 17. *If A, B are two spheres of radii r_1, r_2 respectively and centers μ_1, μ_2 then $A + B$ is a sphere of radius $r_1 + r_2$ and centered at $\mu_1 + \mu_2$.*

A gaussian density function μ is log-concave: for every pair of points $x_1, x_2 \in \Re^n$ and a $\lambda \in [0,1]$, it satisfies

$$\log \mu(\lambda x_1 + (1-\lambda)x_2) \geq \lambda \log \mu(x_1) + (1-\lambda)\log \mu(x_2).$$

The following is a consequence of (the more general) Leindler-Prékopa inequality [12, 14, 15] for log-concave measures.

FACT 18. *If A, B are convex bodies, $\mathcal{D}$ is a gaussian distribution, and $\lambda \in (0,1)$ then*

$$Vol_{\mathcal{D}}(\lambda A + (1-\lambda)B) \geq (Vol_{\mathcal{D}}(A))^{\lambda}(Vol_{\mathcal{D}}(B))^{1-\lambda}.$$