Learning mixtures of arbitrary gaussians

(STOC 2001 submission)

Sanjeev Arora*
Princeton University

Ravi Kannan†
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 some cases, 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.

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. (By contrast, we will allow spheres to overlap, and also consider gaussians of arbitrary shape.)

In this paper we design a variety of polynomial-time algorithms for learning gaussians. Each algorithm has its own strong points, and we do not yet have a single algorithm that combines these strong points. In general, if the gaussians in the mixture are restricted to be spherelike (possibly of different

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

†kannan-ravindran@cs.yale.edu

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 is 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 is, of course, poor for an arbitrary point set. But if the points are 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 coordinate 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. If the σ_i's are not "too different", then moment-generating methods (see Lemma 8 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

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

be thought of as *spherelike.* The sharpness of this concentration depends roughly on the quantity $\sigma_{\max}/R$, which we call the *eccentricity* [2]. The larger the eccentricity, the more "smeared" the gaussian is. Our algorithm will deal with both types of gaussians.

Section 3 will also use a related parameter, called the *median* radius of the gaussian. This is the radius R' such that $1/2$ the probability mass of the gaussian lies in a ball of radius R'. It is related to radius defined above, for example $R' \leq \sqrt{2}R$. Other relationships between the two follow from Lemma 8

2.2 Likelihoods, classification, etc.

Let $(w_1, F_1, w_2, F_2, \ldots, w_m, F_m)$ be a mixture of gaussians in $\Re^n$. With any point $x \in \Re^n$, one can associate m numbers $(F_i(x))_{i=1,\ldots,m}$ 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. 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.)
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.

Our results. Our results may be summarized as follows.

A. PTAS for Problem 4 when the gaussians are spherical. In fact, for spherical gaussians one can find a classification whose loglikelihood is within a factor $(1 + 1/n^{1/4-\epsilon})$ of the optimum.

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), which is NP-hard. 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).

Although a clustering is also sought in parts 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). Section 3 describes an algorithm using distance-based clustering. The algorithm 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

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

from that information to find a better clustering. Both sections omit many details for lack of space but they appear in the Appendix.

We note that Dasgupta 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.

3 Distance based Clustering

In this section, we give an algorithm for Problem 1 —find the correct labelling of all samples— of Section 2.2. The gaussians in the mixture have to satisfy a certain separation condition. The algorithm 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.

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. It is easy to come up with a pair of nearby gaussians so that even if one knew the gaussians exactly, the maximum-likelihood labeling (as defined in Section 2.2) is incorrect whp. 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 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 $\sigma_{i,max} \approx R_i/\sqrt{n}$. Then the separation required is:

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

It can be shown that this separation is essentially necessary (within a factor of $\sqrt{\log n}$) for the solution to the Classification Problem to be unique with high probability; see Lemma 21 in the Appendix.

Now we describe why distance-based clustering works. Under our separation condition, 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.

Such distance concentration results were known prior to our work, at least for spherical or spherelike gaussians. Here we prove such results for general Gaussians using a new approach: Isoperimetric inequalities (see Theorem 1). Note that results about concentration about the median/mean value can

be proven using the more traditional moment-generating function approach —indeed, we do this in Lemma 8, which is stronger than the results of this section and useful in Section 4. However, the approach using Isoperimetric inequality has two advantages. First, it applies to a more general class of distributions than just Gaussians (which we do not use in this draft but plan to in the final version). Second, they provide concentration results 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 moment-generating function approach.

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 seems 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 $O(1/f^2)$. 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*

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

Ravi: What is the + sign below ???

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 $1 - 2e^{-t}$,*

$$(R + t\sigma)^2 + |z-p|^2 + 2\sqrt{t}|z-p|\sigma \quad \geq |x-z|^2 \quad \geq ((R - t\sigma)^+)^2 + |z-p|^2 - 2\sqrt{t}|z-p|\sigma$$

Proof: We have

$$|x-z|^2 = ((x-p)+(p-z)) \cdot ((x-p)+(p-z)) = |x-p|^2 + |p-z|^2 + 2(x-p)\cdot(p-z).$$

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. □

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*

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

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.) In this case, $\sigma_{i,max} \approx R_i/\sqrt{n}$, where the error is small enough that our calculations below are valid. Choosing $t = \Omega(\log(nk/\delta))$, it is easy to see that there are positive constants c, c' such that (with high probability),

$$|x-y|^2 \leq 2R_i^2(1 + \frac{ct}{\sqrt{n}}) \forall x, y \in S_i \qquad \forall i \tag{4}$$

$$|x-y|^2 \geq 2\min(R_i^2, R_j^2) + \frac{c't(R_i+R_j)^2}{\sqrt{n}} \forall x \in S_i, \forall y \in S_j, \quad \forall i \neq j. \tag{5}$$

We will assume that c, c' are high enough. Now we describe the algorithm for this case. 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 $|x_i - x_l|$, $\left|x_j - x_l\right|$ are within a multiplicative factor $(1 + 6t/\sqrt{n})$ of each other. The following claim follows from (4) and (5).
Claim: *With high probability over the choice of the sample, the following is true. Nodes corresponging to the smallest gaussian form an isolated clique. 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.

3.3 The general case

Let $\delta > 0$ be the probability of failure allowed. In what follows, we choose

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

The Algorithm

1. Initialization : $T \leftarrow S$.
2. 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.]
3. 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}.]
4. 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].

$$B(x, \alpha + s\nu) \cap T = B(x, \alpha + (s-1)\nu) \cap T.$$

5. 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.]
6. 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.]
7. Repeat until T is empty.

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 all should all extend to give results similar to the Gaussian case. This will be explored in the final version.

4 Classification algorithm 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 poly(n) — then this sample complexity is close to the information-theoretic lowerbound of $\Omega(n \log n/w_{\min})$. (This lowerbound follows from 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. Section F.3 mentions some generalizations of the algorithm.

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 $R/\sigma_{\max} > (\log k)^3 \log 1/\gamma$. 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 Section A.1 in the Appendix.) Note that for a spherelike gaussian, the γ-radius is essentially the same as its radius (see Lemma 8).

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, and $1/\gamma$ is the sample complexity of the algorithm, which as mentioned is $O(\text{poly}(k)n \log n/w_{\min})$.

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 7 *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 give only 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 reconstruct in $\Re^n$ (See Section F.2 in the Appendix.)

[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.

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 14. It will be at most $1/k^{10}$

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 (details appear in the appendix). First, one can show using standard VC dimension arguments —see Theorem 14— that gaussians with mixing weight at least $1/2k$ are "discovered" by the enumeration and they do not lose any of their samplepoints during the FILTER procedure (see Fact 23). For gaussians that are not spherelike, this is totally trivial since Definition 2 says that every other gaussian has negligible probability mass inside its γ-ball. For a spherelike gaussian the argument is slightly more nontrivial and relies on the observation that every other gaussian that "pollutes" $\mathrm{Ann}(G)$ is somewhat smaller, so the distance concentration results imply that Step 1 of FILTER would remove all these polluting points whp, without affecting the points that "belong" there.

Next, one has 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 (Theorem 24). This has two cases: either G is a 2-circuloid (Lemma 25) or not (Lemma 26). The first case is proved easily using concentration results. 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. 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 $\lambda A + (1-\lambda)B$ is devoid of samplepoints. Using Leindler-Prekopa we can show a lowerbound on its probability mass that implies that the inbetween ball is light (in the sense defined in the FILTER procedure). Thus the fake gaussian G gets rejected.

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 is why our final step is to take 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 $\delta k \leq 1/k^9$.

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 $\sum_i w_i = 1$)

$$\hat{\sigma}^2 = \frac{1}{Mn}\sum_j \left|x_j - p_{c(j)}\right|^2, \tag{6}$$

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 (6) involves minimizing the familiar objective function from the sum-of-square version of the k-median problem. The 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 are now known.

Interpreting our earlier algorithms in terms of likelihoods. We showed how to solve the sample classification problem when the component gaussians are well-separated. What if the component gaussians are not separated? The algorithm may output inbetween gaussians. One can show for spherical gaussians however that one still gets a $(1+\epsilon)$-approximation to the optimum loglikelihood. Details are omitted.

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 (Lemma 21). 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.

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.

A Concentration results for sum of gaussian variables

We prove results regarding concentration (around the mean) of sums of gaussian variables. These are stronger than the results in Section 3 and are useful in Section 4.

Lemma 8 *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*

$$Prob\left(Y \geq R^2 + t\right) \leq \exp\left(-\frac{t^2}{16\sum_{i=1}^n \sigma_i^4}\right) + \exp\left(-\frac{t}{8\sigma_{max}^2}\right) \quad (7)$$

$$Prob\left(Y \leq R^2 - t\right) \leq \exp\left(-\frac{t^2}{4\sum_{i=1}^n \sigma_i^4}\right). \quad (8)$$

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

$$R^2 - 2s\sqrt{\log M + \log \frac{1}{\epsilon}} \leq Y \leq R^2 + 8(\log M + \log \frac{1}{2\epsilon})\sigma_{max}^2 + 4s\sqrt{\log M + \log \frac{1}{2\epsilon}} \tag{9}$$

Proof: Note that (9) follows directly from (7) and (8).

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

PROOF OF (7) Let λ be a positive real number to be specified later. It will be chosen such that $\lambda\sigma_{max}^2 \leq 1/4$, which impliess that all the integrals below are defined (i.e., are finite) as the reader may check. We have (using the independence of the y_i),

$$E(e^{\lambda Y}) = \prod_{i=1}^{n} E(e^{\lambda\sigma_i^2 y_i^2})$$

$$= \prod_{i=1}^{n} \int_{y=-\infty}^{\infty} \frac{e^{-y^2/2}}{\sqrt{2\pi}} e^{\lambda\sigma_i^2 y^2}\, dy$$

$$= \prod_i \frac{1}{\sqrt{1 - 2\lambda\sigma_i^2}} \text{ using the substitution } z = \sqrt{1 - 2\lambda\sigma_i^2}\, y.$$

So we have

$$\text{Prob}\left(Y \geq R^2 + t\right) = \text{Prob}\left(e^{\lambda Y - \lambda(R^2+t)} \geq 1\right)$$

$$\leq E\left(e^{\lambda Y - \lambda(R^2+t)}\right) \leq e^{-\lambda(R^2+t)} \prod_{i=1}^{n} \frac{1}{\sqrt{1 - 2\lambda\sigma_i^2}}.$$

Taking logs of both sides, we have (using $\lambda\sigma_{max}^2 \leq 1/4$),

$$\log\left(\text{Prob}\left(Y \geq R^2 + t\right)\right)$$

$$\leq -\lambda R^2 - \lambda t - \frac{1}{2}\sum_i \log(1 - 2\lambda\sigma_i^2)$$

$$\leq -\lambda t + 4\lambda^2 \sum_i \sigma_i^4 \text{ using } -\log(1 - x) \leq x + 2x^2 \text{ for } 0 \leq x \leq 1/2.$$

In the case that $t \leq 2\sum \sigma_i^4/\sigma_{max}^2$, we will choose

$$\lambda = \frac{t}{8\sum_i \sigma_i^4},$$

(which incidentally minimizes the last expression) to get the upper bound (in fact only the first term of it) asserted.

In the case when $t \geq 2\sum \sigma_i^4/\sigma_{max}^2$, we put $\lambda = 1/(4\sigma_{max}^2)$ and we get the upper bound (only the second term). Thus Part 1 of the lemma is proved.

PROOF OF (8): Let

$$\lambda = \frac{t}{2\sum \sigma_i^4}.$$

We have (using the independence of the y_i),

$$E(e^{-\lambda Y}) = \prod_{i=1}^{n} E(e^{-\lambda\sigma_i^2 y_i^2})$$

$$= \prod_{i=1}^{n} \int_{y=-\infty}^{\infty} \frac{e^{-y^2/2}}{\sqrt{2\pi}} e^{-\lambda\sigma_i^2 y^2}\, dy$$

$$= \prod_i \frac{1}{\sqrt{1 + 2\lambda\sigma_i^2}} \text{ using the substitution } z = \sqrt{1 + 2\lambda\sigma_i^2}\, y.$$

So we have

$$\text{Prob}\left(Y \le R^2 - t\right) = \text{Prob}\left(e^{-\lambda Y + \lambda(R^2-t)} \ge 1\right)$$
$$\le E\left(e^{-\lambda Y + \lambda(R^2-t)}\right) \le e^{\lambda(R^2-t)} \prod_{i=1}^{n} \frac{1}{\sqrt{1+2\lambda\sigma_i^2}}.$$

Taking logs of both sides, we have

$$\log\left(\text{Prob}\left(Y \ge R^2 - t\right)\right)$$
$$\le \lambda R^2 - \lambda t - \frac{1}{2}\sum_i \log(1+2\lambda\sigma_i^2)$$
$$\le -\lambda t + \lambda^2 \sum_i \sigma_i^4 \text{ using } -\log(1+x) \le -x + \frac{x^2}{2} \text{ for } x \ge 0.$$

Now part 2 follows by plugging in λ which incidentally minimizes the last expression.
□

Corollary 9 *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 $c \ge 0$. Let $Vol_{G_1}(A_2) \ge \gamma$, 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.
Proof: Imagine picking a sample x from G_1. Then its distance-squared to p_2 is the random variable

$$|x - p_2|^2 = |x - p_1 + (p_1 - p_2)|^2 = |x - p_1|^2 + d^2 + |(x - p_1) \cdot (p_1 - p_2)|.$$

Note that $(x - p_1) \cdot (p_1 - p_2)$ is distributed as $N(0, cr_1^2/\sqrt{n})$.

Now imagine picking a sample of size $1/\gamma$ from G_1. With probability at least $1 - 1/e$, one of these points, say x, is in A_2, hence

$$|x - p_2|^2 \in r_2^2(1 \pm \Delta).$$

But by Lemma 8, with probability at least 7/8,

$$|x - p_2|^2 \in r_1^2 \pm \frac{r_1^2 \log 1/\gamma}{\sqrt{n}} + |p_1 - p_2|^2 \pm \frac{3cr_1^2 log 1/\gamma}{\sqrt{n}}.$$

Hence with nonzero probability, x satisfies both conditions, implying that the interval $r_2^2(1 \pm \Delta)$ intersects the interval

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

The claim now follows. □

A.1 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 $\sigma_{\max}/\sigma_{\min} \le t$.

Theorem 10 *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: It is easy to check (by picking a random set of directions for example) that if m is a large enough constant, then there exists a set of $m^{m\log m}$ directions in $\Re^m$ —a so-called ϵ-net— such that if the variance of a gaussian in all these directions are within a factor 2 of each other, then the gaussian is a 2-circuloid. We will argue that this condition is true of the projected gaussian.

Let the gaussian in $\Re^n$ be $F_{A,0}$, centered at the origin. A random projection from $\Re^n$ to $\Re^m$ can be viewed as a random rotation U of $\Re^n$ followed by a projection into the first m coordinates. Since U is a random rotation, we may as well assume the gaussian is axis-aligned and $A = diag(1/\sigma_i^2)$. Let $\pi_1, \ldots, \pi_m$ be the variances in the m coordinate directions after the projection. Since each coordinate direction is a random direction in $\Re^n$, we see that each π_i is distributed as $w^T A^{-1} w$, where w is a random unit vector in $\Re^n$. But this is just $\sum_i w_i^2\sigma_i^2$. Note moreover that one way to pick a unit vector w is by using $(y_1, y_2, \ldots, y_n)/\,|y|_2$ where the y_i's are chosen independently from $N(0,1)$. Thus

$$\sum_i w_i^2\sigma_i^2 = \frac{1}{|y|^2}(\sum_i y_i^2\sigma_i^2). \tag{10}$$

Of course, $|y|$ is sharply concentrated about n, so expectation is very close to R^2/n. We now apply the concentration results of Lemma 8 to the numerator.

For each of the "special" directions in $\Re^m$, the variance of the projected gaussian is distributed as (10). Lemma 8 implies that with probability at least $1-\delta$, the variance in *all* $m^{m\log m}$ special directions is within a multiplicative factor 1.3 of the expectation, provided

$$\begin{aligned} 16\sigma_{max}^2(m\log^2 m + \log 1/2\delta) &\leq \sum_i \sigma_i^2 \\ \frac{\sum_i \sigma_i^2}{\sqrt{\sum_i \sigma_i^4}} &\geq 4\sqrt{m\log^2 m + \log 1/2\delta}. \end{aligned}$$

Both these conditions are satisfied when $\sigma_{\max}^2 \leq R^2/128(m\log^2 m + \log(1/2\delta))$ as assumed in the Lemma. □

We have the following lemma as corollary. It says that in a gaussian that is not "spherelike", one finds lots of space with low probability mass. (That is the gaussian is "smeared.")

Lemma 11 *There is a $m_0 > 0$ such that the following is true for all $m \geq m_0$. Every gaussian in $\Re_n$ satisfies at least one of the following two conditions. Either its projection to $\Re^m$ is with probability at least $1/2^m$ a 2-circuloid, or*

$$r' > \frac{\sqrt{6\log n}}{m^3\sqrt{20\log k}} \cdot r'' \tag{11}$$

where r' is the radius of the ball containing $1-1/n^6$ of the probability mass and r'' is the radius of the ball containing $1-1/k^{20}$ of the probability mass.

Proof: Let r be the radius of the gaussian. If $\sigma \leq r/m^3$, then by Theorem 10 a random projection to $\Re^m$ will, with probability at least $1-1/2^m$, be a 2-circuloid.

Now we note two things. First, by Lemma 8 $r'' = O(\sqrt{\log k}r) = O(m^3\sqrt{\log k}\ \sigma_{\max})$. Second, with probability $\geq 1/n^5$, a random sample in the gaussian lies at a distance $> \sqrt{\log n}\ \sigma_{\max}$ from the center (in fact, its component along a single axis has this len

gth). Hence $r' > \sqrt{\log n}\ \sigma_{\max}$. This proves the Lemma. □

B Overlap of oblong gaussians and spherical gaussians

Let F be an oblong gaussian and C_1, C_2 be two spherical gaussians and let $\mathrm{Ann}(\cdot)$ denote the annulus of thickness $t\sqrt{n}$ in the Mahalonobis norm. (Remark: This annulus contains $1-e^{-t^2/2}$ of the probability mass.) Let $\mathrm{Vol}(\cdot)$ denote the usual geometric volume and $\mathrm{Vol}_F()$ denote probability mass under F.

Lemma 12 *For any body G,*

$$Vol_F(G) \leq e^{t^2/2}\frac{Vol(G \cap Ann(F))}{Vol(Ann(F))} + e^{-t^2/2}.$$

Proof: Imagine breaking up $\mathrm{Ann}(F)$ into N equal-volume pieces, where N is very large and each piece is very small (and hence has almost-uniform density). Let μ_i be the probability mass (wrt F) of the ith piece and $\mu_{\min}, \mu_{\max}$ respectively be the minimum and maximum of these. Then

$$\frac{\mu_{\max}}{\mu_{\min}} \leq e^{t^2/2}.$$

It follows that for any body G,

$$\mathrm{Vol}_F(G \cap \mathrm{Ann}(F)) \leq e^{t^2/2}\frac{\mathrm{Vol}(G \cap \mathrm{Ann}(F))}{\mathrm{Vol}(\mathrm{Ann}(F))}.$$

Furthermore, $\mathrm{Vol}_F(G \setminus \mathrm{Ann}(F)) \leq e^{-t^2/2}$, so the lemma is now proved. □

The following Lemma is a corollary.

Lemma 13 *If C_1, C_2 have radii r_1, r_2 respectively where $r_2 \geq r_1$. Suppose each $Vol_{C_i}(Ann(F)) \geq \gamma$, then*

$$\frac{Vol_F(Ann(C_1))}{Vol_F(Ann(C_2))} \leq \frac{e^{3t^2/2}}{\gamma}(\frac{r_1}{r_2})^{n-1} + e^{-t^2/2}.$$

Proof: By Lemma 12 we have

$$\begin{aligned}
\mathrm{Vol}_F(\mathrm{Ann}(C_1)) &\leq e^{t^2/2}\frac{\mathrm{Vol}(\mathrm{Ann}(F) \cap \mathrm{Ann}(C_1))}{\mathrm{Vol}(\mathrm{Ann}(F))} + e^{-t^2/2} \\
&\leq e^{t^2/2}\frac{\mathrm{Vol}(\mathrm{Ann}(C_1))}{\mathrm{Vol}(\mathrm{Ann}(F))} + e^{-t^2/2} \\
&= e^{t^2/2}(\frac{r_1}{r_2})^{n-1}\frac{\mathrm{Vol}(\mathrm{Ann}(C_2))}{\mathrm{Vol}(\mathrm{Ann}(F))} + e^{-t^2/2}
\end{aligned}$$

where the last line used the fact that the volumes of annuli of spherical gaussians vary as the $(n-1)$th power of the radius.

To simplify the above expression, we recall the known condition on $\mathrm{Vol}_{C_2}(\mathrm{Ann}(F))$

$$\begin{aligned}
\gamma &\leq \mathrm{Vol}_{C_2}(\mathrm{Ann}(F)) \\
&\leq e^{t^2/2}\frac{\mathrm{Vol}(\mathrm{Ann}(C_2) \cap \mathrm{Ann}(F))}{\mathrm{Vol}(\mathrm{Ann}(C_2))}
\end{aligned}$$

This gives an upperbound on $\mathrm{Vol}(C_2)$ which we substitute back to obtain

$$\begin{aligned}
\mathrm{Vol}_F(\mathrm{Ann}(C_1)) &\leq e^{t^2/2}(\frac{r_1}{r_2})^{n-1}\frac{e^{t^2/2}\mathrm{Vol}(\mathrm{Ann}(C_2) \cap \mathrm{Ann}(F))}{\gamma \mathrm{Vol}(\mathrm{Ann}(F))} + e^{-t^2/2} \\
&= \frac{e^{t^2}}{\gamma}(\frac{r_1}{r_2})^{n-1}\frac{\mathrm{Vol}(\mathrm{Ann}(C_2) \cap \mathrm{Ann}(F))}{\mathrm{Vol}(\mathrm{Ann}(F))} + e^{-t^2/2} \\
&\leq \frac{e^{t^2}}{\gamma}(\frac{r_1}{r_2})^{n-1}e^{t^2/2}\mathrm{Vol}_F(\mathrm{Ann}(C_2)) + e^{-t^2/2}
\end{aligned}$$

Thus the lemma is proved.

□

C 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 14 *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 $E_1, E_2, \ldots, E_h$,

$$|Vol_{\mathcal{D}}(\cap_i E_i) - Vol_S(\cap_i E_i)| \leq \epsilon. \tag{12}$$

A simple corollary of this theorem is that an expression similar to (12) is also true for any other boolean combination of two ellipsoids instead of $\cap$.

D 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 15 *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 16 *If A, B are convex bodies, $\mathcal{D}$ is a gaussian distribution, and $\lambda \in (0, 1)$ then*

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

E Details of Section 3

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 (13),(14),(15), (16), (17), (18) and (19) 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 $7\delta/10$.**

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 $1 - \delta/10$,

$$|S_i| \geq (0.9) w_i |S| \forall i. \tag{13}$$

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 $1 - \frac{\delta}{10}$,

$$\forall i, 1 \leq i \leq k, \forall x \in S_i, \qquad S_i \setminus B(x, \eta(x)) = \emptyset. \tag{14}$$

We have from lemma (5) that with probability at least $1 - \delta/10$,

$$\forall i, 1 \leq i \leq k, \forall x, y \in S_i, \qquad 2R_i^2 - 8t\sigma_{i,max}R_i \leq |x - y|^2 \leq 2(R_i + 2t\sigma_{i,max})^2. \tag{15}$$

Further, from lemma (6), we have that with probability at least $1 - \frac{\delta}{10}$,

$$\forall i, j, 1 \leq i \neq j \leq k, \forall x \in S_i, \forall y \in S_j,$$
$$|x - y|^2 \geq 2\min(R_i^2, R_j^2) + 60t(R_i + R_j)(\sigma_{i,max} + \sigma_{j,max}) + 30t^2(\sigma_{i,max}^2 + \sigma_{j,max}^2). \tag{16}$$

A (spherical) annulus A is a set of the form $B(x, R_1) \setminus B(x, R_2)$. >From the fact that the Vapnik-Chervonenkis dimension of balls in $\mathbf{R}^n$ is $O(n)$ and standard results, it follows that with probability at least $1 - \delta/4$, **every** annulus in space gets about the right number of points. More precisely, with probability at least $1 - \delta/10$, the following holds : (using (13))

$$||S_i \cap A| - |S_i|F_i(A)| \leq \frac{w_0^2|S_i|}{80w_i} \qquad \forall i, \qquad \forall \text{ annulii } A \text{ anywhere in space .} \tag{17}$$

Also the VC dimension of half spaces is $O(n)$ and it follows that with probability at least $1 - \delta/10$, we have

$$||S_i \cap H| - |S_i|F_i(H)| \leq |S_i|/80 \qquad \forall i \qquad \forall \text{ half spaces } H. \tag{18}$$

>From lemma (20), it follows that with probability at least $1 - \delta/10$, we have

$$\forall \text{ unit length vectors } w, \qquad \forall i, \qquad \frac{1}{|S_i|}\sum_{x \in S_i}(w \cdot (x - p_i))^2 \leq 2\sigma_{i,max}^2. \tag{19}$$

Lemma 17 *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 18 *Suppose $x \in S$ is the center of the ball $B(x, \alpha)$ found in the l th exectution of step (1) of the algorithm and suppose x belongs to S_i (i unknown to us). Then,*

$$B(x, \alpha) \cap S \subseteq S_i \tag{20}$$
$$|x - y|^2 \geq 2R_i^2 + 50t(\sigma_{i,max} + \sigma_{j,max})(R_i + R_j) + 20t^2(\sigma_{i,max}^2 + \sigma_{j,max}^2) \forall y \in S_j \forall j \neq i, j \in P. \tag{21}$$

Proof:

For any $j \in P$, all $y, z \in S_j$, we have from (15) that $|z - y|^2 \leq 2(R_j + 2t\sigma_{j,max})^2$. Thus, a ball of radius $\sqrt{2}(R_j + 2\sigma_{i,max})$ with y as center would qualify in step (1) of the algorithm by (13) . So, by definition of α in that step, we must have

$$\alpha \leq \sqrt{2}(R_j + 2t\sigma_{j,max}) \forall j \in P. \tag{22}$$

If now $B(x, \alpha)$ contains a point z from some S_j, $j \neq i$, by the inductive assumption in lemma (17), we must have $j \in P$. Then by (16), we have

$$\alpha^2 \geq 2\min(R_i^2, R_j^2) + 60t(R_i + R_j)(\sigma_{i,max} + \sigma_{j,max}) + 30t^2(\sigma_{i,max}^2 + \sigma_{j,max}^2),$$

which contradicts (22) (noting that (22) must hold for both i, j). This proves (20).

Now, from the lower bound of (15), it follows that

$$\alpha^2 \geq 2R_i^2 - 8R_i\sigma_{i,max}t.$$

So, from (22), it follows that

$$2R_j^2 \geq 2R_i^2 - 8t(R_i + R_j)(\sigma_{i,max} + \sigma_{j,max}) - 8t^2\sigma_{j,max}^2 \forall j \in P.$$

Thus from (16), we get (21). □

Claim 1 *The β, Q computed in step 2 of the algorithm satisfy*

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

Proof: For any unit length vector w, we have by (19) :

$$\sum_{x\in Q}(w \cdot (x - p_i))^2 \leq \sum_{x\in S_i}(w \cdot (x - p_i))^2 \leq 2|S_i|\sigma_{i,max}^2$$

Since this holds for every w, and the second moment about the mean is less than or equal to the second moment about p_i, we have that $\beta \leq 2\frac{|S_i|}{|Q|}\sigma_{i,max}^2$. This proves the upper bound on β.

Let u be the direction of the maximum variance of F_i. We wish to assert that the variance of Q along u is at least $|Q|\sigma_{i,max}/|S_i|$. To this end, first note that for any reals γ_1, γ_2, with $\gamma_1 > 0$, we have :

$$\begin{aligned}\text{Prob}_{F_i}(\gamma_2 - \gamma_1 \leq x \cdot u \leq \gamma_2 + \gamma_1) &= \frac{1}{2\sqrt{\pi}\sigma_{i,max}}\int_{\gamma_2-\gamma_1}^{\gamma_2+\gamma_1} e^{-r^2/2\sigma_{i,max}^2}dr \\ &\leq \gamma_1/(\sqrt{\pi}\sigma_{i,max}).\end{aligned}$$

Let $\gamma_2 = \frac{1}{|Q|}\sum_{x\in Q}(u \cdot x)$ and let $\gamma_1 = \frac{|Q|}{|S_i|}\sigma_{i,max}$. Then the strip $H = \{x : \gamma_2 - \gamma_1 \leq u \cdot x \leq \gamma_2 + \gamma_1\}$ satisfies $F_i(H) \leq \gamma_1/(\sqrt{\pi}\sigma_{i,max})$. So by (18)

$$|S_i \cap H| \leq 3|Q|/4.$$

So, we have that

$$\begin{aligned}\frac{1}{|Q|}\sum_{x\in Q}\sum_{x\in Q}(u \cdot x - \gamma_2)^2 &\geq \frac{1}{|Q|}\frac{|Q|}{4}\frac{|Q|^2}{|S_i|^2}\sigma_{i,max}^2 \\ &= \frac{1}{4}\sigma_{i,max}^2\frac{|Q|^2}{|S_i|^2},\end{aligned}$$

from which the lower bound on β obviously follows. □

Corollary 19 *The β computed in step 2) of the algorithm satisfies :*

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

Proof: Since $|Q| \geq 3w_0|S|/4$, the Claim above implies the corollary. □

Now since we are increasing the radius in steps of ν which is at most $\sigma_{i,max}$ (Corollary 19), we have using Lemma (18), that

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

and also that the s in step 3) exits.

For real γ, 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 algortihm), we get using (17) that

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

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

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

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

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

This implies from (17) that $|Q'| \geq 0.8|S_i|$ (note that Q' is found in step 4)). Thus from Claim (1), (noting that the proof of that claim works for any subset of S_i), we get that

$$1.6\sigma_{i,max}^2 \geq \beta' \geq 0.16\sigma_{i,max}^2. \tag{23}$$

>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 (15), we have $\alpha + (s-1)\nu \leq \sqrt{2}(R_i + 2t\sigma_{i,max})$. So, we have

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

Thus from (21), 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 (14), 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.

Lemma 20 *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 $\epsilon = 10n(\sqrt{\log n} + \sqrt{\log(1/\delta)})/\sqrt{|L|}$)*

$$\forall \text{ vectors } w,$$
$$E_F((w.(x - E_F(x))^2)(1-\epsilon) \leq E_S((w.(x - E_F(x))^2) \leq E_F((w.(x - E_F(x))^2)(1+\epsilon).$$

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 :

$$\text{forall vectors } w, E_F((w.x)^2)(1-\epsilon) \leq E_S((w.x)^2) \leq E_F((w.x)^2)(1+\epsilon).$$

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

$$\text{forall vectors } u, E_F((u.(Qx))^2)(1-\epsilon) \leq E_S((u.(Qx))^2) \leq E_F((u.(Qx))^2)(1+\epsilon).$$

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 $x_i x_j$ has mean 0 and variance $O(1)$. $E_S(x_i x_j)$ 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 :

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

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 $E_S(x_i x_j)$ 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.

□

E.1 Necessity for separation

We show that t-separation as used in the paper is essentially (upto $\sqrt{\log n}$ factor) the minimum separation for which the sample classification problem has a unique solution for mixtures of spherical gaussians.

Lemma 21 *Suppose all the Gaussians are spherical. Then,*

$$\begin{aligned} & \textit{Prob}\left(\forall i, \forall x \in S_i, \forall j \neq i, \qquad F_i(x) < F_j(x)\right) \leq 1/10 \\ \implies \qquad & |p_i - p_j|^2 \geq (R_i^2 - R_j^2)^- + c'\sqrt{\log(nk/\delta)}\frac{(R_i + R_j)^2}{\sqrt{n}}, \end{aligned}$$

for some constant c'.

Proof: We will only sketch a proof here. Wlg assume that $R_i \leq R_j$. Suppose the conclusion fails; then clearly we must have

$$R_j^2 = R_i^2\left(1 + O\left(\frac{\sqrt{\log(nk/\delta)}}{\sqrt{n}}\right)\right).$$

Suppose x is a random sample picked according to F_i. Then $(x - p_i) \cdot (p_j - p_i)/|p_j - p_i|$ is normal with variance $\sigma_i^2 \approx R_i^2/n$; from this, it follows that with probability at least 1/poly, we have $|(x - p_j) \cdot (p_i - p_j)|/|p_j - p_i|$ is at most $R_i/\sqrt{n}$. Also, the component of $x - p_j$ orthogonal to $p_i - p_j$ is the same as the component of $x - p_i$ orthogonal to $p_i - p_j$ which is normally distributed with mean 0 and variance at most $\frac{n-1}{n}R_i^2$; so with probability 1/poly, we have that x is at distance at most R_i from p_j. From this, it follows by a direct calculation that with this probability $F_j(x) \geq F_i(x)$. Since we have polynomially many samples, the lemma can be proved. □

F Details of Algorithm in Section 4

F.1 The algorithm: learning in $\Re^m$

As already mentioned, the algorithm will enumerate all subsets of size $O(m \log m)$ and compute a variance-covariance matrix for each. Each candidate thus obtained is subjected to the FILTER procedure.

Below, δ is an error parameter obtained by the Lemma on VC-dimension arguments Theorem 14. It will be at most $1/k^{10}$

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 G with the largest S_G, and it produces a random subsample of k points from S_G. The rest of the section will prove the following theorem.

Theorem 22 (Main) *This sample is whp a pure sample from a single gaussian in the mixture whose mixing weight is* $\geq 1/2k$.

Let $V_1, V_2, \ldots, V_k$ be the samplepoints from $E_1, E_2, \ldots, E_k$ in S_1. Note that the V_i's form a partition of S_1. By the law of large numbers, each $|V_i|$ has between $w_i(|S_1| \pm \sqrt{|S_1|} \log k)$ points out of which $w_i(|S| \pm \sqrt{|S|} \log k)$ points are in S. The main idea will be that the "winning" G will look as follows: there is an i such that S_G contains $1 - \delta k^5$ fraction of some V_i and at most δk^5 fraction of S_G will consists of points from $\cup_{p \neq i} V_p$. Thus S_G is "almost pure." Then a moment's thought shows that a random sample of size k from S_G is, with probability at least $1 - k \cdot \delta k^5$, a pure sample from G. This will prove Theorem 22.

Now we prove the claim about the "winning G."

First we show that every good candidate, namely, an E_i whose mixing weight is $\geq 1/2k$, stays in the running. Note it has $\geq w_i |S| /4k$ samples in S, which exceeds $O(m^2 \log^2 m)$. Hence the exhaustive enumeration will produce a 0.1-approximation (in the sense of [2, 17]) to every such ellipsoid. Furthermore if the projected gaussian is not a 2-circuloid, the original ellipsoid in $\Re^n$ was not spherelike. Hence $(c/2, 2^{-\Omega(m)})$-separation implies that $\mathrm{Ann}(G)$ is not "polluted" by points from other gaussians. Hence part 2 of FILTER is justified in using all the samples in Ann$G$ to improve the estimate for this gaussian.

Thus we have proved the following. Without loss of generality, assume F_1 is a gaussian with mixing weight $\geq 1/k$.

Fact 23 *If G is a good approximation to E_1, then S_G at the end of* FILTER *contains all of V_1 and possibly at most $\delta k |S|$ samplepoints from $\cup_{p \neq 1} V_p$.*

Proof: This is obvious if F_1 is not spherelike, because $(c/2, 2^{-\Omega(m)})$-separation implies that no other gaussian will have even $2^{-\Omega(m)}$ probability mass inside the $2^{-\Omega(m)}$-radius of F_1. Hence $\mathrm{Ann}(E_1)$ is uncorrupted by other gaussians whp.

If F_1 is spherelike, then the concentration results in Lemma 8 implies that its points never have distance much less than $\sqrt{2}R$ to each other. Nor do they have small distance to points in $\mathrm{Ann}(E_1)$ from other gaussians, as seen in Section 3.

In both cases, a standard VC dimension argument (Theorem 14) then shows that F_1 populates various spheres with samples proportional (upto an additive error γ) to their probability mass, so G does not lose any points due to FILTER. □

Now we prove that FILTER decimates fake candidates.

Theorem 24 *If a candidate gaussian has $\geq |S_1| /2k$ of the points left in its annulus after* FILTER *then all but $2\delta k$ fraction of these points come from a single gaussian in the mixture.*

The proof follows from Lemmas 25 and 26 which show that viable candidates at the end correspond to those obtained from a "pure" sample from a gaussian.

Lemma 25 *Let G be 2-circuloid with $|S_G| \geq |S_1|/k^5$. Suppose E_j is a projected gaussian that contributes between $1/k^5$ and $1/2$ of the samplepoints of S_G at the start of* FILTER. *Then all pairwise distances among points in $V_j \cap S_G$ are less than $\sqrt{2}R - 6\sqrt{20 \log k}R/\sqrt{m}$ (and hence these points are removed in Part 1 of* FILTER*).*

Proof: There are two cases. If E_j is itself a 2-circuloid, then this follows from Corollary 9.

If E_j is not a 2-circuloid, then the original gaussian F_j is not spherelike. Let r be the radius of F_j and r', r'' respectively be its γ-radius and $1/k^{20}$-radius. Lemma 11 implies $r'' = o(r')$. Let $s = \sqrt{m}/\sqrt{n}$, $s' = r'\sqrt{m}/\sqrt{n}$ and $s'' = r''\sqrt{m}/\sqrt{n}$ be the radii of the projections of these spheres.

Then the distance between the centers of G and E_j is at least $s'' - R$ and at most $s'' + R$. We claim $s'' < R/2$. Suppose not. Then $s' \gg R$, and all of $\mathrm{Ann}(G)$ is contained in the sphere of radius s' about E_j. Then the separation condition implies that no other gaussian can contribute any significant probability mass to $\mathrm{Ann}(G)$, which contradicts the assumption that E_j contributes less than $1/2$ the samplepoints in $\mathrm{Ann}(G)$. □

The proof of the next lemma will use properties of logconcave measures, specifically, the Leindler-Prekopa inequality, Fact 16 in the Appendix.

Lemma 26 *Suppose candidate G is not a 2-circuloid and at least two different E_i's contribute $\geq 1/k^5$ samplepoints to S_G at the start of* FILTER. *Then* FILTER *will find a light sphere.*

Proof: Let μ be the center of G. We prove the existence of the "light sphere" in several steps. Assuming the light sphere does not exist, we prove Claims 1, 2, 3, and show that they imply the existence of a light sphere.

Notation: For a solid A, $\mathrm{Vol}_S(A)$ denotes the fraction of samplepoints from S_1 in A. If H is a gaussian, $\mathrm{Vol}_H(A)$ denotes the probability mass of A under H.

Claim 1: *Suppose E_i is one of the gaussians contributing to $Ann(G)$ and it is not a 2-circuloid. Let ρ be its center and s'' be its $1/k^{20}$-radius. Then $|\mu - \rho| \leq 2s''$.*

Proof: Let s, s', s'' be the various radii for E_i similar to the ones defined in the proof of Lemma 25. We know that $s'' = o(s')$. Let B be the ball of radius s'' around E_i. If $|\mu - \rho| \geq 2s''$ then consider a ball isomorphic to B but shifted towards the center μ of G by a distance $2s''$. Then that ball is inside the ball of radius s' around E_i so well-separatedness implies it has no samplepoints. But log-concavity implies that this ball has even higher probability mass under G, namely, probability mass $\geq 1/k^5$. Thus it is a light sphere. Thus Claim 1 is proved. □

Thus well-separatedness together with Claim 1 implies that at most one E_i that is not a 2-circuloid can contribute significantly to $\mathrm{Ann}(G)$. Now we turn to 2-circuloids.

Claim 2: All t-circuloids that have significant contribution to G have radii that are the same upto a multiplicative factor $k^{O(1/m)}$ ($\approx 1 + O(\log k/m)$).

Proof: If C_1, C_2 are such circuloids then $\mathrm{Vol}_{C_i}(\mathrm{Ann}(G) \cap \mathrm{Ann}(C_i)) \geq 1/k^2$. If no light spheres exist, $\mathrm{Vol}_G(\mathrm{Ann}(G) \cap \mathrm{Ann}(C_i))$ must be significant as well and hence $\mathrm{Vol}_G(\mathrm{Ann}(C_1))$ and $\mathrm{Vol}_G(\mathrm{Ann}(G_2))$ are related within a multiplicative factor $k^{O(1)}$. Hence Lemma 13 implies that their radii are within the claimed factor. This proves Claim 2. □

Claim 3:Suppose no non t-circuloid makes a significant contribution to G. Then at most one t-circuloid does.

Proof: Suppose E_i is such a circuloid. Let its radius be r and center be ρ. We show that the distance $|\rho - \mu|$ is less than $r\sqrt{100 \log k}/\sqrt{m}$, whereupon the claim follows by the separation condition and Claim 2.

Assume for contradiction's sake that the distance $|\rho - \mu| \geq r\sqrt{100 \log k}/\sqrt{m}$. Let A be the ball of radius r centered at μ. Then logconcavity implies $\mathrm{Vol}_G(A) \geq \mathrm{Vol}_G(E_i)$. Let C be the ball of radius r whose center lies on the line joining ρ and μ at a distance $r\sqrt{20 \log k}/\sqrt{m})$ to ρ. It can be viewed as a Minkowski sum of A and E_i, hence Fact 16 implies that $\mathrm{Vol}_G(C) \geq \mathrm{Vol}_G(E_i)$. We claim it has essentially no samplepoints, and thus is a light sphere. After all, where could any such samplepoints come from? But C has negligible overlap with E_i, they cannot come from E_i. They cannot come from a t-circuloid because such a t-circuloid cannot have its center closer than distance cr to E_i, and hence less than

$r(c - \sqrt{20 \log k}/\sqrt{m})$ to the center of C. Then Corollary 9 (applied to C and this t-circuloid) implies that its radius would be approximately $r/\sqrt{1+c^2}$, which is not allowed by Claim 2. Hence we have a contradiction and Claim 3 is proved. □

Now we are ready to prove that Claims 1–3 imply that G has a light sphere, which proves the Lemma. By Claims 1 to 3, such a G can have significant mass contribution from at most one non 2-circuloid E_i of the type described in Claim 1 and one or more 2-circuloids with to G. Let F be such a 2-circuloid, with radius l and center f. By $(c/2, 2^{-\Omega(m)})$-separation and Claim 1, $|\mu - f| \geq s' - 2s''$. But consider the line joining f, μ. All but $1 - 1/k^{20}$ of the probability mass of F lies in a thin slice of thickness $r\sqrt{40 \log k}/\sqrt{m}$ that is normal to this line. Call this slice D. Then $\mathrm{Vol}_G(D) \geq 1/k^5 - 1/k^{12}$. By considering the Minkowski sum of this slice with an identical slice at μ, we conclude that all parallel shifts of this slice (which are themselves contained in spheres) towards μ have probability mass at least as large, namely, $1/k^5 - 1/k^{12}$. Can there be any samplepoints in these shifted slices? By the separation condition they come from another 2-circuloid. They also cannot arise from E_i because all the samples of E_i are within distances $3s''$ of μ, and $s'' = o(s')$. Thus we have proved the existence of a light sphere. □

F.2 Algorithm: Step 2

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 bought 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 in Fact 23 that the pairwise distance among such points is substantially less than $\sqrt{2}R$ (where R is the radius of H). Then we can filter out those points and get with high probability an essentially pure sample from F_i.

F.3 Generalizations

We mention some generalizations of this algorithm.

A) Dasgupta has proposed a "weak" model for gaussian distributions, whereby we are allowed to assume that the density of samplepoints in each sphere or ellipsoid is "about" right (the same as what one would obtain by a VC-dimension argument), but not facts about distance concentration such as Lemma 8, which are too sensitive to noise.

The algorithm can be made to work in this weak model, though its running time rises to $n^{\log n}$. The idea is to do an exhaustive density-check in $O(\log n/c^2)$ dimensions instead of the simple reconstruction in Section F.2.

B) We can learn mixtures of distributions that are products of intervals (i.e., is a box). The reason is that projections of boxes act like gaussians. We require their containing balls to be separate.

C) We can generalize in fact to any other log-concave family of distributions whose each member has a "succinct" description of size poly(dimension) and this succinctness property is preserved under projection. (These conditions allow us to use the technique of projecting data into a lower-dimensional space and using enumeration to generate a list of candidates.) The notion of well-separatedness is as follows. For a component distribution D_1 let r' be all but $1/n^6$ of the probability mass lies within a ball of radius r'. Then we define the *containing ball* to be one whose radius is $2r'$. We require these containing balls to be disjoint.

Then the same ideas as in Section 4 work.