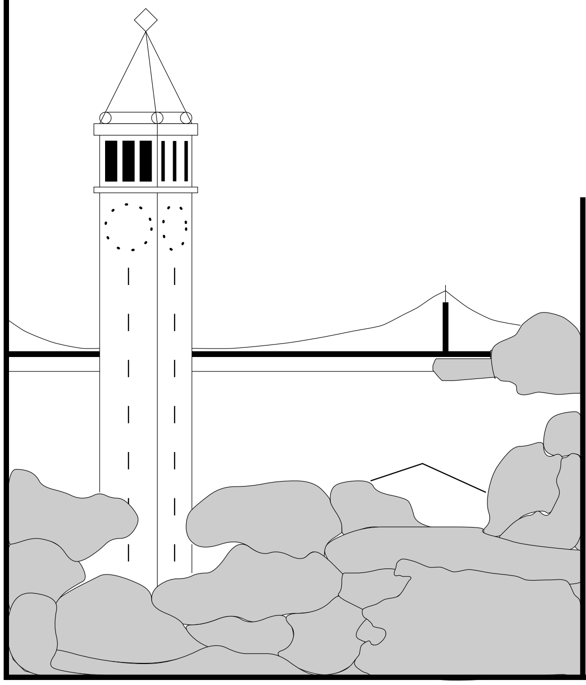
LEARNING MIXTURES OF GAUSSIANS

Part I: Theory

SANJOY DASGUPTA



Report No. UCB/CSD-99-1047

May 1999

Computer Science Division (EECS)
University of California
Berkeley, California 94720

1 Introduction

The mixture of Gaussians is among the most enduring, well-weathered models of applied statistics. A widespread belief in its fundamental importance has made it the object of close theoretical and experimental study for over a century. In a typical application, sample data are thought of as originating from various possible sources, and the data from each particular source is modelled by a Gaussian. This choice of distribution is common in the physical sciences and finds theoretical corroboration in the central limit theorem. Given mixed and unlabelled data from a weighted combination of sources, the goal is to identify the generating mixture of Gaussians, that is, the nature of each Gaussian source – its mean and covariance – and also the ratio in which each source is present, known as its 'mixing weight'.

A brief history of the many uses of mixtures of Gaussians, ranging over fields as varied as psychology, geology, and astrophysics, has been compiled by Titterington, Smith, and Makov (1985). Their book outlines some of the fascinating and idiosyncratic techniques that have been applied to the problem, harking back to days of sharpened pencils and slide rules. Modern methods delegate the bulk of the work to computers, and in their ranks the most popular seems to be the expectation-maximization (EM) algorithm formalized by Dempster, Laird, and Rubin (1977). An explanation of this algorithm, along with helpful remarks about its performance in learning mixtures of univariate Gaussians, can be found in the book of Duda and Hart (1973), in an excellent survey article by Redner and Walker (1984), and in a recent monograph by Lindsay (1995).

The EM algorithm has much to recommend it, but even its most ardent supporters concede a drastic deterioration in performance as the dimension of the data rises, especially if the different clusters overlap. This degradation has been experimentally documented in many places and tends to be regarded as yet another example of 'the curse of dimensionality'.

This paper describes a very simple algorithm for learning an unknown mixture of Gaussians with an arbitrary common covariance matrix and arbitrary mixing weights, in time which scales only linearly with dimension and polynomially with the number of Gaussians. We show that with high probability, it will learn the true centers of the Gaussians to within the precision specified by the user. Previous heuristics have been unable to offer any such performance guarantee, even for highly restricted subcases like mixtures of two spherical Gaussians.

The new algorithm works in three phases. First we prove that it is possible to project the data into a very small subspace without significantly increasing the overlap of the clusters. The dimension of this subspace is independent of the number of data points and of the original dimension of the data. We show, moreover, that after projection general ellipsoidal Gaussians become almost spherical and thereby more manageable. In the second phase, the modes of the low-dimensional distribution are found using a simple algorithm whose performance we rigorously analyze. Finally, the low-dimensional modes are used to reconstruct the original Gaussian centers.

2 Overview

2.1 Background

An n-dimensional Gaussian $N(\mu; \Sigma)$ has density function

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right).$$

Although the density is highest at μ, it turns out that for large n the bulk of the probability mass lies far away from this center. This is the first of many surprises that high-dimensional space will spring upon us. A point $\mathbf{x} \in \mathbb{R}^n$ chosen randomly from a spherical Gaussian $N(0; \sigma^2 I_n)$ has expected squared Euclidean norm $\mathbf{E}(\|\mathbf{x} - \mu\|^2) = n\sigma^2$. The law of large numbers forces the distribution of this squared length to be tightly concentrated around its expected value for big enough n. That is to say, almost the entire distribution lies in a thin shell at distance $\sigma\sqrt{n}$ from the center of the Gaussian! Thus the natural scale of this Gaussian is in units of $\sigma\sqrt{n}$.

It is reasonable to imagine, and is borne out by experience with techniques like EM (Duda & Hart; Redner & Walker), that a mixture of Gaussians is easiest to learn when the Gaussians do not overlap too much. Taking cue from our discussion of $N(\mu; \sigma^2 I_n)$, we adopt the following

Definition Two Gaussians $N(\mu_1, \sigma^2 I_n)$ and $N(\mu_2, \sigma^2 I_n)$ are considered *c-separated* if $\|\mu_1 - \mu_2\| \geq c\sigma\sqrt{n}$. More generally, Gaussians $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ in $\mathbb{R}^n$ are c-separated if

$$\|\mu_1 - \mu_2\| \geq c\sqrt{n \max(\lambda_{max}(\Sigma_1), \lambda_{max}(\Sigma_2))},$$

where $\lambda_{max}(\Sigma)$ is shorthand for the largest eigenvalue of Σ. A mixture of Gaussians is c-separated if its component Gaussians are pairwise c-separated.

A 2-separated mixture corresponds roughly to almost completely separated Gaussians, whereas a mixture that is 1- or $1/2$-separated contains Gaussians which overlap significantly. We will be able to deal with Gaussians that are arbitrarily close together; the running time will, however, inevitably depend upon their radius of separation.

2.2 The problem of dimension

What makes this learning problem difficult? In low dimension, for instance in the case of univariate Gaussians, it is often possible to simply plot the data and visually estimate a solution, provided the Gaussians maintain a respectable distance from one another. This is because a reasonable amount of data conveys a fairly accurate idea of the overall probability density. The high points of this density correspond to centers of Gaussians and to regions of overlap between neighbouring clusters. If the Gaussians are far apart, these modes themselves provide good estimates of the centers.

Easy algorithms of this kind fail dismally in higher dimension. Consider again the Gaussian $N(\mu, \sigma^2 I_n)$. We must pick $2^{O(n)}$ random points from this distribution in order to get just a few which are at distance $\leq 1/2\sigma\sqrt{n}$ from the center! The data in any sample of plausible size, if plotted somehow, would resemble a few scattered specks of dust in an enormous void. What can we possibly glean from such a sample? Such gloomy reflections have prompted researchers to try mapping data into spaces of low dimension.

2.3 Dimensionality reduction

The naive algorithm we just considered requires at least about 2^d data points to learn a mixture of Gaussians in $\mathbb{R}^d$, and this holds true of many other simple algorithms that one might be tempted to concoct. Is it possible to reduce the dimension of the data so dramatically that this requirement actually becomes reasonable?

One popular technique for reducing dimension is principal component analysis (PCA for regulars). It is quite easy to symmetrically arrange a group of k spherical Gaussians so that a PCA projection to any dimension $d < \Omega(k)$ will collapse many of the Gaussians together, and thereby decisively derail any hope of learning. For instance, place the centers of the $(2j-1)^{st}$ and $2j^{th}$ Gaussians along the j^{th} coordinate axis, at positions j and $-j$. The eigenvectors found by PCA will roughly be coordinate axes, and the discarding of any eigenvector will collapse together the corresponding pair of Gaussians. Thus PCA cannot in general be expected to reduce the dimension of a mixture of k Gaussians to below $\Omega(k)$. Moreover, computing eigenvectors in high dimension is a very time-consuming process, fraught with numerical concerns.

A much faster technique for dimensionality reduction, which has received a warm welcome in the theoretical community, is expressed in the Johnson-Lindenstrauss (1984) lemma. The gist is that any M data points in high dimension can be mapped down to $d = \frac{4 \log M}{\epsilon^2}$ dimensions without distorting their pairwise distances by more than $(1 + \epsilon)$. However, for our purposes this reduced dimension is still far too high! According to our rough heuristic, we need 2^d data points, and this exceeds M by many orders of magnitude.

We will show that *for the particular case of mixtures of Gaussians*, by using projection to a randomly chosen subspace as in the Johnson-Lindenstrauss lemma, we can map the data into just $d = O(\log k)$ dimensions. Therefore the amount of data we will need is only polynomial in k.

This might puzzle readers who are familiar with random projection, because the usual motive behind such projections is to approximately preserve relative distances between data points. However, in our situation we expressly do not want this. We want most of the pairwise distances to contract significantly, so that the fraction of points within distance $\Delta\sqrt{d}$ of any Gaussian center in the reduced space $\mathbb{R}^d$ is exponentially greater than the fraction of points within distance $\Delta\sqrt{n}$ of the same center in the original space $\mathbb{R}^n$. At the same time, we do not want the distances between different Gaussians to contract; we must make sure that Gaussians which are initially well-separated remain so when they are projected. These conflicting requirements are accommodated admirably by a projection to just $O(\log k)$ dimensions.

2.4 The algorithm

We are now in a position to present the algorithm. The user thoughtfully supplies: ϵ, the accuracy within which the centers are to be learned; δ, a confidence parameter; and w_{min}, the smallest mixing weight that will be considered. These values will be discussed in full detail in the next section. The parameters M, d, l, p, and q depend upon the inputs, and will be determined later.

Sample S consists of M data points in $\mathbb{R}^n$.

1. Select a random d-dimensional subspace of the original space $\mathbb{R}^n$, and project the data into this space. This takes time only $O(Mdn)$.

2. In the projected space:

   - For each data point $x \in S$, let r_x be the smallest radius such that there are $\geq p$ points within distance r_x of x.
   - Start with $S' = S$.
   - For $i = 1 \dots k$:
     - Let estimate $\widehat{\mu}_i^*$ be the point $x \in S'$ with the lowest r_x.
     - Find the q closest points to this estimated center, and remove them from S'.
   - For each i, let S_i refer to the l points in S which are closest to $\widehat{\mu}_i^*$.

3. Let the (high-dimensional) estimate $\widehat{\mu}_i$ be the mean of S_i in $\mathbb{R}^n$.

This algorithm is very simple to implement.

2.5 Low-dimensional clustering

The data get projected from $\mathbb{R}^n$ to $\mathbb{R}^d$ via a linear map. Since any linear transformation of a Gaussian conveniently remains a Gaussian, we can pretend that the projected data themselves come from a mixture of low-dimensional Gaussians.

The second step of the algorithm is concerned with estimating the means of these projected Gaussians. Regions of higher density will tend to contain more points, and we can roughly imagine the density around any data point x to be inversely related to radius r_x. In particular, the data point with lowest r_x will be near the center of some (projected) Gaussian. If the Gaussians all share the same covariance, then this data point will be close to the center of that Gaussian which has the highest mixing weight.

Once we have a good estimate for the center of one Gaussian, how do we handle the rest of them? The problem is that one Gaussian may be responsible for the bulk of the data if it has a particularly high mixing weight. All the data points with low r_x might come from this one over-represented Gaussian, and need to be eliminated from consideration somehow.

This is done by growing a wide region around the estimated center, and removing from contention all the points in it. The region should be large enough to remove all high-density points in that particular Gaussian, but should at the same time leave intact the high-density points of other Gaussians. The reader may wonder, how can we possibly know how large this region should be if we have no idea of either the covariance or the mixing weights? First, we pick the q points closest to the estimated center rather than using a preset radius; this accomplishes a natural scaling. Second, the probability of encountering a data point at a distance $\leq r$ from the center of the Gaussian grows exponentially with r, and this rapid growth tends to eclipse discrepancies of mixing weight and directional variance.

Both the techniques described – choosing the point with next lowest r_x as a center estimate, and then "subtracting" the points close to it – rely heavily on the accuracy of spherical density estimates. They assume that for any sphere in $\mathbb{R}^d$, the number of data points which fall within that sphere is close to its expected value under the mixture distribution. That this is in fact the case follows from the happy circumstance that the concept class of spheres in $\mathbb{R}^d$ has VC-dimension only $d+1$.

Finally, we mention that this low dimensional part of the algorithm works best on spherical Gaussians. But here our method of projection helps us again, tremendously: even if we start with highly skewed ellipsoidal Gaussians, the random projection will make them almost spherical!

2.6 Mapping back to the original space

At this stage, projected centers in hand, we recall that our actual task was to find the Gaussian means in the original high-dimensional space. Well, this is not too difficult, at least conceptually. For each low-dimensional estimated center $\widehat{\mu}_i^*$, we pick the l data points closest to it in $\mathbb{R}^d$, call them S_i, and then average these same points in $\mathbb{R}^n$. We expect S_i to be relatively uncontaminated with points from other Gaussians (although we cannot of course avoid the odd straggler), and thus its mean should closely approximate μ_i.

We complete our overview with one last clarification. How exactly did the projection help us? It enabled us to find, for each Gaussian, a set of data points drawn mostly from that Gaussian.

2.7 The main results

In the next section we will prove a dimensionality reduction lemma and then demonstrate the correctness of the algorithm in the following simple but instructive case.

Theorem 1 If data is drawn from a mixture of k Gaussians in $\mathbb{R}^n$ which is c-separated, for $c > 1/2$, and if the smallest mixing weight is $\Omega(\frac{1}{k})$, and if the Gaussians are all spherical with unknown covariance matrix $\sigma^2 I_n$, then with probability $> 1-\delta$, all the center estimates returned by the above algorithm are accurate within L_2 distance $\epsilon\sigma\sqrt{n}$. The reduced dimension is $d = O(\log \frac{k}{\epsilon\delta})$ and the amount of data required is $M = k^{O(\log^2 1/(\epsilon\delta))}$.

By building upon this proof, we will in the subsequent section arrive at the more general

Theorem 2 Suppose now that the Gaussians are no longer restricted to being spherical but instead have an unknown common covariance matrix Σ with maximum and minimum eigenvalues $\sigma^2_{max}, \sigma^2_{min}$ respectively, and eccentricity $\varepsilon = \sigma_{max}/\sigma_{min}$. Then with probability $> 1 - \delta$, the center estimates returned by the algorithm are accurate within L_2 distance $\epsilon\sigma_{max}\sqrt{n}$. If the eccentricity $\varepsilon \leq O(\frac{n^{1/2}}{\log k/\epsilon\delta})$, then the reduced dimension is $d = O(\log \frac{k}{\epsilon\delta})$ and the number of data points needed is $M = k^{O(\log^2 1/(\epsilon\delta))}$.

Our algorithm can in fact handle Gaussians which are arbitrarily close together. It is only to curtail the proliferation of symbols that we insist upon $1/2$-separation in these theorems. The mixing weights and eccentricity are similarly unrestricted.

Finally, a word about the inputs: in addition to the usual ϵ (accuracy) and δ (confidence) parameters, the user is expected to supply a lower bound w_{min} on the mixing weights which will be considered.

3 Spherical Gaussians

3.1 Notation

The following notation will be used consistently through the remainder of the paper.

ϵ	Desired accuracy, supplied by user
δ	Desired confidence, supplied by user
ϵ_0	Accuracy of spherical density estimates
M	Overall number of data points
n	Original dimension of data
d	Reduced dimension
k	Number of Gaussians
$N(\mu_i; \Sigma_i)$	The i^{th} Gaussian in $\mathbb{R}^n$
w_i	Mixing weight of the i^{th} Gaussian
w_{min}	Lower bound on the w_i, supplied by user
c, c^*	Lower bound on the separation of Gaussians in the original and reduced spaces, respectively
$N(\mu_i^*, \Sigma_i^*)$	Projection of i^{th} Gaussian into the reduced space $\mathbb{R}^d$
$\pi^*(\cdot)$	Density of the projected mixture of Gaussians
σ	Some standard deviation radius
$\nu_\sigma(\cdot)$	Density of $N(0; \sigma^2 I_d)$ in $\mathbb{R}^d$.
$B(x; r)$	Sphere of radius r centered at x
$B(r'; r)$	$B(x; r)$ for some x at distance r' from the origin
l, p, q	Integer parameters needed by algorithm
ρ	Parameter needed for analysis, related to ϵ

In this section, we will prove the correctness of our algorithm assuming that the Gaussians are spherical with identical covariance matrices $\Sigma_1 = \cdots = \Sigma_k = \sigma^2 I_n$.

3.2 Reducing dimension

Definition For a positive definite matrix Σ, let $\lambda_{max}(\Sigma)$ and $\lambda_{min}(\Sigma)$ refer to its largest and smallest eigenvalues, respectively, and denote by $\varepsilon(\Sigma)$ the *eccentricity* of the matrix, that is, $\sqrt{\lambda_{max}(\Sigma)/\lambda_{min}(\Sigma)}$.

The following dimensionality reduction lemma applies to arbitrary mixtures of Gaussians. Its statement refers to the notion of separation introduced in the overview.

Lemma 1 (Dimensionality Reduction) For any $c > 0$, let $\{(w_i, \mu_i, \Sigma_i)\}$ denote a c-separated mixture of k Gaussians in $\mathbb{R}^n$, and let $\delta_1 > 0$ and $\epsilon_1 > 0$ designate confidence and accuracy parameters, respectively. With probability $> 1 - \delta_1$, the projection of this mixture of Gaussians onto a random d-dimensional subspace yields a $(c\sqrt{1-\epsilon_1})$-separated mixture of Gaussians $\{(w_i, \mu_i^*, \Sigma_i^*)\}$ in $\mathbb{R}^d$, provided

$$d \geq \frac{4}{\epsilon_1^2} \ln \frac{k^2}{\delta_1}.$$

Moreover, $\lambda_{max}(\Sigma_i^*) \leq \lambda_{max}(\Sigma_i)$ and $\lambda_{min}(\Sigma_i^*) \geq \lambda_{min}(\Sigma_i)$. In particular therefore, $\varepsilon(\Sigma_i^*) \leq \varepsilon(\Sigma_i)$.

Proof. Consider a single line segment in $\mathbb{R}^n$, of squared length L. If the original space is projected onto a random d-dimensional subspace, the squared length of this line segment becomes some L^*, of expected value $\mathbf{E}L^* = Ld/n$. It was shown by Johnson and Lindenstrauss (1984) that $\mathbf{P}(L^* < (1-\epsilon)Ld/n) \leq e^{-d\epsilon^2/4}$. Their proof has been simplified by Frankl and Maehara (1988) and most recently by the author and Gupta (1998).

We shall apply this to the line segments joining pairs of Gaussian centers in the original space. There are less than k^2 such segments. By the above discussion, using the value of d specified in the theorem, we find that with probability $> 1 - \delta_1$, in the projected space each new pair of centers μ_i^* and μ_j^* will satisfy

$$\begin{aligned} \|\mu_i^* - \mu_j^*\|^2 &\geq (1-\epsilon_1)\|\mu_i - \mu_j\|^2 d/n \\ &\geq (1-\epsilon_1)(c^2 n \max(\lambda_{max}(\Sigma_i), \lambda_{max}(\Sigma_j)))d/n \\ &\geq c^2(1-\epsilon_1)d \max(\lambda_{max}(\Sigma_i), \lambda_{max}(\Sigma_j)), \end{aligned}$$

where the second line uses the fact that the original Gaussians are c-separated. It follows that the projected mixture is $(c\sqrt{1-\epsilon_1})$-separated, if we can show that $\lambda_{max}(\Sigma_i) \geq \lambda_{max}(\Sigma_i^*)$.

This is straightforward. Write the projection, say P^T, as a $d \times n$ matrix with orthogonal rows. P^T sends Gaussian (μ, Σ) in $\mathbb{R}^n$ to $(P^T\mu, P^T\Sigma P)$ in $\mathbb{R}^d$, whereby

$$\begin{aligned} \lambda_{max}(P^T\Sigma P) &= \max_{u \in \mathbb{R}^d} \frac{u^T(P^T\Sigma P)u}{u^T u} = \max_{v \in \mathbb{R}^n} \frac{(P^T v)^T(P^T\Sigma P)(P^T v)}{(P^T v)^T(P^T v)} \\ &= \max_{v \in \mathbb{R}^n} \frac{(PP^T v)^T\Sigma(PP^T v)}{(PP^T v)^T(PP^T v)} \\ &\leq \max_{v \in \mathbb{R}^n} \frac{v^T\Sigma v}{v^T v} = \lambda_{max}(\Sigma). \end{aligned}$$

(The denominator in the second line uses $P^T P = I_d$.) In similar fashion we can show that $\lambda_{min}(\Sigma_i^*) \geq \lambda_{min}(\Sigma_i)$, completing the proof. ∎

Remarks (1) If two of the Gaussians in the original mixture are particularly far apart, say cf-separated for some $f \geq 1$, then in the projected space they will be $(cf\sqrt{1-\epsilon_1})$-separated. This will be useful to us later. (2) A projection onto a random lower-dimensional subspace will in fact dramatically reduce the eccentricity of Gaussians, as demonstrated in the last lemma of this paper.

Corollary In order to ensure that the projected mixture is at least $1/2$-separated (that is, $c^* \geq 1/2$) with probability $> 1 - \delta/4$, it is enough to choose

$$d \geq \frac{4c^4}{(c^2 - 1/4)^2} \ln \frac{4k^2}{\delta}.$$

3.3 Crude density bounds

We need to ensure that every spherical region in the projected space gets approximately its fair share of data points. This is accomplished effortlessly by VC dimension arguments.

Lemma 2 (Accuracy of density estimates) Let $\nu(\cdot)$ denote any density on $\mathbb{R}^d$. If the number of data points seen satisfies

$$M \geq O\left(\frac{1}{\epsilon_0^2}\left(d\log\frac{1}{\epsilon_0}+\log\frac{1}{\delta}\right)\right)$$

then with probability $> 1-\delta/4$, for every sphere $B \subset \mathbb{R}^d$, the empirical probability of that sphere differs from $\nu(B)$ by at most ϵ_0; that is, the number of points that fall in B is in the range $M\nu(B) \pm M\epsilon_0$.

Proof. For any sphere $B \subset \mathbb{R}^d$, let $1_B(x) = \mathbf{1}(x \in B)$ denote the indicator function for B. The concept class $\{1_B : B \in \mathbb{R}^d \text{ is a sphere}\}$ has VC-dimension $d+1$ (Dudley, 1979). The rest follows from well-known results about sample complexity; details can be found, for instance, in the book by Pach and Agarwal (1995). ∎

We will henceforth assume that M meets the conditions of this lemma and that all spherical density estimates are accurate within ϵ_0. Since all the Gaussians in $\mathbb{R}^n$ have covariance matrix $\sigma^2 I_n$, their projections have covariance exactly $\sigma^2 I_d$. Let $\nu_\sigma(\cdot)$ denote the density of a single Gaussian $N(0;\sigma^2 I_d)$ and let $\pi^*(\cdot)$ denote the density of the entire projected mixture. We now examine a few technical properties of ν_σ. Our first goal is to obtain probability lower bounds which will be used to show that there are many data points near the center of each Gaussian.

Lemma 3 (Crude density lower bounds) If $\tau \leq 1/3$ and $d \geq 10$,
(a) $\nu_\sigma(B(0;\tau\sigma\sqrt{d})) \geq \tau^d$; and
(b) $\nu_\sigma(B(\tau\sigma\sqrt{d};\tau\sigma\sqrt{d})) \geq \tau^d$.

Proof. Let V_d denote the volume of the unit ball in d dimensions. We will use the lower bound

$$V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)} \geq \frac{(2\pi)^{d/2}}{2(d/2)^{d/2}}$$

which follows from the observation $\Gamma(1+k) \leq k^k 2^{-(k-1)}$ for $k \geq 1$. Now center a sphere at the mean of the Gaussian. A crude bound on its probability mass is

$$\nu_\sigma(B(0;\tau\sigma\sqrt{d})) \geq \left(\frac{e^{-(\tau\sigma\sqrt{d})^2/2\sigma^2}}{(2\pi)^{d/2}\sigma^d}\right)(V_d(\tau\sigma\sqrt{d})^d) \geq \frac{\tau^d}{2}(2e^{-\tau^2})^{d/2} \geq \tau^d,$$

Continuing in the same vein, this time for a displaced sphere,

$$\nu_\sigma(B(\tau\sigma\sqrt{d};\tau\sigma\sqrt{d})) \geq \left(\frac{e^{-4(\tau\sigma\sqrt{d})^2/2\sigma^2}}{(2\pi)^{d/2}\sigma^d}\right)(V_d(\tau\sigma\sqrt{d})^d) \geq \frac{\tau^d}{2}(2e^{-4\tau^2})^{d/2} \geq \tau^d$$

provided the stated conditions on τ and d are met. ∎

Next we would like an idea of how fast the probability mass of a sphere decreases as it moves away from the center of a Gaussian, and of how this mass increases as the sphere grows.

Lemma 4 (Relative densities of different spheres)
(a) If z, z' are points for which $\|z\| \geq \tau\sigma\sqrt{d}$ and $\|z'\| = \|z\| + \Delta$, then

$$\nu_\sigma(B(z;\tau\sigma\sqrt{d})) \geq \nu_\sigma(B(z';\tau\sigma\sqrt{d}))e^{\Delta^2/2\sigma^2}.$$

(b) If $r+s \leq \frac{1}{2}\sigma\sqrt{d}$ then

$$\frac{\nu_\sigma(B(0;r+s))}{\nu_\sigma(B(0;r))} \geq \left(\frac{r+s}{r}\right)^{d/2}.$$

Proof. For the first bound, assume without loss of generality that the centers of the two spheres of equal radius lie along the same direction $\widehat{u}$. Pair each point x in $B(z;\tau\sigma\sqrt{d})$ with $x+\Delta\widehat{u}$ in $B(z';\tau\sigma\sqrt{d})$. Writing $x=(x_u,y)$, we find that the density of the former point divided by that of the latter is

$$\frac{e^{-(x_u^2+||y||^2)/2\sigma^2}}{e^{-((x_u+\Delta)^2+||y||^2)/2\sigma^2}} = \exp\left\{\frac{\Delta^2+2x_u\Delta}{2\sigma^2}\right\} \geq e^{\Delta^2/2\sigma^2}$$

since $x_u \geq 0$ by our lower bound on $||z||$.

For the second bound, notice that

$$\nu_\sigma(B(0;r)) \;=\; \int_{x\in B(0;r)}\nu_\sigma(x)dx \;=\; \left(\frac{r}{r+s}\right)^d\int_{y\in B(0;r+s)}\nu_\sigma\left(y\cdot\frac{r}{r+s}\right)\,dy$$

via the change in variable $y = x\cdot\frac{r+s}{r}$. Therefore

$$\frac{\nu_\sigma(B(0;r+s))}{\nu_\sigma(B(0;r))} = \left(\frac{r+s}{r}\right)^d\frac{\int_{y\in B(0;r+s)}\nu_\sigma(y)dy}{\int_{y\in B(0;r+s)}\nu_\sigma(y\cdot\frac{r}{r+s})dy}.$$

We will bound this ratio of integrals by considering a pointwise ratio. For any $y\in B(0;r+s)$,

$$\frac{\nu_\sigma(y)}{\nu_\sigma(y\cdot\frac{r}{r+s})} = \frac{e^{-||y||^2/2\sigma^2}}{e^{-(||y||^2/2\sigma^2)\cdot(r/r+s)^2}} \geq \left(\frac{r}{r+s}\right)^{d/2}$$

under the given conditions on r and s. ∎

3.4 Estimating the projected centers

We are now in a position to prove that for an appropriate choice of the parameters p and q, the algorithm will find one data point close to each projected center. The value ρ used in the analysis that follows will turn out to be proportional to ϵ. To simplify calculations, we will from the outset assume that $\rho \leq 1/3$ and that $d \geq 10$.

Lemma 5 If $\frac{p}{M}+\epsilon_0 \leq w_{min}\rho^d$, then for each i, there is a data point $x\in S$ such that x is at distance at most $\rho\sigma\sqrt{d}$ from μ_i^* and moreover, for any such point, $\geq p$ data points lie within distance $\rho\sigma\sqrt{d}$ of x.

Proof. In light of the fact that all spherical density estimates are accurate within ϵ_0, we need only show that $w_{min}\nu_\sigma(0;\rho\sigma\sqrt{d}) \geq \epsilon_0$ and $w_{min}\nu_\sigma(\rho\sigma\sqrt{d};\rho\sigma\sqrt{d}) \geq \frac{p}{M}+\epsilon_0$. The rest follows from Lemma 3. ∎

This lemma guarantees that in the projected space, there will be many points close to each center, and in fact, that for data points x within distance $\rho\sigma\sqrt{d}$ of any center, $r_x \leq \rho\sigma\sqrt{d}$. We next need to show that r_x will be significantly larger for points x which lie further from the center, at distance $\geq 3\rho\sigma\sqrt{d}$.

Definition $F = 1-e^{-2\rho^2 d}-\frac{1}{w_{min}}e^{-(1-4\rho)^2 d/32}$. Setting $d = O\left(\log\frac{1}{\rho^2 w_{min}}\right)$ guarantees $F \geq \min\{\frac{1}{4},\frac{1}{2}\rho^2 d\}$.

Lemma 6 Suppose that for some radius $r \le \rho\sigma\sqrt{d}$, the sphere $B(x;r)$ contains p data points, where x is a point at distance $\ge 3\rho\sigma\sqrt{d}$ from μ_i^* and distance $\ge \frac{1}{4}\sigma\sqrt{d}$ from the other projected Gaussian centers. Then any point z within distance $\rho\sigma\sqrt{d}$ of μ_i^* must have

$$\pi^*(B(z;r)) \ge 2\epsilon_0 + \pi^*(B(x;r)),$$

provided $\epsilon_0 \le \frac{F}{2-F}\frac{p}{M}$.

Proof.

$$\begin{aligned}\pi^*(B(x;r)) &\le w_i\nu_\sigma(B(3\rho\sigma\sqrt{d};r)) + \nu_\sigma(B(1/4\sigma\sqrt{d};r)) \\ &\le w_i\nu_\sigma(B(\rho\sigma\sqrt{d};r))e^{-2\rho^2 d} + \nu_\sigma(B(\rho\sigma\sqrt{d};r))e^{-(\frac{1}{4}-\rho)^2 d/2} \\ &\le w_i\nu_\sigma(B(z;r))\left(e^{-2\rho^2 d} + \frac{e^{-(1-4\rho)^2 d/32}}{w_{min}}\right) \\ &\le \pi^*(B(z;r))(1-F),\end{aligned}$$

where the second and third lines are supplied by Lemma 4, and the last line uses the definition of F. The stated condition on ϵ_0 then yields

$$\pi^*(B(z;r)) - \pi^*(B(x;r)) \ge \pi^*(B(x;r))\frac{F}{1-F} \ge \left(\frac{p}{M} - \epsilon_0\right)\frac{F}{1-F} \ge 2\epsilon_0,$$

as promised. ∎

In order to satisfy the conditions of these last two lemmas, we adopt the following

Definitions $p = Mw_{min}\rho^d(1-\frac{F}{2})$ and $\epsilon_0 = \frac{p}{M}\frac{F}{2-F}$.

The lemma above can be restated as follows: suppose data point x in the projected space is more than $3\rho\sigma\sqrt{d}$ away from the closest center. Then any data point z within distance $\rho\sigma\sqrt{d}$ of that same center must have $r_z < r_x$. This implies roughly that within any Gaussian, the lowest r_x values come from data points which are within distance $3\rho\sigma\sqrt{d}$ of the center.

A potential problem is that a few of the Gaussians might have much higher mixing weights than the rest and consequently have a monopoly over small r_x values. In order to handle this, after selecting a center estimate we eliminate the q points closest to it, where

Definition $q = w_{min}\nu_\sigma(B(0;\frac{3}{8}\sigma\sqrt{d}))M$. This value is independent of σ and can easily be computed from d, using a lookup table or some simple numerical technique.

It will turn out that the q points closest to each center estimate must include all points within a radius of $\frac{1}{4}\sigma\sqrt{d}$ of the actual projected center and no points which are further than $\frac{1}{2}\sigma\sqrt{d}$ from this center.

Lemma 7 For any point x within distance $3\rho\sigma\sqrt{d}$ of μ_i^*,
(a) $\pi^*(B(x;(\frac{1}{4}+3\rho)\sigma\sqrt{d})) \le \frac{q}{M} - \epsilon_0$; and
(b) $\pi^*(B(x;(\frac{1}{2}-4\rho)\sigma\sqrt{d})) \ge \frac{q}{M} + \epsilon_0$,
provided that $\epsilon_0 \le \frac{q}{2M}$, $\rho \le \frac{1}{96}$, and $d \ge 7\ln\frac{2}{w_{min}}$. In other words, if T denotes the q points closest to x,

$$B(\mu_i^*;\frac{1}{4}\sigma\sqrt{d}) \subseteq T \subseteq B(\mu_i^*;(\frac{1}{2}-\rho)\sigma\sqrt{d}).$$

Proof. In light of the condition on ϵ_0, it is enough to show

$$2\nu_\sigma(B(0;(\tfrac{1}{4}+3\rho)\sigma\sqrt{d})) \le w_{min}\nu_\sigma(B(0;\tfrac{3}{8}\sigma\sqrt{d})) \quad\text{and}\quad \nu_\sigma(B(0;(\tfrac{1}{2}-7\rho)\sigma\sqrt{d})) \ge \tfrac{3}{2}\nu_\sigma(B(0;\tfrac{3}{8}\sigma\sqrt{d})).$$

These bounds are immediate from lemma 4 and the stated conditions on ρ and d. ▮

Remark Let us now assume d, ρ, and ϵ_0 satisfy the various conditions of the above lemmas.

Lemma 8 With probability $> 1 - \delta/2$, each center $\widehat{\mu}_i^*$ chosen by the algorithm is within distance $3\rho\sigma\sqrt{d}$ of a true projected center μ_i^*.

Proof, by induction on the number of centers selected so far.

Referring back to the algorithm, the first center-estimate chosen is the point $x \in S$ with lowest r_x. By Lemma 5, this $r_x \le \rho\sigma\sqrt{d}$. Let μ_i^* be the projected center closest to x. Since the Gaussians are $1/2$-separated, x is at distance at least $\frac{1}{4}\sigma\sqrt{d}$ from all the other projected centers. By Lemma 6, we then see that x must be within distance $3\rho\sigma\sqrt{d}$ of μ_i^*.

Say that at some stage in the algorithm, center-estimates $\widehat{C}$ have already been chosen, $|\widehat{C}| \ge 1$, and that these correspond to true centers C. Select any $y \in \widehat{C}$; by the induction hypothesis there is a j for which $\|y - \mu_j^*\| \le 3\rho\sigma\sqrt{d}$. S' does *not* contain the q points closest to y. By Lemma 7, this removes $B(\mu_j^*; \frac{1}{4}\sigma\sqrt{d})$ from S', yet no point outside $B(\mu_j^*; (\frac{1}{2} - \rho)\sigma\sqrt{d})$ is eliminated from S' on account of y.

Let z be the next point chosen, and let μ_i^* be the center closest to it which is not in C. We have seen that z must be at distance at least $\frac{1}{4}\sigma\sqrt{d}$ from centers in C. Because of the separation of the mixture, z must be at distance at least $\frac{1}{4}\sigma\sqrt{d}$ from all centers but μ_i^*. Again due to the separation of the Gaussians, all points within distance $\rho\sigma\sqrt{d}$ of μ_i^* remain in S', and therefore z is potentially one of these, whereupon, by Lemma 5, $r_z \le \rho\sigma\sqrt{d}$. By Lemma 6 then, $\|z - \mu_i^*\| \le 3\rho\sigma\sqrt{d}$. ▮

3.5 Mapping back into high dimension

We may now safely assume that in $\mathbb{R}^d$, each estimated center $\widehat{\mu}_i^*$ is within $3\rho\sigma\sqrt{d}$ of the corresponding projected center μ_i^*. The set S_i consists of the l data points closest to $\widehat{\mu}_i^*$ in the reduced space. We will choose $l \le p$ so as to constrain S_i to lie within $B(\widehat{\mu}_i^*; \rho\sigma\sqrt{d}) \subseteq B(\mu_i^*; 4\rho\sigma\sqrt{d})$ (by the proof of Lemma 8, each center-estimate has p data points within a $\rho\sigma\sqrt{d}$ radius of it). The final estimate $\widehat{\mu}_i$ in $\mathbb{R}^n$ is the mean of S_i.

The random projection from $\mathbb{R}^n$ to $\mathbb{R}^d$ can be thought of as a composition of two linear transformations: a random rotation in $\mathbb{R}^n$ followed by a projection onto the first d coordinates. Since rotations preserve L_2 distance, we can assume, for the purpose of bounding the L_2 accuracy of our final estimates, that our random projection consists solely of a mapping onto the first d coordinates.

Think of the estimate $\widehat{\mu}_i$ as consisting of two parts: its first d coordinates, which constitute some low-dimensional vector close to $\widehat{\mu}_i^*$, and its remaining $n - d$ coordinates. We have already bounded the error on this first portion. How do we deal with the rest?

Let us fix attention on S_1. We would like it to be the case that this set consists primarily of points chosen from the first Gaussian $G_1 = N(\mu_1, \sigma^2 I_n)$. To this end, we establish the following

Definitions T_j = points in S_1 drawn from the j^{th} Gaussian G_j, and $l_j = |T_j|$. Let A_j be the mean of the points in T_j. And for $j > 1$ let $f_j = \|\mu_j^* - \mu_1^*\|/(\frac{1}{2}\sigma\sqrt{d}) \ge 1$. By the remark after Lemma 1, $\|\mu_j - \mu_1\| \le c f_j \sigma\sqrt{n}$.

We will show that S_1 is relatively uncontaminated by points from other Gaussians, that is, $l_2 + \cdots + l_k$ is small. Those points which do come from G_1 ought to average out to something near its mean μ_1.

Lemma 9 Randomly draw s points $Y_1, \ldots, Y_s$ from Gaussian $N(\mu, \sigma^2 I_n)$. Then for any $\Delta \ge \frac{1}{\sqrt{s}}$,

$$\mathbf{P}\left(\left\|\frac{Y_1 + \cdots + Y_s}{s} - \mu\right\| \ge \Delta\sigma\sqrt{n}\right) \le \left(\frac{e^{s\Delta^2 - 1}}{s\Delta^2}\right)^{-n/2}.$$

Proof. Let $X_i = \frac{Y_i - \mu}{\sigma} \sim N(0, I_n)$. The mean $\frac{1}{s}(X_1 + \cdots + X_s)$ has distribution $N(0, \frac{1}{s} I_n)$, and its squared L_2 norm has moment-generating function $\phi(t) = (1 - \frac{2t}{s})^{-n/2}$. By Markov's inequality,

$$\mathbf{P}\left(\left\|\frac{X_1 + \cdots + X_s}{s}\right\| \geq \Delta\sqrt{n}\right) \leq \left(\left(1 - \frac{2t}{s}\right) e^{2t\Delta^2}\right)^{-n/2} \leq \left(\frac{e^{s\Delta^2 - 1}}{s\Delta^2}\right)^{-n/2},$$

where the last bound is obtained by choosing $t = \frac{s}{2}(1 - \frac{1}{\Delta^2 s})$. ∎

Now we can at last dispatch the proof of Theorem 1.

Lemma 10 With probability $> 1 - \delta$, for all $1 \leq i \leq k$, $\|\widehat{\mu}_i - \mu_i\| \leq \epsilon\sigma\sqrt{n}$, provided that

$$\rho = \frac{\epsilon}{8}, \quad d \geq 10 \log \frac{40ce}{\epsilon w_{min}}, \text{ and } p \geq l \geq \frac{128}{\epsilon} \max\left\{\frac{1}{\epsilon}, \frac{c}{w_{min}} \log \frac{8k^2}{\delta}\right\}.$$

Proof. We will continue to focus upon μ_1. By Lemma 8, all of S_1 lies within distance $4\rho\sigma\sqrt{d}$ of μ_1^* and thus distance at least $(\frac{1}{2} - 4\rho) f_j \sigma\sqrt{d}$ from any other projected center μ_j^*. This implies that for a given point $x \in S_1$, and $j > 1$,

$$\mathbf{P}(x \text{ comes from } G_j) \leq \frac{w_j e^{-(\frac{1}{2} - 4\rho)^2 f_j^2 d/2}}{w_1 e^{-(4\rho)^2 d/2}} \mathbf{P}(x \text{ comes from } G_1) \leq \frac{w_j}{w_{min}} e^{-f_j^2 d/10}. \qquad (\dagger)$$

The mean of the points in S_1 is

$$\frac{l_1 A_1 + \cdots + l_k A_k}{l_1 + \cdots + l_k} = \frac{l_1}{l} A_1 + \sum_{j>1} \frac{l_j}{l} A_j.$$

Assume without loss of generality that $\mu_1 = 0$. We have already seen that the first d coordinates of the points in S_1 lie in $B(0; 4\rho\sigma\sqrt{d})$ and therefore contribute at most $4\rho\sigma\sqrt{d} \leq \frac{1}{2}\epsilon\sigma\sqrt{n}$ to $\|\text{mean}(S_1)\|$. We now turn to the remaining coordinates. Because the Gaussians are spherical, each point in S_1 from G_j looks like a random draw from $N(\mu_j; \sigma^2 I_n)$ as far as its last $n - d$ coordinates are concerned; that is, the values at these coordinates are not correlated with the first d coordinates. Thus we may pretend for our purposes that each point in S_1 is generated by the following process:

- Pick a Gaussian G_i, $1 \leq i \leq k$, according to the probability $(\dagger)$ above.
- Pick a random point from this Gaussian.

The L_2 distance between μ_1 and the mean of S_1, considering only the last $n - d$ coordinates, is then at most

$$\text{Error} \leq \|A_1\| + \sum_{j>1} (\|\mu_j\| + \|A_j - \mu_j\|) \frac{l_j}{l}.$$

We will bound these terms one at a time (bear in mind that we are still only talking about the last $n - d$ coordinates).

(a) For $j > 1$, we know $\|\mu_j\| \leq c f_j \sigma\sqrt{n}$. By Lemma 9, using $s = 1$,

$$\mathbf{P}(\exists j > 1 : l_j > 0 \text{ and } \|A_i - \mu_i\| > 2\sigma\sqrt{n}) \leq k e^{-n/2} < \frac{\delta}{8k}.$$

(b) The probability that a point in S_1 comes from G_j, $j > 1$, is $(\dagger)$ at most $\frac{w_j}{w_{min}} e^{-f_j^2 d/10} < \frac{w_j \epsilon}{40 c f_j^2}$. A quick Chernoff bound then tells us that

$$\mathbf{P}\left(\exists j > 1 : \frac{l_j}{l} > \frac{w_j \epsilon}{40 c f_j^2}(1+f_j)\right) \leq \frac{\delta}{8k}.$$

(c) We have already seen in (b) that that chance that a random point in S_1 is not from G_1 is at most $\frac{e^{-d/10}}{w_{min}} \leq \frac{1}{50}$. Therefore, with the assistance once again of a Chernoff bound, l_1 is at least $\frac{l}{2}$, with probability $> 1 - \frac{\delta}{8k}$. In which case, by Lemma 9, $\mathbf{P}(\|A_1\| \geq \epsilon\sigma\sqrt{n}/4) \leq e^{-n/2} \leq \frac{\delta}{8k}$.
(d) Putting all of these together, with probability $> 1 - \frac{\delta}{2k}$, and since $c \geq 1/2$,

$$\text{Error} \leq \frac{\epsilon\sigma\sqrt{n}}{4} + \sum_{j>1}(f_j c\sigma\sqrt{n} + 2\sigma\sqrt{n})\frac{w_j\epsilon}{40cf_j^2}(1+f_j) \leq \frac{\epsilon\sigma\sqrt{n}}{2},$$

as required. Repeating this for the remaining estimates $\widehat{\mu}_i$ then yields the theorem. ▮

Remark Assuming that $c > 1/2$ and $w_{min} = \Omega(1/k)$, the final choices of reduced dimension and sample complexity are

$$d = O\left(\log \frac{k}{\delta\epsilon}\right) \text{ and } M = k^{O(\log^2 \frac{1}{\delta\epsilon})}.$$

4 General Gaussians

4.1 Notation

The algorithm whose performance on spherical Gaussians we have just settled also works well with arbitrary covariance matrices. The proof of this general case is more involved, but follows approximately the same outline. We start with another battery of notation which builds upon the first.

Σ	Common $n \times n$ covariance matrix
σ_{max}	$\sqrt{\lambda_{max}(\Sigma)}$
σ_{min}	$\sqrt{\lambda_{min}(\Sigma)}$
ε	Eccentricity $\sigma_{max}/\sigma_{min}$
$\Sigma^*, \sigma^*_{max}, \sigma^*_{min}, \varepsilon^*$	Similar, but in the projected space
$\nu(\cdot)$	$N(0; I_d)$
$\nu_\sigma(\cdot)$	As before, $N(0; \sigma^2 I_d)$
$\nu_{\Sigma^*}(\cdot)$	$N(0; \Sigma^*)$
T	A useful linear transformation in $\mathbb{R}^d$
$\|\cdot\|_\Sigma$	Mahalanobis distance, $\|x\|_\Sigma = \sqrt{x^T\Sigma^{-1}x}$
$E(z; r; \Sigma)$	Ellipsoid $\{x : \|x - z\|_\Sigma \leq r\}$

We have already shown that $\sigma^*_{max} \leq \sigma_{max}$, $\sigma^*_{min} \geq \sigma_{min}$, and $\varepsilon^* \leq \varepsilon$. In fact, it will turn out that ε^* is a small constant even if ε is large (depending upon how much larger n is than d), and this will help us tremendously.

4.2 Crude density bounds

The dimensionality reduction lemma of the previous section applies to any mixture of Gaussians and hence needs no revision. The next step is to get bounds on the probability mass assigned to different spherical regions in the projected space.

Since the Gaussians we are now considering have ellipsoidal contours, it is not easy to get tight bounds on the chance that a point will fall in a given sphere. We will content ourselves with rather loose bounds, obtained via the mediation of a linear transformation T which converts ellipses into spheres.

Fix some $d \times d$ covariance matrix Σ^*, and write it as $B^T DB$, where B is orthogonal and D is diagonal with the eigenvalues of Σ^* as entries. Define $T = B^T D^{-1/2} B$; notice that T is its own transpose. The table below hints at the uses to which T will be put.

In $\mathbb{R}^d$ before T is applied	In $\mathbb{R}^d$ after T is applied
Gaussian $N(\mu^*; \Sigma^*)$	Gaussian $N(T\mu^*; I_d)$
Point x such that $\|x\|_{\Sigma^*} = r$	Point Tx such that $\|Tx\| = r$
Ellipse $E(z; r; \Sigma^*)$	Sphere $B(Tz; r)$

Our first step will be to relate the ellipsoidal density ν_{Σ^*} to the more manageable ν. As usual, we are interested in the probability mass assigned to spherical regions. Pick a particular $B(z; r)$ and define s to be $\|z\|_{\Sigma^*}$, the Σ^*-Mahalanobis distance from z to the origin. Since the standard deviation of Σ^* in different directions is constrained to lie in the range $[\sigma^*_{min}, \sigma^*_{max}]$, it is perfectly plausible that

$$\nu(B(s; r/\sigma^*_{max})) \leq \nu_{\Sigma^*}(B(z; r)) \leq \nu(B(s; r/\sigma^*_{min})).$$

This is proved in the following lemma, in a slightly different but equivalent form.

Lemma 11 (Relating ellipsoidal Gaussian density estimates to spherical ones) Pick any point z and any radius r. Writing $s = \|z\|_{\Sigma^*}$,

$$\nu_{\sigma^*_{max}}(B(s\sigma^*_{max}; r)) \leq \nu_{\Sigma^*}(B(z; r)) \leq \nu_{\sigma^*_{min}}(B(s\sigma^*_{min}; r)).$$

Proof. This is easy if T is used appropriately. For instance, because $E(z; r/\sigma^*_{max}; \Sigma^*) \subseteq B(z; r)$ we can write

$$\nu_{\Sigma^*}(B(z; r)) \geq \nu_{\Sigma^*}\left(E\left(z; \frac{r}{\sigma^*_{max}}; \Sigma^*\right)\right) = \nu_{\sigma^*_{max}}(B(s\sigma^*_{max}; r)),$$

where the final equality is a result of applying the transformation $\sigma^*_{max} T$. ▮

Similarly we can bound the relative densities of displaced spheres. Consider two spheres of equal radius r, one close to the center of the Gaussian, at Mahalanobis distance s, and the other at some distance $s + \Delta$. By how much must the probability mass of the closer sphere exceed that of the farther one, given that they may lie in different directions from the center? Although the spheres have equal radius, it might be the case that the closer sphere lies in a direction of higher variance than the farther sphere, in which case its radius is effectively scaled down. The following lemma gives a bound that will work for all spatial configurations of the spheres.

Lemma 12 Pick any point z and set $s = \|z\|_{\Sigma^*}$. If $\|z'\|_{\Sigma^*} \geq s + \Delta$ for some $\Delta > 0$ and if radius $r \leq s\sigma^*_{max}$ then

$$\frac{\nu_{\Sigma^*}(B(z; r))}{\nu_{\Sigma^*}(B(z'; r))} \geq \exp\left\{\frac{(\Delta + 2s)(\Delta - 2s\varepsilon^*)}{2}\right\}.$$

Proof. We will use the fact that Mahalanobis distance satisfies the triangle inequality and that $\|u\|_{\Sigma^*} \leq \|u\|/\sigma^*_{min}$. For any point x in $B(z; r)$,

$$\|x\|_{\Sigma^*} \leq \|z\|_{\Sigma^*} + \|x - z\|_{\Sigma^*} \leq s + \frac{r}{\sigma^*_{min}} \leq s + s\varepsilon^*,$$

where the last inequality follows from our restriction on r. Similarly, for any point x' in $B(z'; r)$,

$$\|x'\|_{\Sigma^*} \geq \|z'\|_{\Sigma^*} - \|x' - z'\|_{\Sigma^*} \geq s + \Delta - \frac{r}{\sigma^*_{min}} \geq \Delta - s(\varepsilon^* - 1).$$

Since $\nu_{\Sigma^*}(y)$ is proportional to $\exp(-\|y\|^2_{\Sigma^*}/2)$ for any point y, the ratio of probabilities of the two spheres must be at least

$$\frac{e^{-(s(1+\varepsilon^*))^2/2}}{e^{-(\Delta - s(\varepsilon^*-1))^2/2}} = \exp\left\{\frac{(\Delta - 2s\varepsilon^*)(\Delta + 2s)}{2}\right\},$$

as anticipated. ▮

Finally we need a bound on the rate at which the probability mass of the sphere $B(0; r)$ grows as its radius increases.

Lemma 13 If radii r and s satisfy $r + s \leq \frac{1}{2}\sigma^*_{min}\sqrt{d}$ then

$$\frac{\nu_{\Sigma^*}(B(0; r+s))}{\nu_{\Sigma^*}(B(0; r))} \geq \left(\frac{r+s}{r}\right)^{d/2}.$$

Proof. The proof for the spherical case can quite readily be adapted to this. Only one bound needs to be changed: for any $y \in B(0; r+s)$, we know $\|y\|_{\Sigma^*} \leq (r+s)/\sigma^*_{min}$ and so

$$\frac{\nu_{\Sigma^*}(y)}{\nu_{\Sigma^*}(y \cdot \frac{r}{r+s})} = \exp\left\{-\frac{\|y\|^2_{\Sigma^*}}{2}\left(1 - \frac{r^2}{(r+s)^2}\right)\right\} \geq \exp\left\{-\frac{(r+s)^2 - r^2}{2\sigma^{*2}_{min}}\right\} \geq \left(\frac{r}{r+s}\right)^{d/2},$$

given the condition on $r + s$. ▮

4.3 Estimating the projected means

The technical lemmas above allow us to approximately follow the outline of the spherical case. Denote by μ^*_i the means of the projected Gaussians and by Σ^* their common covariance matrix. Let π^* be the density of the projected mixture of Gaussians.

Lemma 14 If $\frac{p}{M} + \epsilon_0 \leq w_{min}\rho^d$ then for each i, there is at least one data point x in $E(\mu^*_i; \rho\sqrt{d}; \Sigma^*)$, and for any such x, at least p data points lie in $B(x; \rho\sigma^*_{max}\sqrt{d})$.

Proof. Since all the density estimates are accurate within ϵ_0, we need only show $w_{min}\nu_{\Sigma^*}(E(0; \rho\sqrt{d}; \Sigma^*)) \geq \epsilon_0$ and $w_{min}\nu_{\Sigma^*}(B(x; \rho\sigma^*_{max}\sqrt{d})) \geq \frac{p}{M} + \epsilon_0$ if $\|x\|_{\Sigma^*} \leq \rho\sqrt{d}$. Transformation T and Lemma 11 convert statements about ν_{Σ^*} into statements about ν; in particular,

$$\nu_{\Sigma^*}(E(0; \rho\sqrt{d}; \Sigma^*)) = \nu(B(0; \rho\sqrt{d})) \quad \text{and} \quad \nu_{\Sigma^*}(B(x; \rho\sigma^*_{max}\sqrt{d})) \geq \nu(B(\rho\sqrt{d}; \rho\sqrt{d})).$$

The rest follows from Lemma 3. ▮

Next we make sure that in the projected space, each estimated center is within Mahalanobis distance $(3\varepsilon^* + 1)\rho\sqrt{d}$ of its true value. The first step towards this is showing that data points which lie outside this range, and which are far away from the other Gaussians, have low density spheres around them. We start by defining a quantity which will be needed later.

Definition $F = 1 - \exp\left(-\frac{\varepsilon^*(3\varepsilon^*+2)\rho^2 d}{2}\right) - \frac{1}{w_{min}} \exp\left(-\frac{(\frac{1}{4}+\varepsilon^{*2}\rho)(\frac{1}{4}-\varepsilon^{*2}\rho-2\varepsilon^{*3}\rho)d}{2\varepsilon^{*2}}\right)$. By making sure that $d \geq 10\varepsilon^{*2} \log \frac{3}{w_{min}\rho^2}$ we ensure $F \geq \min\{\frac{1}{4}, \frac{3}{8}\varepsilon^{*2}\rho^2 d\}$.

Lemma 15 Suppose the ball $B(x;r)$ contains $\geq p$ points, for some radius $r \leq \rho\sigma^*_{max}\sqrt{d}$ and some point x such that (1) $\|x - \mu^*_i\|_{\Sigma^*} \geq (3\varepsilon^* + 1)\rho\sqrt{d}$; and (2) $\|x - \mu^*_j\| \geq \frac{1}{4\varepsilon^*}\sigma^*_{min}\sqrt{d}$ for all $j \neq i$.
Pick any point $z \in E(\mu^*_i; \rho\sqrt{d}; \Sigma^*)$; then

$$\pi^*(B(z;r)) \geq 2\epsilon_0 + \pi^*(B(x;r)),$$

provided $\epsilon_0 \leq \frac{F}{2-F}\frac{p}{M}$.

Proof. The conditions on x imply $\|x - \mu^*_j\|_{\Sigma^*} \geq \frac{1}{4\varepsilon^*}\frac{\sigma^*_{min}}{\sigma^*_{max}}\sqrt{d} \geq \frac{\sqrt{d}}{4\varepsilon^{*2}}$ for all $j \neq i$. Therefore, by Lemma 12,

$$\begin{aligned}
\pi^*(B(x;r)) &= w_i\nu_{\Sigma^*}(B(x-\mu^*_i;r)) + \sum_{j\neq i} w_j\nu_{\Sigma^*}(B(x-\mu^*_j;r)) \\
&\leq w_i\nu_{\Sigma^*}(B(z-\mu^*_i;r))e^{-((3\varepsilon^*+2)\rho\sqrt{d})(\varepsilon^*\rho\sqrt{d})/2} \\
&\quad + \nu_{\Sigma^*}(B(z-\mu^*_i;r))e^{-(((1/4\varepsilon^{*2})+\rho)\sqrt{d})(((1/4\varepsilon^{*2})-\rho-2\varepsilon^*\rho)\sqrt{d})/2} \\
&\leq w_i\nu_{\Sigma^*}(B(z-\mu^*_i;r))(1-F) \\
&\leq \pi^*(B(z;r))(1-F)
\end{aligned}$$

The rest follows along the lines of Lemma 6. ▮

After estimating a center in the projected space, we must eliminate from S' all the high-density points in its vicinity. We will simply pick the q points closest to it, and guarantee that this includes at least the central $B(0; \frac{1}{4\varepsilon^*}\sigma^*_{min}\sqrt{d})$ of the Gaussian and nothing outside the central $B(0; \frac{1}{2\varepsilon^*}\sigma^*_{max}\sqrt{d})$. This time round we adopt the following

Definitions $p = Mw_{min}\rho^d(1 - \frac{F}{2})$, $\epsilon_0 = \frac{p}{M}\frac{F}{2-F}$, and $q = w_{min}\nu(B(0; \frac{3}{8\varepsilon^*}\sqrt{d}))M$. As before q can easily be computed, given d/ε^{*2}.

Lemma 16 Pick any point x for which $\|x - \mu^*_i\|_{\Sigma^*} \leq \rho(3\varepsilon^* + 1)\sqrt{d}$. Then
(a) $\pi^*(B(x; (\frac{1}{4\varepsilon^*} + \rho\varepsilon^*(3\varepsilon^*+1))\sigma^*_{min}\sqrt{d})) \leq \frac{q}{M} - \epsilon_0$; and
(b) $\pi^*(B(x; (\frac{1}{2\varepsilon^*} - \rho(3\varepsilon^*+2))\sigma^*_{max}\sqrt{d})) \geq \frac{q}{M} + \epsilon_0$,
provided that $\epsilon_0 \leq \frac{q}{2M}$, $\rho \leq \frac{1}{96\varepsilon^{*3}}$, and $d \geq 8\log\frac{2}{w_{min}}$.
As a consequence, the q points closest to x include all data points within distance $\frac{1}{4\varepsilon^*}\sigma^*_{min}\sqrt{d}$ of μ^*_i and no data point which is more than $(\frac{1}{2\varepsilon^*} - \rho)\sigma^*_{max}\sqrt{d}$ away from μ^*_i.

Proof. We rewrite q as $Mw_{min}\nu_{\Sigma^*}(E(0; \frac{3}{8\varepsilon^*}\sqrt{d}; \Sigma^*))$ and notice that

$$w_{min}\nu_{\Sigma^*}(B(0; \tfrac{3}{8\varepsilon^*}\sigma^*_{min}\sqrt{d})) \leq \frac{q}{M} \leq w_{min}\nu_{\Sigma^*}(B(0; \tfrac{3}{8\varepsilon^*}\sigma^*_{max}\sqrt{d})).$$

Statement (a) would follow from

$$\nu_{\Sigma^*}(B(0; (\tfrac{1}{4\varepsilon^*} + \rho\varepsilon^*(3\varepsilon^*+1))\sigma^*_{min}\sqrt{d})) \leq \frac{w_{min}}{2}\nu_{\Sigma^*}(B(0; \tfrac{3}{8\varepsilon^*}\sigma^*_{min}\sqrt{d})),$$

and for (b) it is enough to check that

$$w_{min}\nu_{\Sigma^*}(B(0;(\tfrac{1}{2\varepsilon^*}-\rho(6\varepsilon^*+3))\sigma^*_{max}\sqrt{d}))\geq\frac{3w_{min}}{2}\nu_{\Sigma^*}(B(0;\tfrac{3}{8\varepsilon^*}\sigma^*_{max}\sqrt{d})).$$

These are direct consequences of Lemma 13 and the stated conditions. ▮

Remark Assume henceforth that the various parameters (p, q, M, d, ρ) are set in accordance with the specifications above.

The proof of Lemma 8 continues to be valid in this more general case, and gives us

Lemma 17 With probability $> 1 - \delta/2$, for every $i \leq k$, $\|\widehat{\mu}_i^* - \mu_i^*\|_{\Sigma^*} \leq (3\varepsilon^* + 1)\rho\sqrt{d}$.

4.4 Back in high-dimensional space

The random projection from $\mathbb{R}^n$ to $\mathbb{R}^d$ can be thought of as a composition of two transformations: a random rotation in $\mathbb{R}^n$ followed by a projection onto the first d coordinates. Since rotations preserve L_2 distance, and our purpose is to bound the accuracy of our center estimates in terms of L_2 distance, we will assume for the next few lemmas that the random projection consists solely of selecting the first d coordinates. We will write high-dimensional points in the form $(x, y) \in \mathbb{R}^d \times \mathbb{R}^{n-d}$, and will assume that each such point is projected down to x.

The covariance matrix Σ can be written in the form

$$\Sigma = \left(\begin{array}{cc} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{array}\right),$$

with $\Sigma_{xx} = \Sigma^*$ being the covariance matrix of the projected Gaussians. What is the correlation between the x and y components of points drawn from Gaussians with covariance Σ?

Fact If a point drawn from $N(0;\Sigma)$ has x as its first d coordinates, then its last $n-d$ coordinates have the distribution $N(Ax; C)$, where $A = \Sigma_{yx}\Sigma_{xx}^{-1}$ and $C = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$. This well-known result can be found, for instance, in Lauritzen's (1996) book on graphical models.

We will need to tackle the question: for a point (x, y) drawn from $N(0;\Sigma)$, what is the expected value of $\|y\|$ given $\|x\|$? In order to answer this, we need to study A a bit more carefully.

Lemma 18 $\|Ax\| \leq \sigma_{max}\|x\|_{\Sigma^*}\sqrt{n/d}$ for any $x \in \mathbb{R}^d$.

Proof. $A = \Sigma_{yx}\Sigma_{xx}^{-1}$ is a $(n-d)\times d$ matrix; divide it into $\frac{n}{d}-1$ square matrices $B_1, \ldots, B_{n/d-1}$ by taking d rows at a time. Fix attention on one such B_i. The rows of B_i correspond to some d consecutive coordinates of y; call these coordinates z. Then we can write $B_i = \Sigma_{zx}\Sigma_{xx}^{-1}$. It is well-known – see, for instance, the textbook by Horn and Johnson (1985), or consider the inverse of the $2d \times 2d$ positive definite covariance matrix of (z, x) – that $(\Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx})$ is positive definite. Therefore, for any $u \in \mathbb{R}^d$,

$$u^T\Sigma_{xx}u > u^T\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}u.$$

For any $v \in \mathbb{R}^d$, choose $u = \Sigma_{xx}^{-1}v$ so that

$$\|v\|^2_{\Sigma^*} = v^T\Sigma_{xx}^{-1}\Sigma_{xx}\Sigma_{xx}^{-1}v = u^T\Sigma_{xx}u > v^TB_i^T\Sigma_{zz}^{-1}B_iv \geq \frac{\|B_iv\|^2}{\lambda_{max}(\Sigma_{zz})} \geq \frac{\|B_iv\|^2}{\sigma^2_{max}}.$$

Therefore $\|B_iv\| \leq \sigma_{max}\|v\|_{\Sigma^*}$. The pieces now come neatly together,

$$\|Ax\|^2 = \|B_1x\|^2 + \cdots + \|B_{n/d-1}x\|^2 \leq \tfrac{n-d}{d}\sigma^2_{max}\|x\|^2_{\Sigma^*},$$

and the lemma is proved. ▮

Define S_i, G_j, T_j, A_j, f_j and l_j as in the spherical case (for the definition of f_j use σ^*_{max} in lieu of σ), and focus attention upon S_1. The y coordinates of points in $T_j \subset S_1$ look roughly like random draws from the distribution $N(A(\widehat{\mu}^*_1 - \mu^*_j); C)$. Can we bound their average?

Lemma 19 Assume $\rho \leq \frac{\epsilon}{20\varepsilon^{*2}}$ and $l \leq p$. For any $j \geq 1$, $A_j - \mu_j$ has the same distribution as $(X, AX + C^{1/2}E_{l_j})$, where X is a random variable with $\|X\| \leq \|\mu^*_1 - \mu^*_j\| + \frac{\epsilon}{4}\sigma^*_{min}\sqrt{d}$, and E_m is the mean of m i.i.d. $N(0; I_{n-d})$ random variables.

Proof. Assume for the sake of convenience that μ_j is zero. In the low-dimensional space, forcing $l \leq p$ guarantees that all of S_1 lies within $\rho\sigma^*_{max}\sqrt{d}$ of $\widehat{\mu}^*_1$, and therefore within $\rho(3\varepsilon^* + 2)\sigma^*_{max}\sqrt{d} \leq 5\varepsilon^*\rho\sigma^*_{max}\sqrt{d} \leq \frac{\epsilon}{4}\sigma^*_{min}\sqrt{d}$ of μ^*_1.

Recall that T_j consists of those points in S_1 which come from Gaussian G_j. For our purposes, we can pretend that each point $(X_i, Y_i) \in T_j$ is generated in the following fashion:
- Pick $X_i \in B(\mu^*_1; \frac{\epsilon}{4}\sigma^*_{min}\sqrt{d}) \subset \mathbb{R}^d$, according to an unknown distribution.
- Choose $Y_i \sim N(AX_i; C)$.
In this manner we choose l_j points $\{(X_i, Y_i)\}$, with mean value some (X, Y). The range of the X_i coordinates constrains $\|X\|$ to be at most $\|\mu^*_1 - \mu^*_j\| + \frac{\epsilon}{4}\sigma^*_{min}\sqrt{d}$. To understand the distribution of Y, we notice $(Y_i - AX_i) \sim C^{1/2}N(0, I_{n-d})$, and taking averages, $Y \sim AX + C^{1/2}E_{l_j}$. ▮

Armed with this result we can finally rework the last lemma of the spherical case.

Lemma 20 With probability $> 1 - \delta$, for all $1 \leq i \leq k$, $\|\widehat{\mu}_i - \mu_i\| \leq \epsilon\sigma_{max}\sqrt{n}$, provided that

$$d \geq 12 \ln \frac{64c^2\varepsilon^{*2}}{\epsilon w_{min}}, \quad \text{and} \quad l \geq \max\left\{\frac{48}{\epsilon^2}, \frac{48}{\epsilon w_{min}} \ln \frac{4k^2}{\delta}\right\}.$$

Proof. We observed in the previous lemma that in low dimension, all of S_1 lies within $5\varepsilon^*\rho\sigma^*_{max}\sqrt{d}$ of μ^*_1, and therefore at distance at least $(\frac{1}{2} - 5\varepsilon^*\rho)f_j\sigma^*_{max}\sqrt{d}$ from any other projected center μ^*_j.

Fix any point $x \in S_1$, and any $j > 1$. Applying the general principle that $\frac{\|u\|}{\sigma^*_{max}} \leq \|u\|_{\Sigma^*} \leq \frac{\|u\|}{\sigma^*_{min}}$, we then know $\|x - \mu^*_1\|_{\Sigma^*} \leq 5\varepsilon^{*2}\rho\sqrt{d}$ and $\|x - \mu^*_j\|_{\Sigma^*} \geq (1/2 - 5\varepsilon^*\rho)f_j\sqrt{d}$ and therefore

$$\mathbf{P}(x \text{ comes from } G_j) \leq \frac{w_j e^{-(\frac{1}{2} - 5\varepsilon^*\rho)^2 f_j^2 d/2}}{w_1 e^{-(5\varepsilon^{*2}\rho)^2 d/2}} \mathbf{P}(x \text{ comes from } G_1) \leq \frac{w_j}{w_{min}} e^{-\frac{f_j^2 d}{12}}. \qquad (\dagger\dagger)$$

The difference between μ_1 and the mean of S_1, which we hope is close to zero, is given by

$$\text{mean}(S_1) - \mu_1 \quad = \quad \left(\sum_{j=1}^{k} A_j \cdot \frac{l_j}{l}\right) - \mu_1 \quad = \quad \sum_{j=1}^{k}(A_j - \mu_j)\frac{l_j}{l} + \sum_{j=2}^{k}(\mu_j - \mu_1)\frac{l_j}{l}.$$

The previous two lemmas immediately bound the L_2 norm of this expression by

$$\begin{aligned} \text{Error} \quad &\leq \quad \sum_{j=1}^{k} \frac{l_j}{l}\left\{\left(\|\mu^*_j - \mu^*_1\| + \frac{\epsilon}{4}\sigma^*_{min}\sqrt{d}\right)\left(1 + \frac{\sigma_{max}}{\sigma^*_{min}}\sqrt{\frac{n}{d}}\right)\right\} + \|C^{1/2}E_l\| + \sum_{j=2}^{k}\|\mu_j - \mu_1\|\frac{l_j}{l} \\ &\leq \quad \|C^{1/2}E_l\| + \frac{\epsilon\sigma_{max}\sqrt{n}}{2} + \left(\sum_{j>1} 8c\varepsilon^* f_j \cdot \frac{l_j}{l}\right)\sigma_{max}\sqrt{n}, \end{aligned}$$

where E_l is, as before, the mean of l i.i.d. $N(0; I_{n-d})$ random variables. We'll bound these terms one at a time.
(a) Since $C = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$ and each of these two right-hand terms is positive semidefinite, $\lambda_{max}(C) \leq \lambda_{max}(\Sigma_{yy}) \leq \sigma_{max}^2$ and therefore $\|C^{1/2}E_l\| \leq \sigma_{max}\|E_l\|$. Lemma 9 assures us that if $l \geq \frac{48}{\epsilon^2}$ then $\mathbf{P}(\|E_l\| > \frac{\epsilon}{4}\sqrt{n}) \leq e^{-n/4} \leq \frac{\delta}{4k}$.
(b) The probability that a point in S_1 comes from G_j is (††) at most $\frac{w_j}{w_{min}} e^{-f_j^2 d/12} < \frac{w_j \epsilon}{64 c^2 \varepsilon^{*2} f_j^2}$. A simple Chernoff bound guarantees that:

$$\mathbf{P}\left(\exists j > 1 : \frac{l_j}{l} > \frac{w_j \epsilon}{16 c \varepsilon^* f_j}\right) \leq \frac{\delta}{4k},$$

given the condition on l.

The lemma follows by applying these two bounds to the error expression. ∎

Remark If $w_{min} = \Omega(\frac{1}{k})$ then we need to use reduced dimension $d = O(\varepsilon^{*2} \log \frac{k}{\epsilon\delta})$ and sample size $M = k^{O(\varepsilon^{*2} \log^2 1/\epsilon\delta)}$.

4.5 Bounding the eccentricity of projected ellipsoids

Our algorithm works best when the projected Gaussians have eccentricity close to one. We will now see that even if the original Gaussians are highly skewed, random projection will make them almost spherical.

Once again, think of the random projection as a random rotation in $\mathbb{R}^n$, represented by some orthogonal matrix U^T, followed by a projection P^T onto the first d coordinates. The high-dimensional covariance matrix Σ has positive eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$, with eccentricity $\varepsilon = \lambda_n/\lambda_1 \geq 1$ and average variance $\lambda = \frac{1}{n}(\lambda_1 + \cdots + \lambda_n)$.

Pick any unit vector $x \in \mathbb{R}^d$, and let $V(x) = x^T \Sigma^* x$ be the variance of the projected Gaussians in direction x. We will show that Σ^* is close to the spherical covariance matrix λI_d by proving $V(x) \approx \lambda$ for all directions x.

Lemma 21 For any unit vector $x \in \mathbb{R}^d$, $V(x)$ has the same distribution as $\sum_{i=1}^n \lambda_i v_i^2$, where v is chosen uniformly at random from the surface of the unit sphere in $\mathbb{R}^n$. Therefore $\mathbf{E}V(x) = \lambda$, over the choice of random projection.

Proof. Let the $d \times n$ matrix P^T represent projection onto the first d coordinates. Then $\Sigma^* = (UP)^T \Sigma (UP)$, and on account of U we may assume Σ is diagonal, specifically $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_n)$. For any direction $x \in \mathbb{R}^d$, $V(x) = x^T \Sigma^* x = (Px)^T (U^T \Sigma U)(Px)$. Since Σ is diagonal,

$$(U^T \Sigma U)_{ij} = \sum_{k=1}^n \lambda_k U_{ki} U_{kj}$$

whereby

$$\begin{aligned} V(x) &= \sum_{i,j=1}^n (Px)_i (Px)_j (U^T \Sigma U)_{ij} \\ &= \sum_{i,j=1}^d x_i x_j \sum_{k=1}^n \lambda_k U_{ki} U_{kj} \quad = \quad \sum_{k=1}^n \lambda_k \left(\sum_{i=1}^d x_i U_{ki}\right)^2 . \end{aligned}$$

We can without loss of generality assume that x lies along some coordinate axis, say the very first one, in which case

$$V(x) = \sum_{i=1}^{n} \lambda_i U_{i1}^2.$$

Since U^T is a random orthogonal matrix, its first row $(U_{11}, \ldots, U_{n1})$ is a random unit vector. ∎

We now have a simple formulation of the distribution of $V(x)$. For any given x, this value is likely to be close to its expectation because it is the sum of n almost-independent bounded random variables. To demonstrate $V(x) \approx \lambda$ simultaneously for all vectors x on the unit sphere in $\mathbb{R}^d$, we will prove uniform convergence for a carefully chosen finite cover of this sphere.

Lemma 22 For any $0 < \epsilon \leq 1$, if $n > O(\frac{\varepsilon^2}{\epsilon^2}(\log \frac{1}{\delta} + d \log \frac{d}{\epsilon}))$, then with probability $> 1-\delta$, the eccentricity ε^* of the projected covariance matrix is at most $1+\epsilon$. In particular, if the high-dimensional eccentricity ε is at most $O(\frac{n^{1/2}}{\log k/\epsilon\delta})$ then with probability at least $1-\delta$, the projected Gaussians have eccentricity $\varepsilon^* \leq 2$.

Proof. By considering moment-generating functions of various gamma distributions as in Lemma 9, we can show that for any particular x and any $\epsilon \in (0,1)$,

$$\mathbf{P}(|V(x) - \lambda| > \epsilon\lambda) \leq \exp\left(-\Omega\left(n\epsilon^2/\varepsilon^2\right)\right).$$

Moreover, $V(y)$ cannot differ too much from $V(x)$ when y lies close to x: using the expression for $V(x)$ found in the previous lemma, with u_i^* as shorthand for $(U_{i1}, \ldots, U_{id})$,

$$\begin{aligned}
|V(x) - V(y)| &\leq \sum_{i=1}^{n} \lambda_i \left|(u_i^* \cdot x)^2 - (u_i^* \cdot y)^2\right| \\
&= \sum_{i=1}^{n} \lambda_i \left|u_i^* \cdot (x+y)\right| \cdot \left|u_i^* \cdot (x-y)\right| \\
&\leq \sum_{i=1}^{n} \lambda_i \, \|u_i^*\|^2 \cdot \|x+y\| \cdot \|x-y\| \\
&\leq 2\,\|x-y\| \left(\sum_{i=1}^{n} \lambda_i \|u_i^*\|^2\right).
\end{aligned}$$

The final parenthesized quantity can be shown to be close to its expectation $d\lambda$ (perhaps we should point out that $\mathbf{E}\|u_i^*\|^2 = \frac{d}{n}$ since u_i^* consists of the first d coordinates of a random unit vector in $\mathbb{R}^n$). Choosing $\|x-y\| \leq O(\frac{\epsilon}{d})$ will then ensure $|V(x) - V(y)| \leq \epsilon\lambda$.

Bounding $V(x)$ effectively bounds $V(y)$ for $y \in B(x; O(\frac{\epsilon}{d}))$. How many points x must be chosen to cover the unit sphere in this way? A geometric argument – see, for instance, Gupta (1999) – shows that $(O(\frac{d}{\epsilon}))^d$ points will do the trick, and completes the proof. ∎

5 In future

We have described an extremely simple and provably correct algorithm for learning the centers of an unknown mixture of Gaussians with shared covariance matrix. This core combinatorial problem having been solved, we will in a companion paper examine some practical issues that arise in the common use of such mixture models. We will show how to estimate the mixing weights and covariance matrix, so as to permit the computation of likelihoods, and then discuss experiments which compare our algorithm to three alternatives: EM by itself, EM with principal component analysis, and a promising new option, EM preceded by random projection.

What important theoretical questions remain?

1. Our algorithm will work when different clusters have differing covariances, provided these matrices have approximately the same trace. Can this qualification be removed so that arbitrary mixtures of Gaussians can be learned?

2. We are able to learn a mixture of k Gaussians within precision ϵ using $k^{O(\log 1/\epsilon^2)}$ data points. Is it possible to improve this sample complexity to just $(\frac{k}{\epsilon})^{O(1)}$, through the clever use of some heuristic like "agglomerative clustering" (Duda & Hart)? A probabilistic analysis of such clustering techniques is long overdue.

3. What happens when the data do not come from a mixture of Gaussians? Our algorithm has to accommodate sampling error and therefore it will perform well on clusters which are close to Gaussian. In more general situations, the problem of finding the centers is of course no longer well-defined. However, Diaconis and Freedman (1984) have shown, roughly, that many natural distributions in high dimension look approximately Gaussian when projected onto a random line. This might make it possible to use our algorithm to cluster data from quite generic non-Gaussian mixture distributions: randomly project the data into a subspace, learn the resulting mixture of almost-Gaussians, and then apply this clustering to the high-dimensional data!

Acknowledgements

The author profusely thanks Peter Bickel, Yoav Freund, Nir Friedman, Anupam Gupta, Michael Jordan, Christos Papadimitriou, Stuart Russell, and Umesh Vazirani.

Literature cited

Dasgupta, S. & Gupta, A. (1999) An elementary proof of the Johnson-Lindenstrauss lemma. Technical Report 99-006, International Computer Science Institute, Berkeley.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum-likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B*, **39**:1-38.

Diaconis, P. & Freedman, D. (1984) Asymptotics of graphical projection pursuit. *Annals of Statistics*, **12**:793-815.

Duda, R.O. & Hart, P.E. (1973) *Pattern Classification and Scene Analysis.* John Wiley, New York.

Dudley, R.M. (1979). Balls in R^k do not cut all subsets of $k + 2$ points. *Advances in Mathematics*, **31**:306-308.

Frankl, P. & Maehara, H. (1988) The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory Ser. B*, **44**:355-365.

Gupta, A. (1999) Embedding tree metrics into low dimensional Euclidean spaces. To appear in *ACM Symposium on Theory of Computing*.

Horn, R.A. & Johnson, C.R. (1985) *Matrix Analysis.* Cambridge University Press.

Johnson, W.B. & Lindenstrauss, J. (1984) Extensions of Lipschitz mapping into Hilbert space. *Contemp. Math.*, **26**:189-206.

Lauritzen, S. (1996). *Graphical models.* Oxford: Oxford University Press.

Lindsay, B. (1995) *Mixture Models: Theory, Geometry, and Applications.* American Statistical Association, Virginia.

Pach, J. & Agarwal, P. (1995) *Combinatorial Geometry.* John Wiley, New York.

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

Titterington, D.M., Smith, A.F.M. & Makov, U.E. (1985) *Statistical Analysis of Finite Mixture Distributions.* John Wiley, New York.