Learning Mixtures of Gaussians

Sanjoy Dasgupta
University of California, Berkeley

Abstract

Mixtures of Gaussians are among the most fundamental and widely used statistical models. Current techniques for learning such mixtures from data are local search heuristics with weak performance guarantees. We present the first provably correct algorithm for learning a mixture of Gaussians. This algorithm is very simple and returns the true centers of the Gaussians to within the precision specified by the user, with high probability. It runs in time only linear in the dimension of the data and polynomial in the number of Gaussians.

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 these 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, spanning fields as varied as psychology, geology, and astrophysics, has been compiled by Titterington, Smith, and Makov (1985). Their authoritative 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 amongst them the most popular appears to be the EM algorithm formalized by Dempster, Laird, and Rubin (1977). EM is a local search heuristic of appealing simplicity. Its principal goal is convergence to a local maximum in the space of Gaussian mixtures ranked by likelihood. An explanation of this algorithm, along with helpful remarks about its performance in learning mixtures of univariate Gaussians, can be found in an excellent survey article by Redner and Walker (1984) and in a recent monograph by Lindsay (1995).

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 more spherical and thereby more manageable. In the second phase, the modes of the low-dimensional distribution are found using a simple new clustering algorithm whose performance we rigorously analyze. Finally, the low-dimensional modes are used to reconstruct the original centers. Each of these stages invokes new technical tools of more general applicability.

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 most 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 $\sigma\sqrt{n}$.

The more general Gaussian $N(0;\Sigma)$ has ellipsoidal contours of equal density. Each such ellipsoid is of the form $\{x : x^T\Sigma^{-1}x = r^2\}$, corresponding to points at a fixed *Mahalanobis distance* $\|x\|_\Sigma = \sqrt{x^T\Sigma^{-1}x}$ from the center of the Gaussian. As in the spherical case, in high dimension the distribution is concentrated around an ellipsoidal shell $\|x\|_\Sigma \approx \sqrt{n}$. The reader should try to reconcile this with the fact that the distribution is also concentrated (perhaps less tightly) around a spherical shell $\|x\| \approx \sqrt{\text{trace}(\Sigma)}$.

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 $\leq \frac{1}{2}\sigma\sqrt{n}$ 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, or PCA. 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.

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 $d = O(\frac{\log M}{\epsilon^2})$ 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*, we can reduce the dimension of the data far more drastically. 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, where k is the number of Gaussians. 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 well-separated remain so when they are projected. These conflicting requirements are accommodated admirably by a projection to just $O(\log k)$ dimensions.

This method of projection has another tremendous benefit: we show that even if the original Gaussians are highly skewed (have ellipsoidal contours of high eccentricity), their projected counterparts will be more spherical and thereby easier to learn! The low-dimensional portion of our algorithm is able to take advantage of this; it works for Gaussians of arbitrary eccentricity, but affords the best guarantees for spherical Gaussians.

2.4 The algorithm

We are now in a position to present the algorithm. The user furnishes: ϵ, the accuracy within which the centers are to be learned; δ, a confidence parameter; k, the number of Gaussians; 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 $x \in S$, 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 $i = 1 \ldots k$:
     - 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.
     - Remove these points from S'.
   - For each i, let S_i denote 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 Spherical density estimates

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 – that of 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. That is, 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 $d + 1$.

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.

The chief technical problem in the reconstruction is to show that small errors in the estimate $\widehat{\mu}_i^*$ are not grossly magnified when carried back into $\mathbb{R}^n$. The core question can be stated quite simply. Given that an unknown point $x \in \mathbb{R}^n$ drawn from Gaussian $N(0;\Sigma)$ gets projected to some $y \in \mathbb{R}^d$, what is the conditional distribution of $||x||$ given $||y||$? A bit of matrix analysis yields the answer.

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 result

In the next section we will prove a dimensionality reduction lemma, the first step towards our main

Theorem Suppose data is drawn from a mixture of k Gaussians in $\mathbb{R}^n$ which is c-separated, for $c > {}^1\!/\!_2$; has smallest mixing weight $\Omega(\frac{1}{k})$; and has (unknown) common covariance matrix Σ with maximum and minimum eigenvalues $\sigma^2_{max}, \sigma^2_{min}$ 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 $M = k^{O(\log^2 1/(\epsilon\delta))}$. The algorithm runs in time $O(M^2 d + M dn)$.

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 this theorem. The mixing weights and eccentricity are similarly unrestricted.

A word about the inputs: in addition to the number of Gaussians k and 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.

In the last section of this paper, we will discuss how the mixing weights and covariance matrix may be estimated, if these are needed. We will also suggest ideas for reducing the sample complexity to $k^{O(\log 1/\delta)}/\epsilon^2$, and for handling more general families of distributions.

3 Reducing dimension

3.1 Maintaining intercluster distances

We start by showing that the dimension of the data can be reduced drastically without significantly increasing the overlap of the clusters.

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, which we parametrize by mixing weights w_i, means μ_i and covariance matrices Σ_i, one per Gaussian. 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 $\delta > 0$ and $\epsilon > 0$ designate confidence and accuracy parameters, respectively. With probability $> 1-\delta$, the projection of this mixture of Gaussians onto a random d-dimensional subspace yields a $(c\sqrt{1-\epsilon})$-separated mixture of Gaussians $\{(w_i, \mu_i^*, \Sigma_i^*)\}$ in $\mathbb{R}^d$, provided $d \geq \frac{4}{\epsilon^2} \ln \frac{k^2}{\delta}$. 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 sketch. 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).

Apply this lemma to the $O(k^2)$ line segments joining pairs of Gaussian centers in the original space. This keeps the centers far apart; to satisfy our definition of separatedness, we must also check that the original Gaussians do not spread out when projected, that is, $\lambda_{max}(\Sigma_i^*) \leq \lambda_{max}(\Sigma_i)$. ∎

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 $(cf\sqrt{1-\epsilon})$-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 next section.

Corollary If $c > {}^1\!/\!_2$, then in order to ensure that the projected mixture is at least ${}^1\!/\!_2$-separated with probability $> 1-\delta$, it is enough to choose $d \geq \frac{4c^4}{(c^2-\frac{1}{4})^2} \ln \frac{k^2}{\delta}$.

3.2 Bounding the eccentricity of projected ellipsoids

The low-dimensional phase of our algorithm works best when the projected Gaussians have eccentricity close to one. We will now see that random projection makes Gaussians more spherical.

Think of the random projection from $\mathbb{R}^n$ to $\mathbb{R}^d$ 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 columns of U^T are an orthonormal basis $\{u_1, \ldots, u_n\}$ of $\mathbb{R}^n$. Denote the restriction of these vectors to their first d coordinates by $u_1^*, \ldots, u_n^*$, respectively. The high-dimensional covariance matrix Σ has eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$, with eccentricity $\varepsilon = \sqrt{\lambda_n/\lambda_1} \geq 1$, and normalized trace $\lambda = \frac{1}{n}(\lambda_1 + \cdots + \lambda_n)$. We will show that the covariance matrix of the projected Gaussians, denoted Σ^*, is close to the spherical covariance matrix λI_d.

Pick any unit vector $x \in \mathbb{R}^d$, and define $V(x)$ to be the variance of the projected Gaussian in direction x.

Lemma 2 (Variance of projected Gaussians) 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. We can write the projected covariance matrix Σ^* as $(UP)^T\Sigma(UP)$, and on account of U we may assume Σ is diagonal, specifically $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_n)$.

Pick any direction $x \in \mathbb{R}^d$. The variance of the projected Gaussian in direction x is $V(x) = x^T\Sigma^* x = (Px)^T(U^T\Sigma U)(Px)$. Since Σ is diagonal,

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

whereby

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

where u_k^* denotes the first d coordinates of the k^{th} row of U.

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 3 (Eccentricity reduction) 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 sketch. By considering moment-generating functions of various gamma distributions, we can show that for any particular x and any $\epsilon \in (0, 1)$, $\mathbf{P}(|V(x) - \lambda| > \epsilon\lambda) \leq e^{-\Omega(n\epsilon^2/\varepsilon^2)}$.

Moreover, $V(y)$ cannot differ too much from $V(x)$ when y lies close to x:

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

The final parenthesized quantity will with high probability 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. ∎

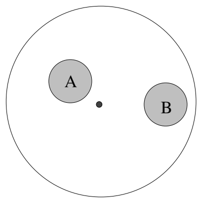
4 Low-dimensional clustering

4.1 Technical overview

Our algorithm for learning the centers of Gaussians in low dimension is one of many that could be used. The technical tools used in its analysis might be helpful in developing other similar routines, and we therefore give a brief overview of them.

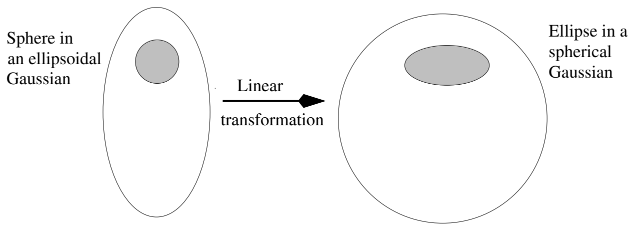
- Using VC bounds, it can be shown that with just $O(d)$ samples, all spheres in $\mathbb{R}^d$ will contain roughly the correct number of points, that is, the expected number under the mixture distribution. This is a convenient and very strong guarantee; we need no other control on sampling error.

- Assume that the Gaussians are spherical. Each point x in the sample is assigned a radius r_x, and we hope that points with low r_x will be close to the centers of the Gaussians. In order to prove this, we must show that in cases such as that depicted below (where the outer sphere conceptually denotes a Gaussian), sphere A has a significantly higher probability mass than sphere B, which has the same radius but is further from the center of the Gaussian. This can be shown easily by a point-wise coupling of A and B.



- Assume the Gaussians are spherical with unit variance. Once a center μ_i^* has been chosen, we will eliminate the q points closest to it, where q is the number of points expected to fall within $B(\mu_i^*; {}^3\!/_8\sqrt{d})$, assuming a mixing weight of w_{min}. It turns out that whatever w_i might actually be, this will eliminate all points in $B(\mu_i^*; {}^1\!/_4\sqrt{d})$ and nothing that does not lie in $B(\mu_i^*; {}^1\!/_2\sqrt{d})$. In effect, it eliminates all the high-density points in the i^{th} Gaussian while leaving intact high-density regions of other Gaussians.

- These arguments seem most naturally suited to spherical Gaussians. They all involve obtaining upper and lower bounds on the probability masses of spherical regions in $\mathbb{R}^d$. In order to extend this to ellipsoidal Gaussians, we use a simple linear transformation which maps a sphere contained in an ellipsoidal Gaussian to an ellipsoid contained in a spherical Gaussian. Bounds on the probability mass of this latter ellipsoid are then obtained by considering its inscribed and circumscribed spheres. These bounds are acceptable because the projected Gaussians have small eccentricity.



4.2 Notation

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

ϵ, δ	Accuracy and 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
$w_i N(\mu_i; \Sigma)$	A mixture component (Gaussian) in $\mathbb{R}^n$
w_{min}	Lower bound on the w_i, supplied by user
c, c^*	Separation of Gaussians in $\mathbb{R}^n, \mathbb{R}^d$
$w_i N(\mu_i^*, \Sigma^*)$	Projection of i^{th} Gaussian into $\mathbb{R}^d$
$\pi^*(\cdot)$	Density of the entire projected mixture
$B(x; r)$	Sphere of radius r centered at x
$B(r'; r)$	$B(x; r)$ for some x with $\|x\| = r'$
l, p, q	Integer parameters needed by algorithm
ρ	Parameter needed for analysis, related to ϵ
$\sigma_{max}, \sigma_{min}$	$\sqrt{\lambda_{max}(\Sigma)}, \sqrt{\lambda_{min}(\Sigma)}$
ε	Eccentricity $\sigma_{max}/\sigma_{min}$
$\sigma^*_{max}, \sigma^*_{min}, \varepsilon^*$	Similar, but in the projected space
$\nu(\cdot)$	$N(0; I_d)$
$\nu_{\Sigma^*}(\cdot)$	$N(0; \Sigma^*)$
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 $\{x : \|x - z\|_\Sigma \leq r\}$

As we have already seen, we can think of ε^* as a small constant even if ε is large, and this will help us tremendously.

4.3 Crude density estimates

Our algorithm relies heavily upon the hope that in the projected space, every spherical region will contain roughly its expected number of points under the mixture distribution. This can shown effortlessly by VC dimension arguments.

Lemma 4 (Accuracy of density estimates) Let $\nu(\cdot)$ denote any density on $\mathbb{R}^d$ from which i.i.d. data is drawn. If the number of data points seen satisfies $M \geq O\left(\frac{d}{\epsilon_0^2} \ln \frac{1}{\delta \epsilon_0}\right)$, then with probability $> 1 - \delta$, 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 closed ball $B \subset \mathbb{R}^d$, let $1_B(x) = \mathbf{1}(x \in B)$ denote the indicator function for B. The concept class $\{1_B : B \subset \mathbb{R}^d \text{ is a sphere}\}$ 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. The next problem we face is that because Gaussians in general have ellipsoidal contours, it is not easy to get tight bounds on the probability mass of a given spherical region. We will content ourselves with rather loose bounds, obtained via the mediation of a linear transformation T which converts ellipsoids into spheres.

Write the $d \times d$ covariance matrix Σ^* 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 shall 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, with $\|x\|_{\Sigma^*} = r$	Point Tx, with $\|Tx\| = r$
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.

Lemma 5 (Relating ellipsoidal Gaussian density estimates to spherical ones) Pick any point z and any radius r. Writing $s = \|z\|_{\Sigma^*}$, the probability mass $\nu_{\Sigma^*}(B(z; r))$ must lie in the range $[\nu(B(s; r/\sigma^*_{max})), \nu(B(s; r/\sigma^*_{min}))]$.

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

$$\begin{aligned} \nu_{\Sigma^*}(B(z; r)) &\geq \nu_{\Sigma^*}(E(z; r/\sigma^*_{max}; \Sigma^*)) \\ &= \nu(B(s; r/\sigma^*_{max})), \end{aligned}$$

where the final equality is a result of applying the transformation 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 6 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 $B(z'; r)$,

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

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 a sphere, under distribution ν_{Σ^*}, grows as its radius increases.

Lemma 7 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. Notice that

$$\begin{aligned} \nu_{\Sigma^*}(B(0; r)) &= \int_{B(0;r)} \nu_{\Sigma^*}(x)dx \\ &= \left(\frac{r}{r + s}\right)^d \int_{B(0;r+s)} \nu_{\Sigma^*}\left(y \cdot \frac{r}{r + s}\right) dy \end{aligned}$$

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

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

We will bound this ratio of integrals by considering a pointwise ratio. For any $y \in B(0; r + s)$, we know $\|y\|_{\Sigma^*} \leq (r + s)/\sigma^*_{min}$ and so

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

given the condition on $r + s$. ∎

We next examine a few technical properties of the unit Gaussian $\nu \sim N(0; I_d)$, as a step towards showing that there are many data points near the centers of projected Gaussians.

Lemma 8 (Crude lower bounds) If $\tau \leq 1/3$ and $d \geq 10$, (a) $\nu(B(0; \tau\sqrt{d})) \geq \tau^d$, and (b) $\nu(B(\tau\sqrt{d}; \tau\sqrt{d})) \geq \tau^d$.

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

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

Continuing in the same vein, this time for a displaced sphere, we get bound (b). ∎

4.4 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 is proportional to ϵ. 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.

Parameters $\rho \leq O(\frac{\epsilon}{\varepsilon^{*2}})$, $d = O(\varepsilon^{*2} \log \frac{1}{w_{min}\rho\delta})$, $\epsilon_0 = w_{min}\rho^d \min\{\frac{1}{8}, \frac{3}{16}\rho^2\varepsilon^{*2}d\}$, $l = p = M(w_{min}\rho^d - \epsilon_0)$, $q = M\nu(B(0; \frac{3}{8\varepsilon^*}\sqrt{d}))$. It is important that all these parameters can easily be computed.

Lemma 9 There is at least one data point within Mahalanobis distance $\rho\sqrt{d}$ of each center. Any such point x has at least p data points close by, in $B(x; \rho\sigma^*_{max}\sqrt{d})$, and thus $r_x \leq \rho\sigma^*_{max}\sqrt{d}$.

Proof. Since all the density estimates are accurate within ϵ_0, we need only show that $w_{min}\nu_{\Sigma^*}(E(0;\rho\sqrt{d};\Sigma^*)) \geq \epsilon_0$ and that $w_{min}\nu_{\Sigma^*}(B(x;\rho\sigma^*_{max}\sqrt{d})) \geq p/M + \epsilon_0$ if $\|x\|_{\Sigma^*} \leq \rho\sqrt{d}$. Transformation T and Lemma 5 convert statements about ν_{Σ^*} into statements about ν, in particular,
$\nu_{\Sigma^*}(E(0;\rho\sqrt{d};\Sigma^*)) = \nu(B(0;\rho\sqrt{d}))$ and
$\nu_{\Sigma^*}(B(x;\rho\sigma^*_{max}\sqrt{d})) \geq \nu(B(\rho\sqrt{d};\rho\sqrt{d}))$.
The rest follows from Lemma 8. ∎

This lemma gives an upper bound on r_x for points x close to a center. We next need to show that r_x will be significantly larger for points further away, at Mahalanobis distance $\geq (3\varepsilon^*+1)\rho\sqrt{d}$ from the center.

Lemma 10 Suppose $r_x \leq \rho\sigma^*_{max}\sqrt{d}$ for some point x which is at Mahalanobis distance $\geq (3\varepsilon^*+1)\rho\sqrt{d}$ from the closest center μ_i^* and at L_2 distance $\geq \frac{1}{4\varepsilon^*}\sigma^*_{min}\sqrt{d}$ from all other centers.
Then any point z within Mahalanobis distance $\rho\sqrt{d}$ of μ_i^* will have $r_z < r_x$.

Proof sketch. The conditions on x imply that
(1) $\|x-\mu_i^*\|_{\Sigma^*} \geq (3\varepsilon^*+1)\rho\sqrt{d}$;
(2) $\|x-\mu_j^*\|_{\Sigma^*} \geq \frac{1}{4\varepsilon^*}\frac{\sigma^*_{min}}{\sigma^*_{max}}\sqrt{d} \geq \frac{\sqrt{d}}{4\varepsilon^{*2}}$ for $j \neq i$; and
(3) $\pi^*(B(x;r_x)) \geq \frac{p}{M} - \epsilon_0$.
The result follows when these three facts are combined using Lemma 6. ∎

This lemma implies roughly that within any Gaussian, the lowest r_x values come from data points which are within distance $(3\varepsilon^*+1)\rho\sqrt{d}$ 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, and guarantee that this knocks out the high-density points near the current center while leaving intact the high-density regions near other centers.

Lemma 11 Let x be any point within Mahalanobis distance $\rho(3\varepsilon^*+1)\sqrt{d}$ of some center μ_i^*. Then the q data points closest to x include all data points in $B(\mu_i^*; \frac{1}{4\varepsilon^*}\sigma^*_{min}\sqrt{d})$ and no point outside $B(\mu_i^*; (\frac{1}{2\varepsilon^*}-\rho)\sigma^*_{max}\sqrt{d})$.

Proof sketch. Rewriting $\frac{q}{M}$ as $w_{min}\nu_{\Sigma^*}(E(0;\frac{3}{8\varepsilon^*}\sqrt{d};\Sigma^*))$, we notice that it lies between $w_{min}\nu_{\Sigma^*}(B(0;\frac{3}{8\varepsilon^*}\sigma^*_{min}\sqrt{d}))$ and $w_{min}\nu_{\Sigma^*}(B(0;\frac{3}{8\varepsilon^*}\sigma^*_{max}\sqrt{d}))$. The first inclusion consists in proving that

$$\pi^*(B(x;(\tfrac{1}{4\varepsilon^*}+\rho\varepsilon^*(3\varepsilon^*+1))\sigma^*_{min}\sqrt{d})) \leq \tfrac{q}{M} - \epsilon_0;$$

this is a direct consequence of Lemma 7 and the lower bound on $\frac{q}{M}$. The second inclusion is shown similarly. ∎

Lemma 12 (Accuracy of low-dimensional center estimates) If the various parameters are set in accordance with the specifications above, then with probability $> 1-\delta$, for every $i \leq k$, $\|\widehat{\mu}_i^* - \mu_i^*\|_{\Sigma^*} \leq (3\varepsilon^*+1)\rho\sqrt{d}$.

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 9, this $r_x \leq \rho\sigma^*_{max}\sqrt{d}$. Let μ_i^* be the projected center closest to x. Since the Gaussians are $1/2$-separated, x is at distance at least $\frac{1}{4}\sigma^*_{min}\sqrt{d}$ from all the other projected centers. By Lemma 10, we then see that x must be within Mahalanobis distance $(3\varepsilon^*+1)\rho\sqrt{d}$ of μ_i^*.

Say that at some stage in the algorithm, center-estimates $\widehat{C}$ have already been chosen, $|\widehat{C}| \geq 1$, and that these correspond to true centers C. Select any $y \in \widehat{C}$; by the induction hypothesis there is a j for which $\|y-\mu_j^*\|_{\Sigma^*} \leq$

$(3\varepsilon^*+1)\rho\sqrt{d}$. S' does *not* contain the q points closest to y. By Lemma 11, this removes $B(\mu_j^*; \frac{1}{4\varepsilon^*}\sigma_{min}^*\sqrt{d})$ from S', yet no point outside $B(\mu_j^*; (\frac{1}{2\varepsilon^*}-\rho)\sigma_{max}^*\sqrt{d})$ 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 $\frac{1}{4\varepsilon^*}\sigma_{min}^*\sqrt{d}$ from centers in C. Because of the separation of the mixture, z must be at distance at least $\frac{1}{4}\sigma_{min}^*\sqrt{d}$ from all centers but μ_i^*. Again due to the separation of the Gaussians, all points within distance $\rho\sigma_{max}^*\sqrt{d}$ of μ_i^* remain in S', and therefore z is potentially one of these, whereupon, by Lemma 9, $r_z \le \rho\sigma_{max}^*\sqrt{d}$. By Lemma 10 then, $||z-\mu_i^*||_{\Sigma^*} \le (3\varepsilon^*+1)\rho\sqrt{d}$. ∎

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

5 Back in high-dimensional space

We may now safely assume that in $\mathbb{R}^d$, each estimated center $\widehat{\mu}_i^*$ is within Mahalanobis distance $(3\varepsilon^*+1)\rho\sqrt{d}$ 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 $B(\widehat{\mu}_i^*; \rho\sigma_{max}^*\sqrt{d}) \subseteq B(\mu_i^*; (3\varepsilon^*+2)\rho\sigma_{max}^*\sqrt{d})$, as per the proof of Lemma 12. 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. 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. We have already bounded the error on the final 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 $G_1 = N(\mu_1, \Sigma)$. To this end, we establish the following

Definitions $T_j =$ points in S_1 drawn from the j^{th} Gaussian G_j and $f_j = ||\mu_j - \mu_1||/(c\sigma_{max}\sqrt{n}) \ge 1$.

We will show that S_1 is relatively uncontaminated by points from other Gaussians, that is, $|T_2|+\cdots+|T_k|$ is small. Those points which do come from G_1 ought to (we hope) average out to something near its mean μ_1. The problem is that the x coordinates could be highly correlated with the y coordinates (depending upon the nature of Σ), and thus a small, unavoidable error in $\widehat{\mu}_1^*$ might potentially cause the set T_1 to lie far from μ_1 in $\mathbb{R}^n$. To dismiss this possibility we need a bit of matrix analysis.

Write covariance matrix Σ in the form $\begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$, 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 the matrix A a bit more carefully.

Lemma 13 $||Ax|| \le \sigma_{max}||x||_{\Sigma^*}\sqrt{n/d}$ for any $x \in \mathbb{R}^d$.

Proof. $A = \Sigma_{yx}\Sigma_{xx}^{-1}$ is a $(n-d)\times d$ matrix; divide it into $n/d - 1$ 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 x; 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 $u \in \mathbb{R}^d$, $u^T\Sigma_{xx}u > u^T\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}u$, and by choosing $u = \Sigma_{xx}^{-1}v$, we find

$$\begin{aligned} ||v||_{\Sigma^*}^2 &= v^T\Sigma_{xx}^{-1}\Sigma_{xx}\Sigma_{xx}^{-1}v > v^T B_i^T \Sigma_{zz}^{-1} B_i v \\ &\ge \frac{||B_i v||^2}{\lambda_{max}(\Sigma_{zz})} \ge \frac{||B_i v||^2}{\sigma_{max}^2}. \end{aligned}$$

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

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

and the lemma is proved. ∎

The y coordinates of points in T_1 look roughly like random draws from the distribution $N(A\widehat{\mu}_1^*; C)$. What bounds can be given for the average of these points?

Lemma 14 Randomly draw s points $Y_1, \ldots, Y_s$ from Gaussian $N(\mu, I_n)$. Then for any $\Delta \ge \frac{1}{\sqrt{s}}$,

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

Proof. Let $Z_i = Y_i - \mu \sim N(0, I_n)$. The mean $(Z_1 + \cdots + Z_s)/s$ has distribution $N(0, (1/s)I_n)$, and its squared L_2 norm has moment-generating function $\phi(t) = (1 - 2t/s)^{-n/2}$. By Markov's inequality,

$$\mathbf{P}\left(\left\|\frac{Z_1 + \cdots + Z_s}{s}\right\| \geq \Delta\sqrt{n}\right) \leq \frac{\phi(t)}{e^{t\Delta^2 n}};$$

the lemma follows by choosing $t = \frac{s}{2}(1 - \frac{1}{\Delta^2 s})$. ∎

We are finally in a position to bound the deviation of mean(T_j) from μ_j and thereby from μ_1. Specifically, the y coordinates of points in $T_j \subset S_1$ look roughly like random draws from the distribution $N(A(\widehat{\mu}_1^* - \mu_j^*); C)$. We will use the previous two lemmas to bound their average.

Lemma 15 For any $j \geq 1$, mean(T_j) has the same distribution as

$$\mu_j + (X, AX + C^{1/2}E_{|T_j|}),$$

where
(1) $X \in \mathbb{R}^d$ is a random variable with $\|X\| \leq \|\mu_1^* - \mu_j^*\| + \frac{\epsilon}{4}\sigma^*_{min}\sqrt{d}$; and
(2) E_m is shorthand for 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 $|T_j|$ 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 $(Y_i - AX_i) \stackrel{d}{=} N(0; C) \stackrel{d}{=} C^{1/2}N(0, I_{n-d})$, and taking averages, $Y \stackrel{d}{=} AX + C^{1/2}E_{|T_j|}$. ∎

Armed with this result we finally prove the main theorem.

Lemma 16 With probability $> 1 - \delta$, for all $1 \leq i \leq k$, $\|\widehat{\mu}_i - \mu_i\| \leq \epsilon\sigma_{max}\sqrt{n}$.

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 $\|x - \mu_j^*\|_{\Sigma^*} \geq (\frac{1}{2} - 5\varepsilon^*\rho)f_j\sqrt{d}$ and therefore

$$\begin{aligned}\frac{\mathbf{P}(x \text{ comes from } G_j)}{\mathbf{P}(x \text{ comes from } G_1)} &\leq \frac{w_j e^{-(\frac{1}{2} - 5\varepsilon^*\rho)^2 f_j^2 d/2}}{w_1 e^{-(5\varepsilon^{*2}\rho)^2 d/2}} \\ &\leq \frac{w_j\epsilon}{64c^2\varepsilon^{*2}f_j^2}.\end{aligned}$$

This inequality effectively bounds the number of outliers $|T_2| + \cdots + |T_k|$. The normed difference between μ_1 and the mean of S_1, which we hope is close to zero, is given by

$$\begin{aligned}&\|\text{mean}(S_1) - \mu_1\| \\ &\leq \sum_{j=1}^{k}\|\text{mean}(T_j) - \mu_j\|\frac{|T_j|}{l} + \sum_{j=2}^{k}\|\mu_j - \mu_1\|\frac{|T_j|}{l} \\ &\leq \|C^{1/2}E_l\| + O\left(\sum_{j>1} c\varepsilon^* f_j \cdot \frac{|T_j|}{l}\right)\sigma_{max}\sqrt{n},\end{aligned}$$

where E_l is, as before, the mean of l i.i.d. $N(0; I_{n-d})$ random variables, and the final inequality uses Lemmas 13 and 15. It remains to bound these two terms.

(a) Since $C = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$ and each of these two right-hand terms is positive semidefinite, $\lambda_{max}(C) \leq \lambda_{max}(\Sigma_{yy}) \leq \sigma^2_{max}$ and therefore $\|C^{1/2}E_l\| \leq \sigma_{max}\|E_l\|$. To bound $\|E_l\|$ we use Lemma 14.

(b) A Chernoff bound shows that $|T_j| \leq O(\frac{lw_j\epsilon}{c\varepsilon^* f_j})$ for $j > 1$, the final piece of the puzzle. ∎

6 Auxiliary issues

6.1 Learning the mixing weights and covariance matrix

The algorithm we have presented solves the core combinatorial problem of learning the centers of a mixture of Gaussians. In some situations, for instance if likelihoods need to be computed, it is also important to learn the mixing weights and the covariance matrix. We now briefly suggest a possible approach.

Assume we have in hand the center-estimates $\widehat{\mu}_i \in \mathbb{R}^n$. Associate each data point with its closest center-estimate. It can be shown that the proportion of misclassified points in this hard clustering will (with high probability) be only $O(k\exp(-\Omega((c - \epsilon)^2 n)))$. The proof is not difficult but requires some care because of the dependence between the data and the estimated centers.

By computing statistics of the points in each cluster, we can obtain estimates for the mixing weights and covariance matrix. Taking averages of the points in each cluster may also give improved estimates of the centers. In fact, this technique might lower the sample complexity to just $k^{O(1)}/\epsilon^2$ instead of $k^{O(\log^2 1/\epsilon)}$ (ignoring terms in δ) – use our algorithm to construct crude estimates of the means, correct to within $O(c\sigma_{max}\sqrt{n})$, and then use the hard clustering to improve this accuracy to $\epsilon\sigma_{max}\sqrt{n}$.

6.2 More general families of distributions

There has been some recent interest in modelling data by distributions which have fatter tails than the Gaussian. For instance, Basu and Micchelli (1998) report experiments which attempt to model speech data using densities

$$p(x) \propto \exp\left(-c((x-\mu)^T\Sigma^{-1}(x-\mu))^\alpha\right)$$

where $\alpha < 1$ (the case $\alpha = 1$ produces a Gaussian). These distributions also have ellipsoidal contours, but the density drops off at a slower rate. This suggests a possible generalization of our results.

Definition A distribution on $\mathbb{R}^n$ is *ellipsoidally symmetric* if it has a density of the form $p(x) = f((x-\mu)^T\Sigma^{-1}(x-\mu))$, where Σ is a positive definite matrix.

Such distributions remain ellipsoidally symmetric when projected, and both our random projection lemmas (maintaining intercluster distance and reducing eccentricity) continue to apply. It should be possible to design a simple low-dimensional clustering algorithm which will work for a wide choice of functions f.

Mixtures of discrete distributions are also commonly used, and various ideas for handling them have recently been suggested by Kearns *et al* (1994) and by Freund and Mansour (1999). It is plausible, as per Diaconis and Freedman (1984), that many families of discrete distributions start looking more Gaussian when randomly projected into low dimension. This suggests an unusual algorithm for learning these mixtures: project the data, apply a low-dimensional Gaussian center estimator, and then perform some sort of high-dimensional reconstruction!

6.3 Odds and ends

One source of concern about our algorithm is that a naive computation of the r_x values would seem to require computing distances between all pairs of points, which is infeasible for enormous data sets. In such cases, acceptable performance might be obtained by computing these values with respect to a small random subset of the data. That is, randomly select a small subset $T \subset S$ of size $O(k)$ and for each data point $x \in S$, let r_x be the smallest radius such that $B(x; r_x)$ contains at least p points of T.

We end with an important open problem. Our algorithm will work when different clusters have differing covariances, provided these matrices have approximately the same trace. It would be a significant advance to remove this qualification.

Acknowledgements

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

Literature cited

Basu, S. & Micchelli, C.A. (1998) Parametric density estimation for the classification of acoustic feature vectors in speech recognition. *Nonlinear Modeling*, eds. J. Suykens and J. Vandewalle. Kluwer, Boston.

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.

Freund, Y. & Mansour, Y. (1999) Estimating a mixture of two product distributions. *ACM Conference on Computational Learning Theory*.

Gupta, A. (1999) Embedding tree metrics into low dimensional Euclidean spaces. *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.

Kearns, M., Mansour, Y., Ron, D., Rubinfeld, R., Schapire, R. & Sellie, L. (1994) On the learnability of discrete distributions. *ACM Symposium on Theory of Computing*.

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.* Wiley.

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.* Wiley.