Journal of Machine Learning Research 8 (2007) 203-226 Submitted 10/05; Revised 9/06; Published 2/07

A Probabilistic Analysis of EM for Mixtures of Separated, Spherical Gaussians

Sanjoy Dasgupta DASGUPTA@CS.UCSD.EDU
University of California, San Diego
9500 Gilman Drive #0404
La Jolla, CA 92093-0404

Leonard Schulman SCHULMAN@CALTECH.EDU
California Institute of Technology
1200 E. California Blvd., MC 256-80
Pasadena, CA 91125

Editor: Leslie Pack Kaelbling

Abstract

We show that, given data from a mixture of k well-separated spherical Gaussians in $\mathbb{R}^d$, a simple two-round variant of EM will, with high probability, learn the parameters of the Gaussians to near-optimal precision, if the dimension is high ($d \gg \ln k$). We relate this to previous theoretical and empirical work on the EM algorithm.

Keywords: expectation maximization, mixtures of Gaussians, clustering, unsupervised learning, probabilistic analysis

1. Introduction

At present the *expectation-maximization* algorithm (Dempster, Laird, and Rubin, 1977; Wu, 1983) is the method of choice for learning mixtures of Gaussians. A series of theoretical and experimental studies over the past three decades have contributed to the collective intuition about this algorithm. We will reinterpret some of these results in the context of a new performance guarantee.

In brief, EM is a hillclimbing algorithm which starts with some initial estimate of the Gaussian mixture and then repeatedly changes this estimate so as to improve its likelihood, until it finally converges to a local optimum of the search space. It is well known to practitioners that the quality of the output can be influenced significantly by the manner in which EM is initialized. Another source of variation is that in practice it is quite common to add or remove Gaussians during the search process, according to various intuitively-motivated criteria.

Given the enormous importance of Gaussian mixture models in applied statistics and machine learning, it is reasonable to wonder how good the EM algorithm is. Among other things, a rigorous analysis of its performance might shed light on some of the unresolved issues in its implementation—for instance, initialization. It is in this spirit that we approach our analysis of EM.

A common form of theoretical study is *worst-case analysis*, which in this case would attempt to bound how far EM can deviate from the optimal log-likelihood, or perhaps the optimal set of mixture parameters. Such an analysis turns out to be trivial, because—as is well known—EM's output can be arbitrarily far from optimal for either of the two criteria above (we will see such an example in Section 2.3). Thus, this line of reasoning does not appear to yield any interesting insights into the algorithm's behavior.

In this paper, we perform what might be called a *best-case analysis*. We assume that the data are the best EM could possibly hope for: i.i.d. samples from a mixture of spherical Gaussians in $\mathbb{R}^d$ which are well separated from one another. At first glance, it seems that we are once again in danger of getting a trivial result, namely that EM will succeed without a hiccup. But this is not the case. In fact, we will see that even in this extremely optimistic setting, many common ways of initializing, and subsequently running, EM will make it fail dramatically. On the other hand, if EM is run in a particular manner which we specify, then within just two rounds, it will identify the correct mixture parameters with near-perfect accuracy.

If the data is expected to have k clusters, it is common for practitioners to start EM with more than k centers, and to later prune some of these extra centers. We present a simple example to demonstrate exactly why this is necessary, and obtain an expression for the number of initial centers which should be used: at least $\frac{1}{w_{min}} \ln k$, where w_{min} is a lower bound on the smallest mixing weight. The typical method of pruning is to remove Gaussian-estimates with very low mixing weight (known as *starved clusters*). Our theoretical analysis shows that this is not enough, that there is another type of Gaussian-estimate, easy to detect, which also needs to be pruned. Specifically, it is possible (and frequently occurs in simulations) that two of EM's Gaussian-estimates share the same cluster, each with relatively high mixing weight. We present a very simple, provably correct method of detecting this situation and correcting it.

It is widely recognized that a crucial issue in the performance of EM is the choice of initial parameters. For the means, we use the popular technique of selecting them randomly from the data set. This is shown to be adequate for the performance guarantee we derive. Our analysis also makes it clear that it is vitally important to pick good initial estimates of the covariances, a subject which has received somewhat less attention. We use an initializer whose origin we are unable to trace but which is mentioned in Bishop's text (1995).

Our analysis is focused on the case when the data are high-dimensional ($d \gg \ln k$), and it brings out some interesting qualitative differences from the low-dimensional case. In particular, it is common to think of EM as assigning "soft" cluster memberships in which each data point is not definitively assigned to a single cluster but, rather, is split between clusters according to its relative proximities to them. Moreover, this is sometimes quoted as a source of EM's effectiveness—that these soft memberships allow cluster boundaries to shift in a smooth and stable manner. In the optimistic high-dimensional scenario we analyze, the behavior is quite different, in that cluster memberships are effectively always "hard". This is because the distances are so large that for any two clusters, there is only a small part of the space in which there is any ambiguity of assignment, and it is unlikely that data points would lie in these zones. Moreover, the phenomenon of smooth transitions between clusterings is nonexistent. Instead, EM quickly snaps into one of several trajectories (loosely defined), and thereafter heads to a nearby local optimum. Initialization is of supreme importance—once EM has snapped into the wrong trajectory, all is lost.

1.1 Relation to Previous Work on EM

A standard criticism of EM is that it converges slowly. Simulations performed by Redner and Walker (1984), and others, demonstrate this decisively for one-dimensional mixtures of two Gaussians. It is also known that given data from a mixture of Gaussians, when EM gets close to the true solution, it exhibits *first-order convergence* (Titterington, Smith, and Makov, 1985). Roughly speaking, the idea is this: given m data points from a mixture with parameters (means, covariances, and mixing weights) θ^*, where m is very large, the likelihood has a local maximum at some set of parameters θ^m close to θ^*. Let $\theta^{\langle t\rangle}$ denote EM's parameter-estimates at time t. It can be shown (cf. Taylor expansion) that when $\theta^{\langle t\rangle}$ is near θ^m,

$$\|\theta^{\langle t+1\rangle} - \theta^m\| \leq \lambda \cdot \|\theta^{\langle t\rangle} - \theta^m\|,$$

where $\lambda \in [0,1)$ and $\|\cdot\|$ is some norm.[1] If the Gaussians are closely packed then λ is close to one; if they are very far from one another then λ is close to zero.

Xu and Jordan (1995) present theoretical results which mitigate some of the pessimism of first-order convergence, particularly in the case of well-separated mixtures, and they note that moreover near-optimal log-likelihood is typically reached in just a few iterations. We also argue in favor of EM, but in a different way. We ask, how close does $\theta^{\langle t\rangle}$ have to be to θ^m for slow convergence to hold? Let $D(\theta_1,\theta_2)$ denote the maximum Euclidean distance between the respective means of θ_1 and θ_2. For one-dimensional data, it can be seen quite easily from canonical experiments (Redner and Walker, 1984) that convergence is slow even if $D(\theta^{\langle t\rangle},\theta^*)$ is large. However, our results suggest that this no longer holds in higher dimension. For reasonably well-separated spherical Gaussians in $\mathbb{R}^d$ (where *separation* is defined precisely in the next section), convergence is very fast until $D(\theta^{\langle t\rangle},\theta^*) \approx e^{-\Omega(d)}$. In fact, we can make EM attain this accuracy in just two rounds. The error $e^{-\Omega(d)}$ is so miniscule for large d that subsequent improvements are not especially important.

At a high level, previous analyses of EM have typically adopted an *optimization-based* viewpoint: they have studied EM by studying the objective function (log-likelihood) that it is ostensibly optimizing. A typical tool in this kind of analysis is to perform a Taylor expansion of the log-likelihood in the vicinity of a local optimum, assuming the data are i.i.d. draws from a mixture of Gaussians, and to thereby get insight into the speed at which EM is likely to move when it gets close to this optimum. A major drawback of this approach is that it only addresses what happens *close to convergence*. It cannot, for instance, give intuition about how to initialize EM, or about whether a local optimum of high quality is attained.

In contrast, we perform a *probabilistic* analysis. We also assume that the data are i.i.d. draws from a mixture of Gaussians, but we focus upon the actual algorithm and ignore the likelihood function altogether. We ask, what will the algorithm do in step one, with high probability over the choice of data? In step two? And so on. This enables us to address issues of initialization and global optimality.

1.2 Results

Performance guarantees for clustering will inevitably involve some notion of the *separation* between different clusters. There are at least two natural ways of defining this. Take for simplicity the case of two d-dimensional Gaussians $N(\mu_1, I_d)$ and $N(\mu_2, I_d)$. If each coordinate (attribute) provides a little bit of discriminative information between the two clusters, then on each coordinate μ_1 and μ_2 differ by at least some small amount, say δ. The L_2 distance between μ_1 and μ_2 is then at least $\delta\sqrt{d}$. As further attributes are added, the distance between the centers grows, and the two clusters become more clearly distinguishable from one another. This is the usual rationale for using high-dimensional data: the higher the dimension, the easier (in an information-theoretic sense) clustering

[1] This might not seem so bad, but contrast it with *second-order convergence*, in which $\|\theta^{\langle t+1\rangle} - \theta^m\| \leq \lambda \cdot \|\theta^{\langle t\rangle} - \theta^m\|^2$.

should be. The only problem then, is whether there are algorithms which can efficiently exploit the tradeoff between this high information content and the curse of dimensionality. This viewpoint suggests that the Euclidean distance between the centers of d-dimensional clusters can reasonably be measured in units of $\sqrt{d}$, and that it is most important to develop algorithms which work well under the assumption that this distance is some constant times $\sqrt{d}$. On the other hand, it should be pointed out that if $\|\mu_1 - \mu_2\| = \delta\sqrt{d}$ for some constant $\delta > 0$, then for large d the overlap in probability mass between the two Gaussians is miniscule, exponentially small in d. Therefore, it should not only be interesting but also possible to develop algorithms which work well when the L_2 distance between centers of clusters is some constant, regardless of the dimension (as opposed to a constant times $\sqrt{d}$).

Where does EM fall in this spectrum of separation? It lies somewhere in between: we show that it works well when the distance between d-dimensional clusters is bigger than $d^{1/4}$.

Our central performance guarantee requires that the clusters actually look spherical-Gaussian, more specifically that the data points are drawn i.i.d. from some (unknown) mixture of spherical Gaussians, which could potentially have different variances. We show that if the clusters are reasonably well-separated (in the sense we just defined), and if the dimension $d \gg \ln k$ then only two rounds of EM are required to learn the mixture to within near-optimal precision, with high probability $1 - k^{-\Omega(1)}$. Our measure of accuracy is the function $D(\cdot,\cdot)$ introduced above. The precise statement of the theorem can be found in the next section, and applies not only to EM but also to other similar schemes, including for instance some of the variants of EM and k-means introduced by Kearns, Mansour, and Ng (1997).

Several recent papers have suggested alternative algorithms for learning the parameters of a mixture of well-separated Gaussians (or other distributions with similar concentration properties), given data drawn i.i.d. from that distribution. The first in this series, by Dasgupta (1999), requires the Gaussians to be “sphere-like” and separated by a distance of $\Omega(\sqrt{d})$. Arora and Kannan (2004) handle more general Gaussians, and reduce the separation requirement to $\Omega(d^{1/4})$ in the spherical case (as in this paper). Vempala and Wang (2004) use spectral projection to bring the separation constraint for spherical Gaussians down to just $\Omega((k \log d)^{1/4})$, where k is the number of clusters. Vempala, Kannan, and Salmasian (2005) and Achlioptas and McSherry (2005) give extensions of these latter results to ellipsoidal clusters. The last three results are especially relevant to the current paper because the amount of intercluster separation we require can likely be substantially reduced if spectral projection is used as a preprocessing step.

In the final section of the paper, we discuss a crucial issue: what features of our main assumption (that the clusters are high-dimensional Gaussians) make our guarantees for EM possible? This assumption is also the basis of the other theoretical results mentioned above, but can real data sets reasonably be expected to satisfy it? If not, in what way can it usefully be relaxed?

2. Statement of Results

To motivate our model, we start by examining some properties of high-dimensional Gaussians.

2.1 High-dimensional Gaussians

A spherical Gaussian $N(\mu, \sigma^2 I_d)$ assigns to point $x \in \mathbb{R}^d$ the density

$$p(x) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right),$$

$\|\cdot\|$ being Euclidean distance. If $X = (X_1, \ldots, X_d)$ is randomly chosen from $N(0, \sigma^2 I_d)$ then its coordinates are i.i.d. $N(0, \sigma^2)$ random variables. Each coordinate has expected squared value σ^2 so $\mathbb{E}\|X\|^2 = \mathbb{E}(X_1^2 + \cdots + X_d^2) = \sigma^2 d$. It then follows by a large deviation bound that $\|X\|^2$ will be tightly concentrated around $\sigma^2 d$:

$$\mathbb{P}(|\|X\|^2 - \sigma^2 d| > \varepsilon\sigma^2 d) \leq e^{-d\varepsilon^2/8}.$$

This bound and others like it will be proved in Section 3. It means that almost the entire probability mass of $N(0, \sigma^2 I_d)$ lies in a thin shell at a radius of about $\sigma\sqrt{d}$ from the origin. The density of the Gaussian is highest at the origin; however, the surface area at distance r from the origin, $0 \leq r \leq \sigma\sqrt{d}$, increases faster than the density at distance r decreases (Bishop, 1995, exercise 1.4).

It is natural therefore to think of a Gaussian $N(\mu, \sigma^2 I_d)$ as having *radius* $\sigma\sqrt{d}$. We say two Gaussians $N(\mu_1, \sigma_1^2 I_d), N(\mu_2, \sigma_2^2 I_d)$ in $\mathbb{R}^d$ are *c-separated* if

$$\|\mu_1 - \mu_2\| \geq c \max\{\sigma_1, \sigma_2\}\sqrt{d},$$

that is, if they are c radii apart. A mixture of Gaussians is c-separated if the Gaussians in it are pairwise c-separated. In general we will let c_{ij} denote the separation between the i^{th} and j^{th} Gaussians, and $c = \min_{i \neq j} c_{ij}$. We can reasonably expect that the difficulty of learning a mixture of Gaussians increases as c decreases.

A 2-separated mixture of Gaussians contains clusters with almost no overlap. For large d, this is true even of a $\frac{1}{100}$-separated mixture, because for instance, the two balls $B(0, \sqrt{d})$ and $B(\frac{1}{100}\sqrt{d}, \sqrt{d})$ in $\mathbb{R}^d$ share only a tiny fraction of their volume. One useful way of thinking about a pair of c-separated Gaussians is to imagine that on each coordinate their means differ by c. If c is small, then the projection of the mixture onto any one coordinate will look unimodal. This might also be true of a projection onto a few coordinates. But for large d, when all coordinates are considered together, the distribution will cease to be unimodal. This is precisely the reason for using high-dimensional data.

What values of c can be expected of real-world data sets? This will vary from case to case. As an example, we analyzed a canonical data set consisting of handwritten digits collected by USPS. Each digit is represented as a vector in $[-1, 1]^{256}$. We fit a mixture of ten (non-spherical) Gaussians to this data set, by doing each digit separately, and found that it was 0.63-separated.

2.2 The EM algorithm

A mixture of k spherical Gaussians in $\mathbb{R}^d$ is specified by a set of mixing weights w_i (which sum to one and represent the proportions in which the various Gaussians are present) and by the individual Gaussian means $\mu_i \in \mathbb{R}^d$ and variances σ_i^2.

Given a data set $S \in \mathbb{R}^d$, the EM algorithm works by first choosing starting values $w_i^{\langle 0 \rangle}, \mu_i^{\langle 0 \rangle}, \sigma_i^{\langle 0 \rangle}$ for the parameters, and then updating them iteratively according to the following two steps (at time t).

E step Let $\tau_i \sim N(\mu_i^{\langle t\rangle}, \sigma_i^{\langle t\rangle 2} I_d)$ denote the density of the i^{th} Gaussian-estimate. For each data point $x \in S$, and each $1 \leq i \leq k$, compute

$$p_i^{\langle t+1\rangle}(x) = \frac{w_i^{\langle t\rangle}\tau_i(x)}{\sum_j w_j^{\langle t\rangle}\tau_j(x)},$$

the conditional probability that x comes from the i^{th} Gaussian with respect to the current estimated parameters.

M step Now update the various parameters in an intuitive way. Let m be the size of S.

$$\begin{aligned} w_i^{\langle t+1\rangle} &= \frac{1}{m}\sum_{x\in S} p_i^{\langle t+1\rangle}(x), \\ \mu_i^{\langle t+1\rangle} &= \frac{1}{m w_i^{\langle t+1\rangle}}\sum_{x\in S} x\, p_i^{\langle t+1\rangle}(x), \\ \sigma_i^{\langle t+1\rangle 2} &= \frac{1}{m w_i^{\langle t+1\rangle} d}\sum_{x\in S} \|x-\mu_i^{\langle t+1\rangle}\|^2\, p_i^{\langle t+1\rangle}(x). \end{aligned}$$

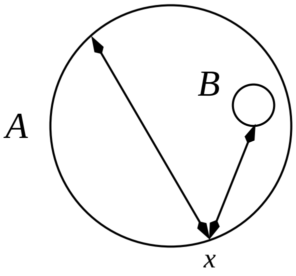
2.3 The Main Issues

We now give a high-level description of some fundamental issues that arise in our analysis of EM, and that dictate our key design decisions.

2.3.1 TIGHT CONCENTRATION OF INTERPOINT DISTANCES

In a high-dimensional space $\mathbb{R}^d$, the distances between data points—whether sampled from the same Gaussian or from different Gaussians—are tightly concentrated around their expected values. In particular, if the Gaussians happen to have the same variance $\sigma^2 I_d$, and if the distances between their centers are $\gg \sigma d^{1/4}$, then the chance that two points from different Gaussians are closer together than two points from the same Gaussian, is tiny, $e^{-\Omega(poly(d))}$. Therefore, an examination of interpoint distances is enough to almost perfectly cluster the data. A variety of different algorithms will work well under these circumstances, and EM is no exception.

What if the Gaussians have different variances σ_i? Once again, interpoint distances are close to their expected values, but now a new complication is introduced. If a small-variance cluster B is close to the center of a larger-variance cluster A, then it is quite possible for points $x \in A$ to be closer to all of B than to any other point in A:



We expressly rule out this case by requiring the separation between any two clusters i and j to satisfy

$$\|\mu_i - \mu_j\|^2 \geq |\sigma_i^2 - \sigma_j^2| d.$$

2.3.2 INITIALIZATION

Suppose the true number of Gaussians, k, is known. Let S denote the entire data set, with S_i being the points drawn from the i^{th} true Gaussian $N(\mu_i, \sigma^2 I_d)$. A common way to initialize EM is to pick l data points at random from S, and to use these as initial *center-estimates* $\mu_i^{\langle 0 \rangle}$. How large should l be? It turns out that if these l points include at least one point from each S_i, then EM can be made to perform well. This suggests $l = \Omega(k \ln k)$. Conversely, if the initial centers miss some S_i, then EM might perform poorly.

Here is a concrete example (Figure 1). Let d denote some high dimension, and place the k true Gaussians $N(\mu_1, I_d), \ldots, N(\mu_k, I_d)$ side by side in a line, leaving a distance of at least $3\sqrt{d}$ between consecutive Gaussians. Assign them equal mixing weights. As before let S_i be the data points from the i^{th} Gaussian, and choose EM's initial center-estimates from the data. Suppose the initial centers contain nothing from S_1, one point from S_2, and at least one point from S_3. The probability of this event is at least some constant. Then no matter how long EM is run, it will assign just one Gaussian-estimate to the first two true Gaussians. In the first round of EM, the point from S_2 (call it $\mu_1^{\langle 0 \rangle}$) will move between μ_1 and μ_2. It will stay there, right between the two true centers. None of the other center-estimates $\mu_i^{\langle t \rangle}$ will ever come closer to μ_2; their distance from it is so large that their influence is overwhelmed by that of $\mu_1^{\langle t \rangle}$. This argument can be formalized easily using the large deviation bounds that we will introduce in the next section.

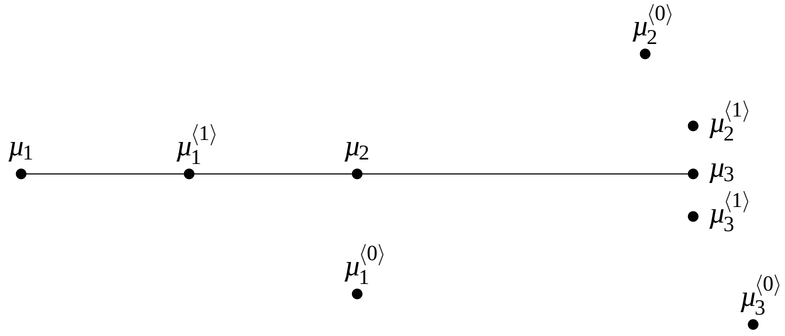

$\mu_2^{\langle 0 \rangle}$
$\mu_2^{\langle 1 \rangle}$
μ_1
$\mu_1^{\langle 1 \rangle}$
μ_2
μ_3
$\mu_3^{\langle 1 \rangle}$
$\mu_1^{\langle 0 \rangle}$
$\mu_3^{\langle 0 \rangle}$

Figure 1: For this mixture, the positions of the center-estimates do not move much after the first step of EM.

How about the initial choice of variance? In the case when the Gaussians have the same spherical covariance, this is not all that important, except that a huge overestimate might cause slower convergence. In the case when the Gaussians have different variances, however, the initial estimates are crucially important, and so we will use a fairly precise estimator, a variant of which is mentioned in Bishop's text (1995).

2.3.3 AFTER THE FIRST ROUND OF EM

After one round of EM, the center-estimates are pruned to leave exactly one per true Gaussian. This is accomplished in a simple manner. First, remove any center-estimates with very low mixing weight (this is often called "cluster starvation"). Any remaining center-estimate (originally chosen, say,

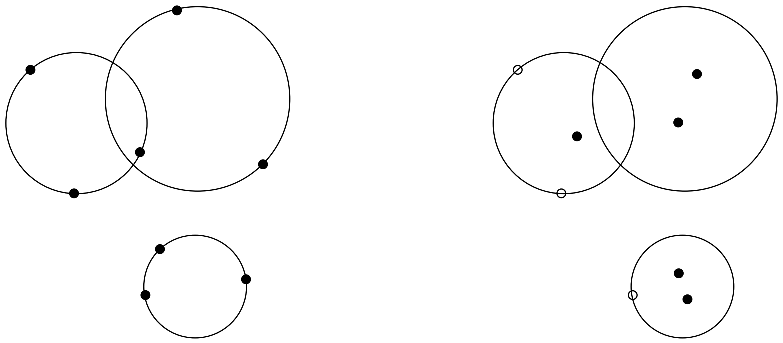


Figure 2: The large circles are true clusters, while the dots (solid or hollow) are EM's center-estimates. *Left:* after initialization, we have at least one center-estimate per true cluster. *Right:* After the first round of EM, each center-estimate has either been "starved" (shown as hollow) or has moved closer to the corresponding true center.

from S_i) has relatively high mixing weight, and we can show that as a result of the first EM iteration, it will have moved close to μ_i (Figure 2). A trivial clustering heuristic, due to Gonzalez (1985), is then good enough to select one center-estimate near each μ_i.

With exactly one center-estimate per (true) Gaussian, a second iteration of EM will accurately retrieve the means, variances, and mixing weights. In fact the clustering of the data (the fractional labels assigned by EM) will be almost perfect, that is to say, each fractional label will be close to zero or one, and will in almost all cases correctly identify the generating Gaussian. Therefore further iterations will not help much: these additional iterations will move the center-estimates around by at most $e^{-\Omega(d)}$.

2.4 A Two-round Variant of EM

Here is a summary of the modified algorithm, given m data points in $\mathbb{R}^d$ which have been generated by a mixture of k Gaussians. The value of l will be specified later; for the time being it can be thought of as $O(k \ln k)$.

Initialization Pick l data points at random as starting estimates $\mu_i^{\langle 0 \rangle}$ for the Gaussian centers. Assign them identical mixing weights $w_i^{\langle 0 \rangle} = \frac{1}{l}$. For initial estimates of the variances use

$$\sigma_i^{\langle 0 \rangle 2} = \frac{1}{2d} \min_{j \neq i} \|\mu_i^{\langle 0 \rangle} - \mu_j^{\langle 0 \rangle}\|^2.$$

EM Run one round of EM. This yields modified estimates $w_i^{\langle 1 \rangle}, \mu_i^{\langle 1 \rangle}, \sigma_i^{\langle 1 \rangle}$.

Pruning Remove all center-estimates whose mixing weights are below $w_T = \frac{1}{4l}$. Then, prune the remaining center-estimates down to just k by using the following adaptation of an algorithm of Gonzalez (1985):

- Compute distances between center-estimates:

$$d(\mu_i^{\langle 1\rangle},\mu_j^{\langle 1\rangle}) = \frac{\|\mu_i^{\langle 1\rangle}-\mu_j^{\langle 1\rangle}\|}{\sigma_i^{\langle 1\rangle}+\sigma_j^{\langle 1\rangle}}.$$

- Choose one of these centers arbitrarily.
- Pick the remaining $k-1$ iteratively as follows: pick the center farthest from the ones picked so far. (The distance from a point x to a set S is $\min_{y\in S} d(x,y)$.)

Call the resulting center-estimates $\tilde{\mu}_i^{\langle 1\rangle}$ (where $1\le i\le k$), and let $\tilde{\sigma}_i^{\langle 1\rangle 2}$ be the corresponding variances. Set the mixing weights to $\tilde{w}_i^{\langle 1\rangle}=\frac{1}{k}$.

EM Run one more step of EM, starting at the $\{\tilde{w}_i^{\langle 1\rangle},\tilde{\mu}_i^{\langle 1\rangle},\tilde{\sigma}_i^{\langle 1\rangle}\}$ parameters and yielding the final estimates $w_i^{\langle 2\rangle},\mu_i^{\langle 2\rangle},\sigma_i^{\langle 2\rangle}$.

2.5 The Main Result

Now that the notation and algorithm have been introduced, we can state the main theorem.

Theorem 1 *Say m data points are generated from a mixture of k Gaussians in $\mathbb{R}^d$,*

$$w_1 N(\mu_1,\sigma_1^2 I_d)+\cdots+w_k N(\mu_k,\sigma_k^2 I_d),$$

where the intercenter distances satisfy the inequality $\|\mu_i-\mu_j\|^2\ge|\sigma_i^2-\sigma_j^2|d$.

Define c_{ij} to be the separation between the i^{th} and j^{th} Gaussians—that is, $\|\mu_i-\mu_j\|=c_{ij}\max(\sigma_i,\sigma_j)\sqrt{d}$ — and $c=\min_{i\neq j}c_{ij}$ to be the overall separation. Let S_i denote the points from the i^{th} Gaussian, and let $w_{min}=\min_i w_i$. For any $\delta>0$ and $\varepsilon>0$, if

1. *parameter $l=\Omega(\frac{1}{w_{min}}\ln\frac{1}{\delta w_{min}})$,*
2. *dimension $d=\Omega(\max(1,c^{-4})\ln\frac{\max(1,c^{-4})}{\delta w_{min}})$,*
3. *number of samples $m=\Omega(l\max(1,c^{-2}))$,*
4. *separation $c^2 d=\Omega(\ln\frac{1}{\varepsilon w_{min}})$,*

then with probability at least $1-\delta$, the variant of EM described above will produce final center-estimates which (appropriately permuted) satisfy

$$\|\mu_i^{\langle 2\rangle}-\mu_i\| \quad\le\quad \|\mathrm{mean}(S_i)-\mu_i\|+\varepsilon\sigma_i\sqrt{d}.$$

This theorem will be proved over the remainder of the paper, but a few words about it are in order at this stage. First of all, the constants which have been left out of the theorem statement are given in Section 4. Second, the best that can be hoped is that $\mu_i^{\langle 2\rangle}=\mathrm{mean}(S_i)$; therefore, the final error bound on the center-estimates becomes very close to optimal as $c^2 d$ increases. Third, similarly strong bounds can be given for the final mixing weights and variances; see Theorem 17. Finally, notice that the bounds require $c\gg d^{-1/4}$; in other words, the distance between the centers of Gaussians i and j must be $\gg\max(\sigma_i,\sigma_j)d^{1/4}$.

d	dimension
k	true number of clusters
w_i, μ_i, σ_i^2	true mixture parameters
σ_{ij}	shorthand for $\max(\sigma_i, \sigma_j)$
w_{min}	lower bound on the mixing weights: $w_i \geq w_{min}$
l	number of clusters with which EM is started; $l > k$
$w_i^{\langle t\rangle}, \mu_i^{\langle t\rangle}, \sigma_i^{\langle t\rangle 2}$	EM's parameter-estimates at time t
$p_i^{\langle t\rangle}(x)$	EM's soft-assignment (to point x, from cluster i) at time t
w_T	threshold used to prune "starved clusters": $w_T = 1/4l$
c_{ij}	separation between Gaussians i and j; $\lVert \mu_i - \mu_j \rVert = c_{ij}\sigma_{ij}\sqrt{d}$
c	overall separation, $c = \min_{i \neq j} c_{ij}$
S	data points
S_i	data points sampled from the i^{th} Gaussian
m	number of data points
ε_o	concentration of interpoint distances and dot products
C_i	center-estimates from S_i which survived the first round: $C_i = \{\mu_{i'}^{\langle 1\rangle} : \mu_{i'}^{\langle 0\rangle} \in S_i, w_{i'}^{\langle 1\rangle} \geq w_T\}$
$d(\mu_i^{\langle 1\rangle}, \mu_j^{\langle 1\rangle})$	weighted distance between centers, used by pruning procedure

Figure 3: Notation.

3. Concentration Properties of Gaussian Samples

Our analysis hinges crucially upon concentration effects: specifically, that interpoint distances and means of subsets of points are likely to be close to their expected values.

The most basic such property is that the squared length of a point drawn from a high-dimensional Gaussian is tightly concentrated. The proof of this well-known fact is repeated here for easy reference.

Lemma 2 *Pick X from the distribution $N(0, I_d)$.*

(a) (Very large deviations) For any $\lambda \geq 1$, we have $\mathbf{P}(\|X\|^2 \geq \lambda d) \leq (e^{\lambda - 1 - \ln \lambda})^{-d/2}$.

(b) (Modest deviations) For any $\varepsilon \in (0,1)$, we have $\mathbf{P}(|\|X\|^2 - d| \geq \varepsilon d) \leq 2e^{-d\varepsilon^2/8}$.

Proof. $\|X\|^2$ has a χ^2 distribution with expectation d, variance $2d$, and moment-generating function $\phi(t) = \mathbf{E}e^{t\|X\|^2} = (1-2t)^{-d/2}$.

(a) By Markov's inequality, for $t \in [0, \frac{1}{2}]$,

$$\mathbf{P}\left(\|X\|^2 \geq \lambda d\right) \leq \frac{\phi(t)}{e^{t\lambda d}};$$

the first assertion of the lemma follows by choosing $t = \frac{1}{2}(1 - \frac{1}{\lambda})$.

(b) Fix any $\varepsilon \in (0,1)$. Applying Markov's inequality tells us that for any $t \in [0, \frac{1}{2}]$,

$$\mathbf{P}(\|X\|^2 \geq (1+\varepsilon)d) = \mathbf{P}(e^{t\|X\|^2} \geq e^{t(1+\varepsilon)d}) \leq \frac{\phi(t)}{e^{t(1+\varepsilon)d}}$$

and for any $t \geq 0$,

$$\mathbf{P}(\|X\|^2 \leq (1-\varepsilon)d) = \mathbf{P}(e^{-t\|X\|^2} \geq e^{-t(1-\varepsilon)d}) \leq \phi(-t)e^{t(1-\varepsilon)d}.$$

Using $t = \frac{\varepsilon}{2(1+\varepsilon)}$ in the first case and $t = \frac{\varepsilon}{2(1-\varepsilon)}$ in the second yields bounds $e^{-d\varepsilon^2/4(1+\varepsilon)}$ and $e^{-d\varepsilon^2/4}$, respectively. ▮

Lemma 2 immediately implies that the distance between two points from the mixture is sharply concentrated around its conditional expected value. Here are the details.

Lemma 3 *If X and Y are chosen independently from $N(\mu_i, \sigma_i^2 I_d)$ and $N(\mu_j, \sigma_j^2 I_d)$ respectively, then for any $\varepsilon \in (0,1)$, the probability that $\|X-Y\|^2$ does not lie in the range*

$$\|\mu_i - \mu_j\|^2 + (\sigma_i^2 + \sigma_j^2)d(1 \pm \varepsilon) \pm 2\|\mu_i - \mu_j\|\sqrt{\sigma_i^2 + \sigma_j^2} \cdot \varepsilon d^{1/2}$$

is at most $2e^{-\varepsilon^2 d/8} + e^{-\varepsilon^2 d/2}$.

Proof. The sum of independent normals is itself normal. Specifically,

$$X - Y \overset{d}{=} N(\mu_i - \mu_j, (\sigma_i^2 + \sigma_j^2)I_d) \overset{d}{=} (\mu_i - \mu_j) + \sqrt{\sigma_i^2 + \sigma_j^2}\, W,$$

where W is a random variable with distribution $N(0, I_d)$. Therefore

$$\|X-Y\|^2 \overset{d}{=} \|\mu_i - \mu_j\|^2 + (\sigma_i^2 + \sigma_j^2)\|W\|^2 + 2\sqrt{\sigma_i^2 + \sigma_j^2}\, (\mu_i - \mu_j) \cdot W.$$

Lemma 2(b) gives a bound on $\|W\|^2$; for the last term we use

$$(\mu_i - \mu_j) \cdot W \overset{d}{=} \|\mu_i - \mu_j\|\, Z$$

where Z is standard normal and thus satisfies $\mathbf{P}(|Z| > \varepsilon d^{1/2}) \leq e^{-\varepsilon^2 d/2}$ (Durrett, p.7). ▮

Thus, for i.i.d. data from a high-dimensional mixture of spherical Gaussians, we have a very good idea of how interpoint distances will be distributed. Similarly, we can bound the number of points drawn from each Gaussian, and also the sizes of certain angles (dot products) formed by data points and cluster centers. The following lemma will be used repeatedly in the analysis.

Lemma 4 (Bounds on interpoint distances and angles, and cluster sizes) *Draw m data points from a c-separated mixture of k spherical Gaussians with smallest mixing weight at least w_{min}. Let $\sigma_{ij} = \max\{\sigma_i, \sigma_j\}$ and write the separation between Gaussians as $\|\mu_i - \mu_j\| = c_{ij}\sigma_{ij}\sqrt{d}$. Let S_i denote the points from the i^{th} Gaussian. Pick any $\varepsilon_o \in (0,1)$. Then, with probability at least $1 - (m^2+m)e^{-\varepsilon_o^2 d/8} - (\frac{1}{2}m^2 + km + km^2)e^{-\varepsilon_o^2 d/2} - ke^{-mw_{min}/8}$,*

(a) for any $x, y \in S_j$,

$$\|x-y\|^2 = 2\sigma_j^2 d(1 \pm \varepsilon_o);$$

(b) for $x \in S_i, y \in S_j, i \neq j$,

$$\|x-y\|^2 = (\sigma_i^2 + \sigma_j^2 + c_{ij}^2\sigma_{ij}^2)d(1 \pm 2\varepsilon_o).$$

(c) for any data point $y \in S_j$,

$$\|y-\mu_j\|^2 = \sigma_j^2 d(1 \pm \varepsilon_o)$$

while for $i \neq j$,

$$\|y-\mu_i\|^2 = (\sigma_j^2 + c_{ij}^2\sigma_{ij}^2)d(1 \pm 2\varepsilon_o);$$

(d) For any $1 \leq i, j, g \leq k$ and any $x \in S_i, y \in S_g$,

$$|(x-\mu_i)\cdot(y-\mu_j)| \leq \sigma_i\varepsilon_o d \cdot \sqrt{(\sigma_g^2 + c_{jg}^2\sigma_{jg}^2)(1+2\varepsilon_o)}.$$

(e) each $|S_i| \geq \frac{1}{2}mw_i$.

Proof. Part (a) follows immediately from the previous lemma. For (b), we start with

$$\|x-y\|^2 = (c_{ij}^2\sigma_{ij}^2 + \sigma_i^2 + \sigma_j^2)d \pm (\sigma_i^2 + \sigma_j^2 + 2c_{ij}\sigma_{ij}\sqrt{\sigma_i^2+\sigma_j^2})\varepsilon_o d$$

from the previous lemma and then simplify using $2c_{ij}\sigma_{ij}\sqrt{\sigma_i^2+\sigma_j^2} \leq c_{ij}^2\sigma_{ij}^2 + \sigma_i^2 + \sigma_j^2$. For the third claim, notice that $y - \mu_j \overset{d}{=} N(0, \sigma_j^2 I_d)$, so we can bound $\|y-\mu_j\|^2$ using Lemma 2. For the second half of (c),

$$\|y-\mu_i\|^2 = \|\mu_i-\mu_j\|^2 + \|y-\mu_j\|^2 - 2(\mu_i-\mu_j)\cdot(y-\mu_j).$$

As was done in the proof of Lemma 3, we can show that $(\mu_i-\mu_j)\cdot(y-\mu_j)$ has the same distribution as $\|\mu_i-\mu_j\|$ times a $N(0,\sigma_j^2)$ random variable. Putting the pieces together, $\|y-\mu_i\|^2 = (c_{ij}^2\sigma_{ij}^2 + \sigma_j^2)d \pm (\sigma_j^2 + 2c_{ij}\sigma_{ij}\sigma_j)\varepsilon_o d$, finishing up with the inequality $2c_{ij}\sigma_{ij}\sigma_j \leq c_{ij}^2\sigma_{ij}^2 + \sigma_j^2$.

For part (d), imagine that first y is chosen, then x. Since $x - \mu_i \overset{d}{=} N(0,\sigma_i^2 I_d)$, the component of $x-\mu_i$ in the direction of $y-\mu_j$ (after y is fixed) has distribution $N(0,\sigma_i^2)$, and thus has absolute value $\leq \sigma_i\varepsilon_o d^{1/2}$ with probability at least $1 - e^{-\varepsilon_o^2 d/2}$. Whereupon

$$|(x-\mu_i)\cdot(y-\mu_j)| \;\leq\; \|y-\mu_j\|\sigma_i\varepsilon_o d^{1/2} \;\leq\; \sqrt{(\sigma_g^2 + c_{jg}^2\sigma_{jg}^2)d(1+2\varepsilon_o)}\cdot\sigma_i\varepsilon_o d^{1/2}.$$

Finally, (e) follows from the Chernoff bound; see, for instance, the appendix of Kearns and Vazirani (1991):

$$\mathbf{P}(|S_i| \leq \tfrac{1}{2}mw_i) \;\leq\; e^{-mw_i/8}.$$

▌

Next, we turn to concentration properties of means of subsets of points. We'll start by showing that there is no large subset of S_i whose average is far from μ_i.

Lemma 5 (Averages of subsets) *Pick a set of n points randomly from $N(0,I_d)$. Choose any integer $1 \leq t \leq n$. Then with probability at least $1-e^{-d/2}$, no subset of t or more of the points has mean of norm greater than $\varepsilon\sqrt{d}$, where*

$$\varepsilon = \sqrt{6\max\left(\frac{1}{t}, \frac{1}{d}\ln\frac{ne}{t}\right)}.$$

Proof. First observe that it suffices to prove the statement for subsets of size exactly t.

Fix any set of t indices. The mean of the corresponding points, call it μ, is distributed according to $N(0, \frac{1}{t}I_d) \overset{d}{=} t^{-1/2}N(0, I_d)$. In particular, $\mathbf{E}\|\mu\|^2 = d/t$. Lemma 2 tells us that for any $\lambda > 1$,

$$\mathbf{P}(\|\mu\|^2 > \lambda \cdot d/t) \leq \exp(-d(\lambda - 1 - \ln\lambda)/2).$$

We will choose $\lambda = t\varepsilon^2$, where ε is defined in the lemma statement. This guarantees that $\lambda \geq 6$ and thus $1 + \ln\lambda \leq \frac{1}{2}\lambda$, whereupon

$$\mathbf{P}(\|\mu\|^2 > \varepsilon^2 d) \;\leq\; e^{-d\lambda/4} \;=\; e^{-dt\varepsilon^2/4}.$$

The number of possible choices of t indices is $\binom{n}{t} \leq (ne/t)^t$. Summing over these,

$$\mathbf{P}(\exists \text{ subset of } t \text{ points with } \|\text{mean}\|^2 > \varepsilon^2 d) \;\leq\; \left(\frac{ne}{t}\right)^t e^{-dt\varepsilon^2/4} \;\leq\; e^{-dt\varepsilon^2/12} \;\leq\; e^{-d/2},$$

by the particular choice of ε. ▮

This will enable us to show that if one of EM's cluster-estimates overlaps substantially with a true cluster S_i, then the mean of the overlapping points will be close to μ_i. The only technical difficulty is that EM has soft assignments for data points, and therefore we need to also deal with *weighted* averages.

More generally, we consider the following problem: suppose you are allowed to distribute a specific amount of weight over the elements of S_i, where each element receives a weight between 0 and 1. What is the *worst* soft assignment of this kind, the one whose weighted average is furthest from μ_i? It is not difficult to see that the worst assignment is a *hard* assignment, whereupon we can apply the previous lemma.

Lemma 6 (Weighted averages) *For any finite set of points $S \subset \mathbb{R}^d$, with associated $[0,1]$-valued weights $\{w_x : x \in S\}$, there is a subset $T \subset S$ such that*

1. $|T| = \lfloor \sum_{x\in S} w_x \rfloor$*; and*
2. $\|$*mean*$(T)\| \geq \|$*weighted-mean*$(S)\|$*, that is,*

$$\left\| \frac{1}{|T|} \sum_{x\in T} x \right\| \;\geq\; \left\| \frac{\sum_{x\in S} w_x x}{\sum_{x\in S} w_x} \right\|.$$

Proof. Let μ_S denote the weighted mean of S. Order points $x \in S$ by increasing $x \cdot \mu_S$, and let T consist of the last $\lfloor \sum_{x\in S} w_x \rfloor$ points in this ordering. Then, the component of mean(T) in the direction of μ_S is least as large as that of any weighted mean of the points in which the weights are in the range $[0,1]$ and sum to $\geq \lfloor \sum_{x\in S} w_x \rfloor$. Letting $\mu_T = \text{mean}(T)$, in particular $\mu_T \cdot \mu_S \geq \mu_S \cdot \mu_S = \|\mu_S\|^2$; and thus $\|\mu_T\|^2$ is at least this large. ▮

Remark In what follows we will assume that all the large deviation bounds of this section—Lemmas 4, 5, and 6—hold for the particular sample S we have drawn, for some $\varepsilon_o \in (0,1)$. For Lemma 5, we will use $t = \frac{1}{2}mw_T$.

From these various concentration properties, we see that points from the same Gaussian, say the i^{th} one, are at distance about $\sqrt{2\sigma_i^2 d}$ from each other while points from different Gaussians $i \neq j$ are at distance about $\sqrt{(\sigma_i^2+\sigma_j^2+c_{ij}^2\sigma_{ij}^2)d}$ from each other. These estimates are accurate to within $O(\sigma_{ij}d^{1/4})$. Therefore, in the case where all Gaussians have the same variance, it is sufficient to have $c_{ij}^2\sigma_{ij}^2 d \gg \sigma_{ij}^2 d^{1/2}$—more simply, $c \gg d^{-1/4}$—for the interpoint distances to reveal enough information for clustering. In particular, it should be possible to make EM work well. The general case, in which the Gaussians have different variances, requires more careful treatment but yields the same conclusion.

4. Conditions

Various parts of the analysis require assumptions on the sample size, dimensionality, and separation. To simplify the exposition, we summarize these conditions up front.

(C1) $\|\mu_i - \mu_j\|^2 \geq |\sigma_i^2 - \sigma_j^2| d$ for all i, j .

(C2) $\varepsilon_o \leq \frac{1}{96}\min(1, c^2)$.

(C3) $d \geq 864 \max(1, c^{-2}) \ln 8el$.

(C4) $m \geq 6912 l \max(1, c^{-2})$.

The constants are astronomical but are doubtless much larger than they need to be, as no attempt has been made to optimize them. For (C2), recall that ε_o is the extent to which squared interpoint distances are concentrated: by Lemma 4, these are all within a multiplicative factor $1 \pm O(\varepsilon_o)$ of their expected values.

We start by establishing properties of the initial choice of centers and variances.

5. Initialization

As we saw earlier, it is crucial that every cluster is represented in the initial center-estimates, and that the variance-estimates are fairly accurate. We now confirm these conditions.

Lemma 7 (Properties of the initial parameters) *If $l > k$ and each $w_i \geq w_{min}$ and condition (C1) holds, then with probability at least $1 - k(l+1)e^{-lw_{min}} - ke^{-lw_{min}/12}$,*
(a) every Gaussian is represented at least twice in the initial center-estimates;
(b) the i^{th} Gaussian provides at most $\frac{3}{2}lw_i$ initial center-estimates, for all $1 \leq i \leq k$; and
(c) if the r^{th} center-estimate is drawn from S_i, then $\sigma_i^2(1-2\varepsilon_o) \leq \sigma_r^{\langle 0 \rangle 2} \leq \sigma_i^2(1+\varepsilon_o)$.

Proof. Assume the center-estimates are simply the first l (randomly chosen) data points. The chance that these do not touch a particular S_i at least twice is $(1-w_i)^l + lw_i(1-w_i)^{l-1} \leq (l+1)(1-w_{min})^l \leq (l+1)e^{-lw_{min}}$. Similarly, since the number of center-estimates chosen from S_i has expectation lw_i, a Chernoff bound tells us that

$$\mathbf{P}(\text{there are more than } \tfrac{3}{2}lw_i \text{ initial centers from } S_i) \;\leq\; e^{-lw_i/12}.$$

For the bound on $\sigma_r^{\langle 0\rangle}$, we have already established that there is at least one other center-estimate from the same S_i. If this is the closest center-estimate to $\mu_r^{\langle 0\rangle}$, then by Lemma 4(a) the squared distance between the two is $2\sigma_i^2 d(1\pm\varepsilon_o)$. On the other hand, the closest center-estimate to $\mu_r^{\langle 0\rangle}$ might be from some other cluster S_j, in which case, by Lemma 4(b), the squared distance is at least $(\sigma_i^2+\sigma_j^2+c_{ij}^2\sigma_{ij}^2)d(1-2\varepsilon_o)\geq 2\sigma_i^2 d(1-2\varepsilon_o)$, since by (C1), $c_{ij}^2\sigma_{ij}^2\geq|\sigma_i^2-\sigma_j^2|$. These two cases give the upper and lower bounds on $\sigma_r^{\langle 0\rangle 2}$. ∎

Remark All the theorems of the following sections are made under the additional hypothesis that the high-probability events of Lemma 7 hold.

6. The First Round of EM

What happens during the first round of EM? The first thing we clarify is that although in principle EM allows "soft" assignments in which each data point is fractionally distributed over various clusters, in practice for large d every data point will give almost its entire weight to center-estimates from one cluster. This is because in high dimension, the distances between clusters are so great that there is just a very narrow region between two clusters where there is any ambiguity of assignment, and the probability that points fall within this region is miniscule.

Lemma 8 (Soft assignments) *Suppose $\mu_{i'}^{\langle 0\rangle}\in S_i$ and $\mu_{j'}^{\langle 0\rangle}\in S_j$. If condition (C2) holds, then for any data point $x\in S_i$, the ratio between the probabilities assigned to x by Gaussian-estimates $p_{i'}^{\langle 1\rangle}$ and $p_{j'}^{\langle 1\rangle}$ is*

$$\frac{p_{i'}^{\langle 1\rangle}(x)}{p_{j'}^{\langle 1\rangle}(x)}\geq\exp\left(\frac{c_{ij}^2\sigma_{ij}^2}{\sigma_j^2}\cdot\frac{d}{8}\right).$$

Proof. The following calculations make occasional use of Lemma 4, along with several inequalities that exploit the bound (C2) on ε_o.

$$\begin{aligned}
\frac{p_{i'}^{\langle 1\rangle}(x)}{p_{j'}^{\langle 1\rangle}(x)} &= \frac{\sigma_{j'}^{\langle 0\rangle d}}{\sigma_{i'}^{\langle 0\rangle d}}\exp\left\{\frac{\|x-\mu_{j'}^{\langle 0\rangle}\|^2}{2\sigma_{j'}^{\langle 0\rangle 2}}-\frac{\|x-\mu_{i'}^{\langle 0\rangle}\|^2}{2\sigma_{i'}^{\langle 0\rangle 2}}\right\}\\
&\geq \exp\left\{\frac{d}{2}\ln\frac{\sigma_j^2(1-2\varepsilon_o)}{\sigma_i^2(1+\varepsilon_o)}+\frac{(\sigma_i^2+\sigma_j^2+c_{ij}^2\sigma_{ij}^2)d\cdot(1-2\varepsilon_o)}{2\sigma_j^2(1+\varepsilon_o)}-\frac{2\sigma_i^2 d(1+\varepsilon_o)}{2\sigma_i^2(1-2\varepsilon_o)}\right\}\\
&\geq \exp\left\{d\left(\frac{1}{2}\ln\frac{\sigma_j^2}{\sigma_i^2}(1-3\varepsilon_o)+\frac{\sigma_i^2+\sigma_j^2+c_{ij}^2\sigma_{ij}^2}{2\sigma_j^2}(1-3\varepsilon_o)-1-\frac{7\varepsilon_o}{2}\right)\right\}\\
&\geq \exp\left\{d\left(\ln(1-3\varepsilon_o)+\frac{c_{ij}^2\sigma_{ij}^2}{2\sigma_j^2}(1-3\varepsilon_o)-5\varepsilon_o\right)\right\}\quad\geq\ \exp\left(\frac{c_{ij}^2\sigma_{ij}^2}{\sigma_j^2}\cdot\frac{d}{8}\right).
\end{aligned}$$

The second-last step uses $\ln(\sigma_j^2/\sigma_i^2)+(\sigma_i^2/\sigma_j^2)(1-3\varepsilon_o)\geq 1+\ln(1-3\varepsilon_o)$, which can be obtained easily from the more familiar $x\geq 1+\ln x$. ∎

In the case where all the Gaussians have the same variance, this lemma says that each data point is given weight at most $e^{-c^2d/8}$ by any center-estimate from a different cluster. When the Gaussians do not have the same variance, the lemma is even stronger. In particular, if there is a small cluster S_i right near a large one, S_j, we can conclude that center-estimates from the small cluster assign weight at most $e^{-c^2\sigma_j^2 d/\sigma_i^2}$ to points from the big cluster. We need this stronger bound in our subsequent analysis: the two clusters could be at such different scales that a point from the large cluster, because it has substantial norm, could significantly throw off the center-estimate from the small cluster.

We are now in a position to assess what happens during the first round of EM. At the end of this round, let C_j denote the center-estimates originally from S_j which have high mixing weight, that is, $C_j = \{\mu_{j'}^{\langle 1\rangle} : \mu_{j'}^{\langle 0\rangle} \in S_j, w_{j'}^{\langle 1\rangle} \geq w_T\}$. We will see that such center-estimates move quite a bit closer to their respective true centers μ_j as a result of the first EM update.

Lemma 9 *Under conditions (C2), (C3), and (C4), any "non-starved" center-estimate $\mu_{i'}^{\langle 1\rangle} \in C_i$ has*

$$\|\mu_{i'}^{\langle 1\rangle} - \mu_i\| \leq \frac{1}{8}\min(1,c)\sigma_i\sqrt{d}.$$

Proof. Bound $\|\mu_{i'}^{\langle 1\rangle} - \mu_i\|$ by the sum of two terms:

$$\begin{aligned}
\|\mu_{i'}^{\langle 1\rangle} - \mu_i\| &= \left\| \frac{\sum_{x\in S} p_{i'}^{\langle 1\rangle}(x)(x-\mu_i)}{\sum_{x\in S} p_{i'}^{\langle 1\rangle}(x)} \right\| \\
&\leq \frac{\|\sum_{x\in S_i} p_{i'}^{\langle 1\rangle}(x)(x-\mu_i)\| + \|\sum_{x\notin S_i} p_{i'}^{\langle 1\rangle}(x)(x-\mu_i)\|}{\sum_{x\in S} p_{i'}^{\langle 1\rangle}(x)} \\
&\leq \frac{\|\sum_{x\in S_i} p_{i'}^{\langle 1\rangle}(x)(x-\mu_i)\|}{\sum_{x\in S_i} p_{i'}^{\langle 1\rangle}(x)} + \frac{\|\sum_{j\neq i}\sum_{x\in S_j} p_{i'}^{\langle 1\rangle}(x)(x-\mu_i)\|}{\sum_{x\in S} p_{i'}^{\langle 1\rangle}(x)}.
\end{aligned}$$

The first term can be bounded using Lemma 5, provided $p_{i'}^{\langle 1\rangle}(S_i)$ is substantial. By Lemma 8, for any $x \in S_j, j \neq i$, we have $p_{i'}^{\langle 1\rangle}(x) \leq e^{-c_{ij}^2\sigma_{ij}^2 d/8\sigma_i^2} \leq e^{-c^2 d/8}$. Thus

$$\sum_{x\in S_i} p_{i'}^{\langle 1\rangle}(x) \;\geq\; \sum_{x\in S} p_{i'}^{\langle 1\rangle}(x) - \sum_{j\neq i}\sum_{x\in S_j} p_{i'}^{\langle 1\rangle}(x) \;\geq\; mw_T - me^{-c^2 d/8} \;\geq\; \tfrac{1}{2}mw_T + 1$$

using (C3) and (C4). Lemmas 5 and 6, together with (C4), then give

$$\frac{\|\sum_{x\in S_i} p_{i'}^{\langle 1\rangle}(x)(x-\mu_i)\|}{\sum_{x\in S_i} p_{i'}^{\langle 1\rangle}(x)} \;\leq\; \sigma_i\sqrt{d}\cdot\sqrt{6\max\left(\frac{2}{mw_T}, \frac{1}{d}\ln\frac{2e|S_i|}{mw_T}\right)} \;\leq\; \frac{1}{12}\min(1,c)\sigma_i\sqrt{d}.$$

For the second half of the $\|\mu_{i'}^{\langle 1\rangle} - \mu_i\|$ expression, we observe that for any $x \in S_j, j \neq i$, by Lemma 4 $\|x-\mu_i\| \leq \sqrt{(\sigma_j^2 + c_{ij}^2\sigma_{ij}^2)d(1+2\varepsilon_o)}$. Using the conditions on d and ε_o, this can be upper-bounded by $e^{c_{ij}^2\sigma_{ij}^2 d/16\sigma_i^2}\sigma_i\sqrt{d}$. Thus $p_{i'}^{\langle 1\rangle}(x)\|x-\mu_i\| \leq e^{-c^2 d/16}\sigma_i\sqrt{d}$, and

$$\begin{aligned}
\frac{\|\sum_{j\neq i}\sum_{x\in S_j} p_{i'}^{\langle 1\rangle}(x)(x-\mu_i)\|}{\sum_{x\in S} p_{i'}^{\langle 1\rangle}(x)} &\leq \frac{1}{mw_T}\sum_{j\neq i}\sum_{x\in S_j} p_{i'}^{\langle 1\rangle}(x)\|x-\mu_i\| \\
&\leq \frac{1}{w_T}e^{-c^2 d/16}\sigma_i\sqrt{d} \;\leq\; \frac{1}{24}\min(1,c)\sigma_i\sqrt{d},
\end{aligned}$$

completing the proof. ▮

We also need to analyze the variance-estimates. These started off excellent, and so we mostly need to check that they don't degrade too much during the first round of EM. The difficulty with the usual formula for variance is that it involves terms of the form $\|x-\mu_i^{\langle 1\rangle}\|^2$, whereas we only have tight concentration bounds for terms like $\|x-\mu_i\|^2$. To cope with this, we first derive an alternative expression for the variance.

Lemma 10 (Alternative formula for variance) *For any i, and any choice of $\mu \in \mathbb{R}^d$, the formula for $\sigma_i^{\langle t\rangle 2}$ can be rewritten thus:*

$$\sigma_i^{\langle t\rangle 2} \;=\; \frac{\sum_x p_i^{\langle t\rangle}(x)\|x-\mu_i^{\langle t\rangle}\|^2}{d\sum_x p_i^{\langle t\rangle}(x)} \;=\; \frac{\sum_x p_i^{\langle t\rangle}(x)\|x-\mu\|^2}{d\sum_x p_i^{\langle t\rangle}(x)} - \frac{\|\mu-\mu_i^{\langle t\rangle}\|^2}{d}.$$

Proof. Consider the distribution over S which assigns point $x \in S$ a probability mass proportional to $p_i^{\langle t\rangle}(x)$. Taking expectations over X drawn from this distribution, we have $\mathbb{E}X = \mu_i^{\langle t\rangle}$, and for any $\mu \in \mathbb{R}^d$,

$$\mathbb{E}\|X-\mu\|^2 \;=\; \mathbb{E}\|X\|^2 + \|\mu\|^2 - 2\mu\cdot\mathbb{E}X \;=\; \mathbb{E}\|X\|^2 + \|\mu\|^2 - 2\mu\cdot\mu_i^{\langle t\rangle}$$

and similarly

$$\mathbb{E}\|X-\mu_i^{\langle t\rangle}\|^2 \;=\; \mathbb{E}\|X\|^2 + \|\mu_i^{\langle t\rangle}\|^2 - 2\mu_i^{\langle t\rangle}\cdot\mathbb{E}X \;=\; \mathbb{E}\|X\|^2 - \|\mu_i^{\langle t\rangle}\|^2.$$

Subtracting,

$$\mathbb{E}\|X-\mu\|^2 - \mathbb{E}\|X-\mu_i^{\langle t\rangle}\|^2 \;=\; \|\mu\|^2 - 2\mu\cdot\mu_i^{\langle t\rangle} + \|\mu_i^{\langle t\rangle}\|^2 \;=\; \|\mu-\mu_i^{\langle t\rangle}\|^2,$$

which is a paraphrase of the lemma statement. ▮

It is now simpler to bound the variances at the end of first round of EM.

Lemma 11 (Variance estimates in round one) *Under conditions (C2), (C3), and (C4), for any $i' \in C_i$,*

$$\sigma_i^2\left(1-\frac{3}{64}\min(1,c^2)\right) \;\le\; \sigma_{i'}^{\langle 1\rangle 2} \;\le\; \sigma_i^2\left(1+\frac{1}{32}\right).$$

Proof. By Lemma 10,

$$\sigma_{i'}^{\langle 1\rangle 2} \;=\; \frac{\sum_{x\in S} p_{i'}^{\langle 1\rangle}(x)\|x-\mu_i\|^2}{d\sum_{x\in S} p_{i'}^{\langle 1\rangle}(x)} - \frac{\|\mu_i-\mu_{i'}^{\langle 1\rangle}\|^2}{d}.$$

First let us lower-bound this, using Lemmas 4(c), 8, and 9.

$$\begin{aligned}
\sigma_{i'}^{\langle 1\rangle 2} &\geq \frac{\sum_{x\in S_i} p_{i'}^{\langle 1\rangle}(x)\|x-\mu_i\|^2}{d\sum_{x\in S} p_{i'}^{\langle 1\rangle}(x)} - \frac{\|\mu_i - \mu_{i'}^{\langle 1\rangle}\|^2}{d} \\
&\geq \frac{\sum_{x\in S_i} p_{i'}^{\langle 1\rangle}(x)\cdot\sigma_i^2 d(1-\varepsilon_o)}{d\sum_{x\in S} p_{i'}^{\langle 1\rangle}(x)} - \frac{\min(1,c^2)}{64}\sigma_i^2 \\
&\geq \sigma_i^2(1-\varepsilon_o)\cdot\frac{p_{i'}^{\langle 1\rangle}(S) - \sum_{j\neq i} p_{i'}^{\langle 1\rangle}(S_j)}{p_{i'}^{\langle 1\rangle}(S)} - \frac{\min(1,c^2)}{64}\sigma_i^2 \\
&\geq \sigma_i^2(1-\varepsilon_o)\cdot\left(1-\frac{me^{-c^2d/8}}{mw_T}\right) - \frac{\min(1,c^2)}{64}\sigma_i^2.
\end{aligned}$$

For the upper bound, we again use Lemmas 4(c) and 8, along with the conditions on d and ε_o, to assert that for points $x\in S_j$, $j\neq i$, we have $p_{i'}^{\langle 1\rangle}(x)\|x-\mu_i\|^2 \leq e^{-c_{ij}^2\sigma_{ij}^2 d/8\sigma_i^2}(\sigma_j^2 + c_{ij}^2\sigma_{ij}^2)d\cdot(1+2\varepsilon_o) \leq e^{-c^2d/16}\sigma_i^2 d$, and thus

$$\begin{aligned}
\sigma_{i'}^{\langle 1\rangle 2} &\leq \frac{\sum_{x\in S_i} p_{i'}^{\langle 1\rangle}(x)\|x-\mu_i\|^2 + \sum_{j\neq i}\sum_{x\in S_j} p_{i'}^{\langle 1\rangle}(x)\|x-\mu_i\|^2}{d\sum_{x\in S} p_{i'}^{\langle 1\rangle}(x)} \\
&\leq \frac{\sum_{x\in S_i} p_{i'}^{\langle 1\rangle}(x)\cdot\sigma_i^2 d(1+\varepsilon_o)}{dp_{i'}^{\langle 1\rangle}(S)} + \frac{\sum_{j\neq i}\sum_{x\in S_j} e^{-c^2d/16}\sigma_i^2 d}{dp_{i'}^{\langle 1\rangle}(S)} \\
&\leq \sigma_i^2(1+\varepsilon_o) + \frac{me^{-c^2d/16}\sigma_i^2}{mw_T}.
\end{aligned}$$

The rest follows by substituting in conditions (C2) and (C3). ∎

Remark Henceforth we will assume that the conclusions of Lemmas 9 and 11 hold.

7. Pruning

We now know each center-estimate in C_j is accurate within $\frac{1}{8}\min(1,c)\sigma_j\sqrt{d}$. A simple clustering heuristic due to Gonzalez (1985), described in Section 2.4, is used to choose k points from $\cup_j C_j$.

Lemma 12 *Under condition (C3), the sets C_i obey the following properties.*
(a) Each C_i is non-empty.
(b) There is a real value $\Delta > 0$ such that if $x\in C_i$ and $y,z\in C_j$ $(i\neq j)$ then $d(y,z)\leq\Delta$ and $d(x,y)>\Delta$.
(c) The pruning procedure identifies exactly one member of each C_i.

Proof. (a) From Lemmas 4 and 7 we know that $|S_i|\geq\frac{1}{2}mw_i$, and that at most $\frac{3}{2}lw_i$ initial center-estimates are chosen from S_i. It was seen in Lemma 9 that each point in S_i gives weight at least $1-le^{-c^2d/8}$ to center-estimates from S_i. It follows that at the end of the first round of EM, at least one of these center-estimates must have mixing weight at least

$$\frac{(\frac{1}{2}mw_i)(1-le^{-c^2d/8})}{m\cdot\frac{3}{2}lw_i} = \frac{1}{3l}\cdot(1-le^{-c^2d/8}) \geq \frac{1}{4l} = w_T$$

and therefore C_i cannot be empty.
(b) Pick $\mu_{i'}^{\langle 1\rangle} \in C_i$ and $\mu_{j'}^{\langle 1\rangle}, \mu_{j''}^{\langle 1\rangle} \in C_j$ for any pair $i \neq j$. Then, from Lemmas 9 and 11,

$$d(\mu_{j'}^{\langle 1\rangle}, \mu_{j''}^{\langle 1\rangle}) \;=\; \frac{\|\mu_{j'}^{\langle 1\rangle} - \mu_{j''}^{\langle 1\rangle}\|}{\sigma_{j'}^{\langle 1\rangle} + \sigma_{j''}^{\langle 1\rangle}} \;<\; \frac{2 \cdot \frac{c}{8}\sigma_j\sqrt{d}}{2 \cdot \frac{61}{64}\sigma_j} \;<\; \frac{c}{6}\sqrt{d}.$$

Call this value Δ. Meanwhile,

$$\begin{aligned} d(\mu_{i'}^{\langle 1\rangle}, \mu_{j'}^{\langle 1\rangle}) \;=\; \frac{\|\mu_{i'}^{\langle 1\rangle} - \mu_{j'}^{\langle 1\rangle}\|}{\sigma_{i'}^{\langle 1\rangle} + \sigma_{j'}^{\langle 1\rangle}} \;&>\; \frac{c_{ij}\sigma_{ij}\sqrt{d} - \frac{c}{8}\sigma_i\sqrt{d} - \frac{c}{8}\sigma_j\sqrt{d}}{\frac{33}{32}\sigma_i + \frac{33}{32}\sigma_j} \\ &\geq\; \frac{c_{ij} \cdot \frac{1}{2}(\sigma_i + \sigma_j)\sqrt{d} - \frac{c}{8}(\sigma_i + \sigma_j)\sqrt{d}}{\frac{33}{32}(\sigma_i + \sigma_j)} \;\geq\; \frac{c}{3}\sqrt{d}, \end{aligned}$$

strictly greater than Δ.
(c) There are k true clusters and the pruning procedure picks exactly k center-estimates. It will not pick two from the same true cluster because these must be at distance $\leq \Delta$ from each other, whereas there must be some untouched cluster containing a center-estimate at distance $> \Delta$ from all points selected thus far. ▮

8. The Second Round of EM

We now have one center-estimate $\mu_i^{\langle 1\rangle}$ per true cluster (for convenience permute their labels to match the S_i), each with mixing weight $\frac{1}{k}$ and covariance $\sigma_i^{\langle 1\rangle 2} I_d$. Furthermore each $\mu_i^{\langle 1\rangle}$ is within distance $\frac{1}{8}\min(1,c)\sigma_i\sqrt{d}$ of the corresponding true Gaussian center μ_i. Such favorable circumstances make it easy to show that the subsequent round of EM achieves near-perfect clustering.

There is just one tricky issue. As in the first round, in order to bound EM's soft assignments, we'd like to assert that the distances $\|x - \mu_i^{\langle 1\rangle}\|^2$ are tightly concentrated around certain values. What is different this time, however, is the statistical dependency between data points x and center-estimates $\mu_i^{\langle 1\rangle}$. To avoid having to manage this dependency, we instead recall that the $\mu_i^{\langle 1\rangle}$ are just weighted averages of data points, and as far as possible, we rewrite expressions like $\|x - \mu_i^{\langle 1\rangle}\|^2$ as weighted sums of expressions involving only data points, such as $\|x - y\|^2$ or $x \cdot y$. There is one particular kind of dot product which will be an especially useful building block, and we start by analyzing it.

Lemma 13 *Pick any $1 \leq i, j \leq k$. Under conditions (C2) and (C3), for any $x \in S_i$,*

$$|(x - \mu_i) \cdot (\mu_j^{\langle 1\rangle} - \mu_j)| \leq \frac{3}{2}\varepsilon_o\sigma_{ij}^2 d.$$

Proof. We could use the fact that $\mu_j^{\langle 1\rangle}$ is reasonably close to μ_j, but this doesn't give a very good bound. Instead we notice that $(x - \mu_i) \cdot (\mu_j^{\langle 1\rangle} - \mu_j)$ is a weighted average of terms of the form $(x - \mu_i) \cdot (y - \mu_j)$, as y varies over S. When $y \in S_j$, these dot products are small, as seen in Lemma 4(d). And when $y \notin S_j$, the weight of the term is small.

Specifically, suppose $y \in S_g$ for $g \neq j$. By Lemmas 4(d) and 8,

$$p_j^{\langle 1\rangle}(y)\,|(x-\mu_i)\cdot(y-\mu_j)| \leq e^{-c_{jg}^2\sigma_{jg}^2 d/8\sigma_j^2}\cdot\sqrt{(\sigma_g^2+c_{jg}^2\sigma_{jg}^2)(1+2\varepsilon_o)}\cdot\sigma_i\varepsilon_o d.$$

By the conditions on d, this is at most $e^{-c^2d/16}\sigma_i\sigma_j\varepsilon_o d$. Hence

$$\begin{aligned}
|(x-\mu_i)\cdot(\mu_j^{\langle 1\rangle}-\mu_j)| &= \left|\frac{\sum_y p_j^{\langle 1\rangle}(y)(x-\mu_i)\cdot(y-\mu_j)}{p_j^{\langle 1\rangle}(S)}\right| \\
&\leq \frac{\sum_{y\in S_j} p_j^{\langle 1\rangle}(y)|(x-\mu_i)\cdot(y-\mu_j)|}{p_j^{\langle 1\rangle}(S_j)} + \frac{\sum_{g\neq j}\sum_{y\in S_g} p_j^{\langle 1\rangle}(y)|(x-\mu_i)\cdot(y-\mu_j)|}{p_j^{\langle 1\rangle}(S)} \\
&\leq \sigma_i\sigma_j\varepsilon_o d\sqrt{1+2\varepsilon_o} + \frac{1}{mw_T}\sum_{g\neq j}\sum_{y\in S_g} e^{-c^2d/16}\sigma_i\sigma_j\varepsilon_o d \\
&\leq \varepsilon_o\sigma_{ij}^2 d\cdot\left(1+\varepsilon_o+\frac{1}{w_T}e^{-c^2d/16}\right).
\end{aligned}$$

Under (C2) and (C3), the term in parentheses is at most $3/2$. ∎

We now develop a counterpart of Lemma 8, a bound on how "soft" EM's assignments can be.

Lemma 14 *Under conditions (C1), (C2), and (C3), for any $1 \leq i \neq j \leq k$, and for any $x \in S_i$,*

$$\frac{p_i^{\langle 2\rangle}(x)}{p_j^{\langle 2\rangle}(x)} \;\geq\; \exp\left(\frac{c_{ij}^2\sigma_{ij}^2 d}{8\sigma_j^2}\right).$$

Proof. This is mostly a matter of confirming that $\|x-\mu_i^{\langle 1\rangle}\|, \|x-\mu_j^{\langle 1\rangle}\|, \sigma_i^{\langle 1\rangle}, \sigma_j^{\langle 1\rangle}$ are not too far from $\|x-\mu_i\|, \|x-\mu_j\|, \sigma_i, \sigma_j$. Making use of Lemmas 4(c), 9, and 13,

$$\begin{aligned}
\|x-\mu_i^{\langle 1\rangle}\|^2 &= \|x-\mu_i\|^2+\|\mu_i-\mu_i^{\langle 1\rangle}\|^2+2(x-\mu_i)\cdot(\mu_i-\mu_i^{\langle 1\rangle}) \\
&\leq \sigma_i^2 d(1+\varepsilon_o)+\frac{\min(1,c^2)}{64}\sigma_i^2 d+3\varepsilon_o\sigma_i^2 d \\
&\leq \sigma_i^2 d\left(1+\frac{1}{16}\min(1,c^2)\right)
\end{aligned}$$

In much the same vein,

$$\begin{aligned}
\|x-\mu_j^{\langle 1\rangle}\|^2 &= \|x-\mu_j\|^2+\|\mu_j-\mu_j^{\langle 1\rangle}\|^2+2((x-\mu_i)+(\mu_i-\mu_j))\cdot(\mu_j-\mu_j^{\langle 1\rangle}) \\
&\geq (\sigma_i^2+c_{ij}^2\sigma_{ij}^2)d(1-2\varepsilon_o)-3\varepsilon_o\sigma_{ij}^2 d-2c_{ij}\sigma_{ij}d^{1/2}\cdot\frac{\min(1,c)}{8}\sigma_j d^{1/2} \\
&\geq \sigma_i^2 d\left(1-\frac{1}{48}\min(1,c^2)\right)+c_{ij}^2\sigma_{ij}^2 d\left(1-\frac{29}{96}\right).
\end{aligned}$$

Using these and the variance bound from Lemma 11, the rest follows in the same manner as Lemma 8. ∎

Henceforth, assume this condition holds true. The best we can now hope for is that each final estimate $\mu_i^{\langle 2\rangle}$ is exactly the mean of S_i. In fact, it will turn out that the error $\|\mu_i^{\langle 2\rangle}-\mu_i\|$ is not too different from $\|\text{mean}(S_i)-\mu_i\|$.

Lemma 15 *Under conditions (C1)–(C4), for each i,*

$$\|\mu_i^{\langle 2\rangle} - \mu_i\| \;\; \leq \;\; \|\mathrm{mean}(S_i) - \mu_i\| + \frac{5}{w_{min}} e^{-c^2 d/16} \sigma_i \sqrt{d}.$$

Proof. As usual, we start by separating points in S_i from those outside.

$$\begin{aligned}
\|\mu_i^{\langle 2\rangle} - \mu_i\| &= \left\| \frac{\sum_{x\in S} p_i^{\langle 2\rangle}(x)(x-\mu_i)}{\sum_{x\in S} p_i^{\langle 2\rangle}(x)} \right\| \\
&\leq \frac{1}{p_i^{\langle 2\rangle}(S)} \left\| \sum_{x\in S_i} p_i^{\langle 2\rangle}(x)(x-\mu_i) \right\| + \frac{1}{p_i^{\langle 2\rangle}(S)} \sum_{j\neq i} \sum_{x\in S_j} p_i^{\langle 2\rangle}(x) \|x-\mu_i\|.
\end{aligned}$$

With Lemma 14 in hand, these two terms are straightforward to bound. For instance, we know that for any $x \in S_i$, $p_i^{\langle 2\rangle}(x) \geq 1 - ke^{-c^2d/8}$. Hence the first term is at most

$$\begin{aligned}
&\frac{1}{p_i^{\langle 2\rangle}(S)} \left(\left\| \sum_{x\in S_i} (1-ke^{-c^2d/8})(x-\mu_i) \right\| + \left\| \sum_{x\in S_i} (p_i^{\langle 2\rangle}(x) - (1-ke^{-c^2d/8}))(x-\mu_i) \right\| \right) \\
\leq\;& \frac{\left\| \sum_{x\in S_i} (1-ke^{-c^2d/8})(x-\mu_i) \right\|}{|S_i| \cdot (1-ke^{-c^2d/8})} + \frac{1}{p_i^{\langle 2\rangle}(S)} \sum_{x\in S_i} ke^{-c^2d/8} \|x-\mu_i\| \\
\leq\;& \|\mathrm{mean}(S_i) - \mu_i\| + \frac{|S_i| \cdot ke^{-c^2d/8}}{|S_i| \cdot (1-ke^{-c^2d/8})} \sigma_i \sqrt{d(1+2\varepsilon_o)} \\
\leq\;& \|\mathrm{mean}(S_i) - \mu_i\| + 2ke^{-c^2d/8} \sigma_i \sqrt{d}.
\end{aligned}$$

while the second term is

$$\begin{aligned}
\frac{1}{p_i^{\langle 2\rangle}(S)} \sum_{j\neq i} \sum_{x\in S_j} p_i^{\langle 2\rangle}(x) \|x-\mu_i\| &\leq \frac{1}{p_i^{\langle 2\rangle}(S)} \sum_{j\neq i} \sum_{x\in S_j} e^{-c_{ij}^2 \sigma_{ij}^2 d/8\sigma_i^2} \sqrt{(\sigma_j^2 + c_{ij}^2 \sigma_{ij}^2)(1+2\varepsilon_o)d} \\
&\leq \frac{1}{p_i^{\langle 2\rangle}(S)} \sum_{x\notin S_i} e^{-c^2d/16} \sigma_i \sqrt{d} \\
&\leq \frac{m}{|S_i| \cdot (1-ke^{-c^2d/8})} e^{-c^2d/16} \sigma_i \sqrt{d} \;\leq\; \frac{3}{w_i} e^{-c^2d/16} \sigma_i \sqrt{d}.
\end{aligned}$$

For the very last bound we use $|S_i| \geq \frac{1}{2} m w_i$ (Lemma 4(e)). ▮

Here's a summary of everything we have so far.

Theorem 16 *Suppose that $l > k$, that $w_i \geq w_{min}$ for all i, and that conditions (C1)–(C4) hold. With probability at least $1 - 2m^2 e^{-\varepsilon_o^2 d/8} - 2m^2 k e^{-\varepsilon_o^2 d/2} - 2ke^{-lw_{min}/12} - k(l+1)e^{-lw_{min}}$, the center-estimates returned after two rounds of EM satisfy (for each i):*

$$\|\mu_i^{\langle 2\rangle} - \mu_i\| \;\; \leq \;\; \|\mathrm{mean}(S_i) - \mu_i\| + \frac{5}{w_{min}} e^{-c^2 d/16} \sigma_i \sqrt{d}.$$

Choosing $c^2 d \geq 16 \ln \frac{5}{\varepsilon w_{min}}$ and $\varepsilon_o = \frac{1}{96} \min(1, c^2)$ gives Theorem 1. We can also prove bounds on the final mixing weights and variance.

Theorem 17 *To the results of Theorem 16 it can be added that if $c^2 d \geq 16 \ln \frac{5}{\varepsilon w_{min}}$ (for some $\varepsilon > 0$), then for any i,*

$$\frac{|S_i|}{m} \cdot (1-\varepsilon) \;\leq\; w_i^{\langle 2\rangle} \;\leq\; \frac{|S_i|}{m} + \varepsilon$$

and

$$(1-\varepsilon)\mathrm{var}(S_i) - \frac{\|\mu_i^{\langle 2\rangle} - \mathrm{mean}(S_i)\|^2}{d} \;\leq\; \sigma_i^{\langle 2\rangle 2} \;\leq\; (1+\varepsilon)\mathrm{var}(S_i) + \varepsilon\sigma_i^2 + \frac{\varepsilon\|\mu_i - \mathrm{mean}(S_i)\|^2}{d}$$

(where $\mathrm{var}(S_i)$ *is the empirical variance of cluster S_i).*

Proof. The bounds on $w_i^{\langle 2\rangle}$ come directly from writing

$$w_i^{\langle 2\rangle} \;=\; \frac{1}{m}\sum_{x\in S} p_i^{\langle 2\rangle}(x) \;=\; \frac{1}{m}\left(\sum_{x\in S_i} p_i^{\langle 2\rangle}(x) + \sum_{x\notin S_i} p_i^{\langle 2\rangle}(x)\right)$$

and then using our old bounds $p_i^{\langle 2\rangle}(x) \geq 1 - ke^{-c^2 d/8}$ for $x \in S_i$ and $p_i^{\langle 2\rangle}(x) \leq e^{-c^2 d/8}$ for $x \notin S_i$ (Lemma 14).

For the variance, we again exploit the alternative formulation of Lemma 10, but this time in a slightly different way. Let $\mu = \mathrm{mean}(S_i)$.

$$\sigma_i^{\langle 2\rangle 2} \;=\; \frac{\sum_{x\in S} p_i^{\langle 2\rangle}\|x-\mu\|^2}{d\, p_i^{\langle 2\rangle}(S)} - \frac{\|\mu_i^{\langle 2\rangle} - \mu\|^2}{d}.$$

Thus:

$$\begin{aligned}
\sigma_i^{\langle 2\rangle 2} \;&\geq\; \frac{\sum_{x\in S_i}(1-ke^{-c^2 d/8})\|x-\mu\|^2}{d(|S_i| + me^{-c^2 d/8})} - \frac{\|\mu_i^{\langle 2\rangle} - \mu\|^2}{d} \\
&=\; \frac{\sum_{x\in S_i}\|x-\mathrm{mean}(S_i)\|^2}{d|S_i|} \cdot \frac{1-ke^{-c^2 d/8}}{1+me^{-c^2 d/8}/|S_i|} - \frac{\|\mu_i^{\langle 2\rangle} - \mu\|^2}{d},
\end{aligned}$$

the first term of which is recognizable as $\mathrm{var}(S_i)$. Similarly,

$$\begin{aligned}
\sigma_i^{\langle 2\rangle 2} \;&\leq\; \frac{1}{d|S_i|(1-ke^{-c^2 d/8})}\left\{\sum_{x\in S_i}\|x-\mu\|^2 + \sum_{x\notin S_i} e^{-c^2 d/8}\cdot 2(\|x-\mu_i\|^2 + \|\mu_i-\mu\|^2)\right\} \\
&\leq\; \frac{\mathrm{var}(S_i)}{1-ke^{-c^2 d/8}} + \frac{2m}{|S_i|(1-ke^{-c^2 d/8})}\left(e^{-c^2 d/16}\sigma_i^2 + e^{-c^2 d/8}\frac{\|\mu_i-\mu\|^2}{d}\right),
\end{aligned}$$

from which the claim follows directly. ∎

9. Concluding Remarks

This paper provides a principled basis for answering some important questions surrounding EM: how many clusters should be used, how the parameters ought to be initialized, and how pruning should be carried out. These results may be of interest to practitioners of EM.

But what about the claim that EM can be made to work in just two rounds? This requires what we call the

> **Strong Gaussian assumption.** The data are i.i.d. samples from a true mixture of Gaussians.

This assumption is the standard setting for other theoretical results about EM, but is it reasonable to expect of real data sets? We recommend instead the

> **Weak Gaussian assumption.** The data looks like it comes from a particular mixture of Gaussians in the following sense: for any sphere in $\mathbb{R}^d$, the fraction of the data that falls in the sphere is the expected fraction under the mixture distribution, $\pm\varepsilon_0$, where ε_0 is some term corresponding to sampling error and will typically be proportional to $m^{-1/2}$, where m is the number of samples. Some other concept class of polynomial VC dimension can be used in place of spheres.

The strong assumption immediately implies the weak assumption (with high probability) by a large deviation bound, since the concept class of spheres in $\mathbb{R}^d$ has VC dimension just $d+1$ (Dudley, 1979; Haussler, 1992). What kinds of conclusions can we draw from the strong assumption but not the weak one? Here is an example: "if two data points are drawn from $N(0,I_d)$ then with overwhelming probability they are separated by a distance of at least $\sqrt{d}$". The weak assumption does not support this; with just two samples, in fact, the sampling error is so high that it does not allow us to draw any non-trivial conclusions at all.

It is often argued that the Gaussian is the most natural model of a cluster because of the central limit theorem. However, central limit theorems, more specifically Berry-Esséen theorems (Feller, 1966), yield Gaussians in the sense of the weak assumption, not the strong one. For the same reason, the weak Gaussian assumption arises naturally when we take random projections of mixtures of product distributions (Diaconis and Freedman, 1984). Ideally therefore, we could provide performance guarantees for EM under just this condition. It might be possible to extend our analysis appropriately; for an example of what needs to be changed in the algorithm, consider that the weak assumption allows $\sqrt{m}$ data points to be placed arbitrarily in space, and therefore an outlier removal procedure is probably necessary.

Acknowledgments

The first author is indebted to Daniel Hsu for suggesting many simplifications to the analysis, to the reviewers for significantly improving the presentation, and to the NSF for support under grant IIS-0347646.

References

D. Achlioptas and F. McSherry. On spectral learning of mixtures of distributions. In *Proceedings of the 18th Conference on Learning Theory*, 458–469, 2005.

S. Arora and R. Kannan. Learning mixtures of separated nonspherical Gaussians. *Annals of Applied Probability*, 15(1A):69–92, 2005.

C. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, New York, 1995.

S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, 634–644, 1999.

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

P. Diaconis and D. Freedman. Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12:793–815, 1984.

R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* John Wiley, New York, 1973.

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

R. Durrett. *Probability: Theory and Examples.* Duxbury, Belmont, California, 1996.

W. Feller. *An Introduction to Probability Theory and its Applications*, vol. II. John Wiley, New York, 1966.

T. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

D. Haussler. Decision-theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

M. Kearns, Y. Mansour, and A. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, 1997.

M. Kearns and U.V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, Massachusetts, 1994.

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

D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley, London, 1985.

S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. *Journal of Computer and Systems Sciences*, 68(4):841–860, 2004.

S. Vempala, R. Kannan, and H. Salmasian. The spectral method for general mixture models. In *Proceedings of the 18th Conference on Learning Theory*, 2005.

C.F.J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

L. Xu and M.I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8:129–151, 1996.