# When is Clustering Hard?

**Nathan Srebro***
Department of Computer Science, University of Toronto
nati@cs.toronto.edu

**Gregory Shakhnarovich**
Computer Science and Artificial Intelligence Laboratory, MIT
gregory@csail.mit.edu

**Sam Roweis**
Department of Computer Science, University of Toronto
roweis@cs.toronto.edu

Abstract of Work in Progress
PASCAL Workshop on Statistics and Optimization of Clustering, July 2005

We propose questions regarding the informational and computational limits of learning a mixture of Gaussians—the sample sizes necessary in order to recover the generating mixture with unbounded, and bounded, computation. We would like to quantify the gap between the two, i.e. the excess information required for tractable estimation. We report on an empirical study attempting to shed light on these questions.

## 1 Introduction

Consider the optimization problem of clustering a collection $n$ of points in $\mathbb{R}^d$ into $k$ clusters, e.g. by fitting a mixture-of-Gaussians model for the data. Viewed as an optimization problem of optimizing an objective function, such as the likelihood or the sum of squared distances to cluster centers as in $k$-means, this problem is "hard" in the traditional worst-case complexity sense. That is, as the number of samples increase, the worst-case running time of algorithms that are guaranteed to exactly optimize the objective increases sharply.[1] On the other hand, when the data really is clustered, and enough data is available, local search methods typically succeed in optimizing the objective and recovering the clustering. This leads to the conventional wisdom that *"clustering is not hard—it is either easy, or not interesting"*. Our goal is to evaluate this statement quantitatively and establish whether there is a regime in which clustering is "hard" even though an "interesting" cluster structure does exist in the data.

Lately, a series of results established that if the data is generated from adequately separated mixture of Gaussians, and enough data is available, then clustering is in fact easy—polynomial time algorithms exist that can recover, with high probability, the correct clustering, and so also the correct centers up to some requested error. These results provide an upper bound on the **computational limit** of clustering—the minimum required seperation and minimum required sample size for which recovering the clustering is tractable.

When not enough samples are available, even when the data is generated from a well-separated mixture of Gaussians, the correct model cannot be recovered, simply because

*Presenting author

[1] More precisely: exponentially for any known algorithm, and assuming $P \neq NP$, super-polynomial at least for any algorithm minimizing the $k$-means objective.

there is not enough information in the data. Even if the objective is optimized on the training samples (e.g., the maximum likelihood clustering is found), the solution in terms of cluster assignments or centers may be quite different from the model that actually generated the data. Loosley speaking, there are enough incorrect clusterings, and enough random variation in the likelihood of incorrect clusterings, such that one of them happens to have likelihood higher than the correct clustering. However, we also know that for any mixture model, when enough data is available, the maximum likielhood solution *will* converge to the true generating model, while other local maxima of the likelihood function will be lower then it [RW84].

Ignoring computational issues, one can ask: When is there enough information in the data to recover the correct clustering? Focusing on the likelihood, how many data points are necessary for the maximum likelihood model to be close to the correct clustering—what is the **informational limit** of clustering?

We would like to study these two limits and the relationship between them. Is the computational limit higher then the informational limit? If so, can one quantify the *excess information* needed for computational tractablitiy?

The discussion thus far refers to a scenario in which the data is sampled from the well-separated mixture of Gaussians. It is possible that the "true" generating process does not exactly follow this model, but the data is still separated enough into localized clusters. In such cases it is still possible to recover the clustering by fitting a Gaussian mixture model, and local search methods typically suffice if enough data is available. One can therefore hope to extend the analysis also to such scenarios, characterizing the properties of the true clustering that make it recoverable with a large enough sample.

Ultimately, we would like to obtain sharp theoretical quantitative evaluations of these limits, as functions of the model parameters, such as number of clusters, dimensionality and separation, rather then loose asymptotic bounds. Our long-term goal is to obtain a computational limit that is independent of the algorithm, such that above this limit we know how the clustering can be recovered, and beneath the limit no algorithm can recover the clustering with high probability (under some complexity assumptions). We would like to understand the nature of the transition at this limit: how quickly do problems change from hard to easy? We would also like an informational limit that applies to any (possibly intractable) estimation criterion, not only to maximum-likelihood estimation.

We are yet far from answering these questions. In order to begin addressing them, we start with a limited empirical study which we describe here. We hope that this study can serve as a staring point and guide us in useful directions.

We focus on the simplest possible setting—a uniform mixture of isotropic Gaussian with equal variance, centered at the vertices of a simplex (i.e. all pairs of centers are at the same distance from each other). The mixing probabilities, variances and the number of clusters in this scenario are known, and only the centers need to be estimated. We generate random samples from such models, and then try to recover the clustering using standard techniques. We also attempt to establish the true maximum likelihood solution utilizing our knowledge of the true model generating the data. We study the samples sizes, with respect to the various parameter settings (dimensionality, number of clusters and seperation) that are sufficient for the suspected maximum-likelihood solution to be close to the correct clustering, and the sample sizes which are sufficient for "fair" methods not utilizing knowledge of the correct centers to find this solution. We observe a clear gap between the two thresholds.

## 2 Background

We briefly survey here relevant aspects of recent work describing algorithms which are guaranteed to recover a well-separated Gaussians mixture model, given enough data. We discuss the results in the context of a uniform mixture of $k$ spherical unit-variance $d$-dimensional Gaussians, with a separation (minimum distance between centers) of $s$.[2]

When the Gaussians are well-separated, the modes of the mixture are at the centers of the Gaussians. Unfortunately, in high dimensions, a very large sample is required in order to identify the modes of a distribution. Dasgupta [Das99] suggested projecting a $d$-dimensional sample to a random subspace of dimension $\Theta(\log k)$, and showed that if the separation between Gaussians is $s > \frac{1}{2}d^{1/2}$, then the modes of the distribution in this subspace still correspond to the centers, and can be identified, with probability $1-\delta$, from a sample of size $k^{\Omega(\log^2 1/\delta)}$. Arora and Kannan [SK01] later improved the minimum required separation to $s = \Omega(d^{1/4}\log(d))$, using either random projections, or a method based on fact that with this separation, distances between points in the same cluster are lower then distances between points in different clusters.

Vempala and Wang [VW04] show that using spectral projections (i.e. projecting the data to the top principal components, as in PCA) instead of random projections allows identifying much less separated Gaussians. They show that with a separation of $s = \Omega(k^{1/4}\log^{1/4} dk)$ and a sample of size of $\Omega(d^3k^2 \log dks/\delta)$, a $k$-dimensional spectral projection of the data preserves enough separation between centers of spherical Gaussians such that after such a projection, the Gaussians can be identified by methods similar to those discussed above. These techniques have recently been extended also to non-spherical Gaussians, where the required separation is $s = \Omega(k^{5/2})$ [SKV04, AM05].

The main thrust of the above results is providing conditions under which the Gaussians are well-separated and easily identifiable (perhaps after a projection), such that even the simplest algorithms can recover them. In general, the results depend on all (or most) distances (after the projection) between points in the same cluster being smaller then distances between points in different clusters. In such extreme cases, local search methods, such as the popular Expectation-Maximization (EM) algorithm can also easily recover the clustering. In fact, Dasgupta and Schulman [DS00] showed that with a separation of $\Omega(d^{1/4})$ and a number of samples polynomial in $k$, two rounds of EM are enough in order to get fairly close to the correct centers. This is provided that instead of searching over models with $k$ centers, the first round of EM uses $\Theta(k \log k)$ centers, and those are then pruned down to $k$ far-away, but well used, centers.

The precise distance-based or mode-based methods suggested by the above results should therefore not be regarded as alternatives to local search heuristics, but rather as analyzable methods. The methods also often involve many parameters that need to be carefully selected, and the theory does not provide for an optimal choice of the parameters for finite sample sizes. In any case, Dasgupta and Schulman's [DS00] result suggests that, perhaps after a spectral projections, we might as well use EM, initially allowing for $O(k \log k)$ centers.

All the results described above require, beyond a large enough sample, also a minimal separation between the Gaussians. A mixture of Gaussians that is not well separated might not correspond to a reasonable "clustering", but a maximum likelihood estimate will still converge to the correct model with enough samples, for any separation. An open issue is whether some minimum separation is required in order for the estimation problem to be tractable, regardless of the sample size. If so, how does this limit compare to the minimal separation in which the mixture corresponds to a "clustering" in some sense (e.g. the modes

[2] Most bounds have a dependence on the eccentricity of the Gaussians, the relationships between covariances of different Gaussians, and the magnitude of the mixing proportions, but we ignore these dependencies for the time being.

are still at the cluster centers, or components of points can be identified with reasonable accuracy when the centers are known)?.

# 3 Methodology

In each experiment, we generate a random data set of $n$ points in $\mathbb{R}^d$ from a "true" mixture of $k$ isotropic unit-variance Gaussian components with equal mixing weights. The true centers $\mu_1, \ldots, \mu_k$ are the vertexes of a $k$-dimensional simplex, so that $\|\mu_i - \mu_j\| = s$ for any $i \neq j$.

We record, for each sample point, the label of the Gaussian component from which it was drawn. The accuracy of an estimated clustering is assessed by comparing the clustering to these "true" labels. Specifically, we measure the label edge-error, that is, the fraction of pairs of points that have the same labels in one clustering, but different labels in the other clustering. Although we estimate mixture models, we derive a clustering from each such estimated model by assigning each point to the nearest center (recall that the priors and variances are uniform and fixed). Note that even the true model often has a positive label edge-error, since for small separations $s$, not all points are closest to the center that generated them.

We attempt to reconstruct the true model using the EM method, ignoring our knowledge of the true centers. We fix the covariance matrices to the true (identity) covariance matrices and the priors to be uniform, and estimate only the centers. We initialize each repetition of EM to a random subsets of points from the sample, and iterate to convergence. We experiment with running EM in the original, high dimensional, space, as well as with running EM after a $k$-dimensional spectral projection (PCA). In the latter case, we first run EM until convergence in the $k$-dimensional space, then lift back up to the original high dimensional space, and continue running EM there until convergence.

We also experiment with running EM with more then $k$ centers. We run EM until convergence with $\lambda k \log(k)$ (for $\lambda = 1 \ldots 5$, where log is the natural logarithm). We then prune the centers using the method suggested by Dasgupta and Schulman [DS00], and run EM again until convergence. Combined with a spectral projection, we experiment with pruning both before and after lifting back to the original space.

We repeat each of the above variations of EM ten times, with different initializations, and consider the solution with the highest likelihood, as well as the distribution of the likelihoods over the different runs.

In order to attempt to find the true maximum likelihood solution, we also run EM where the centers are initialized to the true centers. We do so both with, and without, a spectral projection. This method is by no means guarantees to find the true maximum likelihood solution, and in fact often finds solutions with much lower likelihood then EM from a random initialization. What we find with this method is a peak of the likelihood near the true solutions[3]. At the very least we can say that if this peak is higher then the peaks found by the EM repetitions, then EM failed at finding the maximum likelihood solution. Beyond this, if different iterations of EM converged to different solutions, and this peak is much higher then the spread of likelihoods EM converged to, we have some indication this this might indeed be the maximum likelihood solution.

We refer to the model with the highest likelihood, of all the models we encounter as the "suspected maximum likelihood" solution. We refer to the model with the maximum likelihood of all "fair" runs of EM (not using knowledge of the true centers) as the "EM

[3]This is likely the local maximum likelihood estimate that is guaranteed to be found in the neighborhood of the true solution, and which approaches it with a rate given by the Fisher information [RW84]

maximum likelihood" solution. We compare the sample size required for the label edge-error of the suspected maximum likelihood solution to be close to the label edge-error of the true solution, to the sample size required for label edge-error the EM maximum likelihood solution to be similarly low.

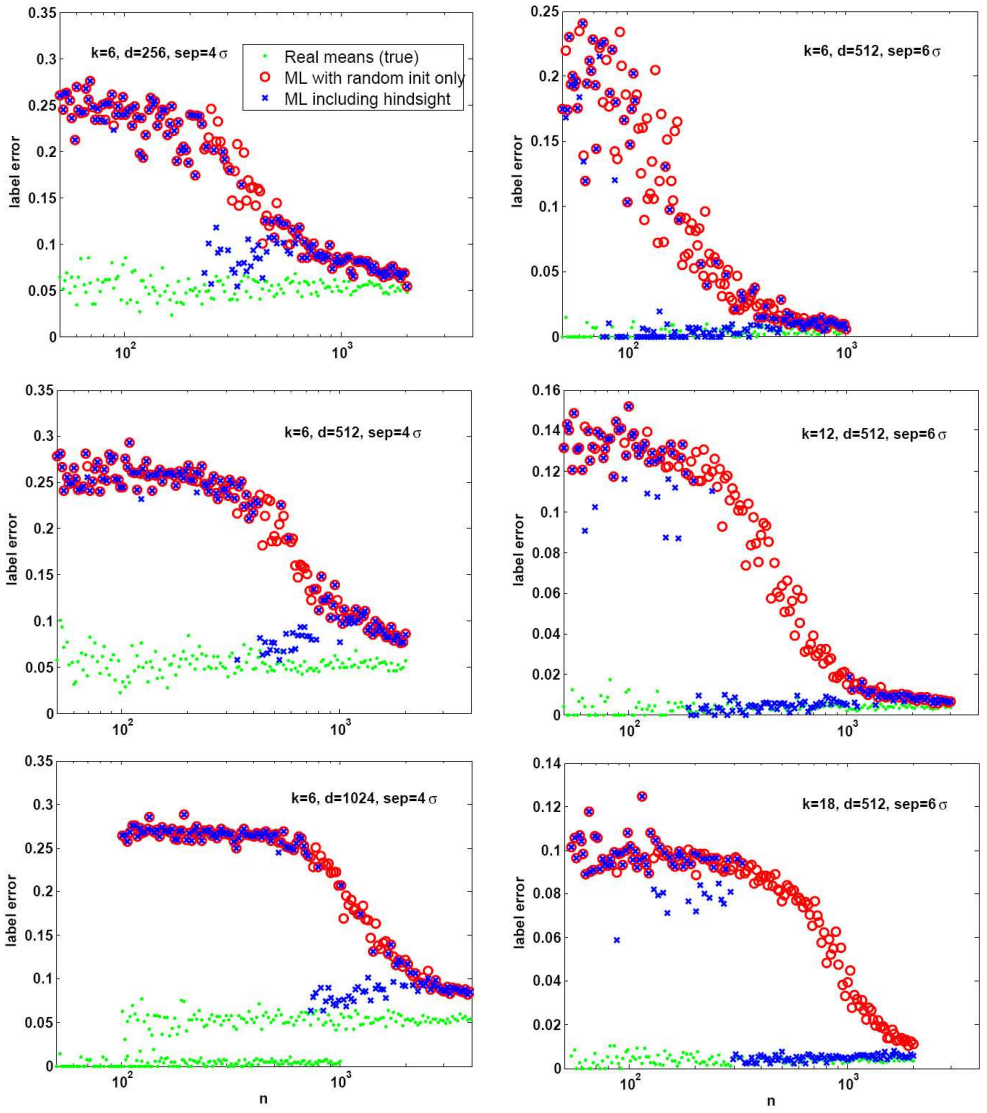


Figure 1: Label edge-error for the true model, suspected maximum likelihood model (with hindsight) and EM maximum likelihood model (no hindsight) for experiments with different number of clusters $k$, dimensionality $d$ and separation (different plots) and sample sizes (horizontal axis)

Figure 1 demonstrates the type of results obtained. Each of the six plots shows the label edge-error for EM maximum likelihood solution and the suspected maximum likelihood solution. Note that with $k$ clusters, only $1/k$ of the pairs of points are in the same cluster in the true clustering, and so a trivial clustering has a label error of $1/k$. A gap can be observed between the sample size in which the suspected maximum likelihood solution achieves low

error, to the sample size in which the EM maximum likelihood solution achieves similar error rates. Observe the increase in the required sample size as the dimensionality (left column) and number of clusters (right column) increase, and as the separation decreases (center row). In order to better understand the relationship between the EM maximum likelihood solution and the local maximum likelihood solution near the true model, we plot, in Figure 2, the difference in log-likelihood between the solution obtained by each of the ten runs of EM (with a spectral projection and starting with $k \log k$ clusters), and the log-likelihood of the EM run starting from the true centers. For a small sample size, the true peak is not tall enough, and EM easily finds many false peaks which are better. As the sample size increases, the true peak becomes taller then the "random" peaks the EM finds. When enough samples are available, the true peak is pronounced enough that EM successfully finds it for any random initialization.

# References


[AM05] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. `http://research.microsoft.com/~optas`, 2005.

[Das99] Sanjoy Dasgupta. Learning mixtures of gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, 1999.

[DS00] Sanjoy Dasgupta and Leonard Schulman. A two-round variant of em for gaussian mixtures. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence*, 2000.

[RW84] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

[SK01] Arora Sanjeev and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, 2001.

[SKV04] Hadi Salmasian, Ravindran Kannan, and Santosh Vempala. The spectral method for mixture models. *Electronic Colloquium on Computational Complexity (ECCC)*, (067), 2004.

[VW04] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci.*, 68(4):841–860, 2004.


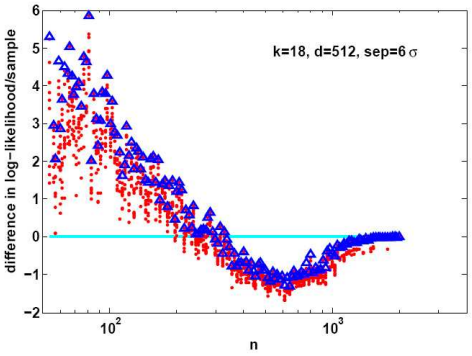


Figure 2: Log-likelihood of the model EM (with a spectral projection, and pruning components) converged to from random initialization minus log-likelihood of the model EM converged to from the true centers, divided by the sample size. Each point represents one random initialization, and the triangles are the best of ten runs for a single data set.