

The Spectral Method for general Mixture Models

R. Kannan, S. Vempala, H. Salmasian

Unknown probability density over \mathbf{R}^n :

$$F = w_1 F_1 + w_2 F_2 + \dots + w_k F_k,$$

$$w_i \geq 0, \sum w_i = 1$$

F_i are from a known class - for example general Gaussians (arbitrary variance-covariance matrix)
Or More generally, log-concave densities.

Given polynomially many (in n, k) samples drawn according to mixture F , break the samples up into k sub-sets each drawn according to one of the F_i .

$$n \rightarrow \infty \quad k \in O(1).$$

[“Weaker” Question : Find best fit from class.
Not considered here.]

Some notation : μ_i centroid of F_i .

σ_i is the maximum variance of F_i in any direction.

$\sigma_i^2 =$ maximum eigenvalue of var-covar matrix.

Separation :

$$\text{Min}_{i,j} \frac{|\mu_i - \mu_j|}{(\sigma_i + \sigma_j)}.$$

How many standard deviations apart are the means ?

Better Question : By how many standard deviations in the direction joining them are the means separated ? - **Cannot handle at present.**

Observation 1 If Separation is small, cannot hope to break up the given samples into the component k parts in a unique manner. So must assume separation not too small.

Observation 2 For spherical Gaussians of the same radius, at separation of $o(1)$, the probability clouds are still “merged”; cannot separate into components uniquely.

Observation 3 For $k = 2$ and two spherical Gaussians of same radius, at $\Omega(1)$ separation, the **central** hyper-plane separates the probability clouds; so if we find the direction of the line joining the two μ_i , we are done.

Result 3 Same case. For separation in $\Omega(n^{1/4})$, “**samples nearest to a sample come from the same F_i** ”. So, distance based separation works.

Basic Techniques

- Distance based clustering - Idea - classify close-by points (or points close to one center) into one cluster. **Fails if distances blur.**
- Random Projection. Project to a random sub-space of low (perhaps k) dimensional sub-space. Cluster the projection.

The centers and hence the clouds may come closer in the projection.

- Project to the k -dimensional sub-space which minimizes the sum of distances squared to the sample points. (SVD sub-space) **CAN BE FOUND IN P-TIME !!** For spherical Gaussians, the SVD sub-space (of the densities) passes exactly through the centers.

Problem cases : Two large flat pancakes
on top of each other.

Previous Results

| Authors | Sep | Assumptions |
|------------------|--------------------------------------|--|
| Dasgupta [?] | $n^{\frac{1}{2}}$ | Gaussians, bounded variance and $w_i = \Omega(1/k)$ |
| D-Schulman [?] | $n^{\frac{1}{4}}$ | Spherical Gaussians |
| Arora-Kannan [?] | $n^{\frac{1}{4}}$ | Gaussians |
| Vempala-Wang [?] | $k^{\frac{1}{4}}$ | Spherical Gaussians |
| This paper | $\frac{k^{\frac{3}{2}}}{\epsilon^2}$ | Logconcave distributions |

To Come : Achlioptas and McSherry

“For spherical Gaussians, the SVD sub-space W passes thro’ the centers” - Simple to prove conceptually :

Average distance squared of one Gaussian from a sub-space M of dimension k ([Pythagorous Thm](#))

= Distance squared of μ from M + $(n - k)\sigma^2$.

For general Gaussians, not necessarily exactly true, but approximately.

$$\sum_i w_i \text{dist}^2(W, \mu_i) \leq k \sum w_i \sigma_i^2.$$

Not sufficient.

Theorem Let $S = S_1 \cup S_2 \dots \cup S_k$ be a sample from a mixture F with k components such that S_i is from the i th component F_i and let W be the SVD subspace of S . For each i , let μ_i^S be the mean of S_i and $\hat{\sigma}_{i,W}^2(S)$ be the maximum variance of S_i along any direction in W . Then,

$$\sum_{i=1}^k |S_i| d(\mu_i^S, W)^2 \leq k \sum_{i=1}^k |S_i| \hat{\sigma}_{i,W}^2(S).$$

Main Difference from intuitive statement : (i) Both means and variances are of the sample. (ii) RHS has variances only in W .

Idea of Proof :

Let M be the space spanned by the real means
- μ_i .

Upper bound $a =$ sum of distances squared to
 M .

Let $b =$ sum of distances squared to W . Note
 $b \leq a$.

So, in a weighted average sense, the SVD sub-space is not far away from the means. Thus, for “large” F_i , their μ_i must be close to the sub-space. In fact, we show that no other (large or small) F_j has its “cloud” “contaminating” the cloud of a large F_i in the projection onto SVD sub-space W .

Gaussian F_i large if $|S_i|\hat{\sigma}_i^2 \geq \frac{\epsilon^3}{100k} \max_j |S_j|\hat{\sigma}_j^2$.

For i large and any other j ,

$$|\hat{\mu}_i - \hat{\mu}_j| > \frac{k^{3/2}}{\epsilon^2}(\sigma_i + \sigma_j).$$

Find a large component - peel it off and repeat.

How does one find a large component ?

Assume each $w_i \geq \epsilon$. For each projected sample point p , let $T(p)$ be its nearest $\epsilon N/4$ neighbors. (N total number of samples) Then, we show that for large i , for every sample p picked according to F_i , all samples in $T(p)$ must be from the same component.

Let T be a (large) set of i.i.d. samples picked according to a Gaussian and T' be an arbitrary subset of T with $|T'| = |T|/10$ (T' is picked from T by an adversary.) Then the maximum variance of T' in any direction is within a constant factor of the maximum variance of T in any direction.

So, far we have solved the clustering problem. How can we actually find the densities (to complete “learning the mixture”) ?

Log-concave densities generalize uniform distributions on convex sets and learning general convex sets is not yet solved even under uniform densities. But, at least :

Lovász, Vempala : Given $O^*(n \ln n / \epsilon^4)$ i.i.d. samples from a log-concave density in \mathbf{R}^n , the variance-covariance matrix of the density can be estimated to **relative error** ϵ .

(Estimate matrix) $^{-1} \times$ (var-covar Matrix) has all eigenvalues $\in (1 - \epsilon, 1 + \epsilon)$.