Mixture of Gaussians :-

Last time, we saw a method for labelling points drawn from a mixture of Gaussians with which component of the mixture they were drawn from. The method was based on the observation that when if $X \sim N(\mu_1, I_d)$ and $Y \sim N(\mu_2, I_d)$, are independent then w.h.p.

$$\|X-Y\|^2 \approx (\mu_1-\mu_2)^2 + 2d \pm O(\sqrt{d})$$

This required $\|\mu_1-\mu_1\| \geq O(\sqrt{d^{1/4}\log n})$ for n samples from the mixture.

We however also want to be handle the case where the separation does not grow with the dimension.

A possible idea: Project to a smaller dimension.

— Random projections 'preserve distances' upto scaling.

— But the scaling factor is the same for the inter-centre distances and the variances of the component Gaussians

Suppose we could project to the dimensions spanned by the μ_i:

— The inter-center distances do not change.

— The variances get scaled down by a factor of $\frac{k}{d}$ [k is the no. of components] !

We now make this precise. Let $X \sim \mathcal{D}$ where

$\mathcal{D}$ is a mixture of $\mathcal{D}_i = N(\mu_i, I_d)$ with weights w_i, $i \in [k]$.

Let V be any subspace of dimension $r \leq d$. Then we compute

$$\mathbb{E}\left[\|\text{proj}_V(X)\|_2^2\right] = \sum_{i=1}^{k} w_i \, \mathbb{E}\left[\|\text{proj}_V(Y_i)\|_2^2\right]$$

where $Y_i \sim N(\mu_i, I_d)$

$$= r + \sum_{i=1}^{k} w_i \|\text{proj}_V(\mu_i)\|_2^2 .$$

Observation:- k-dimensional vector spaces V maximizing $\mathbb{E}[\|\text{proj}_V(X)\|_2^2]$ are precisely those containing $\text{span}(\mu_1, \ldots, \mu_k)$.

[There is a unique such space if the μ_i are linearly independent.]

Suppose now that we have n independent samples $X_1, \ldots X_n$ from $\mathcal{D}$ which we write as a $d \times n$ matrix A.

Given a space V, we denote by $\text{proj}_V(A)$ the matrix obtained by projecting each column of A to V.

We also have

$$\mathbb{E}\left[\|\text{proj}_V(A)\|_F^2\right] = nr + \sum_{i=1}^{k} n\omega_i \|\text{proj}_V(\mu_i)\|_2^2$$

where $\|A\|_F^2 = \sum_{i=1}^{n} \|A_{\cdot i}\|_2^2 = \sum_{i=1}^{n}\sum_{j=1}^{d} |A_{ji}|^2$

From the observation we know that the k-dimensional V maximizing $\mathbb{E}[\|\text{proj}_V(A)\|_F^2]$ are precisely those which contain span$(\mu_1, \ldots \mu_k)$.

If we had the above observation for $\|\text{proj}_V(A)\|_F^2$ (as opposed

to $E[\|\text{proj}_V(A)\|_F^2]$) this would imply that one only needs to take V to be the subspace of the first k (left) singular vectors of A.

— However, one might still expect that with enough samples the first few singular vectors of A might still be sufficient. This is what Vempala and Wang show

Theorem (Vempala and Wang '04)
Let $A \in \mathbb{R}^{d \times n}$ consist of n independent samples from $\mathcal{D}$. Let $\varepsilon = \max\{k, C_1 \log \frac{4n}{\delta}\}$

and suppose that

$$n \geq \frac{C_2}{\varepsilon^2 \omega_{min}} \left(d \log \frac{d \max \|\mu_i\|^2}{\varepsilon} + \frac{\log(k/\delta)}{d-r} \right)$$

Let V be the space spanned by top r left sing. vectors of A, and let U be any r dimensional space containing $\text{span}(\mu_1, \ldots \mu_k)$.

Then with prob. at least $1-\delta$,

$$\sum_{i=1}^{k} \omega_i \left(\|\text{proj}_U(\mu_i)\|^2 - \|\text{proj}_V(\mu_i)\|^2 \right)$$

$$= \sum_{i=1}^{k} \omega_i \left(\|\mu_i - \text{proj}_V(\mu_i)\|_2^2 \right)$$

$$\leq \varepsilon (n-r).$$

$\times$

In particular, when applied with $\varepsilon \approx \frac{\hat{\varepsilon}}{(n-r)\omega_{min}}$, the above guarantee implies $\|\mu_i - \text{proj}_V(\mu_i)\|_2 \leq \varepsilon$ for all i.

— Thus projecting to the span of

the top r left singular vectors of A does not decrease the distances between the means by much.

However, one cannot claim that after the projection to V, the columns of A can be treated as iid samples from $N(\text{proj}_V(\mu_i), I_r)$

[Why ?]

A fix :- Use only a fraction, say 1/10, of the columns of A to compute V.

Now the random subspace V remains independent of the rest of the columns. In particular, conditioning on V, it remains the case that the

projections of the remaining columns of A onto V are still independent Gaussians with correlation matrix I_n.