Supose we have a dataset with n observations $\mathbf{X} = \{\mathbf{x}_1, ...\mathbf{x}_n\}$ and a (log) likelihood function:

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) \tag{1}$$

I can get the Maximum Likehood Estimators of the parameter, which I call $\hat{\boldsymbol{\theta}}_0$. I denote the achieved likelihood $\mathcal{L}_0$.

Now I want to fit the same model using mixtures, assuming that there are K groups of points with parameters $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K$. For that, we introduce the a matrix $\mathbf{Z}$ of latent variables where $z_{nk} = 1$ if point n belongs to cluster k.

The likelihood can the be re-expressed as:

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Because optimizing with that sum is usually hard, we do a couple of tricks:

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} = \ln \mathbb{E}_q \left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right)$$

which is the logarithm of a expected value. By Jennsenn inequality, we know that:

$$\ln \mathbb{E}[x] \geq \mathbb{E}[\ln(x)]$$

and then:

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \geq \overbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}}^{\mathcal{L}}$$

The equality holds if the function inside the logarithm is constant. And this happens when $q(\mathbf{Z})$ is the posterior of $\mathbf{Z}$. Therefore we have the general EM expression of the lower bound:

$$\mathcal{L} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \ln \overbrace{\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}}^{p(\mathbf{X}|\boldsymbol{\theta})} = \overbrace{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}^{Q} - \overbrace{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}^{H(Z) \text{ (entropy)}}$$

k=1

If there is only one cluster, $p(\mathbf{Z}|\cdot)$ can be thought of as a constant, and

$$\mathcal{L} = \ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}_0$$

and if the plug the estilmators $\hat{\boldsymbol{\theta}}_0$ we obtain:

$$\ln p(\mathbf{X}|\hat{\boldsymbol{\theta}}_0) = \mathcal{L}_0$$

K overlapped clusters

The likelihood of K clusters completely overlapped, all sharing the same parameters $\hat{\boldsymbol{\theta}}_0$ we have that $p(\mathbf{X}|\boldsymbol{\theta})$ does not depend on $\mathbf{Z}$, therefore

$$\mathcal{L} = \ln p(\mathbf{X}|\boldsymbol{\theta}) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) = \ln p(\mathbf{X}|\boldsymbol{\theta})$$

Let us demonstrate it again using Q and $H(\mathbf{Z})$.

$$\mathcal{L} = \overbrace{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}^{Q} - \overbrace{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}^{H(Z)\ \text{(entropy)}}$$

Since the $\boldsymbol{\theta}$ are equal for every cluster, then:

$$\begin{aligned}\mathcal{L} &= \overbrace{\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \underbrace{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}_{1}}^{Q} - \overbrace{\ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \underbrace{\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}_{1}}^{H(Z)\ \text{(entropy)}} \\ &= \overbrace{\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}^{Q} - \overbrace{\ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}^{H(Z)\ \text{(entropy)}} = \ln p(\mathbf{X}|\boldsymbol{\theta})\end{aligned}$$