CO-495 Coursework 1

Ilaria Manco

February 2, 2018

Exercise 1.

a) From left to right, the form of the covariance matrix is:

- isotropic, because the points are spherically distributed, meaning the covariance is the same for all the random variables.
- anisotropic, because the points appear distributed within an ellipse with its major axis stretched. Furthermore, this covariance is not a diagonal matrix since the ellipse's axes are inclined with respect to the x- and y-axis.
- anisotropic, for the same reason as above, but with the major axis more inclined (approximately 45°) with respect to the horizontal axis.
- isotropic, as in the first case, but with a wider spread. The eigenvalues of this covariance matrix are therefore bigger than in the first case.

b) Since there are multiple local maxima of the log likelihood function, the EM algorithm does not necessarily find the largest of these (i.e. the global maximum), when we initialise the parameters randomly, since we might be too far from the global maximum. Hence the algorithm incorrectly estimates the parameters, as shown in Fig. 1a.
In order to avoid this, we can initialise the parameters by using the k-means algorithm. The means $\boldsymbol{\mu_k}$ and the covariance matrices $\boldsymbol{\Sigma_k}$ can then be initialised to the centres and the sample covariances of the clusters respectively, while the mixing coefficients π_k can be set to the proportion of data points within the respective cluster k. This leads to an improved outcome, as shown in 1b

Exercise 2. In this case the parameters must satisfy an additional constraint:

$$\boldsymbol{\Sigma_k} = \boldsymbol{\Sigma} \qquad \forall k.$$

Therefore, the maximisation of the expectation value of the log-likelihood function

$$G(\theta) = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{-\frac{1}{2}(\mathbf{x_n}-\boldsymbol{\mu_k})^T\boldsymbol{\Sigma_k}^{-1}(\mathbf{x_n}-\boldsymbol{\mu_k}) - \frac{1}{2}(F\ln 2\pi + \ln|\boldsymbol{\Sigma_k}|) + \ln\pi_k\right\}$$

with respect to the covariance matrix becomes

$$\frac{dG(\theta)}{d\boldsymbol{\Sigma}} = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{(\mathbf{x_n}-\boldsymbol{\mu_k})(\mathbf{x_n}-\boldsymbol{\mu_k})^T - \boldsymbol{\Sigma}\right\} = \mathbf{0}$$

$$\implies \boldsymbol{\Sigma} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})(\mathbf{x_n}-\boldsymbol{\mu_k})(\mathbf{x_n}-\boldsymbol{\mu_k})^T}{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})}.$$

Noting that

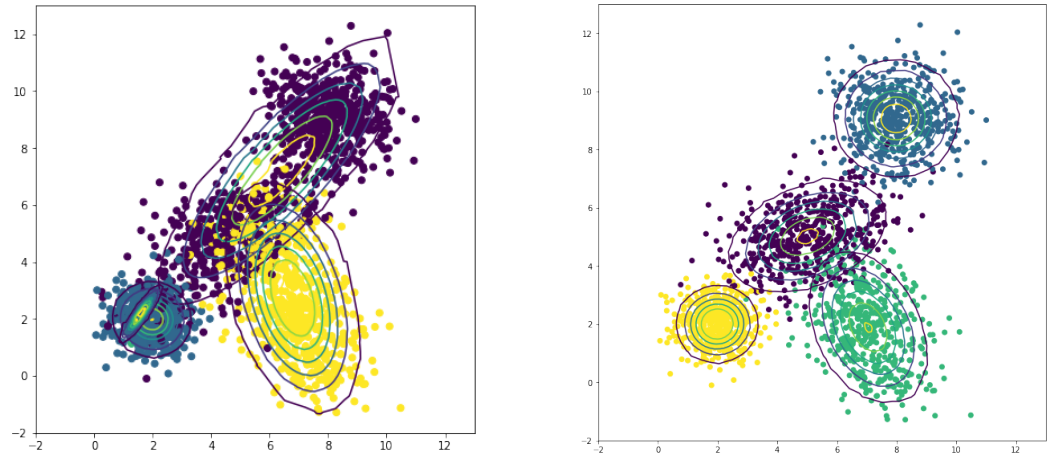
$$\sum_{k=1}^{K}\gamma(z_{nk}) = 1,$$

the final equation for the covariance matrix takes the form

$$\boldsymbol{\Sigma} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})(\mathbf{x_n}-\boldsymbol{\mu_k})(\mathbf{x_n}-\boldsymbol{\mu_k})^T}{N}.$$

This is different to the equation for the case with the unconstrained covariance matrices, since now we need to keep the sum over k. Previously, this would "collapse" to one term, because, for any $\tilde{k}$, the derivative $\frac{dG(\theta)}{d\boldsymbol{\Sigma}_k}$ would go to zero for $k \neq \tilde{k}$.

The maximisation with respect to the other two parameters $\boldsymbol{\mu_k}$ and π_k is unchanged.



(a) Random initialisation (b) k-means initialisation

Figure 1: Samples from the joint distribution in which the four components of the mixtures are depicted in yellow, purple, green and blue. Each data point $\boldsymbol{x}_n$ is assigned to a cluster with a probability equal to the corresponding responsibility $\gamma(z_{nk})$.