Computational Statistics II
Assignment 3: a simple EM

Alessia Pini

Aim of this assignment is to write in a simple case, the code for performing an EM algorithm on a mixture of bi-variate Gaussians with $K = 3$ components of the mixture, and to apply it for performing a simple clustering.

1. Simulate a data set of sample size $n = 100$ from a mixture of gaussians with 3 components in the two following scenarios:

   A $\mu_1 = (0, 0)'$, $\mu_2 = (4, 3)'$, $\mu_3 = (3, -1)'$,
   $\pi_1 = \pi_2 = \pi_3$,
   $\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

   B $\mu_1 = (0, 0)'$, $\mu_2 = (4, 3)'$, $\mu_3 = (3, -1)'$,
   $\pi_1 = \pi_2 = \pi_3$,
   $\Sigma_1 = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1.5 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 0.45 \end{pmatrix}$, $\Sigma_3 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 0.9 \end{pmatrix}$

2. In both scenarios, initialize the parameters of the EM using $\pi_1 = \pi_2 = \pi_3 = 1/3$, $\Sigma_k = I$, and choose some random values for the three means. Perform one E-step and one M-step. Visualize the results of this first iteration. Note that for applying the EM, you can either write down your own code, or apply one of the existing R packages.

3. Appky the EM to the two data sets (scenarios A and B), iterating through E-step and M-step. Use the convergence of the log likelihood as stopping criterion.

4. Use a threshold on the estimated responsibilities to assign with EM each data point to a cluster. Compare the results with what was obteined from the k-means in terms of rate of misclassified points in both scenarios.

5. **Bonus.** Use a MC simulation based on 100 runs to estimate the rate of misclassified points in the two scenarios.