Homework 5 – Part II: 25 points

Exercise J-4.2: Suppose that $y_1, y_2, \cdots, y_n$ is a set of data from a three-component mixture density

$$\alpha f_1(y; \boldsymbol{\theta_1}) + \beta f_2(y; \boldsymbol{\theta_2}) + (1 - \alpha - \beta) f_3(y; \boldsymbol{\theta_3}),$$

where $f_i(y; \boldsymbol{\theta_i}) = (2\pi\sigma^2)^{-1/2} e^{-(y-\mu_i)^2/2\sigma^2}$ is the $N(\mu_i, \sigma^2)$ density. Here $\boldsymbol{\theta_i} = (\mu_i, \sigma^2)$ for $i = 1,2,3$.

a) [3 points] Plot a density histogram of the variable y in the dataset ExJ42, and superimpose it by a Kernel smoother.

b) [5 points] Derive the EM algorithm for estimating the mixture parameters α and β and the parameters $\mu_1, \mu_2, \mu_3,$ and σ^2. Write your pseudo algorithm.

c) [10 points] Write an R function to code the EM algorithm in part (b) with the following input variables:

i. Data: $\boldsymbol{y} = (y_1, \cdots, y_n)$, an $n \times 1$ vector of data
ii. Parameter initial values: $\boldsymbol{\theta} = (\alpha, \beta, \mu_1, \mu_2, \mu_3, \sigma^2)$, a 6 by 1 vector
iii. Maximum number of iterations: An upper-bound for the maximum number of iterations
iv. Tolerance: for the maximum relative error (MRE) stopping criterion as a stopping rule.

Your function should output the iteration process including iteration number, value of the log-likelihood at each iteration, and the MRE. At the end the MLE estimates for α and $\beta, \mu_1, \mu_2, \mu_3$, and σ^2should be printed. Then, use your function to fit the three-component mixture model to the data given in the dataset ExJ42.

d) [3 points] Again graph a density histogram of the data and superimpose the histogram by the three-component normal mixture density with parameters obtained as in part (c).

e) [4 points] Compute the posterior probability that each case belongs to group 1, 2, and 3, and classify each case to one of the three groups based on the highest posterior probability. Then show your result graphically as follows: Draw a graph, with the x-axis indicating case number, and the y-axis including the group numbers 1, 2, 3 that the cases were assigned to. Then use red, green, and blue dots to indicate cases that are assigned to Groups 1, 2, and 3 respectively on the graph. Do you see a pattern, and if so, what does the pattern indicate about the classification of the cases relative to the case numbers?