Assignment 3

STAT 440/840 - CM 761

Due Tuesday July 16 at 9am - to be submitted through crowdmark

---

A G component finite mixutre of multivariate-normal is given by

$$g\left(\mathbf{x} \mid \boldsymbol{\theta}\right) = \sum_{g=1}^{G} \pi_g \, \phi_p\left(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right).$$

Note, the parameters are

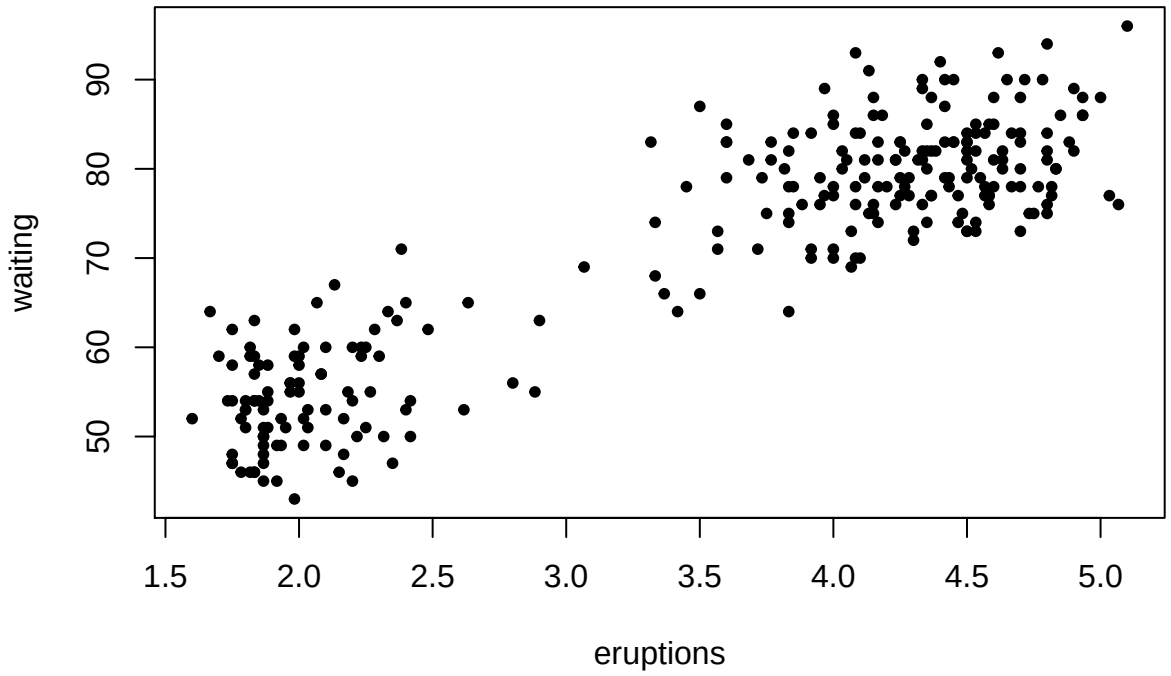
$$\boldsymbol{\theta} = \left(\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1 \ldots, \boldsymbol{\Sigma}_G\right).$$

1. **[8 Marks]** Properties of the model
   a) (1 Mark) To apply the EM we take the component membership for each observation as missing data denoted by Z_{ig}. What is the marginal distribution of the missing data?
   b) (1 Mark) What is the conditional distribution of the observed data given the missing data?
   c) (2 Marks) What is the distribution of the missing data given the observed data.
   d) (1 Mark) What is the expected value of the missing data given the observed data.
   e) (3 Marks) Give an algorithm and a `R` function to generate data from a G component finite mixture of a multivariate-normals.

2. **[20 Marks]** An EM algorithm,
   a) (1 Mark) Give the observed log-likelihood function.
   b) (2 Marks) Write a `R` function which takes the parameters and data as an input and ouput the observed log-likelihood.
   c) (1 Marks) To begin the derivation of the EM algorithm, give the complete data log-likelihood.
   d) (4 Marks) E-step: Derive the expected complete data log-likelihood, denoted by $\mathcal{Q}$.
   e) (4 Marks) Write a `R` function which takes the parameters and data as an input and ouput the expected value of the missing data given the observed data.
   f) (4 Marks) M-step: Find the parameter updates by maximizing the expected complete data log-likelihood.
   g) (4 Marks) Write a `R` function which takes the expected value of the missing data and returns the updated parameter values.

3. **[24 Marks]** Implementing a EM algorithm, for the the old faithful dataset in `R`.

```
data(faithful)
plot(faithful, pch=20)
```



a) (12 Marks) Implement a EM algorithm to fit a two component ($G = 2$) multivariate-normal finite mixture to the old faithful dataset in `R`. Use the following starting value

$$\pi_1 = \frac{1}{10}, \quad \pi_2 = \frac{9}{10},$$

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 2 \\ 60 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 2 \\ 50 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}, \qquad \text{and} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix}.$$

As part of the results give

- `R` code for the EM,
- the MLE,
- plot the observed log-likelihood from each iteration,
- a contour plot of the fitted density along with the points,
- comment on the fit.

b) (2 Marks) Write a `R` function to generate random parameters values to use as starting values.

c) (6 Marks) Start the EM from a 100 different starting values. As part of the results

- report the solution with the highest log-likelihood,
- give contour plot of the fitted density along with the points and
- comment why this is different than the solution in 3a).

d) (4 Marks) When fitting finite mixture models it often of interest to know the predicted component memembership are given by the a posteriori probabilities (expected values). These are typically done with we using maximum *a posteriori* probabilities (MAP) given by

$$\text{MAP}(\widehat{z}_{ig}) = \begin{cases} 1 & \text{if } \max_h\{\widehat{z}_{ih}\} \text{ occurs in component } g, \\ 0 & \text{otherwise.} \end{cases}$$

Provide `R` code to calculate the MAP estimates. Then plot the data and colour the different MAP estimates.

4. **[12 Marks]** EM extensions,

a) (6 marks) Implement the Incremental EM with $m = 20$. Using the starting value given in 3a) perform 100 iterations and plot the observed log-likelihood from each iteration. Comment on the result.

b) (6 marks) Implement the Stochastic EM. Using the starting value given in 3a) perform 100 iterations and plot the observed log-likelihood from each iteration. Comment on the result.