Coding Assignment 4

Due Monday, April 05

Implement the EM algorithm for a p-dimensional Gaussian mixture model with G components:

$$\sum_{k=1}^{G} p_k \cdot \mathsf{N}(x; \mu_k, \Sigma).$$

Store the estimated parameters as a list in `R` with three components

- `prob`: G-dimensional probability vector $(p_1, \ldots, p_G)$
- `mean`: p-by-G matrix with the k-th column being μ_k, the p-dimensional mean for the k-th Gaussian component;
- `Sigma`: p-by-p covariance matrix Σ shared by all G components.

Structure of your code should look like the following.

```
Estep <- function(data, G, para){
  # Return the n-by-G probability matrix
  }

Mstep <- function(data, G, para, post.prob){
  # Return the updated parameters
  }

myEM <- function(data, itmax, G, para){
  for(t in 1:itmax){
        post.prob <- Estep(data, G, para)
        para <- Mstep(data, G, para, post.prob)
      }
  return(para)
  }
```

Test your code on the `faithful` data from `R` package `mclust` with $G = 2$ and $G = 3$. The estimated parameters from your algorithm and the ones from `mclust` after 20 iterations should be the same.

Implement all the computation by your own code; do not use any libraries except loading the test data from `mclust`.

What you need to submit?

An R Markdown file in HTML format, which should contain all code used to produce your results.

Name your file starting with `Assignment_4_xxxx_netID` where "`xxxx`" is the last 4-dig of your University ID.

In addition to necessary R/Python code and output, your submission should include

- expression of the marginal (or the incomplete) likelihood function $\prod_{i=1}^{n} p(x_i \mid p_{1:G}, \mu_{1:G}, \Sigma)$ or its log, which is the objective function we aim to maximize;
- expression of the complete likelihood function $\prod_{i=1}^{n} p(x_i, Z_i \mid p_{1:G}, \mu_{1:G}, \Sigma)$ or its log, which is the function we work with in the EM algorithm;
- expression of the distribution of Z_i's at the E-step;
- expression of the objective function you aim to maximize (or minimize) at the M-step;
- derivation and the updating formulae for $p_{1:G}$, $\mu_{1:G}$, and Σ at the M-step.

You can write your derivation for $G = 2$ if needed.

I like to use upper case letters for latent variables, such as Z_i, but it's just a personal preference; feel free to use lower case letters.

If you do not know how to include math formulae in R Markdown, you can write your derivation on a piece of paper, take a photo, and then insert it into your HTML file or save it as a separate PDF file.

Name your second file, if applicable, starting with `Assignment_4_Supp_xxxx_netID` where "`xxxx`" is the last 4-dig of your University ID.