STAT 740 - Spring 2004 - Homework 6

Due: Friday, April 9th

Consider a random sample $\underline{y} = y_1, \ldots. y_m$ where each observation has probability α of being drawn from a $N(\mu_1,\sigma^2)$ distribution and probability $(1-\alpha)$ of being drawn from a $N(\mu_2,\sigma^2)$ distribution. The goal is to use the EM algorithm to estimate the four parameters $\theta = (\mu_1, \mu_2, \sigma^2, \alpha)$.

This can be done by using the unobserved data $\underline{x} = x_1, \ldots. x_m$ where x_i is one if y_i was chosen from the first normal distribution and 0 otherwise.

1) Find the pdf $f(\underline{y}, \underline{x}| \mu_1, \mu_2, \sigma^2, \alpha)$ of the complete data, and its logarithm.

2) Briefly explain why finding the expectation in $Q(\theta|\theta_n)$ can be reduced to finding the $E(x_i| \mu_{1n}, \mu_{2n}, \sigma^2_n, \alpha_n, y_i)$, and why this is identical to finding the $p_{in}=P(x_i=1| \mu_{1n}, \mu_{2n}, \sigma^2_n, \alpha_n, y_i)$. Find the p_{in} in terms of $\mu_{1n}, \mu_{2n}, \sigma^2_n, \alpha_n$.

3) Find $Q(\theta|\theta_n)$ as a function of $\mu_1, \mu_2, \sigma^2, \alpha$, and the p_{in} .

4) Find $\mu_{1(n+1)}, \mu_{2(n+1)}, \sigma^2_{n+1}, \alpha_{n+1}$.

5) Write the R-code to perform this algorithm (you can write it to run the given number of iterations and don't have to have it check for convergence). The function `dnorm` could be useful.

6) Use an initial seed of 100 to generate a sample `y<-c(rnorm(200,8,4),rnorm(100,-2,4))`. Find the parameter estimates.