STA 4102/5106: Homework Assignment #5

(Wednesday, September 24)
Due: Wednesday, October 1

1. Derive an EM algorithm to find the maximum likelihood estimate of θ where θ is a parameter in the multinomial distribution:

$$(x_1, x_2, x_3, x_4) \sim M(n; 0.25\theta, 0.25(2+\theta), 0.5(1-2\theta), 0.5\theta)$$

Similar to the case covered in the class, choose a variable for the missing data and derive the EM algorithm for iteratively estimating θ. Implement this algorithm in Matlab and test it on the dataset $(x_1, x_2, x_3, x_4) = (6,52,28,14)$ and $n = 100$.
(hint: partition $0.25(2 + \theta)$ to 0.5 and 0.25θ).

2. (STA5106 Students Only) Let Y be a continuous random variable with probability density function:

$$Y \sim \alpha_1 f_1(y; \mu_1, \sigma_1^2) + \alpha_2 f_2(y; \mu_2, \sigma_2^2),$$

where f_1 and f_2 are two Gaussian density functions with means μ_1, μ_2 and variances σ_1^2, σ_2^2, respectively. Also, $0 \le \alpha_1, \alpha_2 \le 1$, such that $\alpha_1 + \alpha_2 = 1$. Given n observations of Y, our goal is to find the maximum likelihood estimate of

$$\theta = (\alpha_1, \mu_1, \sigma_1^2, \alpha_2, \mu_2, \sigma_2^2)$$

We will use the EM algorithm for this estimation. Let $\theta^{(m)}$ be the current values of the unknown. Then, the update for $\theta^{(m+1)}$ is given by:

$$\alpha_l^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} P(l \mid \theta^{(m)}, Y_i),$$

$$\mu_l^{(m+1)} = \frac{\sum_{i=1}^{n} Y_i P(l \mid \theta^{(m)}, Y_i)}{\sum_{i=1}^{n} P(l \mid \theta^{(m)}, Y_i)},$$

$$\sigma_l^{(m+1)} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \mu_l^{(m+1)})^2 P(l \mid \theta^{(m)}, Y_i)}{\sum_{i=1}^{n} P(l \mid \theta^{(m)}, Y_i)}},$$

where

$$P(l \mid \theta^{(m)}, Y_i) = \frac{\alpha_l^{(m)} f_l(Y_i; \mu_l^{(m)}, (\sigma_l^{(m)})^2)}{\sum_{l=1}^{2} \alpha_l^{(m)} f_l(Y_i; \mu_l^{(m)}, (\sigma_l^{(m)})^2)}.$$

Download two datasets from the class website to apply to this problem. For each data:

(a) Plot a histogram of the data using the **hist** function in Matlab.
(b) Using some initial values guessed from the histogram, apply EM algorithm to estimate the unknown parameters.
(c) Plot the evolution of the observed data log-likelihood function versus the iteration index. At the m-th iteration, the observed data log-likelihood function is:

$$\sum_{i=1}^{n} \log[\alpha_1^{(m)} f_1(Y_i; \mu_1^{(m)}, (\sigma_1^{(m)})^2) + \alpha_2^{(m)} f_2(Y_i; \mu_2^{(m)}, (\sigma_2^{(m)})^2)]$$