STA 410/2102, Spring 2003 — Assignment #3

Due at **start** of class on April 1. Worth 18% of the final mark.

Note that this assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own.

In this assignment, you will solve two maximum likelihood estimation problems using the EM algorithm.

For each problem, you should hand in your derivation of the formulas to use for the E and M steps. You should also write an R function to implement the EM algorithm, and test it on the data provided, as well as other data as appropriate. Your program should run for a specified number of iterations (you don't need to detect convergence automatically), and it should have a "debug" option that prints the parameter values and the log likelihood at each iteration. (Note that the log likelihood should never decrease.) The other arguments to your EM function should be the data and the initial values for the parameters.

You should hand in your derivations of the algorithms, your R functions, the commands and output for your tests, and a discussion of how well the EM algorithm worked. In particular, you should discuss whether or not there appear to be multiple local maxima for the likelihood in these problems.

Problem 1:

Suppose that the values $Z_1, \ldots, Z_n$ are generated from a normal distribution with mean μ and variance one, where μ is an unknown parameter greater than zero. We don't observe the Z_i values, however, but only their absolute values, $X_1, \ldots, X_n$, with $X_i = |Z_i|$. Given these observed values, you should find the maximum likelihood estimate for μ using an EM algorithm that you derive, in which the unobserved data are the signs of $Z_1, \ldots, Z_n$.

You should test your function on the data available from the web page and in the file `/u/radford/data1` on the utstat and CQUEST systems.

Problem 2:

Suppose that an ornithologist has observations on when female birds return to their nests after expeditions to collect food for their chicks. For each female bird, the time when the chicks hatch is known, and all time measurements are relative to this time. It is believed that the birds return from food collection expeditions regularly, at times θ, 2θ, 3θ, etc. after hatching, with some random variation, assumed here to be normal with mean zero and variance one. The unknown parameter θ is what the ornithologist wishes to estimate. The birds are assumed to behave independently.

To avoid disturbing the birds at a critical time, the time of return from the first expedition was not observed. Instead, for each bird, the ornithologist has observed the time of return from *either* the second or the third expedition. (The birds are hard to spot, so some returns are missed.) The times of these returns constitute the data, $X_1, \ldots, X_n$. Unfortunately, the ornithologist does not know whether each X_i is the time of return for the second expedition or for the third. The probability of observing a second return versus a third return is not known.

The data can be modeled as a mixture of the $N(2\theta, 1)$ and $N(3\theta, 1)$ distributions, with mixing probabilities p and $1-p$. In other words, the density function for an observation is

$$f(x) \ = \ p\,(1/\sqrt{2\pi})\ \exp(-(x-2\theta)^2/2) \ + \ (1-p)\,(1/\sqrt{2\pi}))\ \exp(-(x-3\theta)^2/2)$$

Given the data $X_1, \ldots, X_n$, you should find the maximum likelihood estimates for θ and p, using an EM algorithm that you derive. The missing data should be the indicators of whether each observation was of the second or the third return.

You should test your function on the data available from the web page and in the file `/u/radford/data2` on the utstat and CQUEST systems.