STA 410/2102, Fall 2000 — Assignment #4

Due in class on December 8. Worth 16% of the final mark.

Suppose that the time from when a machine is manufactured to when it fails is is exponentially distributed (a common, though simplistic, assumption). However, suppose that some machines have a manufacturing defect that causes them to be more likely to fail early than machines that don't have the defect.

Let the probability that a machine has the defect be p, the mean time to failure for machines without the defect be μ_g, and the mean time to failure for machines with the defect be μ_d. The probability density for the time to failure will then be the following mixture density:

$$p\,(1/\mu_d)\exp(-x/\mu_d) \;+\; (1-p)\,(1/\mu_g)\exp(-x/\mu_g)$$

Suppose that you have a number of independent observations of times to failure for machines, and that you wish to find maximum likelihood estimates for p, μ_g, and μ_d. Write an S/R program to find these estimates using the EM algorithm, with the unobserved data being the indicators of whether or not each machine is defective. Note that the model is not identifiable — swapping μ_d and μ_g while replacing p with $1-p$ has no effect on the density. This isn't really a problem; you can just interpret whichever mean is smaller as the mean for the defective machines.

You may write your program so that it simply runs for however many iterations you specify (ie, you don't have to come up with a convergence test). However, your program should have the option of printing the parameter estimates and the log likelihood at each iteration, so that you can manually see whether it has converged. (This will also help debugging.)

You should test your program on two data sets, each with 1000 observations, which are in `/u/radford/ass4a.data` and `/u/radford/ass4b.data` on utstat and CQUEST, and which are also available from the course web page. You can read this data with a command like

```
> x <- scan("/u/radford/ass4a.data")
```

For both data sets, run your algorithm for as long as you need to to be sure that you have obtained close to the correct maximum likelihood estimates. To be sure, I recommend that you run it for hundreds of iterations or more (this shouldn't take long in R). Discuss how rapidly the algorithm converges on the two data sets.

You may find it useful to generate your own data sets, for which you know the true parameters values, in order to debug your program. You could start with data sets where μ_g and μ_d are very different.

You should hand in your derivation of the formulas neeed for the EM algorithm, your program, the output of your tests, and your discussion of the results.