STA 450/2102 — Assignment #4

Due at the beginning of class on April 7. Worth 18% of the final mark.

Outliers can have a big effect on the regression coefficients found using least squares. One way of handling this is to try to identify outliers that are not informative regarding the relationship of interest, and remove them from the dataset. It's hard to be sure whether an observation is an outlier, however. Another approach is to regard whether an observation is an outlier or not to be an additional piece of data, which unfortunately has not been observed. One can then use EM to find maximum likelihood estimates for the regression coefficients on the basis of the observed data, with these outlier indications being regarded as "missing".

Here is a particular model of this sort. We have n independent observations, indexed by i, for each of which there are two predictor variables, x_{1i} and x_{2i}, and a single response variable, y_i. There is also an outlier indicator variable, z_i, that is 1 if the observation is an outlier, and 0 if it is not. We model this complete data in terms of parameters p, β_0, β_1, β_2, σ_r^2, μ_o, and σ_o^2, as follows:

$$
\begin{aligned}
P(z_i = 1) &= p \\
y_i &= \begin{cases} \beta_0 \ + \ \beta_1 x_{1i} \ + \ \beta_2 x_{2i} \ + \ \epsilon_i & \text{if } z_i = 0 \\ \eta_i & \text{if } z_i = 1 \end{cases}
\end{aligned}
$$

where $\epsilon_i \sim N(0, \sigma_r^2)$ and $\eta_i \sim N(\mu_o, \sigma_o^2)$. In other words, for non-outliers, this is a standard linear regression model with normally-distributed errors, but outliers come from a completely different distribution, that is not related to the predictor variables.

This model might be appropriate if the response variable was measured by some instrument that occasionally goes wrong, delivering a random reading that has nothing to do with the actual value.

You are to write an S/R function to find maximum likelihood estimates for the parameters of this model based on the values in each case of x_{1i}, x_{2i}, and y_i (but not z_i). You should do this using the EM algorithm, starting from some sensible initial parameter values. You will find that doing the M step requires finding weighted least-squares estimates for the regression coefficients — ie, estimates that minimize a weighted sum of the squared residuals, with a possibly different weighting factor for each observation. You can find such estimates with the **lm** function by giving a "weights" argument.

You should test your function on data that you can read with the command:

```
d <- read.table("/u/radford/ass4-data",header=T)
```

This will set `d` to a "data frame", in which `d$x1`, `d$x2`, and `d$y` are the data for 100 observations. You can also get the data from the course web page, at

```
http://www.utstat.toronto.edu/~radford/sta2102/
```

You should hand in your derivation of the correct way to apply EM to this problem, your program implementing EM, and the results of applying your program to the above data. You should also discuss whether there might be more than one local maximum for this problem, but you don't have to exhaustively search for extra maxima, once you have found what you think is a reasonable solution.