STA 410/2102, Fall 2015 — Assignment #2

Due at the start of class on November 19. Please hand it in on 8 1/2 by 11 inch paper, stapled in the upper left, with no other packaging.

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 particular, you should not leave any discussion with someone else with any written notes (either paper or electronic).

In this assignment, you will implement and test EM algorithms for finding maximum likelihood estimates for two problems in which data is modelled as coming from a mixture, with which mixture component a data item comes from sometimes being observed (or partly observed) and sometimes not. The "missing data" handled using EM will be the indicators of which mixture component a data item came from, when this is not known from observation.

The first problem is the same as for Assignment #1, with the same data as for that assignment. The second problem concerns synthetic data on beetles of different species and genera. For the second problem, you should also produce a suitable plot (or plots) that displays the data and the parameter estimates in a way that would be useful to someone interested in the problem.

Problem 1 [30 marks]: For this problem, you will use the same model and data as for Assignment #1, but will find the maximum likelihood estimate for the parameters p_1 and p_2 using the EM algorithm, rather than the methods used in Assignment #1. You should write your program using EM to work for any data, but hand in the results for the same data as before, with $n = 130$, $m_1 = 25$, $m_2 = 25$, $x = 75$, $x_1 = 20$, $x_2 = 6$.

The missing data in this problem is the gender of the x people who answered "yes" on the first survey, and the gender of the $n - x$ people who answered "no" on this survey. In the E step of the algorithm, you need to find the distribution for these two numbers, given the observed data and the current estimates for p_1 and p_2. In the M step, you need to maximize the expected value of what the log likelihood would be if you knew these numbers, with the expectation taken with respect to the distribution found in the E step (which is fixed for the duration of the M step). In the E step, you may find the distribution for the missing data only to the extend needed to do the M step (eg, it might be that only the expectations of some quantities are needed, not the full distribution). You should also output the log likelihood based on the observed data after each EM iteration, and verify that it never decreases (except perhaps slightly, as a result of round-off errors).

Hand in your derivation of what is needed for the E and M steps (which may be hand-written), the source for an R function that does maximum likelihood estimation using EM, and the source and output of a script that tests this function (source and output can be combined if you use `knitr::spin`). You should also briefly comment on how easy (or not) it is to use EM for this problem, and on how quickly it converges.

You may reuse your R code from Assignment #1, or if you prefer, the R code used in the solution to this assignment on the course web page.

Problem 2 [70 marks]: For this problem, you will analyse synthetic data on a set of beetles collected in some region.

For each beetle, indexed by i, we have measurements of its mass (in grams), m_i, the ratio of the length to width of its body, r_i, and whether it was collected in a swampy region, s_i (with $s_i = 1$ if in a swamp, $s_i = 0$ if not).

Some of the beetles have had their species determined, which we assume is a costly procedure. There are ten species that could be present in the collection region, which are identified by integers from 1 to 10. For other beetles, the species has not been determined, but the genus has been determined (which is less costly than determining the exact species). There are four genera, identified by integers from 1 to 4. For the remaining beetles, neither the species nor the genus have been determined (though it is known that they are of one of the ten species). The selection of beetles that had their exact species determined and that had their genus determined was random.

The genus of each species is as follows:

Species:	1	2	3	4	5	6	7	8	9	10
Genus:	1	1	1	2	2	3	3	4	4	4

Data on 500 beetles is available from the course web page. It has a header line, followed by 500 lines of data. If you download it and store it in the directory used by R, you can read it with the following R command:

```
data <- read.table("ass2-data",head=TRUE)
```

The model you should use for this data says that the measurements for different beetles are independent. Furthermore, given the species, the three measurements (m_i, r_i, and s_i) on one beetle are independent (but note that they are not unconditionally independent, when the exact species is not known). The distribution of the log of the mass of a beetle of species s is normal with mean μ_s and standard deviation 0.08. The distribution of the log of the ratio of length to width for a beetle of species s is normal with mean ν_s and standard deviation 0.10. The probability that a beetle of species s is collected from a swamp is ρ_s. Finally, the probability that a beetle collected will be of species s is α_s.

The set of model parameters therefore consists of μ_s, ν_s, ρ_s, and α_s for $s = 1, \ldots, 10$. The μ_s and ν_s parameters are unconstrained, ρ_s must be in $(0, 1)$, and the α_s must be positive and sum to one. (Note that in a real problem the standard deviations would also be parameters rather than being assumed known. I've specified them here to simplify the exercise, but it would be possible to estimate them using EM.)

You should write several R functions that are useful for maximum likelihood estimation for this problem, including one that computes the log likelihood for a set of parameters given the observed data, and one that finds the maximum likelihood estimates using EM (you may manually specify the number of iterations to do, rather than trying to determine this automatically). The EM function should print the log likelihood after each iteration, so that you can verify that it never goes down (except perhaps slightly, due to round-off errors). You may find that defining other functions as

well makes writing these functions easier. You should put all these function definitions in a source file separate from the R script file. This file of function definitions should not refer to the particular data set you are analysing.

In a separate R script file, you should run the EM function to find the maximum likelihood estimates for the data from the course web page. You should also produce a plot (or plots) that display the data and the estimates in a way that would provide insight to someone interested in this problem. To do this, you may find it useful to change the plotting symbol with the `pch` argument to `plot`, or to change the colour with the `col` argument, so that you can display two variables with horizontal and vertical position and another discrete variable or variables with symbol or colour. Note that you can put additional information on top of a plot using the `points` and `lines` functions.

You should hand in your derivations of the E and M steps of the EM algorithm (which may be hand-written), along with the source for your functions, the source and output of your script (which may be one file, if you use `knitr::spin`), and a brief discussion of how well EM seems to work for this problem (eg, how fast it converges), and how the problem might be solved in some other way (though you don't have to actually try any other methods).