# STA 410/2102, Spring 2004 — Assignment #3

Due at **start** of class on March 26. Worth 20% 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 use the EM algorithm to solve the same maximum likelihood estimatation problem as you did for Problem 3 of the last assignment. As in that assignment, the model uses a "wrapped normal" distribution for circular data, in which each data point is an angle. Angles will be measured in radians, ranging from 0 to $2\pi$. The density function for the wrapped normal distribution is as follows:

$$f(x) \;=\; \sum_{i=-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-2\pi i-\mu)^2\right)$$

If we have a set of i.i.d. data points, $x_1, \ldots, x_n$, we could model them directly, with the density function above, as in the last assignment. Alternatively, we could imagine that the amount of wrapping is an unobserved variable, with integer values $i_1, \ldots, i_n$ for the $n$ data points. If we observed these variables, as well as $x_1, \ldots, x_n$, the likelihood would be

$$\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_j-2\pi i_j-\mu)^2\right)$$

The maximum likelihood estimates for $\mu$ and $\sigma$ would simply be the sample mean of the $x_j - 2\pi i_j$ values and the sample standard deviation of these values (dividing by $n$ rather than $n-1$). Of course, we haven't actually observed $i_1, \ldots, i_n$. But we can use this view of the problem to create an EM algorithm to find the maximum likelihood estimate given just the observed $x_1, \ldots, x_n$.

Accordingly, you should write an R function to find a maximum likelihood estimate for both $\mu$ and $\sigma$, based on $x_1, \ldots, x_n$, using the EM algorithm. Unlike the case for the last assignment, there is no reason to transform $\sigma$ by taking its logarithm. Your function should take as arguments a vector of circular data points, $x$, and initial guesses for $\mu$ and $\rho$. It should return a list with elements called `mu` and `sigma` containing the maximum likelihood estimates. You may fix the number of iterations, rather than trying to determine convergence automatically. You should investigate how well this algorithm works — eg, does it converge? if so how fast? does the result depend on the starting point?. You should use the same data sets used for the last assignment, as well as any other data sets you think would be interesting to try.

You should hand in your derivations of the E and M steps for the algorithm, your program, the output of the tests, and your discussion of the results.