Mid-semester Test
STAT8178: Statistical Computing

Session 1, 2020 Due Friday 24 April, via *iLearn*, by 11:55 p.m.

General instructions

For the question below, you are expected to document your answer fully. Try to work within scripts or functions, rather than issuing commands at the command line, or manually manipulating spreadsheets, so as to leave a repeatable record of what you've done. If you *do* decide to manipulate data manually, please fully describe what you've done in your assignment submission, and submit any files (such as manipulated spreadsheets) that may clarify your work.

Where you are required to write code, please state and interpret the output, as well as submitting the code itself.

Also, some of your answers should be mathematical expressions. The best way to write these is LaTeX, but it has a long learning curve, and is probably not worth learning if you don't expect to do much technical writing after this unit. *Word*'s mathematical typesetting functionality has improved a great deal in the last decade; a quick intro to get you started appears at `https://en.wikibooks.org/wiki/Typing_Mathematics_in_Microsoft_Word`. If even that seems too hard and you want to just handwrite your equations, photograph them, and paste them into your answer submission, you may; I won't actually deduct marks, but you should feel vaguely ashamed, as if you've come to class with your shirt inside-out or something.

For convenience, the questions may describe working with either *Matlab* or *R*. As always, if you believe you can do equivalent computations on a different platform to the one described, you are free to do so, and I am happy to discuss it if you're not sure. Regardless of which programming language you're using, you will probably need to look things up. You should freely use the built-in documentation and internet searches to learn the details of how to use each function or command that you need.

Question 1: The EM algorithm (26 marks)

In this question, we will use the EM algorithm to fit a 'mixed' variant of an AR(1) process.

The sequence (X_n) of random variables is Markovian, following the rule

$$\begin{aligned} Z_n &= \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{otherwise,} \end{cases} \\ X_{n+1} &\sim \begin{cases} N(X_n/2, \sigma) & \text{if } Z_n = 1, \\ N(X_n, \sigma) & \text{if } Z_n = 0. \end{cases} \end{aligned} \tag{1}$$

The parameters p and σ are unknown, and will be estimated at the kth iteration by $p^{(k)}$ and $\sigma^{(k)}$. We assume that the sequence (X_n) is observed, but (Z_n) is unobserved.

Throughout these questions, you may reduce clutter (and make it easier for both you and the marker) by using f to represent the standard normal density, i.e.,

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, \tag{2}$$

so that a variable $X \sim N(\mu, \sigma)$ has density

$$f_X(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right).$$

(a) State the density of the conditional distribution of $X_{n+1} = x_{n+1}$ given that $X_n = x_n$ and $Z_n = z_n$. **(4 marks)**

(b) Suppose a sequence $(x_1, \ldots, x_{N+1})$ is observed from the above model. State ℓ_c, the complete-data log-likelihood. **(5 marks)**

(c) Letting $e_n^{(k)}$ abbreviate $E[Z_n|X_n = x_n, X_{n+1} = x_{n+1}, p = p^{(k)}, \sigma = \sigma^{(k)}]$, write an expression for the "Q" function, $Q^{(k)}(p, \sigma) = E[\ell_c(p, \sigma)|x_1, \ldots, x_{n+1}, p^{(k)}, \sigma^{(k)}]$. **(2 marks)**

(d) Taking $e_n^{(k)}$ and the sequence (x_n) as fixed, determine

$$(p^{(k+1)}, \sigma^{(k+1)}) = \arg\max Q^{(k)}(p, \sigma).$$

Hint: when you're differentiating Q with respect to σ, recall that $\ln f(t) = -t^2/2 - \ln(2\pi)/2$. **(5 marks)**

(e) Determine the value of $e_n^{(k)} = E[Z_n|X_n = x_n, X_{n+1} = x_{n+1}, p, \sigma]$. A couple of hints: since Z_n is a binary variable, its expectation is just the probability that it equals 1. Also, you can freely ignore the distinction between probability and density in this question; for instance, it's okay to write

$$P[X_{n+1} = x_{n+1}, Z_n = 1|X_n = x_n, p, \sigma] = P[Z_n = 1|X_n = x_n, p, \sigma]P[X_{n+1} = x_{n+1}|Z_n = 1, X_n = x_n, p, \sigma]$$

even though the middle P is a probability and the other two are probability densities. **(3 marks)**

(f) Write code to fit this model via the EM algorithm. Execute the code on the sequence found in either of the files `2020MidSem.txt` or `2020MidSem.mat`. Report and interpret the results. **(7 marks)**