STA 437/1005, Fall 2008 — Assignment #1

Due on October 6, at start of lecture. Worth 10% of the course grade.

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 of this assignment with any written notes or other recordings, nor receive any written or other material from anyone else by other means such as email.

For all questions, show both the final answer and how you obtained it. Wherever possible, your answers should take the form of specific numbers (fractions or decimal numbers), not just a formula that could produce these numbers.

Question 1: Let X, Y, and Z be independent random variables, all with the $N(1,1)$ distribution. Define the random variables A and B as follows:

$$A = X + 2Y \quad (1)$$
$$B = X + Y + Z \quad (2)$$

Finally define C as the random vector with A and B as components (ie, $C = [A \;\; B]'$).

a) What is the mean vector of C?

b) What is the covariance matrix of C?

c) What is the conditional distribution of A given that $B = 1$?

Question 2: Suppose X is a random vector of length p with covariance matrix Σ_X. Define $Y = QX$, where Q is some $p \times p$ orthogonal matrix, and let Σ_Y be the covariance matrix of Y.

a) Find a simple expression for Σ_Y.

b) Suppose that e is an eigenvector of Σ_X with eigenvalue λ. Prove that Qe is an eigenvector of Σ_Y, and find what eigenvalue is associated with it.

Question 3: Recall the spectral decomposition theorem: If A is a $k \times k$ symmetric real matrix, it is possible to find a set of k eigenvectors of A that are orthogonal and have length one, and if $e_1, \ldots, e_k$ are any such set of eigenvectors, with eigenvalues $\lambda_1, \ldots, \lambda_k$, then $A = \lambda_1 e_1 e_1' + \cdots + \lambda_k e_k e_k'$.

Use this theorem to prove that if A is a symmetric matrix with eigenvectors $e_1, \ldots, e_k$ that are orthogonal and have length one, with non-zero eigenvalues $\lambda_1, \ldots, \lambda_k$, then $B = e_1 e_1'/\lambda_1 + \cdots + e_k e_k'/\lambda_k$ is the inverse of A. Note that although there may be several ways of proving this, for this question you should prove it by multiplying A and B and verifying that the result is the identity matrix, using the spectral decomposition theorem.

Question 4: The effect (if any) of air pollution on mortality has been studied for many years, and has large implications for public policy. From the course web page,

`http://www.utstat.toronto.edu/~radford/sta437`

you can get a file containing daily data on weather, air pollution, and deaths in Toronto from 1992 to 1997. This data file contains 2192 lines, one per day, in time order, with each line containing the values of 10 variables. There is also a header line at the front with the names of the variables.

The variables are as follows:

`year`	Year, from 1992 to 1997
`month`	Month, from 1 (January) to 12 (December)
`day`	Day of the month, from 1 to 31
`deaths`	Number of deaths in Toronto
`pressure`	Average air pressure, in kilopascals
`temperature`	Average temperature, in degrees Celcius
`humidity`	Average relative humidity, percent
`so2`	Average level of sulfer dioxide, ppb
`ozone`	Average level of ozone, ppb
`pm10`	Average level of 10 micron particulate matter, micrograms per cubic meter

The `pm10` variable is observed on only some days, with the value for other days being set to `NA`.

In interpreting this data, it is important to note that the weather itself is known to have an effect on mortality, and that the weather also has an effect on the level of pollutants (`so2`, `ozone`, and `pm10`).

Read this data into R, look at it, and report any conclusions you may find about how these variables are related, whether they have normal distributions, and how they might be transformed to have distributions closer to normal. In your report, include a small number of plots or other R output that justifies your conclusions.

In your report, you should also discuss to what extent this data can be regarded as a random sample from a distribution that is of interest regarding the question of whether air pollution has an effect on mortality.

For this assignment, you need not perform any formal statistical tests. You should just make informal assessments based on plots and sample statistics, and using your common sense knowledge.