Math 6020-1, Spring 2015; Assignment 2

Due on Wednesday March 4, 2015

1. Read Chapter 8.

2. Consider the following lower half of the sample correlation matrix for $(x_1, \ldots, x_7)$,[1]

$$\mathbf{R} = \begin{pmatrix} 1.000 & & & & & & \\ 0.402 & 1.000 & & & & & \\ 0.396 & 0.618 & 1.000 & & & & \\ 0.301 & 0.150 & 0.321 & 1.000 & & & \\ 0.305 & 0.135 & 0.289 & 0.846 & 1.000 & & \\ 0.339 & 0.206 & 0.363 & 0.759 & 0.797 & 1.000 & \\ 0.340 & 0.183 & 0.345 & 0.661 & 0.800 & 0.736 & 1.000 \end{pmatrix},$$

   where $x_1, \ldots, x_7$ respectively denote the following:[2]

   (a) x_1 = head length;

   (b) x_2 = head width;

   (c) x_3 = face width;

   (d) x_4 = left finger length;

   (e) x_5 = left forearm length;

   (f) x_6 = left foot length; and

   (g) x_7 = height.

   [W.R. MacDonnell (1902) *Biometrika*, vol. 1, pp. 177–277].

   (a) Perform a principle components analysis of the correlation matrix $\mathbf{R}$. If we call the correlation data CrimCorr,[3] then in R you type

   ```
   PcaOutput <-  princomp(covmat = CrimCorr).
   ```

   The command

[1] The upper half can be filled in by symmetry.

[2] These are some of the physical characteristics taken from a sample of 3000 convicted criminals. But this fact is not germaine to our present discussion.

[3] It would be not a good idea to call it $\mathbf{R}$ in R!

```
summary(PcaOutput , loadings = TRUE)
```

will then produce:

i. A table of the 7 principle components against their [sample] SDs, proportion of variance caused by each PC, and their cumulative proportions; together with

ii. A table of loadings. These are the sample correlations between the original data and the principle components.

(b) Learn about the R commands princomp and summary. In particular, learn how you can access the output variables of princomp.[4]

(c) Create a "scree diagram" of your data. That is, plot i versus $\widehat{\lambda}_i$ for $i = 1, \ldots, 7$, where $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_7$ denote the [ordered] eigenvalues of $\mathbf{R}$. Look for "elbows" and match the elbows—if there are any—with the proportions of the total variance caused by each eigenvalue. Do you see any? Which might be a good candidate?

(d) Does it make sense to perform dimension/variable reduction in this problem?

(e) Repeat the preceding, but use Farmer's log-scree diagram wherein you plot i versus $\log \widehat{\lambda}_i$ instead of $\widehat{\lambda}_i$. Look for "elbows," etc.

(f) Interpret as many of the principle components as you can. [I.e., can you attach a "meaning" to the PCs?]

(g) How does your analysis teach us about the population from which this particular data set might have come from? You may assume that the data is in fact representative of that population [e.g., the sample was a simple random sample, etc.].

[4] For instance, here, PcaOutput$sdev^2 will show you the sample standard deviations of the principle components. Are there other output variables? If there are, then how do you access them?