Assignment #2 STA437H1S/2005H1S

due Friday February 26, 2016

Instructions: Students in STA437S do problems 1 through 3; those in STA2005S do all 4 problems.

1. Andrews curves (conceived the University of Toronto's own David Andrews) represent an interesting approach to multivariate visualization. The idea is to represent each multivariate observation $(x_{i1}, \cdots, x_{ip})$ (which is possibly normalized) by a sinusoidal function on $[0, 1]$:

$$g_i(t) = \frac{x_{i1}}{\sqrt{2}} + x_{i2}\sin(2\pi t) + x_{i3}\cos(2\pi t) + x_{i4}\sin(4\pi t) + x_{i5}\cos(4\pi t) + \cdots$$

Observations that are similar will have similar Andrews curves while outlying observations will often have curves that are distinctively different.
On Blackboard, there is a file `andrews.txt`, which contains a function `andrews` that computes Andrews curves for a data matrix whose columns are variables and rows are observations; for example,

```
> source("andrews.txt")  # read the function into R
> x <- cbind(rnorm(100),rnorm(100),rnorm(100),rnorm(100),rnorm(100))
> r <- andrews(x,scale=T) # scales columns to have mean 0 and variance 1
```

The file `testdata.txt` contains $100 - k$ observations from a 10-variate normal distribution and k outliers generated from another distribution (where $k \leq 15$).
(a) Look at the data using Andrews curves. How many clear outliers do there seem to be?
(b) Using the information from the Andrews curves as well as pairwise scatterplots, principal components etc, give an estimate of how many outliers are in the data.

2. (a) If $\{g_i(t)\}$ are the Andrews curves defined in question 1, show that

$$2\int_0^1 [g_i(t) - g_j(t)]^2\, dt = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2.$$

(b) If $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$, what is the Andrews curve of $\bar{\boldsymbol{x}}$?
(c) Suppose that $\boldsymbol{x}_k$ lies on a line between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, that is, $\boldsymbol{x}_k = \lambda\boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_j$ for some $0 < \lambda < 1$. What can you say about the Andrews curve of $\boldsymbol{x}_k$ relative to those of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$?

3. In Assignment #1, you looked at two dimensional scatterplots of data on two species of rock crabs; here, you will do a principal components analysis of these data.
As before, the data are in a file `crabs.txt` on Blackboard; the columns of the file are species (B or O), sex (M or F), index (1-50 within each species-sex combination), width of the frontal lip (LP), the rear width of the shell (RW), length along the midline of the shell (CL), the maximum width of the shell (CW), and the body depth (BD).
The data can be read into R using the following code:

```
> x <- scan("crabs.txt",skip=1,what=list("c","c",0,0,0,0,0,0))
> colour1 <- ifelse(x[[1]]=="B","blue","orange") # species colours
> colour2 <- ifelse(x[[2]]=="M","black","red") # sex colours
> sex <- x[[2]]
> FL <- x[[4]]
> RW <- x[[5]]
> CL <- x[[6]]
> CW <- x[[7]]
> BD <- x[[8]]
```

(a) Using the correlation matrix, do a principal component analysis of the 5 variables.

```
> r <- princomp(~FL+RW+CL+CW+BD,cor=T)
> summary(r,loadings=T)
```

Give an interpretation of the first two principal components based on their loadings.
(b) Look at pairwise scatterplots of the 5 principal components using `colour1` to distinguish the two species:

```
> pairs(r$scores,col=colour1)
```

Which pairs of principal components seem to separate the two species?
(c) Now look at pairwise scatterplots of the 5 principal components using `colour2` to distinguish the two sexes:

```
> pairs(r$scores,col=colour2)
```

Which pairs of principal components seem to separate the two sexes?
(d) Suppose you are given the following measurements for the 5 variables: FL = 18.7, RW = 15.0, CL = 35.0, CW = 40.3, BD = 16.6. What is your prediction of the species and sex of this crab?

4. In projection pursuit and independent component analysis, kurtosis is often used as a measure or index of non-normality. The rationale for this is that if we add independent random variables, the kurtosis of the sum will be closer to the kurtosis of a normal distribution than the kurtosis of each of the random variables in the sum.
In this problem, suppose that X_1 and X_2 are independent random variables with mean 0 and variance 1.
(a) Suppose that a_1 and a_2 are constants with $a_1^2 + a_2^2 = 1$. Show that $E(a_1X_1 + a_2X_2) = 0$ and $\text{Var}(a_1X_1 + a_2X_2) = 1$.
(b) Show that $E[(a_1X_1 + a_2X_2)^4] = a_1^4E(X_1^4) + 6a_1^2a_2^2 + a_2^4E(X_2^4)$.
(c) Show that

$$E[(a_1X_1 + a_2X_2)^4] - 3 = a_1^4\left\{E(X_1^4) - 3\right\} + a_2^4\left\{E(X_2^4) - 3\right\}$$

and so

$$\left|E[(a_1X_1 + a_2X_2)^4] - 3\right| \leq a_1^4 \left|E(X_1^4) - 3\right| + a_2^4 \left|E(X_2^4) - 3\right|.$$

(Hint: Write $3 = 3(a_1^2 + a_2^2)$ and $6a_1^2a_2^2 = 3a_1^2a_2^2 + 3a_1^2a_2^2$.)

(d) Show that when a_1 and a_2 are both non-zero

$$\left|E[(a_1X_1 + a_2X_2)^4] - 3\right| < \max\left\{\left|E(X_1^4) - 3\right|, \left|E(X_2^4) - 3\right|\right\}$$

unless $E(X_1^4) = E(X_2^4) = 3$ (in which case $E[(a_1X_1 + a_2X_2)^4] = 3$).