## Multivariate Statistical Analysis
## Statistics 4223/5223 — Spring 2018

## Assignment 4

*Reading:*

By Tuesday, March 20, read Chapters 10–11 and 16 of *Applied Multivariate Statistical Analysis, fourth edition*; by Wolfgang Härdle and Léopold Simar.

For Thursday, March 22, read Chapters 12 and 14 of Härdle and Simar.

*Homework 4:*

The following problems are due before class on Tuesday, March 27. Homework can also be submitted to the course mailbox in Room 904 SSW, any time before 5:00pm on Wednesday, April 4.

1. The data in `Number_Parity.csv` (file can be found in the `Data` folder on Courseworks) were collected to test a psychological model of numerical cognition: How does the processing of numbers depend on the way the numbers are presented (words versus Arabic digits)?

   Thirty-two subjects were required to make a series of quick numerical judgments about two numbers presented either as two words (two vs. four, for example) or two single Arabic digits (2 vs. 4). The subjects were asked to respond "same" if the two numbers had the same numerical parity (both even or both odd) and "different" if the two numbers had a different parity (one even one odd). For each of the four combinations of parity and format, the median reaction times for correct responses were recorded for each subject.

$$\begin{aligned} x_1 &= \textit{WordDiff} &&= \text{reaction time for word format, different parity} \\ x_2 &= \textit{WordSame} &&= \text{reaction time for word format, same parity} \\ x_3 &= \textit{Num\_Diff} &&= \text{reaction time for Arabic numeral, different parity} \\ x_4 &= \textit{Num\_Same} &&= \text{reaction time for Arabic numeral, same parity} \end{aligned}$$

   Conduct a repeated measures analysis on these data.

   (a) Assess the reasonableness of assuming the data are a random sample from a multivariate normal population.

   (b) Test the null hypothesis of no treatment effect. That is, find and interpret the $p$-value for a test of $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$.

(c) Recall that simultaneous $1-\alpha$ confidence intervals for all contrasts $\boldsymbol{b}'\boldsymbol{\mu}$ are given by

$$\boldsymbol{b}'\bar{\boldsymbol{x}} \pm \sqrt{\frac{(n-1)(p-1)}{n-p+1}F_{1-\alpha;p-1,n-p+1}}\sqrt{\frac{1}{n}\boldsymbol{b}'\mathbf{S}\boldsymbol{b}}$$

Compute and interpret simultaneous 95% confidence intervals for

i. the contrast for parity effect (different vs. same), averaged over word format and Arabic digits;

ii. the contrast for format effect (word vs. numeral), averaged over same and different parity; and

iii. an *interaction contrast* measuring the difference in parity effect for word format versus parity effect given Arabic digits (or equivalently, the difference in format effect for different parity versus format effect given same parity).

2. The data file `Turtles.csv`, in the `Data` folder on Courseworks, contains measurements of carapace (shell) dimensions for 24 female and 24 male painted turtles: $x_1$ = length, $x_2$ = width and $x_3$ = height, all in millimeters.

Assume the female and male turtles are independent random samples from trivariate normal distributions with a common covariance matrix; denote the mean vector for female turtles by $\boldsymbol{\mu}_1$ and that of male turtles by $\boldsymbol{\mu}_2$.

(a) Test for equality of the two population mean vectors, $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$; report and interpret a $p$-value.

(b) Use the Bonferroni method to find simultaneous 95% confidence intervals for the component mean differences. Interpret your intervals.

3. The data file `Track_Records.csv`, in the `Data` folder on Courseworks, contains the national track records for women in $n = 54$ countries, in $p = 7$ different running events.

| Variable | Event |
|---|---|
| $x_1$ | 100 meter (sec) |
| $x_2$ | 200 meter (sec) |
| $x_3$ | 400 meter (sec) |
| $x_4$ | 800 meter (min) |
| $x_5$ | 1500 meter (min) |
| $x_6$ | 3000 meter (min) |
| $x_7$ | Marathon (min) |

(a) Obtain the sample correlation matrix **R** for these data, and determine its eigenvalues and eigenvectors.

(b) Calculate the proportion of total (standardized) sample variance explained by each (normalized) principal component, and prepare a graphical summary in the form of a *scree plot.* Also find the proportion of (standardized) variance explained by the first $k$ (normalized) principal components for $k = 1, 2, \ldots, 7$. How many NPCs should we retain if our goal is to account for 90% of total (standardized) variance?

(c) Interpret the first two NPCs.

(d) Rank the nations based on their score on the first (normalized) principal component. Does this ranking correspond with your previous notion of athletic excellence for the various countries?

(e) Make a scatterplot of the first two (normalized) principal components. Identify the points corresponding to Samoa, the Cook Islands, and North Korea, and explain what about those countries makes them stand out in the plot as they do.

4. Continue with the women's national track records data from the previous exercise.

(a) Convert each record to an average speed for the race, measured in meters per second. Notice that the records for 800 m, 1500 m, 3000 m, and the marathon are given in minutes. The marathon is 26.2 miles, or 42,195 meters, long.

(b) Perform a principal components analysis using the covariance matrix **S** of the speed data. Again find the proportion of variance explained by each of the first $k$ principal components for $k = 1, 2, \ldots, 7$. How many principal components should we retain if our goal is to account for 90% of total sample variance?

(c) Interpret the first two principal components. Are these interpretations similar to those of the first two NPCs in the previous exercise?

(d) If the nations are ranked on the basis of their first principal component score, does the subsequent ranking differ notably from that in the previous exercise?

(e) Make a scatterplot of the first two principal component scores, and label the points by the countries' abbreviations. Comment on the difference between this plot and the corresponding plot for the previous exercise.