Due date: 16 October, 2018 11:59pm via Blackboard

Instruction: Please prepare your answers in a Word document and clearly label all your answers and output for their corresponding questions. Use Word equation editor (or any equation editor of your choosing) to write all equations.
All necessary output should be placed within the question or placed in the appendix and must be referred in the discussion.

Objectives of this assignment: To provide students with the opportunity to reinforce taught concepts including correlation analysis, PCA and FA in real world applications.

Question 1 (10 Marks)

A naturalist for the Alaska Fish and Game Department studies grizzly bears with the goal of maintaining a healthy population. Measurements on n=61 bears provided the following covariance matrix:

$$
\begin{bmatrix}
3266.46 & 1343.97 & 731.54 & 1175.50 & 162.68 & 238.37 \\
1343.97 & 721.91 & 324.25 & 537.35 & 80.17 & 117.73 \\
731.54 & 324.25 & 179.28 & 281.17 & 39.15 & 56.80 \\
1175.50 & 537.35 & 281.17 & 474.98 & 63.73 & 94.85 \\
162.68 & 80.17 & 39.15 & 63.73 & 9.95 & 13.88 \\
238.37 & 117.73 & 56.80 & 94.85 & 13.88 & 21.26
\end{bmatrix}
$$

Where the variables are X1: Weight, X2: body length (cm), X3: Neck (cm), X4: Girth (cm), X5: Head length (cm) and X6: Head width (cm)

a. Perform a principal component analysis using the covariance matrix? Can the data be effectively summarized in fewer than six dimensions? Comment on it. **(3 Marks)**
b. Perform a principal component analysis using the correlation matrix **(4Marks)**
c. Comment on the similarities and differences between the two analyses. **(3 Marks)**

Question2 (30 Marks)

The file *wine.csv* contains data on concentrations of 13 different chemicals in wines grown in the same region in Italy that are derived from three different cultivars. The dataset is available in Blackboard.

    a. Download the dataset and prepare the dataset for the analysis. (investigate any outliers or missing values and produce summary statistics) **(5 Marks)**
    b. Produce the correlation matrix and see whether it is suitable for principle component analysis **(5 Marks)**
    c. Compute the mean and standard deviation for the 13 chemical concentrations in the sample wine and comment on **(5 Marks)**
    d. Perform PCA and see how many PCs need to retain **(5 Marks)**
    e. Obtain the scree plot and perform the parallel analysis to confirm your choice. **(5 Marks)**
    f. Write down the selected PCAs as the linear combination of the original variables. **(5 Marks)**

Question 3 (40 Marks)

Dataset *Household.csv* contains the responses of a questionnaire on housing. In the questionnaire, participants were asked to rate their satisfaction on a scale from 1 to 4 for 30 related questions, with 4 indicating most satisfaction and 1 indicating most dissatisfaction. The dataset is available in Blackboard.

    a. Download the dataset and prepare the dataset for the analysis. (investigate any outliers or missing values and produce summary statistics) **(4 Marks)**
    b. Provide a correlation matrix of all items analysed. What information can you gather from the correlation matrix? **(3 Marks)**

    c. Discuss the suitability of the dataset for factor analysis using KMO and Bartle test. **(3 Marks)**

    d. Run a factor analysis (using 'minres' method) using the eigenvalue criterion as an initial guide for the number of factors to be extracted and confirm using the parallel analysis.**(5 Marks)**

    e. Rerun factor analysis using based on your number of factors that you have chooses from part(d) **(5 Marks)**

    f. Discuss whether you have obtained the simple structure **(5 Marks)**

    g. Apply factor rotation(varimax and oblimin) methods and see whether the possibility of simple structure. Which one is a better fit? Why? **(5Marks)**

    h. Produce residual analysis and discuss whether the selected model fit the data well. **(5 Marks)**

    i. Rerun analysis until you have simple structure removing the items **(5 Marks)**