Multivariate Analysis: Midterm Exam

Due 5/10 in class

Announcement: Please follow the guidelines to write your report:

- The presentation should be as formal as possible (with a title page, names and IDs, and necessary outputs and plots);
- Please (i) clearly address the problem; (ii) clearly describe the basic ideas of each method you use; (iii) clearly introduce every step required to approach the answer; (iv) carefully interpret the results.
- **Do not** attach any R code in your report;
- Writing either in English or Chinese is fine.

1) Download "European_Jobs.txt" and "European_Jobs_Description.txt" for the data set and its detailed variable description.
   Q1: Perform a complete **Principal Components Analysis** for this data and interpret the result.
   Note: The number of PCs must be determined by a formal statistical hypothesis test, while the relationships among objects and variables can be interpreted by using a 2D plot.

2) Suppose a human resource researcher would like to study the relationships between the employment proportions of various types of industry in Europe. He divides the variables in the data "European_Jobs.txt" into two groups: (Agr, Min, Man, PS, Con) and (SI, Fin, SPS, TC). The first group represents the industries that are more labor-intensive, while the second group represents the industries that are less labor-intensive.
   Q2: Perform a complete **Canonical Correlation Analysis** for these two groups of variables and interpret the result.
   Note: The number of canonical variates must be determined by a formal statistical hypothesis test, while the required model assumptions need to be validated.

3) The data set "glass.dat" was collected by the department of criminological investigation, it consists of 114 observations with 6 types, where each observation is described by 9 attributes (please refer to "glass_description.txt" for variable description). At the scene of the crime, the glass left can be used as evidence, if its type is correctly identified.
   Q3: Construct the decision rules for **classifying the types of glass** using (i) Classification Tree; (ii) LDA; (iii) QDA; (v) Nearest Neighbor; and (vi) Logistic discrimination. Compare all your resulting decision rules and explain which one you will best recommend.
   Q4: Based on the control of "Classification Tree" obtained in **Q3**, develop two decision rules by using the strategies of (i) random forest and (ii) boosting. How do these two strategies compare with the decision rules in Q3 in terms of prediction accuracy?

4) The data "new_wages.txt" contains information regarding types of wages and some categorical characteristics from a random sample of 534 persons. The description of all variables is given in "new_description.txt".
   Q5) Construct a suitable decision rule to **classify the types of wages** based on the methods introduced in our class. Also, comment on the performance of your decision rule in terms of prediction accuracy.