Homework 5 Due date: March 30

1. Given the car price data (https://www.kaggle.com/hellbuoy/car-price-prediction),
   (a) Determine the sample principal components and their variances with all the continuous variables except for *price* (*wheelbase, carlength, carwidth, carheight, curbweight, enginesize, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg*). Please standardize each variable before conducting the analysis.
   (b) Please choose the proper number of principal components that can best explain the variation of the 13 variables in (a) according to the variance explained and the scree plot. Interpret the principal components.
   (c) Calculate statistical distances $(\boldsymbol{x_j} - \bar{\boldsymbol{x}})' S^{-1} (\boldsymbol{x_j} - \bar{\boldsymbol{x}})$ for all observations with the 13 variables in (a). Determine the proportion of observations $\boldsymbol{x_j}$ falling within the squared distance 5^2.
   (d) For the vectors $\boldsymbol{v_j}$ formed by the first two principal component scores, Sketch the contour $(\boldsymbol{v_j} - \bar{\boldsymbol{v}})' \Lambda^{-1} (\boldsymbol{v_j} - \bar{\boldsymbol{v}}) = 5^2$, where Λ is a diagonal matrix with the first two eigenvalues of S in the diagonal. Highlight the points selected in (c) with the red color. Discuss the consistency/inconsistency of the contour and the highlighted points.
   (e) Compute the correlation coefficient between the first two principal component scores and the variable *price*.
   (f) Retrieve the first two sample principal components with all the continuous variables except for *price* (*wheelbase, carlength, carwidth, carheight, curbweight, enginesize, boreratio, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg*). Make a biplot for the data. Color the points according to the categorical variable *fuelsystem* and interpret the results. You should report which of the two coordinate systems you use in this report.
2. Prove the following results in the note (p37).
   (a) The squared Euclidean distance between any two points in the principal component space is the same as the Mahalanobis distance between them in the original space.
   (b) The squared length of the vector from the origin to the coordinates representing a particular variable reflects the variance of that variable.
   (c) The correlation of two variables is reflected by the angle between the two

corresponding vectors for the two variables.

3. Given the gene expression data from 廖冠儒 in Homework1, please first calculate the mean expression profiles with the 11 genes under all combinations of TNBC Status, STAGE, and pCR. You should get a table with dimension 16x11. Calculate the correlations between any pairs of rows and use 1-|correlation| as the distance matrix to conduct a classical MDS analysis. Color your points according to the treatment response (pCR vs RD).
4. Watch the following video for additional lectures. https://drive.google.com/file/d/1k6bu-In8NS18axvaQeACu4nQ7uErOoLR/view?usp=sharing
5. Watch the following video for additional lectures. https://drive.google.com/file/d/1h_jMPLwP8FlSbS0ffORcmgTL7xTgieXe/view?usp=sharing