# **BS806 --- Homework 10**

**Reading**. Chapter 10 of (ISLR) Introduction to statistical learning. Pages 385—401, 404--413.

**Question 1.**

1. Simulate a sample of 300 observations which include 4 variables from 3 different clusters (each cluster includes 100 observations) with centroids defined as **[10 pts]:**

We will use this simulated dataset (with dimension of 300 × 4) for the following analysis. Standardize the data if appropriate.

1. Generate principal components of the variables, and plot the first two PCs. Does the plot of the first two PCs suggest that there are 3 clusters in the data? **[20pts]**
2. Use k-means clustering to discover groups in the data. **[25pts]**
   1. Analyze the within cluster sum of squares (WSS) for different numbers of clusters (i.e. vary the number of k) to decide what is the most likely number of clusters. How many potential clusters in the data?

We will assume there are two clusters for b and c.

* 1. Generate the cluster centroids assuming there are two clusters.
  2. Visualize the clusters using the first two PCs and different color representing different clusters assuming there are two clusters.

1. Use hierarchical clustering with complete linkage to generate a dendrogram of the data. **[10pts]**
2. Use the resampling technique shown in class to decide whether there are evident cluster. **[10pts]**
3. Generate a heatmap of the data and show the significant clusters based on resampling approach from Question 5 at 1% level of significance (hints: use cluster argument within annHeatmap2() function). **[15pts]**
4. Compare the agreement for the clusters discovered with k-means clustering and hierarchical clustering based on resampling approach at 1% level of significance. **[10pts]**