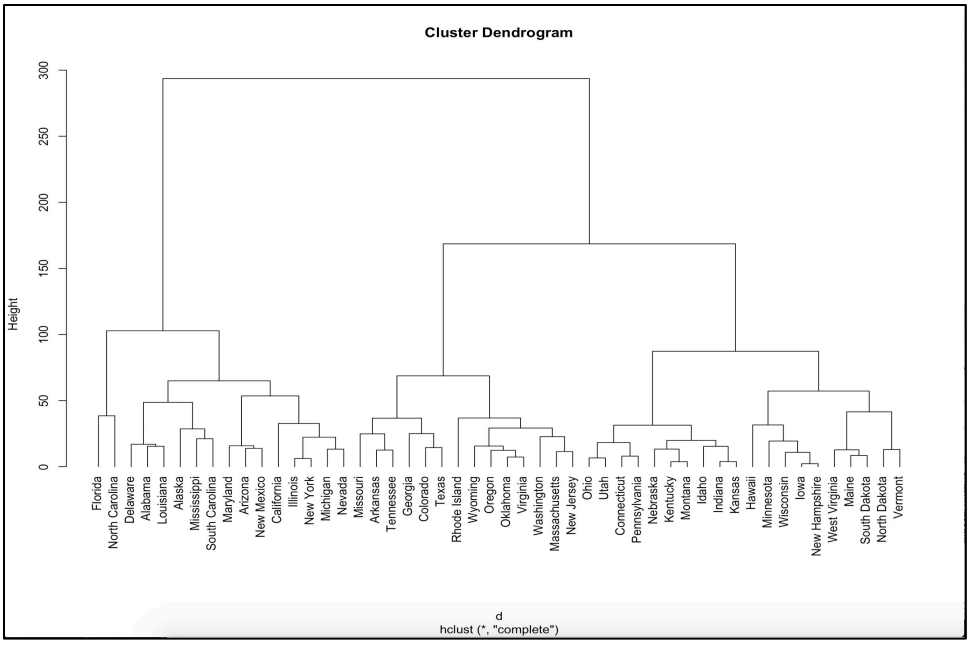
Assignment 2

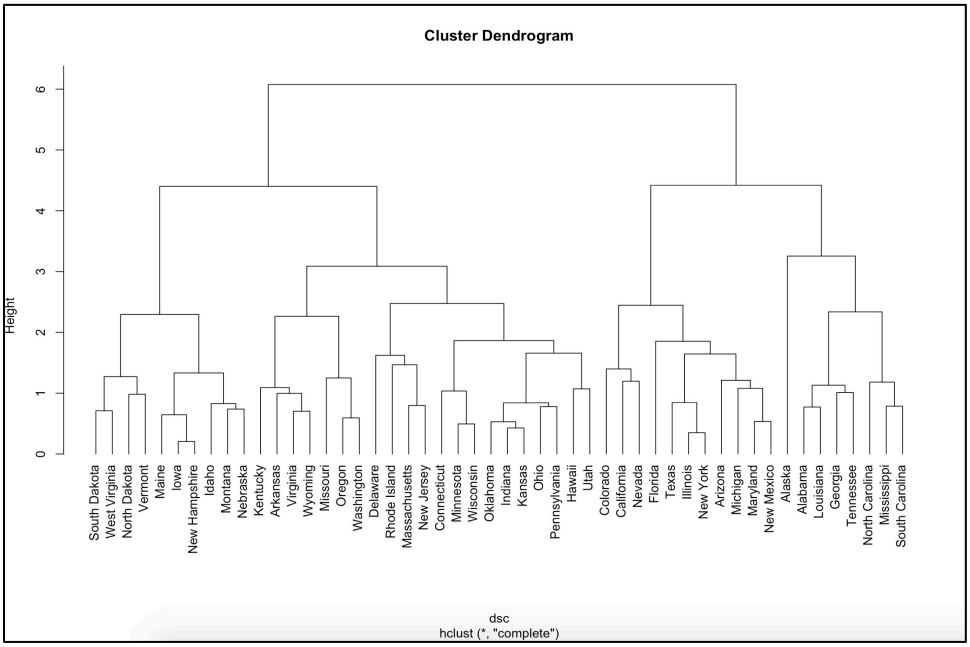
Question 1

Consider the USArrests data. We will now perform hierarchical clustering on the states.
(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.
(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?
(c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.
(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed?
Provide a justification for your answer.

Post hierarchical clustering below is the dendrogram obtained



Dendogram post scaling the variables



Clustering result pre and post scaling

State	No Scaling	Scaling	State	No Scaling	Scaling	State	No Scaling	Scaling
Alabama	1	1	Louisiana	1	1	Ohio	3	3
Alaska	1	1	Maine	3	3	Oklahoma	2	3
Arizona	1	2	Maryland	1	2	Oregon	2	3
Arkansas	2	3	Massachusetts	2	3	Pennsylvania	3	3
California	1	2	Michigan	1	2	Rhode Island	2	3
Colorado	2	2	Minnesota	3	3	South Carolina	1	1
Connecticut	3	3	Mississippi	1	1	South Dakota	3	3
Delaware	1	3	Missouri	2	3	Tennessee	2	1
Florida	1	2	Montana	3	3	Texas	2	2
Georgia	2	1	Nebraska	3	3	Utah	3	3
Hawaii	3	3	Nevada	1	2	Vermont	3	3
Idaho	3	3	New Hampshire	3	3	Virginia	2	3
Illinois	1	2	New Jersey	2	3	Washington	2	3
Indiana	3	3	New Mexico	1	2	West Virginia	3	3
Iowa	3	3	New York	1	2	Wisconsin	3	3
Kansas	3	3	North Carolina	1	1	Wyoming	2	3
Kentucky	3	3	North Dakota	3	3			

Hierarchical clustering stats pre and post scaling

Clustering Post Scaling -->		1	2	3
Clustering Pre Scaling	1	6	9	1
	2	2	2	10
	3	0	0	20

The data should be scaled before processing since every measure can have different units and order of magnitude.

Question 2

In this problem, you will generate simulated data, and then perform PCA and K-means clustering on the data.

(a) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables.
Hint: There are a number of functions in R that you can use to generate data. One example is the rnorm() function; runif() is another option. Be sure to add a mean shift to the observations in each class so that there are three distinct classes.

(b) Perform PCA on the 60 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, then continue on to part (c). If not, then return to part (a) and modify the simulation so that there is greater separation between the three classes. Do not continue to part (c) until the three classes show at least some separation in the first two principal component score vectors.

(c) Perform K-means clustering of the observations with K = 3. How well do the clusters that you obtained in K-means clustering compare to the true class labels?
Hint: You can use the table() function in R to compare the true class labels to the class labels obtained by clustering. Be careful how you interpret the results: K-means clustering will arbitrarily number the clusters, so you cannot simply check whether the true class labels and clustering labels are the same.

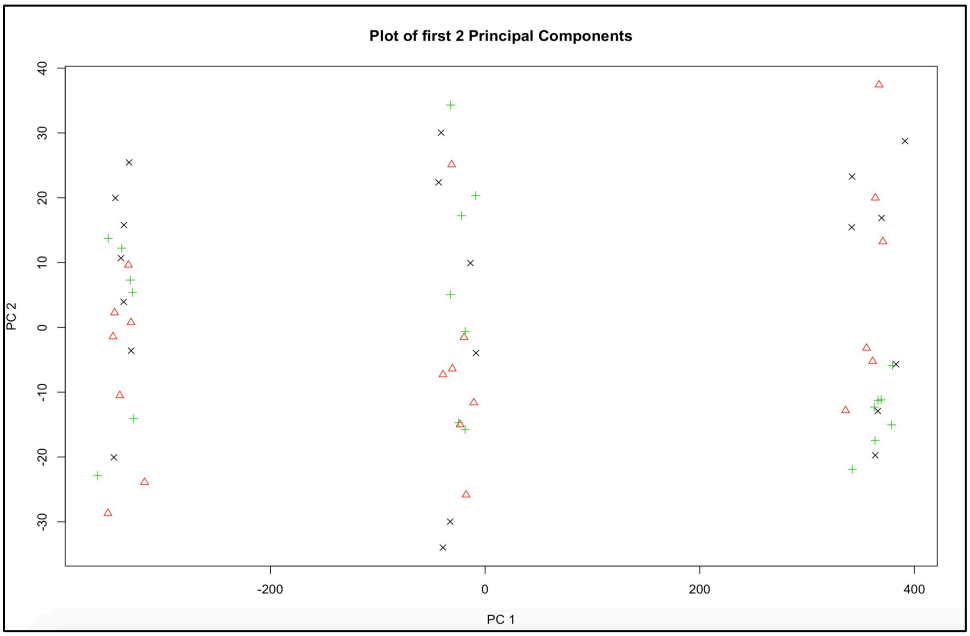
(d) Perform K-means clustering with K = 2. Describe your results.

(e) Now perform K-means clustering with K = 4, and describe your results.

(f) Now perform K-means clustering with K = 3 on the first two principal component score vectors, rather than on the raw data. That is, perform K-means clustering on the 60 × 2 matrix of which the first column is the first principal component score vector, and the second column is the second principal component score vector. Comment on the results.

(g) Using the scale() function, perform K-means clustering with K = 3 on the data after scaling each variable to have standard deviation one. How do these results compare to those obtained in (b)? Explain.

Part b:
Post PCA on the data generated from uniform distribution below is the plot of first 2 principal components.



As per the plot we can see clear separation between three classes.

Part c:
Performing K-means clustering with K=3 below is the result of the clustering.

True Labels -->		1	2	3
Clustering Labels	1	20	0	0
	2	0	20	0
	3	0	0	20

Part d:
Performing K-means clustering with K=2 below is the result of the clustering.

True Labels -->		1	2	3
Clustering Labels	1	0	20	0
	2	20	0	20

Part e:
Performing K-means clustering with K=2 below is the result of the clustering.

True Labels -->		1	2	3
Clustering Labels	1	0	0	20
	2	10	0	0
	3	10	0	0
	4	0	20	0

Part f:
Performing K-means clustering with K=3 using first 2 Principal Components below is the result of the clustering.

True Labels -->		1	2	3
Clustering Labels	1	20	0	0
	2	0	20	0
	3	0	0	20

Part g:
Performing K-means clustering with K=3 using scaled data below is the result of the clustering.

True Labels -->		1	2	3
Clustering Labels	1	0	0	20
	2	0	20	0
	3	20	0	0

Question 3

On the book website, www.StatLearning.com, there is a gene expression data set (Ch10Ex11.csv) that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

(a) Load in the data using read.csv(). You will need to select header=F.

(b) Apply hierarchical clustering to the samples using correlation based distance, and plot the dendrogram. Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used?

(c) Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to answer this question, and apply it here.

Inline comments in code