Homework 2
Applied Machine Learning
Fall 2017
CSCI-P 556/INFO-I 526

Soutri Mukherjee
soumukh@iu.edu

October 6, 2017

Problem 1 [20 points]

$$\begin{pmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{pmatrix}$$

a)Let the points be A,B,C,D Hence the dissimilarity matrix can be written as

$$\begin{pmatrix} & A & B & C & D \\ A & 0 & 0.3 & 0.4 & 0.7 \\ B & 0.3 & 0 & 0.5 & 0.8 \\ C & 0.4 & 0.5 & 0 & 0.45 \\ D & 0.7 & 0.8 & 0.45 & 0 \end{pmatrix}$$

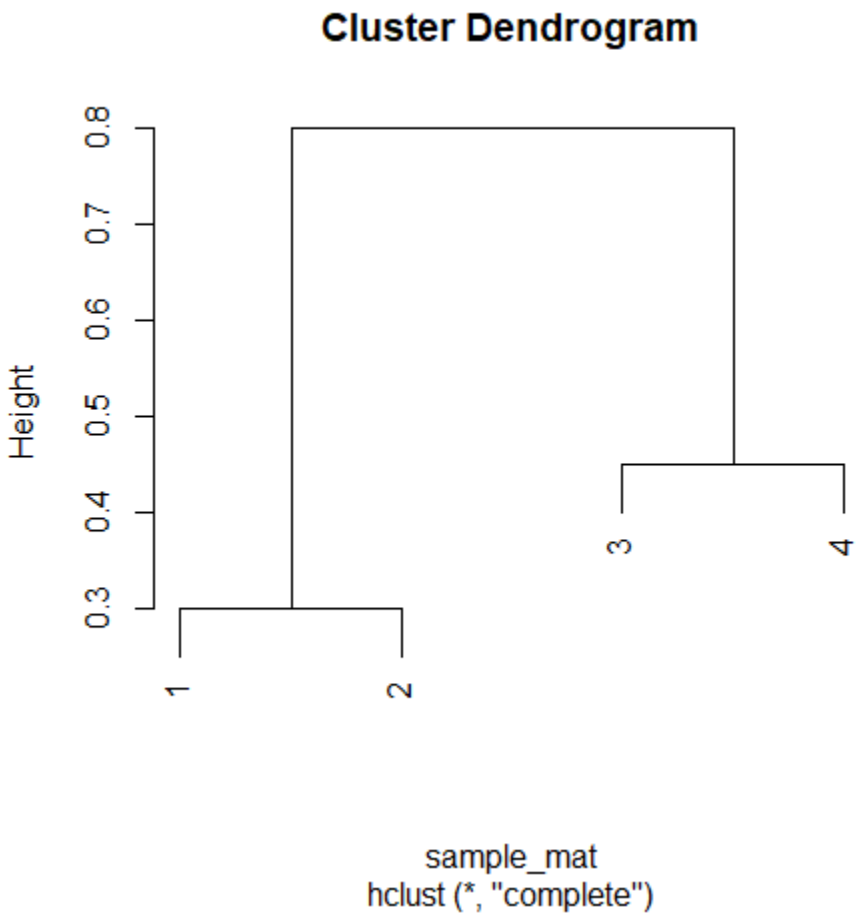
Look for the samples that are most similar i.e having the lowest dissimilarity.
In this case, points A,B has the lowest dissimilarity i.e 0.3.
Thus we merge A,B into a single cluster at height 0.3.
Also,In complete linkage , the dissimilarity between merged pair and others will be maximum of the pair of dissimilarities in each case.
Thus,the new dissimilarity matrix is as follows:

$$\begin{pmatrix} & (A,B) & C & D \\ (A,B) & 0 & 0.5 & 0.8 \\ C & 0.5 & 0 & 0.45 \\ D & 0.8 & 0.45 & 0 \end{pmatrix}$$

Following the same procedure , we merge D,C and the matrix is :

$$\begin{pmatrix} & (A,B) & (C,D) \\ (A,B) & 0 & 0.8 \\ (C,D) & 0.8 & 0 \end{pmatrix}$$

Thus, finally we merge (A,B) and (C,D).
The dendogram for the above Complete Linkage is:



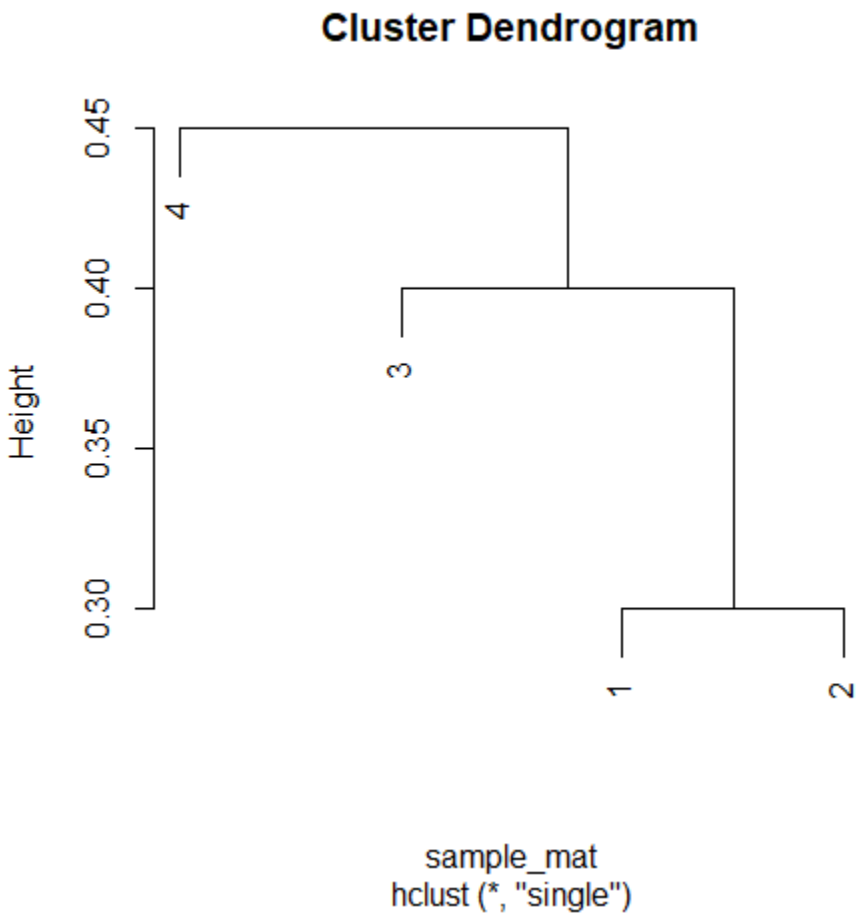
b) Look for the samples that are most similar i.e having the lowest dissimilarity.
In this case, points A,B has the lowest dissimilarity i.e 0.3.
Thus we merge A,B into a single cluster at height 0.3.
In single linkage , the dissimilarity between merged pair and others will be minimum of the pair of dissimilarities in each case.
Thus,the new dissimilarity matrix is as follows:

$$\begin{pmatrix} & (A,B) & C & D \\ (A,B) & 0 & 0.4 & 0.7 \\ C & 0.4 & 0 & 0.45 \\ D & 0.7 & 0.45 & 0 \end{pmatrix}$$

In this case (A,B) and C has the least dissimilarity, so we merge (A,B) and C to a single cluster at height 0.4
The new dissimilarity matrix will be:

$$\begin{pmatrix} & ((A,B),C) & D \\ ((A,B),C) & 0 & 0.45 \\ D & 0.45 & 0 \end{pmatrix}$$

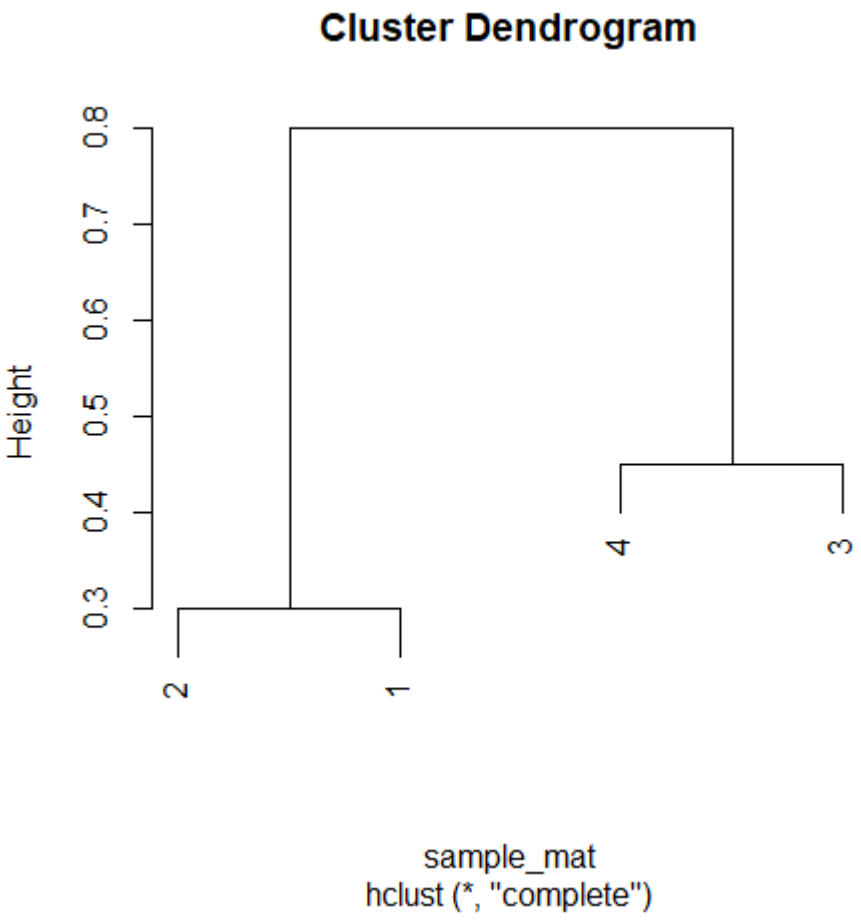
Finally we, merge ((A,B),C) and D at a height 0.45.
The dendogram for the above Single Linkage is:



c)If we cut the dendogram obtained in Complete linkage in such a way that two clusters are obtained, then the two clusters will be (A,B) and (C,D).

d)If we cut the dendogram obtained in Single linkage in such a way that two clusters are obtained, then the two clusters will be (A,B,C) and (D).

e)



Problem 2 [50 points]

Implement expectation-maximization algorithm for Gaussian mixture models (see the EM algorithm below) in R and call this program G_k. As you present your code explain your protocol for

3.1 initializing each Gaussian:- In my Program, in order to initialize each I have taken random observations from the dataset as the mean.

3.2 maintaining k Gaussian:- In order maintain the gausians , I have done that at two stages i.e 1)Expectation 2)Maximization.At the expectation step, I have calculated the posterior probability using dnvnorm. While at maximization step , I have updated my mean matrix, covariance matrix and prior matrix. Covariance matrix is updated using cov.wt.

3.3 deciding ties:- EM performs soft clustering .Hence, There is a very slight chance of having a tie.

3.4 stopping criteria:-The stoppping criteria is determined by the square of the euclidian distance between the previous mean of that gaussian and the present mean of that gaussian.This value keeps on decreasing with every iteration which determines that the algorithm is converging. I have also set a threshold stopping criteria.Upon reaching that threshold value, we deduce that the algorithm has converged.

Problem 3 [70 points]

In this questions, you are asked to run your program, G_k, against the Ringnorm and Ionosphere data sets and compare G_k with C_k (k-means algorithm from previous homework). Click on the below links to download the data sets.

- Ringnorm Data Set
- Ionosphere Data Set

Answer the following questions:

3.1 Initialize G_k and C_k with the same set of initial points (initial centroids for C_k and μ_i-s for G_k are identical) and run them for $k = 2, \ldots, 5$ for 20 runs each. Report error rates and iteration counts for each k using whisker plots that reveal comparison of C_k and G_k. An example of whisker plot is given below. A simple error rate can be calculated as follows:

- If $k = 2$: C_k and G_k will predict two clusters. Error calculation is trivial for two clusters.
- If $k > 2$: after C_k and G_k converge, combine the clusters as follows to ended up with two clusters: since the true clusters are known for a given arbitrary blocks number, final clusters are determined by measuring the Euclidean (this is the easiest choice) distances between true cluster centers and predicted cluster centers.

In other words, you will always calculate the error for $k = 2$ since there are only 2 clusters in the given data sets. Below is an example of error calculation for Ionosphere data set. You can similarly calculate an error rate for Ringnorm data set.

For each centroid C_i, and each Gaussian G_k form two counts (over Ionosphere Data Set) :

$$g_i \leftarrow \sum_{\delta \in c_i.B} [\delta.C = \text{“g”}], \quad \text{good}$$
$$b_i \leftarrow \sum_{\delta \in c_i.B} [\delta.C == \text{“b”}], \quad \text{bad}$$

where $[x = y]$ returns 1 if True, 0 otherwise. For example, $[2 = 3] + [0 = 0] + [34 = 34] = 2$

The centroid C_i and Gaussian G_k is classified as good if $g_i > b_i$ and bad otherwise. We can now calculate a simple error rate. Assume C_i is good. Then the error is:

$$error(C_i) = \frac{b_i}{b_i + g_i} \quad [\text{same for error}(G_i)]$$

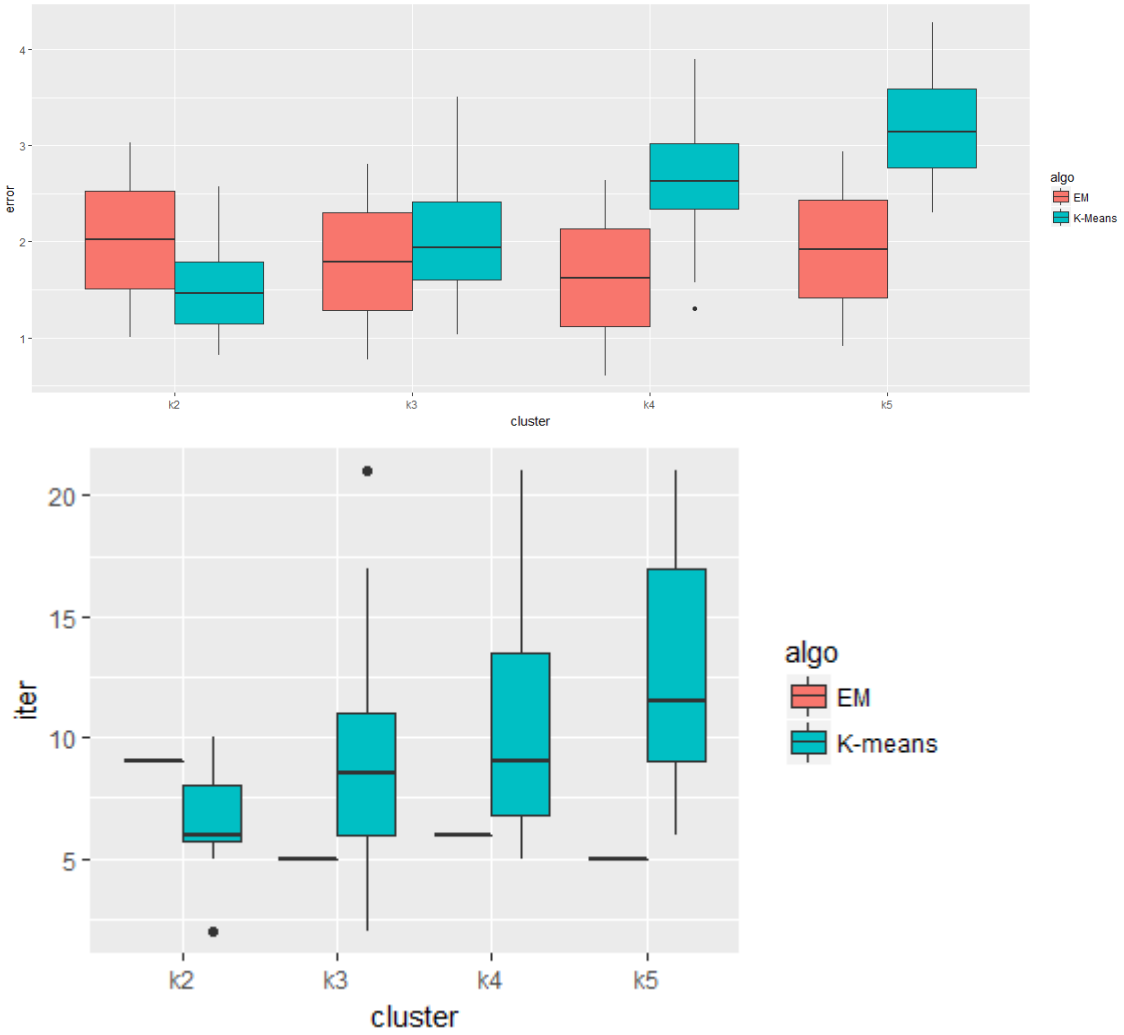
We can find the total error rate easily:

$$Error(\{C_1, C_2\}) = \sum_{i=1}^{2} error(C_i)$$

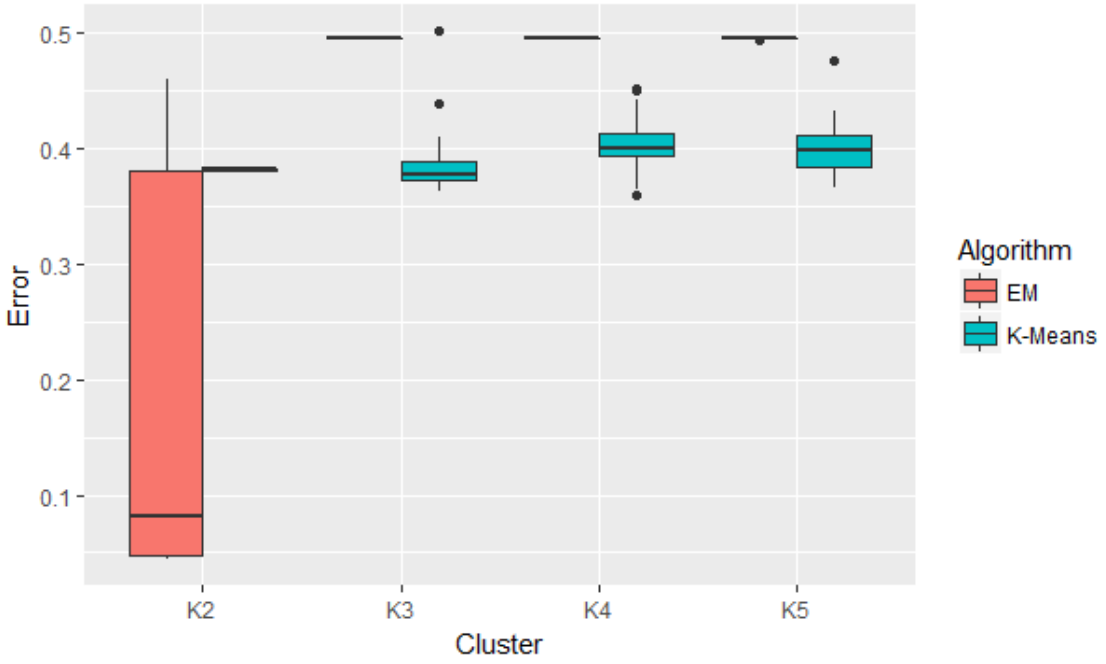
Discuss your results, i.e., which one performs better.

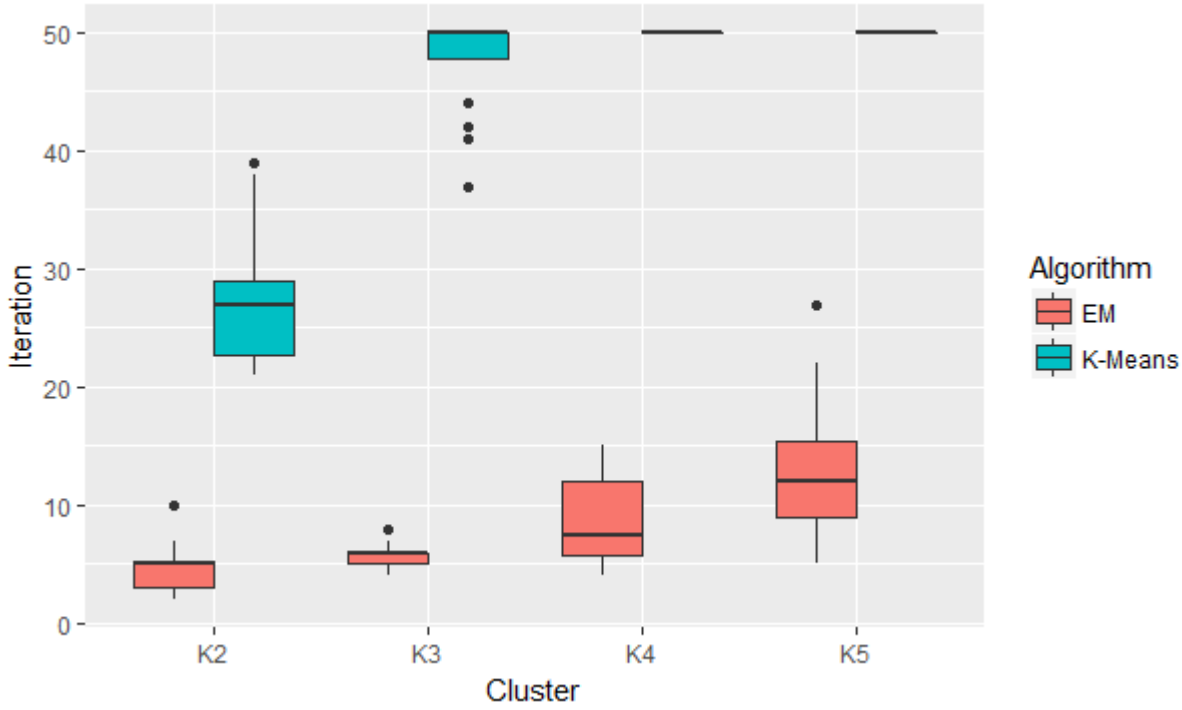
3.2 In this question, we will run your G_k with fixing the variances to ones and the priors to be uniform. Do not update the variances and priors throughout iterations. As explained in question 3.1, compare your new G_k and C_k using whisker plots. Discuss your results, i.e., which one performed better.

The following are the whisker plots for ionosphere dataset when Ck and Gk are run with eack having 20 iterations. From the graph, we can see that EM takes lesser number of iterations to converge than K-means. Also, the error rats are slightly lower than Kmeans.This is shows that the EM is a better clustering algorithm than Kmeans.



The following are the whisker plots for ringnorm dataset when Ck and Gk are run with each having 20 iterations. From the graph, we can see that EM takes lesser number of iterations to converge than K-means. Also, the error rats are slightly lower than Kmeans.This is shows that the EM is a better clustering algorithm than Kmeans.





When we don't update the matrix, then

Problem 4 [50 points]

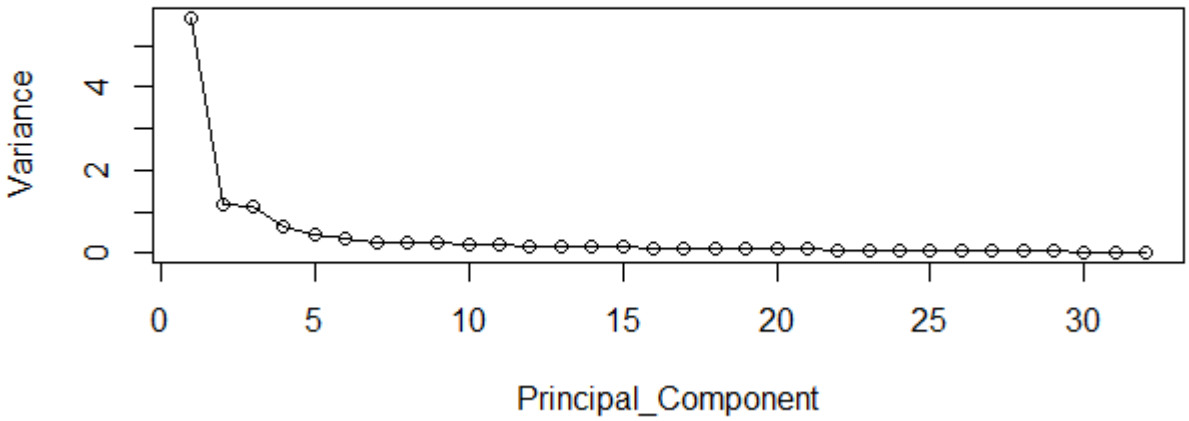
In this question, you will first perform principal component analysis (PCA) over Ionosphere and Rignorm data sets and then cluster the reduced data sets using G_k (from question 3.1) and C_k. You are allowed to use R packages for PCA. Ignore the class variables (35th and 1st variables for Ionosphere and Ringnorm data sets, respectively) while performing PCA. Answer the questions below:

4.1 Make a scatter plot of PC1 and PC2 for both data sets. Discuss principal components (The first and second principal components). What are PC1 and PC2?

bh.png

Principal component analysis (PCA) is used for analyzing variance when you are dealing with multivariate data. PC1 and PC2 are first and second principle components of the dataset that provides us with maximum variance.The variance reduces with every PCA.

4.2 Create scree plots after PCA and explain the plots. The plot shows that the variance decreases with with every Principal Component.But after a point the variance hardly decreases and becomes constant.Thus, we deduce that PCA maximize the variance of projection along each component.

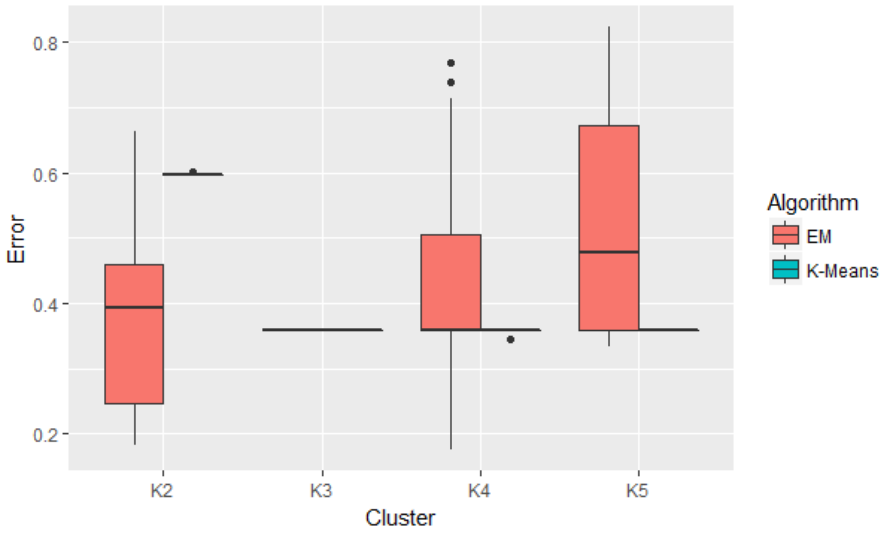


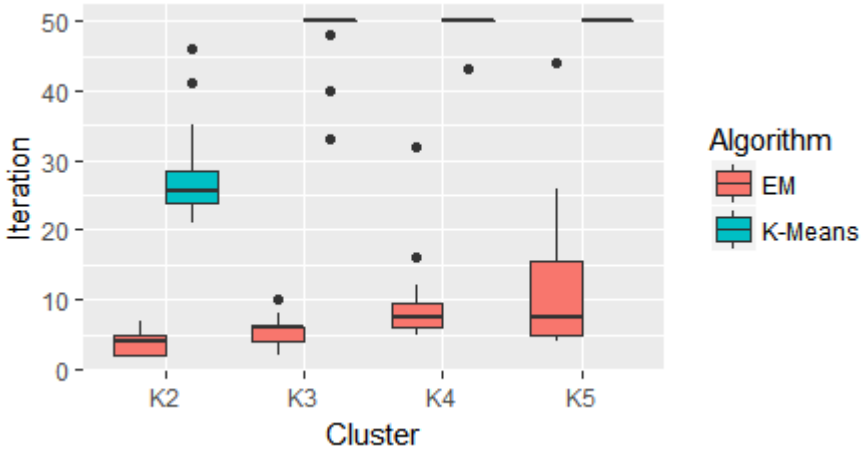
4.3 Observe the loadings using prcomp() or princomp() functions in R and discuss loadings in PCA?i.e., how are principal components and original variables related?

The loadings that we obtain in prcomp() or princomp are nothing but the eigen vectors that are used to obtain Principal Components of a Dataset. The Prncipal Components are just the lower dimensional reprensentation of a higher dimension original variable.

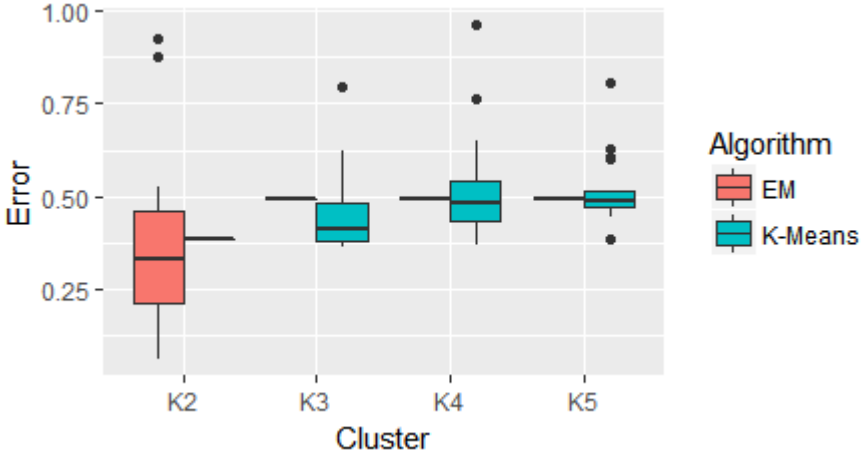
4.4 Keep 90% of variance after PCA and reduce Ionosphere and Rignorm data sets. Run C_k and G_k with the reduced data sets and compare them using whisker plots as shown in question 3.1

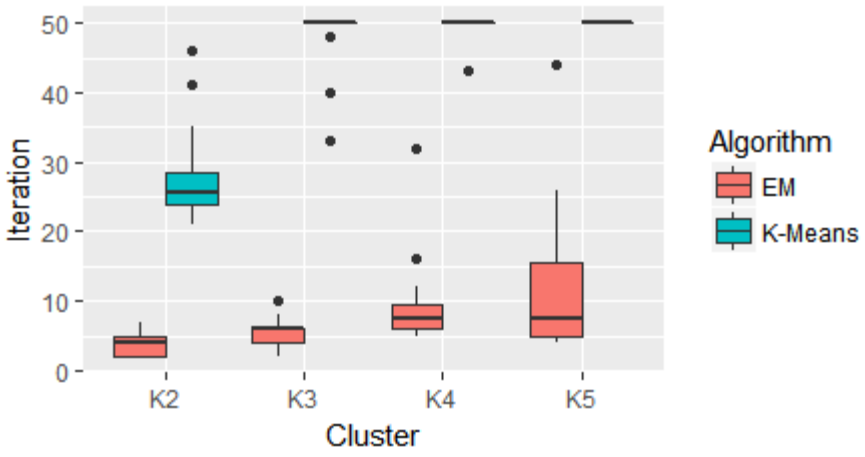
The following are the whisker plot for ionosphere dataset





The following are the whisker plot for ringnorm dataset





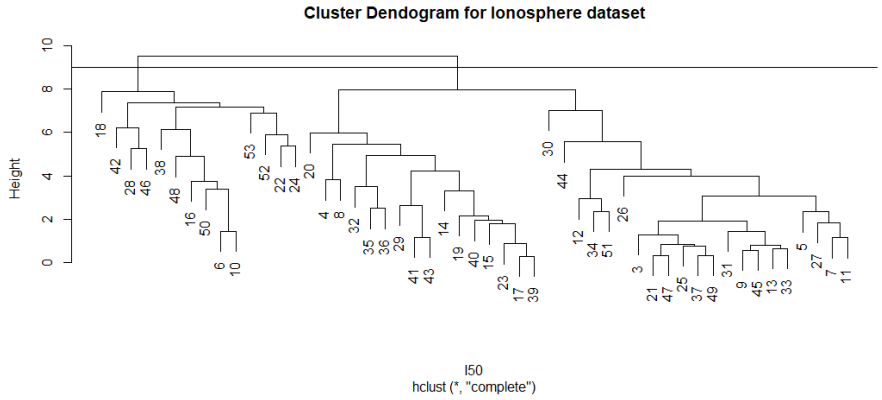
4.5 Discuss that how PCA affects the performance of C_k and G_k.

Problem 5 [50 points]

Randomly choose 50 points from Ionosphere data set (call this data set I_{50}) and perform hierarchical clustering. You are allowed to use R packages for this question. (Ignore the class variable while performing hierarchical clustering.)

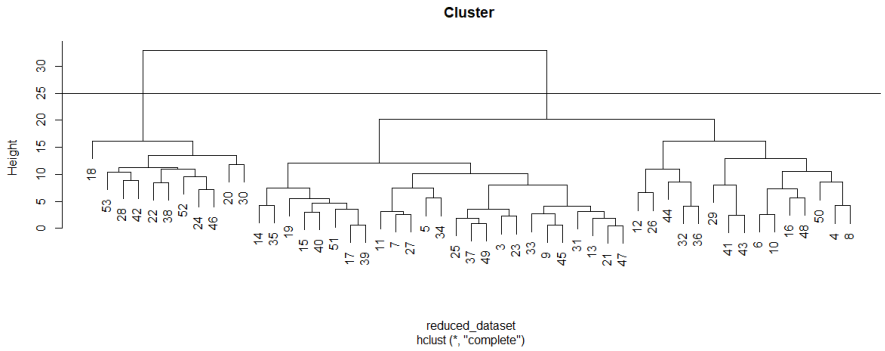
5.1 Using hierarchical clustering with complete linkage and Euclidean distance cluster I_{50}. Plot the dendrogram.

5.2 Cut the dendrogram at a height that results in two distinct clusters. Calculate an error rate.



5.3 First, perform PCA on I_{50} (Keep 90% of variance). Then hierarchically cluster the reduced data using complete linkage and Euclidean distance. Plot the dendrogram

5.4 Cut the dendrogram at a height that results in two distinct clusters. Calculate an error rate. How did



PCA affect hierarchical clustering?