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

Shradha Baranwal
sbaranwa@iu.edu

October 3, 2017

Problem 1 [20 points]

a) Complete linkage hierarchical clustering
Dissimilarity Matrix :

$$M = \begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

Dissimilarity after first fusion :

$$M = \begin{bmatrix} & 0.5 & 0.8 \\ 0.5 & & 0.45 \\ 0.8 & 0.45 & \end{bmatrix}$$

Dissimilarity after second fusion :

$$M = \begin{bmatrix} & 0.8 \\ 0.8 & \end{bmatrix}$$

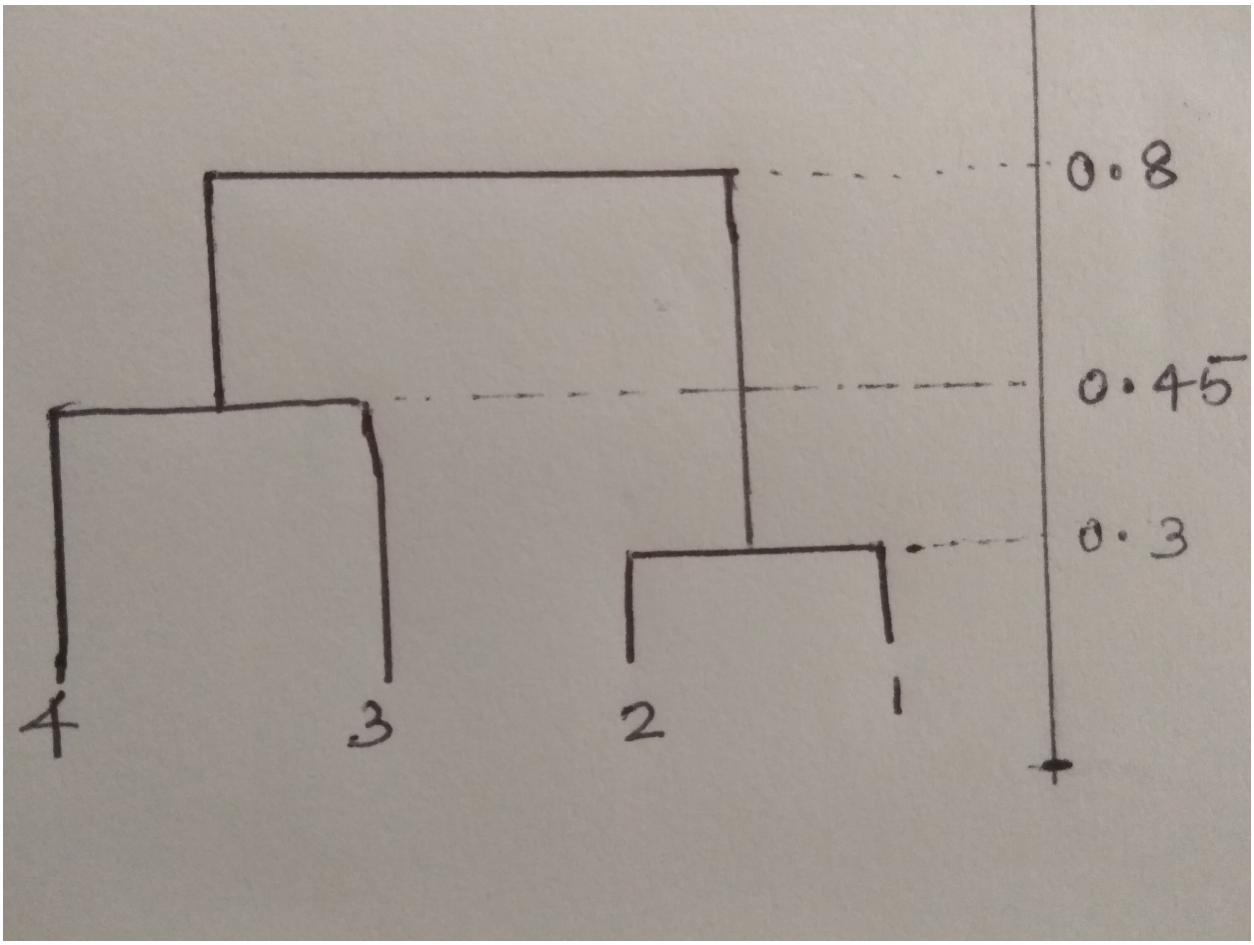

Figure 1: Complete Linkage

b) Single linkage hierarchical clustering
Dissimilarity Matrix :

$$M = \begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

Dissimilarity after first fusion :

$$M = \begin{bmatrix} & 0.4 & 0.7 \\ 0.4 & & 0.45 \\ 0.7 & 0.45 & \end{bmatrix}$$

Dissimilarity after second fusion :

$$M = \begin{bmatrix} & 0.45 \\ 0.45 & \end{bmatrix}$$

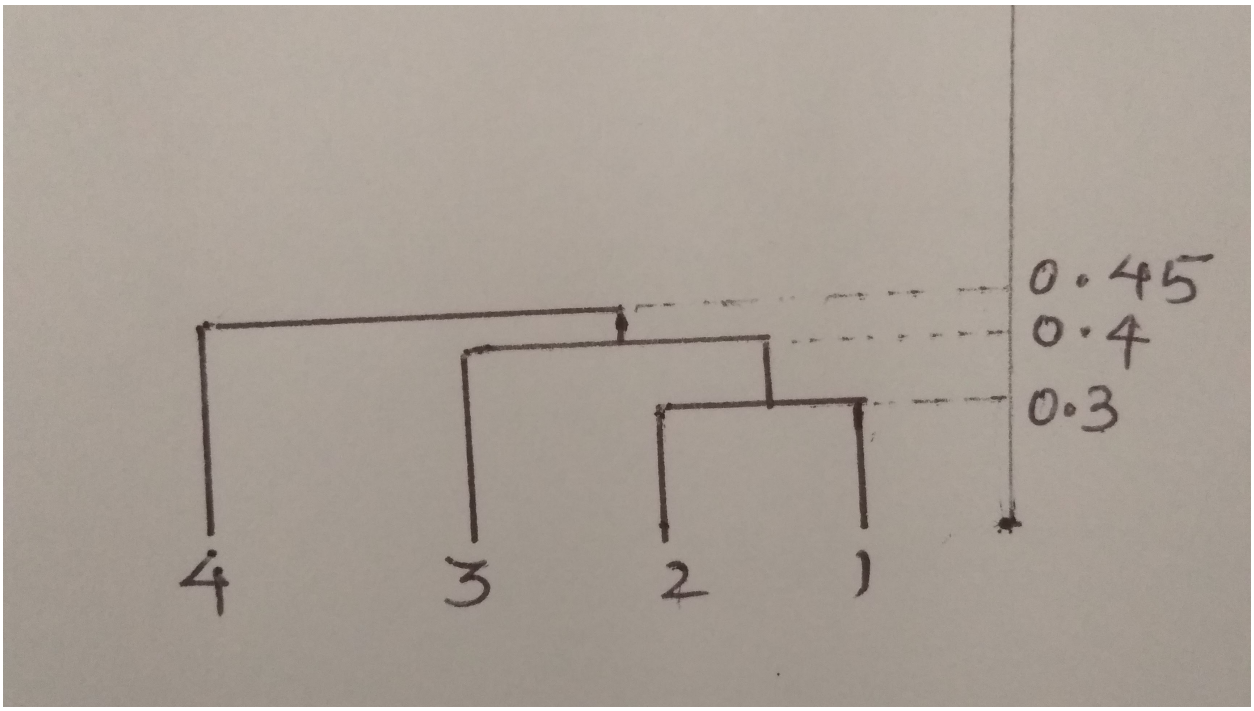

Figure 2: Single Linkage

c) Cutting the dendogram(a) to form two clusters
In the above answer (a), on cutting the dendogram to form two clusters, the observations in each cluster are : (4,3), (1,2)

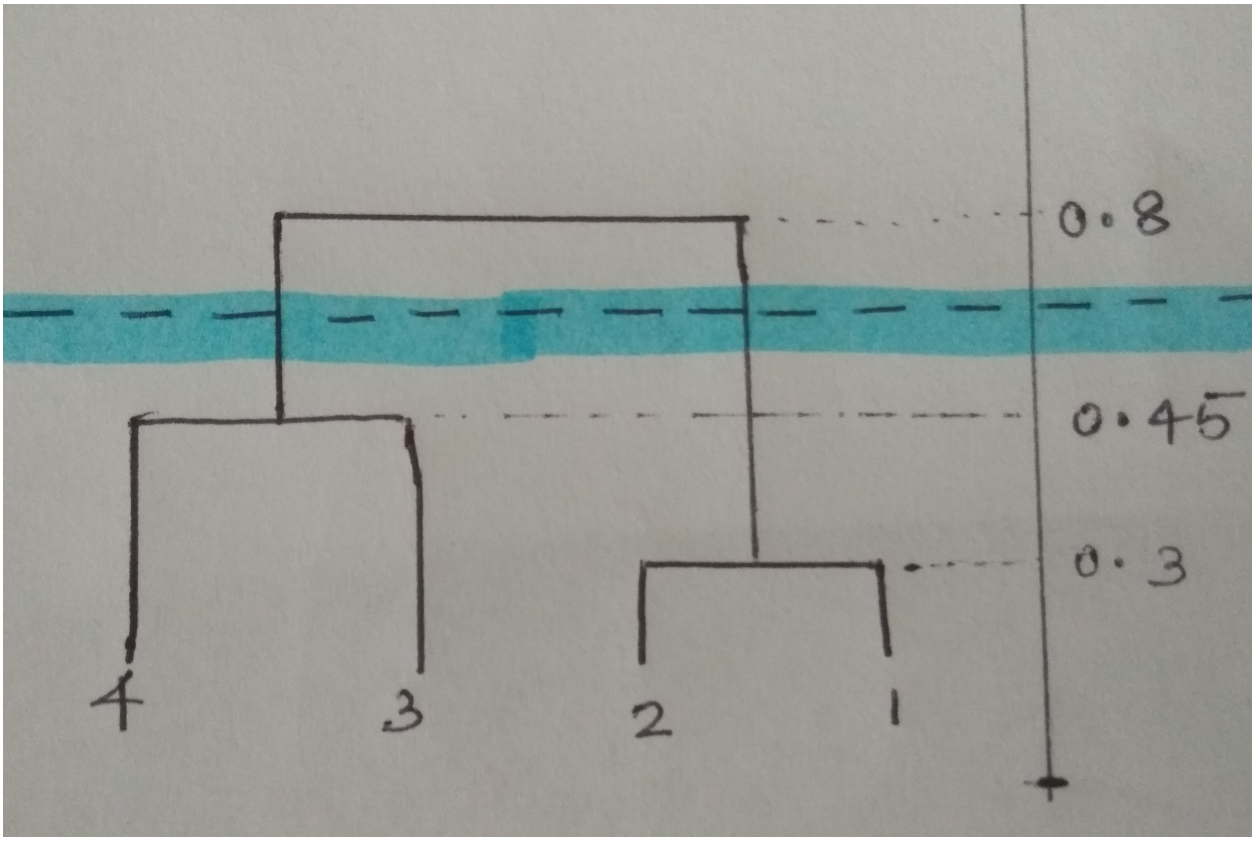

Figure 3: Cluster Complete Linkage

d) Cutting the dendogram(b) to form two clusters
In the above answer (b), on cutting the dendogram to form two clusters, the observations in each cluster are : (4), (1,2,3)

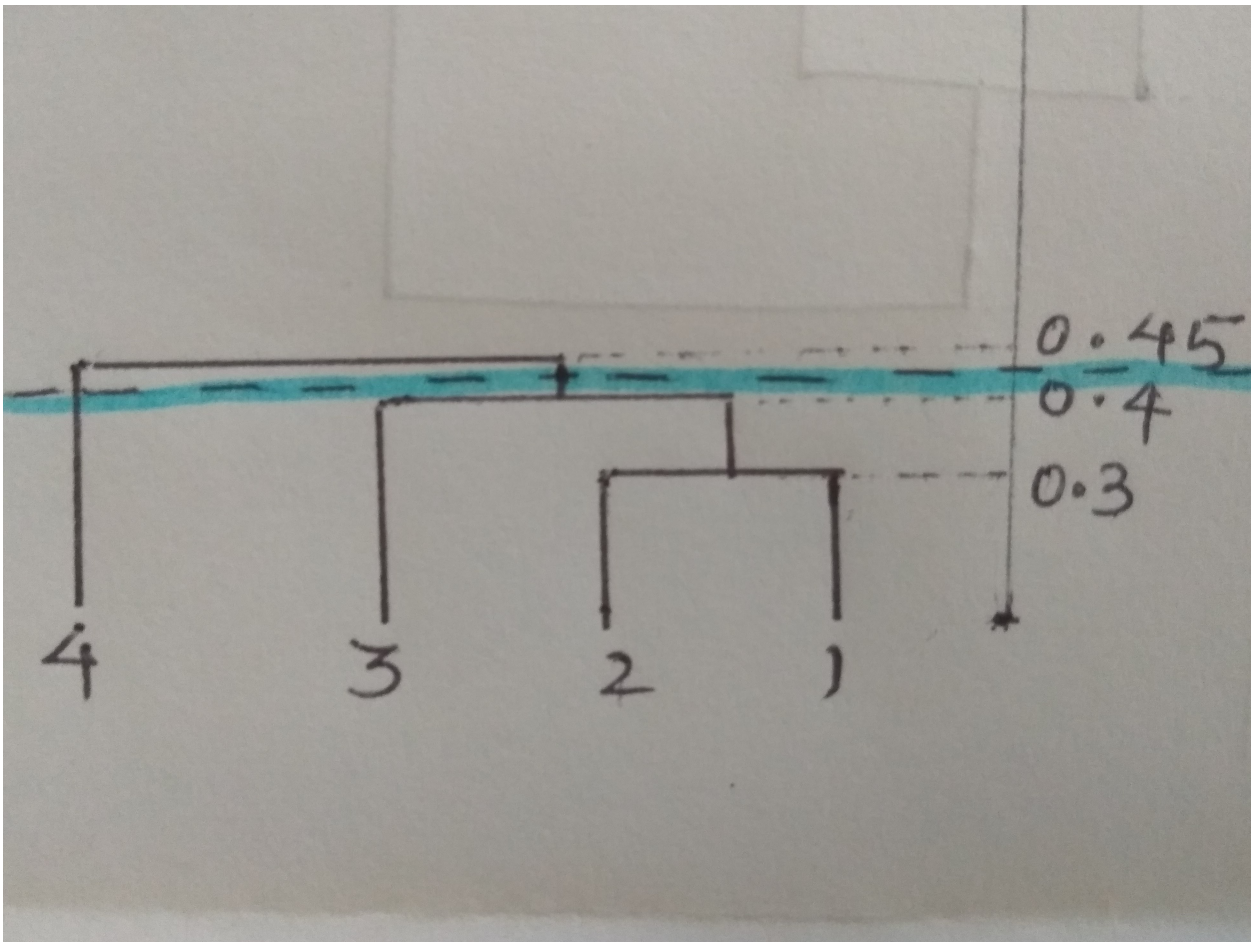

Figure 4: Cluster single Linkage

e) Reposition of leaves

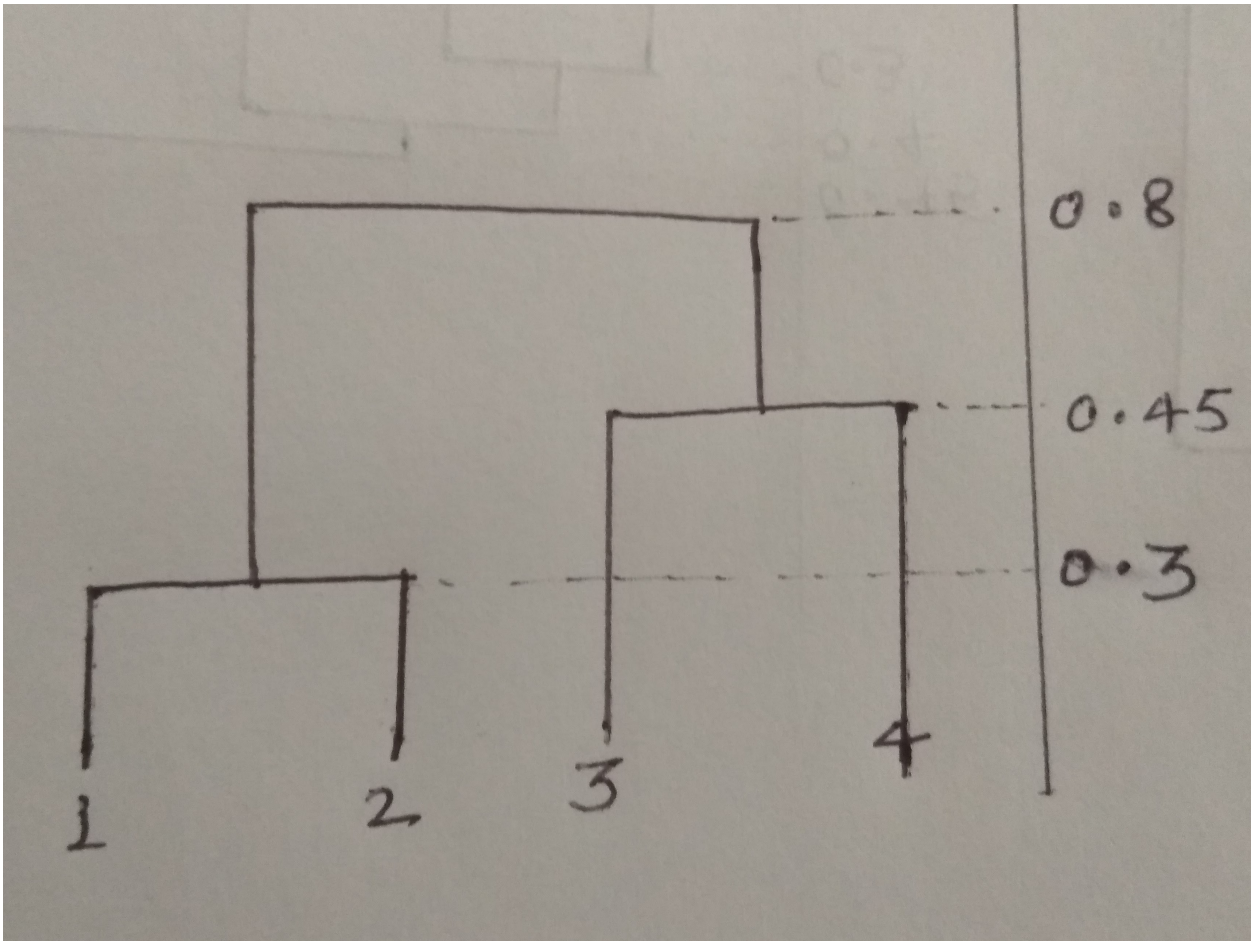

Figure 5: Reposition of leaves

Since the information of cluster dissimilarity lies on the vertical axis, repositioning of leaves in fusion doesn't change the meaning of the dendogram.

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

2.1 Initializing each Gaussian
The three variables of a probability distribution are mean, covariance matrix and priors.
Mean is initialized to random points from data set.
Covariance matrix is initialized to identity matrix which is independent variables. Priors are initialized to uniform which is 1/k.

2.2 maintaining k Gaussian
The gaussians are updated by calculating the probabilities of each data point on each gaussian distribution.
Using the normalized probabilities as weights, the mean, covariances and priors are recalculated and the cycle continues.
The final gaussian distribution is decided when the sum of squared distances between means is lesser than a threshold value.

2.3 deciding ties
In the EM implementation, I have used which.max() function to find the gaussian distribution in which the probability of a particular data point is maximum. This function returns the first index which contains the maximum value. This is basically selecting a cluster at random as the order of index for each cluster is random.

2.4 stopping criteria
The threshold value for the sum of squared differences of the mean is passed as an argument to the function which is used as stopping criteria. In case of singular covariance matrix for jth cluster during the convergence, the mean for jth cluster is randomly selected and covariance is set back to identity matrix.

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). Answer the following questions:

3.1 Comparison between G_k with C_k for ionosphere and ringnorm data sets.

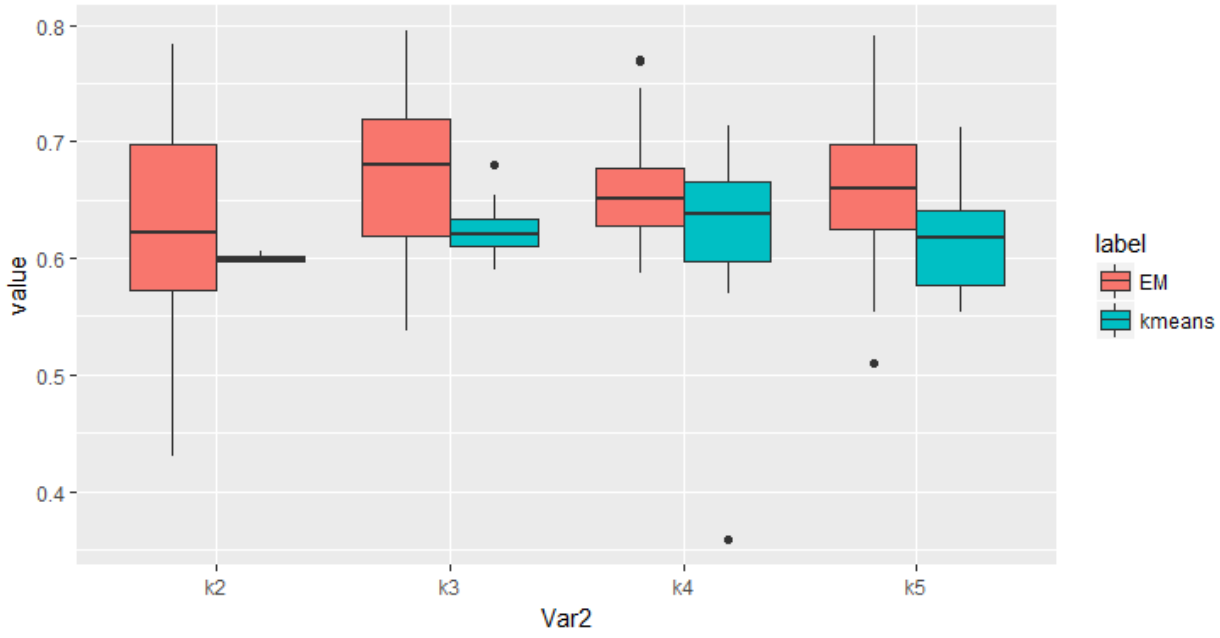

Figure 6: Error comparison between EM and K means algorithm for Ionosphere dataset

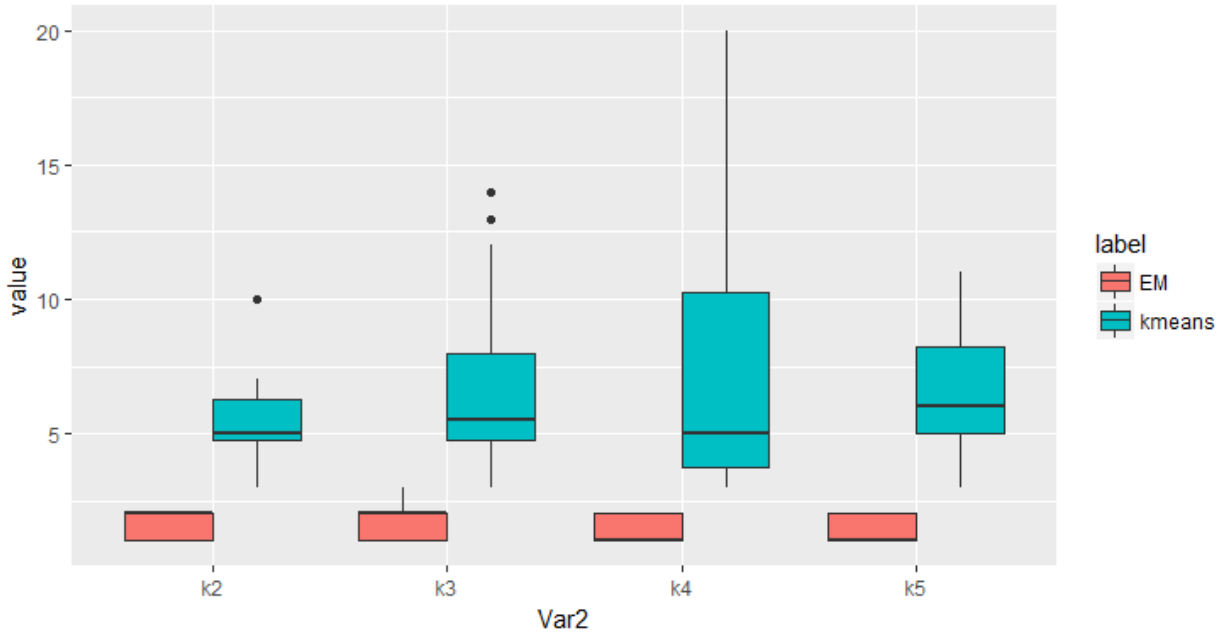

Figure 7: Comparison of number of iterations for K means and EM algorithms Ionosphere dataset

Looking at the error comparison graph between G_k with C_k for Ionosphere dataset, the mean of the error calculated is more for EM algorithm as compared to k means algorithm.
The number of iterations, however, is lesser for G_k as compared to C_k for the same threshold as well as same set of initial centroids.

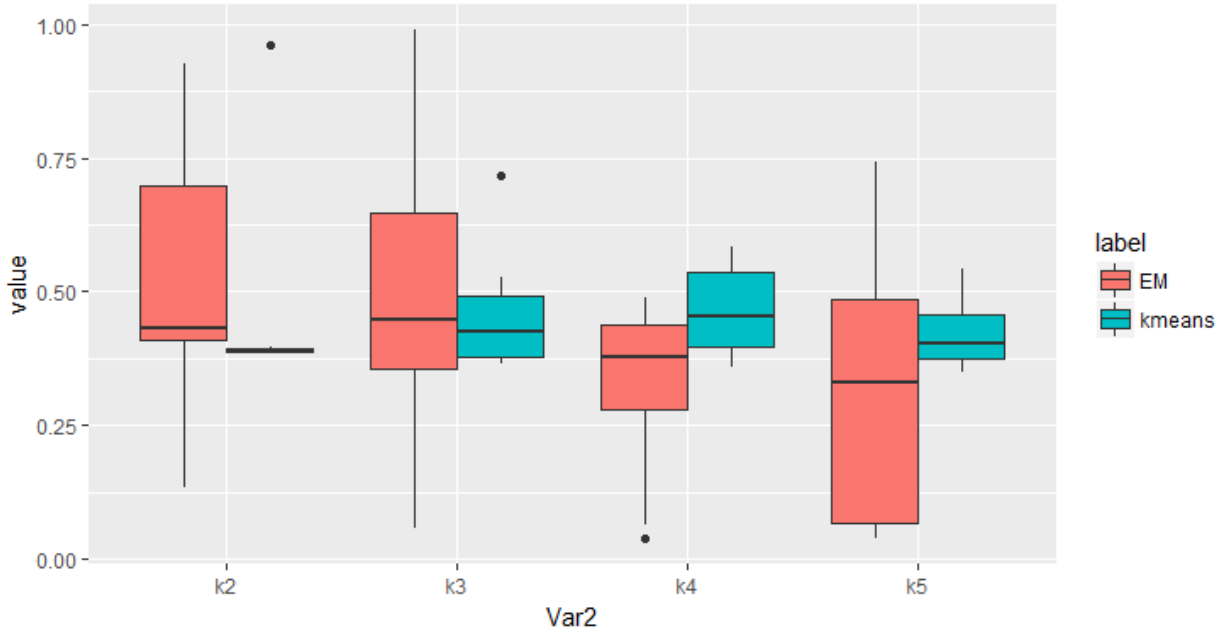

Figure 8: Error comparison between EM and K means algorithm for Ringnorm dataset

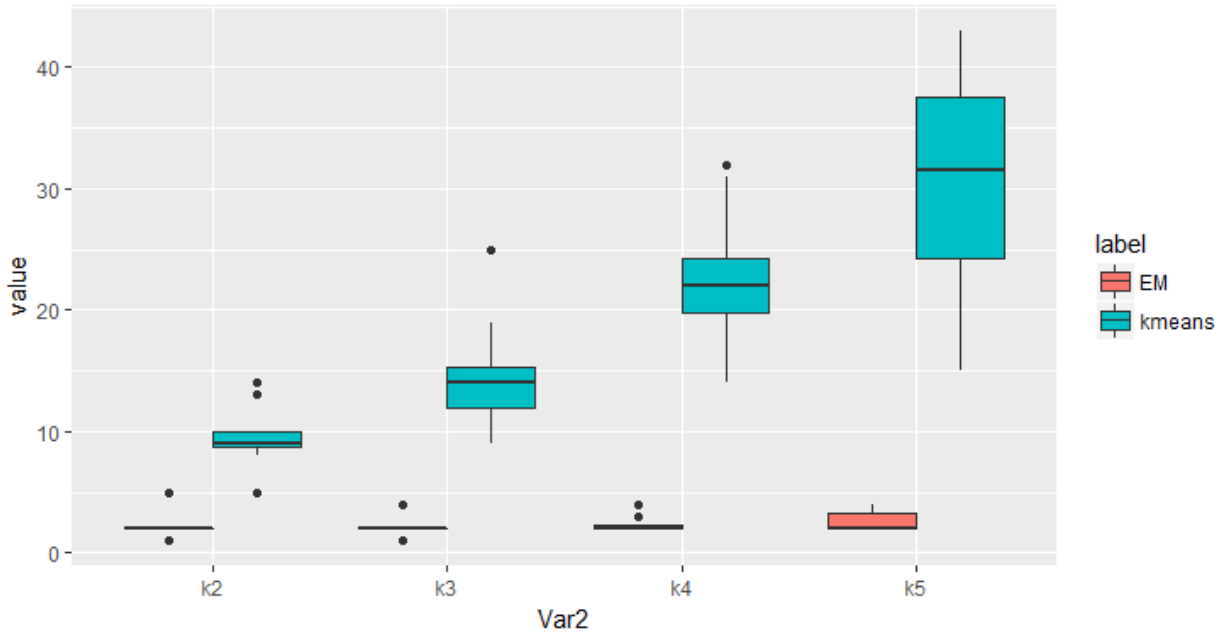

Figure 9: Comparison of number of iterations for K means and EM algorithms Ringnorm dataset

The graphs for ringnorm dataset comparing both the algorithms is similiar to the graphs generated for ionosphere data with error rate averaging higher as compared to kmeans average error.
The number of iterations for EM in case of ringnorm data is lower as compared to kmeans.

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.

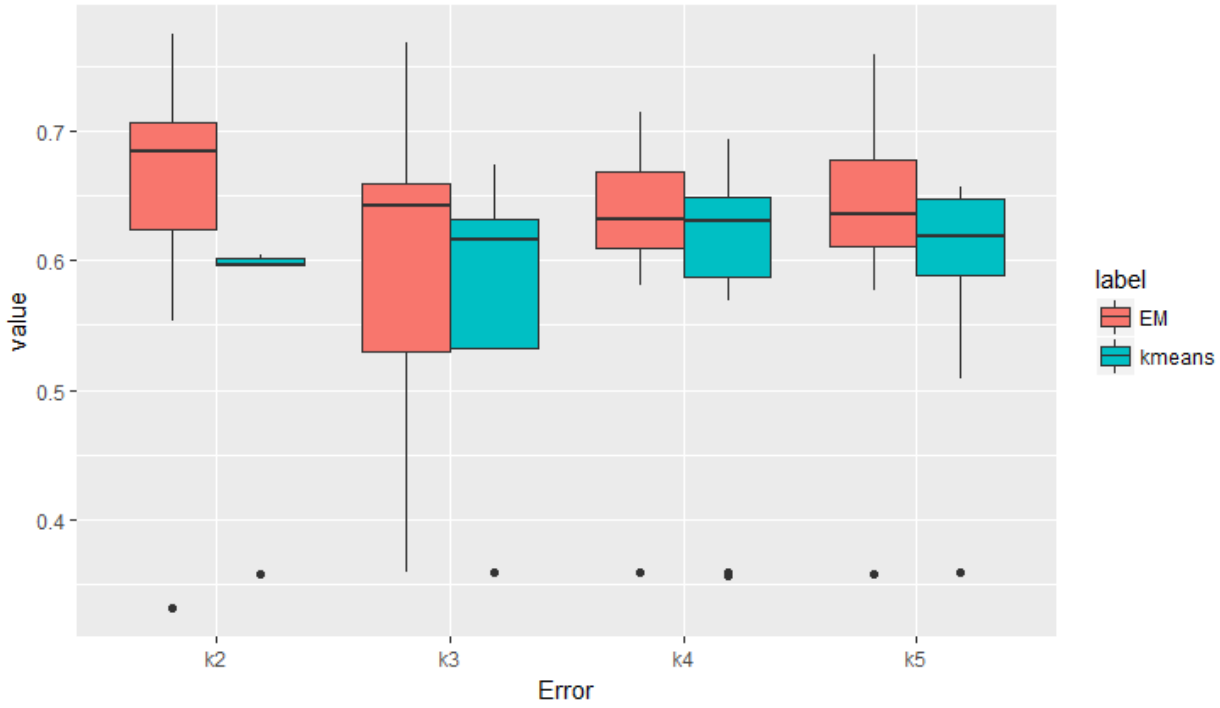

Figure 10: Error comparison between EM and K means algorithm for Ionosphere dataset

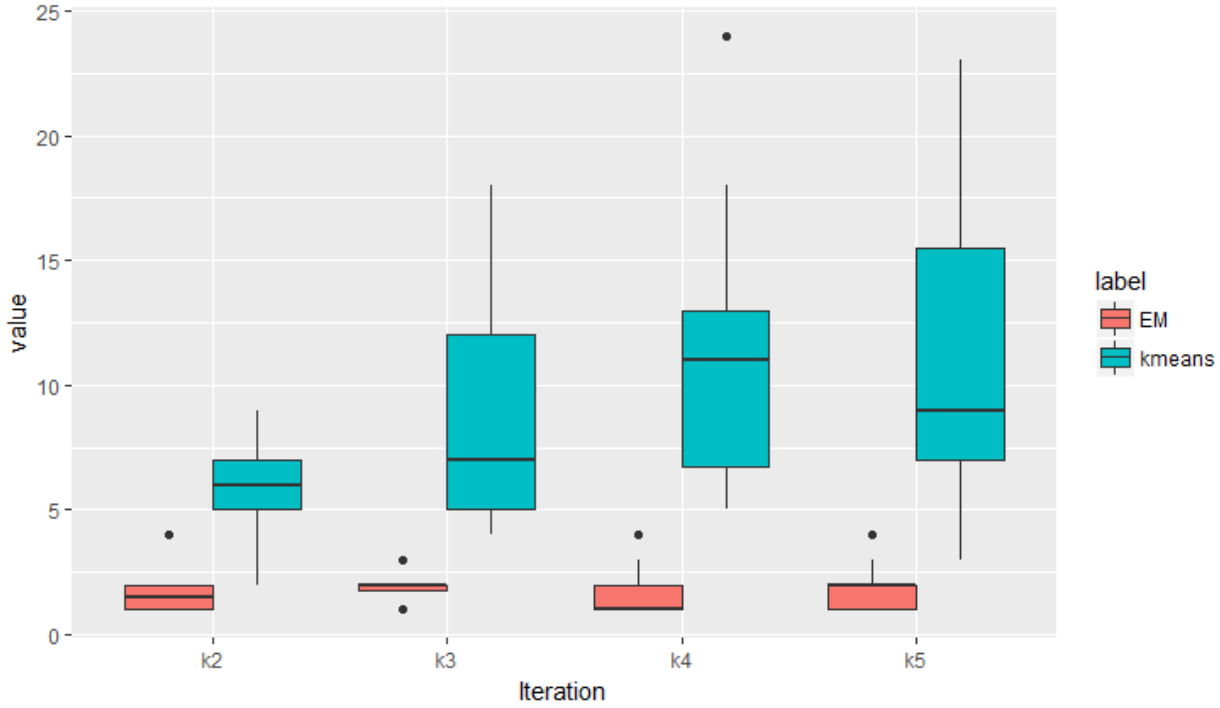

Figure 11: Comparison of number of iterations for K means and EM algorithms

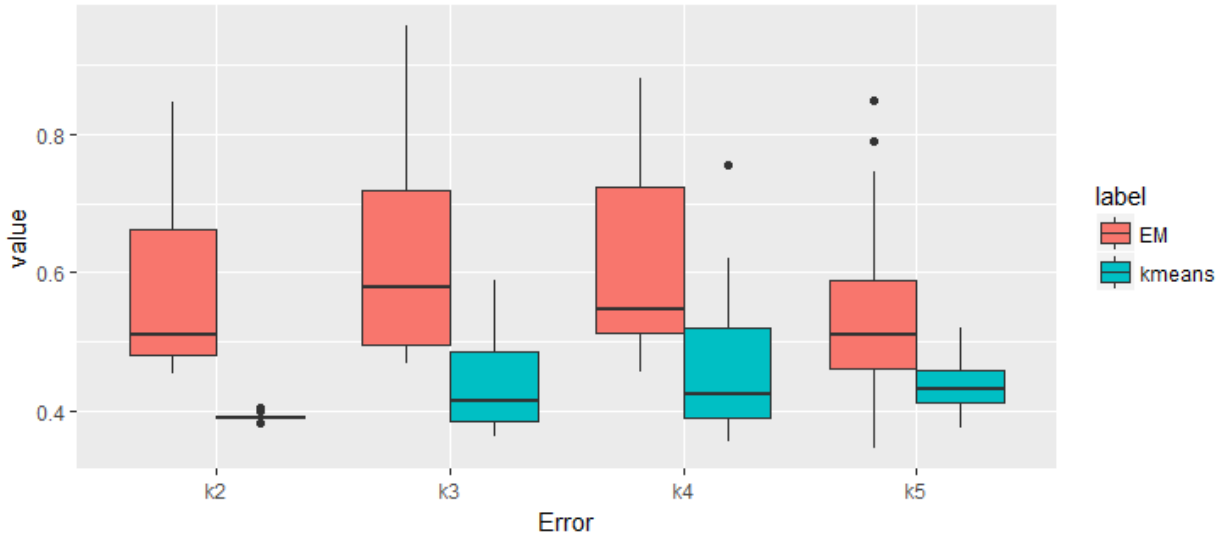

Figure 12: Error comparison between EM and K means algorithm for Ringnorm dataset

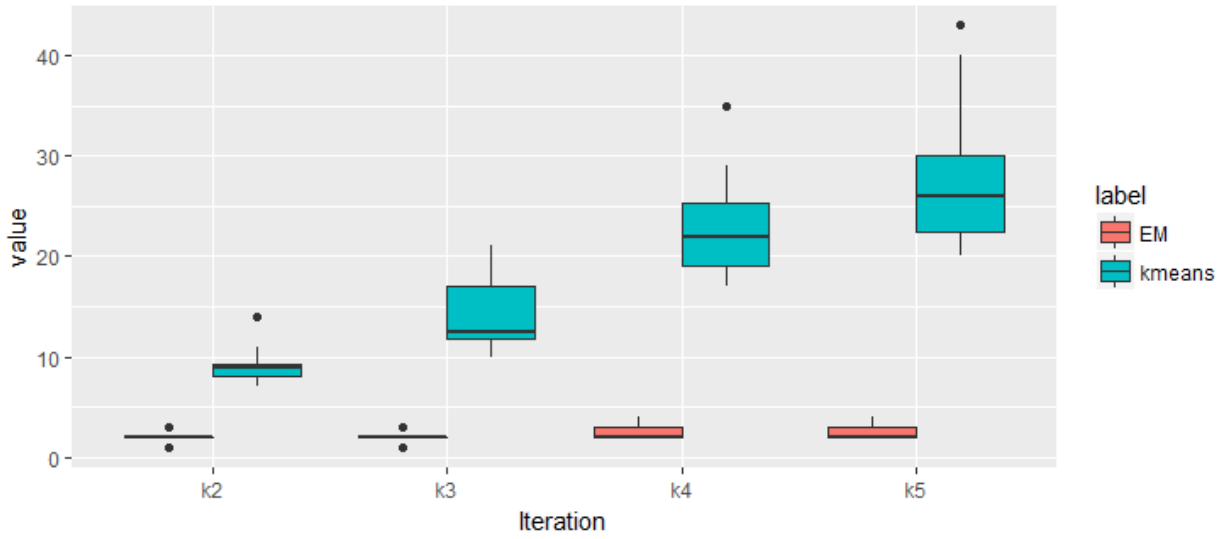

Figure 13: Comparison of number of iterations for K means and EM algorithms

After keeping constant the co variances and mean of gaussian distributions in EM algorithm, there was not much difference in the error and number of iterations graphs for Ionosphere dataset. However, in case of ringnorm dataset, the mean error for ringnorm dataset increased for EM increased as compared to the last time. However, the number of iterations remained lower than the k means.

Problem 4 [50 points]

In this question, you will first perform principal component analysis (PCA) over Ionosphere and Ringnorm 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?

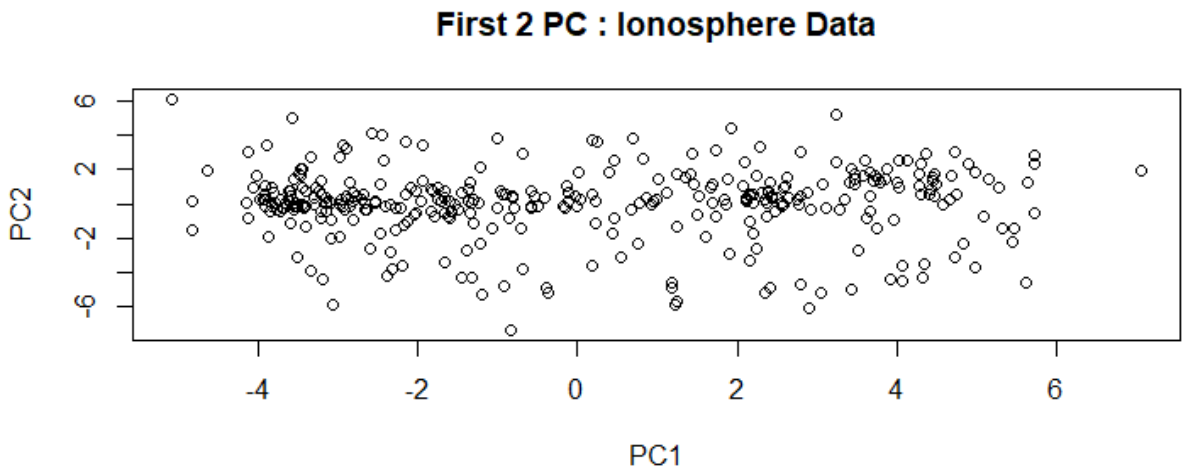

Figure 14: Scatter plot of first 2 principal components for ionosphere data

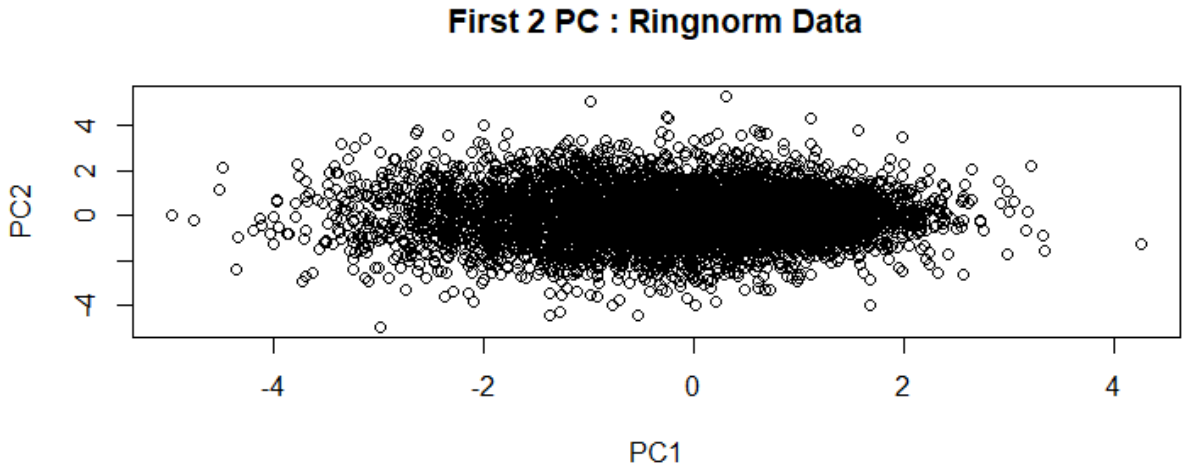

Figure 15: Scatter plot of first 2 principal components for ringnorm data

The first two principal component vectors of any data set gives the plane from which the data points are the closest. Hence, the variance maintained by these 2 principal component vectors is the maximum of all the principal components. The first two principal component score vectors are the projections of each data point on the plane principal component vectors 1 and 2.

4.2 Create scree plots after PCA and explain the plots.

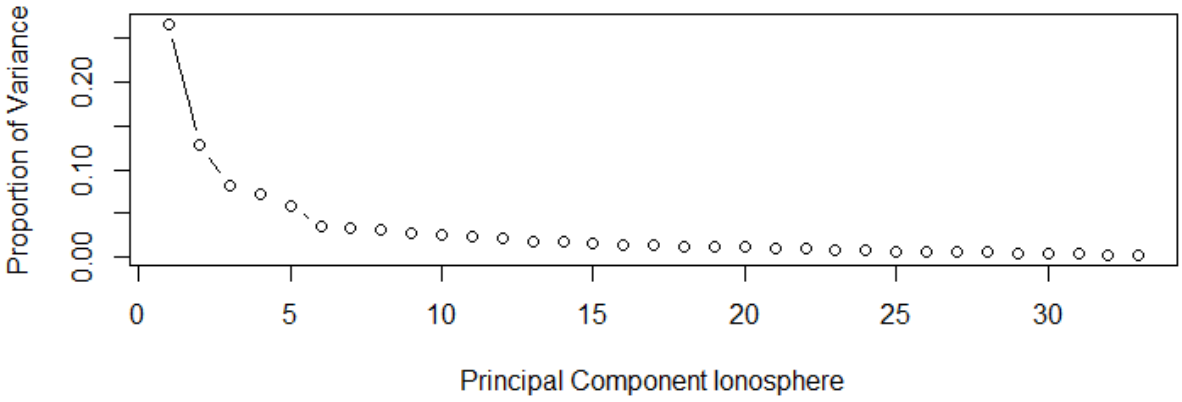

Figure 16: Scree plot Ionosphere Data

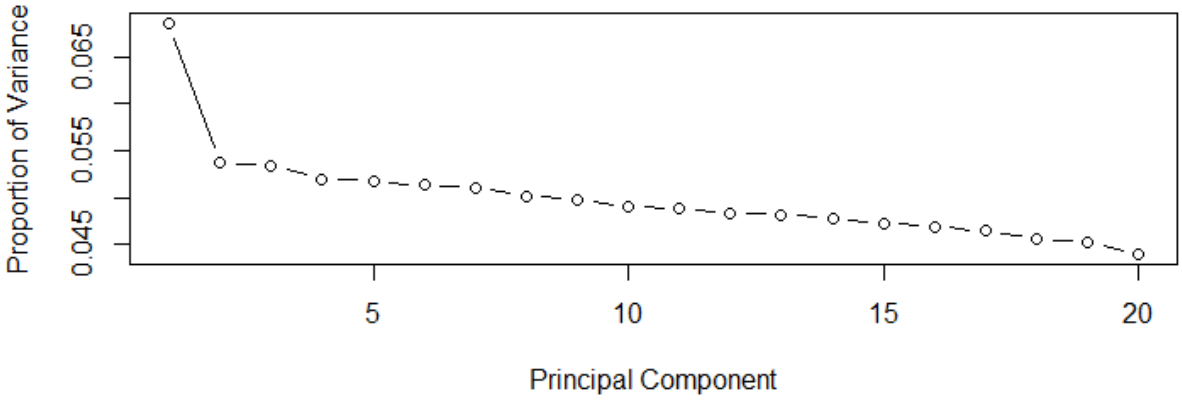

Figure 17: Scree plot Ringnorm Data

The above two figures gives the percentage of variance of each pc for both ionosphere and ringnorm data set. The variances in ionosphere dataset is a monotonically decreasing graph wherein the variance help by the last few pc are tending to zero.
In case of ringnorm data, however, the percentage of variance is pretty small in the first pc and decreases almost linearly.

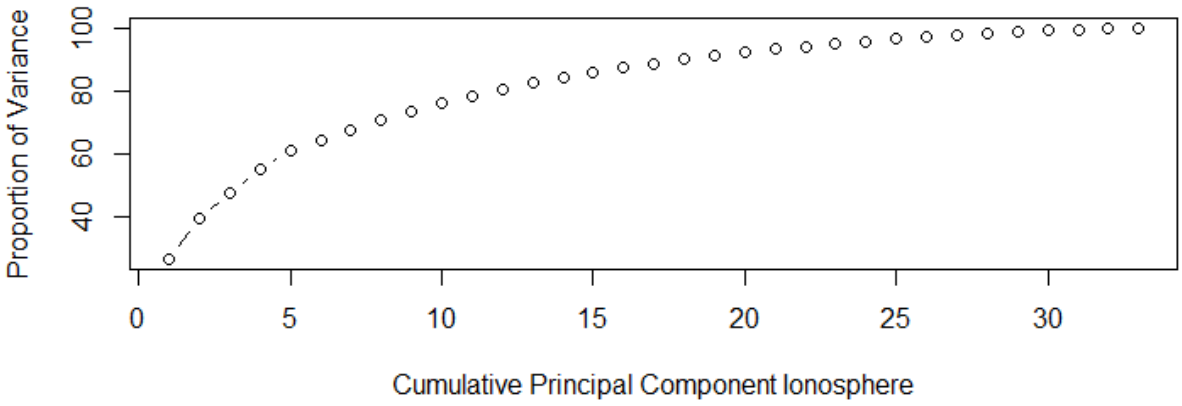

Figure 18: Scree plot Ionosphere Data

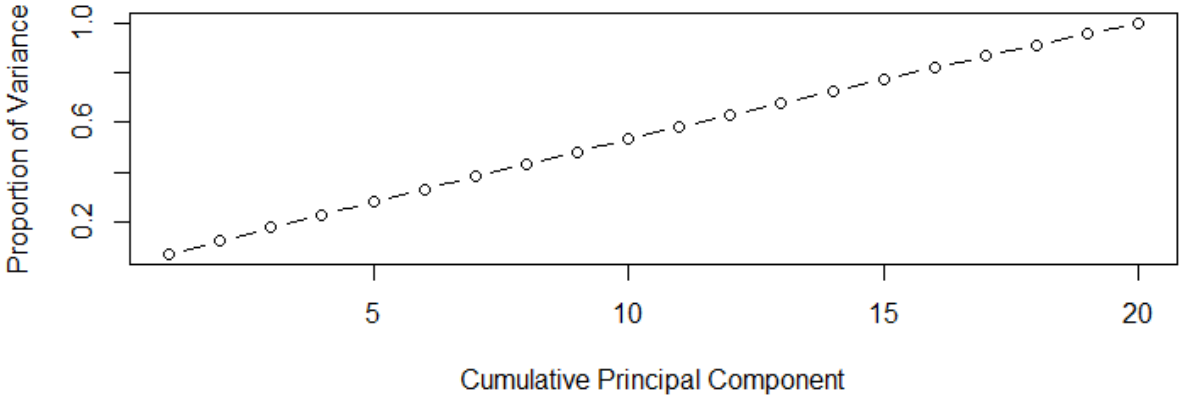

Figure 19: Scree plot Ringnorm Data

The above figures show the cumulative variance of each principal component for both ionosphere and ringnorm data sets. The variance of each principal component in ionosphere dataset is decreasing. However, in case of ringnorm the variance of each principal component is almost same. Hence, to maintain 90% of the variance in both the data sets, 12 PCs can be dropped reducing the number of variables to 19. However, in case of ringnorm, to maintain the same percentage of variance, only few PCs could be excluded.

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 principal loadings vectors are the directions in the feature space with the maximum variance of data. There can be a maximum of d (number of dimensions and $d < n$) principal components. They

are the planes which are closest to the data using Euclidean distance metric.
The loadings are vectors which are used to calculate the projections of data in a that direction. The projections are the principal component score vectors. Each principal component gives the variance of the data maintained in that component in the decreasing order.
Thus, you can say that the principal components score vectors are the projections of original variables in the direction of principal components vectors.

4.4 Keep 90% of variance after PCA and reduce Ionosphere and Ringnorm 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

Below are the comparison graphs of error and iteration for ringnorm and ionosphere datasets after reducing the variables by keeping 90% variance of data :

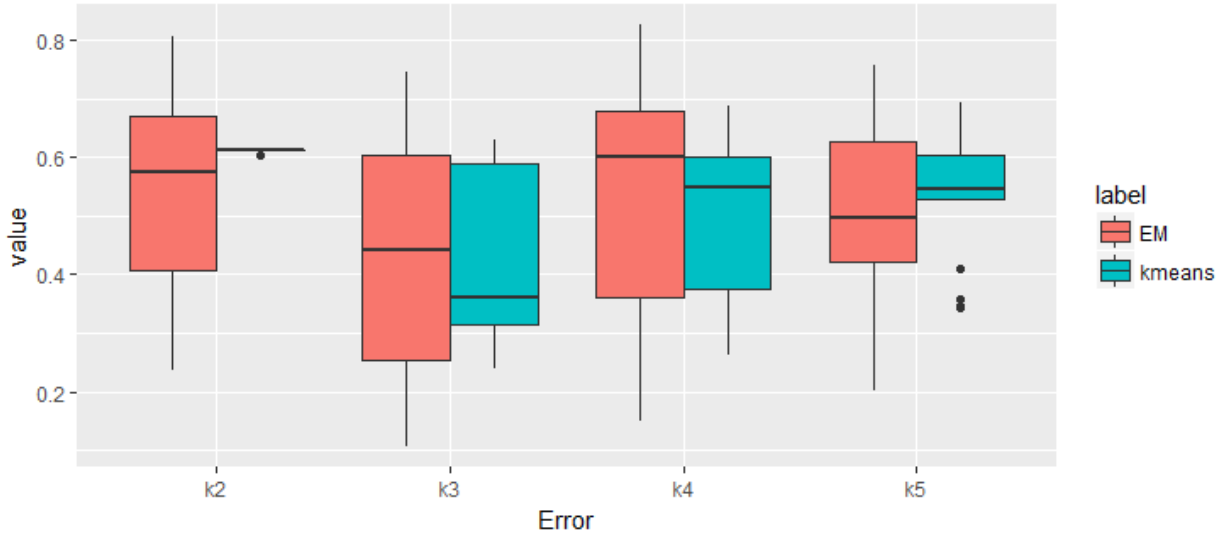

Figure 20: Error comparison between EM and K means algorithm for Ionosphere dataset

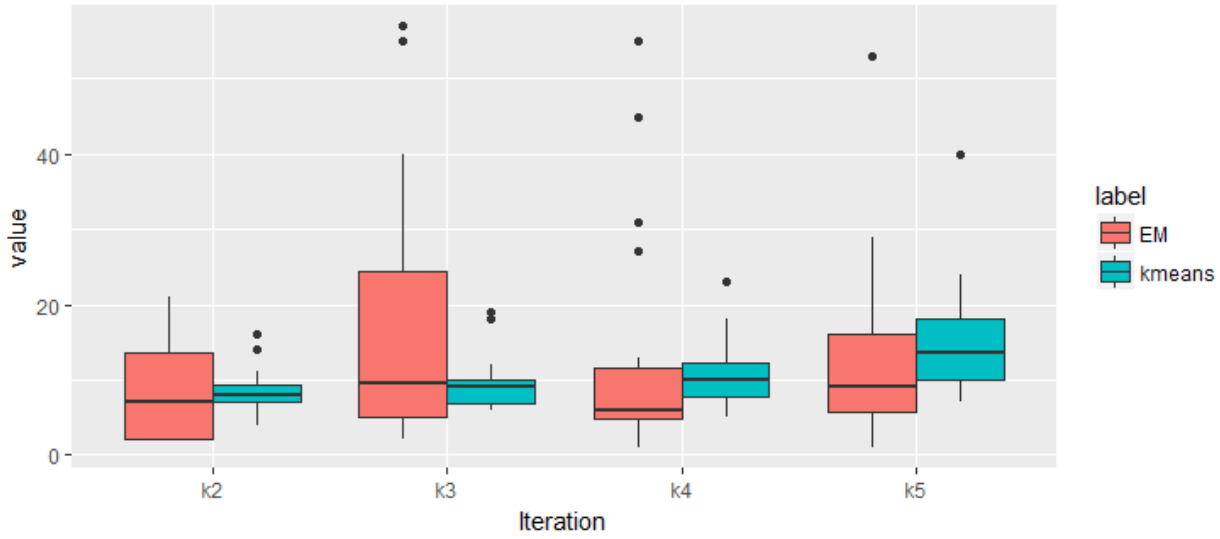

Figure 21: Comparison of number of iterations for K means and EM algorithms

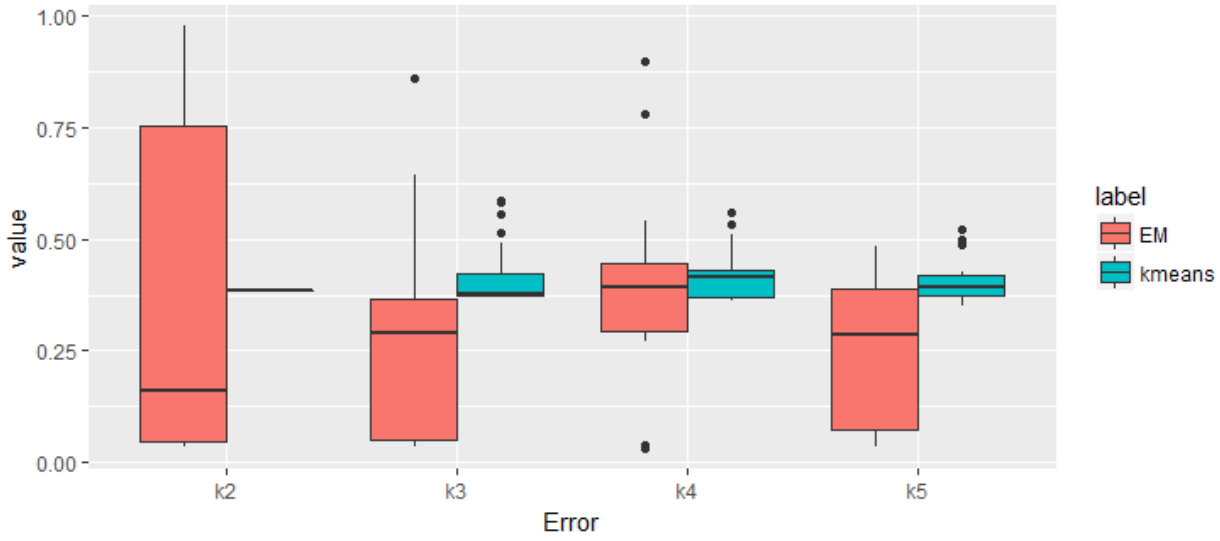

Figure 22: Error comparison between EM and K means algorithm for Ringnorm dataset

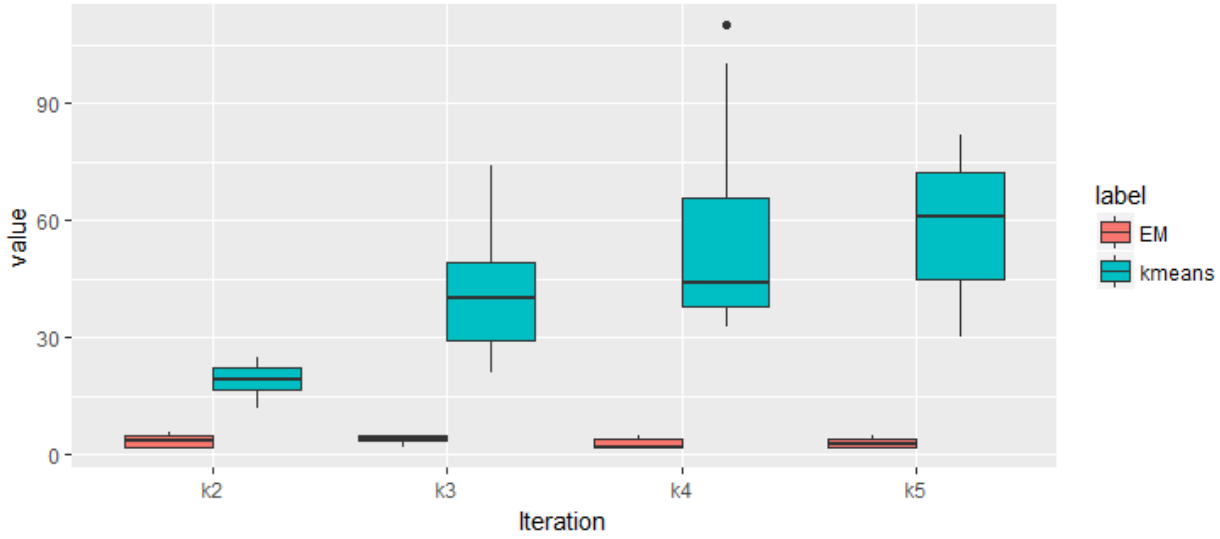

Figure 23: Comparison of number of iterations for K means and EM algorithms

In case of ionosphere, the number of iterations has reduced significantly. This might be due to the reduction of number of variables significantly. However, the graph for number of iterations for ringnorm has remained the same as without pca. This is because the number of variables has not reduced much in case of ringnorm data.

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.

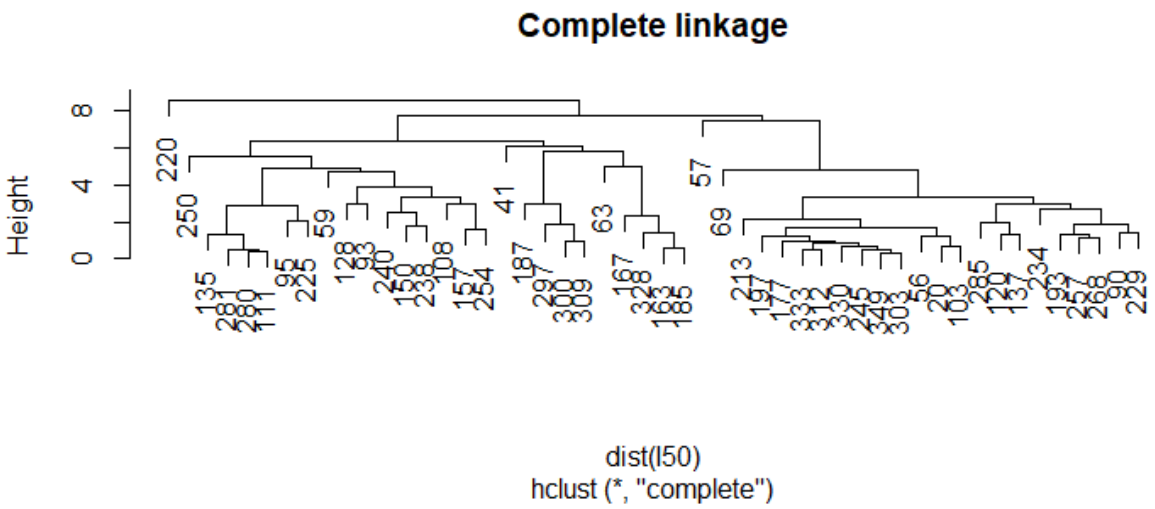

Figure 24: Complete linkage for I_{50}

5.2 Cut the dendrogram at a height that results in two distinct clusters. Calculate an error rate.
Cutting the dendogram for 2 clusters, resulted in a cluster with one element and another cluster with 49 elements. The error rate for these cluster is : 0.3061224

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.

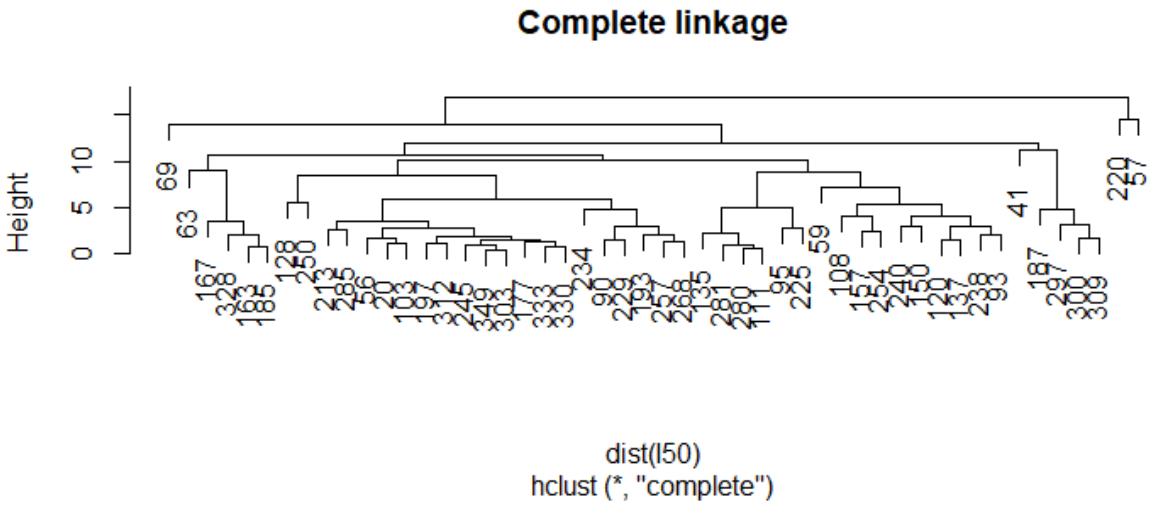

Figure 25: Complete linkage 90% variance of I_{50}

5.4 Cut the dendrogram at a height that results in two distinct clusters. Calculate an error rate. How did PCA affect hierarchical clustering? Cutting the dendogram for 2 clusters, resulted in a cluster with two element and another cluster with 48 elements. The error rate for these cluster is : 0.2916667 PCA resulted in the reduction of error rate after hierarchical clustering.

Extra credit [60 points]

This part is optional.

1 Improve the EM algorithm through initialization. k-means ++ is an extended k-means clustering algorithm and induces non-uniform distributions over the data that serve as the initial centroids. Read the paper and implement this idea to improve your G_k program (from question 3.1). Run your new G_k and old one (question 3.1) for $k = 2, \ldots, 5$ and compare the results using whisker plots. [30 points]

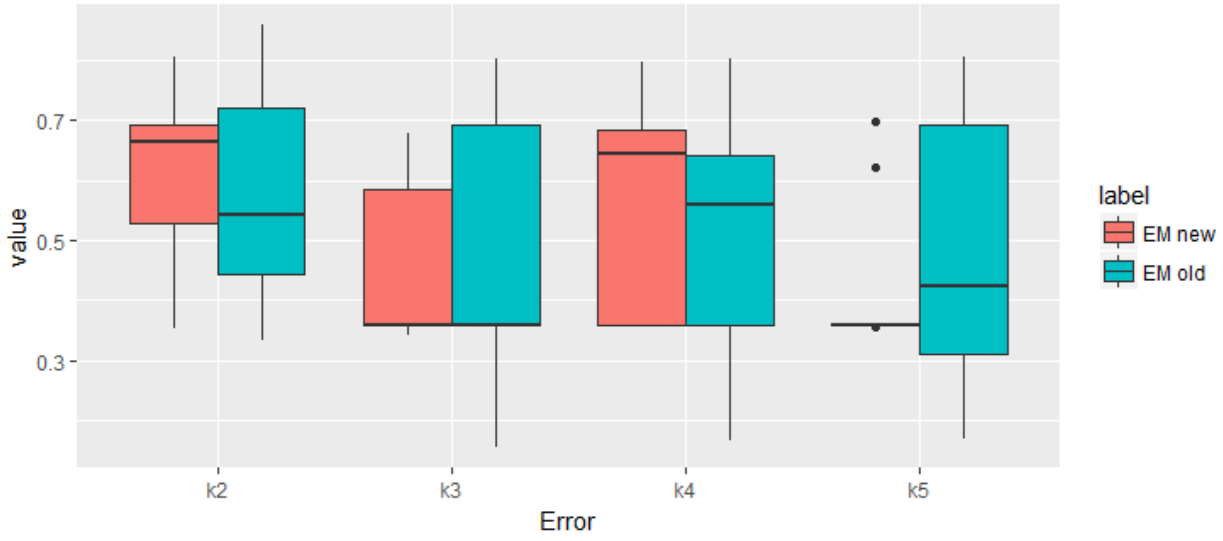

Figure 26: Error comparison between EM and K means algorithm for Ionosphere dataset

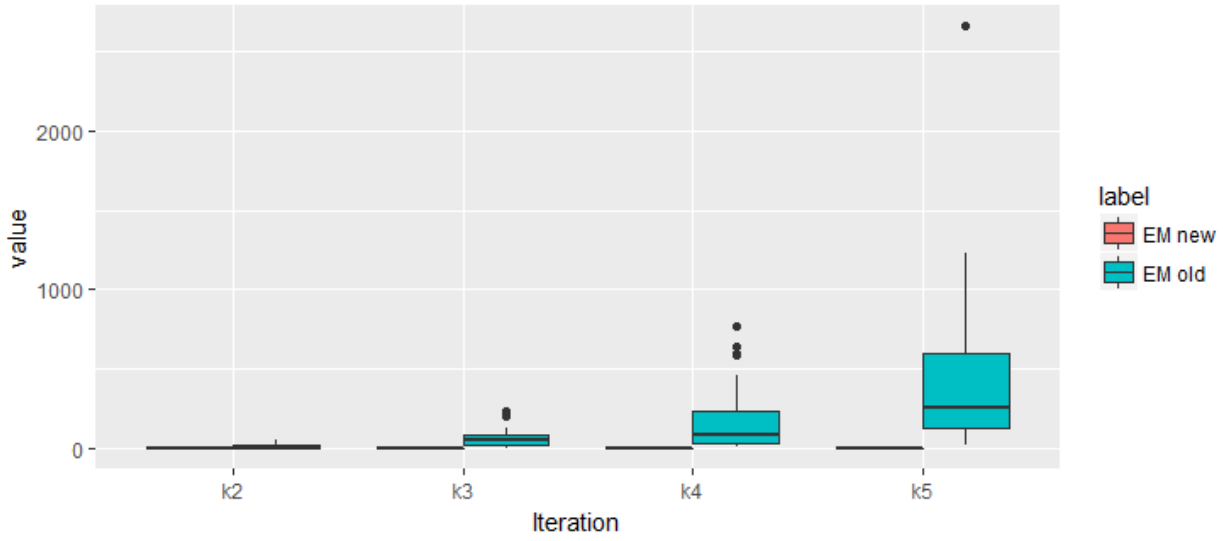

Figure 27: Comparison of number of iterations for K means and EM algorithms

In the above whisker plots for error and iteration comparison between EM and EM++, the mean error difference between both the algorithms is not much. However, the number of iterations for EM++ is less. Also, the number of iterations has almost remained constant as the number of clusters increased unlike for EM in which the number of iterations increased with number of clusters.

2 Run the EM algorithm for different mixture models, i.e., Poisson, and against different data sets. [30 points]