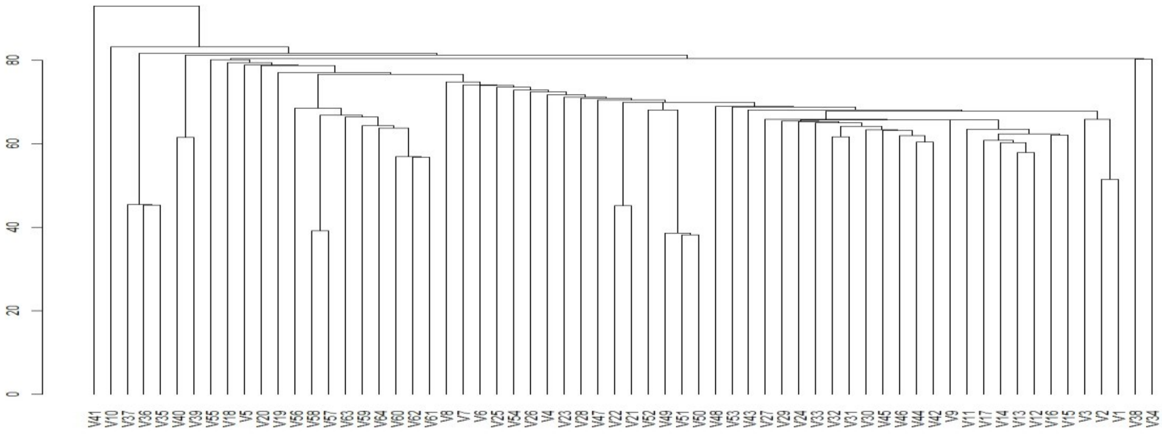
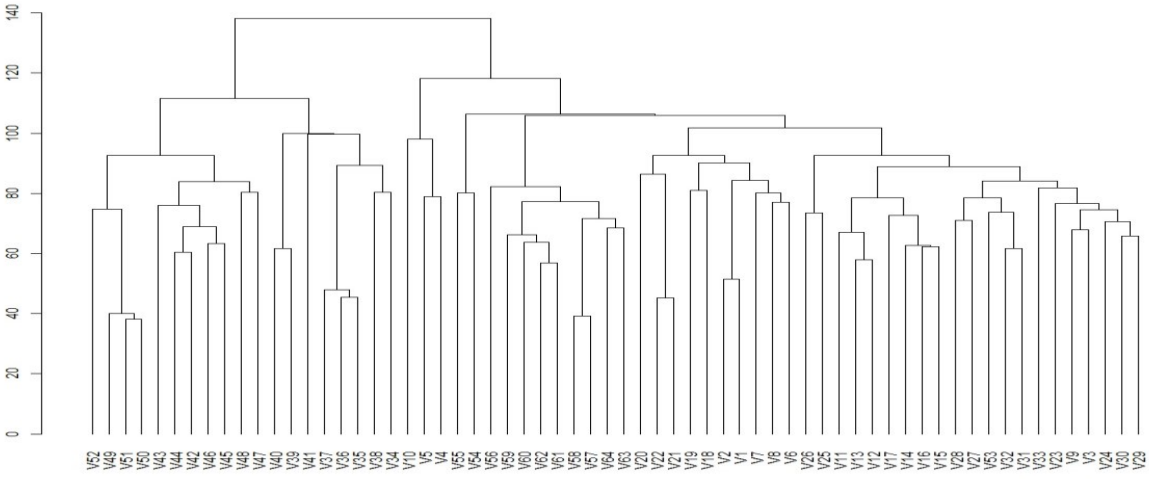
3. Discuss the performance of hierarchical agglomerative clustering when using different linkage functions.

1. Using Single



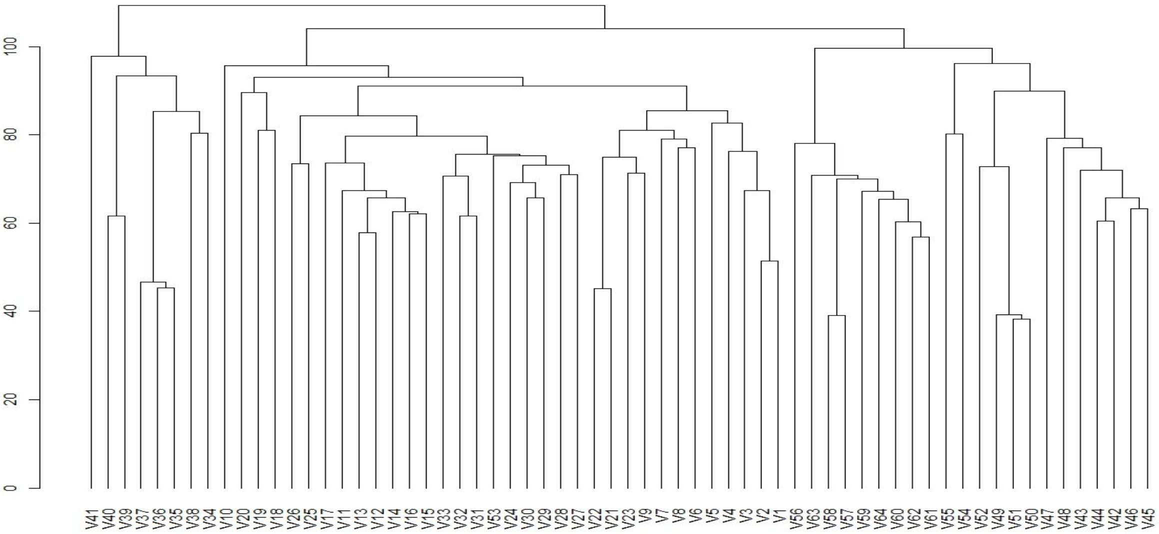
In the case of Single linkage, plot tend to show that attributes merged from right to left.

2. Using complete



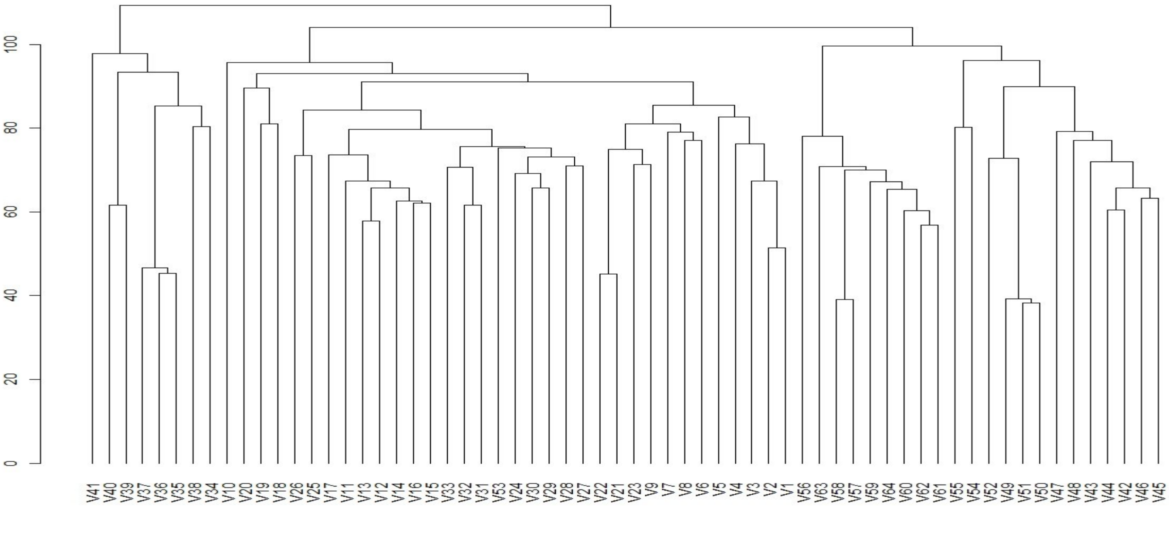
In the case of complete linkage, Since each observation is a merge from a maximum distance to the smallest distance, it tends to show that it goes through several merge processes more than single linkage.

3. Using average



In the case of average, the distance between observations is calculated by the average distance.

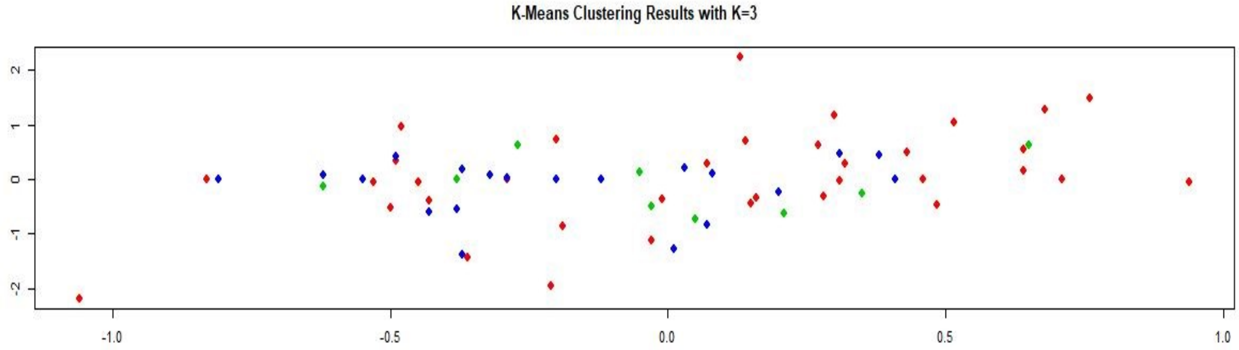
4. Using centroid

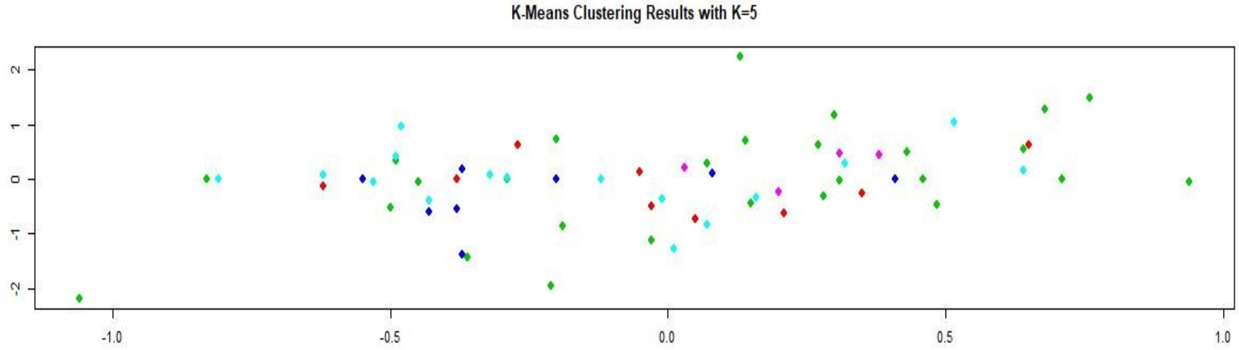


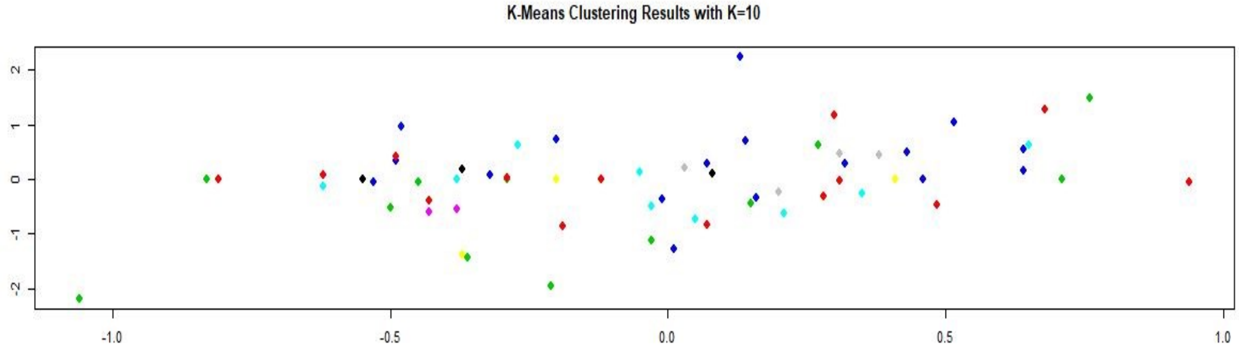
In my own program, plot shows the same as average.

4. Apply the R function kmeans() to the above NCI microarray data set

with different K and discuss its performance.


K-Means Clustering Results with K=3


K-Means Clustering Results with K=5


K-Means Clustering Results with K=10

Seeing the above picture, when K is 10, performance seems like to be better than others.

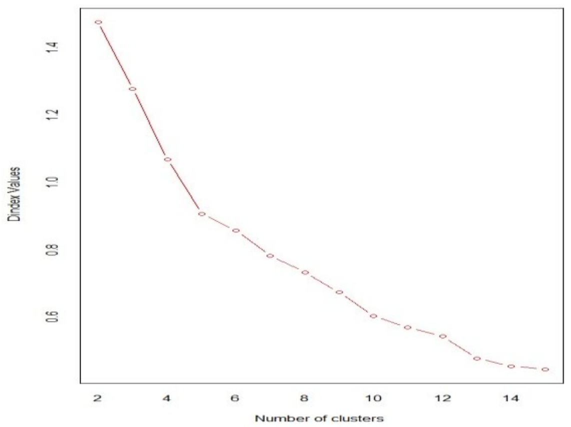
5 Compare and contrast the performance of K-means and hierarchical agglomerative clustering.

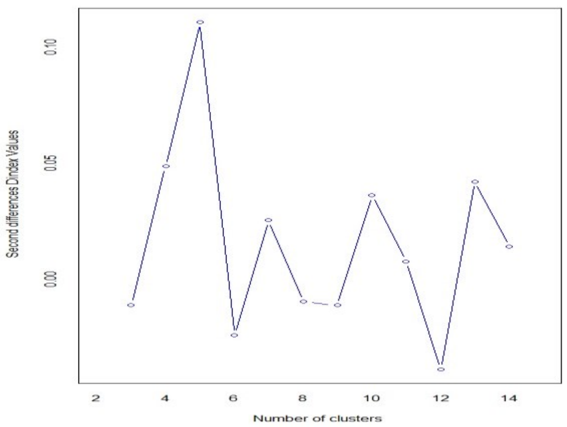
- The hierarchical cluster shows detail information. For example, after running the hierarchical cluster program, it can confirm which attributes merged and where has been clustered. Additionally, the hierarchical algorithm has several methods. Thus, it can be indicated slightly different results according to methods. However, this algorithm is much slower than K-means. In other hands, K-means depends on K. Therefore, the results might be much different.

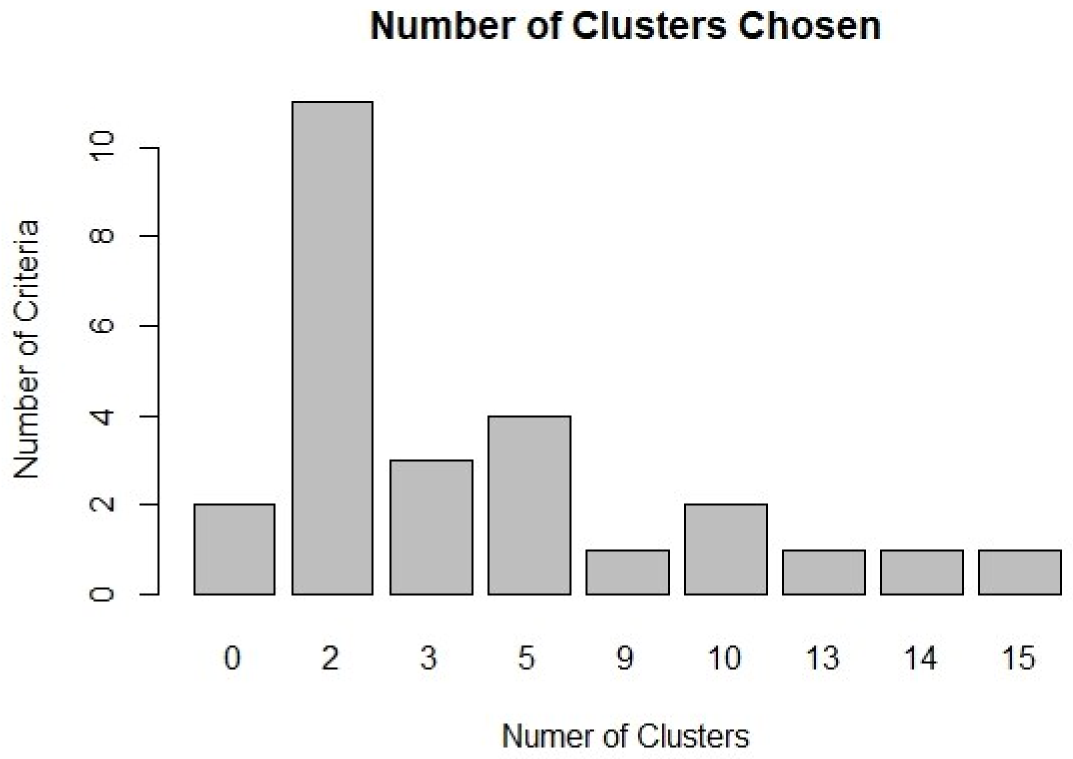
6 Discuss how to choose the number of clusters in the K-means and hierarchical agglomerative clustering.

There is no definitive answer to this question because the user should generate the number of clusters. Thus, the optimal number of clusters is somehow subjective. However, there are several methods in order to find the optimal number of clusters. Among these methods, it can be used Nbclust function provided in R package.

When using Nbclust function, function shows charts like below the picture.







This picture shows which the number of clusters is the best cluster.

As like this, there are no right answers for the optimal number of clusters but if using nbclust, it can find the best number of clusters.