# Rioja Use Cases

*Parismita Das*

*26 February 2017*

## Using Given Dataset

This is a example to obtain constrained HAC using dataset extracted from dissimilarity.txt We start with conveting the dataset given in format of .txt to distance class.

```
#reading dataset from txt file
data <- read.table("/home/parismita/dissimilarity.txt")
#converting to numeric list
n<-as.numeric(unlist(data))
#converting to dissimilarity matrix
dmat<-matrix(c(n),nrow = 77,ncol = 77)
dmat<-as.dist(dmat)
```

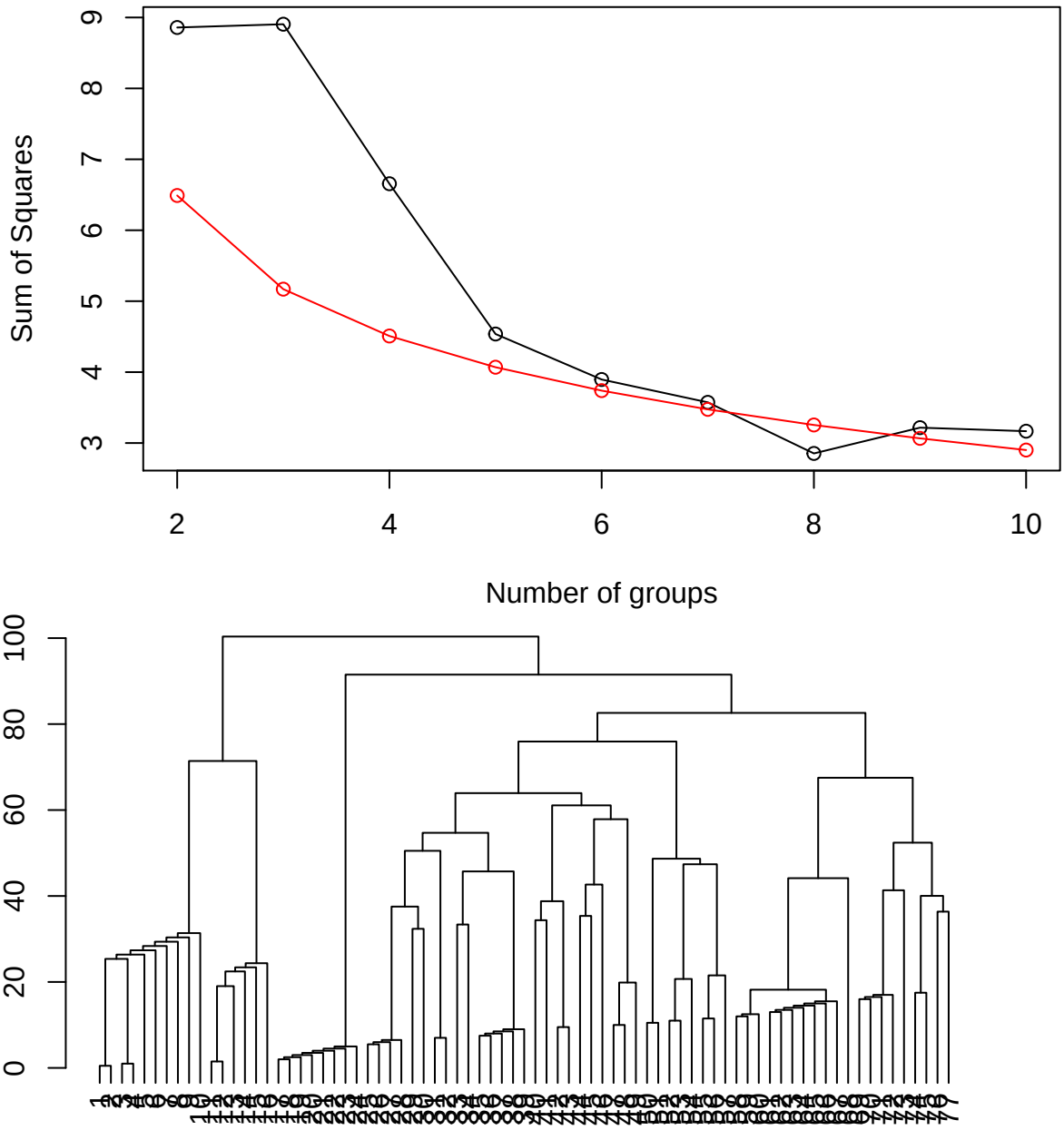
Computing the clusters using chclust. We have two methonds in rioja package

1) slink
2) ward's

```
#chclust function for Constrained hierarchical clustering method
clust <- chclust(diss)                          #using ward's
clust <- chclust(dmat,method = "conslink")    #using slink
```

Plotting the Bstick representation of the data for ward's method

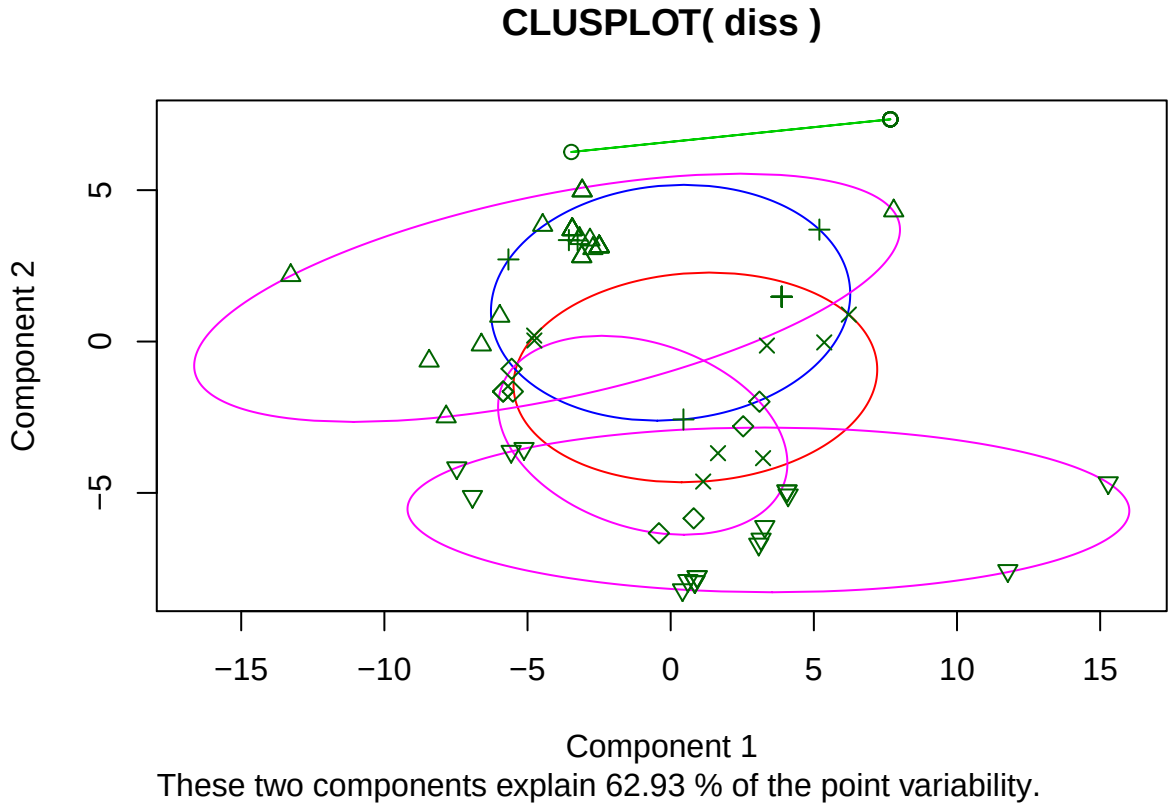
```
## This is rioja 0.9-9
```





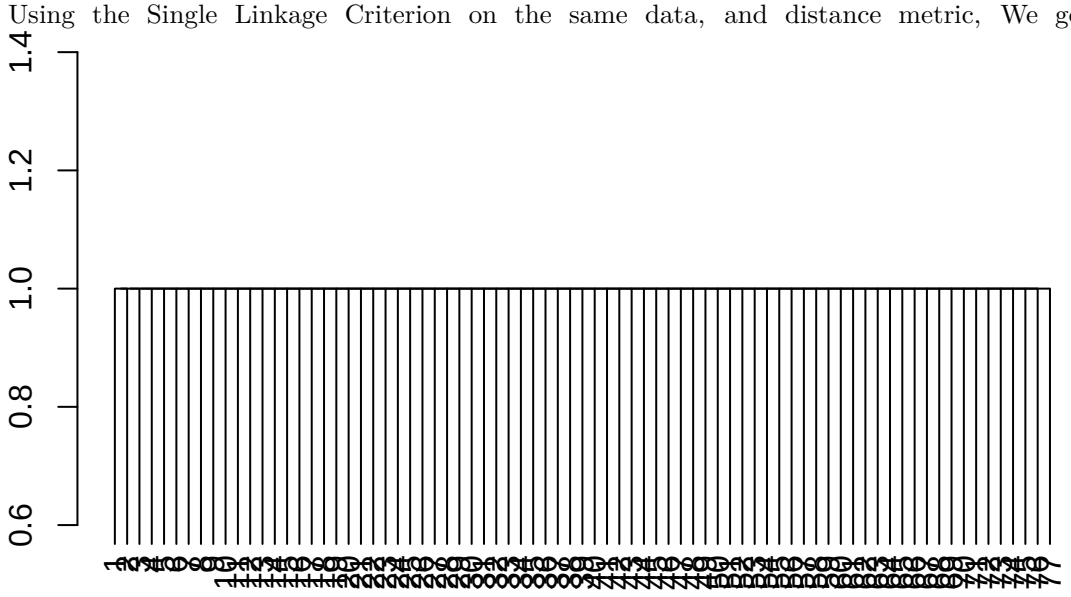


Plotting the dendogram diagram which we obtain after clustering, according to its distance.




Showing the Clusters we get using using the function hclust, Using Ward's Method.

Using the Single Linkage Criterion on the same data, and distance metric, We get the plot as:




Hence this shows the Hierarchical clustering using our own data.

## Problems with Non-Euclidian Data

Non-Euclidean distances can only be approximately embedded in a Euclidean space and not accurately. And Pseudo-Euclidean (PE) embedding can be found for any symmetric dissimilarity matrix.

For ward's linkage: the data can be misleading as the varience of the clusters are calculated by euclidean method

$$d_{ij} = d(\{X_i\}, \{X_j\}) = \|X_i - X_j\|^2.$$

For SLINK method: The data have no varience hence cant seperate into clusters as all clusters are at same height so even if we cut the dendrogram to obtain clusters, it gives same number of elements