Rioja Use Cases

Parismita Das

26 February 2017

Using Implemented Method

The inbuilt dataset RLGH(Diatom stratigraphic data) is used for showing clustering via constrained HAC method which is done by chclust function of Rioja Package. In the code

```
data(RLGH)
```

The distance Matrix is calculated by dist function of R, by default it calculates Euclidean Distance Metric

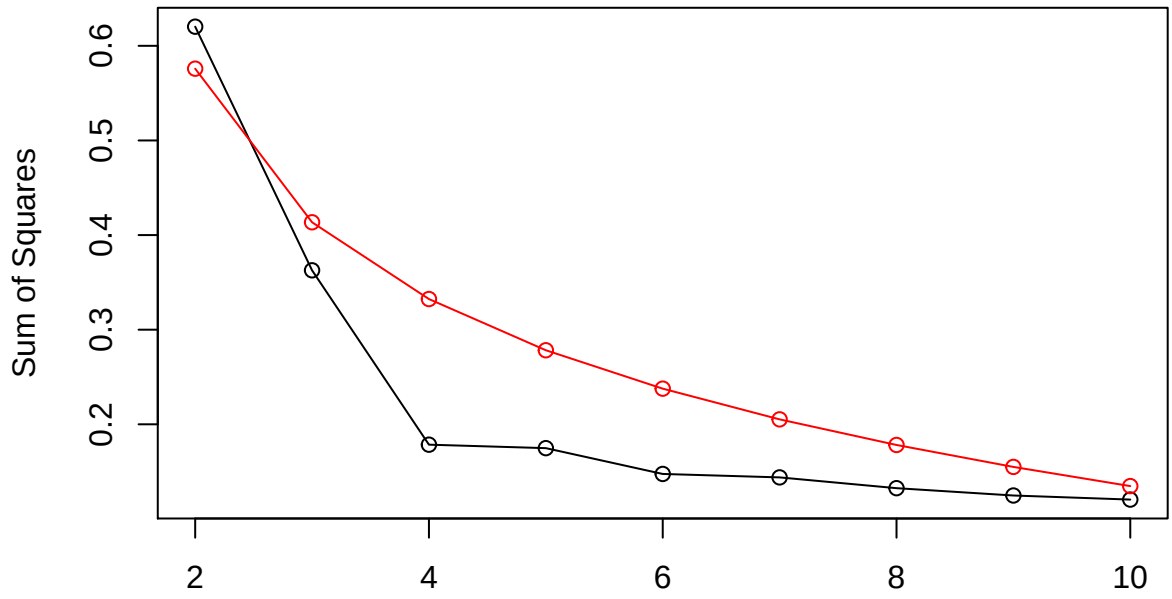
```
#diss is computed distance matrix of diatomic species relative abundance
diss <- dist(sqrt(RLGH$spec/100))
```

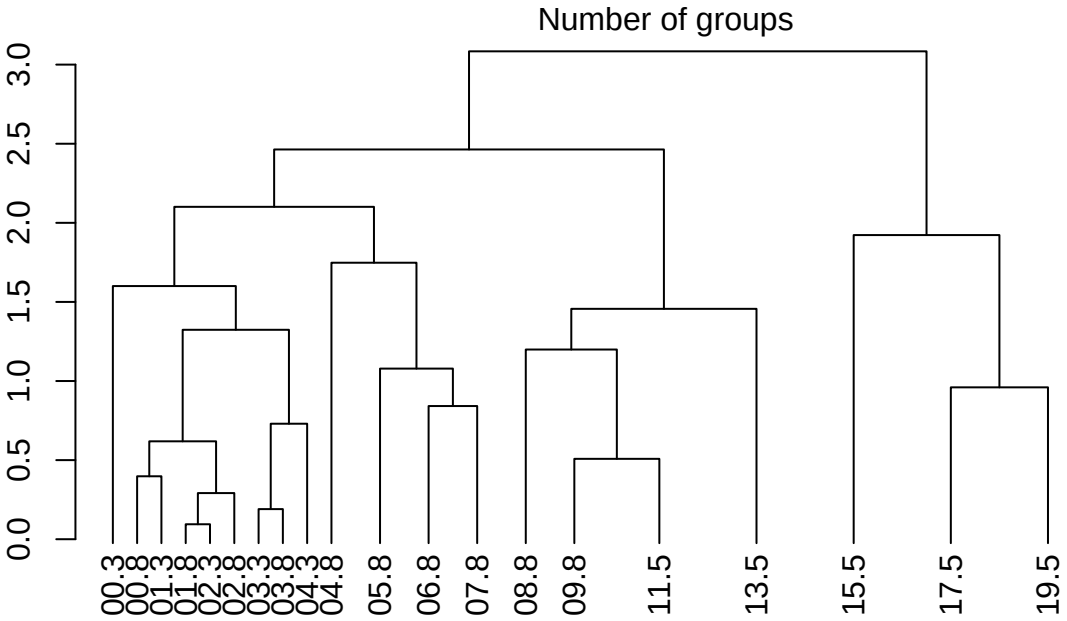
Using chclust function for Constrained hierarchical clustering by coniss method, and comparing the dispersion of a hierarchical classification to that obtained from a broken stick model using bstick

```
clust <- chclust(diss)
```

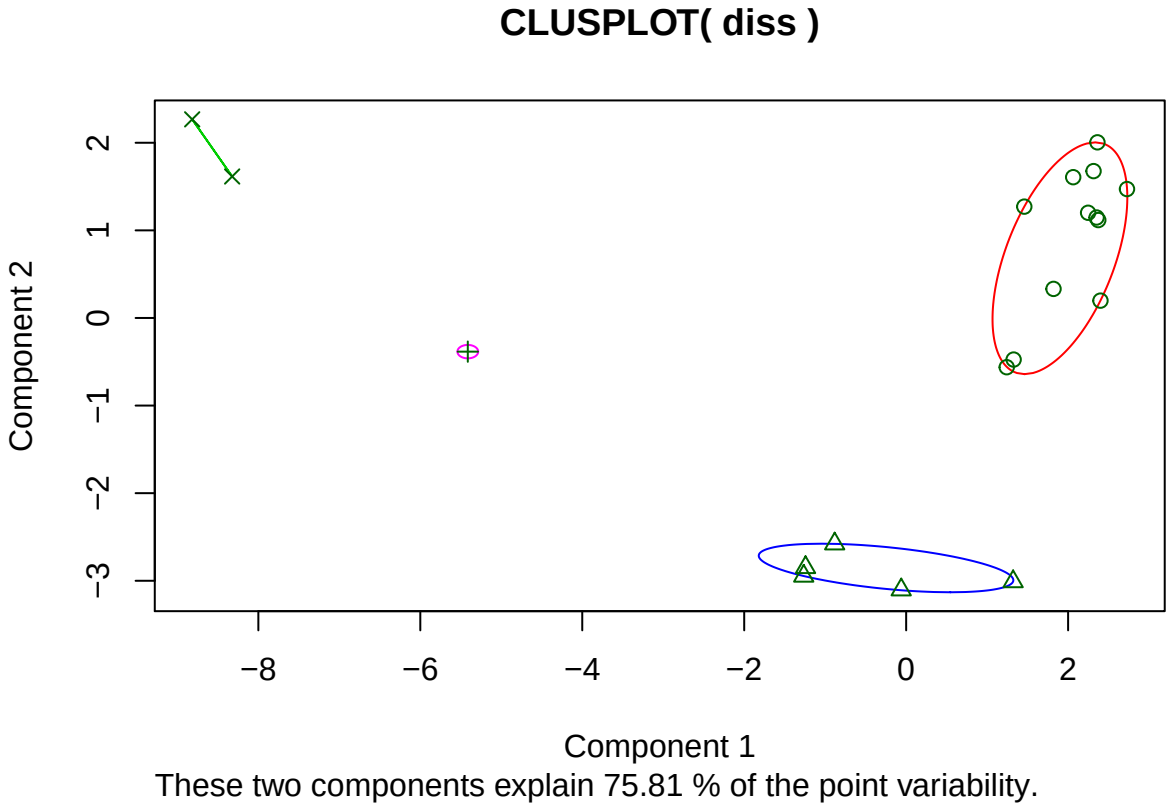
bstick function compares the dispersion of a hierarchical classification to that obtained from a broken stick model. Hence plotting variances of ordination axes/components and overlaying broken stick distributions

```
## 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.

Hence this shows the Hierarchical clustering using the Example given in the Rioja Package

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 Matrix form

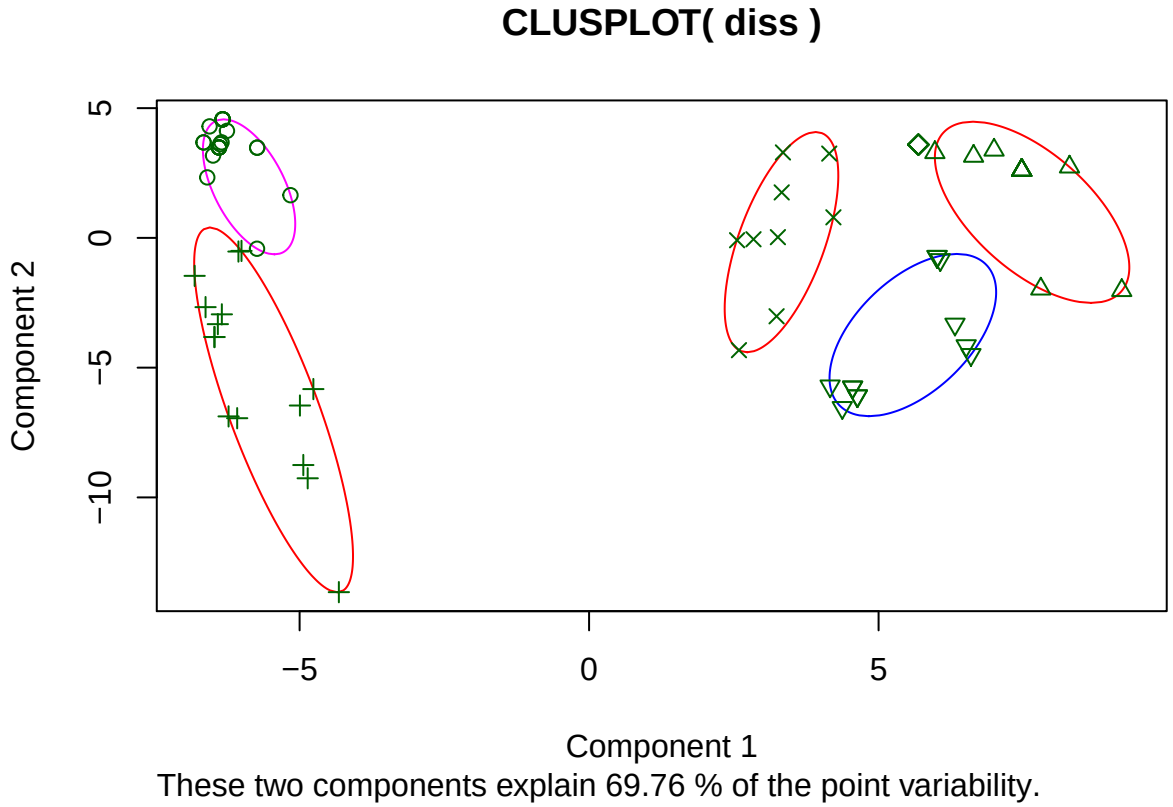
```
#reading dataset from txt file
data <- read.table("/home/parismita/dissimilarity.txt")
#converting to numeric list
n<-as.numeric(unlist(data))
#converting to dissimilarity matrix
dmat<-matrix(c(n),nrow = 77,ncol = 77)
```

As it is not euclidian dissimilarity matrix we have to find distance matrix by other methods. After trying out various Methods, The Canberra Method of calculating distance Matrix Works Best, The Canberra Method is a special case of Manhattan Method. And Ward's clustering to obtain the dendrogram.

```
#diss is computed distance matrix of dissimilarity matrix power of p >1
diss <- dist(dmat, method='canberra')
#chclust function for Constrained hierarchical clustering method
clust <- chclust(diss)
```

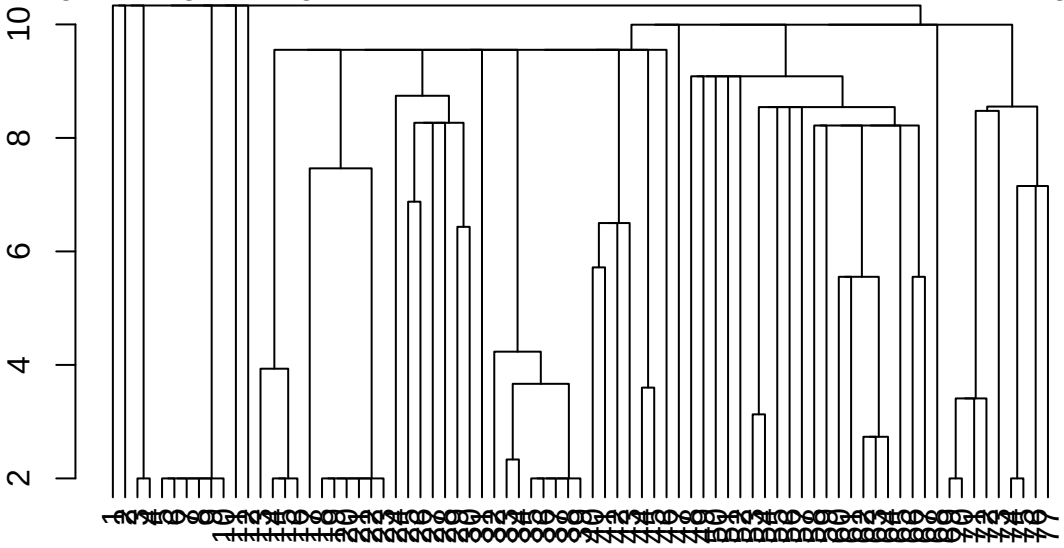
Plotting the Bstick representation of the data

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.

We do not calculate Pseudo-Euclidean (PE) embedding here, instead we calculate the distacne metric by ourself and fit it in rioja inbuilt functions

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

For SLINK method: The data have low varience hence cant seperate into clusters