M4\_L3\_RomilShah

Romil Shah

June 19, 2016

## Read Data and additional packages

require(ggplot2)

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.2.5

require(mclust)

## Loading required package: mclust

## Warning: package 'mclust' was built under R version 3.2.5

## Package 'mclust' version 5.2

## Type 'citation("mclust")' for citing this R package in publications.

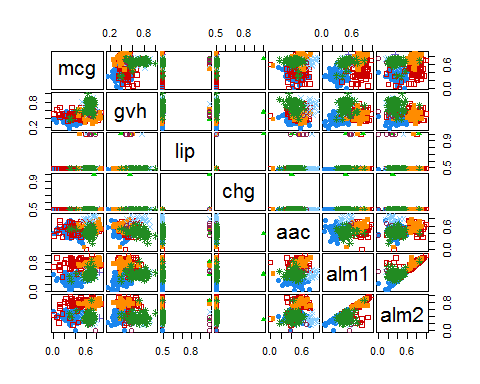
dataframe <- read.table("C:/Users/rams1/Desktop/DSCS6030/Module\_04/ecoli.data")  
colnames(dataframe) <- c("SeqNames","mcg","gvh","lip","chg","aac","alm1","alm2","Class")  
ecoli <- dataframe[2:8]  
class = dataframe$Class  
table(class)

## class  
## cp im imL imS imU om omL pp   
## 143 77 2 2 35 20 5 52

head(ecoli)

## mcg gvh lip chg aac alm1 alm2  
## 1 0.49 0.29 0.48 0.5 0.56 0.24 0.35  
## 2 0.07 0.40 0.48 0.5 0.54 0.35 0.44  
## 3 0.56 0.40 0.48 0.5 0.49 0.37 0.46  
## 4 0.59 0.49 0.48 0.5 0.52 0.45 0.36  
## 5 0.23 0.32 0.48 0.5 0.55 0.25 0.35  
## 6 0.67 0.39 0.48 0.5 0.36 0.38 0.46

# cluster Pairs  
clPairs(ecoli,class)



## Expectation Maximization Clustering

### Obtaining the EM Model

mod1<-Mclust(ecoli)  
mod1

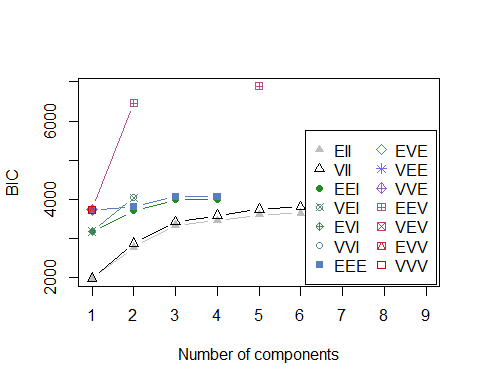
## 'Mclust' model object:  
## best model: ellipsoidal, equal volume and shape (EEV) with 5 components

summary(mod1)

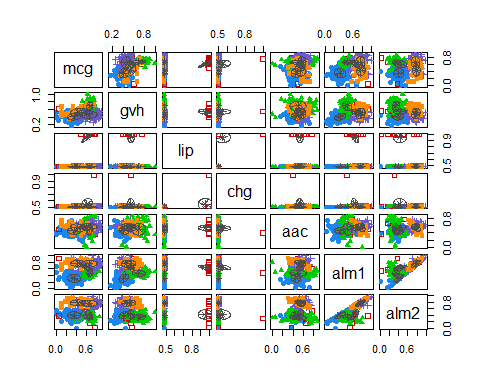
## ----------------------------------------------------  
## Gaussian finite mixture model fitted by EM algorithm   
## ----------------------------------------------------  
##   
## Mclust EEV (ellipsoidal, equal volume and shape) model with 5 components:  
##   
## log.likelihood n df BIC ICL  
## 3882.357 336 151 6886.33 6841.295  
##   
## Clustering table:  
## 1 2 3 4 5   
## 147 11 75 75 28

### Plots for EM produces following plots

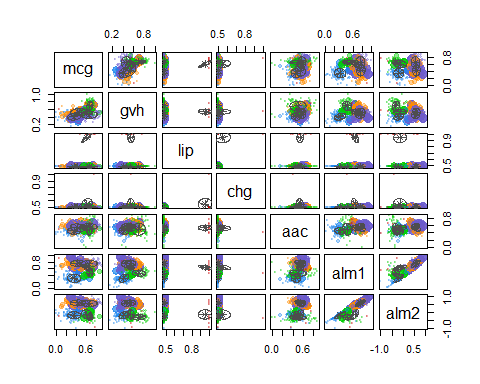
# BIC  
plot(mod1,what="BIC")



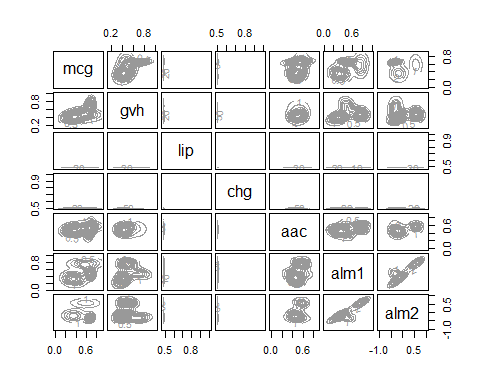
# Classification  
plot(mod1, what = "classification")



# Uncertainty  
plot(mod1, what = "uncertainty")



# Density  
plot(mod1, what = "density")



# Only BIC for EM initialized by model based clustering  
BIC = mclustBIC(ecoli)  
summary(BIC)

## Best BIC values:  
## EEV,5 EEV,2 EEE,3  
## BIC 6886.33 6461.3391 4068.066  
## BIC diff 0.00 -424.9909 -2818.264

## Answers:

Answers to the questions:

#### A(1)

The data consists of various relations between the clusters. The EM model was chosen such that the clustering by EM algorithm assigns the different clusters tothe data. As we have about 7 classes, it becomes easier to cluster the data such that the E step estimates the data of the Ecoli database and the M step will maximize the cluster allotment based on minimum error and maximum aposteriori.

#### A(2)

EM clustering is very similar to the clustering by k-means or PAM or hierarchical. But the major difference is that the overlap of the gaussians can be possible in EM which is not possible in -means. Thus the clustering gaussians can overlap so that the data is ordered into proper clusters which is not possible in k-means.