M4L3\_Assignment

Xueyi Fan

June 13, 2016

# Assignment:

* First, you will use the same dataset you chose for previous assignment (M04 Lesson 02 for the partition (k-means, PAM) and hierarchical clustering) from the he the UC Irvine Machine Learning Repository at <https://archive.ics.uci.edu/ml/>
* Next, cluster some of your data using EM based clustering and answer the following questions:

1. How did you choose a model for EM? Evaluate the model performance.
2. Cluster some of your data using EM based clustering that you also used for k-means, PAM, and hierarchical clustering. How do the clustering approaches compare on the same data?

In assignment M4L2, I choose the [Breast Cancer Wisconsin (Prognostic) data set](https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/)

Loading the data:

data\_url <- 'https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.data'  
  
cancer\_data <- read.table(url(data\_url), sep = ',')  
#cancer\_data<- read.table('wpbc.data.txt', sep=",")  
  
names(cancer\_data) <- c('ID number', 'Outcome','Time','radius\_mean','texure\_mean','perimeter\_mean','area\_mean','smoothness\_mean','compactness\_mean','concavity\_mean','concave\_points\_mean','symmetry\_mean','fractal\_dimension\_mean', 'radius\_SE','texure\_SE','perimeter\_SE','area\_SE','smoothness\_SE','compactness\_SE','concavity\_SE','concave\_points\_SE','symmetry\_SE','fractal\_dimension\_SE','radius\_worst','texure\_worst','perimeter\_worst','area\_worst','smoothness\_worst','compactness\_worst','concavity\_worst','concave\_points\_worst','symmetry\_worst','fractal\_dimension\_worst','tumor\_size','lymph\_node\_status')   
cancer <- data.frame(cancer\_data[,4:13])  
head(cancer)

## radius\_mean texure\_mean perimeter\_mean area\_mean smoothness\_mean  
## 1 18.02 27.60 117.50 1013.0 0.09489  
## 2 17.99 10.38 122.80 1001.0 0.11840  
## 3 21.37 17.44 137.50 1373.0 0.08836  
## 4 11.42 20.38 77.58 386.1 0.14250  
## 5 20.29 14.34 135.10 1297.0 0.10030  
## 6 12.75 15.29 84.60 502.7 0.11890  
## compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean  
## 1 0.1036 0.1086 0.07055 0.1865  
## 2 0.2776 0.3001 0.14710 0.2419  
## 3 0.1189 0.1255 0.08180 0.2333  
## 4 0.2839 0.2414 0.10520 0.2597  
## 5 0.1328 0.1980 0.10430 0.1809  
## 6 0.1569 0.1664 0.07666 0.1995  
## fractal\_dimension\_mean  
## 1 0.06333  
## 2 0.07871  
## 3 0.06010  
## 4 0.09744  
## 5 0.05883  
## 6 0.07164

str(cancer)

## 'data.frame': 198 obs. of 10 variables:  
## $ radius\_mean : num 18 18 21.4 11.4 20.3 ...  
## $ texure\_mean : num 27.6 10.4 17.4 20.4 14.3 ...  
## $ perimeter\_mean : num 117.5 122.8 137.5 77.6 135.1 ...  
## $ area\_mean : num 1013 1001 1373 386 1297 ...  
## $ smoothness\_mean : num 0.0949 0.1184 0.0884 0.1425 0.1003 ...  
## $ compactness\_mean : num 0.104 0.278 0.119 0.284 0.133 ...  
## $ concavity\_mean : num 0.109 0.3 0.126 0.241 0.198 ...  
## $ concave\_points\_mean : num 0.0706 0.1471 0.0818 0.1052 0.1043 ...  
## $ symmetry\_mean : num 0.186 0.242 0.233 0.26 0.181 ...  
## $ fractal\_dimension\_mean: num 0.0633 0.0787 0.0601 0.0974 0.0588 ...

In the breast cancer dataset, there are 35 columns. Two of them are factors and others are number. Here I use column 4-13 as the clustering data.

library("mclust")

## Package 'mclust' version 5.2

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

em\_clust <- Mclust(cancer)  
em\_clust

## 'Mclust' model object:  
## best model: ellipsoidal, equal shape and orientation (VEE) with 2 components

### 1. How did you choose a model for EM? Evaluate the model performance.

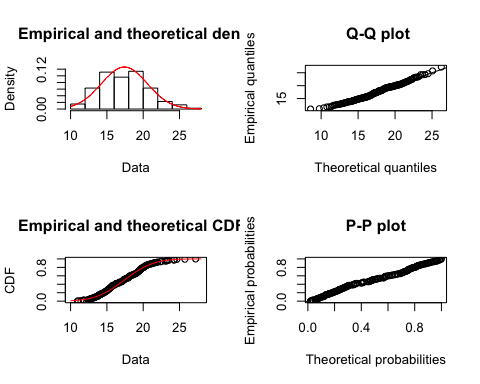
According to the tutorial of [mages' blog](http://www.magesblog.com/2011/12/fitting-distributions-with-r.html), I use fitdistrplus package to fit distributions

library("fitdistrplus")

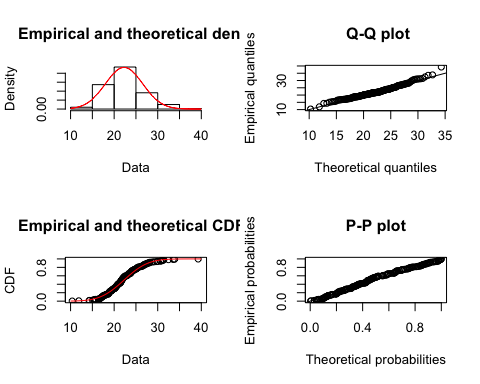
## Loading required package: MASS

for (i in 1:10){  
 fit <- fitdist(cancer[,i], distr = "norm", method = "mle", discrete = F)  
 cat("This is column ",i)  
 plot(fit)  
}

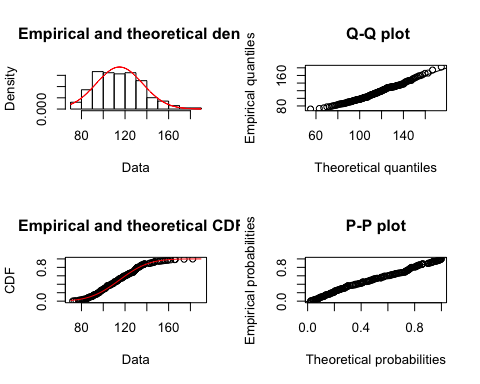
## This is column 1



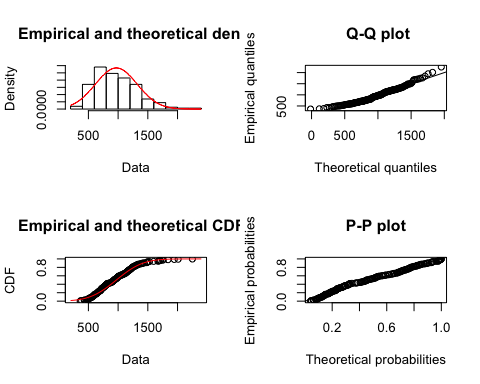
## This is column 2



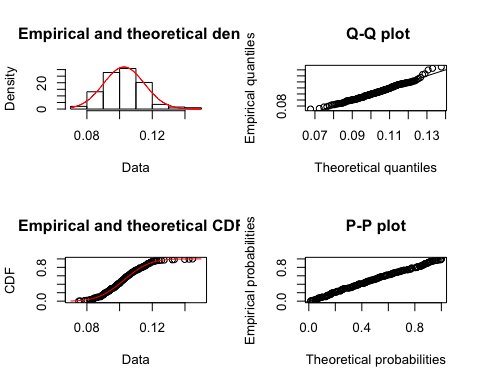
## This is column 3



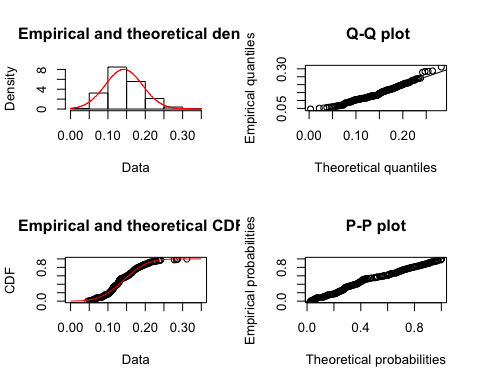
## This is column 4



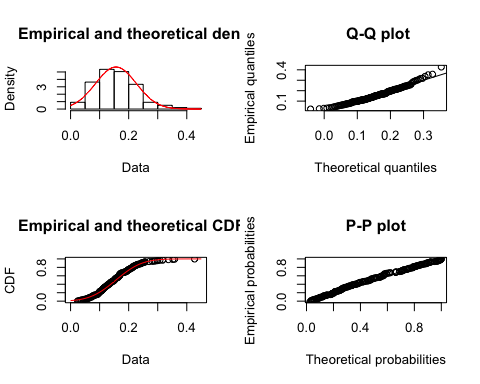
## This is column 5



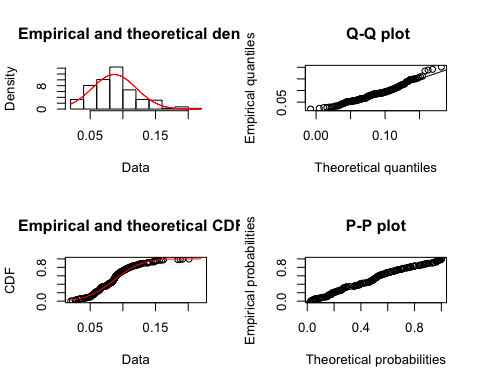
## This is column 6



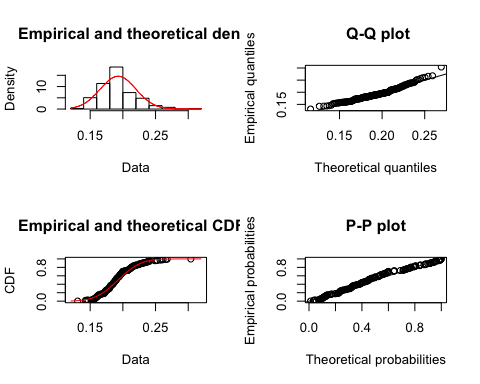
## This is column 7



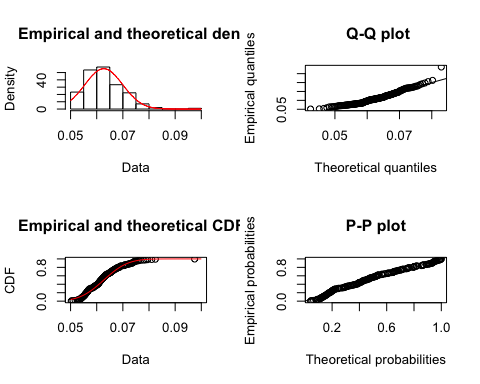
## This is column 8



## This is column 9



## This is column 10

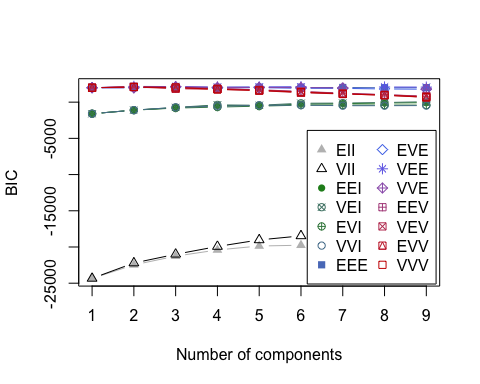


The function fitdist will created four graphs: the density plot, Q-Q plot, CDF, P-P plot. According to the results of Q-Q plots, the data from these 10 columns are from normal distribution. So I will choose mixture model which is a mixtrure of Gasussians.

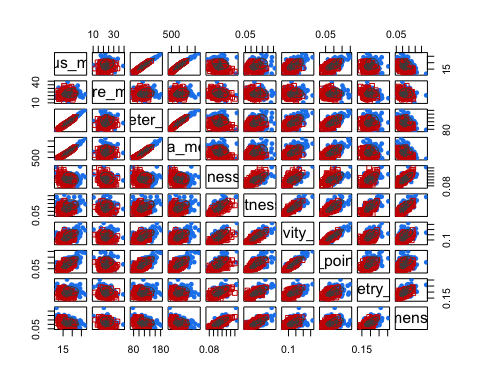
#evaluate the model performance  
summary(em\_clust)

## ----------------------------------------------------  
## Gaussian finite mixture model fitted by EM algorithm   
## ----------------------------------------------------  
##   
## Mclust VEE (ellipsoidal, equal shape and orientation) model with 2 components:  
##   
## log.likelihood n df BIC ICL  
## 1290.914 198 77 2174.631 2158.937  
##   
## Clustering table:  
## 1 2   
## 50 148

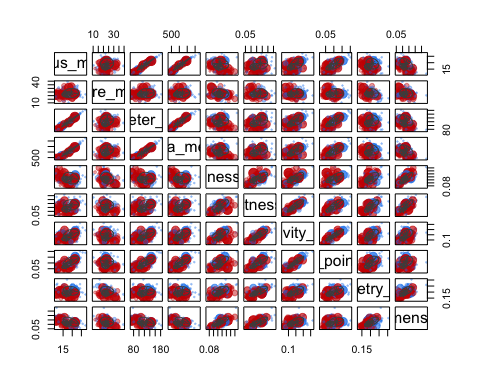
# BIC  
plot(em\_clust, what = "BIC")



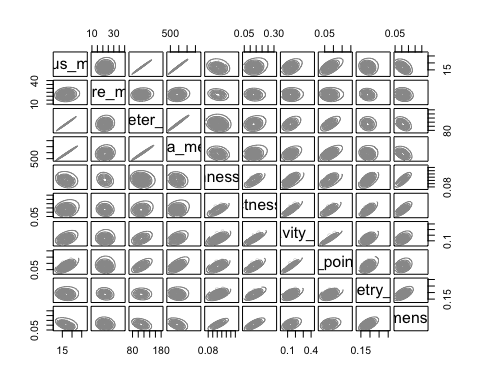
#classification  
plot(em\_clust, what = "classification")



#uncertainty  
plot(em\_clust, what = "uncertainty")



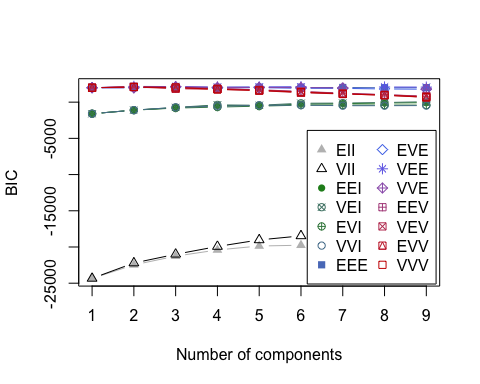
#density  
plot(em\_clust, what = "density")



#BIC  
BIC = mclustBIC(cancer)  
summary(BIC)

## Best BIC values:  
## VEE,2 VEE,3 VEE,6  
## BIC 2174.631 2151.12864 2141.82450  
## BIC diff 0.000 -23.50198 -32.80612

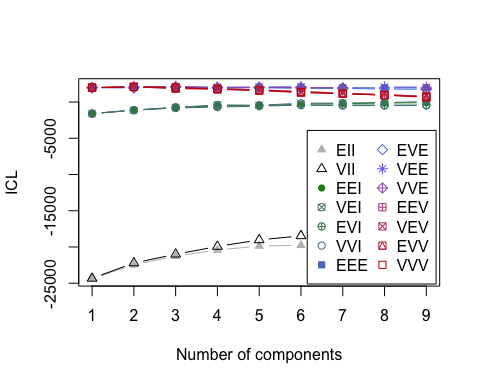
plot(BIC)



#ICL   
ICL = mclustICL(cancer)  
summary(ICL)

## Best ICL values:  
## VEE,2 VEE,6 VEE,3  
## ICL 2158.937 2122.20671 2120.16562  
## ICL diff 0.000 -36.73064 -38.77172

plot(ICL)



Based on the result of BIC anD ICL, the EII model of mixtrure of Gasussians has the best performance which has the lowest score of BIC or ICL.

### 2. Cluster some of your data using EM based clustering that you also used for k-means, PAM, and hierarchical clustering. How do the clustering approaches compare on the same data?

#EM  
table(cancer\_data$Outcome, em\_clust$classification)

##   
## 1 2  
## N 39 112  
## R 11 36

#k-means  
cancer.2.kmeans <- kmeans(cancer, centers = 2)  
table(cancer\_data$Outcome, cancer.2.kmeans$cluster)

##   
## 1 2  
## N 53 98  
## R 25 22

#PAM  
library("cluster")

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

cancer.2.pam <- pam(cancer,2)  
table(cancer\_data$Outcome, cancer.2.pam$clustering)

##   
## 1 2  
## N 63 88  
## R 31 16

#hierarhical clustering  
  
cancer.h.clust <- hclust(d=dist(cancer))  
hclust.2 <- cutree(cancer.h.clust, k=2)  
table(cancer\_data$Outcome,hclust.2)

## hclust.2  
## 1 2  
## N 139 12  
## R 41 6

According to the confusion matrix, the EM based clustering method and hierarchical clustering method have better performance than k-means and PAM. Most of nonrecur or recur are clustered together well in the EM and hierarchical clustering method.