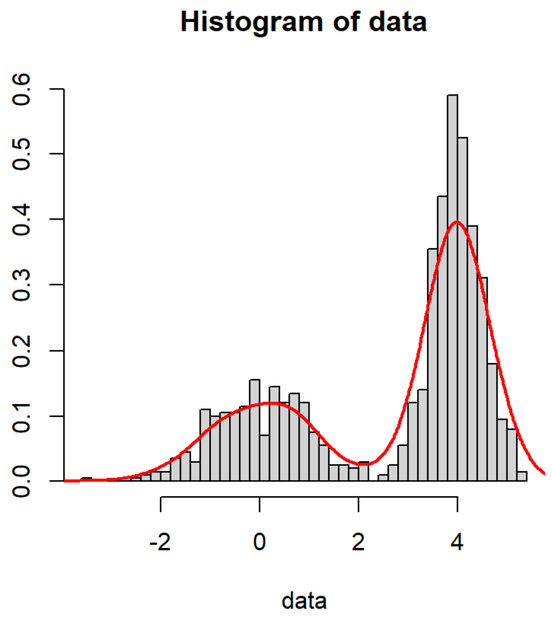
Fiche MAD 2

TP 3 - Mixture Model

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians.
In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs. Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. However, while problems associated with "mixture distributions" relate to deriving the properties of the overall population from those of the sub-populations, "mixture models" are used to make statistical inferences about the properties of the sub-populations given only observations on the pooled population, without sub-population identity information.
Our goal here is to create a one dimensional mixture of 2 gaussians with given parameters, we will use also the *K-Means* algorithm to find 2 cluster and apply the E-M algorithm.

```
n<-1000
Proportions<-c(1,2)/3
parameters<-list()
parameters[[1]]<-list(mean=0,sd=1)
parameters[[2]]<-list(mean=4,sd=1/2)

simu-mixture<-function(n=100,parameters,Proportions)
z<-t(rmultinom(n,1,prob = Proportions))
z<-apply(z,1,which.max)
x<-matrix(0,n,1)
for (i in 1:length(z))
x[i,1]<-rnorm(1,mean=parameters[[z[i]]]DOLLARmean,
sd=parameters[[z[i]]]DOLLARsd)
return(list(x=x,z=z))

hist(simu-mixture(n=1000,  parameters=parameters,
Proportions = Proportions)DOLLARx,breaks=50)
```

By doing this we create a vector with all of ours parameters and we can now use a histogram to plot our values :



We can see here that we have two major values that are highlight here, 0 and 4 the two means given for this mixture. Now we are using *K-Means* algorithm :

```
res<-kmeans(simulationDOLLARx,2,nstart = 30)
```

With those results, we are trying to find the kmeans output (classification) estimate the parameters of the mixture, so we build 2 vectors with each part of the *K-Means* separation and we just look trough them with *mean* and *sd*.

```
X-g1=simulationDOLLx[resDOLLcluster==1,]
X-g2=simulationDOLLx[resDOLLcluster==2,]
mean(X-g1)
sd(X-g1)
```

With those commands we should be able to get the mixture model parameters.

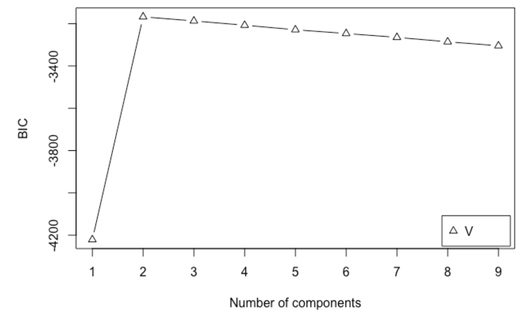
We are now using *Mclust* library to work with the EM algorithm. First by doing the E-step and then the M-step and combine them together.

EM-Algorithm : The EM algorithm is an iterative approach that cycles between two modes. The first mode attempts to estimate the missing or latent variables, called the estimation-step or E-step. The second mode attempts to optimize the parameters of the model to best explain the data, called the maximization-step or M-step.
E-Step : Estimate the missing variables in the dataset.
M-Step : Maximize the parameters of the model in the presence of the data.
The EM algorithm can be applied quite widely, although is perhaps most well known in machine learning for use in unsupervised learning problems, such as density estimation and clustering. Perhaps the most discussed application of the EM algorithm is for clustering with a mixture model.

So we do apply our Algorithm EV, we apply it with constant variance, here and see what we got :

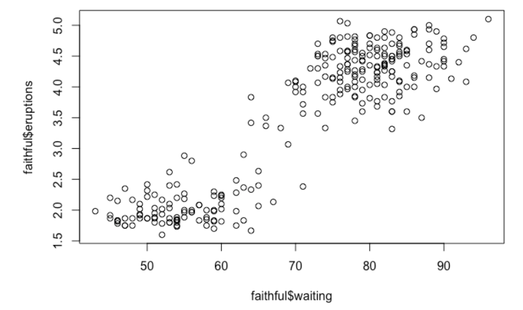
```
model-Mclust-E <- Mclust(simulationDOLLARx,
G=2, modelNames="E") equal variance
model-Mclust-EDOLLARparameters

model-Mclust-V <- Mclust(simulationDOLLARx,
G=2, modelNames="V") equal variance
model-Mclust-VDOLLARparameters

plot(model-Mclust-VDOLLARBIC)
```

So after applying our algorithm, our goal is to compare each one of the 3 answers we got and try to understand why do we got same or close values. We do plot BIC, and see the difference between the number of composents :



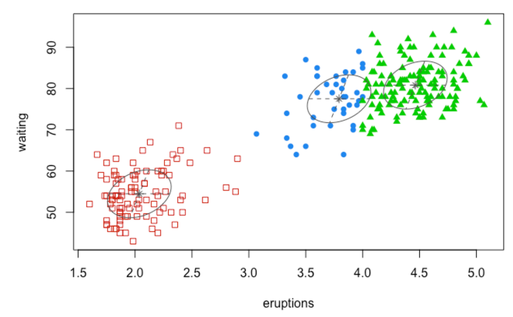
We are now working on faithful data, and our goal is to create some cluster and analyse them. The first idea is to look directly into the dataset by plotting some some values and see if we do have some cluster who get out of this plot. *Mclust* choose the hyper parameters with the previous criteria and choose the number of cluster.

```
data(faithful)
summary(faithful)
plot(faithfulDOLLARwaiting,
faithfulDOLLAReruptions)
```



We look into this plot and we try to see if there is any cluster (group) that come to us easily (it can be a bait), so maybe there is likely 2 group here. So we continue to work on our model :

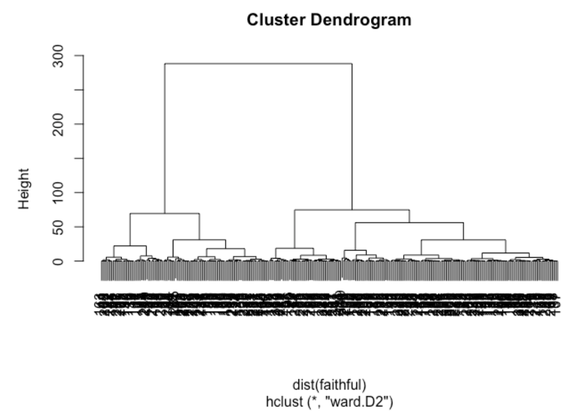
```
model <- Mclust(faithful)
plot(model)
```



We are now looking up to run the hclust on the data using the Ward Criterion and compare the clustering of hclust and the clustering of Mclust for two clusters.

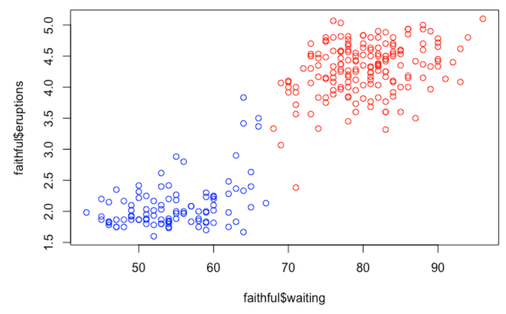
Care here - is tiret8

```
hclust-ward.D2 = hclust(dist(faithful),method =
"ward.D2")
plot(hclust-ward.D2)

cut-ward.D2-2 <- cutree(hclust-ward.D2,k = 2)
model <- Mclust(faithful,G=2)
table("Mixture model"=
modelDOLLARclassification,
"Hierarchical model" = cut-ward.D2-2)
```

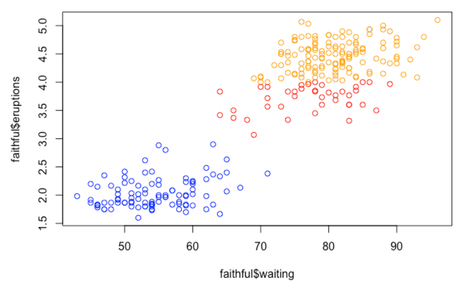


And now for 3 :

```
cut-ward.D2-3 <- cutree(hclust-ward.D2,k = 3)
model <- Mclust(faithful,G=3)
table("Mixture model"=
modelDOLLARclassification,
"Hierarchical model" = cut-ward.D2-3)
plot(faithfulDOLLARwaiting,
faithfulDOLLAReruptions,
col=c("Red","Blue") [cut-ward.D2-2])
plot(faithfulDOLLARwaiting,
faithfulDOLLAReruptions,
col=c("Red","Blue","Orange")  [modelDOLLARclassification])
```

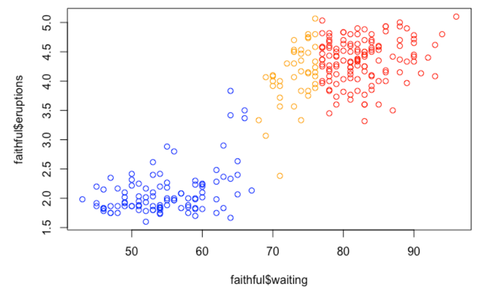


Plot 2 Cluster



Plot 3 Cluster

```
plot(faithfulDOLLARwaiting,
faithfulDOLLAReruptions,
col=c("Red","Blue","Orange") [cut-ward.D2-3])
```



Plot 3 Cluster

As we can see here, we have a method that give us 2 clusters. After we plot 2 differents plots of 3 clusters, two of them are supper-posing themselves and they have 2 different distribution. It s because clustering is not always giving the same answer, because they start by picking out K individuals in the data they are given.