mclust analysis of the Old Faithful Dataset

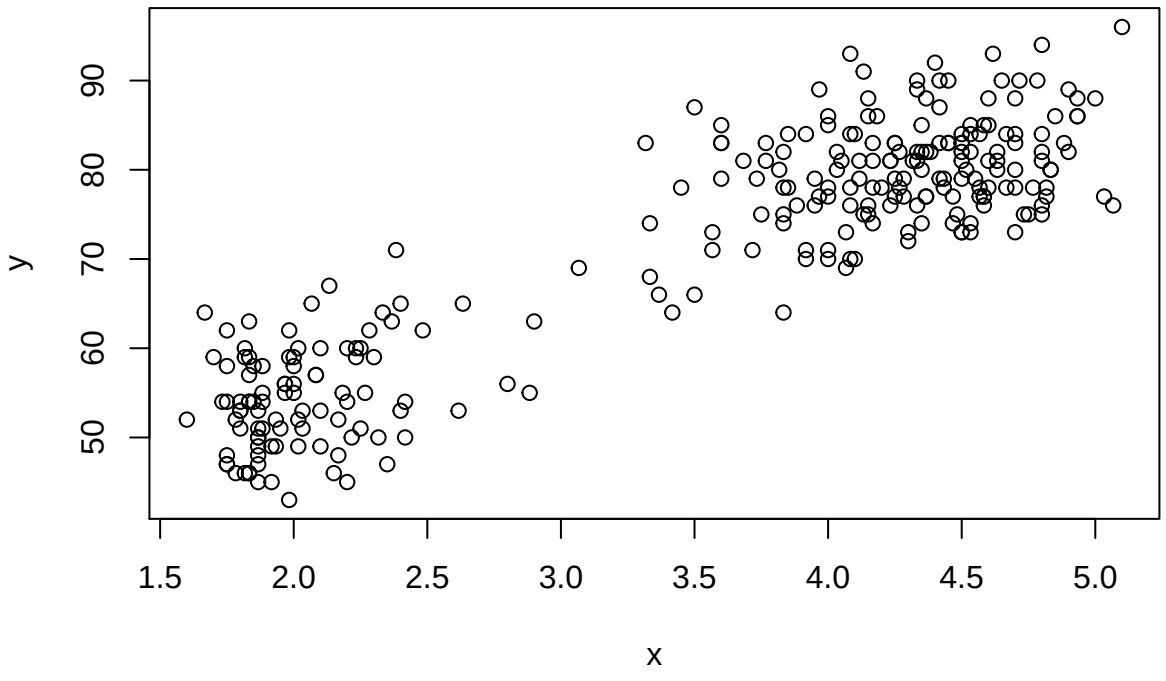
Jon Kinsey

Wed Dec 3 20:33:47 2014

```
# mclust provides functionality for cluster analysis combining model-based
# hierarchical clustering, EM for Gaussian mixture models, and BIC.
# First the cluster analysis estimates the number of clusters that best
# represents this data set and also the covariance structure of the spread points.
# This is performed through the technique called Bayesian Information Criterion (BIC)
# that varies the number of cluster from 1 to 9. The BIC is the value of the
# maximized loglikelihood measured with a penalty for the number of parameters
# in the model. Then it's executed the hierarchical clustering technique (HC),
# which doesn't require a initialization phase. The output of the HC, that is, the
# cluster that each element belongs, is used to initialize the
# Expectation-Maximization technique (EM).
#
library(mclust)
```

```
## Package 'mclust' version 4.4
## Type 'citation("mclust")' for citing this R package in publications.
```

```
#
x = faithful[,1] # get the first column of the faithful data set
y = faithful[,2] # get the second column of the faithful data set
plot(x,y) # plot before clustering
```



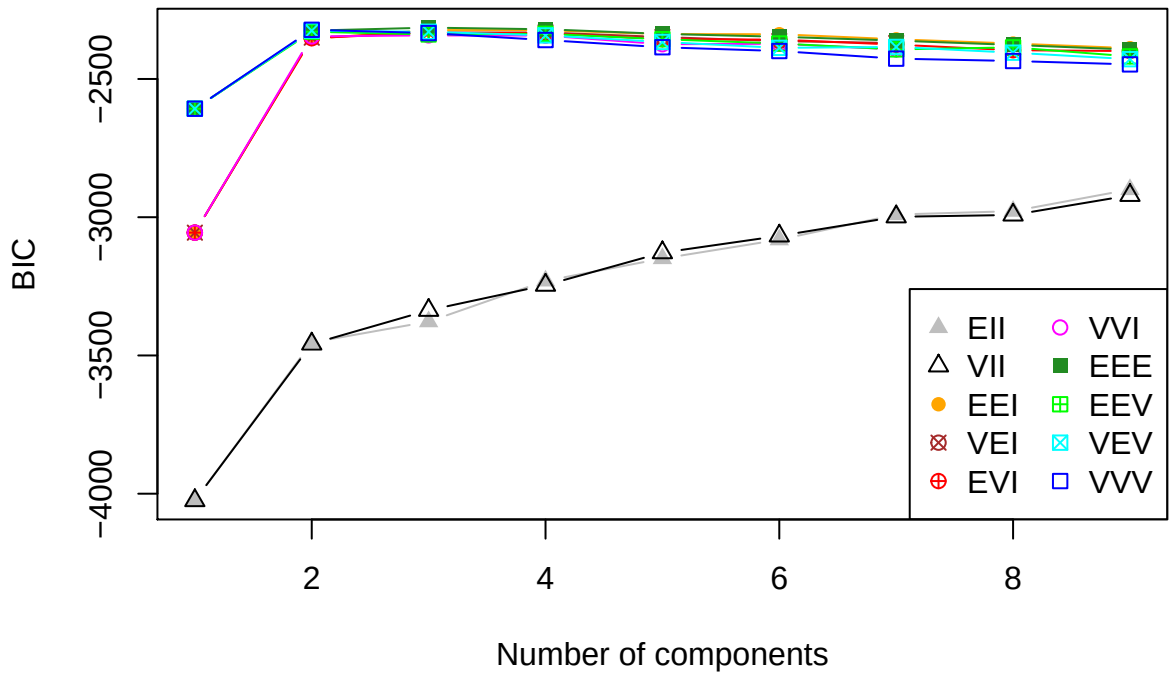

```
#
# # estimate the number of cluster (BIC), initialize (HC) and clusterize (EM)
faithfulMclust <- Mclust(faithful)
summary(faithfulMclust)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEE (elliposidal, equal volume, shape and orientation) model with 3 components:
##
##  log.likelihood   n df   BIC   ICL
##           -1126 272 11 -2314 -2361
##
## Clustering table:
##   1   2   3
## 130  97  45
```

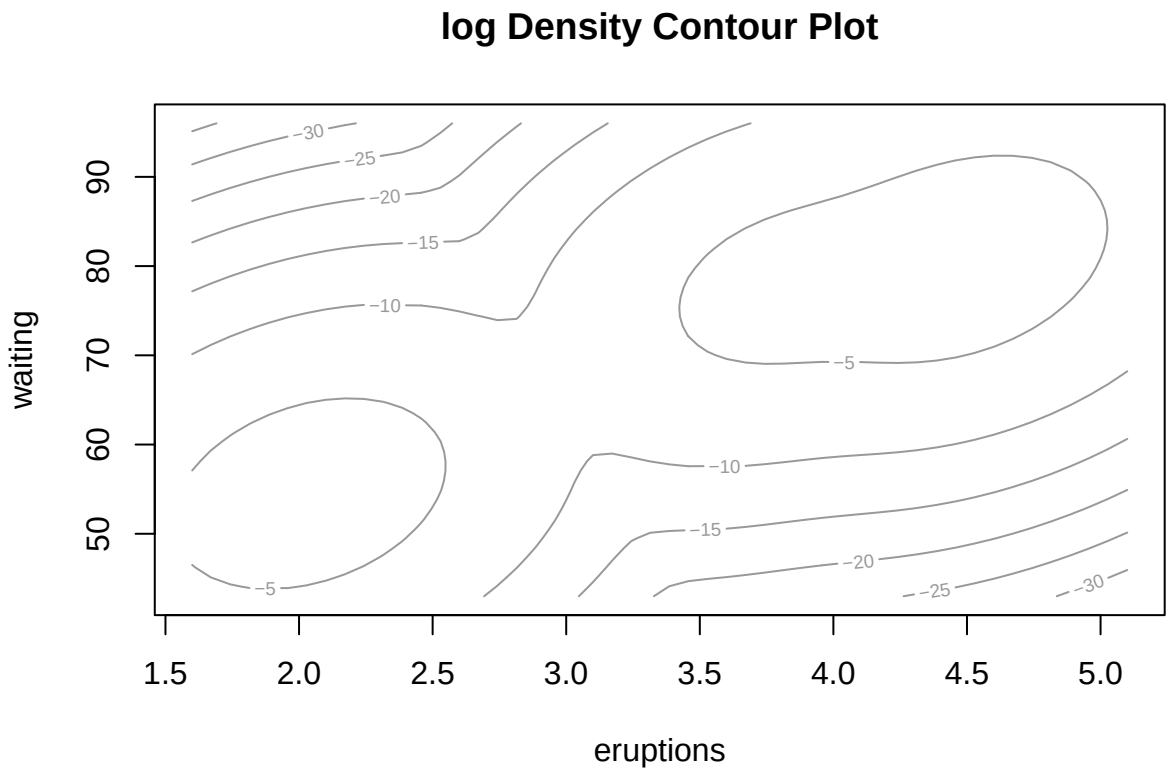
```r
# In this case, the best model according to BIC is an equal-covariance model
# with 3 components or clusters. A more detailed summary including the
# estimated parameters can be obtained with the following code:
summary(faithfulMclust, parameters = TRUE)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEE (elliposidal, equal volume, shape and orientation) model with 3 components:
##
##  log.likelihood   n df   BIC   ICL
##           -1126 272 11 -2314 -2361
##
## Clustering table:
##   1   2   3
## 130  97  45
##
## Mixing probabilities:
##      1      2      3
## 0.4619 0.3565 0.1816
##
## Means:
##             [,1]   [,2]  [,3]
## eruptions  4.476  2.038  3.82
## waiting   80.892 54.493 77.67
##
## Variances:
## [,,1]
##           eruptions waiting
## eruptions   0.07728  0.4765
## waiting     0.47650 33.7485
## [,,2]
##           eruptions waiting
## eruptions   0.07728  0.4765
## waiting     0.47650 33.7485
## [,,3]
##           eruptions waiting
## eruptions   0.07728  0.4765
## waiting     0.47650 33.7485
```

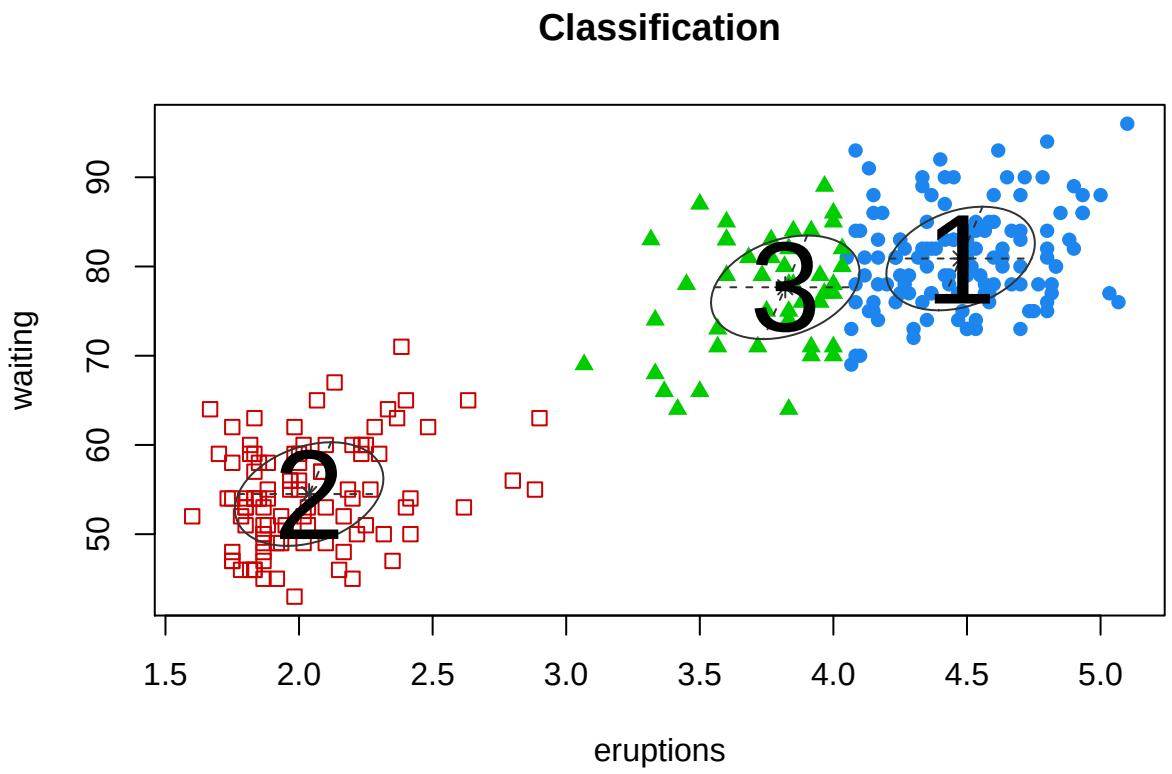
```
#
plot(faithfulMclust,what="BIC")
```




```
plot(faithfulMclust,what="density")
```



```
plot(faithfulMclust,what="classification")
clustmeans<-faithfulMclust$parameters$mean
text(clustmeans[1,], clustmeans[2,], seq.int(ncol(clustmeans)), cex=4)
```




```
#
# If the next day you experience 2 eruptions with waiting 50 and want to
# classify that value using the existing model, you would use
pp <- predict(faithfulMclust, newdata=data.frame(eruptions=2, waiting=50))
pp$classification
```

```
## [1] 2
```

```
# which gives cluster number 2
# what if you have 4 eruptions with a waiting time of 70 then
pp <- predict(faithfulMclust, newdata=data.frame(eruptions=4, waiting=70))
pp$classification
```

```
## [1] 3
```

```
# this gives cluster 3 which agrees with the plot
#
# The object produced by Mclust is a list with a components describing the
# estimated model. The names of these components can be displayed as follows:
names(faithfulMclust)
```

```
##  [1] "call"           "data"           "modelName"      "n"
```

```
##  [5] "d"              "G"              "BIC"            "bic"
##  [9] "loglik"         "df"             "hypvol"         "parameters"
## [13] "z"              "classification" "uncertainty"
```

```
#
```