D2

(Unsupervised learning) clustering

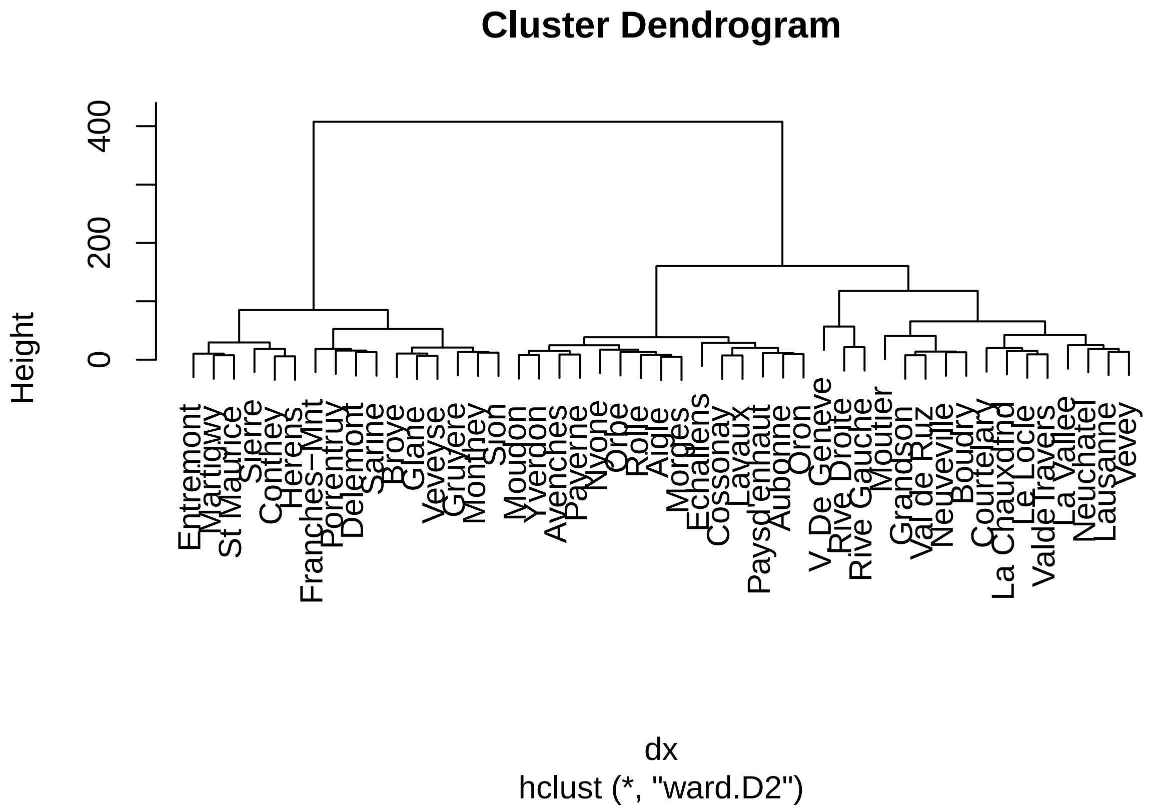
The hierarchical clustering

this method is implemented in R within the 'class' package and the appropriate method is names 'hclust'

> exercise: cluster the 'swiss' data with 'hclust'

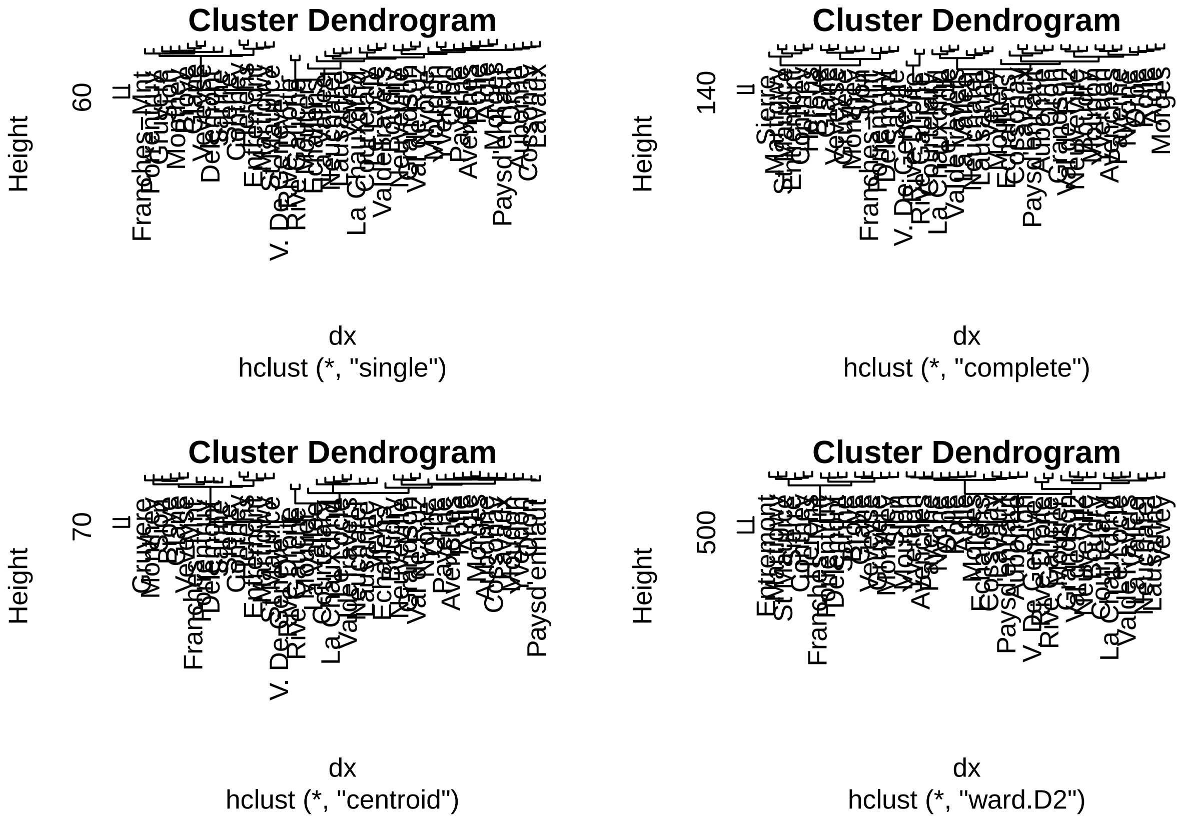
```
library(class)
x=swiss
#?hclust
#need to build distance matrixs
dx=dist(x) #default: method"euclidean"

#ward distance
cluster<-hclust(dx,method="ward.D2")
#clearly 2 groups
plot(cluster)
```



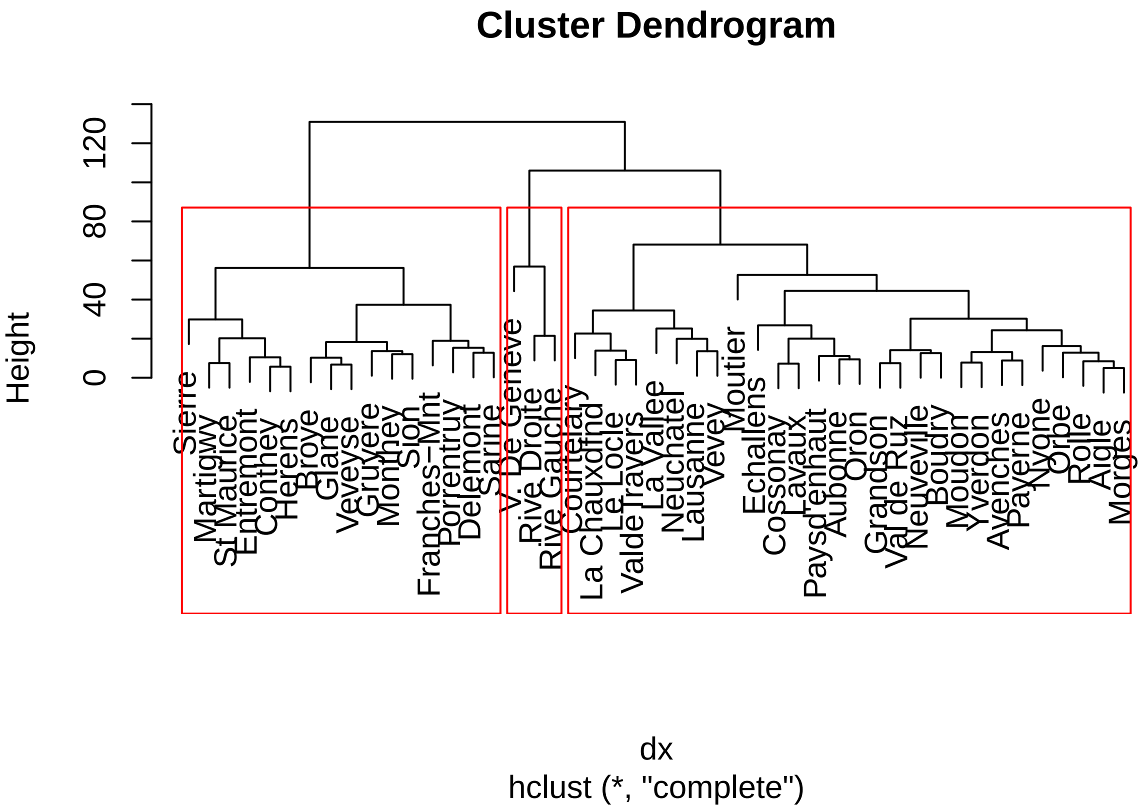
now try 4 differenct method

```r
par(mfrow=c(2,2))
#single distance
cluster1<-hclust(dx,method="single")
plot(cluster1)
#complete distance
outComplete<-hclust(dx,method="complete")
plot(outComplete)
#centroiddistance
cluster3<-hclust(dx,method="centroid")
plot(cluster3)
#ward distance
outWard<-hclust(dx,method="ward.D2")
plot(outWard)
```

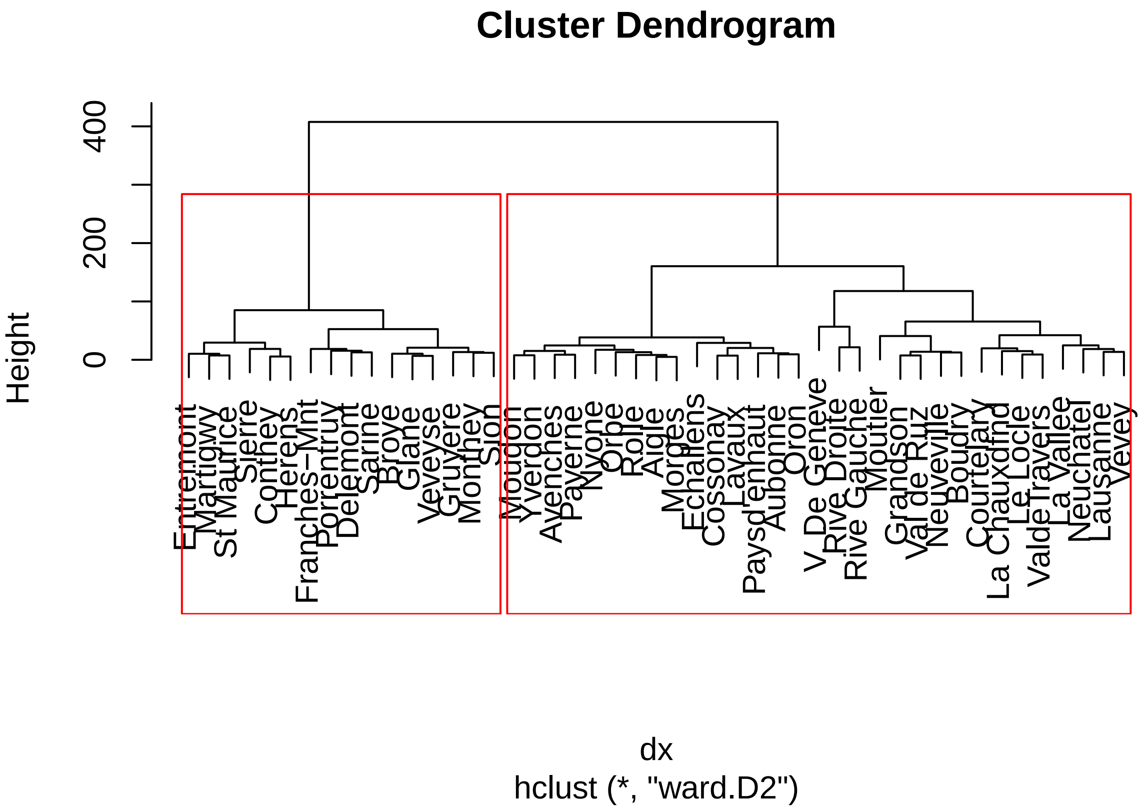


- now we choose only 2 to compare: complete and ward at this point we don't have yet the assignment to the clustering: we need cutree

```r
plot(outComplete)
k1=3
#get clustering
res1=cutree(outComplete,k1)
#visualize for cluster
rect.hclust(outComplete,k1)
```

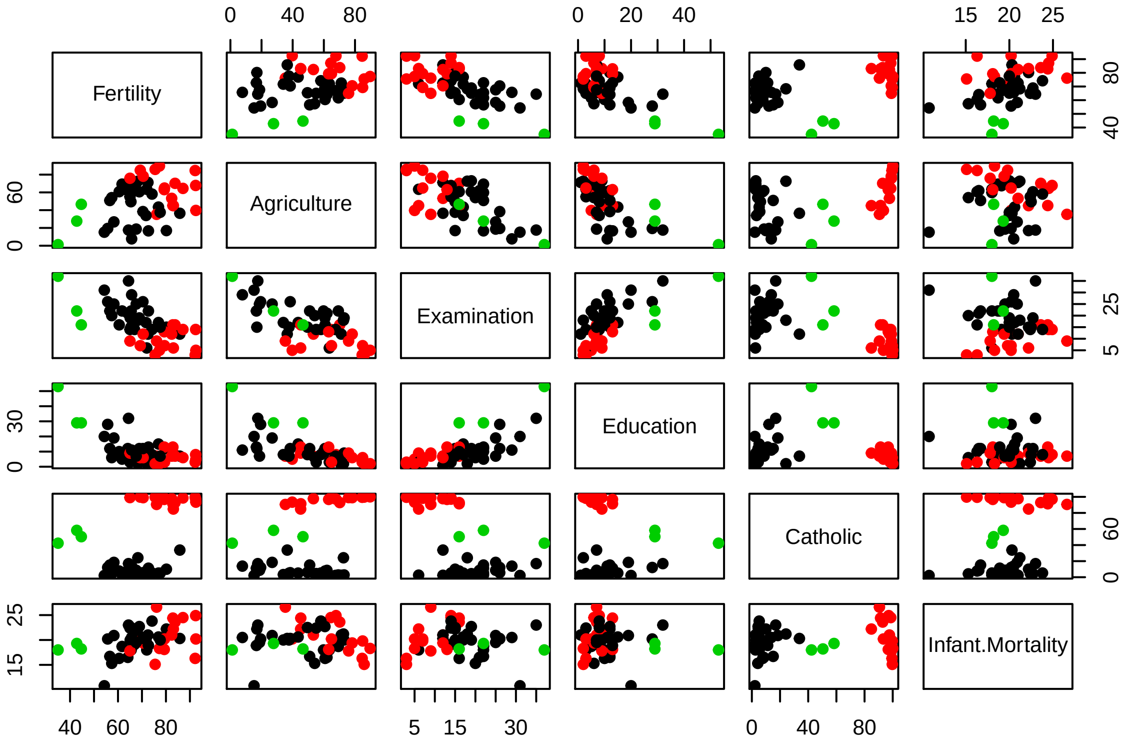


```
plot(outWard)
k2=2
res2=cutree(outWard,k2)
rect2=rect.hclust(outWard,k2)
```

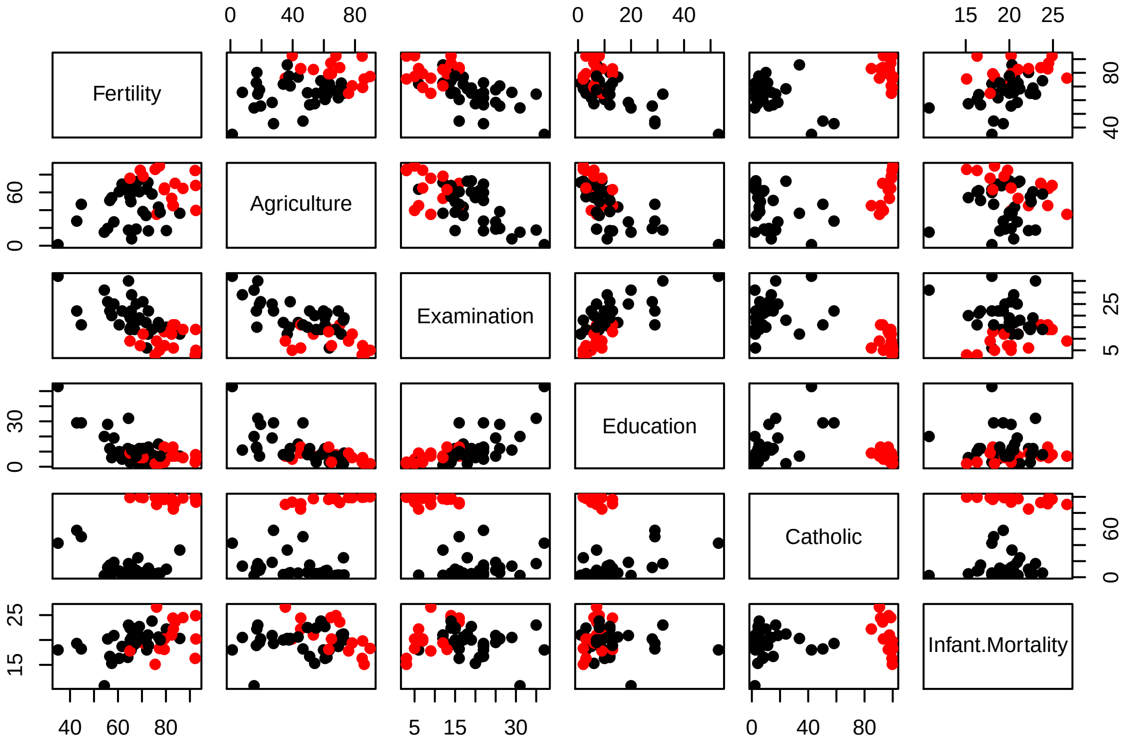


make a pair to see the variable result compaire to the clustering

```
pairs(swiss,col=res1,pch=19)
```



```
pairs(swiss,col=res2,pch=19)
```



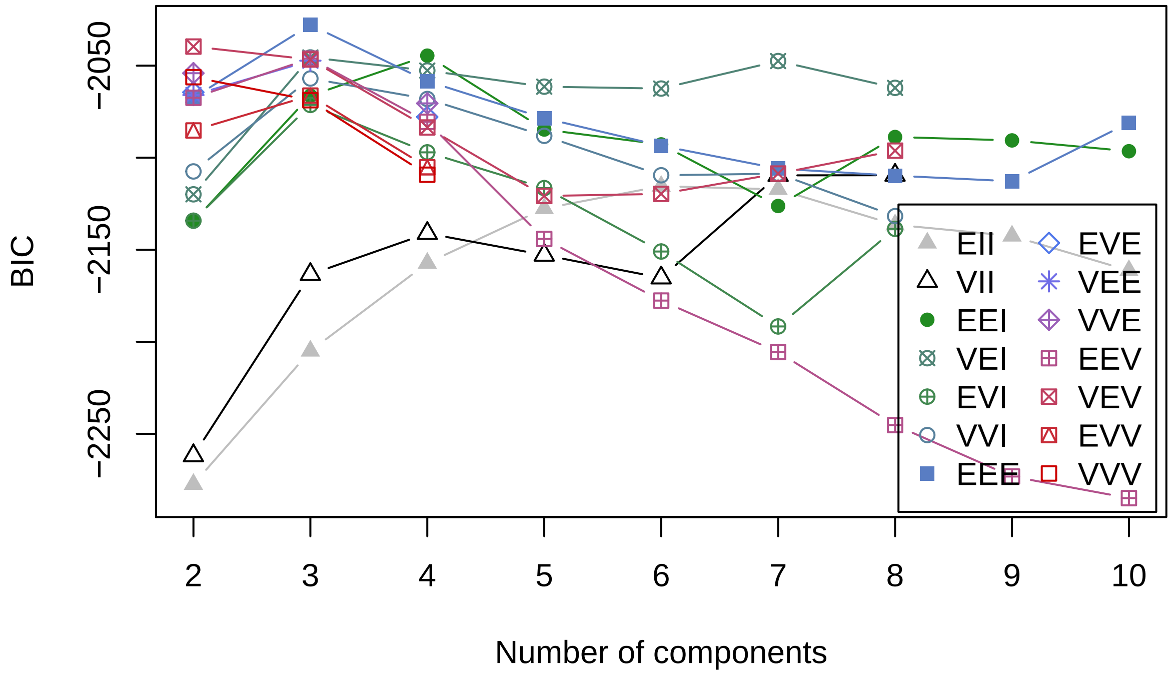
The mixture model and the EM algorithm

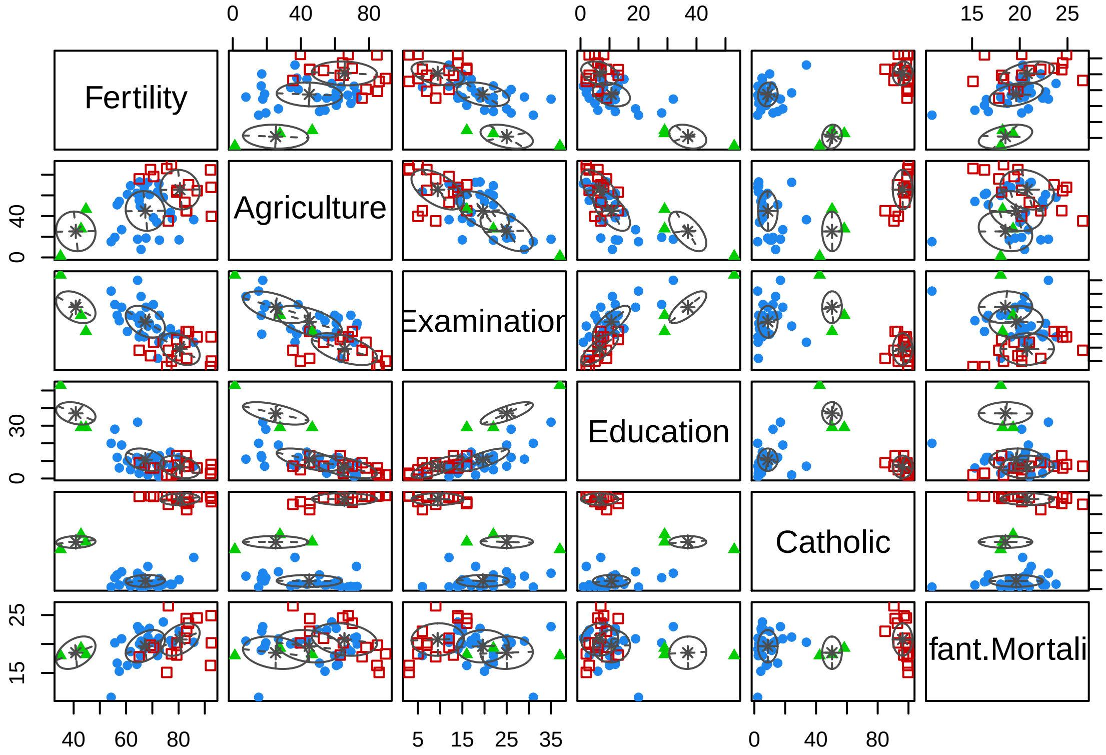
the ‘mclust’ package (Raftery et al.) allow to cluster some data with GGM and the EM algorithm

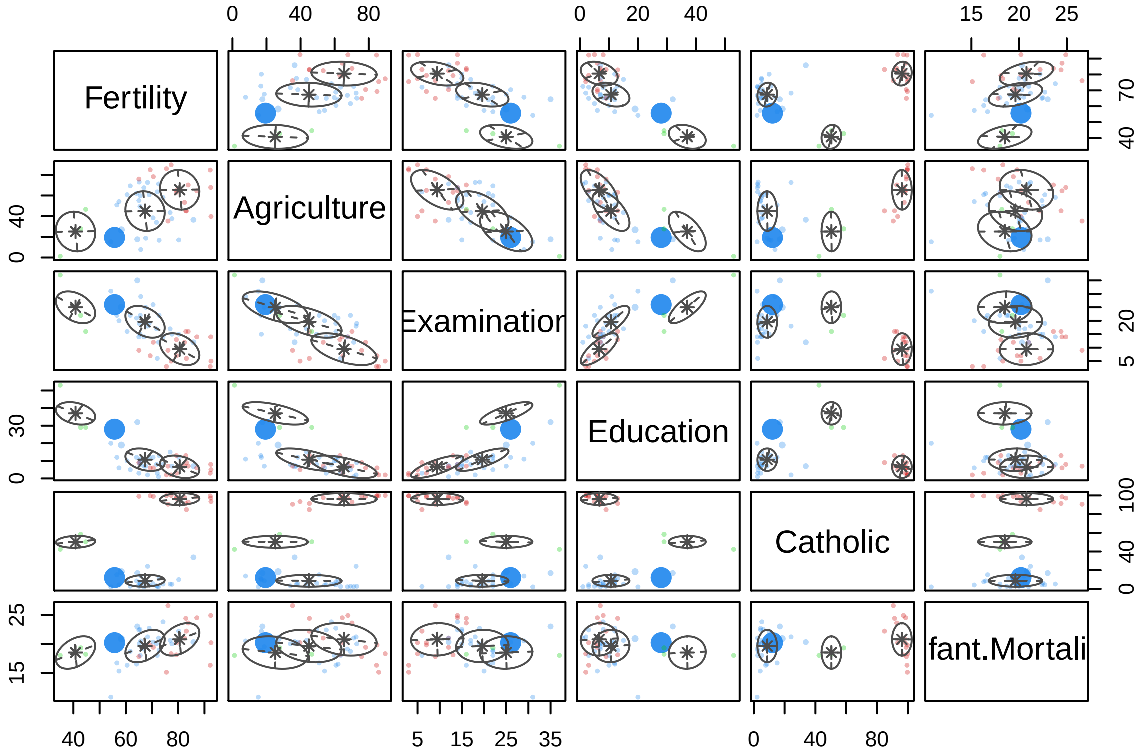
```
#install.packages('mclust')
library(mclust)
```

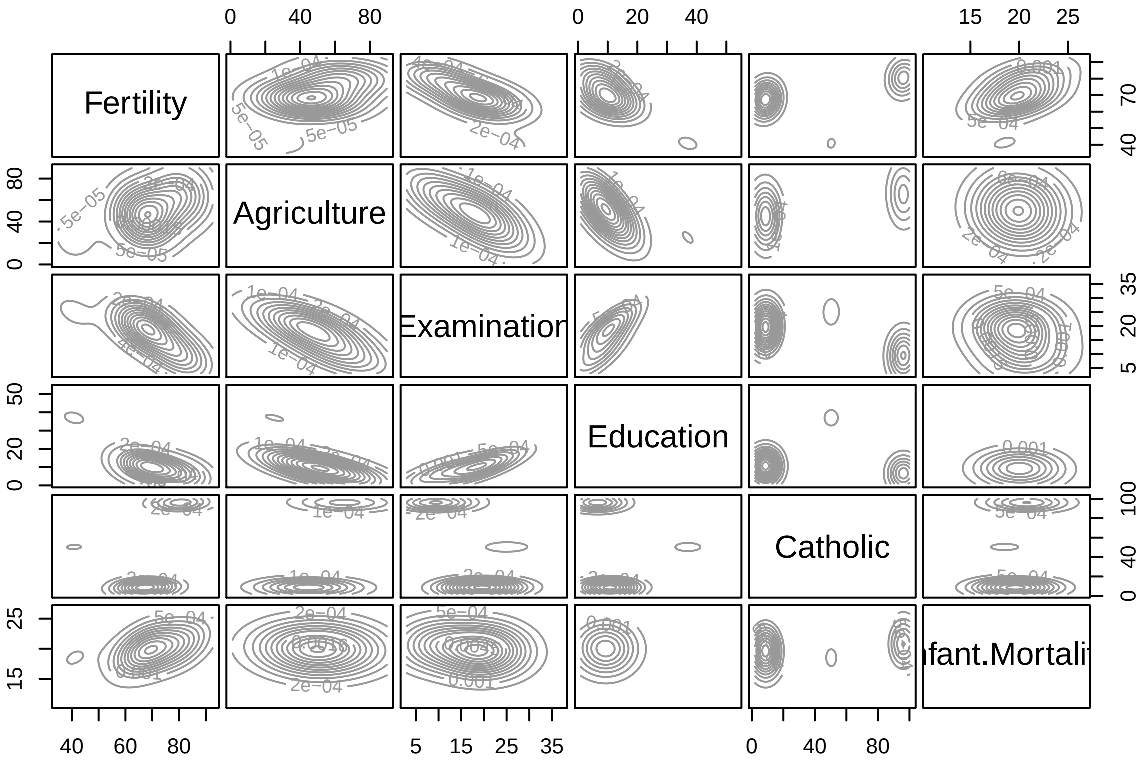
```
## Package 'mclust' version 5.4.3
## Type 'citation("mclust")' for citing this R package in publications.
```

```
data("swiss")
#G=number of group
out=Mclust(swiss, G=2:10)
plot(out)
```


BIC
−2050
−2150
−2250
2
3
4
5
6
7
8
9
10
Number of components
EII
VII
EEI
VEI
EVI
VVI
EEE
EVE
VEE
VVE
EEV
VEV
EVV
VVV


Fertility
Agriculture
Examination
Education
Catholic
fant.Mortali
0
40
80
0
20
40
15
20
25
70
40
0
40
20
5
30
0
100
40
0
25
15
40
60
80
5
15
25
35
0
40
80


Fertility
Agriculture
Examination
Education
Catholic
fant.Mortali
0
40
80
0
20
40
15
20
25
70
40
40
0
20
5
30
0
100
40
0
25
15
40
60
80
5
15
25
35
0
40
80



```
# plot 1:
#get the highest point of BIC (3 groups)
#here best model is EEE, which is exactly K-mean


# plot 2:
# we see that the result is exactly the same as HC (complete)

#plot 3:
# the larger is the point the larger is the uncertainty
```

```
out$modelName
```

```
## [1] "EEE"
```

```
out$parameters$mean
```

```
##                       [,1]     [,2]     [,3]
## Fertility        67.335714 80.55000 40.83333
## Agriculture      44.900000 65.51875 25.16667
## Examination      19.607143  9.43750 25.00000
## Education        10.678571  6.62500 37.00000
## Catholic          8.723571 96.15000 50.36667
## Infant.Mortality 19.621429 20.77500 18.50000
```

```r
out$parameters$pro
```

```
## [1] 0.59574468 0.34042553 0.06382979
```

```r
out$parameters$variance
```

```
## $modelName
## [1] "EEE"
## 
## $d
## [1] 6
## 
## $G
## [1] 3
## 
## $sigma
## , , 1
## 
##                   Fertility Agriculture Examination    Education
## Fertility         56.324063  -11.930674 -16.8778115 -14.42933131
## Agriculture      -11.930674  368.415980 -61.5283245 -73.35079786
## Examination      -16.877812  -61.528324  34.9492781  28.40615501
## Education        -14.429331  -73.350798  28.4061550  40.76291793
## Catholic           4.581123   -2.658348   0.7508359   0.02387538
## Infant.Mortality   8.648693  -11.600266   0.7853343   0.84984802
##                     Catholic Infant.Mortality
## Fertility         4.58112258        8.6486930
## Agriculture      -2.65834751      -11.6002660
## Examination       0.75083587        0.7853343
## Education         0.02387538        0.8498480
## Catholic         40.66932150       -0.0764924
## Infant.Mortality -0.07649240        7.8731307
## 
## , , 2
## 
##                   Fertility Agriculture Examination    Education
## Fertility         56.324063  -11.930674 -16.8778115 -14.42933131
## Agriculture      -11.930674  368.415980 -61.5283245 -73.35079786
## Examination      -16.877812  -61.528324  34.9492781  28.40615501
## Education        -14.429331  -73.350798  28.4061550  40.76291793
## Catholic           4.581123   -2.658348   0.7508359   0.02387538
## Infant.Mortality   8.648693  -11.600266   0.7853343   0.84984802
##                     Catholic Infant.Mortality
## Fertility         4.58112258        8.6486930
## Agriculture      -2.65834751      -11.6002660
## Examination       0.75083587        0.7853343
## Education         0.02387538        0.8498480
## Catholic         40.66932150       -0.0764924
## Infant.Mortality -0.07649240        7.8731307
## 
## , , 3
## 
```

```
##                   Fertility Agriculture Examination    Education
## Fertility         56.324063  -11.930674 -16.8778115 -14.42933131
## Agriculture      -11.930674  368.415980 -61.5283245 -73.35079786
## Examination      -16.877812  -61.528324  34.9492781  28.40615501
## Education        -14.429331  -73.350798  28.4061550  40.76291793
## Catholic           4.581123   -2.658348   0.7508359   0.02387538
## Infant.Mortality   8.648693  -11.600266   0.7853343   0.84984802
##                     Catholic Infant.Mortality
## Fertility         4.58112258        8.6486930
## Agriculture      -2.65834751      -11.6002660
## Examination       0.75083587        0.7853343
## Education         0.02387538        0.8498480
## Catholic         40.66932150       -0.0764924
## Infant.Mortality -0.07649240        7.8731307
##
##
## $Sigma
##                   Fertility Agriculture Examination    Education
## Fertility         56.324063  -11.930674 -16.8778115 -14.42933131
## Agriculture      -11.930674  368.415980 -61.5283245 -73.35079786
## Examination      -16.877812  -61.528324  34.9492781  28.40615501
## Education        -14.429331  -73.350798  28.4061550  40.76291793
## Catholic           4.581123   -2.658348   0.7508359   0.02387538
## Infant.Mortality   8.648693  -11.600266   0.7853343   0.84984802
##                     Catholic Infant.Mortality
## Fertility         4.58112258        8.6486930
## Agriculture      -2.65834751      -11.6002660
## Examination       0.75083587        0.7853343
## Education         0.02387538        0.8498480
## Catholic         40.66932150       -0.0764924
## Infant.Mortality -0.07649240        7.8731307
##
## $cholSigma
##                  Fertility Agriculture Examination Education    Catholic
## Fertility         7.504936    -1.58971   -2.248895 -1.922646  0.61041462
## Agriculture       0.000000    19.12822   -3.403527 -3.994478 -0.08824476
## Examination       0.000000     0.00000   -4.278756 -2.450949 -0.42611701
## Education         0.000000     0.00000    0.000000 -3.886303  0.05130754
## Catholic          0.000000     0.00000    0.000000  0.000000 -6.33282877
## Infant.Mortality  0.000000     0.00000    0.000000  0.000000  0.00000000
##                  Infant.Mortality
## Fertility              1.15240065
## Agriculture           -0.51067390
## Examination           -0.38302481
## Education             -0.02234927
## Catholic               0.15586488
## Infant.Mortality      -2.47241061
```

the Rmix package also allows to use the GGM+EM

```
#install.packages('Rmixmod')
library(Rmixmod)
```

```
## Loading required package: Rcpp
## Rmixmod v. 2.1.2.2 / URI: www.mixmod.org
```

```
out=mixmodCluster(swiss,2:10)
# 2:10 = means that it would choose the best group between it
#default is 1 to 9. if it's 1 it means that there's no need to do clustering
#plot(out) # type in the console
```