430_hw9

Enguang Fan

4/14/2020

1

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.6.2
```

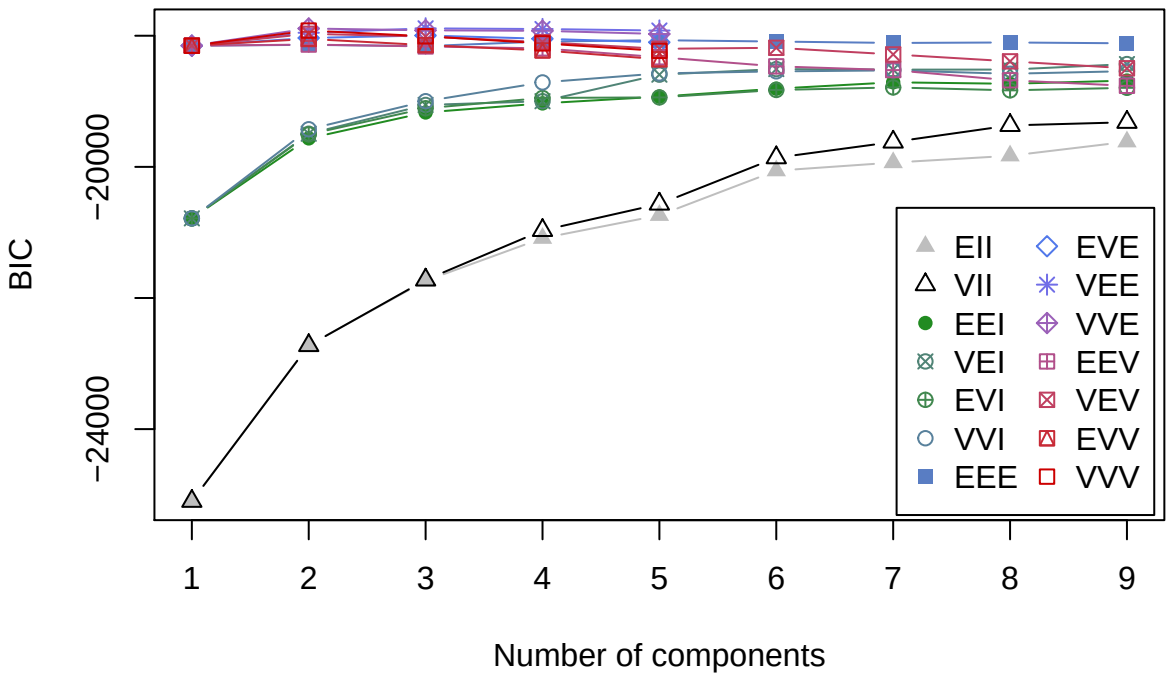
```
library(mclust)
```

```
## Warning: package 'mclust' was built under R version 3.6.3
```

```
## Package 'mclust' version 5.4.5
## Type 'citation("mclust")' for citing this R package in publications.
```

```
data(Hitters)
Hitters=Hitters[,c(1:7)]
Hitters=na.omit(Hitters)
```

```
BIC=mclustBIC(Hitters)
plot(BIC)
```




```
summary(BIC)
```

```
## Best BIC values:
##             VEE,3         VVE,2        VEE,4
## BIC      -17887.3 -17889.643907 -17897.62015
## BIC diff      0.0     -2.342643    -10.31888
```

The VEE assumption with clusters = 3 has the highest BIC

##2

```
mod1=Mclust(Hitters,G=3,modelNames=c("VEE"))
summary(mod1,parameters=TRUE)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEE (ellipsoidal, equal shape and orientation) model with 3 components:
##
##  log-likelihood   n df      BIC       ICL
##       -8790.625 322 53 -17887.3 -17929.44
##
## Clustering table:
##   1   2   3
```

```
##  90  42 190
##
## Mixing probabilities:
##         1         2         3
## 0.2756318 0.1341151 0.5902531
##
## Means:
##             [,1]      [,2]       [,3]
## AtBat 222.519309 260.30951 482.307917
## Hits   55.379851  67.88255 129.870260
## HmRun   4.924292   6.66402  14.433045
## Runs   25.915239  31.15061  67.071410
## RBI    24.576176  33.56510  62.265474
## Walks  19.864074  25.47573  50.572181
## Years   4.002625  14.21989   7.511602
##
## Variances:
## [,,1]
##              AtBat        Hits      HmRun       Runs        RBI      Walks
## AtBat 4017.617015 1165.645903 98.5993866 541.905650 511.343920 243.148769
## Hits  1165.645903  386.970464 28.9314407 175.863389 159.222044  70.097084
## HmRun   98.599387   28.931441 19.6643820  22.200037  41.776038  11.368800
## Runs   541.905650  175.863389 22.2000374 114.656412  80.776030  53.908211
## RBI    511.343920  159.222044 41.7760377  80.776030 145.521109  44.832168
## Walks  243.148769   70.097084 11.3687998  53.908211  44.832168  98.251513
## Years   -1.832119   -1.018391  0.9533656  -2.237751   3.274865   2.867119
##            Years
## AtBat -1.8321189
## Hits  -1.0183909
## HmRun  0.9533656
## Runs  -2.2377512
## RBI    3.2748651
## Walks  2.8671190
## Years  6.0844751
## [,,2]
##              AtBat        Hits      HmRun       Runs        RBI      Walks
## AtBat 4543.366021 1318.183383 111.502192 612.819915 578.258849 274.967437
## Hits  1318.183383  437.609770  32.717435 198.877032 180.057985  79.270052
## HmRun  111.502192   32.717435  22.237681  25.105154  47.242888  12.856531
## Runs   612.819915  198.877032  25.105154 129.660454  91.346454  60.962689
## RBI    578.258849  180.057985  47.242888  91.346454 164.564134  50.698946
## Walks  274.967437   79.270052  12.856531  60.962689  50.698946 111.108795
## Years   -2.071872   -1.151658   1.078124  -2.530585   3.703417   3.242313
##            Years
## AtBat -2.071872
## Hits  -1.151658
## HmRun  1.078124
## Runs  -2.530585
## RBI    3.703417
## Walks  3.242313
## Years  6.880695
## [,,3]
##               AtBat        Hits      HmRun        Runs        RBI     Walks
## AtBat 14721.208735 4271.118076 361.284350 1985.631322 1873.64812 890.93703
```

```
## Hits   4271.118076 1417.923349 106.009551  644.392348  583.41572 256.84723
## HmRun   361.284350  106.009551  72.053526   81.344584  153.07426  41.65715
## Runs   1985.631322  644.392348  81.344584  420.119929  295.97664 197.52854
## RBI    1873.648122  583.415724 153.074265  295.976642  533.21325 164.27243
## Walks   890.937032  256.847230  41.657150  197.528542  164.27243 360.00968
## Years    -6.713185   -3.731552   3.493288   -8.199488   11.99964  10.50559
##           Years
## AtBat -6.713185
## Hits  -3.731552
## HmRun  3.493288
## Runs  -8.199488
## RBI   11.999644
## Walks 10.505595
## Years 22.294516
```

```
names(mod1)
```

```
##  [1] "call"           "data"           "modelName"      "n"
##  [5] "d"              "G"              "BIC"            "bic"
##  [9] "loglik"         "df"             "hypvol"         "parameters"
## [13] "z"              "classification" "uncertainty"
```

```
cluster1 = subset(Hitters,mod1$classification ==1)
cluster2 = subset(Hitters,mod1$classification ==2)
cluster3 = subset(Hitters,mod1$classification ==3)
```

```
cov(cluster1)
```

```
##            AtBat        Hits       HmRun       Runs        RBI      Walks
## AtBat 6582.29663 1732.166292 132.2382022 757.737079 785.110112 408.197753
## Hits  1732.16629  489.149688  39.0531835 213.467915 224.029838 107.057803
## HmRun  132.23820   39.053184  16.2876404  20.982772  39.435206  11.780524
## Runs   757.73708  213.467915  20.9827715 117.587516  98.868664  55.823471
## RBI    785.11011  224.029838  39.4352060  98.868664 168.752684  62.997878
## Walks  408.19775  107.057803  11.7805243  55.823471  62.997878  82.896504
## Years   12.50787    5.005119   0.3363296   1.246192   2.952684   2.027091
##            Years
## AtBat 12.5078652
## Hits   5.0051186
## HmRun  0.3363296
## Runs   1.2461923
## RBI    2.9526841
## Walks  2.0270911
## Years  5.2920100
```

```
cov(cluster2)
```

```
##            AtBat        Hits      HmRun       Runs        RBI      Walks
## AtBat 3710.46690 1080.804878 26.5888502 421.459930 393.703833 217.658537
## Hits  1080.80488  354.630081 10.4552846 137.804878 124.195122  62.065041
## HmRun   26.58885   10.455285 11.9140534  14.337979  17.728223  12.227642
## Runs   421.45993  137.804878 14.3379791  98.613240  48.710801  55.292683
```

```
## RBI    393.70383  124.195122 17.7282230  48.710801  95.344948  29.317073
## Walks  217.65854   62.065041 12.2276423  55.292683  29.317073  99.056911
## Years  -14.95470   -4.333333 -0.9140534  -7.167247   1.491289   2.723577
##             Years
## AtBat -14.9547038
## Hits   -4.3333333
## HmRun  -0.9140534
## Runs   -7.1672474
## RBI     1.4912892
## Walks   2.7235772
## Years  10.3530778
```

```
cov(cluster3)
```

```
##            AtBat        Hits      HmRun        Runs        RBI     Walks
## AtBat 9785.33949 3199.669173 328.190086 1548.471206 1464.93378 440.56452
## Hits  3199.66917 1236.619326  96.338346  571.829017  497.37900 150.62016
## HmRun  328.19009   96.338346  84.378056   86.437650  175.86862  33.98195
## Runs  1548.47121  571.829017  86.437650  403.871930  280.28577 168.19332
## RBI   1464.93378  497.379003 175.868616  280.285770  547.10677 122.25464
## Walks  440.56452  150.620162  33.981955  168.193317  122.25464 369.17694
## Years  -10.97845   -6.264272   5.399722   -7.241882   14.88271  11.57889
##            Years
## AtBat -10.978446
## Hits   -6.264272
## HmRun   5.399722
## Runs   -7.241882
## RBI    14.882707
## Walks  11.578892
## Years  19.647898
```

```
summary(mod1,parameters=TRUE)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEE (ellipsoidal, equal shape and orientation) model with 3 components:
##
##  log-likelihood   n df      BIC       ICL
##       -8790.625 322 53 -17887.3 -17929.44
##
## Clustering table:
##   1   2   3
##  90  42 190
##
## Mixing probabilities:
##         1         2         3
## 0.2756318 0.1341151 0.5902531
##
## Means:
##             [,1]      [,2]       [,3]
## AtBat 222.519309 260.30951 482.307917
```

```
## Hits   55.379851  67.88255 129.870260
## HmRun   4.924292   6.66402  14.433045
## Runs   25.915239  31.15061  67.071410
## RBI    24.576176  33.56510  62.265474
## Walks  19.864074  25.47573  50.572181
## Years   4.002625  14.21989   7.511602
##
## Variances:
## [,,1]
##             AtBat        Hits      HmRun       Runs        RBI      Walks
## AtBat 4017.617015 1165.645903 98.5993866 541.905650 511.343920 243.148769
## Hits  1165.645903  386.970464 28.9314407 175.863389 159.222044  70.097084
## HmRun   98.599387   28.931441 19.6643820  22.200037  41.776038  11.368800
## Runs   541.905650  175.863389 22.2000374 114.656412  80.776030  53.908211
## RBI    511.343920  159.222044 41.7760377  80.776030 145.521109  44.832168
## Walks  243.148769   70.097084 11.3687998  53.908211  44.832168  98.251513
## Years   -1.832119   -1.018391  0.9533656  -2.237751   3.274865   2.867119
##            Years
## AtBat -1.8321189
## Hits  -1.0183909
## HmRun  0.9533656
## Runs  -2.2377512
## RBI    3.2748651
## Walks  2.8671190
## Years  6.0844751
## [,,2]
##             AtBat        Hits      HmRun       Runs        RBI      Walks
## AtBat 4543.366021 1318.183383 111.502192 612.819915 578.258849 274.967437
## Hits  1318.183383  437.609770  32.717435 198.877032 180.057985  79.270052
## HmRun  111.502192   32.717435  22.237681  25.105154  47.242888  12.856531
## Runs   612.819915  198.877032  25.105154 129.660454  91.346454  60.962689
## RBI    578.258849  180.057985  47.242888  91.346454 164.564134  50.698946
## Walks  274.967437   79.270052  12.856531  60.962689  50.698946 111.108795
## Years   -2.071872   -1.151658   1.078124  -2.530585   3.703417   3.242313
##           Years
## AtBat -2.071872
## Hits  -1.151658
## HmRun  1.078124
## Runs  -2.530585
## RBI    3.703417
## Walks  3.242313
## Years  6.880695
## [,,3]
##              AtBat        Hits      HmRun        Runs        RBI     Walks
## AtBat 14721.208735 4271.118076 361.284350 1985.631322 1873.64812 890.93703
## Hits   4271.118076 1417.923349 106.009551  644.392348  583.41572 256.84723
## HmRun   361.284350  106.009551  72.053526   81.344584  153.07426  41.65715
## Runs   1985.631322  644.392348  81.344584  420.119929  295.97664 197.52854
## RBI    1873.648122  583.415724 153.074265  295.976642  533.21325 164.27243
## Walks   890.937032  256.847230  41.657150  197.528542  164.27243 360.00968
## Years    -6.713185   -3.731552   3.493288   -8.199488   11.99964  10.50559
##           Years
## AtBat -6.713185
## Hits  -3.731552
```

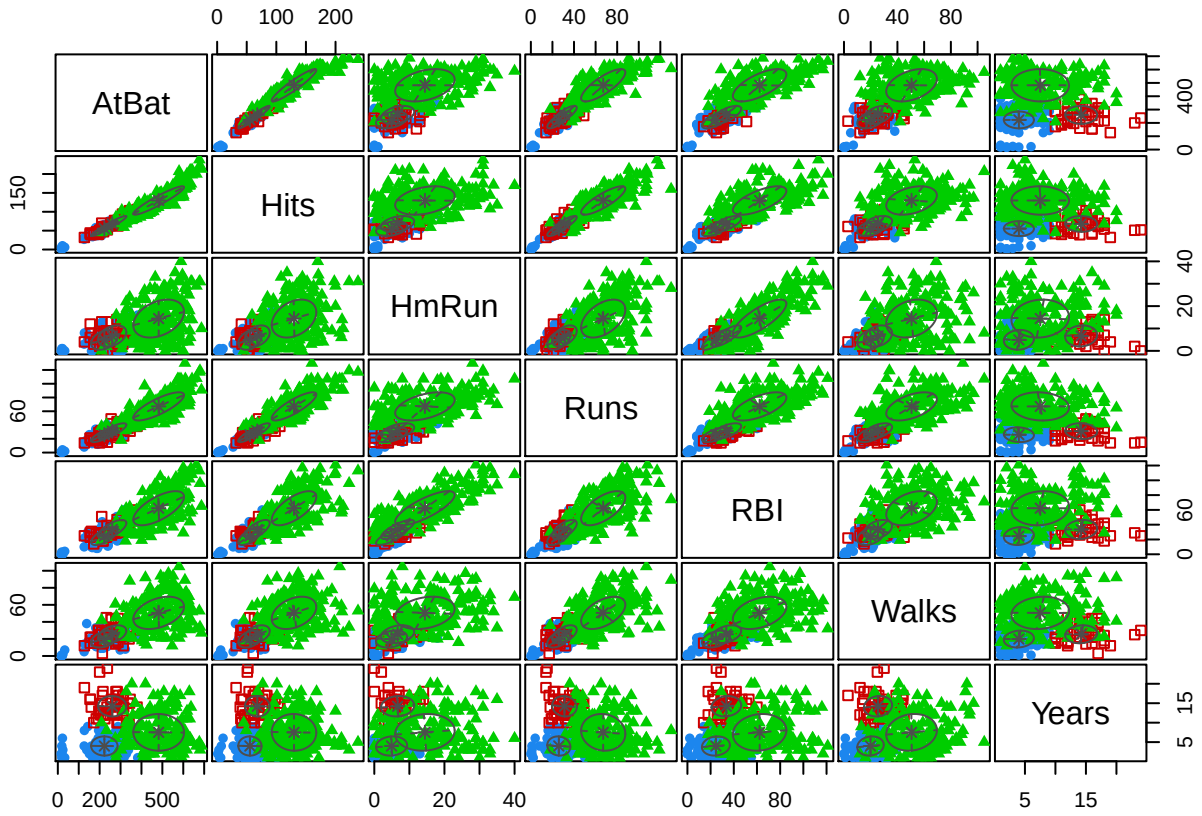
```
## HmRun  3.493288
## Runs  -8.199488
## RBI   11.999644
## Walks 10.505595
## Years 22.294516
```

Conclusion: The variance matrix from first and second cluster are quite similar yet the third cluster's variance matrix is somewhat different with other clusters. This is consistent to what we we observe from the scatter plot from #3, the red and blue clusters have similar variance while green cluster has much bigger variance.

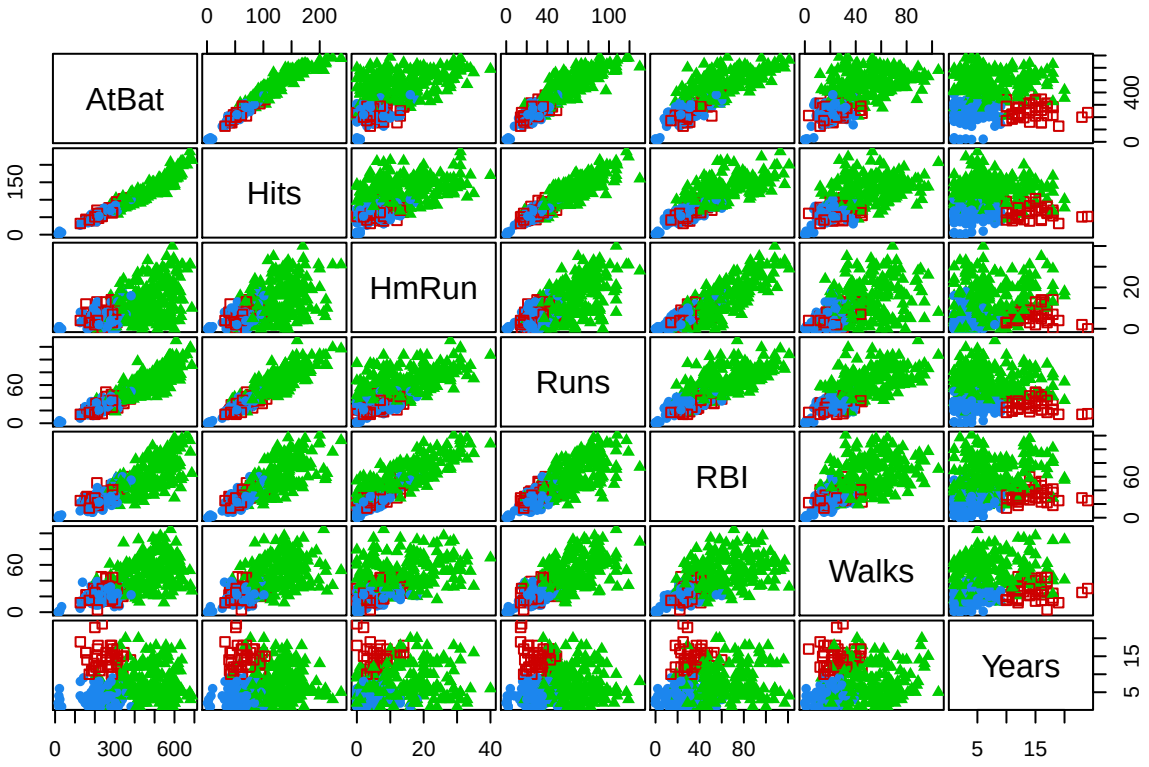
Interestingly, we find that the variance matrix derived from calculating each cluster separately is slightly different to the one we get from summary table, it could due to the rounding problem, hence we still use the one we get from summary table(though the mean value is the same).

##3

```
plot(mod1,what = "classification")
```



```
clPairs(Hitters[1:7],mod1$classification)
```



The scatter plot provide us with some intuitive evidence about how the clustes are classified.

The blue clusters represent the players who play relatively less years(around 0~10years), also their performance in 1986 are mediocre, the bat number, hit number, run and homerun number are less than other clusters. In all, they are just mediocre players in league.

The red clusters represent the players who play much longer years than other two cluster(around 10~20 years), yet their performance are not so good, with quite similar performance in 1986 with the blue clusters in term of bat number, hit number, run and homerun number . In all, these play longer time in league but has normal performance in 1986, possibly because the physical problems like injury caused by aging problem.

The green cluster has most players, who has best performance based on their bat number, hit number, run and homerun numbers. Most of them has not very long career(around 0-10years) yet some of them play a much longer career around 15 years. This make sense cause the peak of a player will not last very long yet some of star players did have better performance and longer career.

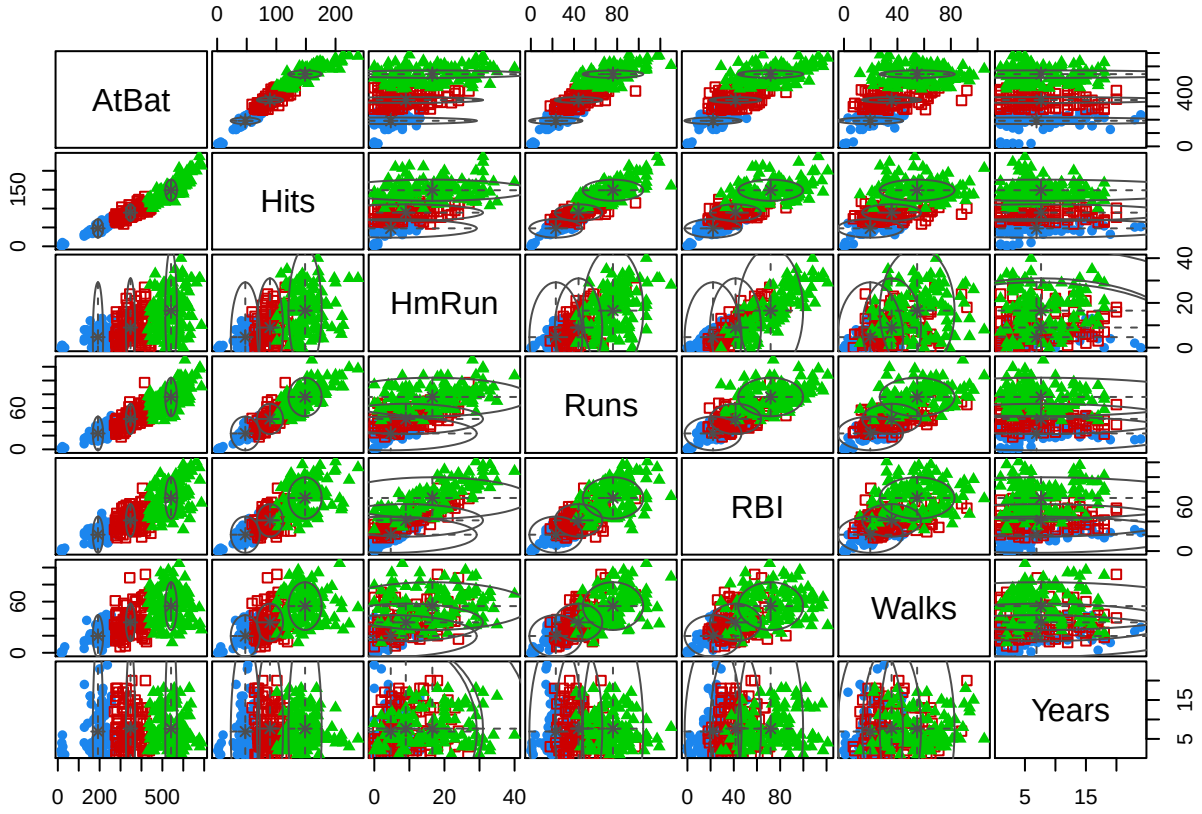
##4

```
mod2=Mclust(Hitters,G=3,modelNames=c("VII"))
summary(mod2,parameters=TRUE)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VII (spherical, varying volume) model with 3 components:
```

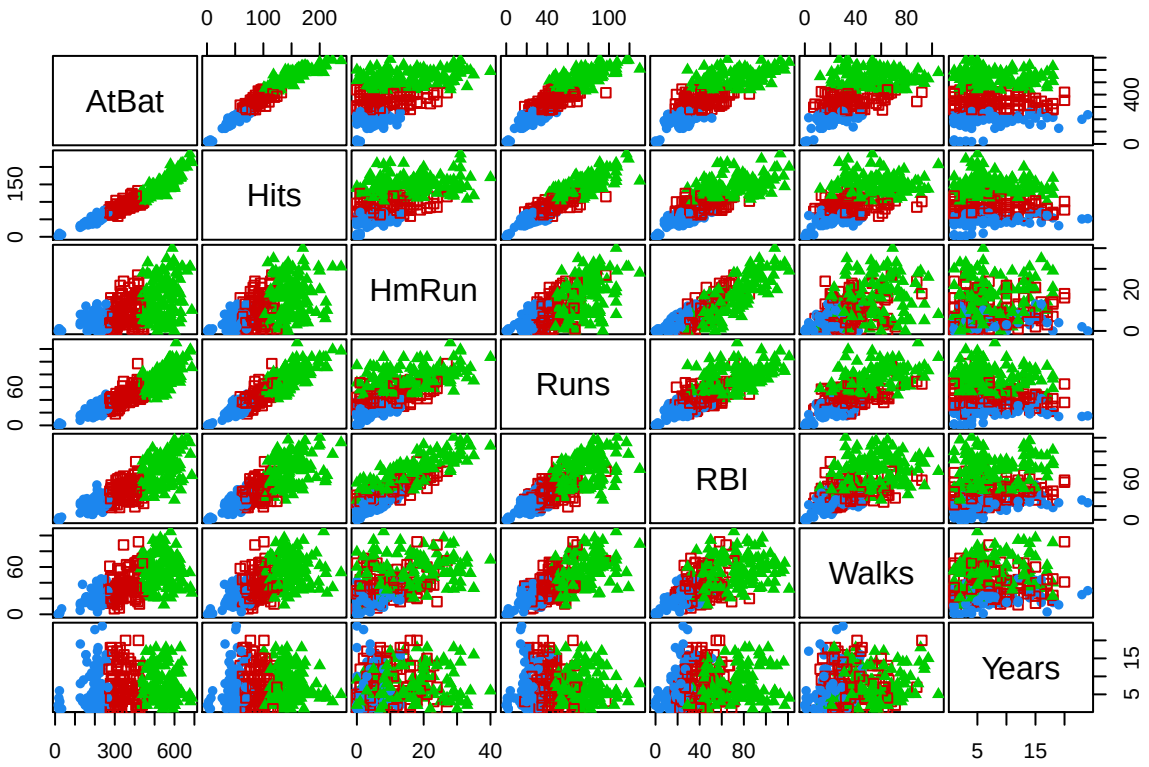
```
##
##  log-likelihood   n df       BIC       ICL
##       -10786.97 322 26 -21724.07 -21732.61
##
## Clustering table:
##   1   2   3
##  87 109 126
##
## Mixing probabilities:
##         1         2         3
## 0.2740905 0.3346137 0.3912958
##
## Means:
##             [,1]       [,2]       [,3]
## AtBat 192.009065 347.716614 541.661732
## Hits   47.571054  89.236127 148.548587
## HmRun   4.694030   8.983489  16.554225
## Runs   23.129395  44.198340  76.108712
## RBI    22.112778  41.491413  71.770389
## Walks  19.656004  35.658069  54.748946
## Years   6.897593   7.663883   7.638963
##
## Variances:
## [,,1]
##          AtBat     Hits    HmRun     Runs      RBI    Walks    Years
## AtBat 601.7438   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
## Hits    0.0000 601.7438   0.0000   0.0000   0.0000   0.0000   0.0000
## HmRun   0.0000   0.0000 601.7438   0.0000   0.0000   0.0000   0.0000
## Runs    0.0000   0.0000   0.0000 601.7438   0.0000   0.0000   0.0000
## RBI     0.0000   0.0000   0.0000   0.0000 601.7438   0.0000   0.0000
## Walks   0.0000   0.0000   0.0000   0.0000   0.0000 601.7438   0.0000
## Years   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 601.7438
## [,,2]
##          AtBat     Hits    HmRun     Runs      RBI    Walks    Years
## AtBat 484.4234   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
## Hits    0.0000 484.4234   0.0000   0.0000   0.0000   0.0000   0.0000
## HmRun   0.0000   0.0000 484.4234   0.0000   0.0000   0.0000   0.0000
## Runs    0.0000   0.0000   0.0000 484.4234   0.0000   0.0000   0.0000
## RBI     0.0000   0.0000   0.0000   0.0000 484.4234   0.0000   0.0000
## Walks   0.0000   0.0000   0.0000   0.0000   0.0000 484.4234   0.0000
## Years   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 484.4234
## [,,3]
##          AtBat     Hits    HmRun     Runs      RBI    Walks    Years
## AtBat 785.1825   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000
## Hits    0.0000 785.1825   0.0000   0.0000   0.0000   0.0000   0.0000
## HmRun   0.0000   0.0000 785.1825   0.0000   0.0000   0.0000   0.0000
## Runs    0.0000   0.0000   0.0000 785.1825   0.0000   0.0000   0.0000
## RBI     0.0000   0.0000   0.0000   0.0000 785.1825   0.0000   0.0000
## Walks   0.0000   0.0000   0.0000   0.0000   0.0000 785.1825   0.0000
## Years   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000 785.1825
```

```
plot(mod2,what = "classification")
```



1. The number of three clusters are very similar, which is 87 109 126. 2. The covariance between different variables are all 0, this is because of the VII model 3. In separate clusters, different variables have same variance, making diagnol variance matrix, this is because of the VII model

```
clPairs(Hitters[1:7],mod2$classification)
```



##5

```
mod2$bic
```

```
## [1] -21724.07
```

```
mod1$bic
```

```
## [1] -17887.3
```

```
mod1$df
```

```
## [1] 53
```

```
mod2$df
```

```
## [1] 26
```

The new model has smaller df than best model, so best model has fewer parameters.