Principal Component Analysis (PCA) with FactoMineR (decathlon dataset)

François Husson & Magalie Houée-Bigot

Import data (data are imported from internet)

```
decathlon <- read.table("http://www.agrocampus-ouest.fr/math/RforStat/decathlon.csv",
      header=TRUE, sep=";", row.names=1, check.names=FALSE)
```

`header=TRUE` : indicates that the file contains the names of the variables

`sep=";"` : indicates the fields separator (usually ";" or "," for csv files)

`row.names=1` : indicates the column of the table which contains the row names

`check.names=FALSE` : indicated that the names of the variables in the data frame are unchecked

It is important to check that the import is well done

```
summary(decathlon)
```

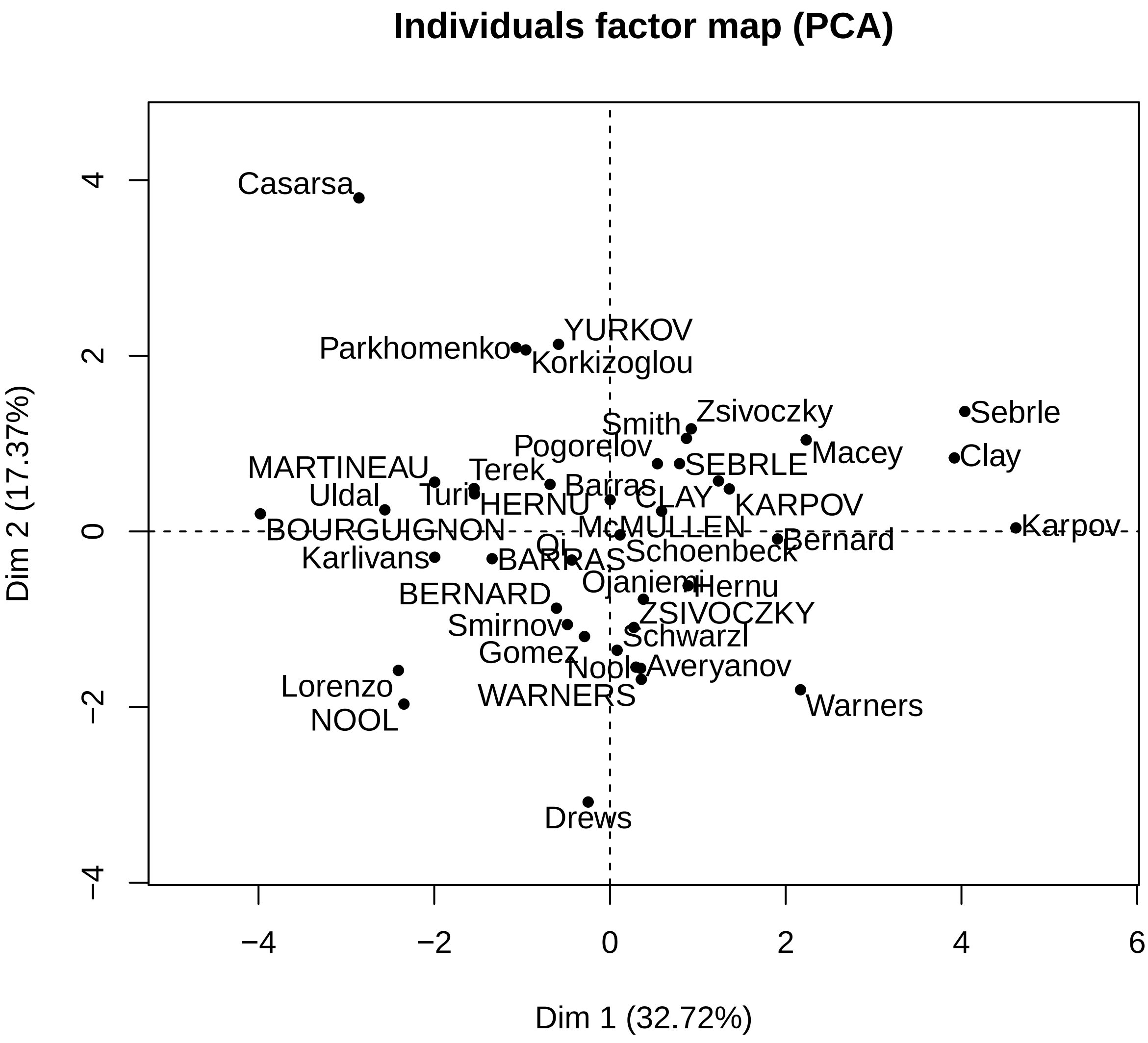
```
##       100m          Long jump       Shot put       High jump
##  Min.   :10.44   Min.   :6.61   Min.   :12.68   Min.   :1.850
##  1st Qu.:10.85   1st Qu.:7.03   1st Qu.:13.88   1st Qu.:1.920
##  Median :10.98   Median :7.30   Median :14.57   Median :1.950
##  Mean   :11.00   Mean   :7.26   Mean   :14.48   Mean   :1.977
##  3rd Qu.:11.14   3rd Qu.:7.48   3rd Qu.:14.97   3rd Qu.:2.040
##  Max.   :11.64   Max.   :7.96   Max.   :16.36   Max.   :2.150
##       400m           110m H          Discus        Pole vault
##  Min.   :46.81   Min.   :13.97   Min.   :37.92   Min.   :4.200
##  1st Qu.:48.93   1st Qu.:14.21   1st Qu.:41.90   1st Qu.:4.500
##  Median :49.40   Median :14.48   Median :44.41   Median :4.800
##  Mean   :49.62   Mean   :14.61   Mean   :44.33   Mean   :4.762
##  3rd Qu.:50.30   3rd Qu.:14.98   3rd Qu.:46.07   3rd Qu.:4.920
##  Max.   :53.20   Max.   :15.67   Max.   :51.65   Max.   :5.400
##     Javeline          1500m            Rank          Points
##  Min.   :50.31   Min.   :262.1   Min.   : 1.00   Min.   :7313
##  1st Qu.:55.27   1st Qu.:271.0   1st Qu.: 6.00   1st Qu.:7802
##  Median :58.36   Median :278.1   Median :11.00   Median :8021
##  Mean   :58.32   Mean   :279.0   Mean   :12.12   Mean   :8005
##  3rd Qu.:60.89   3rd Qu.:285.1   3rd Qu.:18.00   3rd Qu.:8122
##  Max.   :70.52   Max.   :317.0   Max.   :28.00   Max.   :8893
##    Competition
##  Decastar:13
##  OlympicG:28
##
##
##
##
```

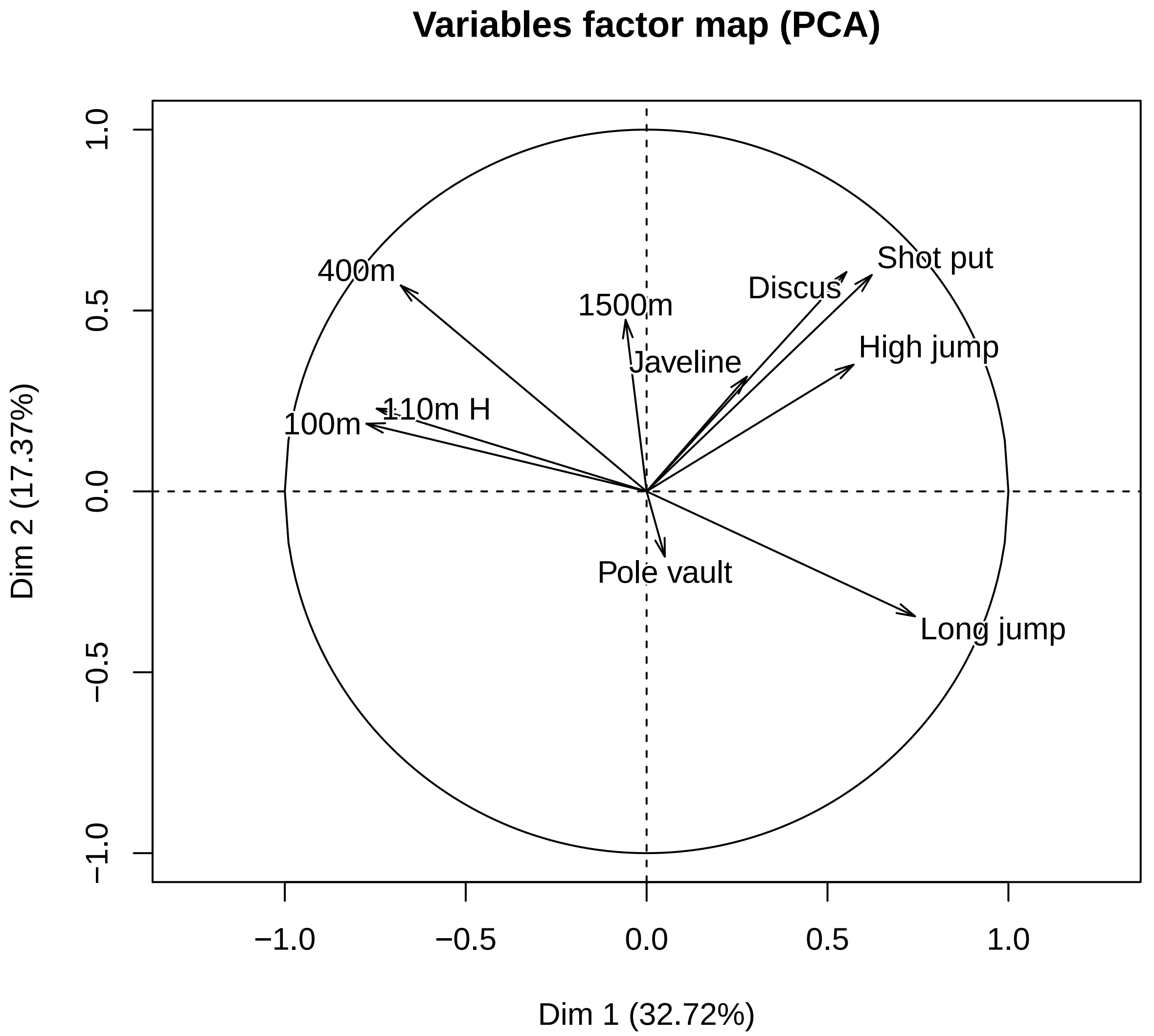
Loading FactoMineR

```
library(FactoMineR)
```

PCA with only active elements as active

```
res <- PCA(decathlon[,1:10])
```





Outputs can be summarized with the function `summary`.

```
summary(res)
```

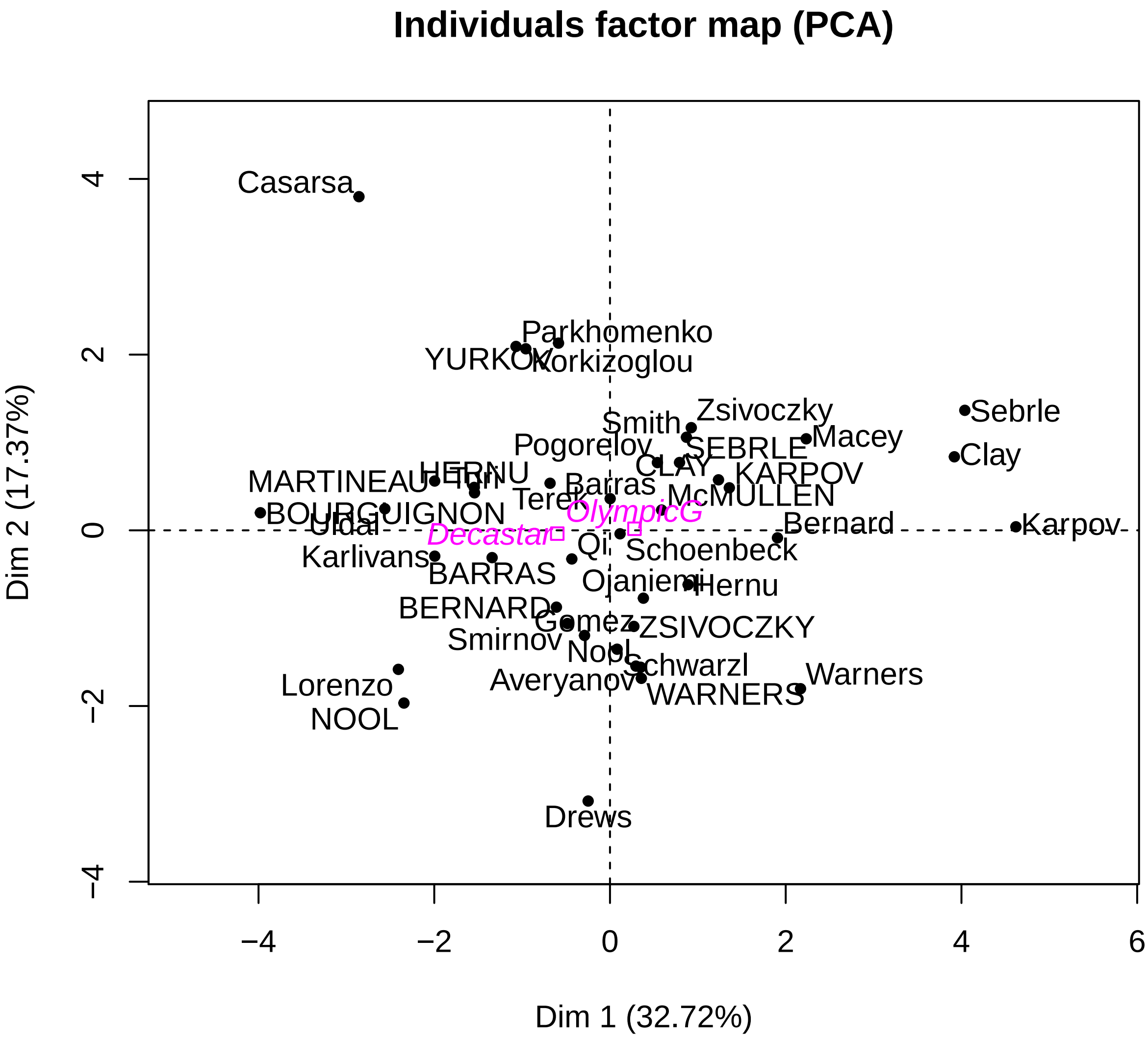
Outputs are given for the first 2 dimensions (by default 3 dimensions are given).

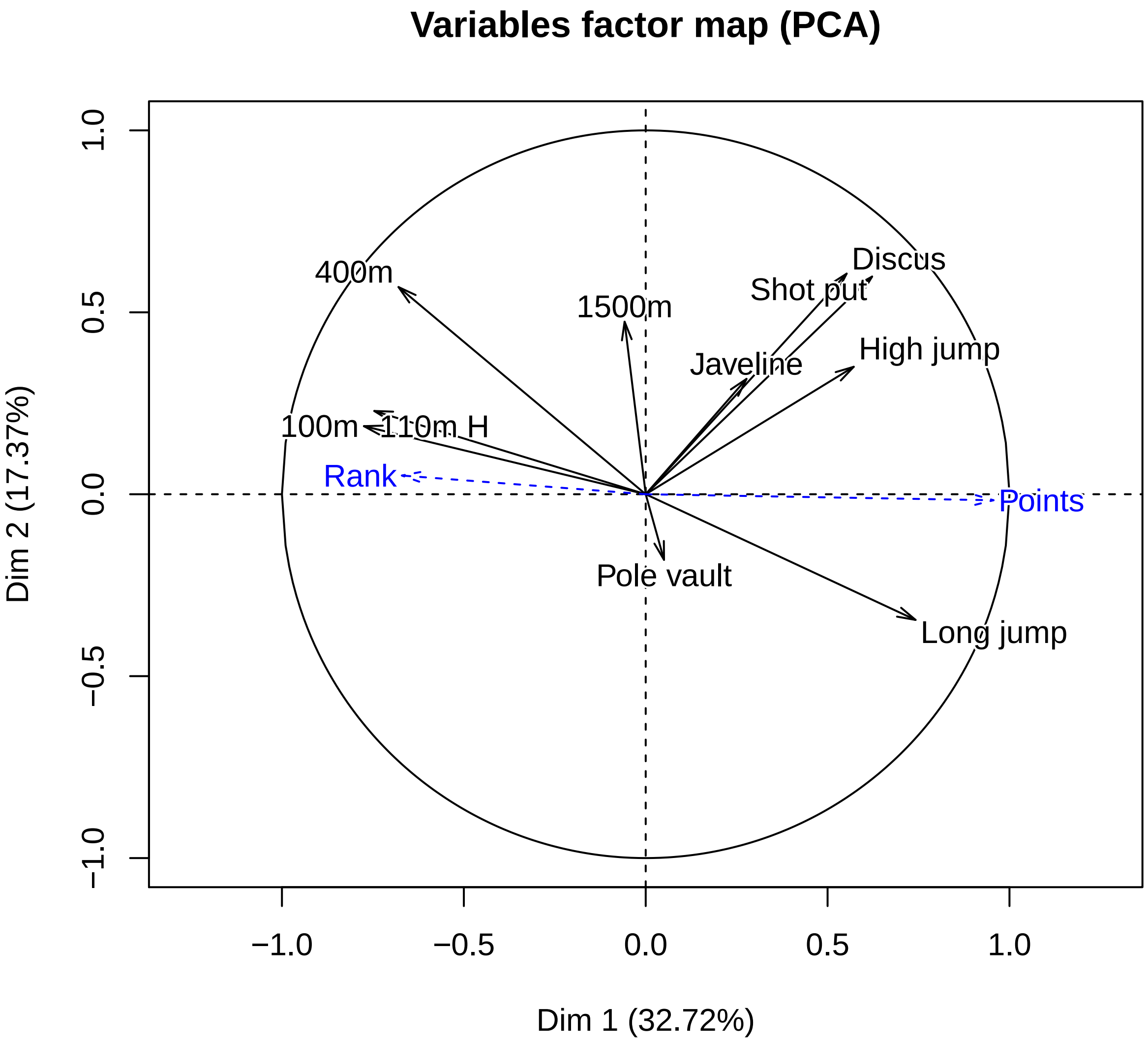
```
summary(res, ncp=2)
```

```
##
## Call:
## PCA(X = decathlon[, 1:10])
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
## Variance               3.272   1.737   1.405   1.057   0.685   0.599
## % of var.             32.719  17.371  14.049  10.569   6.848   5.993
## Cumulative % of var.  32.719  50.090  64.140  74.708  81.556  87.548
##                        Dim.7   Dim.8   Dim.9  Dim.10
## Variance               0.451   0.397   0.215   0.182
## % of var.              4.512   3.969   2.148   1.822
## Cumulative % of var.  92.061  96.030  98.178 100.000
##
## Individuals (the 10 first)
##                Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2  
## Sebrle     |  4.843 |  4.038 12.158  0.695 |  1.366  2.619  0.080 |
## Clay       |  4.647 |  3.919 11.451  0.711 |  0.837  0.984  0.032 |
## Karpov     |  5.006 |  4.620 15.911  0.852 |  0.040  0.002  0.000 |
## Macey      |  3.434 |  2.233  3.719  0.423 |  1.042  1.524  0.092 |
## Warners    |  2.979 |  2.168  3.505  0.530 | -1.803  4.565  0.366 |
## Zsivoczky  |  2.566 |  0.925  0.638  0.130 |  1.169  1.918  0.207 |
## Hernu      |  1.824 |  0.889  0.589  0.238 | -0.618  0.537  0.115 |
## Nool       |  3.098 |  0.295  0.065  0.009 | -1.546  3.354  0.249 |
## Bernard    |  2.827 |  1.906  2.709  0.455 | -0.086  0.010  0.001 |
## Schwarzl   |  1.971 |  0.081  0.005  0.002 | -1.353  2.572  0.472 |
##
## Variables
##               Dim.1    ctr   cos2    Dim.2    ctr   cos2  
## 100m        | -0.775 18.344  0.600 |  0.187  2.016  0.035 |
## Long jump   |  0.742 16.822  0.550 | -0.345  6.869  0.119 |
## Shot put    |  0.623 11.844  0.388 |  0.598 20.607  0.358 |
## High jump   |  0.572  9.998  0.327 |  0.350  7.064  0.123 |
## 400m        | -0.680 14.116  0.462 |  0.569 18.666  0.324 |
## 110m H      | -0.746 17.020  0.557 |  0.229  3.013  0.052 |
## Discus      |  0.552  9.328  0.305 |  0.606 21.162  0.368 |
## Pole vault  |  0.050  0.077  0.003 | -0.180  1.873  0.033 |
## Javeline    |  0.277  2.347  0.077 |  0.317  5.784  0.100 |
## 1500m       | -0.058  0.103  0.003 |  0.474 12.946  0.225 |
```

PCA with supplementary variables

```r
res <- PCA(decathlon, quanti.sup=11:12, quali.sup=13)
```


Individuals factor map (PCA)
Casarsa
Parkhomenko
YURKOV
Korkizoglou
Zsivoczky
Sebrle
Smith
Macey
Pogorelov
SEBRLE
Clay
CLAY
HERNU
KARPOV
MARTINEAU
Turi
Barras
Terek
McMULLEN
OlympicG
BOURGUIGNON
Uldal
Bernard
Karpov
Decastar
Qi
Schoenbeck
Karlivans
BARRAS
Ojaniemi
Hernu
BERNARD
Gomez
ZSIVOCZKY
Smirnov
Nool
Schwarzl
Warners
Lorenzo
Averyanov
WARNERS
NOOL
Drews
4
2
0
−2
−4
−4
−2
0
2
4
6
Dim 1 (32.72%)
Dim 2 (17.37%)



```
summary(res, ncp=2, nbelements=Inf)
```

```
##
## Call:
## PCA(X = decathlon, quanti.sup = 11:12, quali.sup = 13)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
## Variance               3.272   1.737   1.405   1.057   0.685   0.599
## % of var.             32.719  17.371  14.049  10.569   6.848   5.993
## Cumulative % of var.  32.719  50.090  64.140  74.708  81.556  87.548
##                        Dim.7   Dim.8   Dim.9  Dim.10
## Variance               0.451   0.397   0.215   0.182
## % of var.              4.512   3.969   2.148   1.822
## Cumulative % of var.  92.061  96.030  98.178 100.000
```

```
##
## Individuals
##                 Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2  
## Sebrle      |  4.843 |  4.038 12.158  0.695 |  1.366  2.619  0.080 |
## Clay        |  4.647 |  3.919 11.451  0.711 |  0.837  0.984  0.032 |
## Karpov      |  5.006 |  4.620 15.911  0.852 |  0.040  0.002  0.000 |
## Macey       |  3.434 |  2.233  3.719  0.423 |  1.042  1.524  0.092 |
## Warners     |  2.979 |  2.168  3.505  0.530 | -1.803  4.565  0.366 |
## Zsivoczky   |  2.566 |  0.925  0.638  0.130 |  1.169  1.918  0.207 |
## Hernu       |  1.824 |  0.889  0.589  0.238 | -0.618  0.537  0.115 |
## Nool        |  3.098 |  0.295  0.065  0.009 | -1.546  3.354  0.249 |
## Bernard     |  2.827 |  1.906  2.709  0.455 | -0.086  0.010  0.001 |
## Schwarzl    |  1.971 |  0.081  0.005  0.002 | -1.353  2.572  0.472 |
## Pogorelov   |  2.383 |  0.540  0.217  0.051 |  0.771  0.834  0.105 |
## Schoenbeck  |  1.797 |  0.114  0.010  0.004 | -0.040  0.002  0.000 |
## Barras      |  2.224 |  0.002  0.000  0.000 |  0.360  0.182  0.026 |
## Smith       |  3.536 |  0.870  0.565  0.061 |  1.059  1.576  0.090 |
## Averyanov   |  2.521 |  0.349  0.091  0.019 | -1.559  3.411  0.382 |
## Ojaniemi    |  2.338 |  0.380  0.108  0.026 | -0.772  0.838  0.109 |
## Smirnov     |  2.021 | -0.485  0.175  0.057 | -1.061  1.580  0.275 |
## Qi          |  1.764 | -0.434  0.141  0.061 | -0.326  0.149  0.034 |
## Drews       |  3.423 | -0.249  0.046  0.005 | -3.082 13.334  0.811 |
## Parkhomenko |  3.486 | -1.069  0.853  0.094 |  2.093  6.152  0.361 |
## Terek       |  3.282 | -0.682  0.347  0.043 |  0.536  0.403  0.027 |
## Gomez       |  2.613 | -0.290  0.063  0.012 | -1.197  2.011  0.210 |
## Turi        |  3.069 | -1.542  1.772  0.252 |  0.427  0.256  0.019 |
## Lorenzo     |  3.510 | -2.409  4.324  0.471 | -1.583  3.518  0.203 |
## Karlivans   |  2.704 | -1.994  2.965  0.544 | -0.294  0.122  0.012 |
## Korkizoglou |  3.975 | -0.958  0.684  0.058 |  2.066  5.995  0.270 |
## Uldal       |  2.946 | -2.562  4.894  0.757 |  0.245  0.085  0.007 |
## Casarsa     |  4.921 | -2.857  6.085  0.337 |  3.798 20.252  0.596 |
## SEBRLE      |  2.369 |  0.792  0.467  0.112 |  0.772  0.836  0.106 |
## CLAY        |  3.507 |  1.235  1.137  0.124 |  0.575  0.464  0.027 |
## KARPOV      |  3.396 |  1.358  1.375  0.160 |  0.484  0.329  0.020 |
## BERNARD     |  2.763 | -0.610  0.277  0.049 | -0.875  1.074  0.100 |
## YURKOV      |  3.018 | -0.586  0.256  0.038 |  2.131  6.376  0.499 |
## WARNERS     |  2.428 |  0.357  0.095  0.022 | -1.685  3.986  0.482 |
## ZSIVOCZKY   |  2.563 |  0.272  0.055  0.011 | -1.094  1.680  0.182 |
## McMULLEN    |  2.561 |  0.588  0.257  0.053 |  0.231  0.075  0.008 |
## MARTINEAU   |  3.742 | -1.995  2.968  0.284 |  0.561  0.442  0.022 |
## HERNU       |  2.794 | -1.546  1.782  0.306 |  0.488  0.335  0.031 |
## BARRAS      |  1.952 | -1.342  1.342  0.472 | -0.311  0.136  0.025 |
## NOOL        |  3.734 | -2.345  4.099  0.394 | -1.966  5.429  0.277 |
## BOURGUIGNON |  4.299 | -3.979 11.802  0.857 |  0.200  0.056  0.002 |
##
## Variables
##                Dim.1    ctr   cos2    Dim.2    ctr   cos2  
## 100m        | -0.775 18.344  0.600 |  0.187  2.016  0.035 |
## Long jump   |  0.742 16.822  0.550 | -0.345  6.869  0.119 |
## Shot put    |  0.623 11.844  0.388 |  0.598 20.607  0.358 |
## High jump   |  0.572  9.998  0.327 |  0.350  7.064  0.123 |
## 400m        | -0.680 14.116  0.462 |  0.569 18.666  0.324 |
## 110m H      | -0.746 17.020  0.557 |  0.229  3.013  0.052 |
## Discus      |  0.552  9.328  0.305 |  0.606 21.162  0.368 |
## Pole vault  |  0.050  0.077  0.003 | -0.180  1.873  0.033 |
## Javeline    |  0.277  2.347  0.077 |  0.317  5.784  0.100 |
## 1500m       | -0.058  0.103  0.003 |  0.474 12.946  0.225 |
##
## Supplementary continuous variables
##                Dim.1   cos2    Dim.2   cos2
## Rank        | -0.671  0.450 |  0.051  0.003 |
## Points      |  0.956  0.914 | -0.017  0.000 |
##
## Supplementary categories
##                 Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test
## Decastar    |  0.946 | -0.600  0.403 -1.430 | -0.038  0.002 -0.123 |
## OlympicG    |  0.439 |  0.279  0.403  1.430 |  0.017  0.002  0.123 |
```

In order to print the results in a file:

```
summary(res, nbelements=Inf, file="summaryResult.txt")
```

Description of the dimensions

```
dimdesc(res)
```

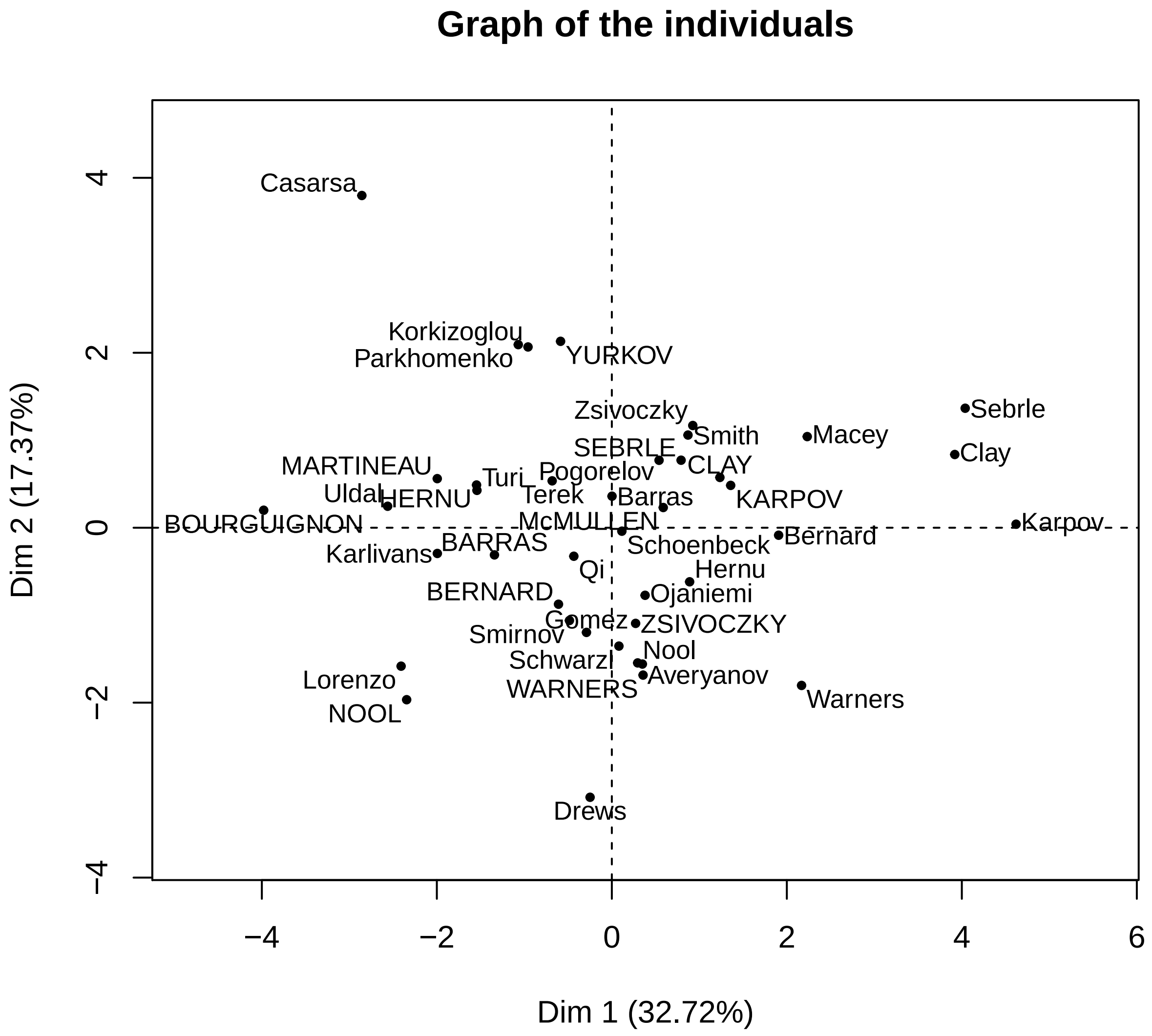
```
## $Dim.1
## $Dim.1$quanti
##           correlation      p.value
## Points      0.9561543 2.099191e-22
## Long jump   0.7418997 2.849886e-08
## Shot put    0.6225026 1.388321e-05
## High jump   0.5719453 9.362285e-05
## Discus      0.5524665 1.802220e-04
## Rank       -0.6705104 1.616348e-06
## 400m       -0.6796099 1.028175e-06
## 110m H     -0.7462453 2.136962e-08
## 100m       -0.7747198 2.778467e-09
##
##
## $Dim.2
## $Dim.2$quanti
##           correlation      p.value
## Discus      0.6063134 2.650745e-05
## Shot put    0.5983033 3.603567e-05
## 400m        0.5694378 1.020941e-04
## 1500m       0.4742238 1.734405e-03
## High jump   0.3502936 2.475025e-02
## Javeline    0.3169891 4.344974e-02
## Long jump  -0.3454213 2.696969e-02
##
##
## $Dim.3
## $Dim.3$quanti
##            correlation      p.value
## 1500m        0.7821428 1.554450e-09
## Pole vault   0.6917567 5.480172e-07
## Javeline    -0.3896554 1.179331e-02
```

```
dimdesc(res, proba=0.2) # change the significance threshold to characterize the dimension
```

```
## $Dim.1
## $Dim.1$quanti
##           correlation      p.value
## Points      0.9561543 2.099191e-22
## Long jump   0.7418997 2.849886e-08
## Shot put    0.6225026 1.388321e-05
## High jump   0.5719453 9.362285e-05
## Discus      0.5524665 1.802220e-04
## Javeline    0.2771108 7.942460e-02
## Rank       -0.6705104 1.616348e-06
## 400m       -0.6796099 1.028175e-06
## 110m H     -0.7462453 2.136962e-08
## 100m       -0.7747198 2.778467e-09
##
## $Dim.1$quali
##                     R2   p.value
## Competition 0.05110487 0.1552515
##
## $Dim.1$category
##            Estimate   p.value
## OlympicG  0.4393744 0.1552515
## Decastar -0.4393744 0.1552515
##
##
## $Dim.2
## $Dim.2$quanti
##           correlation      p.value
## Discus      0.6063134 2.650745e-05
## Shot put    0.5983033 3.603567e-05
## 400m        0.5694378 1.020941e-04
## 1500m       0.4742238 1.734405e-03
## High jump   0.3502936 2.475025e-02
## Javeline    0.3169891 4.344974e-02
## 110m H      0.2287933 1.501925e-01
## Long jump  -0.3454213 2.696969e-02
##
##
## $Dim.3
## $Dim.3$quanti
##            correlation      p.value
## 1500m        0.7821428 1.554450e-09
## Pole vault   0.6917567 5.480172e-07
## High jump   -0.2595119 1.013160e-01
## Javeline    -0.3896554 1.179331e-02
```

Graph of the individuals with a title and a smaller size for the labels

```
plot(res, cex=0.8, invisible="quali", title="Graph of the individuals")
```

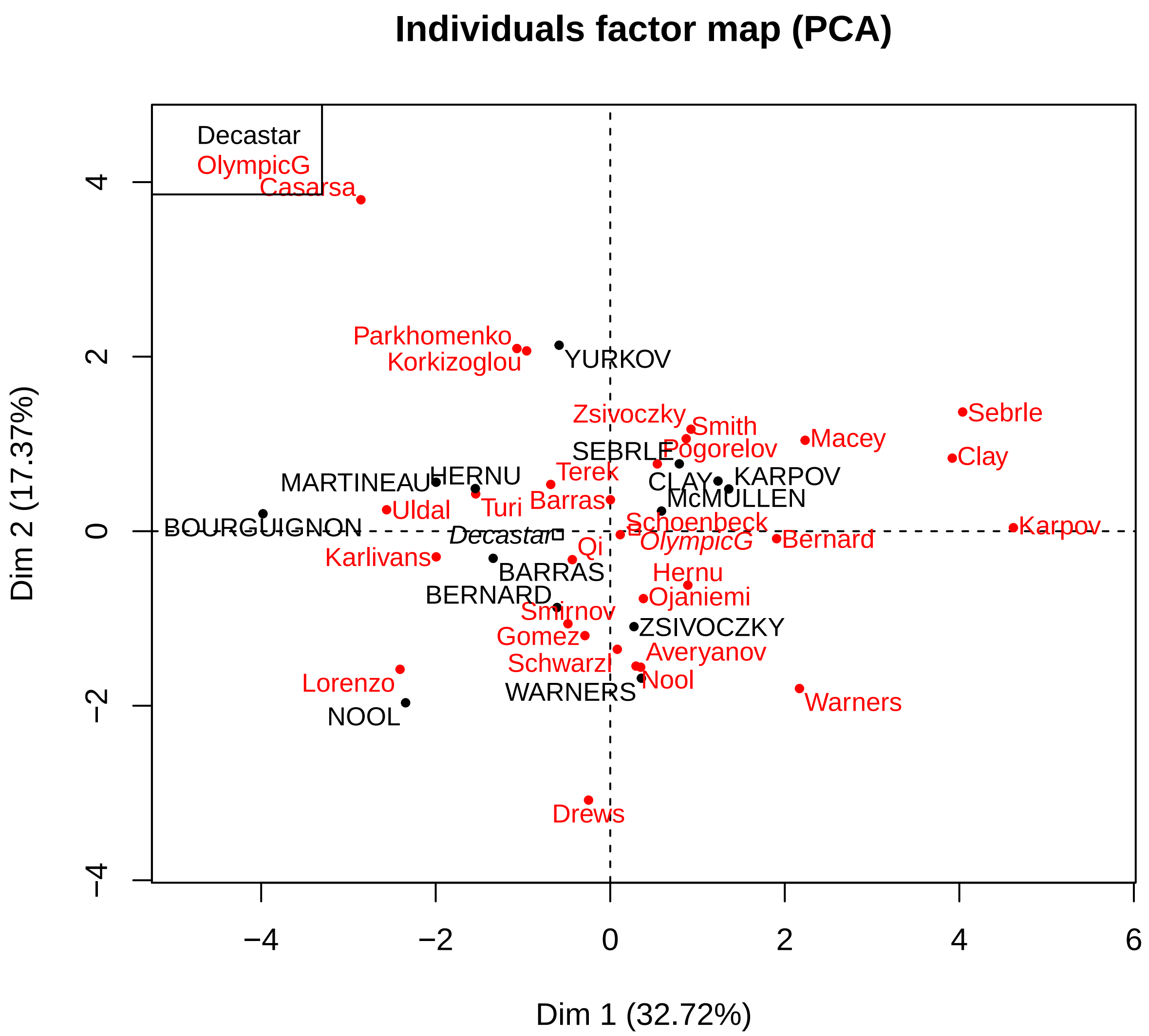


With many individuals and if the labels are not useful, one can suppress the labels with the argument label="none".

```
plot(res, cex=0.8, invisible="quali", label="none", title="Graph of the individuals")
```

Drawing individuals according to the competition

```
plot(res, cex=0.8, habillage="Competition")
```

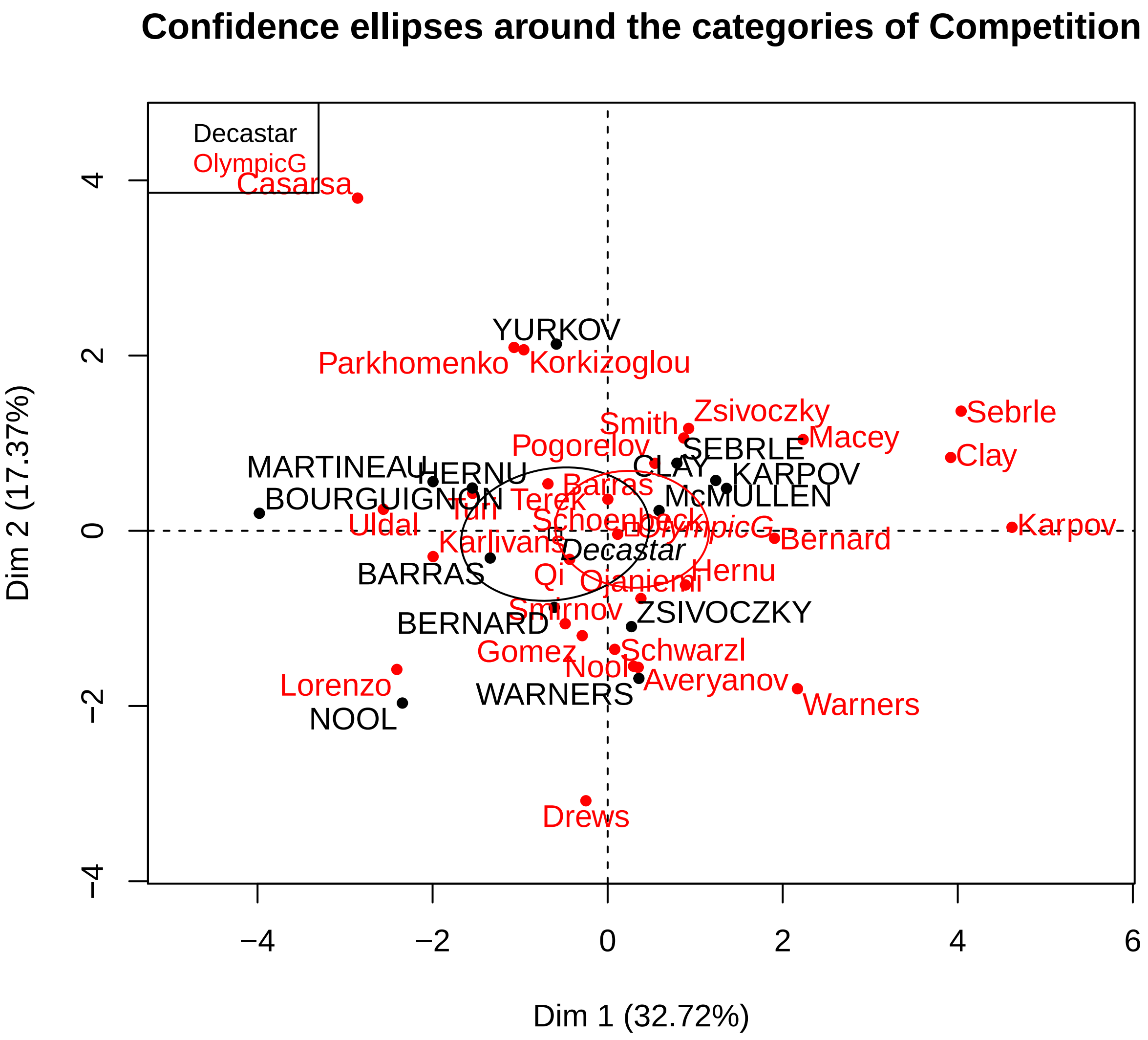


We could have written:

```
plot(res, cex=0.8, habillage=13)
```

Confidence ellipses around the categories

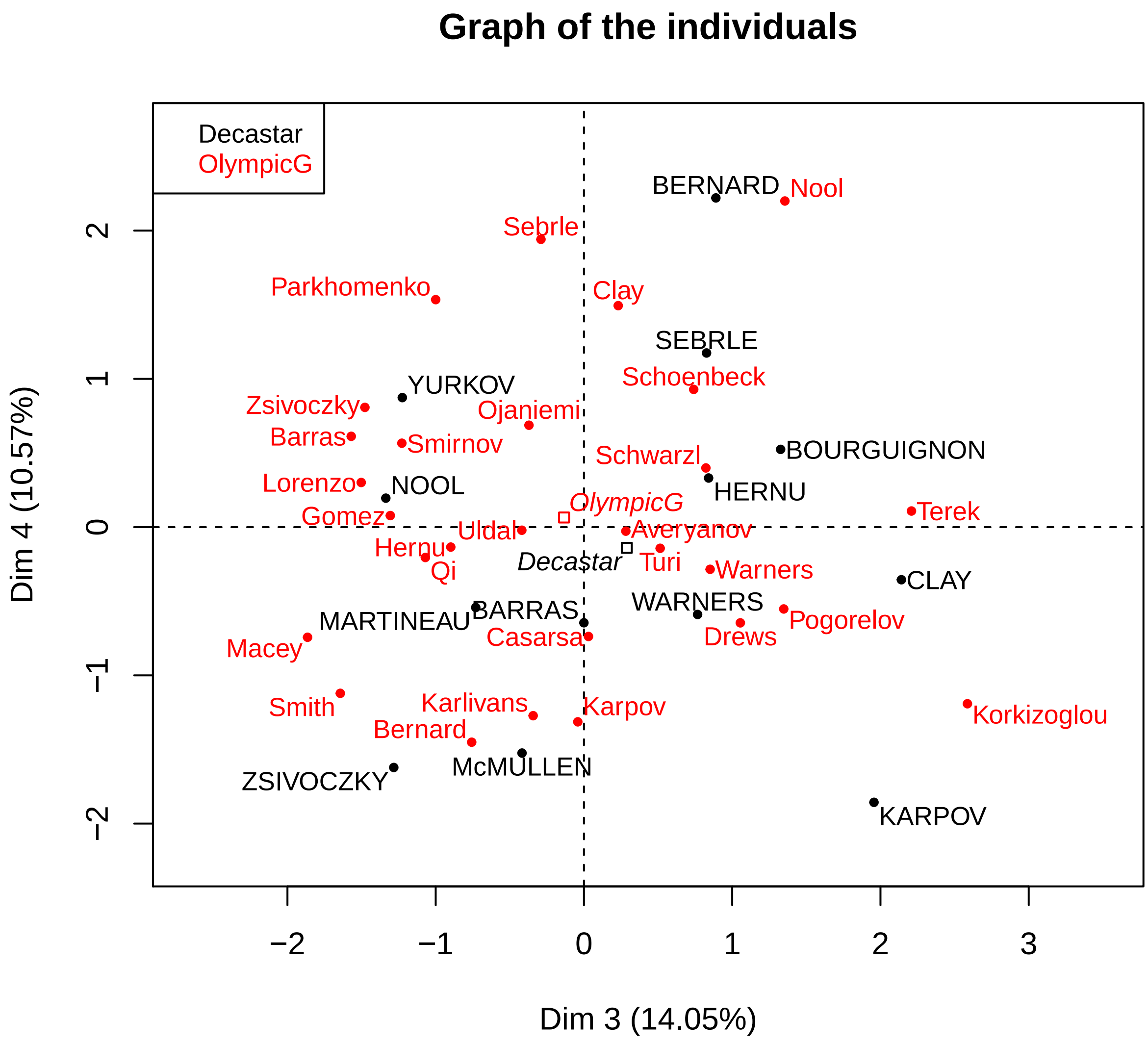
```
plotellipses(res)
```



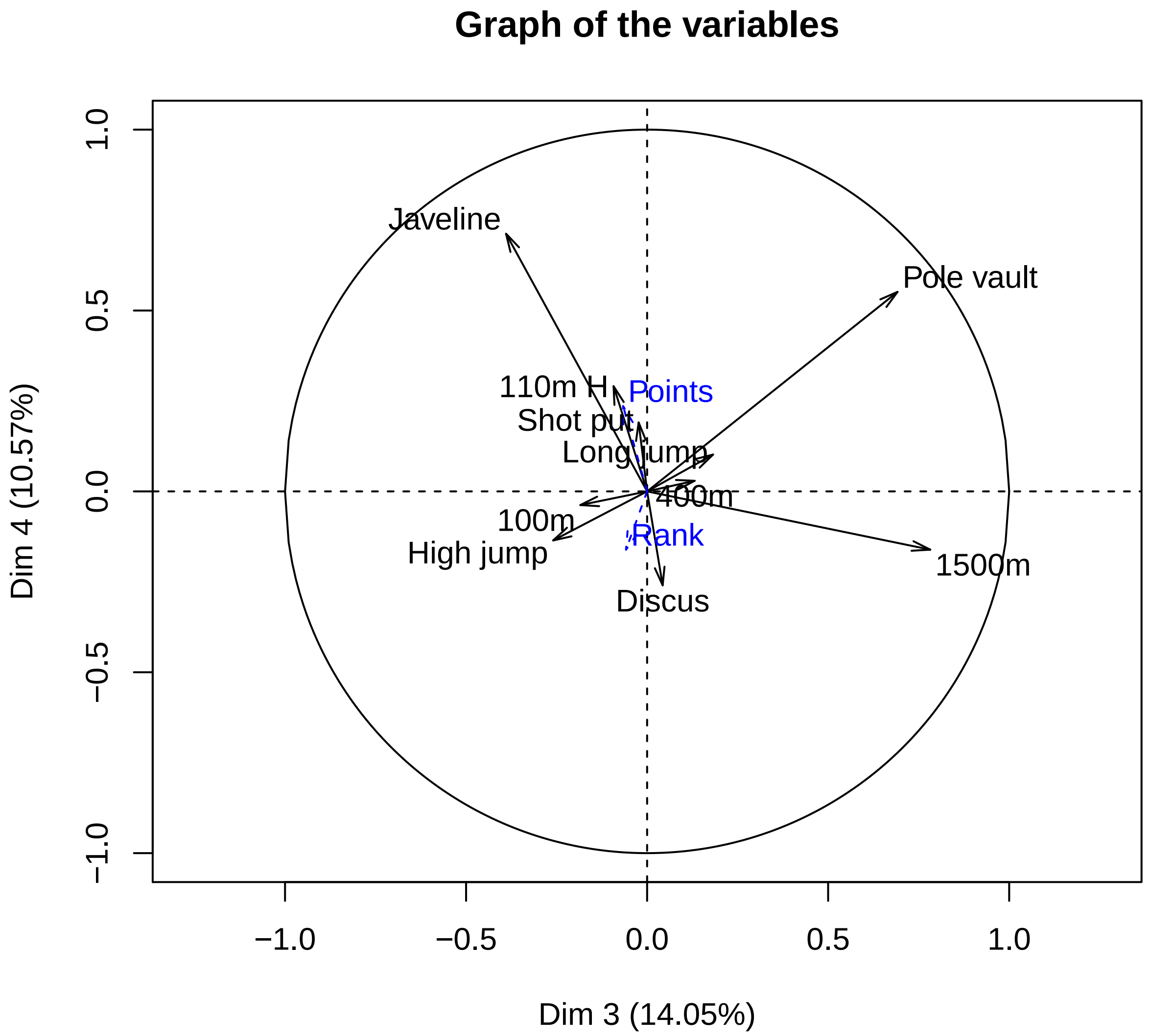
If several qualitative variables are available, there will be as many graphs as qualitative variables. And on each graph the confidence ellipses around the categories of a categorical variable.

Graph for dimensions 3 and 4

```
plot(res, choix="ind", cex=0.8, habillage=13, title="Graph of the individuals", axes=3:4)
```



```
plot(res, choix="var", title="Graph of the variables", axes=3:4)
```


Graph of the variables
Javeline
Pole vault
110m H
Points
Shot put
Long jump
400m
100m
High jump
Rank
1500m
Discus
1.0
0.5
0.0
−0.5
−1.0
Dim 4 (10.57%)
−1.0
−0.5
0.0
0.5
1.0
Dim 3 (14.05%)

Selecting individuals

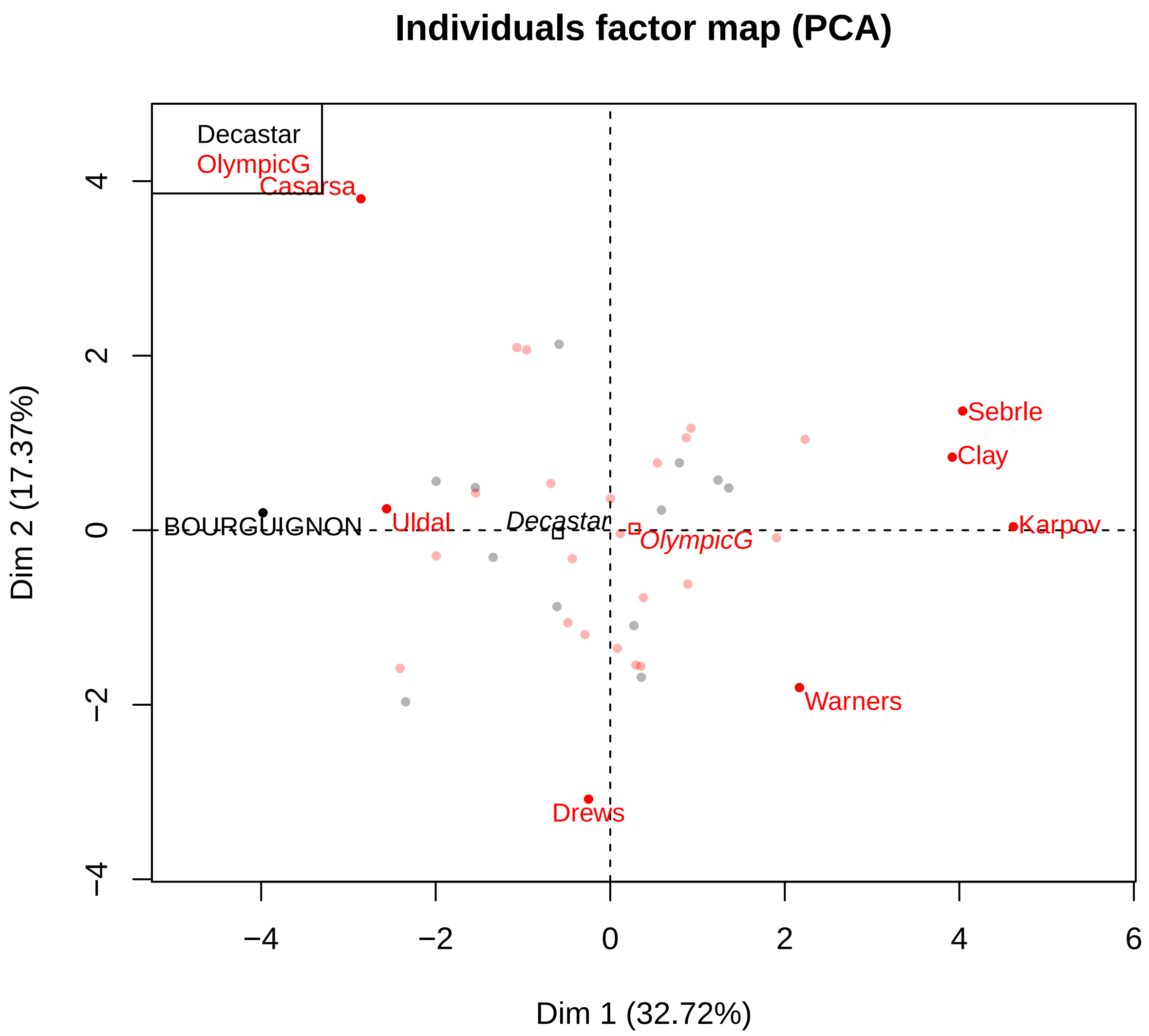
`select="cos2 0.7"` : select the individuals that have a quality of representation on the map greater than 0.7

`select="cos2 5"` : select the 5 individuals that have the best quality of representation on the map

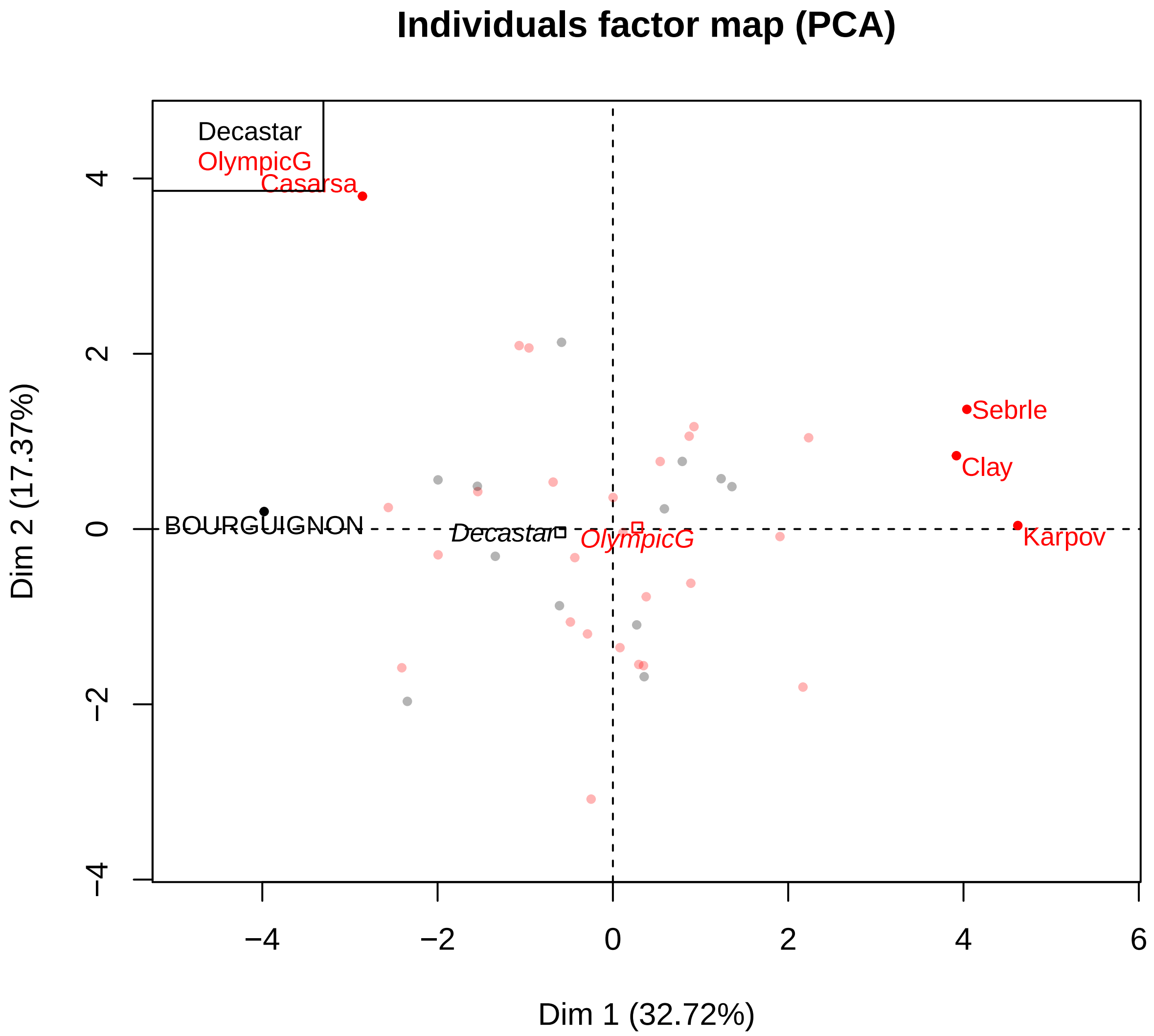
`select="contrib 5"` : select the 5 individuals that contribute the most to the construction of the map

`select=c("nom1","nom2")` : select the individuals by their name

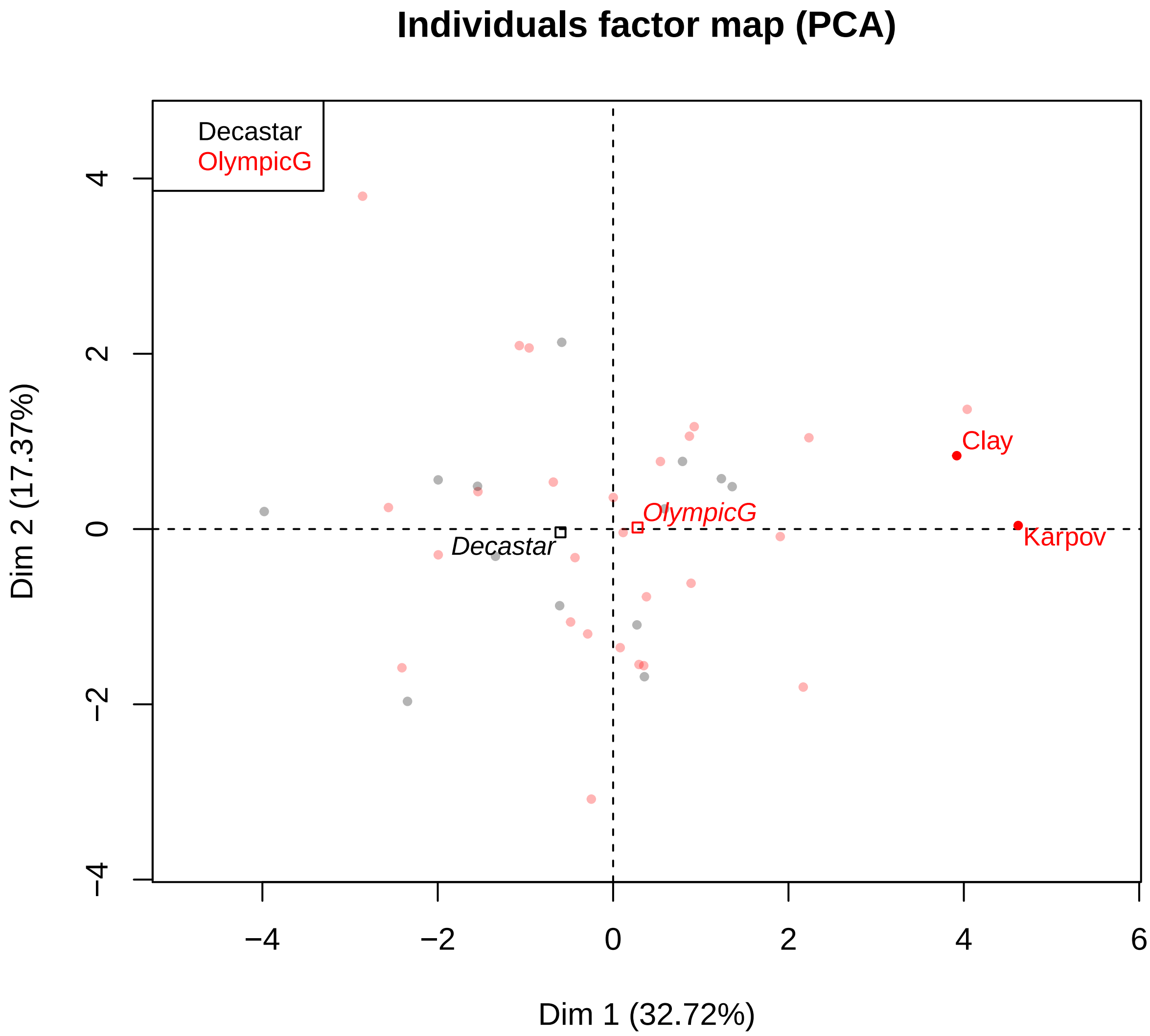
```
plot(res, cex=0.8, habillage=13, select="cos2 0.7")
```



```
plot(res, cex=0.8, habillage=13, select="contrib 5")
```

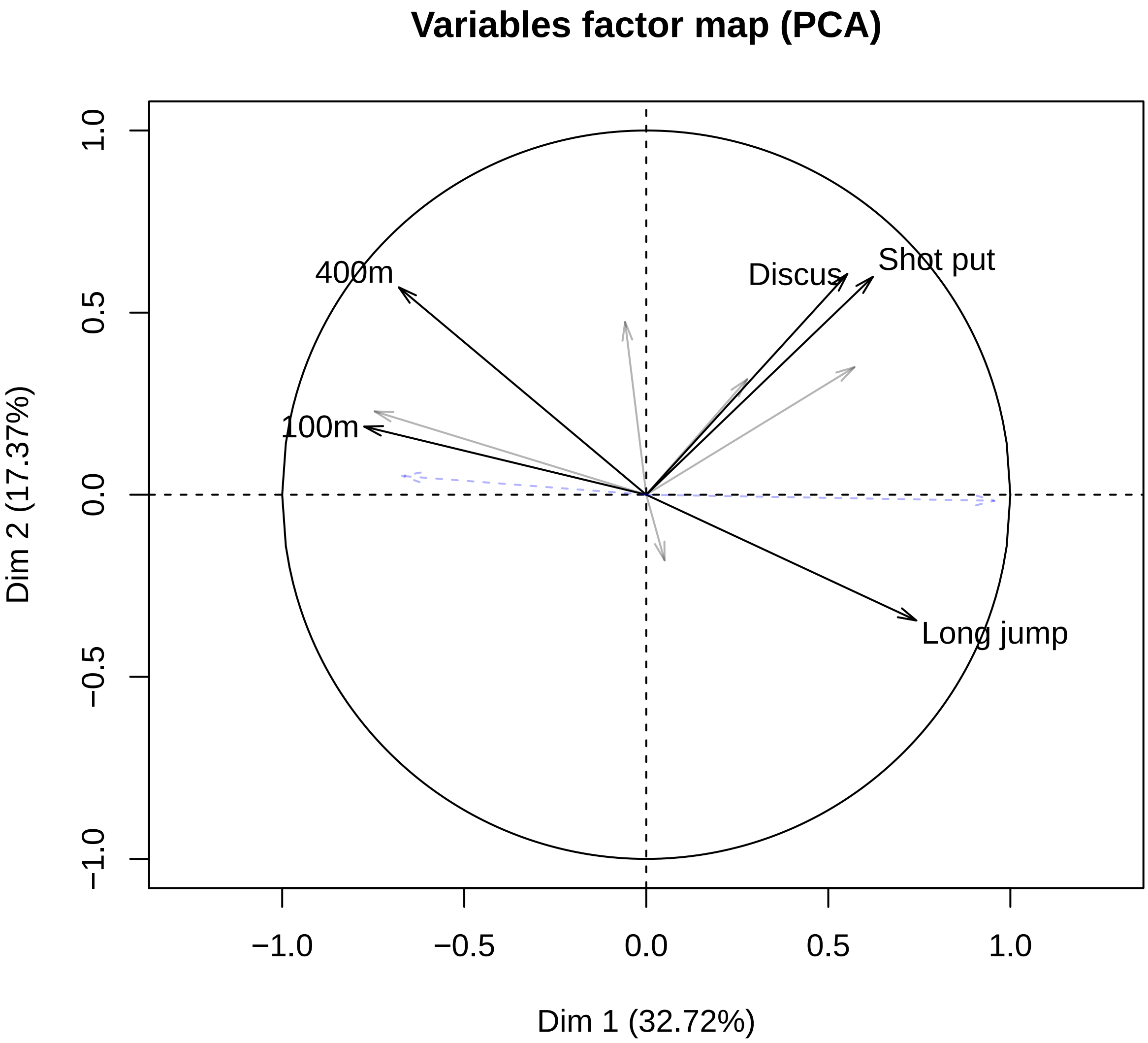


```r
plot(res, cex=0.8, habillage=13, select=c("Clay","Karpov"))
```



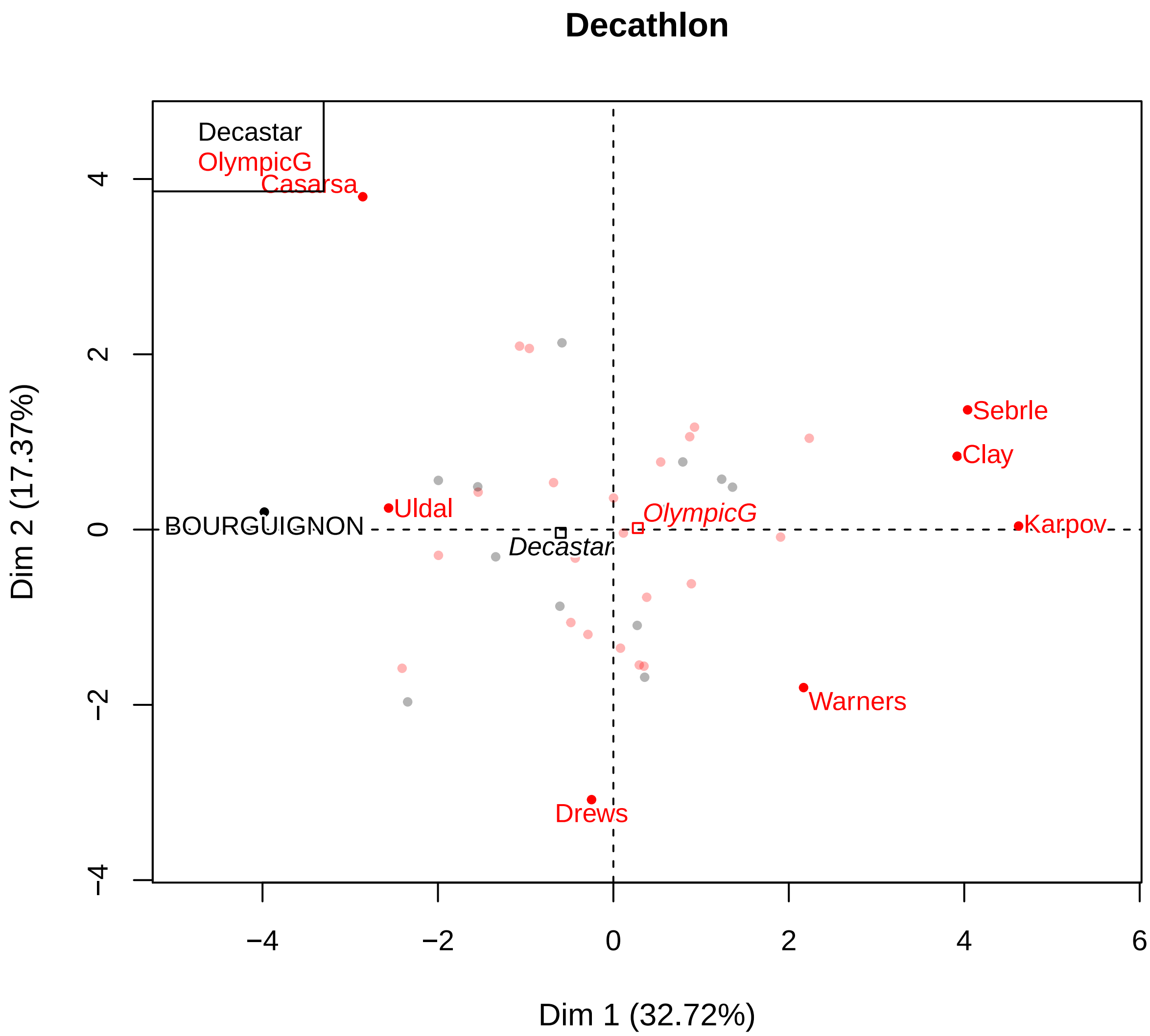
Selecting variables

```
plot(res, choix="var", select="contrib 5")
```


Variables factor map (PCA)
400m
100m
Discus
Shot put
Long jump
Dim 1 (32.72%)
Dim 2 (17.37%)
−1.0
−0.5
0.0
0.5
1.0
−1.0
−0.5
0.0
0.5
1.0

Graph with different options

```r
plot(res, cex=0.8, habillage=13, select="cos2 0.7", title="Decathlon",
     cex.main=1.1, cex.axis=0.9, shadow=TRUE, auto="y")
```


Decathlon
Decastar
OlympicG
Casarsa
Sebrle
Clay
Karpov
Uldal
BOURGUIGNON
OlympicG
Decastar
Warners
Drews
Dim 2 (17.37%)
Dim 1 (32.72%)