Analyse en composantes principales (ACP) avec FactoMineR sur les données décathlon

François Husson

Importation du jeu de données

Vous pouvez importer le jeu de données après l'avoir sauvegardé sur votre ordinateur ou bien directement à partir su lien suivant http://www.agrocampus-ouest.fr/math/livreR/decathlon.csv.

```
decathlon <- read.table("http://www.agrocampus-ouest.fr/math/livreR/decathlon.csv",
      header=TRUE, sep=";",dec=".", row.names=1, check.names=FALSE, fileEncoding="latin1")
```

`header=TRUE` : précise que le nom des variables est présent

`sep=";"` : précise que le séparateur de colonnes est le point-virgule (fréquent dans les fichiers csv, pour une tabulation il faudrait écrire `sep="\t"`)

`dec="."` : le séparateur de décimale est le point (parfois dans Excel on trouve la virgule)

`row.names=1` : précise que le nom des individus est dans la première colonne du tableau

`check.names=FALSE` : impose que le nom des colonnes soit pris tel que dans le fichier (sinon les espaces sont remplacés par des points et des X sont mis avant les nombres)

Il est important de s'assurer que l'importation a bien été effectuée, et notamment que les variables quantitatives sont bien considérées comme quantitatives et les variables qualitatives bien considérées comme qualitatives

```
summary(decathlon)
```

```
##       100m          Longueur        Poids          Hauteur
##  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          Disque          Perche
##  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
##     Javelot          1500m         Classement        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
```

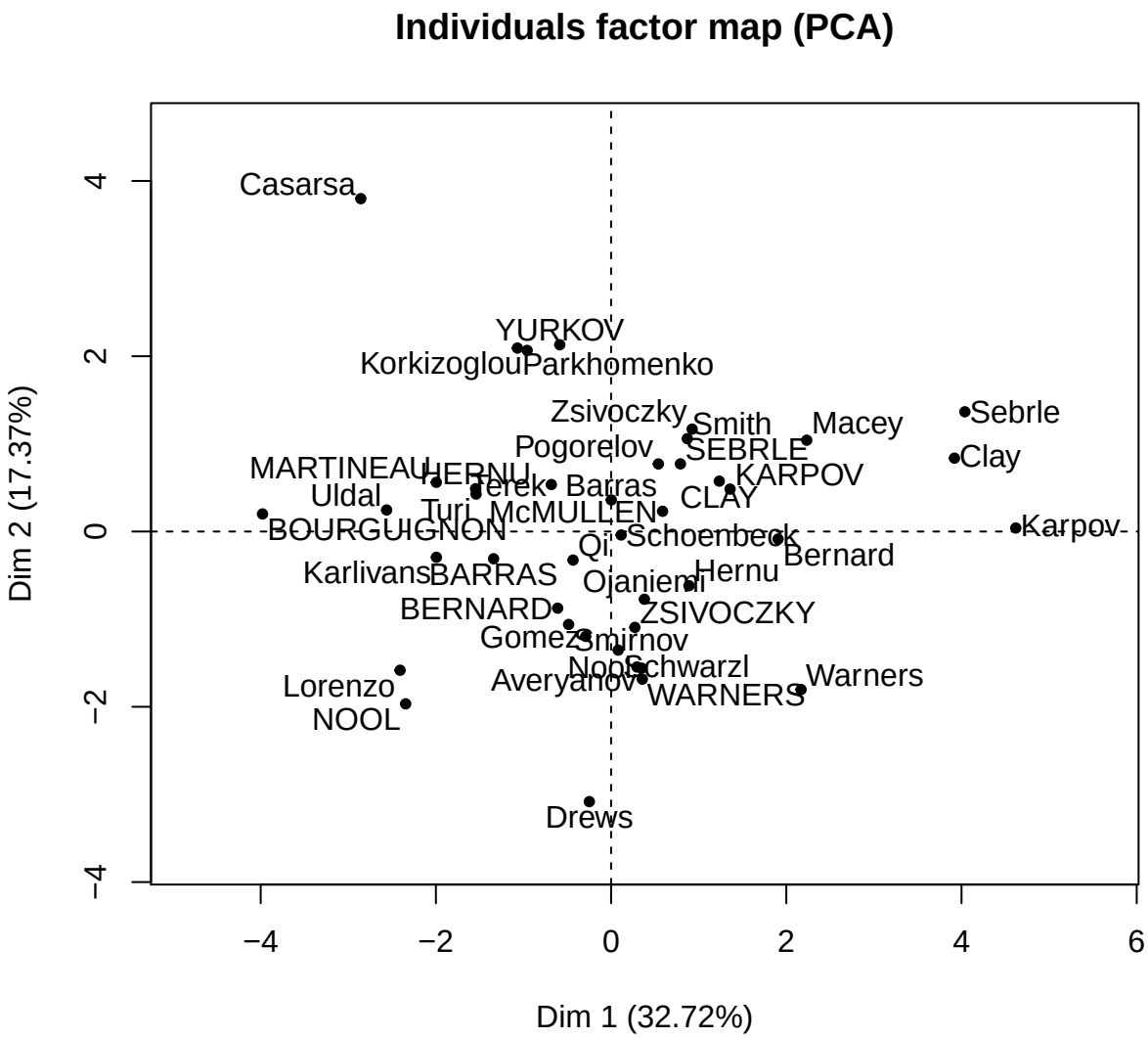
```
##  Decastar:13
##  JO      :28
##
##
##
##
```

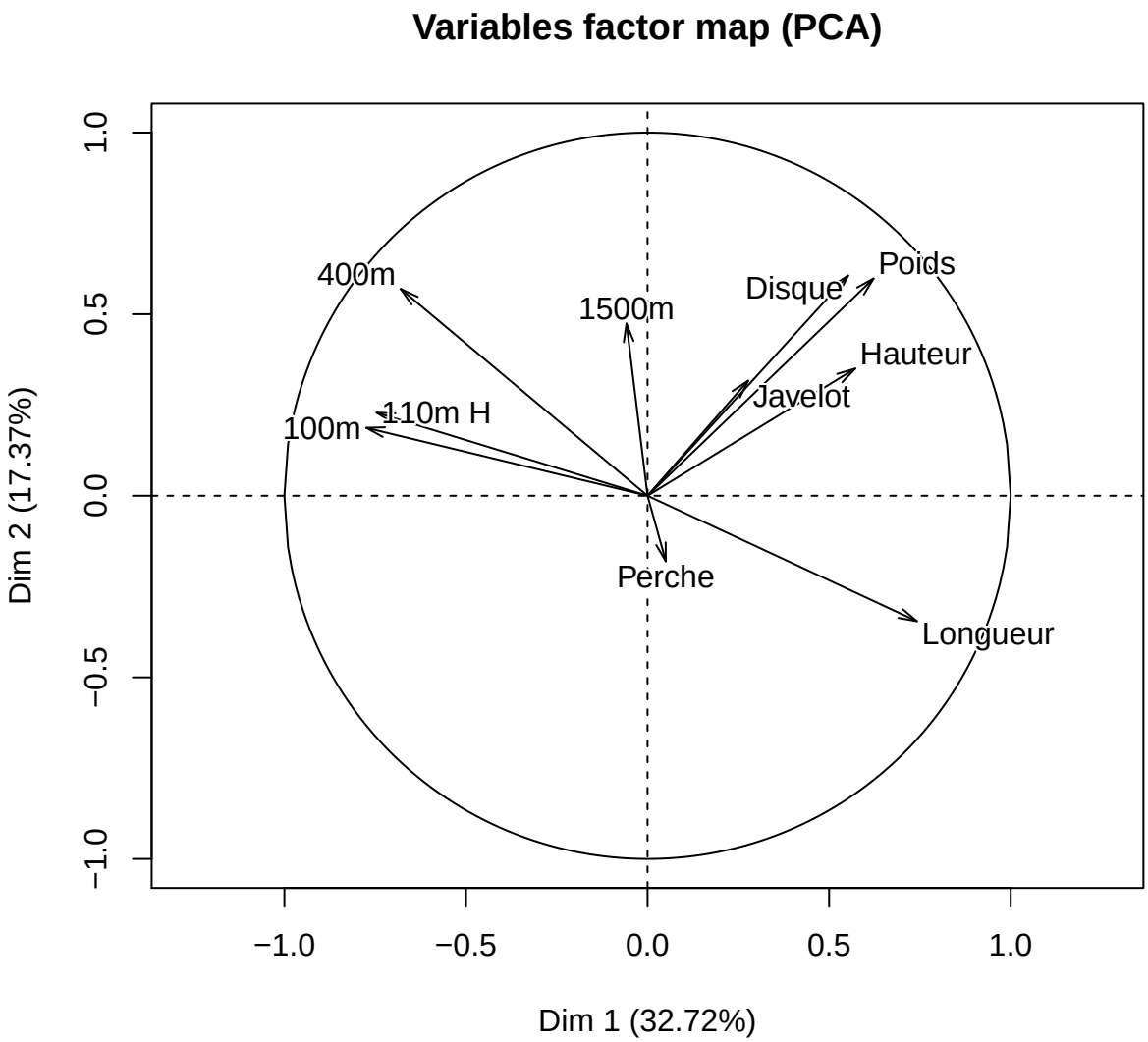
Chargement de FactoMineR

```
library(FactoMineR)
```

L'ACP avec uniquement des éléments (lignes et variables) actifs

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


Individuals factor map (PCA)
Casarsa
YURKOV
KorkizoglouParkhomenko
Zsivoczky
Smith
Macey
Sebrle
Pogorelov
SEBRLE
Clay
MARTINEAU
HERNU
Terek
Barras
KARPOV
Uldal
CLAY
Turi
McMULLEN
Karpov
BOURGUIGNON
Qi
Schoenbeck
Bernard
KarlivansBARRAS
Ojaniemi
Hernu
BERNARD
ZSIVOCZKY
GomezSmirnov
NoolSchwarzl
Warners
Lorenzo
Averyanov
WARNERS
NOOL
Drews
Dim 1 (32.72%)
Dim 2 (17.37%)
−4
−2
0
2
4
6



On peut obtenir un résumé des principaux résultats en utilisant la fonction `summary`.

```
summary(res)
```

Nous demandons ici à avoir les résultats sur les 2 premières dimensions pour éviter d'avoir des tableaux trop grands (par défaut, la fonction retourne les résultats des 3 premières dimensions).

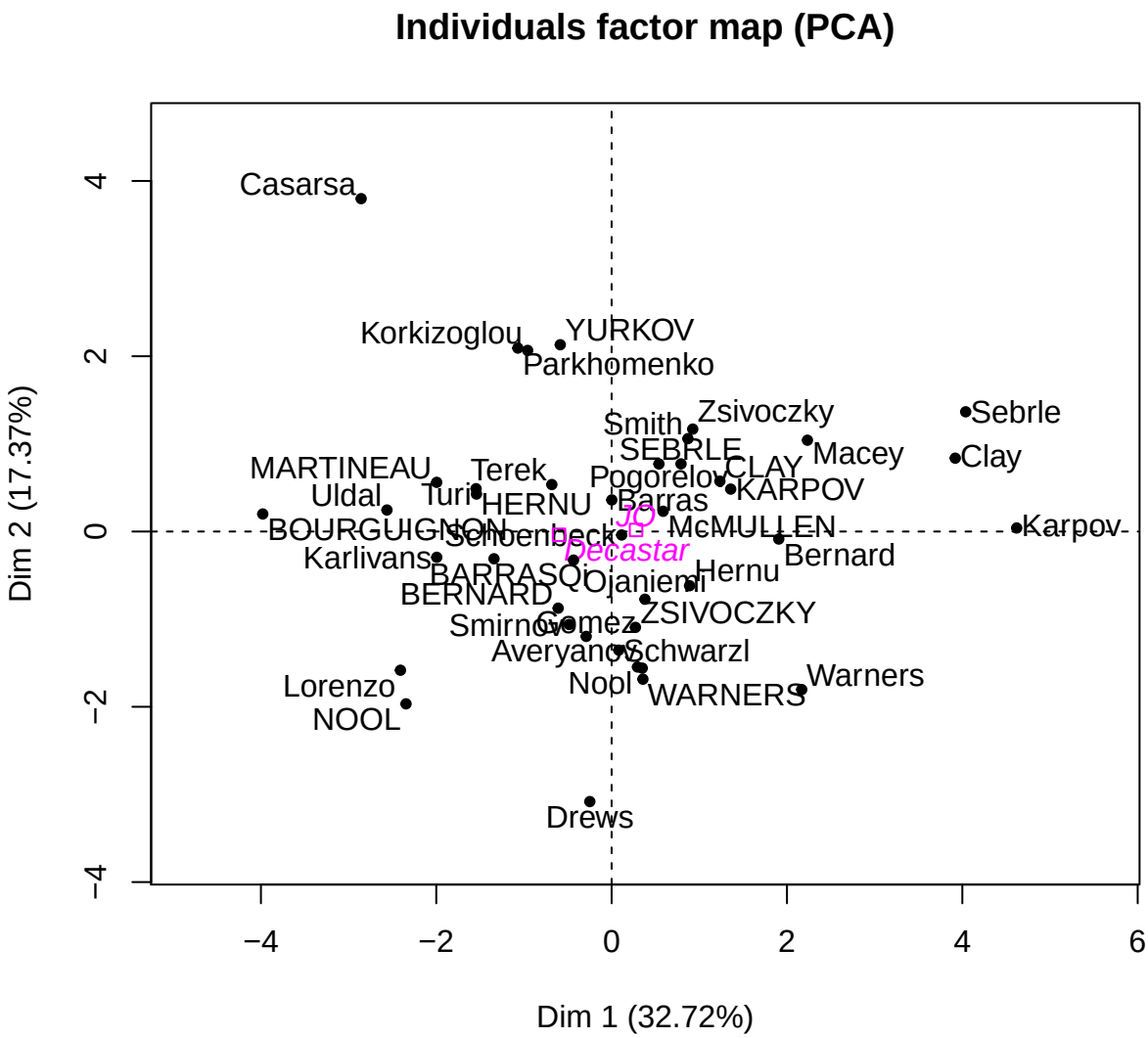
```
summary(res, ncp=2)
```

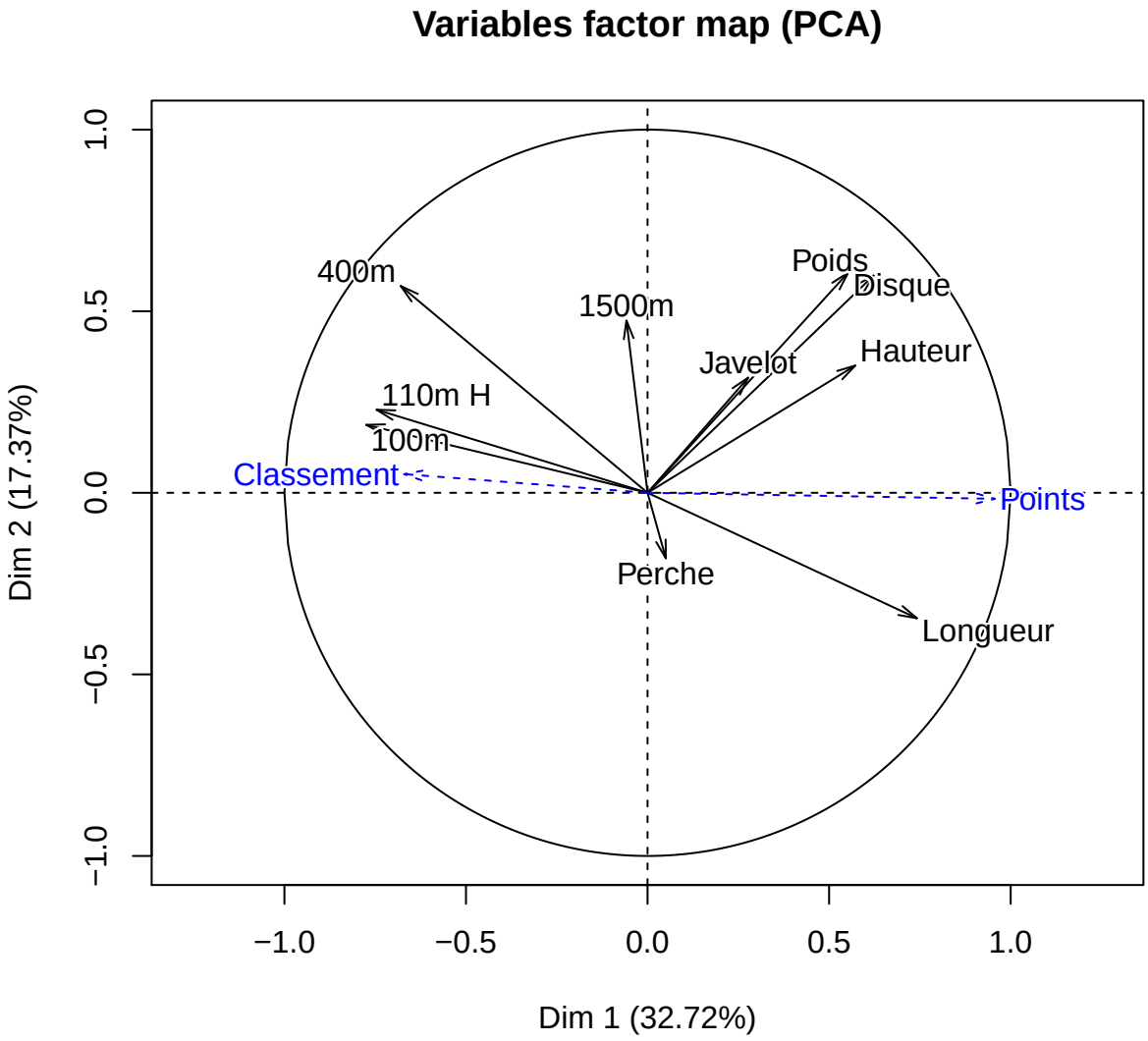
```
##
## Call:
## PCA(X = decathlon[, 1:10])
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
```

```
## 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 |
## Longueur  |  0.742 16.822  0.550 | -0.345  6.869  0.119 |
## Poids     |  0.623 11.844  0.388 |  0.598 20.607  0.358 |
## Hauteur   |  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 |
## Disque    |  0.552  9.328  0.305 |  0.606 21.162  0.368 |
## Perche    |  0.050  0.077  0.003 | -0.180  1.873  0.033 |
## Javelot   |  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 |
```

L'ACP avec des variables supplémentaires

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


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




```
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 |
## Longueur    |  0.742 16.822  0.550 | -0.345  6.869  0.119 |
## Poids       |  0.623 11.844  0.388 |  0.598 20.607  0.358 |
## Hauteur     |  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 |
## Disque      |  0.552  9.328  0.305 |  0.606 21.162  0.368 |
```

```
## Perche      |  0.050  0.077  0.003 | -0.180  1.873  0.033 |
## Javelot     |  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
## Classement  | -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 |
## JO          |  0.439 |  0.279  0.403  1.430 |  0.017  0.002  0.123 |
```

Pour imprimer les résultats dans un fichier :

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

Description des dimensions

```
dimdesc(res)
```

```
## $Dim.1
## $Dim.1$quanti
##             correlation      p.value
## Points        0.9561543 2.099191e-22
## Longueur      0.7418997 2.849886e-08
## Poids         0.6225026 1.388321e-05
## Hauteur       0.5719453 9.362285e-05
## Disque        0.5524665 1.802220e-04
## Classement   -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
## Disque      0.6063134 2.650745e-05
## Poids       0.5983033 3.603567e-05
## 400m        0.5694378 1.020941e-04
## 1500m       0.4742238 1.734405e-03
## Hauteur     0.3502936 2.475025e-02
## Javelot     0.3169891 4.344974e-02
## Longueur   -0.3454213 2.696969e-02
##
##
## $Dim.3
## $Dim.3$quanti
```

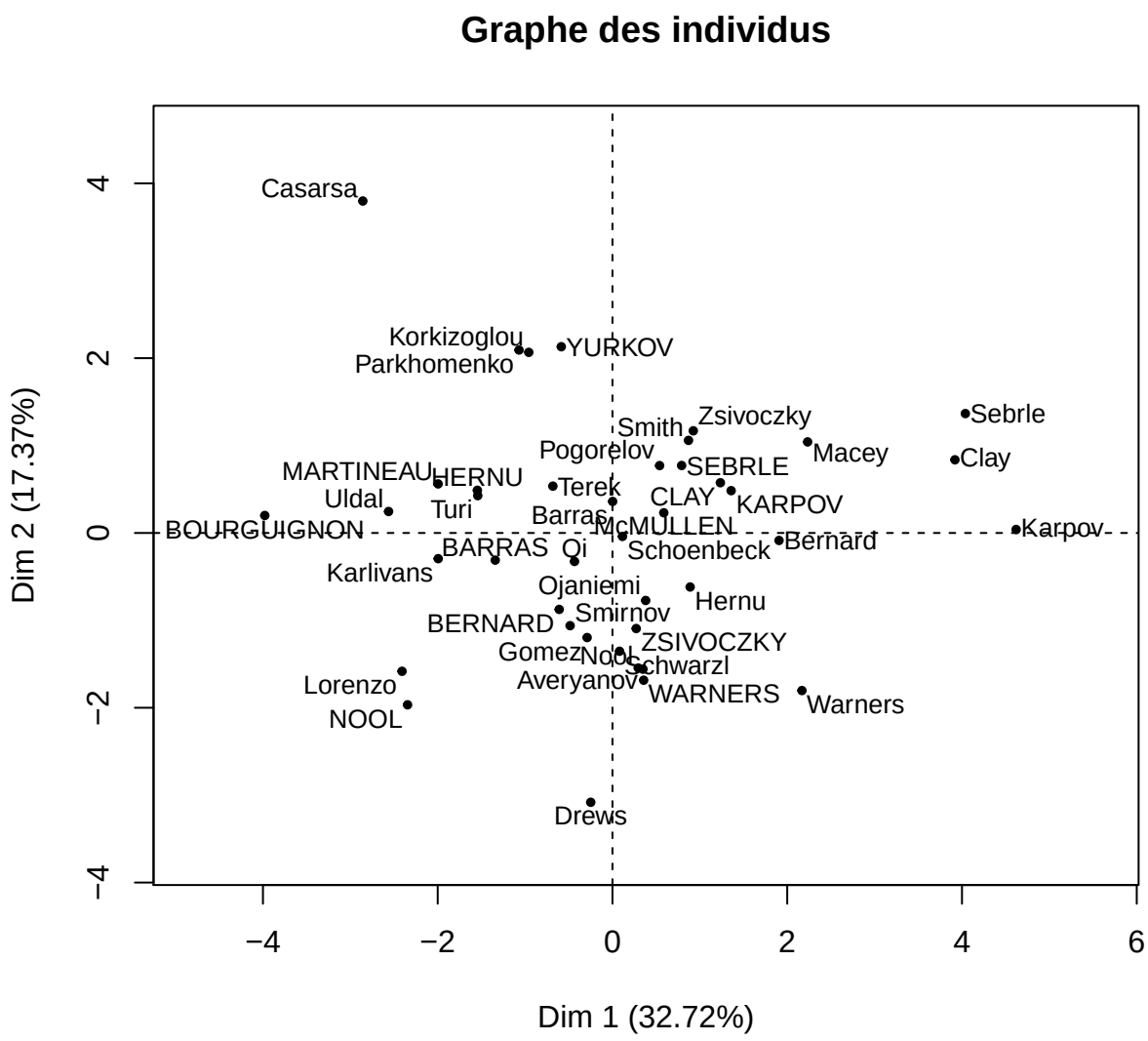
```
##         correlation      p.value
## 1500m     0.7821428 1.554450e-09
## Perche    0.6917567 5.480172e-07
## Javelot  -0.3896554 1.179331e-02
```

```
dimdesc(res, proba=0.2)
```

```
## $Dim.1
## $Dim.1$quanti
##            correlation      p.value
## Points       0.9561543 2.099191e-22
## Longueur     0.7418997 2.849886e-08
## Poids        0.6225026 1.388321e-05
## Hauteur      0.5719453 9.362285e-05
## Disque       0.5524665 1.802220e-04
## Javelot      0.2771108 7.942460e-02
## Classement  -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
## JO        0.4393744 0.1552515
## Decastar -0.4393744 0.1552515
##
##
## $Dim.2
## $Dim.2$quanti
##          correlation      p.value
## Disque     0.6063134 2.650745e-05
## Poids      0.5983033 3.603567e-05
## 400m       0.5694378 1.020941e-04
## 1500m      0.4742238 1.734405e-03
## Hauteur    0.3502936 2.475025e-02
## Javelot    0.3169891 4.344974e-02
## 110m H     0.2287933 1.501925e-01
## Longueur  -0.3454213 2.696969e-02
##
##
## $Dim.3
## $Dim.3$quanti
##         correlation      p.value
## 1500m     0.7821428 1.554450e-09
## Perche    0.6917567 5.480172e-07
## Hauteur  -0.2595119 1.013160e-01
## Javelot  -0.3896554 1.179331e-02
```

Graphe des individus avec des libellés de police plus petite et avec un titre

```
plot(res, cex=0.8, invisible="quali", title="Graphe des individus")
```

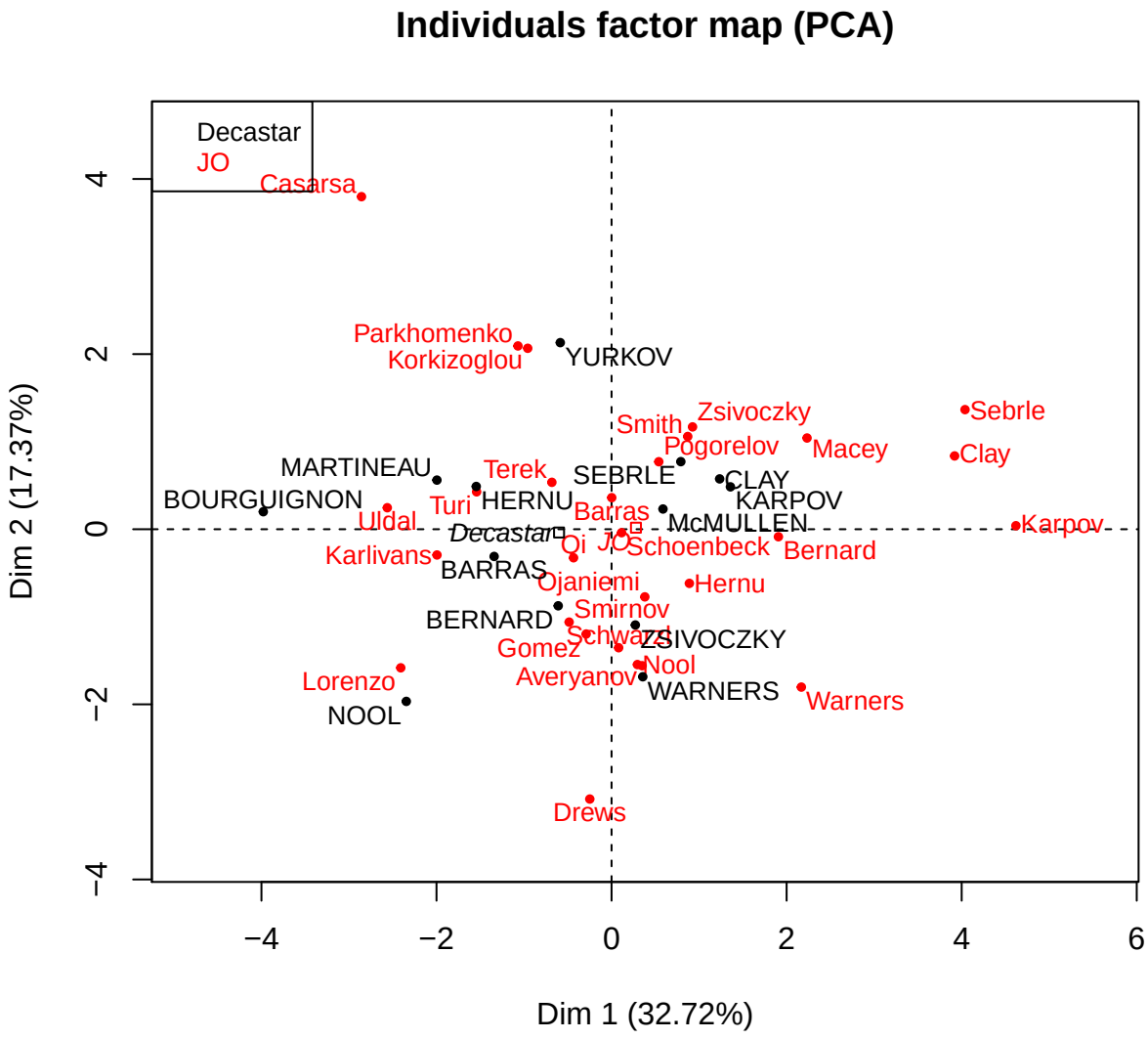


Si on a beaucoup d'individus et que les libellés des individus ne sont pas explicites (des numéros par exemple), on peut supprimer les noms des libellés tout en laissant les points avec l'argument label="none".

```
plot(res, cex=0.8, invisible="quali", label="none", title="Graphe des individus")
```

Coloriage des individus en fonction de leur modalité

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

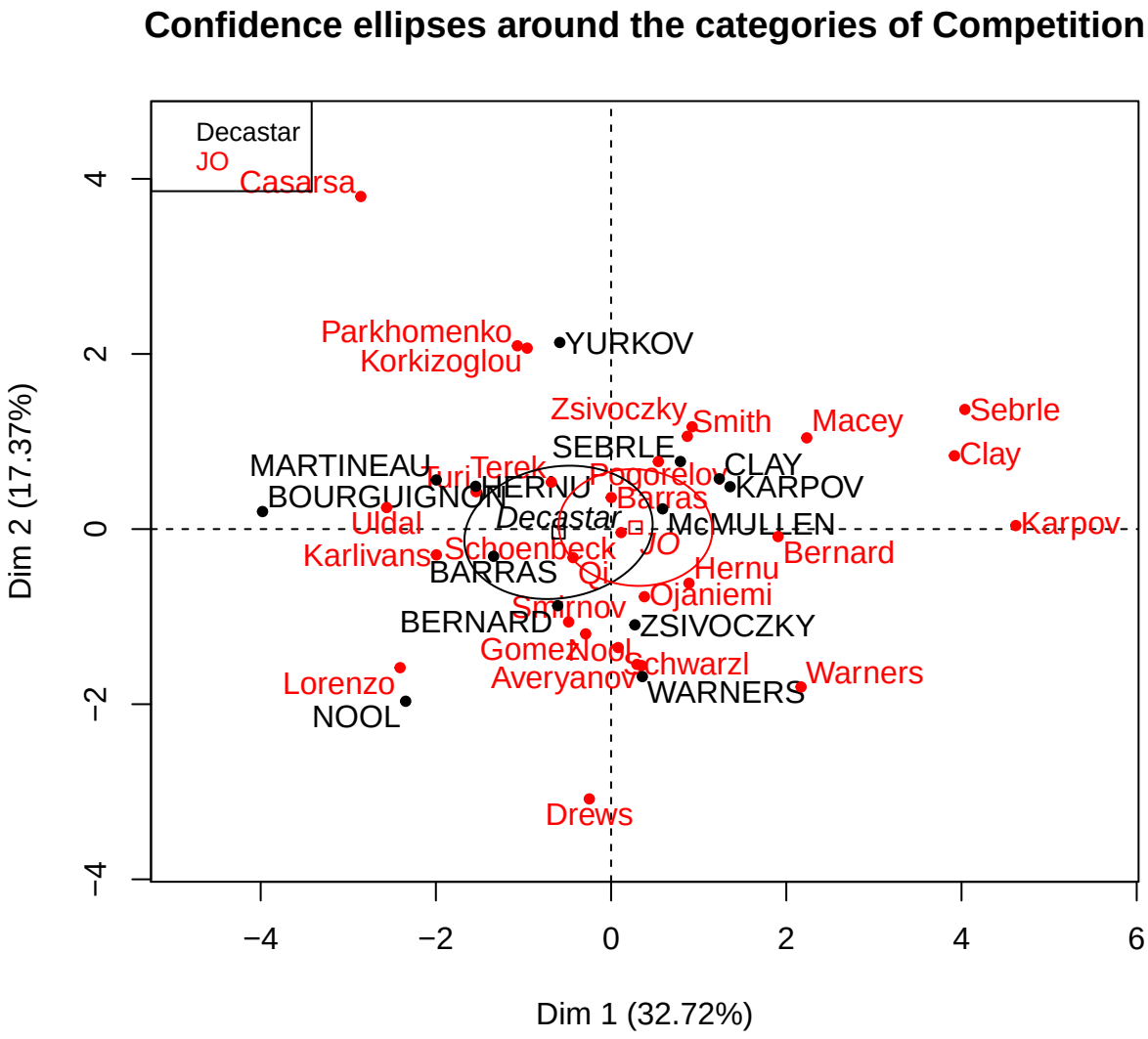


On aurait pu écrire :

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

Ellipses de confiance autour des modalités

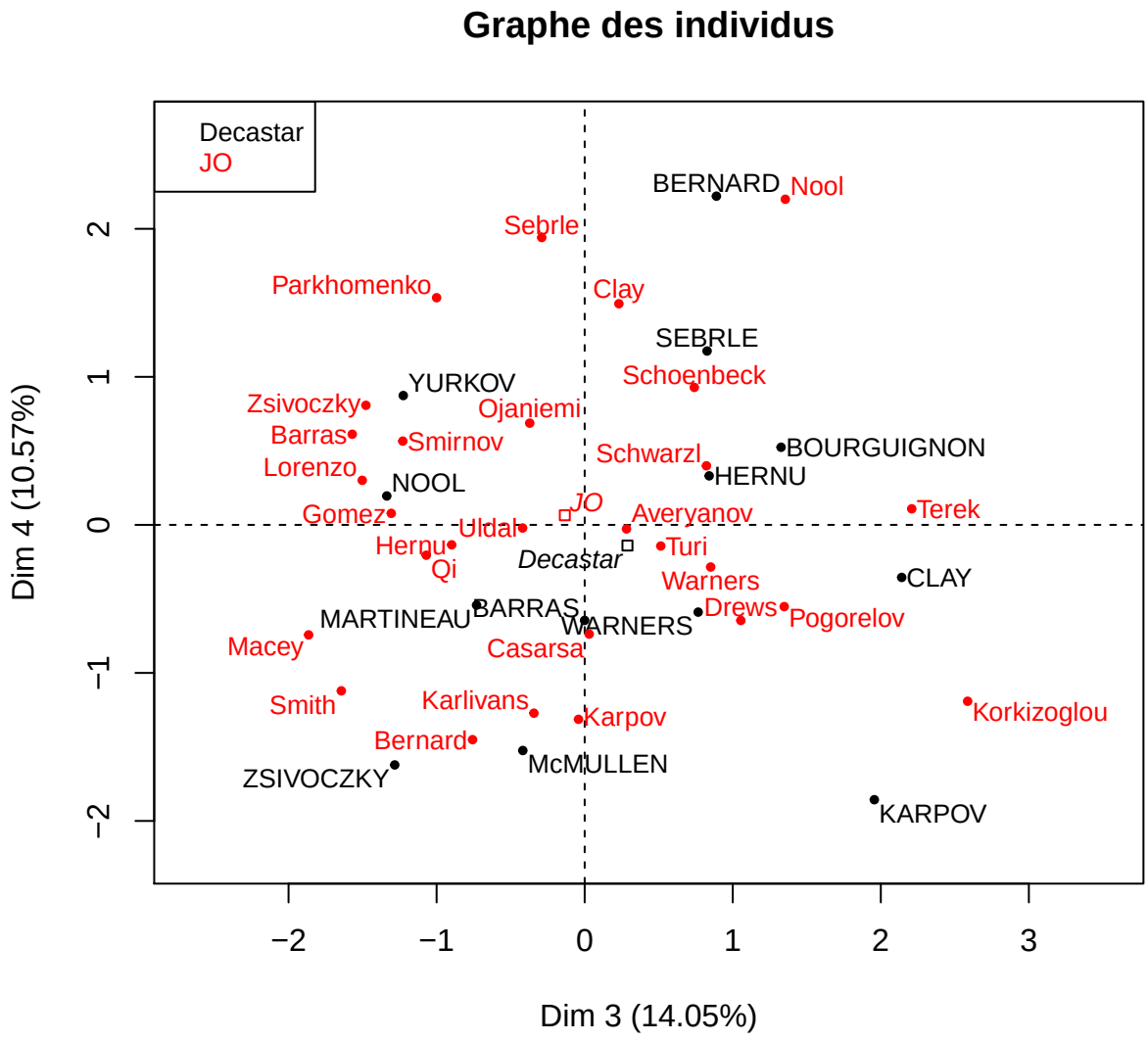
```
plotellipses(res)
```



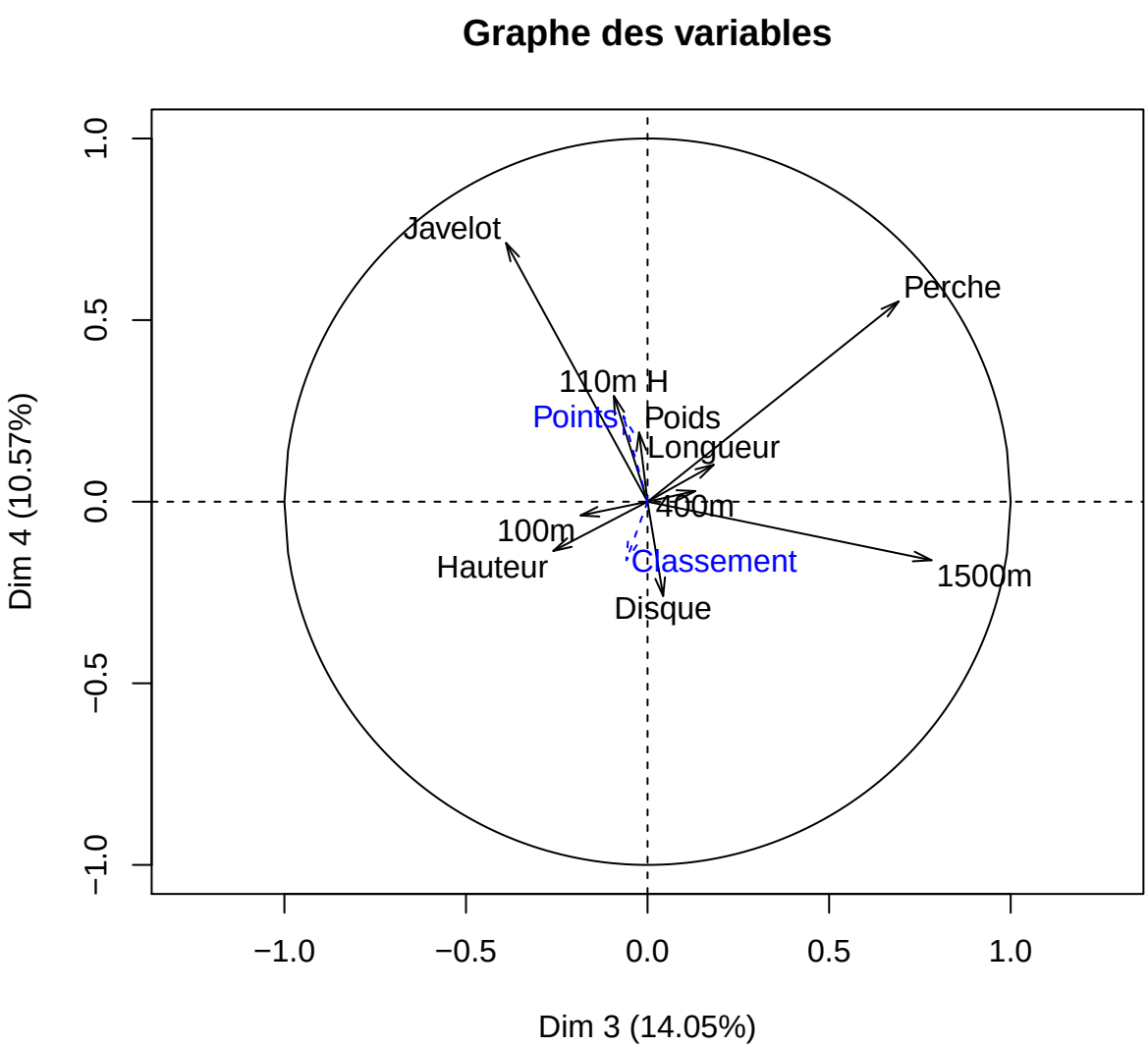
Si on a plusieurs variables qualitatives, on aura autant de graphes que de variables qualitatives. Avec sur chaque graphe, les ellipses de confiance des modalités de la variable qualitative en question.

Graphes sur les dimensions 3 et 4

```
plot(res, choix="ind", cex=0.8, habillage=13, title="Graphe des individus", axes=3:4)
```




```
plot(res, choix="var", title="Graphe des variables", axes=3:4)
```


Graphe des variables
Javelot
Perche
110m H
Points
Poids
Longueur
400m
100m
Hauteur
Classement
1500m
Disque
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%)

Graphe avec sélection des individus

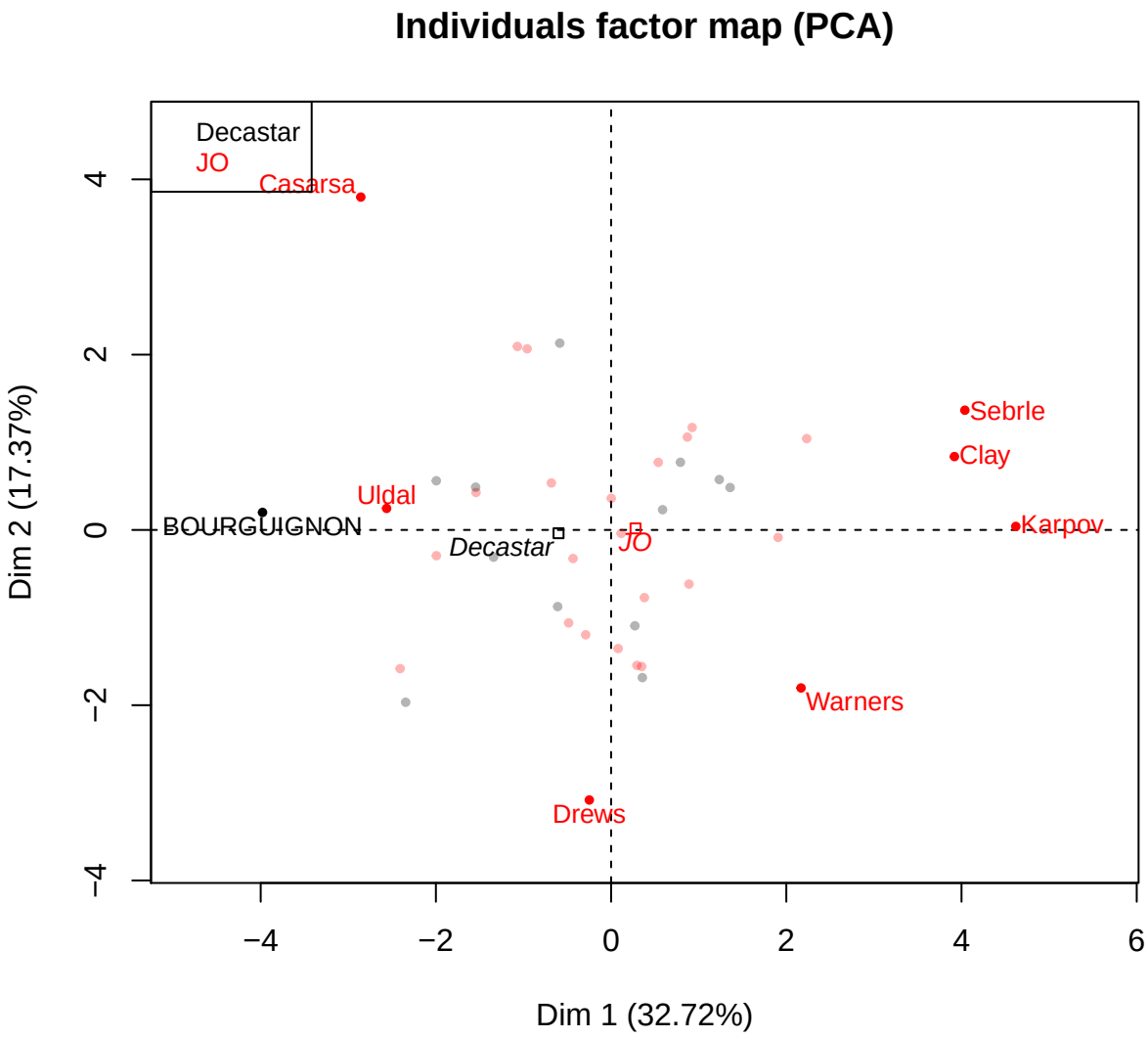
`select="cos2 0.7"` : sélectionne les individus qui ont, sur le plan tracé, une qualité de projection supérieure à 0.7

`select="cos2 5"` : sélectionne les 5 individus qui ont la meilleure qualité de projection sur le plan tracé

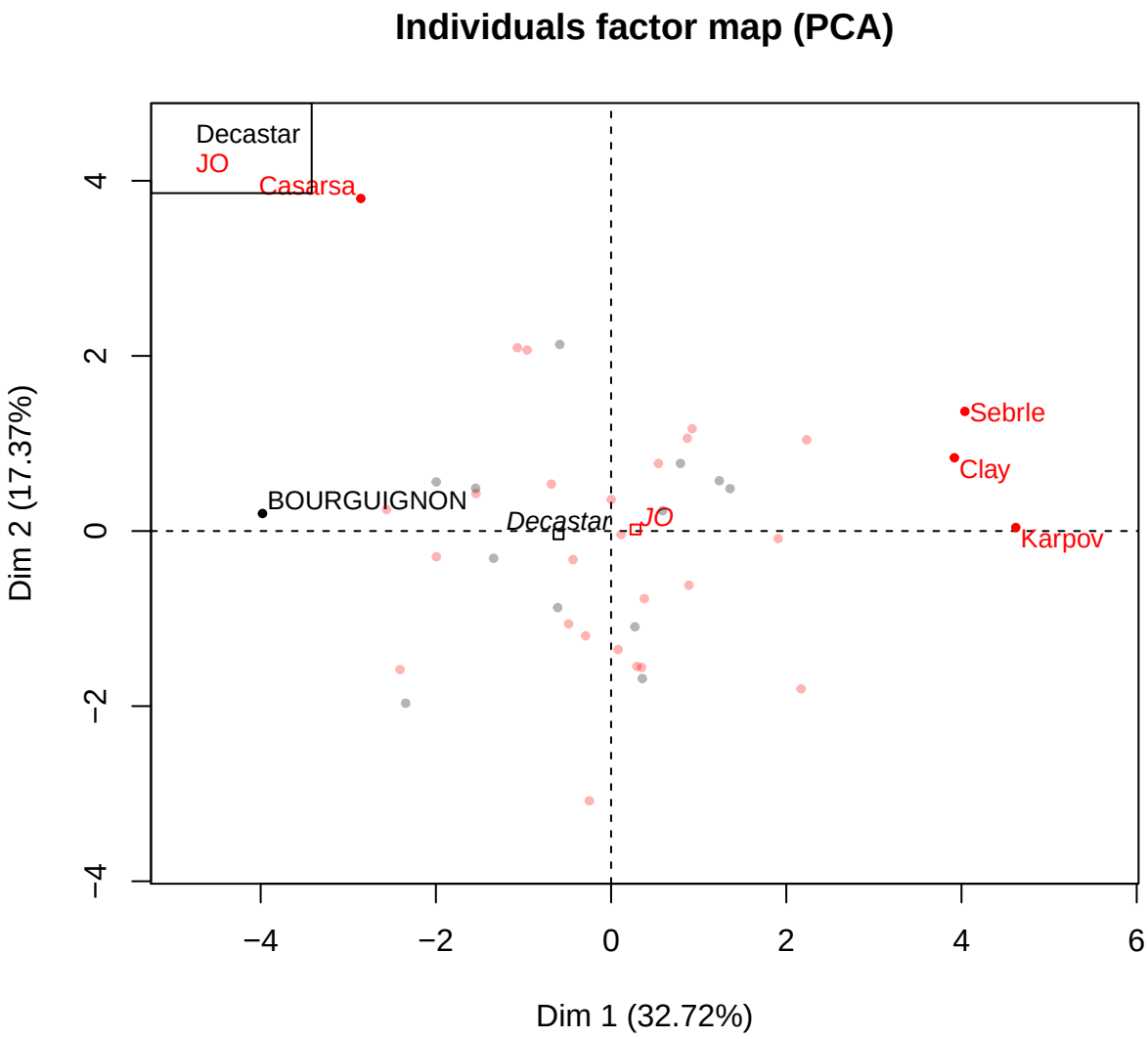
`select="contrib 5"` : sélectionne les 5 individus qui ont le plus contribué à la construction du plan tracé

`select=c("nom1","nom2")` : sélectionne les individus par leur nom

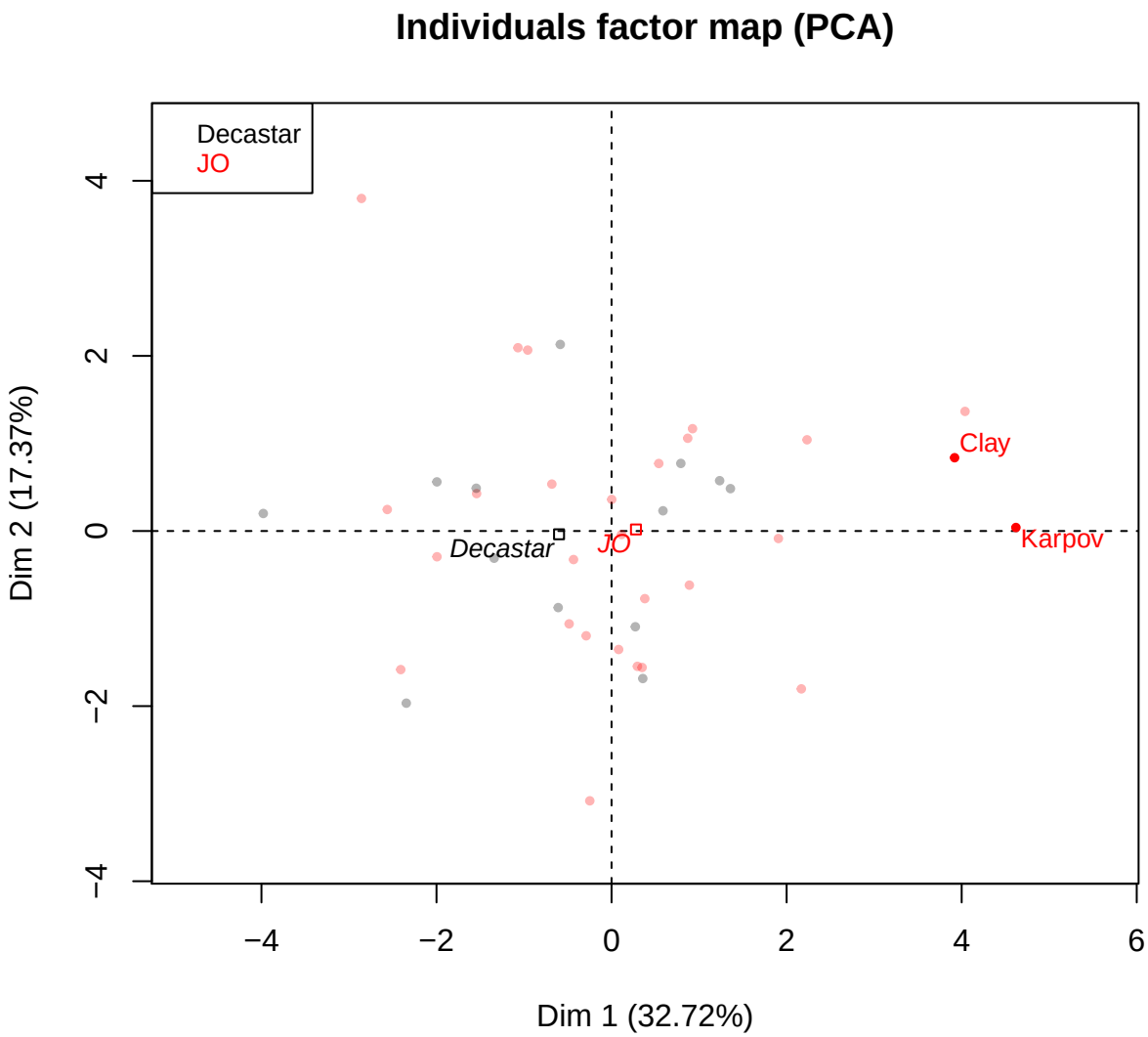
```
plot(res, cex=0.8, habillage=13, select="cos2 0.7")
```




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

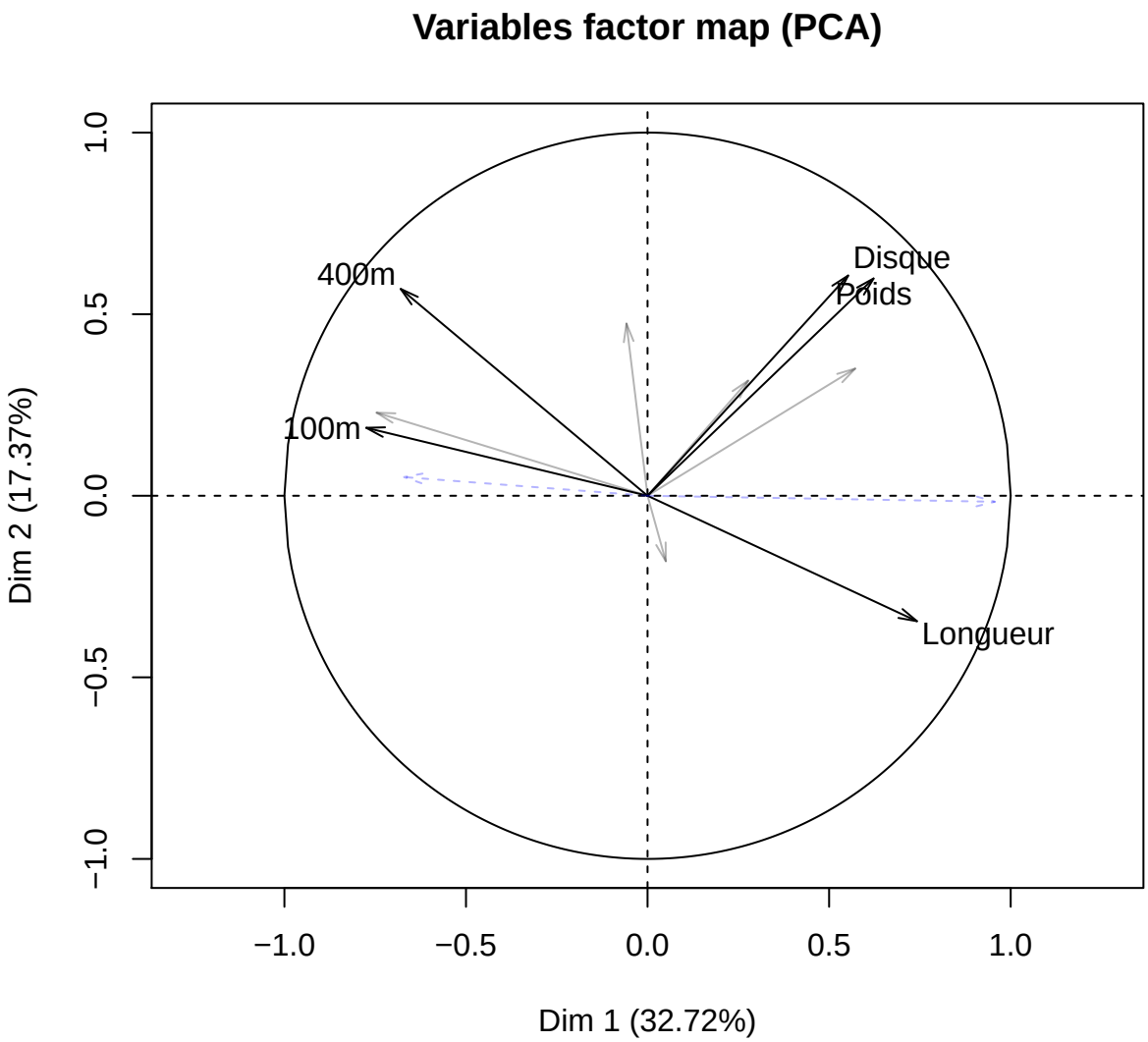



```
plot(res, cex=0.8, habillage=13, select=c("Clay","Karpov"))
```



Graphe du cercle des corrélations avec sélection des variables

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


Variables factor map (PCA)
400m
100m
Disque
Poids
Longueur
Dim 1 (32.72%)
Dim 2 (17.37%)
−1.0
−0.5
0.0
0.5
1.0

Graphe avec sélection des individus, des tailles de police différentes pour les titres, des ombres sous les points

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