

# Principal Component Analysis

*Dataset decathlon*

This dataset contains 41 individuals and 13 variables, 2 quantitative variables are considered as illustrative, 1 qualitative variable is considered as illustrative.

---

## 1. Study of the outliers

The analysis of the graphs does not detect any outlier.

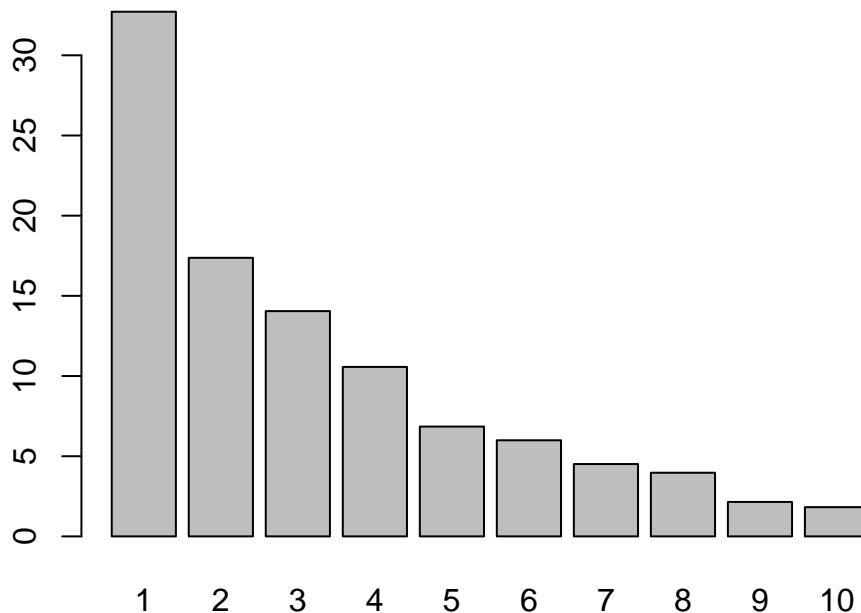
---

## 2. Inertia distribution

The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied.

The first two dimensions of PCA express **50.09%** of the total dataset inertia ; that means that 50.09% of the individuals (or variables) cloud total variability is explained by the plane. This percentage is relatively high and thus the first plane well represents the data variability. This value is greater than the reference value that equals **37.87%**, the variability explained by this plane is thus significant (the reference value is the 0.95-quantile of the inertia percentages distribution obtained by simulating 1552 data tables of equivalent size on the basis of a normal distribution).

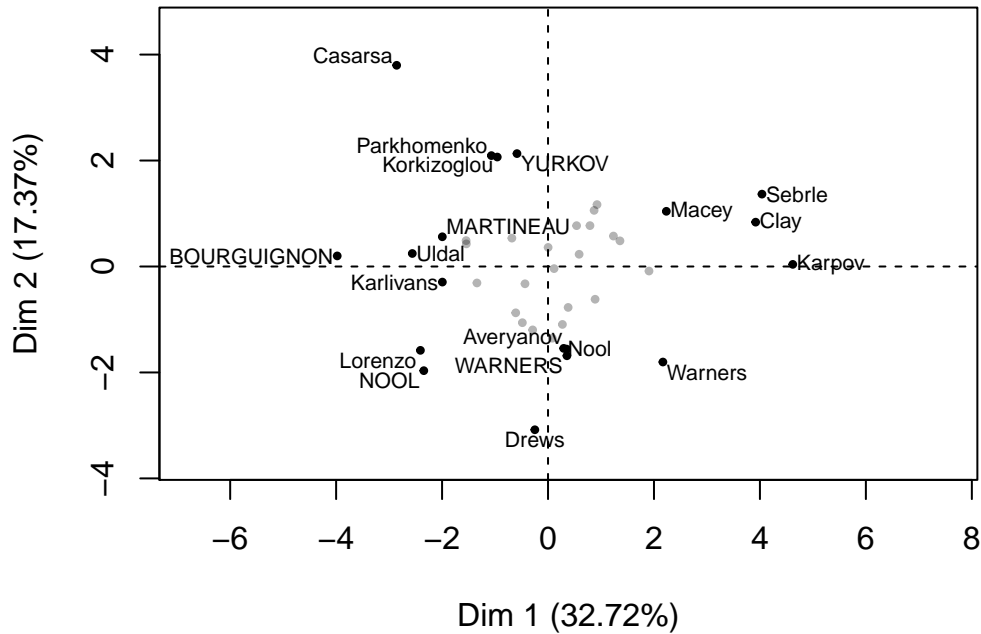
From these observations, it should be better to also interpret the dimensions greater or equal to the third one.



## Figure 2 - Decomposition of the total inertia on the components of the PCA

An estimation of the right number of axis to interpret suggests to restrict the analysis to the description of the first 3 axis. These axis present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (64.14% against 51.62%). This observation suggests that only these axis are carrying a real information. As a consequence, the description will stand to these axis.

### 3. Description of the plane 1:2

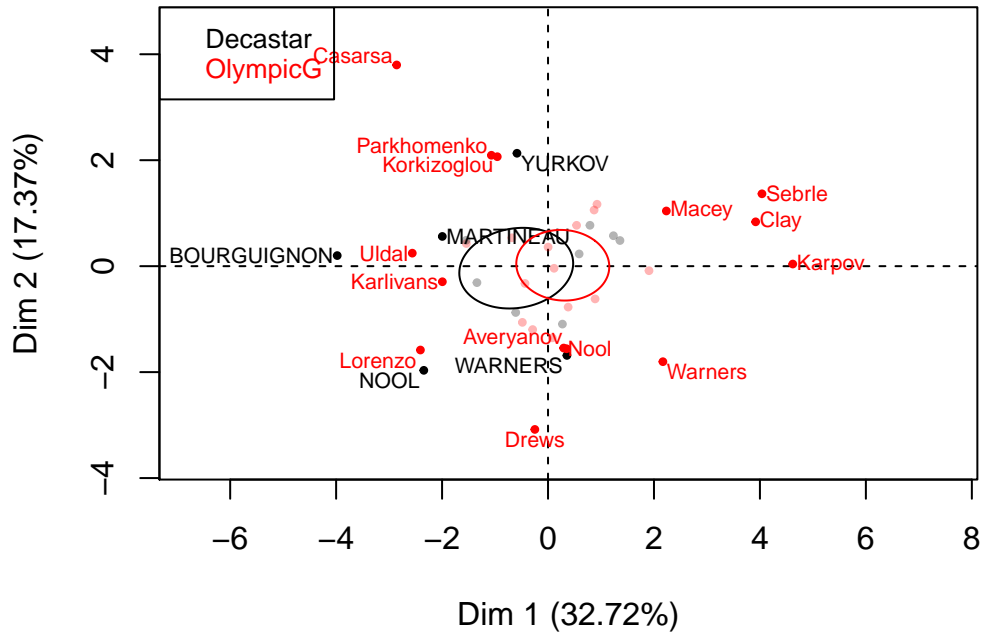


**Figure 3.1 - Individuals factor map (PCA)** The labeled individuals are those with the higher contribution to the plane construction.

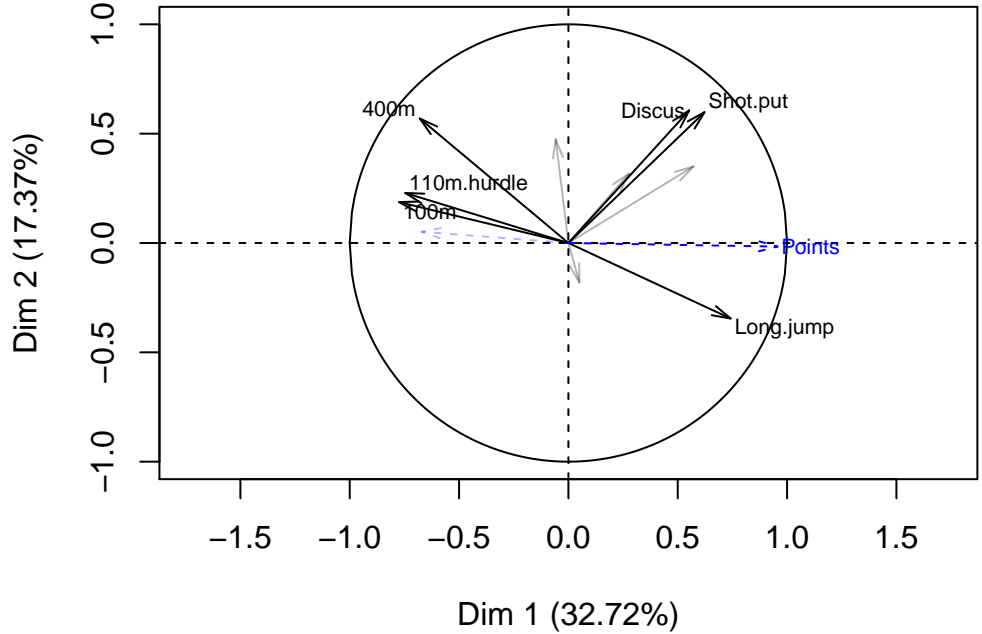
The Wilks test p-value indicates which variable factors are the best separated on the plane (i.e. which one explain the best the distance between individuals).

Competition  
0.366311

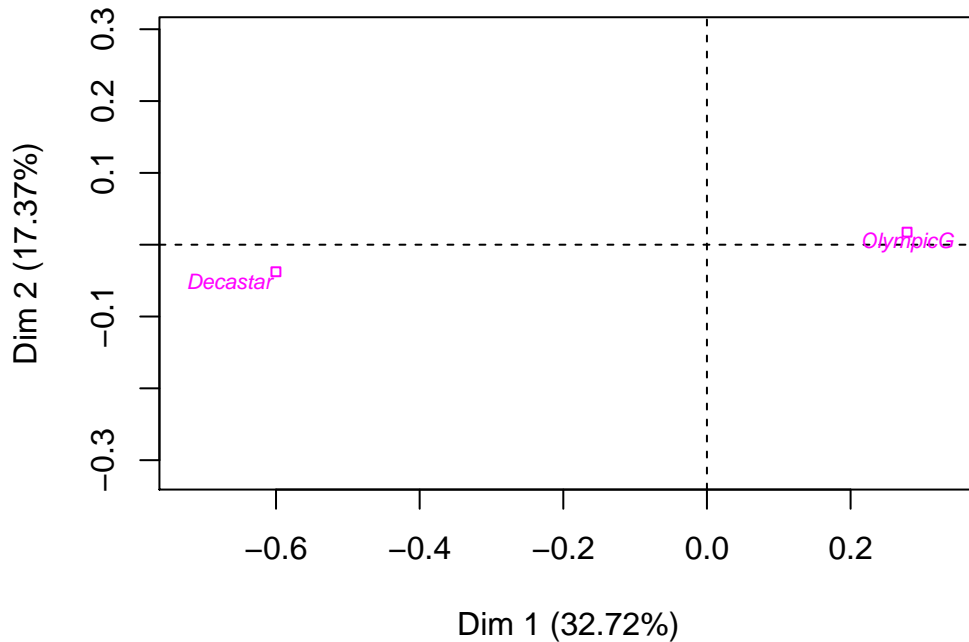
There only is one possible qualitative variable to illustrate the distance between individuals : *Competition*.



**Figure 3.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction. The individuals are coloured after their category for the variable Competition.*



**Figure 3.3 - Variables factor map (PCA)** *The variables in black are considered as active whereas those in blue are illustrative. The labeled variables are those the best shown on the plane.*



**Figure 3.4 - Qualitative factor map (PCA)** The labeled factors are those the best shown on the plane.

---

The **dimension 1** opposes individuals such as *Karpov*, *Sebrle*, *Clay* and *Macey* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *BOURGUIGNON*, *Uldal*, *Lorenzo*, *NOOL* and *Karlivans* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *Karpov*, *Sebrle*, *Clay* and *Macey* stand (characterized by a positive coordinate on the axis) is sharing :

- high values for the variables *Points*, *High.jump*, *Discus*, *Shot.put* and *Long.jump* (variables are sorted from the strongest).
- low values for the variables *400m*, *110m.hurdle*, *Rank* and *100m* (variables are sorted from the weakest).

The group in which the individuals *BOURGUIGNON*, *Uldal*, *Lorenzo*, *NOOL* and *Karlivans* stand (characterized by a negative coordinate on the axis) is sharing :

- high values for the variables *110m.hurdle*, *100m* and *Rank* (variables are sorted from the strongest).
- low values for the variables *Discus*, *High.jump*, *Points* and *Shot.put* (variables are sorted from the weakest).

Note that the variable *Points* is highly correlated with this dimension (correlation of 0.91). This variable could therefore summarize itself the dimension 1.

---

The **dimension 2** opposes individuals such as *Casarsa*, *YURKOV* and *Parkhomenko* (to the top of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *Warners*, *Dreus* and *WARNERS* (to the bottom of the graph, characterized by a strongly negative coordinate on the axis).

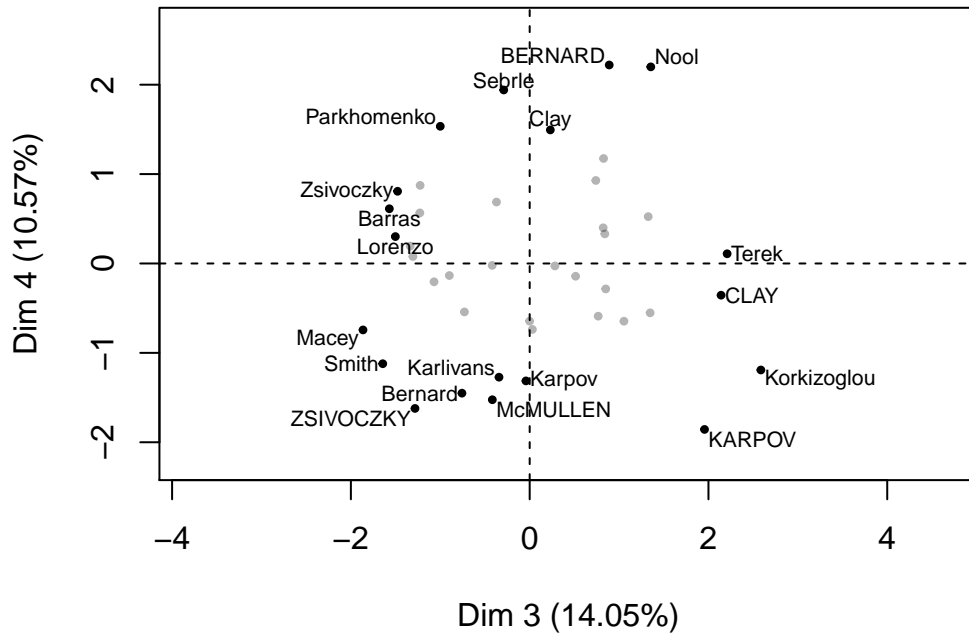
The group in which the individuals *Casarsa*, *YURKOV* and *Parkhomenko* stand (characterized by a positive coordinate on the axis) is sharing :

- high values for the variable *400m*.
- low values for the variable *Long.jump*.

The group in which the individuals *Warners*, *Drews* and *WARNERS* stand (characterized by a negative coordinate on the axis) is sharing :

- high values for the variable *Pole.vault*.
- low values for the variable *110m.hurdle*.

#### 4. Description of the dimension 3

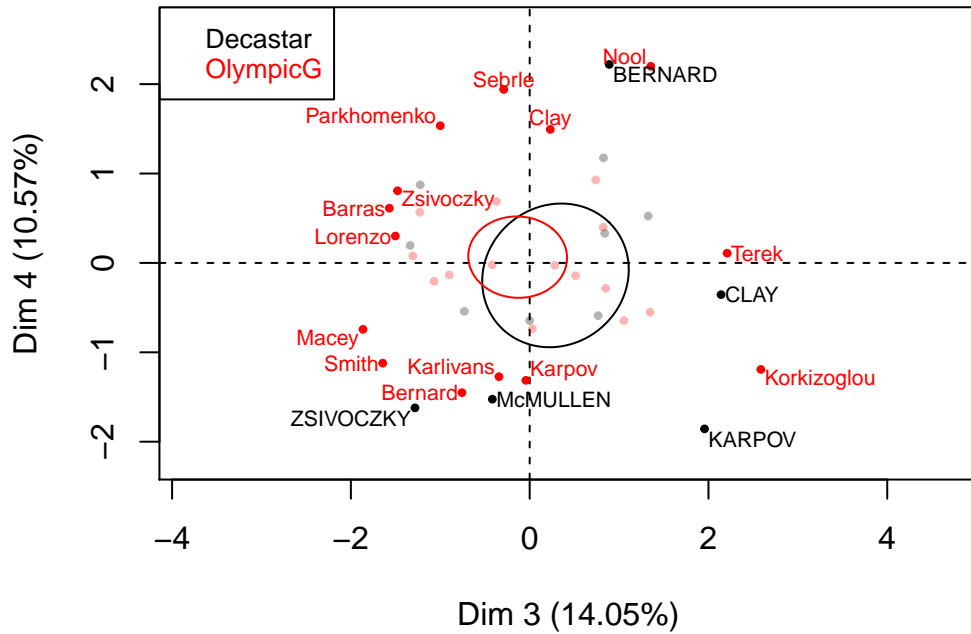


**Figure 4.1 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

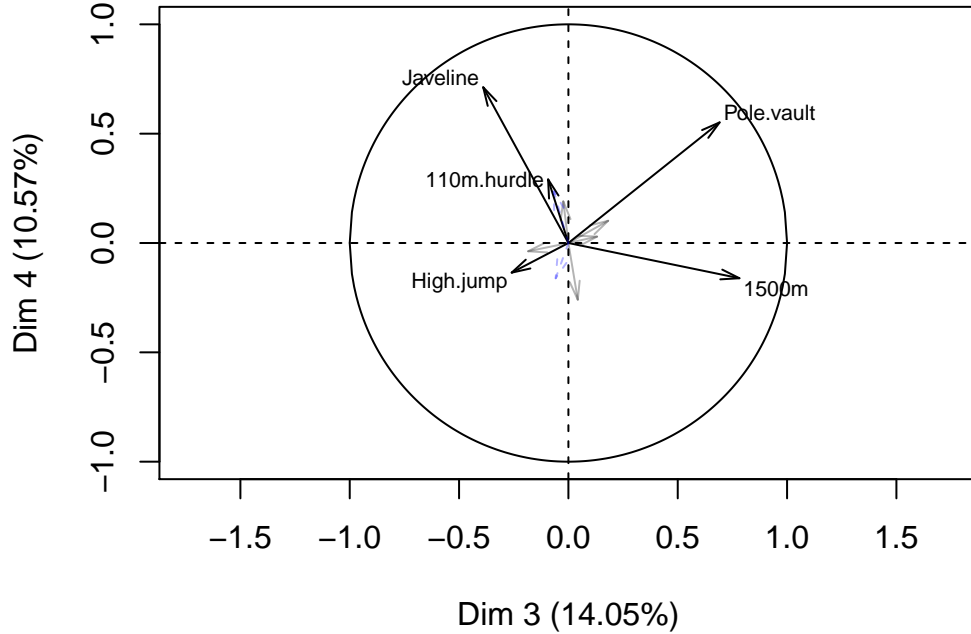
The Wilks test p-value indicates which variable factors are the best separated on the plane (i.e. which one explain the best the distance between individuals).

Competition  
0.496903

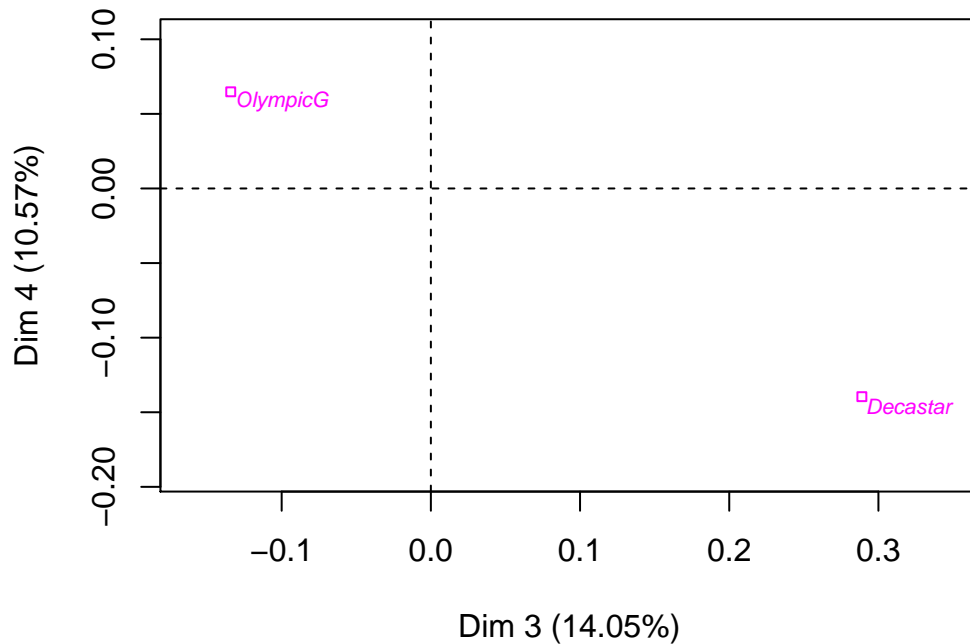
There only is one possible qualitative variable to illustrate the distance between individuals : *Competition*.



**Figure 4.2 - Individuals factor map (PCA)** The labeled individuals are those with the higher contribution to the plane construction. The individuals are coloured after their category for the variable Competition.



**Figure 4.3 - Variables factor map (PCA)** The variables in black are considered as active whereas those in blue are illustrative. The labeled variables are those the best shown on the plane.



**Figure 4.4 - Qualitative factor map (PCA)** *The labeled factors are those the best shown on the plane.*

---

The **dimension 3** opposes individuals such as *KARPOV*, *Korkizoglou*, *Terek* and *CLAY* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *ZSIVOCZKY*, *Barras*, *Zsivoczky*, *McMULLEN*, *Macey*, *Bernard* and *Smith* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

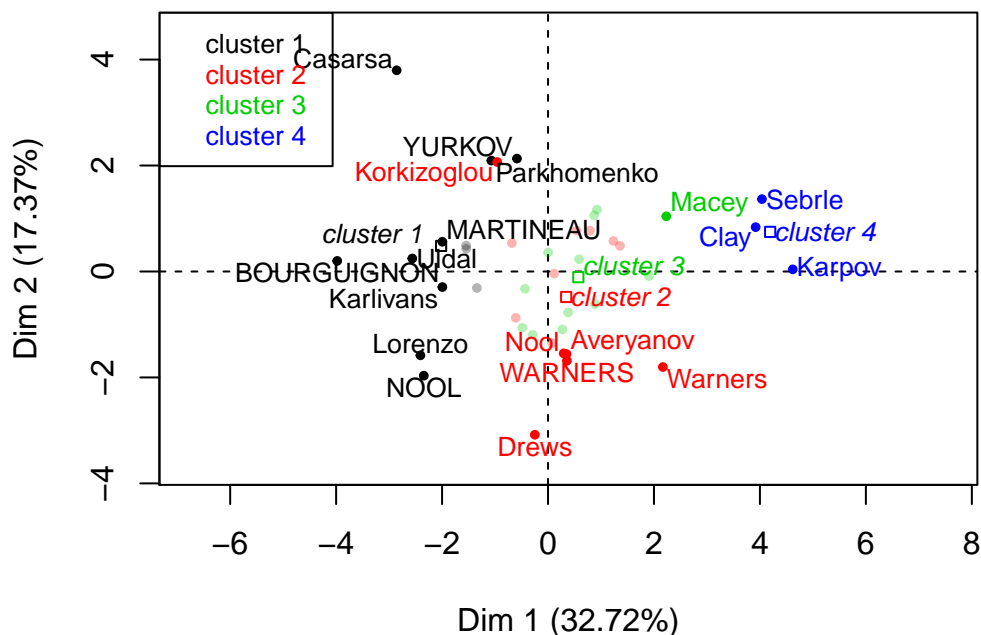
The group in which the individuals *KARPOV*, *Korkizoglou*, *Terek* and *CLAY* stand (characterized by a positive coordinate on the axis) is sharing :

- high values for the variable *1500m*.
- low values for the variable *Javeline*.

The group in which the individuals *ZSIVOCZKY*, *Barras*, *Zsivoczky*, *McMULLEN*, *Macey*, *Bernard* and *Smith* stand (characterized by a negative coordinate on the axis) is sharing :

- low values for the variables *1500m* and *Pole.vault* (variables are sorted from the weakest).
-

## 5. Classification



**Figure 5 - Ascending Hierarchical Classification of the individuals.** The classification made on individuals reveals 4 clusters.

The **cluster 1** is made of individuals such as *YURKOV*, *MARTINEAU*, *NOOL*, *BOURGUIGNON*, *Parkhomenko*, *Lorenzo*, *Karlivans*, *Udal* and *Casarsa*. This group is characterized by :

- high values for the variables *100m*, *110m.hurdle*, *400m* and *Rank* (variables are sorted from the strongest).
- low values for the variables *Shot.put*, *Long.jump* and *Points* (variables are sorted from the weakest).

The **cluster 2** is made of individuals such as *WARNERS*, *Warners*, *Nool*, *Averyanov*, *Drews* and *Korkizoglou*. This group is characterized by :

- high values for the variables *Pole.vault* and *1500m* (variables are sorted from the strongest).
- low values for the variables *100m* and *Javeline* (variables are sorted from the weakest).

The **cluster 3** is made of individuals such as *Macey*. This group is characterized by :

- low values for the variables *1500m* and *Pole.vault* (variables are sorted from the weakest).

The **cluster 4** is made of individuals such as *Sebrle*, *Clay* and *Karpov*. This group is characterized by :

- high values for the variables *Points*, *Long.jump*, *Discus*, *Shot.put*, *Javeline* and *High.jump* (variables are sorted from the strongest).
- low values for the variables *110m.hurdle*, *Rank*, *400m* and *100m* (variables are sorted from the weakest).



## Annexes

```
dimdesc(res, axes = 1:3)
```

```
$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.hurdle -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
```

Figure 6 - List of variables characterizing the dimensions of the analysis.

```
res.hcpc$desc.var
```

```
$quanti.var
      Eta2      P-value
Points      0.7438620 4.908988e-11
100m        0.6581552 9.668613e-09
Pole.vault  0.5712977 5.972228e-07
Long.jump   0.5293255 3.246949e-06
110m.hurdle 0.4455078 6.229435e-05
400m        0.4425235 6.859144e-05
Shot.put    0.2869412 5.393490e-03
Discus      0.2777274 6.745695e-03
Rank        0.2693094 8.250830e-03
```

1500m	0.2602251	1.022178e-02
Javeline	0.2500899	1.293207e-02
High.jump	0.2255099	2.250558e-02

\$quanti

\$quanti\$`1`

	v.test	Mean in category	Overall mean	sd in category
100m	4.741585	11.300833	10.99805	0.1445947
110m.hurdle	3.964894	15.060000	14.60585	0.3798903
400m	3.822084	50.686667	49.61634	1.0702051
Rank	2.667277	17.250000	12.12195	7.7041655
Shot.put	-2.100392	14.056667	14.47707	0.8698116
Long.jump	-3.406381	6.998333	7.26000	0.2586450
Points	-4.131722	7661.916667	8005.36585	196.1718882
	Overall sd	p.value		
100m	0.2597956	2.120526e-06		
110m.hurdle	0.4660000	7.342867e-05		
400m	1.1392975	1.323286e-04		
Rank	7.8217805	7.646858e-03		
Shot.put	0.8143118	3.569438e-02		
Long.jump	0.3125193	6.583024e-04		
Points	338.1839416	3.600552e-05		

\$quanti\$`2`

	v.test	Mean in category	Overall mean	sd in category
Pole.vault	4.295481	5.021429	4.762439	0.1919024
1500m	2.602164	285.612857	279.024878	12.7576030
100m	-1.969217	10.885714	10.998049	0.1539812
Javeline	-2.125561	56.091429	58.316585	4.5043580
	Overall sd	p.value		
Pole.vault	0.2745887	1.743152e-05		
1500m	11.5300118	9.263766e-03		
100m	0.2597956	4.892823e-02		
Javeline	4.7675931	3.353983e-02		

\$quanti\$`3`

	v.test	Mean in category	Overall mean	sd in category
1500m	-2.893339	270.825000	279.024878	5.8957039
Pole.vault	-3.715512	4.511667	4.762439	0.1635967
	Overall sd	p.value		
1500m	11.5300118	0.0038117012		
Pole.vault	0.2745887	0.0002027925		

\$quanti\$`4`

	v.test	Mean in category	Overall mean	sd in category
Points	4.242103	8812.66667	8005.365854	68.78145745
Long.jump	3.468581	7.87000	7.260000	0.06480741
Discus	3.107539	50.16000	44.325610	1.19668988
Shot.put	2.974272	15.84000	14.477073	0.46568945
Javeline	2.586808	65.25667	58.316585	6.87867397
High.jump	2.289003	2.09000	1.976829	0.02449490
110m.hurdle	-2.119695	14.05000	14.605854	0.06531973
Rank	-2.299627	2.00000	12.121951	0.81649658
400m	-2.333955	48.12000	49.616341	0.98634004

100m	-2.745523	10.59667	10.998049	0.18080069
	Overall sd	p.value		
Points	338.18394159	2.214348e-05		
Long.jump	0.31251927	5.232144e-04		
Discus	3.33639725	1.886523e-03		
Shot.put	0.81431175	2.936847e-03		
Javeline	4.76759315	9.686955e-03		
High.jump	0.08785906	2.207917e-02		
110m.hurdle	0.46599998	3.403177e-02		
Rank	7.82178048	2.146935e-02		
400m	1.13929751	1.959810e-02		
100m	0.25979560	6.041458e-03		

```
attr("class")
[1] "catdes" "list "
```

**Figure 7 - List of variables characterizing the clusters of the classification.**