Principal Component Analysis

Dataset dataset

This dataset contains 166 individuals and 27 variables, 1 qualitative variable is considered as illustrative.

### 1. Study of the outliers

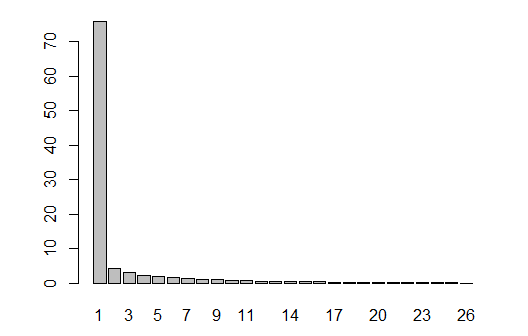
The analysis of the graphs does not detect any outlier.

### 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 **80.39%** of the total dataset inertia ; that means that 80.39% of the individuals (or variables) cloud total variability is explained by the plane. This percentage is high and thus the first plane represents an important part of the data variability. This value is strongly greater than the reference value that equals **14.44%**, the variability explained by this plane is thus highly significant (the reference value is the 0.95-quantile of the inertia percentages distribution obtained by simulating 1007 data tables of equivalent size on the basis of a normal distribution).

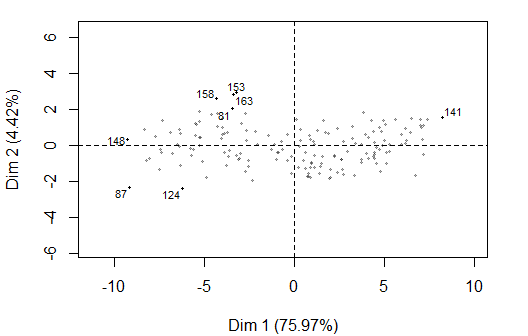
From these observations, it is probably not useful to interpret the next dimensions.



**Figure 2 - Decomposition of the total inertia on the components of the PCA** *The first factor is major: it expresses itself 75.97% of the data variability.* *Note that in such a case, the variability related to the other components might be meaningless, despite of a high percentage.*

An estimation of the right number of axis to interpret suggests to restrict the analysis to the description of the first 1 axis. These axis present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (75.97% against 7.63%). This observation suggests that only this axis is carrying a real information. As a consequence, the description will stand to these axis.

### 3. Description of the dimension 1

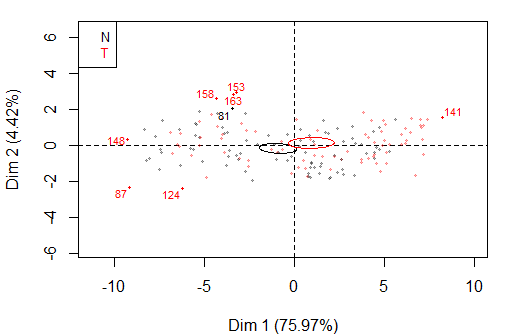


**Figure 3.1 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

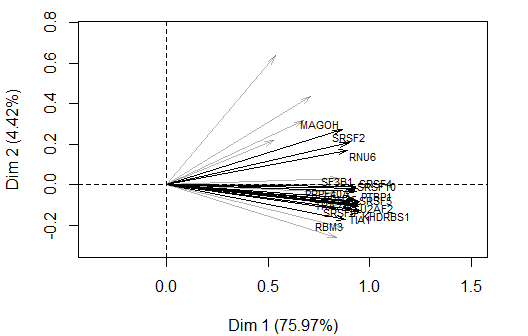
The Wilks test p-value indicates which variable factors are the best separated on the plane (i.e. which one explain the best the distance between individuals).

## Class   
## 0.004917444

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



**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* Class.



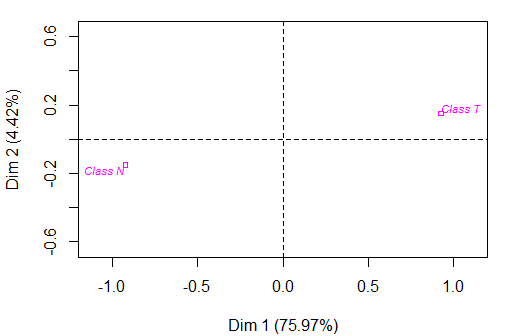
**Figure 3.3 - Variables factor map (PCA)** *The labeled variables are those the best shown on the plane.*

## Warning in if (grepl("contrib", select)) selection <-  
## (rev(order(res.pca$ind$contrib[, : the condition has length > 1 and only  
## the first element will be used

## Warning in if (grepl("dist", select)) selection <-  
## (rev(order(res.pca$ind$dist)))[1:min(nrow(res.pca$ind$coord), : the  
## condition has length > 1 and only the first element will be used

## Warning in if (grepl("coord", select)) selection <-  
## (rev(order(apply(res.pca$ind$coord[, : the condition has length > 1 and  
## only the first element will be used

## Warning in if (grepl("cos2", select)) {: the condition has length > 1 and  
## only the first element will be used



**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 *141* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *1241481588715381* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group 1 (characterized by a positive coordinate on the axis) is sharing :

* high values for variables like *SRSF3*, *KHDRBS1*, *RBM45*, *SRSF4*, *RBM3*, *TIA1*, *U2AF2*, *SRSF10*, *SRSF9* and *SRSF5* (variables are sorted from the strongest).

The group in which the individual *141* stands (characterized by a positive coordinate on the axis) is sharing :

* high values for variables like *ESRP1*, *PRPF40A*, *ESRP2*, *SF3B1*, *SND1*, *PTBP1*, *MAGOH*, *TRA2A*, *SRRM1* and *SRSF2* (variables are sorted from the strongest).

The group in which the individuals *124*, *148* and *87* stand (characterized by a negative coordinate on the axis) is sharing :

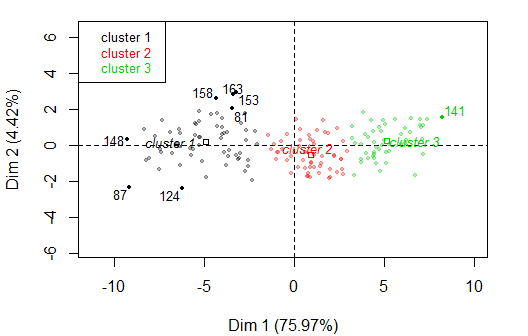
* low values for variables like *SND1*, *SRSF2*, *PRPF40A*, *SRSF4*, *SF3B1*, *SRSF10*, *SRSF5*, *RNU6*, *KHDRBS1* and *SRSF9* (variables are sorted from the weakest).

The group in which the individuals *158*, *153* and *81* stand (characterized by a negative coordinate on the axis) is sharing :

* low values for variables like *SRRM1*, *TIA1*, *U2AF2*, *RBM45*, *KHDRBS1*, *NOVA1*, *SRSF4*, *PTBP1*, *SRSF9* and *SRSF5* (variables are sorted from the weakest).

Note that the variables *KHDRBS1*, *PTBP1*, *RBM45*, *SRSF9* and *TIA1* are highly correlated with this dimension (respective correlation of 0.92, 0.91, 0.91, 0.9, 0.91). These variables could therefore summarize themselve the dimension 1.

### 4. Classification



**Figure 4 - Ascending Hierarchical Classification of the individuals.** *The classification made on individuals reveals 3 clusters.*

The **cluster 1** is made of individuals such as *81*, *87*, *124*, *148*, *153*, *158* and *163*. This group is characterized by :

* low values for variables like *U2AF2*, *TIA1*, *RBM45*, *SRSF10*, *SRSF4*, *KHDRBS1*, *SRSF5*, *PTBP1*, *SRSF9* and *RBM3* (variables are sorted from the weakest).

The **cluster 2** is made of individuals sharing :

* high values for variables like *SRSF3*, *SRSF10*, *SRSF4*, *U2AF2*, *RBM45*, *RNU4*, *RBM3*, *TIA1*, *SRSF5* and *NOVA1* (variables are sorted from the strongest).

The **cluster 3** is made of individuals such as *141*. This group is characterized by :

* high values for variables like *SF3B1*, *KHDRBS1*, *SRSF2*, *SRSF9*, *PTBP1*, *TIA1*, *MAGOH*, *RNU6*, *PRPF40A* and *SRRM1* (variables are sorted from the strongest).

## Annexes

dimdesc(res, axes = 1:1)

$Dim.1  
$Dim.1$quanti  
 correlation p.value  
KHDRBS1 0.9584411 3.014040e-91  
TIA1 0.9565263 1.120263e-89  
PTBP1 0.9538029 1.461324e-87  
RBM45 0.9524784 1.405727e-86  
SRSF9 0.9500585 7.474649e-85  
U2AF2 0.9482971 1.192877e-83  
SRSF5 0.9447794 2.279526e-81  
SRSF4 0.9438857 8.197343e-81  
SRSF10 0.9360125 2.808573e-76  
SF3B1 0.9335646 5.514297e-75  
PRPF40A 0.9224009 1.180114e-69  
TRA2A 0.9187433 4.426157e-68  
SRSF2 0.9014697 1.571696e-61  
RNU6 0.8953710 1.673374e-59  
RBM3 0.8857288 1.534415e-56  
SND1 0.8796668 8.222501e-55  
SRRM1 0.8726961 6.187399e-53  
MAGOH 0.8673266 1.455959e-51  
ESRP2 0.8531581 3.253765e-48  
NOVA1 0.8525604 4.424311e-48  
SRSF3 0.8408525 1.403480e-45  
RNU4 0.8325370 6.367680e-44  
ESRP1 0.7109606 7.305416e-27  
SNRNP200 0.6765196 1.477647e-23  
RNU12 0.5393992 6.474615e-14  
SRRM4 0.5280980 2.635540e-13  
  
$Dim.1$quali  
 R2 p.value  
Class 0.04348444 0.007015347  
  
$Dim.1$category  
 Estimate p.value  
T 0.923996 0.007015347  
N -0.923996 0.007015347

**Figure 5 - List of variables characterizing the dimensions of the analysis.**