Principal Component Analysis

Dataset clusterMyData

This dataset contains 18 individuals and 6 variables, 1 qualitative variable is considered as illustrative.

### 1. Study of the outliers

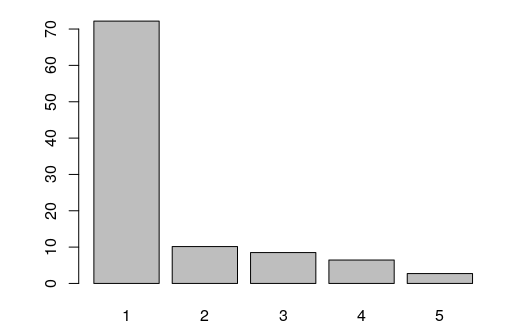
The analysis of the graphs does not detect any outlier.

### 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 **82.33%** of the total dataset inertia ; that means that 82.33% 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 greater than the reference value that equals **71.89%**, 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 1588 data tables of equivalent size on the basis of a normal distribution).

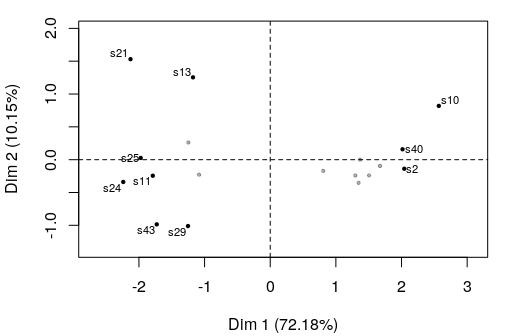
From these observations, it is probably not useful to interpret the next dimensions.



**Figure 2 - Decomposition of the total inertia on the components of the PCA** *The first factor is major: it expresses itself 72.18% 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 (72.18% against 47.13%). 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

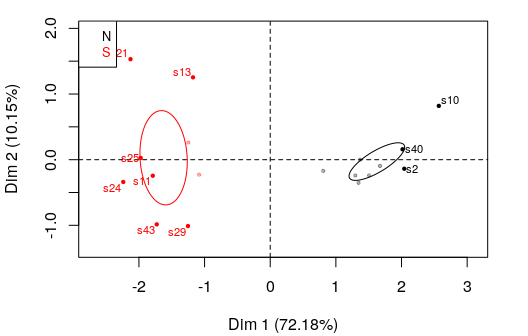


**Figure 3.1 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

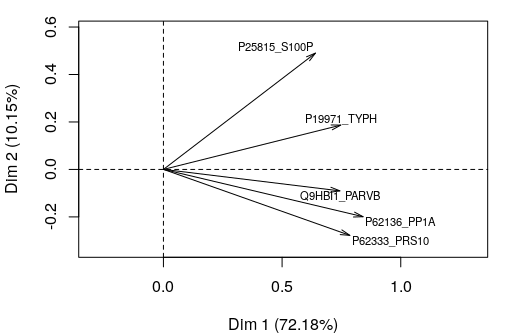
The Wilks test p-value indicates which variable factors are the best separated on the plane (i.e. which one explain the best the distance between individuals).

## Class   
## 2.307357e-09

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.



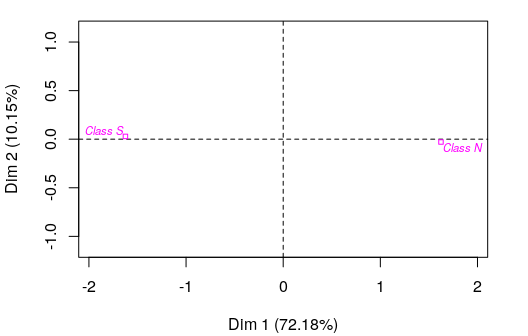
**Figure 3.3 - Variables factor map (PCA)** *The labeled variables are those the best shown on the plane.*

## 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 *s10*, *s2* and *s40* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *s25* and *s24* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *s10*, *s2* and *s40* stand (characterized by a positive coordinate on the axis) is sharing :

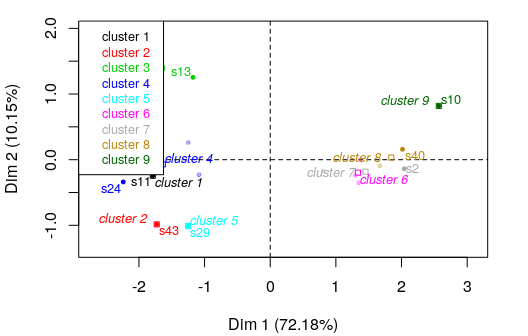
* high values for the variables *P62333\_PRS10*, *P62136\_PP1A*, *Q9HBI1\_PARVB*, *P19971\_TYPH* and *P25815\_S100P* (variables are sorted from the strongest).

The group in which the individuals *s25* and *s24* stand (characterized by a negative coordinate on the axis) is sharing :

* low values for the variables *P25815\_S100P* and *P19971\_TYPH* (variables are sorted from the weakest).

Note that the variables *Class N* and *Class S* are highly correlated with this dimension (respective correlation of 1, 1). 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 9 clusters.*

The **cluster 1** is made of individuals such as *s11*. This group is characterized by :

* variables whose values do not differ significantly from the mean.

The **cluster 2** is made of individuals such as *s43*. This group is characterized by :

* low values for the variable *P25815\_S100P*.

The **cluster 3** is made of individuals such as *s13* and *s21*. This group is characterized by :

* low values for the variable *P62333\_PRS10*.

The **cluster 4** is made of individuals such as *s24* and *s25*. This group is characterized by :

* low values for the variables *P62333\_PRS10*, *P62136\_PP1A* and *P25815\_S100P* (variables are sorted from the weakest).

The **cluster 5** is made of individuals such as *s29*. This group is characterized by :

* variables whose values do not differ significantly from the mean.

The **cluster 6** is made of individuals sharing :

* high values for the variable *P62333\_PRS10*.

The **cluster 7** is made of individuals such as *s2*. This group is characterized by :

* variables whose values do not differ significantly from the mean.

The **cluster 8** is made of individuals such as *s40*. This group is characterized by :

* high values for the variable *P19971\_TYPH*.

The **cluster 9** is made of individuals such as *s10*. This group is characterized by :

* variables whose values do not differ significantly from the mean.

## Annexes

dimdesc(res, axes = 1:1)

$Dim.1  
$Dim.1$quanti  
 correlation p.value  
Q9HBI1\_PARVB 0.9191316 7.119530e-08  
P62136\_PP1A 0.8791340 1.556999e-06  
P19971\_TYPH 0.8482586 8.679887e-06  
P62333\_PRS10 0.8427332 1.134519e-05  
P25815\_S100P 0.7506037 3.319994e-04  
  
$Dim.1$quali  
 R2 p.value  
Class 0.9273467 1.576339e-10  
  
$Dim.1$category  
 Estimate p.value  
N 1.624227 1.576339e-10  
S -1.624227 1.576339e-10

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

res.hcpc$desc.var

$test.chi2  
 p.value df  
Class 0.02122649 8  
  
$category  
$category$`1`  
NULL  
  
$category$`2`  
NULL  
  
$category$`3`  
NULL  
  
$category$`4`  
 Cla/Mod Mod/Cla Global p.value v.test  
Class=S 44.44444 100 50 0.04117647 2.041749  
Class=N 0.00000 0 50 0.04117647 -2.041749  
  
$category$`5`  
NULL  
  
$category$`6`  
NULL  
  
$category$`7`  
NULL  
  
$category$`8`  
NULL  
  
$category$`9`  
NULL  
  
  
$quanti.var  
 Eta2 P-value  
Q9HBI1\_PARVB 0.9670177 8.847889e-06  
P62333\_PRS10 0.9532680 4.099483e-05  
P25815\_S100P 0.9303080 2.334419e-04  
P19971\_TYPH 0.9234568 3.497067e-04  
P62136\_PP1A 0.9085254 7.498578e-04  
  
$quanti  
$quanti$`1`  
NULL  
  
$quanti$`2`  
 v.test Mean in category Overall mean sd in category  
P25815\_S100P -2.134206 -1.910012 -0.08907503 0  
 Overall sd p.value  
P25815\_S100P 0.8532154 0.03282595  
  
$quanti$`3`  
 v.test Mean in category Overall mean sd in category  
P62333\_PRS10 -2.172883 -1.43694 -0.04790587 0.09534456  
 Overall sd p.value  
P62333\_PRS10 0.9318715 0.02978912  
  
$quanti$`4`  
 v.test Mean in category Overall mean sd in category  
P25815\_S100P -2.089511 -0.8980087 -0.08907503 0.4060182  
P62136\_PP1A -2.102555 -1.0214674 -0.10870032 0.3215793  
P62333\_PRS10 -2.117088 -0.9430739 -0.04790587 0.2828247  
 Overall sd p.value  
P25815\_S100P 0.8532154 0.03666174  
P62136\_PP1A 0.9567600 0.03550469  
P62333\_PRS10 0.9318715 0.03425235  
  
$quanti$`5`  
NULL  
  
$quanti$`6`  
 v.test Mean in category Overall mean sd in category  
P62333\_PRS10 2.386444 1.158151 -0.04790587 0.3128199  
 Overall sd p.value  
P62333\_PRS10 0.9318715 0.01701221  
  
$quanti$`7`  
NULL  
  
$quanti$`8`  
 v.test Mean in category Overall mean sd in category  
P19971\_TYPH 2.48387 1.553014 0.0525921 0.1380768  
 Overall sd p.value  
P19971\_TYPH 0.8805703 0.01299633  
  
$quanti$`9`  
NULL  
  
  
attr(,"class")  
[1] "catdes" "list "

**Figure 6 - List of variables characterizing the clusters of the classification.**