Principal Component Analysis

Dataset Lymphocytes

This dataset contains 32 individuals and 55 variables. The quantitative variable “*Edad*” and the other 9 qualitative variables are considered as supplementary.

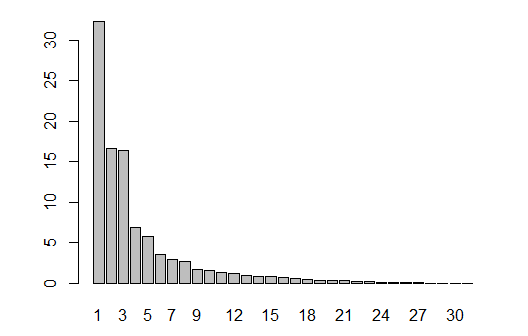
### 1. Study of the outliers

The analysis of the graphs does not detect any outlier.

### 2. Inertia distribution

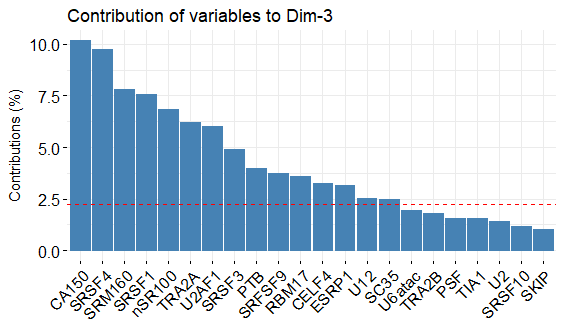
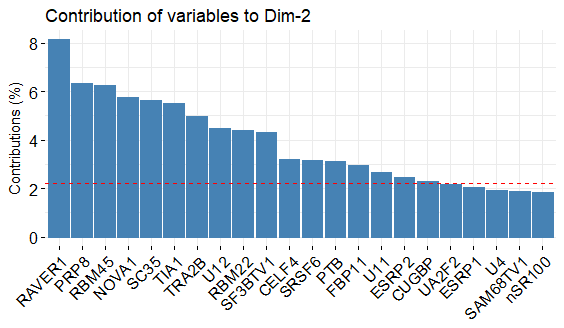
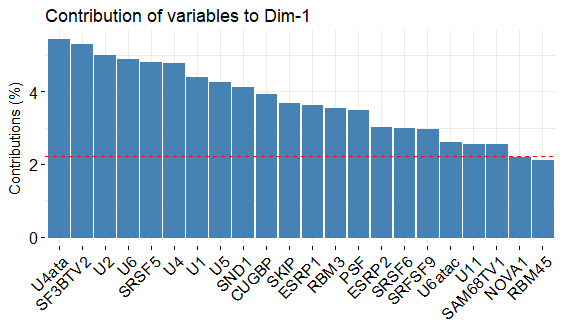
The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied. The first two dimensions of PCA express **49.02%** of the total dataset inertia; that means that 49.02% of the individuals (or variables) cloud total variability is explained by the plane. This is an intermediate percentage and the first plane represents a part of the data variability. This value is strongly greater than the reference value that equals **20.16%**, 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 1202 data tables of equivalent size on the basis of a normal distribution).

From these observations, it may be interesting to consider the next dimensions which also express a high percentage of the total inertia.



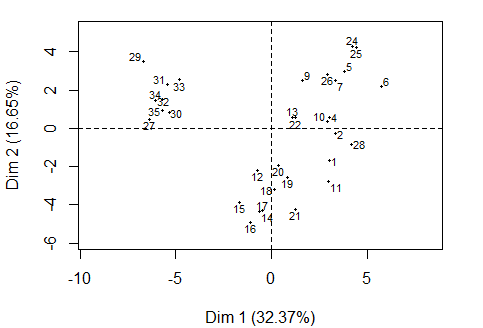
**Figure 2.1 - Decomposition of the total inertia by PCA components.**

An estimation of the right number of axes suggests restricting the analysis to the description of the first 3 axis. These axes present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (65.43% against 28.33%). This observation suggests that only these axes are carrying a real information. Therefore, the description will stand to these axes.

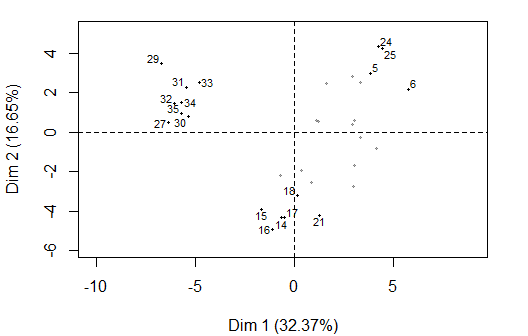


**Figure 2.2 -Contributions of variables to the dimensions.**

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



**Figure 3.1 - Individuals factor map (PCA)** *All individuals are plotted.*

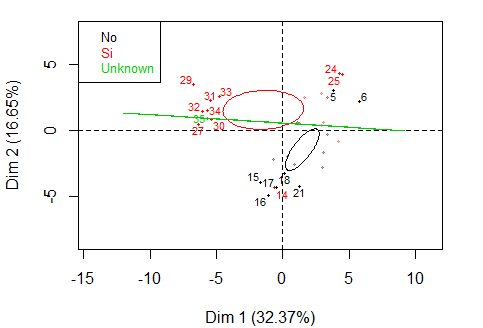


**Figure 3.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

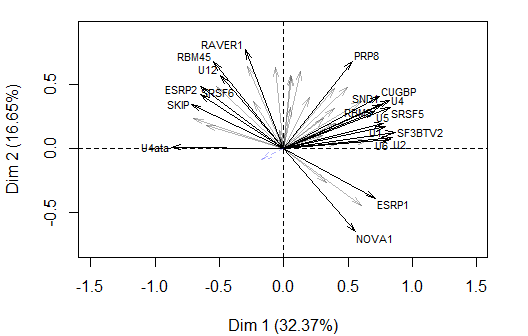
The Wilks test p-value indicates us which variable factors are the best to explain the distance between individuals.

Trombosis ACAIgM ACAIgG ComplicObstet Sexo   
 0.002197769 0.092983456 0.146204030 0.296631943 0.536583278   
 CMIT dsDNA Hta B2GPl   
 0.591791346 0.635666248 0.757410853 0.838454018

The best qualitative variable to illustrate the distance between individuals on this plane is the variable “*Trombosis”*.



**Figure 3.3 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.* *The individuals are coloured by their categories for the variable* Trombosis.



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

The **dimension 1** opposes individuals such as *25*, *6*, *5* and *24* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *32*, *34*, *33*, *27*, *31*, *35* and *29* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *25*, *6*, *5* and *24* stand (characterized by a positive coordinate on the axis) is sharing:

* high values for variables like *CUGBP*, *PTB*, *U4*, *SND1*, *SF3BTV2*, *SRSF5*, *U1*, *SRFSF9*, *RBM3* and *PRP8* (variables are sorted from the strongest).
* low values for the variables *SRSF6*, *SKIP*, *U6atac*, *PSF* and *U4ata* (variables are sorted from the weakest).

The group in which the individuals *32*, *34*, *33*, *27*, *31*, *35* and *29* stand (characterized by a negative coordinate on the axis) is sharing:

* high values for variables like *SKIP*, *SRSF6*, *U4ata*, *ESRP2*, *PSF*, *U6atac*, *CELF4*, *U12*, *U2AF1* and *RBM45* (variables are sorted from the strongest).
* low values for variables like *U5*, *PTB*, *RBM3*, *SND1*, *SRSF5*, *CUGBP*, *U4*, *SRSF10*, *TRA2A* and *RBM17* (variables are sorted from the weakest).

The **dimension 2** opposes individuals such as *25*, *6*, *5* and *24* (to the top of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *17*, *16*, *15* and *21* (to the bottom of the graph, characterized by a strongly negative coordinate on the axis).

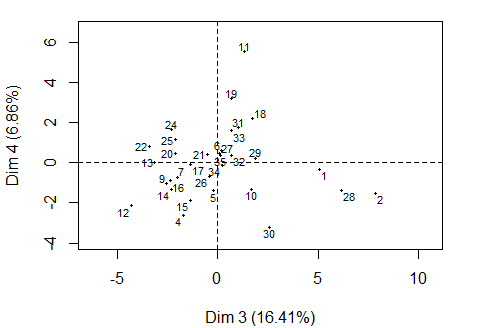
The group in which the individuals *25*, *6*, *5* and *24* stand (characterized by a positive coordinate on the axis) is sharing:

* high values for variables like *CUGBP*, *PTB*, *U4*, *SND1*, *SF3BTV2*, *SRSF5*, *U1*, *SRFSF9*, *RBM3* and *PRP8* (variables are sorted from the strongest)
* low values for the variables *SRSF6*, *SKIP*, *U6atac*, *PSF* and *U4ata* (variables are sorted from the weakest).

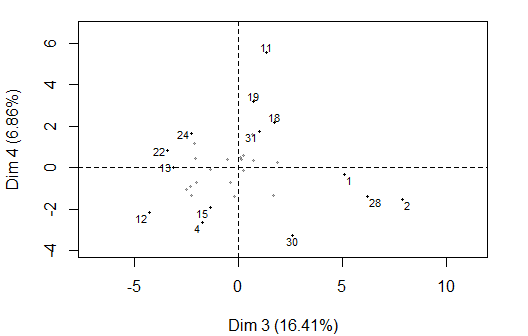
The group in which the individuals *17*, *16*, *15* and *21* stand (characterized by a negative coordinate on the axis) is sharing:

* high values for the variables *NOVA1* and *U11* (variables are sorted from the strongest).
* low values for variables like *PTB*, *SRSF4*, *ESRP2*, *PRP8*, *SRSF6*, *SKIP*, *RBM22*, *SF3BTV1*, *U2AF1* and *CELF4* (variables are sorted from the weakest).

### 4. Description of the dimension 3



**Figure 4.1 - Individuals factor map (PCA)** *All individuals are plotted.*

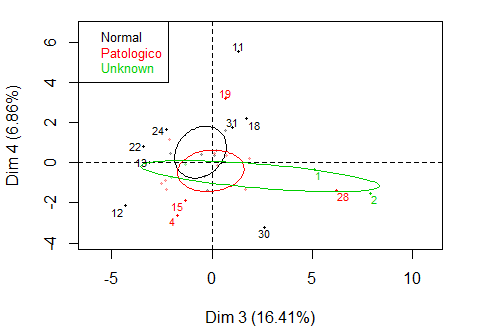


**Figure 4.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

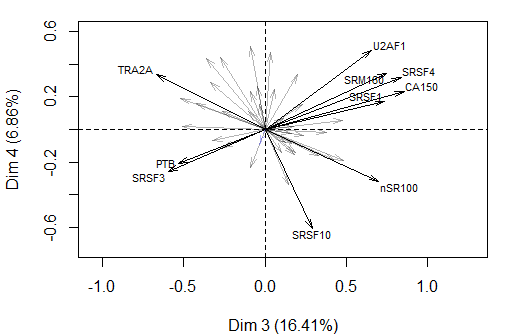
The Wilks test p-value indicates us which variable factors explain best the distance between individuals.

CMIT dsDNA ACAIgG ComplicObstet Hta   
 0.1704376 0.3118033 0.3203695 0.3321779 0.4191616   
 Sexo ACAIgM B2GPl Trombosis   
 0.5222157 0.6489044 0.7212392 0.9000545

The best qualitative variable to illustrate the distance between individuals on this plane is the variable “*CMIT”*.



**Figure 4.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.* *The individuals are coloured by their categories for the variable* “*CMIT”*.



**Figure 4.3 - Variables factor map (PCA)** *The labeled variables are those the best shown on the plane.*

The **dimension 3** opposes individuals such as *2*, *28*, *1* and *30* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *12*, *22*, *13* and *4* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

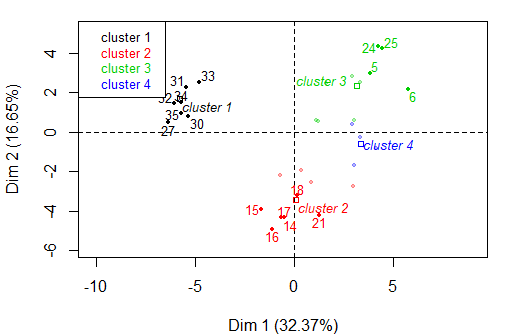
The group in which the individuals *2*, *28*, *1* and *30* stand (characterized by a positive coordinate on the axis) is sharing:

* high values for the variables *nSR100*, *U12*, *CELF4*, *SRSF4*, *SRSF10*, *CA150* and *SRSF1* (variables are sorted from the strongest).
* low values for the variables *PTB*, *SKIP*, *SC35*, *PSF*, *TRA2B*, *SRFSF9* and *TRA2A* (variables are sorted from the weakest).

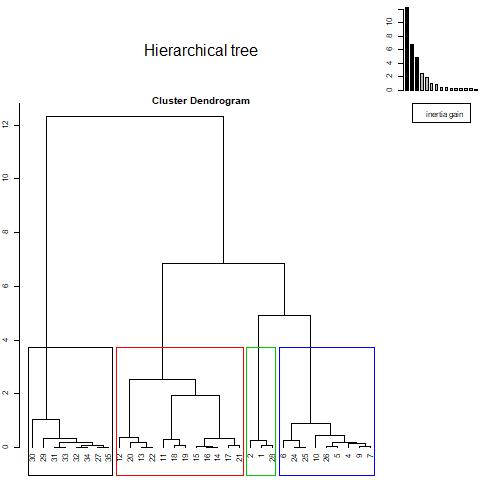
The group in which the individuals *12*, *22*, *13* and *4* stand (characterized by a negative coordinate on the axis) is sharing:

* high values for the variables *PTB*, *SC35*, *U2*, *SRSF3*, *RBM45* and *TRA2A* (variables are sorted from the strongest).
* low values for the variables *U1*, *U2AF1*, *ESRP1*, *SRM160*, *CA150*, *SRSF4* and *SRSF1* (variables are sorted from the weakest).

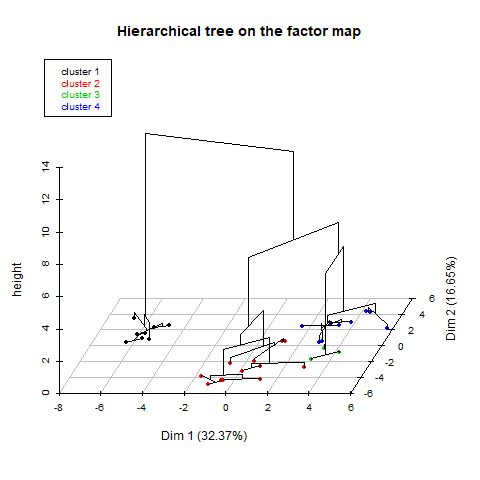
### 5. Classification



**Figure 5.1 - Hierarchical Classification of the individuals.** *The classification over individuals reveals 4 possible clusters.*

**

**Figure 5.2 – Cluster dendrogram***.*



**Figure 5.3 – Hierarchical tree on the factor map.**

The **cluster 1** is made of individuals such as *27*, *29*, *30*, *31*, *32*, *33*, *34* and *35*. This group is characterized by:

* high values for variables like *U4ata*, *ESRP2*, *SRSF6*, *SKIP*, *CELF4*, *U12*, *U6atac*, *RBM45*, *PSF* and *RAVER1* (variables are sorted from the strongest)
* low values for variables like *PTB*, *SAM68TV1*, *RBM17*, *TRA2A*, *SND1*, *RBM3*, *CUGBP*, *U5*, *U4* and *U1* (variables are sorted from the weakest).

The **cluster 2** is made of individuals such as *14*, *15*, *16*, *17*, *18* and *21*. This group is characterized by:

* high values for the variables *NOVA1* and *U11* (variables are sorted from the strongest)
* low values for variables like *U4*, *RBM22*, *SND1*, *SAM68TV1*, *SC35*, *SF3BTV1*, *ESRP2*, *SRSF10*, *CUGBP* and *PRP8* (variables are sorted from the weakest).

The **cluster 3** is made of individuals such as *5*, *6*, *24* and *25*. This group is characterized by:

* high values for variables like *PTB*, *U2*, *PRP8*, *SRSF5*, *CUGBP*, *TRA2A*, *RBM3*, *U4*, *SRSF3* and *SRFSF9* (variables are sorted from the strongest)
* low values for the variables *SRSF4*, *U2AF1* and *U4ata* (variables are sorted from the weakest).

The **cluster 4** is made of individuals sharing:

* high values for variables like *CA150*, *ESRP1*, *SRM160*, *SRSF1*, *SRSF10*, *SRSF4*, *CUGBP*, *U2AF1*, *SND1* and *nSR100* (variables are sorted from the strongest)
* low values for the variables *TRA2B*, *SC35*, *RAVER1*, *TRA2A*, *PSF*, *SRSF6* and *SKIP* (variables are sorted from the weakest).