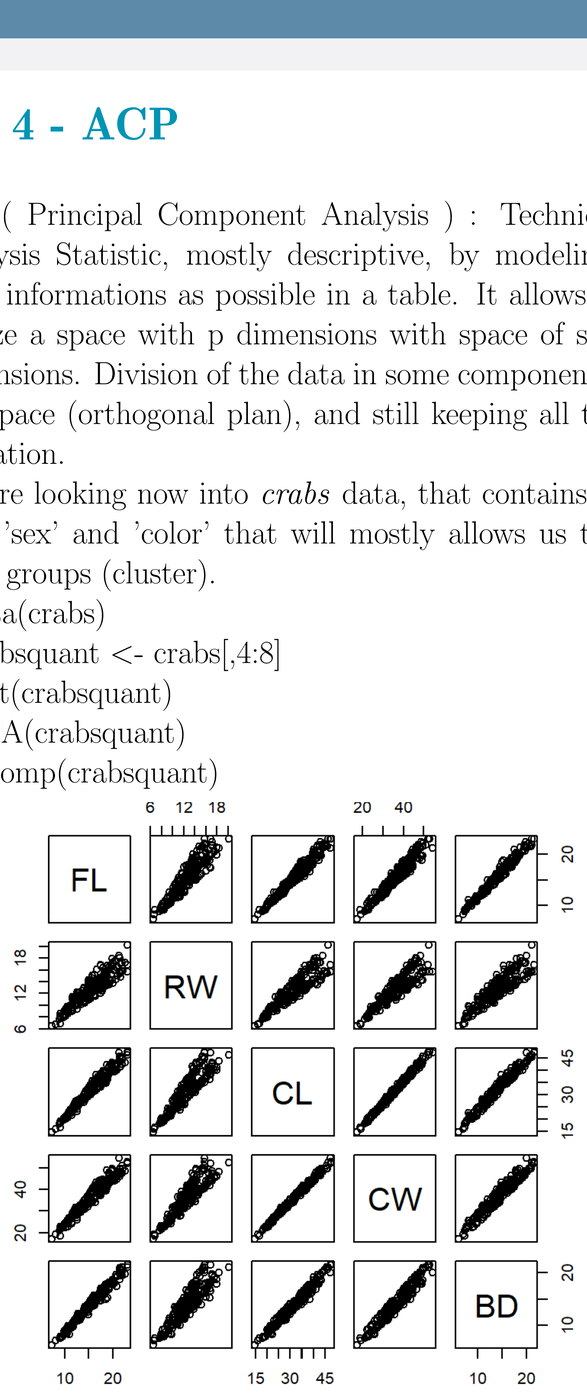
Fiche MAD 3

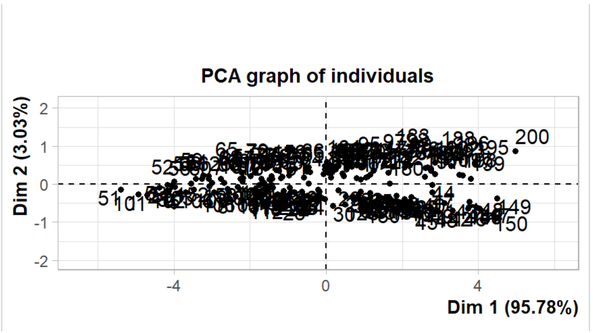
TP 4 - ACP

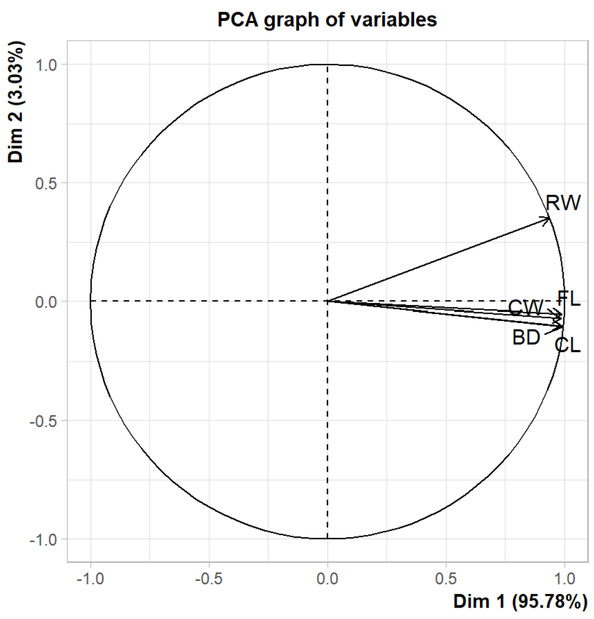
ACP(Principal Component Analysis) : Technique of Analysis Statistic, mostly descriptive, by modeling the most informations as possible in a table. It allows to visualize a space with p dimensions with space of smaller dimensions. Division of the data in some components into sub-space (orthogonal plan), and still keeping all the information.
We are looking now into *crabs* data, that contains a column 'sex' and 'color' that will mostly allows us to find some groups (cluster).

```
data(crabs)
crabsquant <- crabs[,4:8]
plot(crabsquant)
PCA(crabsquant)
prcomp(crabsquant)
```

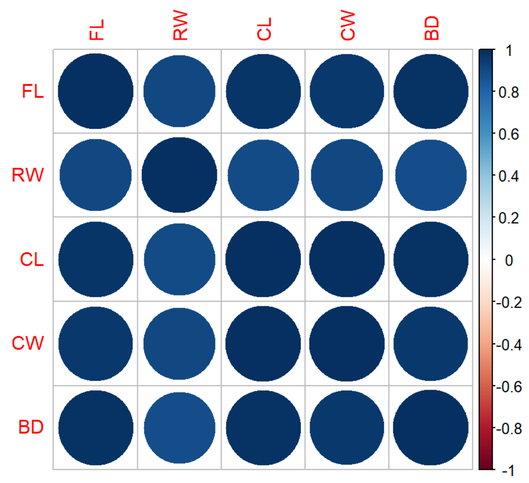


We can now plot our PCA result here :





We can see that one of our axis is explaining about 95 percent of our data. This result is obtained because those variable are correlated, to try to spread out our analysis we need to decorrelate it. So we need to find the most correlated variable to divide all of the others by this one in order to obtain a new dataset to work on, this is in sort of a normalization. In order to obtain this column u can plot the matrix of correlation :

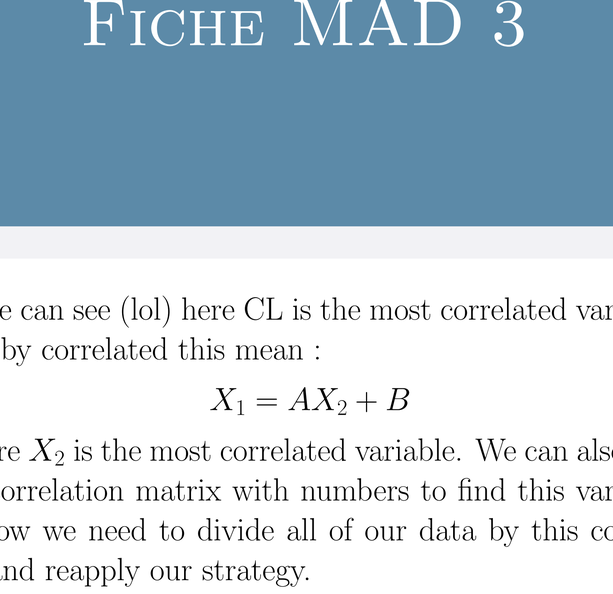


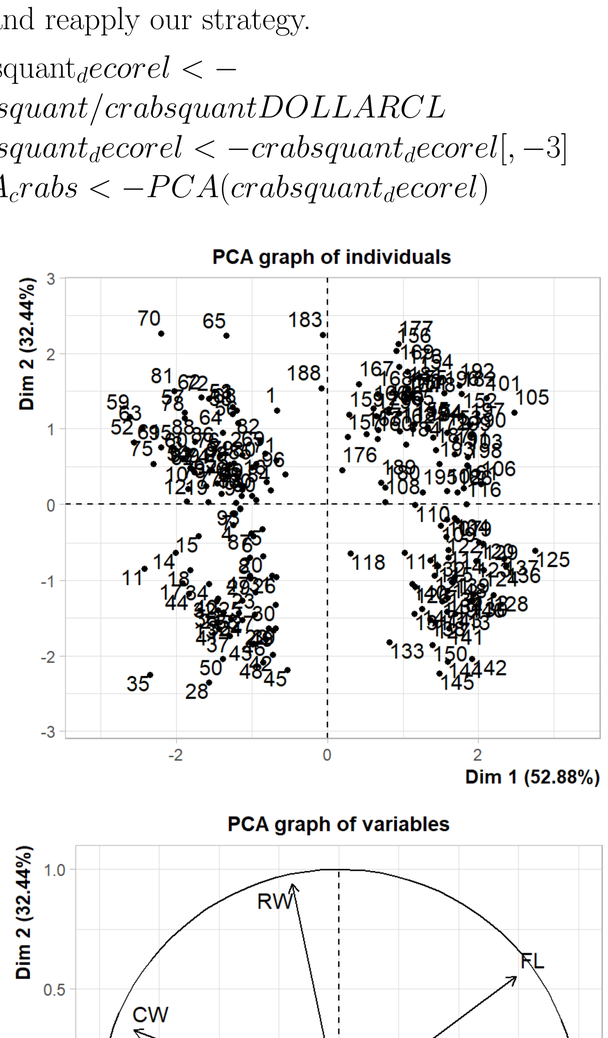
As we can see (lol) here CL is the most correlated variable. And by correlated this mean :

$$X_1 = AX_2 + B \tag{1}$$

Where X_2 is the most correlated variable. We can also plot the correlation matrix with numbers to find this variable. So now we need to divide all of our data by this column CL and reapply our strategy.

crabsquant$_d ecorel <-$
$crabsquant/crabsquant DOLLAR CL$
$crabsquant_d ecorel <- -crabsquant_d ecorel[,-3]$
$PCA_c rabs <- -PCA(crabsquant_d ecorel)$



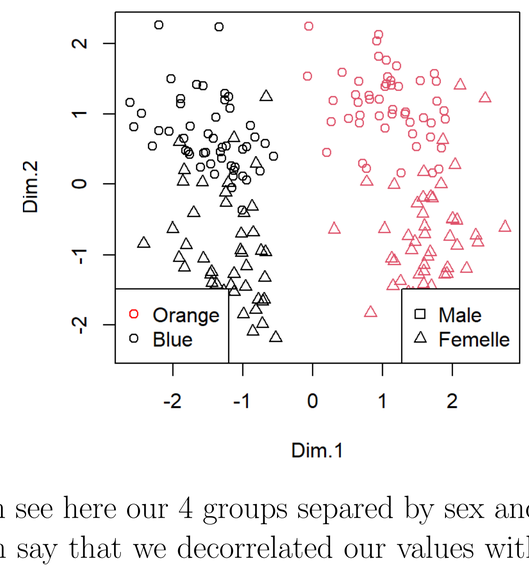


Blue/Male Orange/Male
Blue/Female Blue/Female

Correlation Circle

So here almost 85 percent of the variance is explained and now our variable are more separated into the PCA graph, correlation circle. And after applying this method we see that our plot is explaining our model by "dividing" the model data by 4 groups.

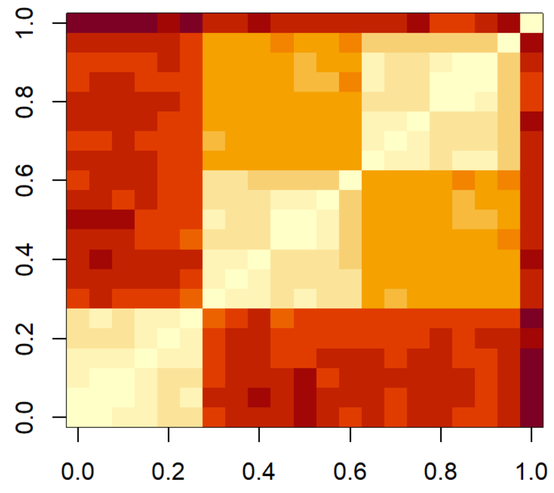
plot(PCA$_c rabsDOLLindDOLLcoord[, 1:2]$,
$col = as.numeric(crabsDOLLsp)$,
$pch = as.numeric(crabsOLLsex))$



We can see here our 4 groups separed by sex and color. We can say that we decorrelated our values with CL that because this is mostly the most important part about crabs, independantly of the sex and color. For example it can be weight or lenght that this species of crabs are almost equal between female and male.

we are working on the globine dataset. We are building D dissimilarity Matrix. This matrix contains the distance between x_j and x_i.
This matrix has a diagonal null, is defined positive and symetric.

d <- read.table("neighbor$_g lobin.dat$")
$d[d < 0] dissimilarittjrpositive$
$sum(as.matrix(d[,-1]) - (t(as.matrix(d[,-1]))))$
$diag(as.matrix(d[,-1]))image(as.matrix(d[,-1]))$



We can see here that we will mostly obtain 3 groups. We now search to obtain a Matrix NxN

$$A_{ij} = \delta(X_i, X_j)$$

, so we search the matrix of square, so we need to square every coefficient of the matrix of dissimilarity Δ.

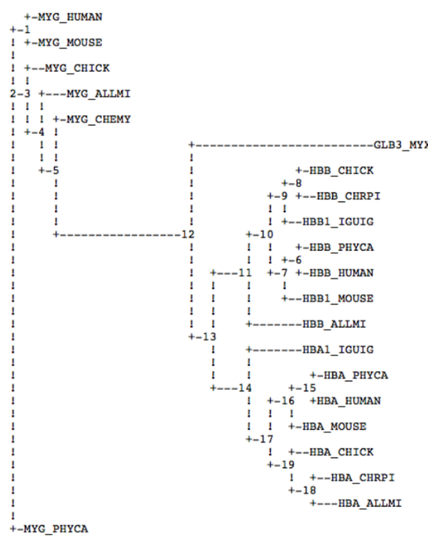


FIG. 2 – *Arbre phylogénétique des globines*

We can find our 3 groups here if we look closely. Matrix of centrage J :

$$J = I - \frac{1}{n} 1_{(n,n)} \tag{2}$$

and then B, the spectral decomposition :

$$B = -\frac{1}{2} J \Delta J \tag{3}$$

```
tmp <- eigen(B)
J <- tmpDOLLvectors
A <- diag(tmpDOLLvalues)
```


In Rstudio, eigen value are ranged by decreasing values and we choose to keep 3 of them because they will mostly explained about 85 percent of the dataset. In order to reduce the dimension and to explain more of the dataset. The eigen vectors with the highest eigen value is the one that does explain the most about the model. Each vector will be orthogonal to the previous in order to create a space with m dimensions.