Devoir n°1

Anne Badel

2019-03-19

Contents

1 Quelques paramètres

```
alpha <- 0.05
```

```
workdir <- getwd()
message("Working directory\t", workdir)
```

```
data.folder <- "/Volumes/Data/badel/Documents/Git/DUBII/module-3-Stat-R/seance_2/data"

meta.file <- "metadata.RDS" #
abond.file <- "microbiota.abundance.log.RDS"
meta.path <- file.path(data.folder, meta.file)
abond.path <- file.path(data.folder, abond.file)

## Load expression table
meta.table <- readRDS(meta.path)
abond.table <- readRDS(abond.path)
```

Le `data.frame` `meta.table` contient les données physiologiques des 237 patients, le `data.frame` `abond.table` contient les données de métagénomiques de ces mêmes patients.

2 Exercice 1 : comparaison de l'âge moyen des patients malades et sains

2.1 identification de la variable aléatoire étudiée

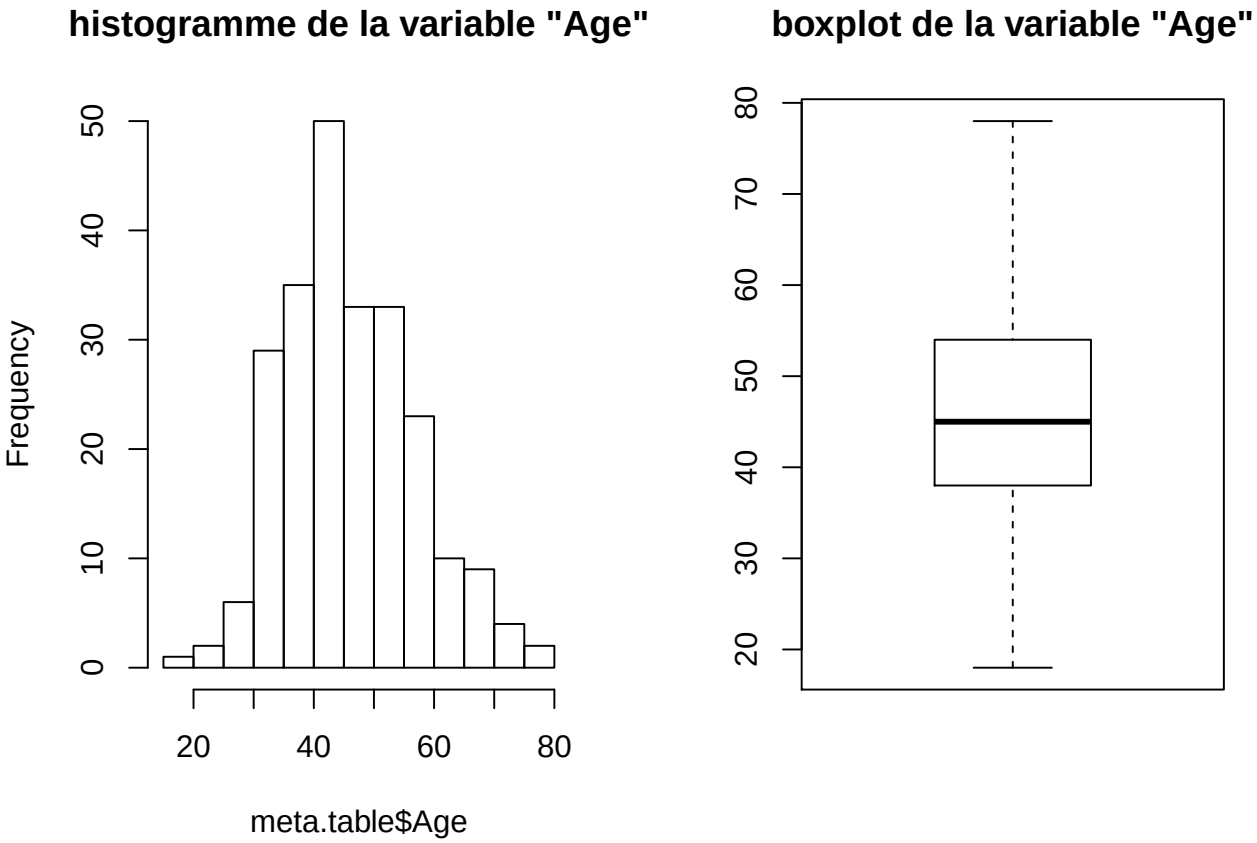
La variable étudiée est l'âge des patients `Age` en fonction du facteur `être ou non malade`

2.1.1 La variable Age

```
summary(meta.table$Age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   38.00   45.00   46.45   54.00   78.00
```

```
par(mfrow=c(1,2))
hist(meta.table$Age, main = "histogramme de la variable \"Age\"")
boxplot(meta.table$Age, main = "boxplot de la variable \"Age\"")
```




```
par(mfrow=c(1,1))
```

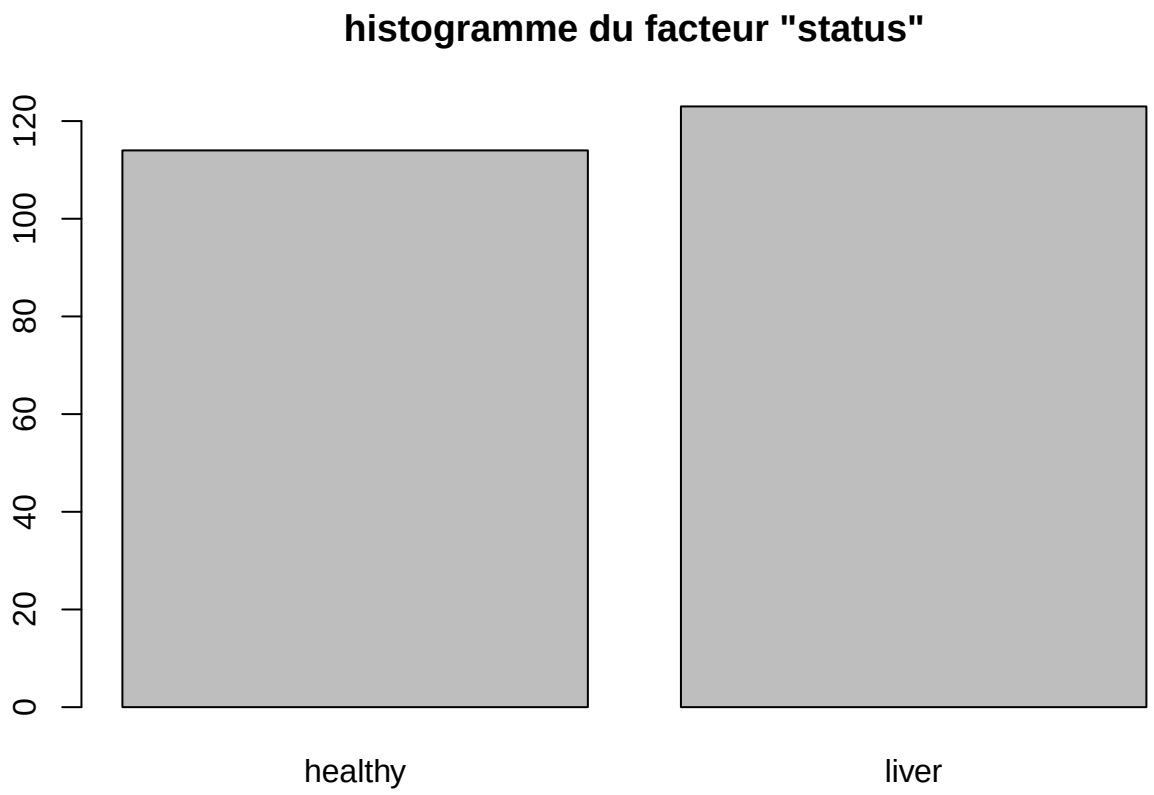
Les individus étudiés ont entre 18 et 78 ans, avec une moyenne de 46,5 ans et une médiane très proche. Il n'y a pas de données manquantes

2.1.2 Le facteur status

```
summary(meta.table$status)
```

```
healthy   liver
    114     123
```

```
plot(meta.table$status, main = "histogramme du facteur \"status\"")
```



2.2 identification du test à utiliser

La variable étudiée, l'`Age` est une variable quantitative de paramètre μ et σ^2 inconnues. Le facteur `status`est un qualitatif à deux niveaux. Nous allons donc faire un test de comparaison de deux moyennes, dont les hypothèses sont les suivantes :

- $H_0 : \mu_{healthy} = \mu_{liver}$
- $H_1 : \mu_{healthy} \neq \mu_{liver}$

Etant donné que nous sommes dans le cas de comparaison de moyennes dans le cas de grands échantillons (n $>$ 30), nous pouvons faire le test directement, sans nous inquiéter des conditions d'application de ce test.

2.3 `Age` versus `status`

```
boxplot(Age ~ status, data = meta.table, main = "\"Age\" versus \"status\"")
```

2.4 calcul de la moyenne et de l'écart-type

Pour calculer moyennes et écart-type dans chacun des groupes, on peut utiliser la fonction `by()` qui renvoit un objet de classe `by`, donc une liste.

```
mean.by <- by(meta.table$Age, meta.table$status, mean)
sd.by <- by(meta.table$Age, meta.table$status, sd)
```

ou la fonction `aggregate()` qui renvoit un `dataa.frame`

```
mean.agg <- aggregate(Age ~ status, data = meta.table, FUN = mean)
sd.agg <- aggregate(Age ~ status, data = meta.table, FUN = sd)
```

Pour afficher les résultats, on peut utiliser la fonction `kable()`

```
res <- data.frame(mean.agg[,2], sd.agg[,2])
colnames(res) <- c("moyenne", "écart-type")
rownames(res) <- c("healthy", "liver")
kable(res)
```

	moyenne	écart-type
healthy	42.59649	9.325689
liver	50.02439	11.157195

2.5 intervalles de confiance

On sait calculer l'intervalle de confiance d'une moyenne grâce à la formule suivante :

```
IC = matrix(nrow = 2, ncol = 2)
rownames(IC) <- c("healthy", "liver")
colnames(IC) <- c("m.inf", "m.sup")
```

$$IC_{1-\alpha}(\mu) = \left[m - z_{1-\frac{\alpha}{2}} \sqrt{\frac{s^2}{n}} \right]$$

soit, pour les individus sains

```
nb.ind <- table(meta.table$status)
m.inf <- res[1,1] - qnorm((1-alpha/2) * sqrt(res[1,2]^2/nb.ind[1]))
m.sup <- res[1,1] + qnorm((1-alpha/2) * sqrt(res[1,2]^2/nb.ind[1]))
IC[1,] <- c(m.inf, m.sup)
```

$\text{IC}(\mu) = [41.5531923, 43.6397902]$

et pour les individus malades

```
nb.ind <- table(meta.table$status)
m.inf <- res[2,1] - qnorm((1-alpha/2) * sqrt(res[2,2]^2/nb.ind[2]))
m.sup <- res[2,1] + qnorm((1-alpha/2) * sqrt(res[2,2]^2/nb.ind[2]))
IC[2,] <- c(m.inf, m.sup)
```

$\text{IC}(\mu) = [47.9525429, 52.0962376]$

Les deux intervalles de confiance sont dans le tableau suivant :

	m.inf	m.sup
healthy	41.55319	43.63979
liver	47.95254	52.09624

2.6 le test statistique

```
(mon.test <- t.test(Age ~ status, data = meta.table, var.equal =T))
```

```
    Two Sample t-test

data:  Age by status
t = -5.5378, df = 235, p-value = 8.175e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.070432  -4.785366
sample estimates:
mean in group healthy   mean in group liver
             42.59649              50.02439
```

```
ma.pvalue <- mon.test$p.value
```

Au risque $\alpha = 0.05$, le test est significatif en effet la p.value,0.00000008, est inférieure à 0.05. On peut donc considérer que l'âge des individus est différent de celui des malades. En moyenne, les individus malades sont plus agés que les individus sains.

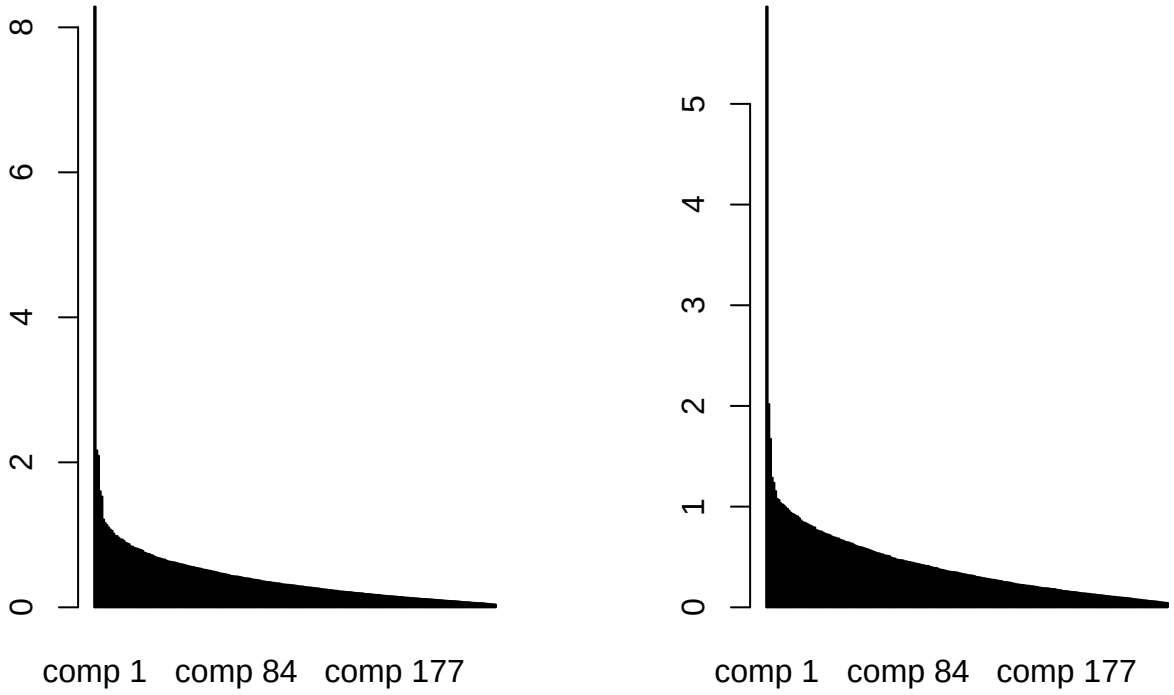
3 Exercice 2 : ACP sur le jeu de données `microbiota.abudance`

```
ma.pca <- PCA(abond.table, graph = F)
ma.pca.nonscale <- PCA(abond.table, graph = F, scale.unit = FALSE)
```

3.1 part de variabilité

Le premier plan factoriel, les deux premières composantes principales, représente 7.984228 % de la variabilité de nos données. Ce n'est pas énorme, et lorsqu'on fait le plot des valeurs propres,

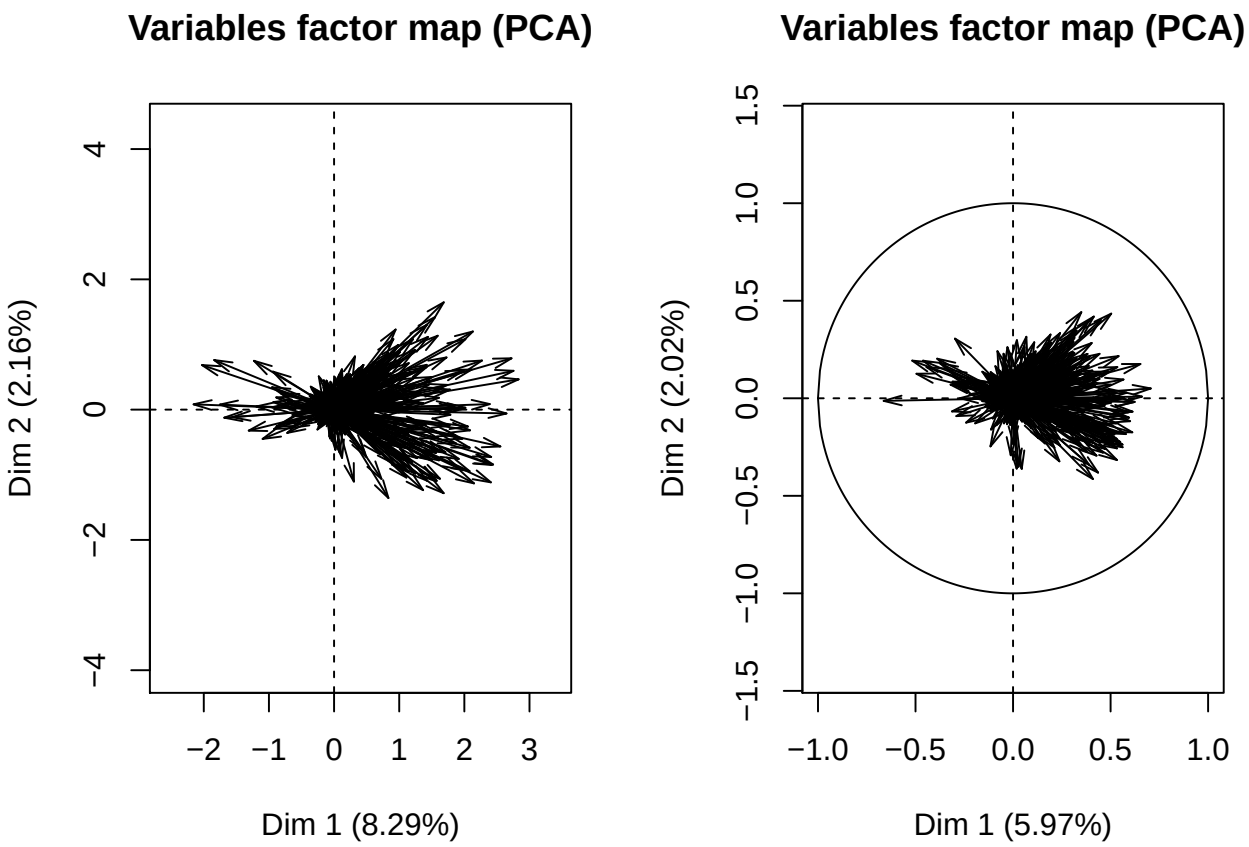
```
par(mfrow=c(1,2))
barplot(ma.pca.nonscale$eig[,2])
barplot(ma.pca$eig[,2])
```




```
par(mfrow=c(1,1))
```

on voit qu'on pourrait aller jusqu'à prendre au moins 163 composantes (valeurs propres > 1).

```
par(mfrow=c(1,2))
plot(ma.pca.nonscale, choix = "var", label = "none")
plot(ma.pca, choix = "var", label = "none")
```




```
par(mfrow=c(1,1))
```

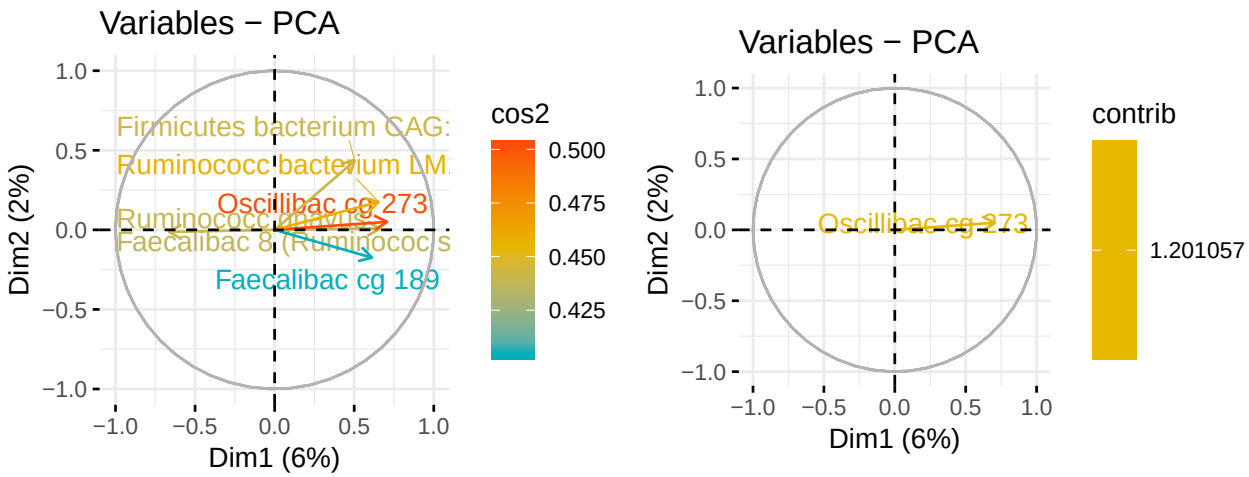
Au vu du nombre important de descripteurs (523), il est normal qu'on ne distingue pas grand chose. Si ce n'est qu'aucune des variables n'est bien représentée, ni n'a une contribution importante.

On peut essayer de représenter seulement les variables les mieux représentées (cos2 élevés).

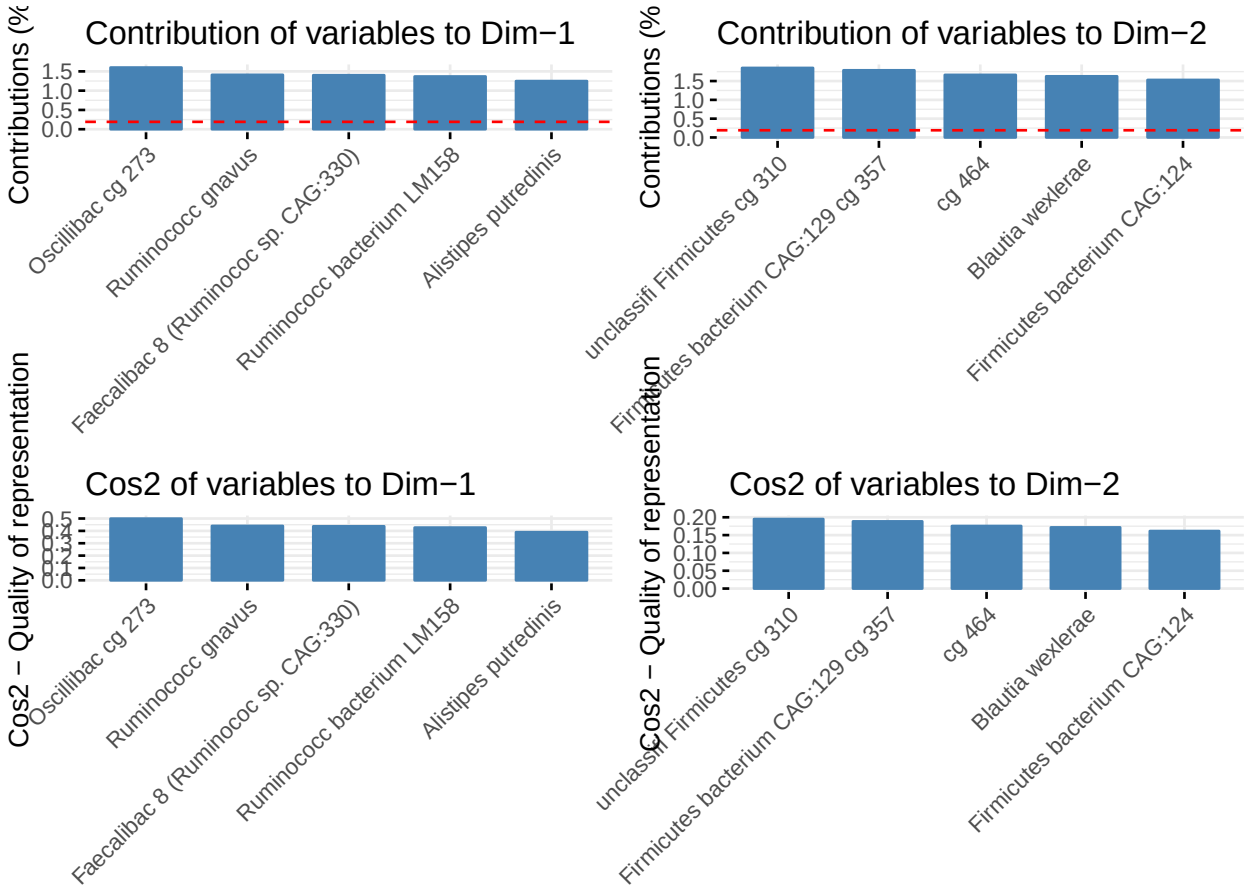
```
cos2.p1 <- fviz_pca_var(ma.pca, col.var="cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE, # Avoid text overlapping
             select.var = list(cos2 = 0.4)
             )
cos2.cp1 <- fviz_cos2(ma.pca, choice = "var", axes = 1, top = 5)
cos2.cp2 <- fviz_cos2(ma.pca, choice = "var", axes = 2, top = 5)
```

Et les descripteurs ayant la meilleure contribution

```
contrib.p1 <- fviz_pca_var(ma.pca, col.var="contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE, # Avoid text overlapping
             select.var = list(contrib = 1)
             )
plot_grid(cos2.p1, contrib.p1, cos2.cp1, cos2.cp2, ncol = 2, nrow = 1)
```




```
contrib.cp1 <- fviz_contrib(ma.pca, choice = "var", axes = 1, top = 5)
c.cp1 <- contrib.cp1$data
c.cp1 <- c.cp1$contrib
names(c.cp1) <- contrib.cp1$data$name
c.cp1 <- sort(c.cp1, decreasing = TRUE)
contrib.cp2 <- fviz_contrib(ma.pca, choice = "var", axes = 2, top = 5)
c.cp2 <- contrib.cp2$data
c.cp2 <- c.cp2$contrib
names(c.cp2) <- contrib.cp1$data$name
c.cp2 <- sort(c.cp2, decreasing = TRUE)
plot_grid(contrib.cp1, contrib.cp2, cos2.cp1, cos2.cp2, ncol = 2, nrow = 2)
```

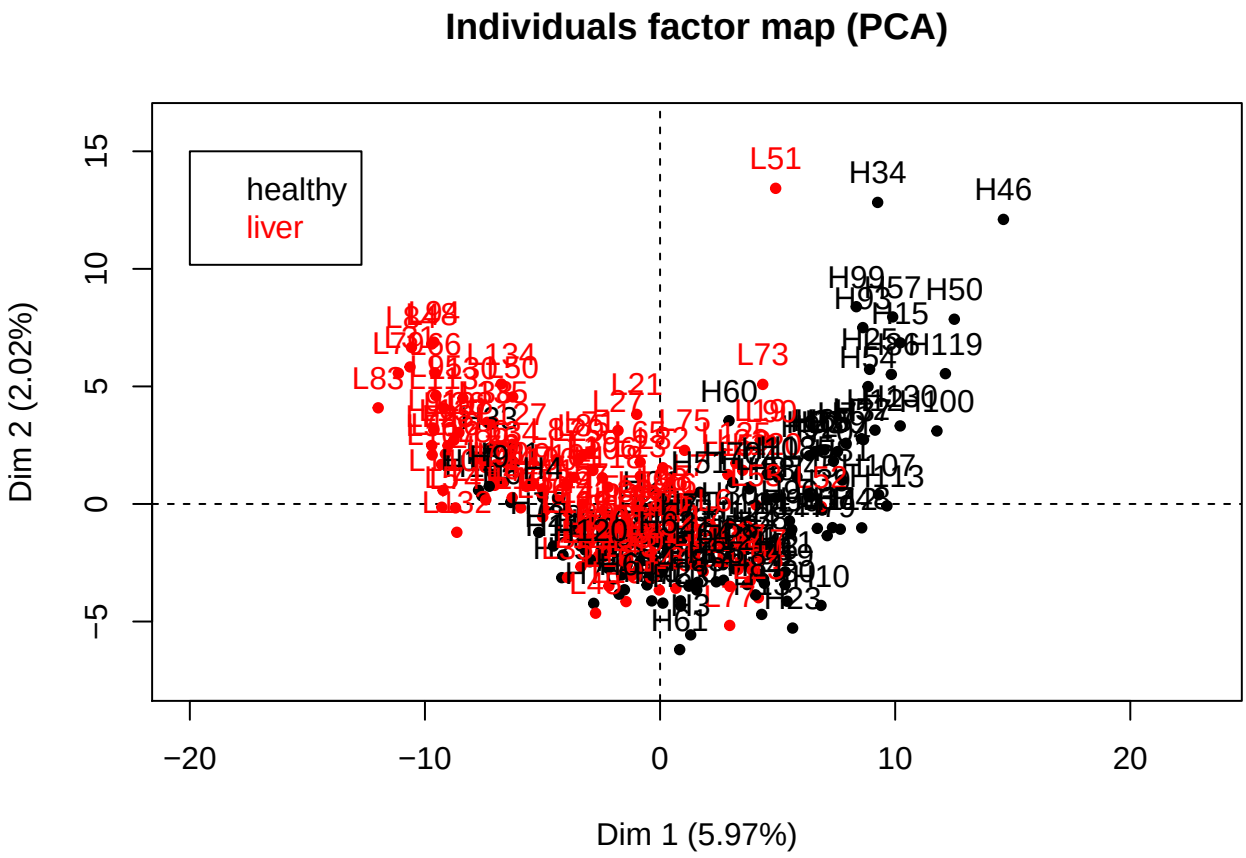


Les descripteurs contribuant le plus à la première composante (CP1) sont : Oscillibac cg 273, Ruminococc gnavus, Faecalibac 8 (Ruminococ sp. CAG:330), Ruminococc bacterium LM158, Alistipes putredinis. Les individus à droite de l'ACP auront des valeurs plus élevées que la moyenne de Oscillibac cg 273, les individus à gauche de l'ACP auront des plus élevées que la moyenne de Ruminococc gnavus. Les descripteurs contribuant le plus à la deuxième composante (CP2) sont : unclassifi Firmicutes cg 310. Et c'est pas grand chose.

3.2 les individus en fonction du statut

D'après ce que nous avons dit précédemment, les individus sains ont en moyenne des valeurs de Oscillibac cg 273, Ruminococc gnavus, Faecalibac 8 (Ruminococ sp. CAG:330), Ruminococc bacterium LM158, Alistipes putredinis plus élevées, alors que les individus malades ont en moyenne des valeurs de Ruminococc gnavus plus élevées.

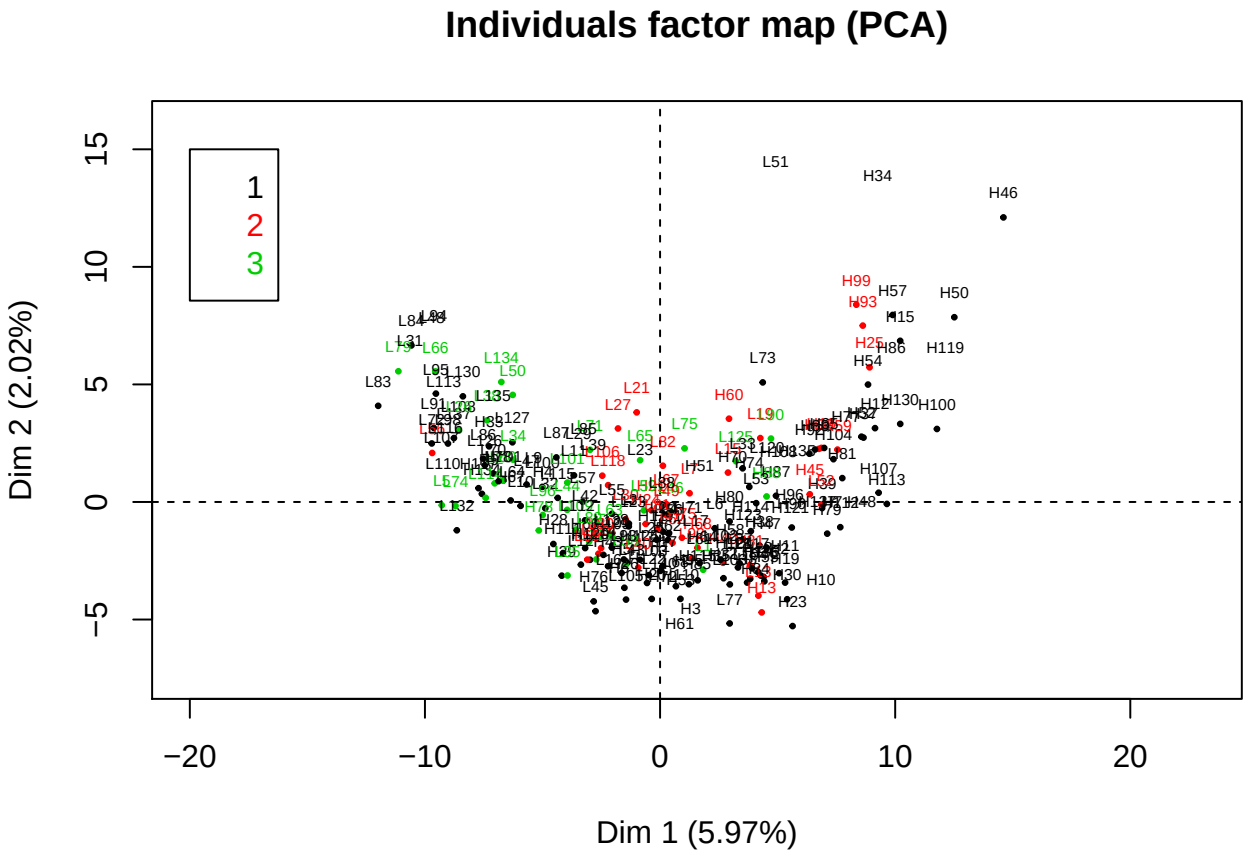
```
plot(ma.pca, choix = "ind", col.ind = meta.table$status)
legend(-20, 15, legend = c("healthy", "liver"), text.col= 1:2)
```



3.3 les individus en fonction de l'enterotype

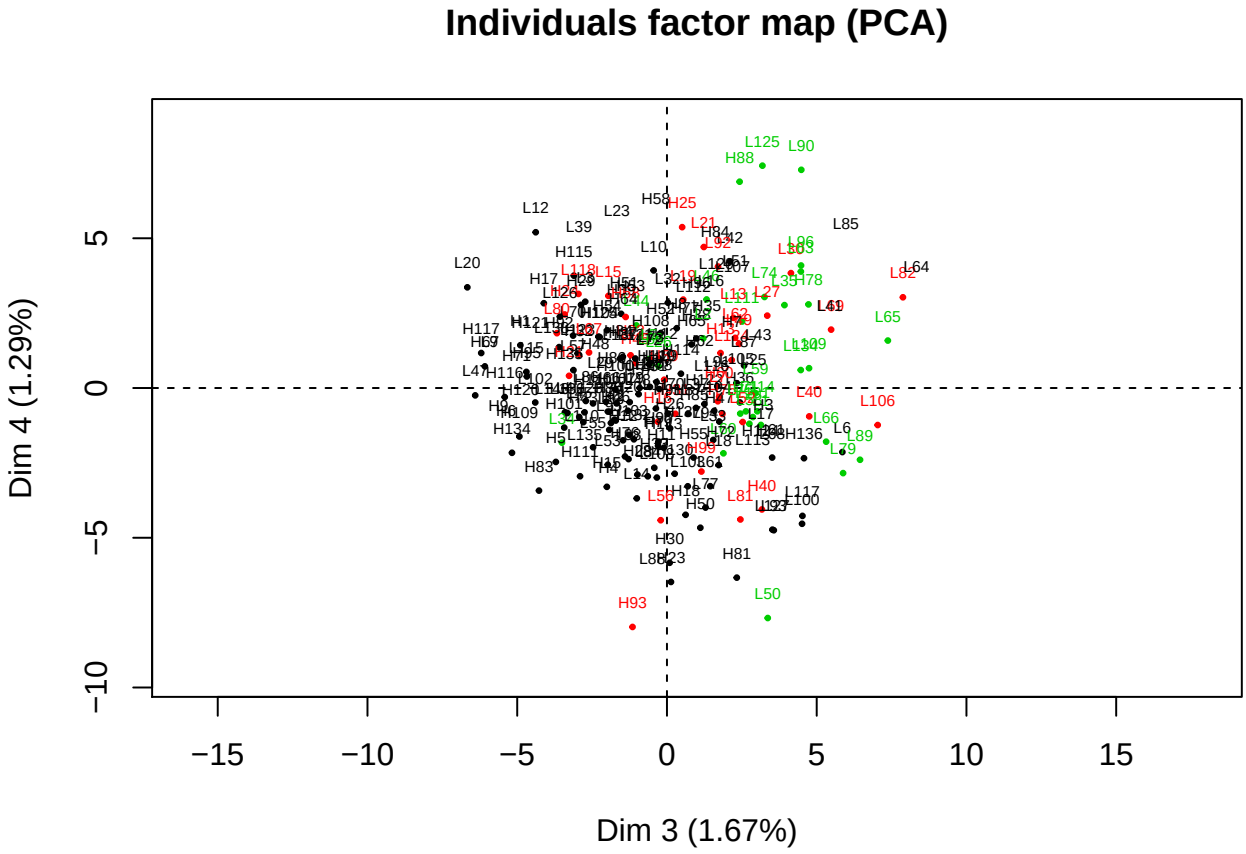
Attention : il y a des données manquantes

```
plot(ma.pca, choix = "ind", col.ind = meta.table$Enterotype, cex = 0.5)
legend(-20, 15, legend = 1:3, text.col= 1:3)
```



Il semble que les variables indentifiées précédemment Oscillibac cg 273, Ruminococc gnavus permettent de différencier les individus de type 2 et 3. Les individus de type 1 sont répartis sur l'ensemble du premier plan factoriel.

```
plot(ma.pca, choix = "ind", col.ind = meta.table$Enterotype, cex = 0.5, axes = c(3,4))
legend(-20, 15, legend = 1:3, text.col= 1:3)
```



Le plan factoriel (3,4) pour seulement 3% de la variabilité semble distingué les types “1”, à gauche des deux autres types.