

Examen de biostatistiques

Barras Clément & Delebarre Bertille

April 1, 2019

Chargement du jeu de données prostate.txt

```
library(factoextra)
```

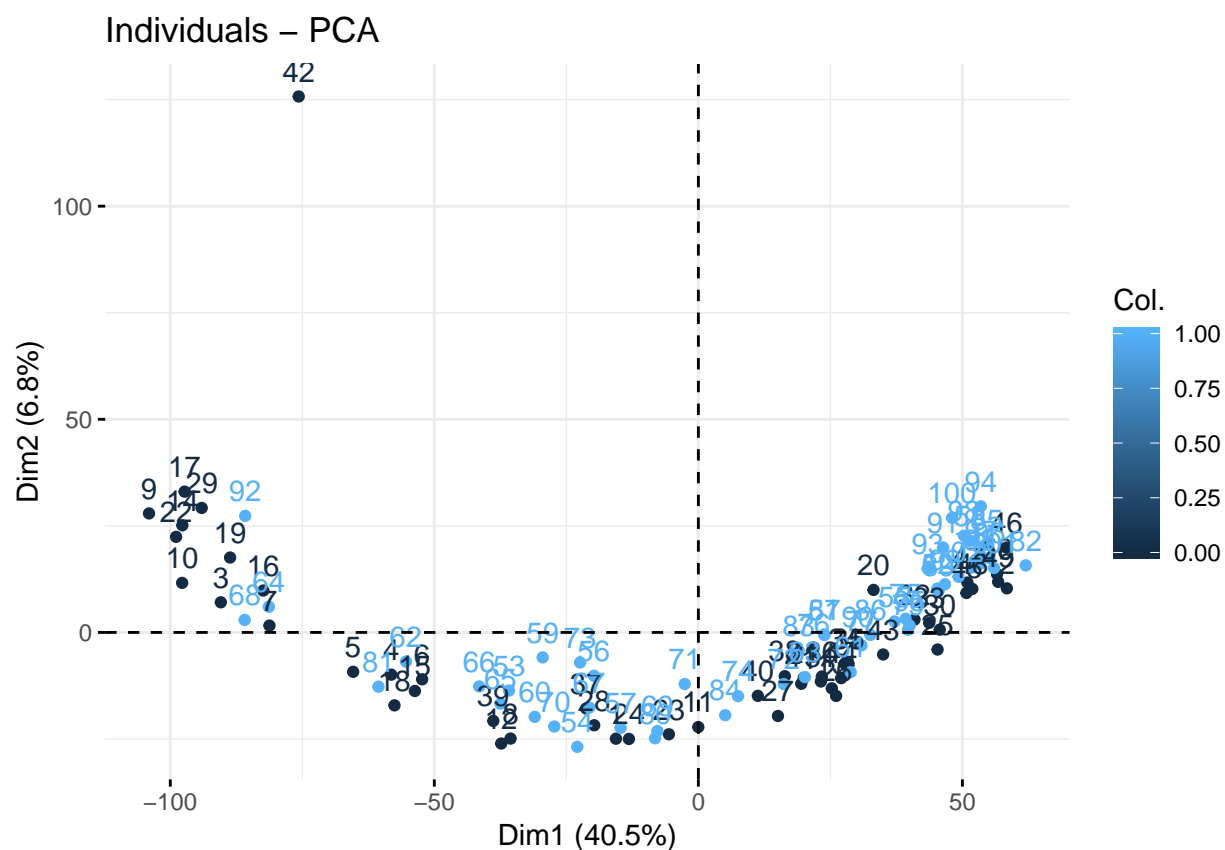
```
## Warning: package 'factoextra' was built under R version 3.5.3
```

```
prostate <- read.delim("../data/prostate.txt", sep=" ")  
n <- nrow(prostate)  
p <- ncol(prostate)  
X = as.matrix(prostate[, 2:p])  
y = as.matrix(prostate$y)
```

Question 1.

On effectue une ACP sur les données, et on les représente sur le premier plan principal :

```
fit.pca = prcomp(X, scale = TRUE, center = TRUE)  
fviz_pca_ind(fit.pca, col.ind = y)
```



Le premier axe principal explique plus de 40% de la variance, ce qui est remarquable au vu du nombre de variables. Cependant, les deux catégories d'individus ne sont pas séparable avec cette projection.

La variable `fit.pca$rotation` contient les vecteurs propres de $\bar{X}^T \bar{X}$ où \bar{X} désigne la matrice des échantillons centrée et réduite.

Afin de représenter quels gènes contribuent le plus aux première et deuxième composantes, on trie les coordonnées des colonnes correspondantes par ordre décroissant en valeur absolue :

```
sort(abs(fit.pca$rotation[,1]), decreasing = TRUE)[1:10]
```

```
##      X5435      X1686      X4784      X5365      X1783      X1321
## 0.01959220 0.01958715 0.01951524 0.01942688 0.01942369 0.01939736
##      X4705      X4813      X2248      X1202
## 0.01938836 0.01938645 0.01938349 0.01937350
```

```
sort(abs(fit.pca$rotation[,2]), decreasing = TRUE)[1:10]
```

```
##      X500      X580      X3712      X881      X5659      X1416
## 0.03921846 0.03803741 0.03797690 0.03782114 0.03767390 0.03754320
##      X5143      X786      X1154      X5616
## 0.03749617 0.03736217 0.03729699 0.03716413
```

Question 2.

Question 3.

Question 4