Examen de biostatistiques

Barras Clément & Delabarre Bertille

April 1, 2019

Chargement du jeu de données `prostate.txt`

```
library(factoextra)
```

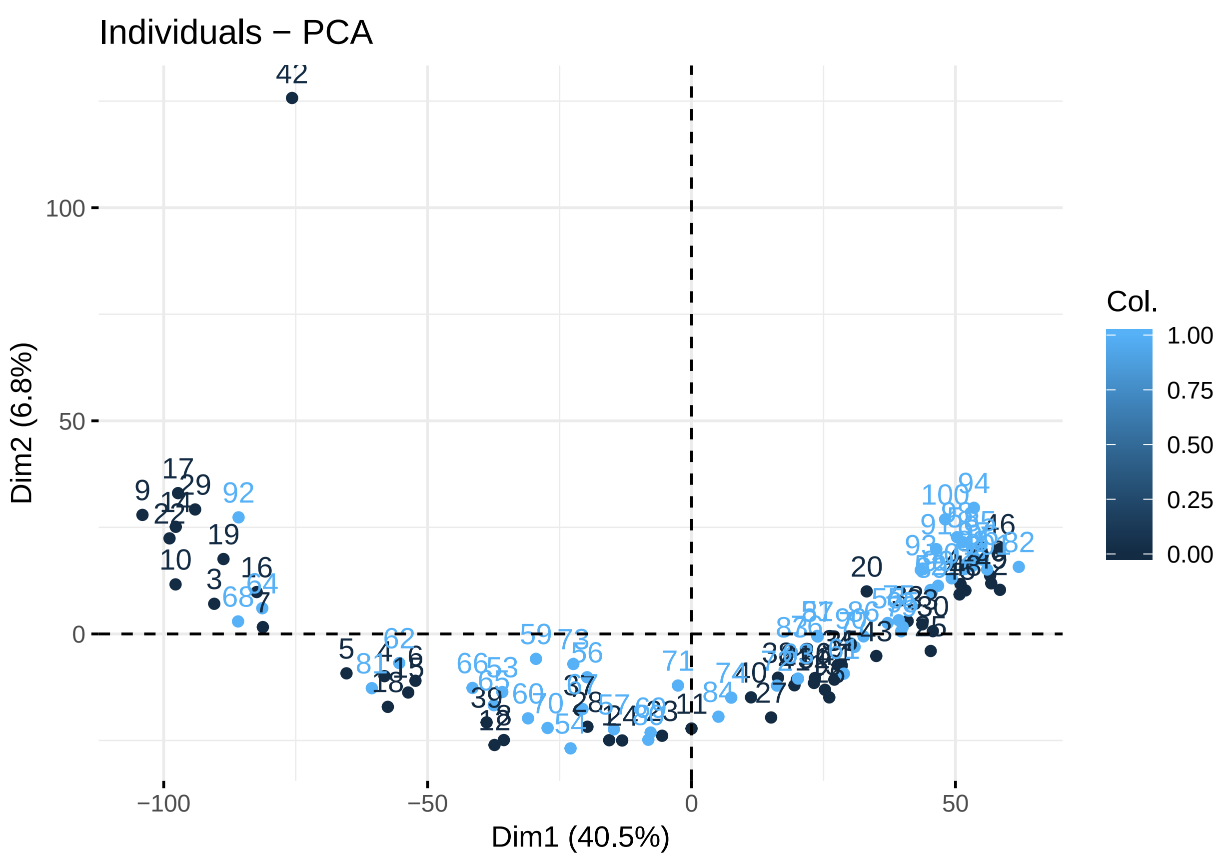
```
## Warning: package 'factoextra' was built under R version 3.5.3
```

```
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 3.

$cov(\mathbf{Xa}, \mathbf{y}) = \mathbf{X}cov(\mathbf{a}, \mathbf{y}) = \mathbf{X}\mathbb{E}[\mathbf{ay}^T]$ car y est centrée. $cov(\mathbf{Xa}, \mathbf{y}) = \frac{1}{n}\langle \mathbf{Xa} \mid \mathbf{y}\rangle = \frac{1}{n}\langle \mathbf{X}^T\mathbf{y} \mid \mathbf{a}\rangle$ Et on sait que $\nabla_{\mathbf{a}} cov(\mathbf{Xa}, \mathbf{y}) = \frac{1}{n}\mathbf{X}^T\mathbf{y}$.

Ce problème est équivalent à la minimisation de $-cov(\mathbf{Xa}, \mathbf{y})$ selon $\mathbf{a}$, avec $\mathbf{a}$ dans un compact (intersection de fermés bornés en dimension finie). Or toute fonction continue sur un compact atteint ses bornes, donc le problème admet une solution $\mathbf{a}^*$ (on ne sait pas si elle est unique). En notant $\mathbf{b} = \frac{1}{n}\mathbf{X}^T\mathbf{y}$, le Lagrangien du problème s'écrit : $\mathcal{L}(\mathbf{a}, \lambda, \mu) = -\mathbf{b}^T\mathbf{a} + \lambda(\sum_{j=1}^{p} |\mathbf{a}_j| - s) + \frac{\mu}{2}(\langle \mathbf{a} \mid \mathbf{a}\rangle - 1)$, avec $\lambda > 0$. La fonction valeur absolue n'est pas différentiable en 0, mais on peut considérer son sous-gradient, qui est $\{1\}$ pour $x > 0$, $\{-1\}$ pour $x < 0$ et $[-1, 1]$ pour $x = 0$. Soit s_i le sous gradient de la fonction valeur absolue évaluée en a_i.

Une condition nécessaire que doit respecter la solution est :

$\forall i \in [|1, p|], 0 \in \partial_{a_i}\mathcal{L}(\mathbf{a}, \lambda, \mu) = -b_i + \lambda s_i + \mu a_i$ Supposons $\mu \neq 0$.

Cas 1: $\mid \frac{b_i}{\mu} \mid > \frac{\lambda}{\mu}$

LA