Analyse en composantes principales

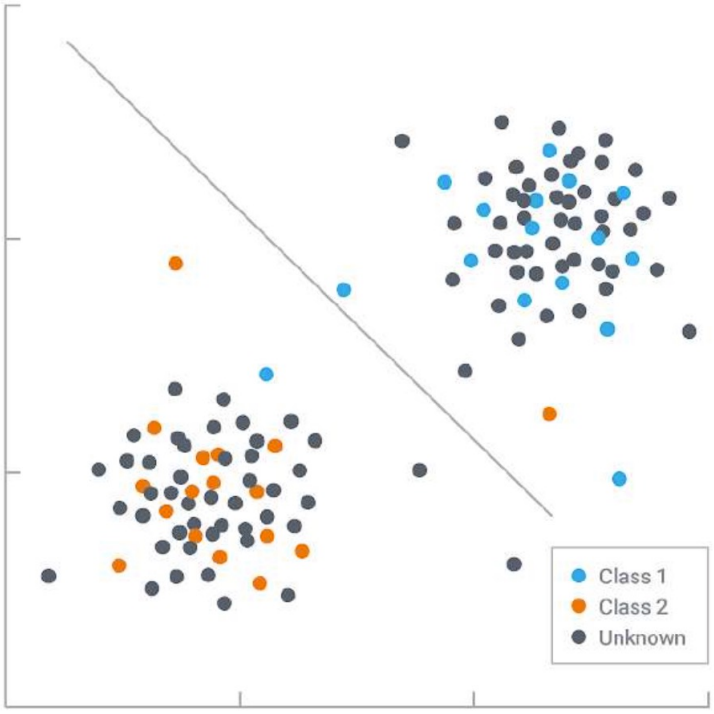
Comment faire de bonnes frites et comment être parcimonieu·x·se

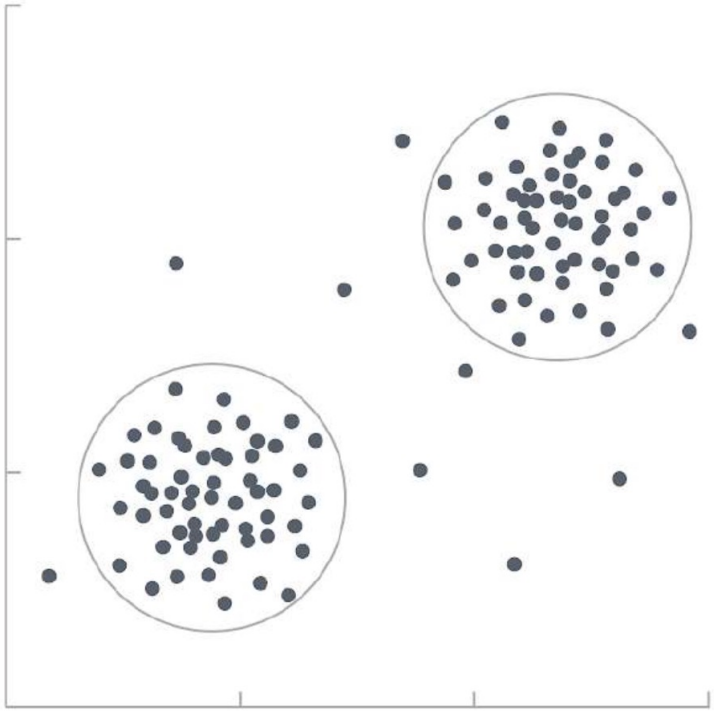

Quiz

Analyse non supervisée/exploratoire vs analyse supervisée

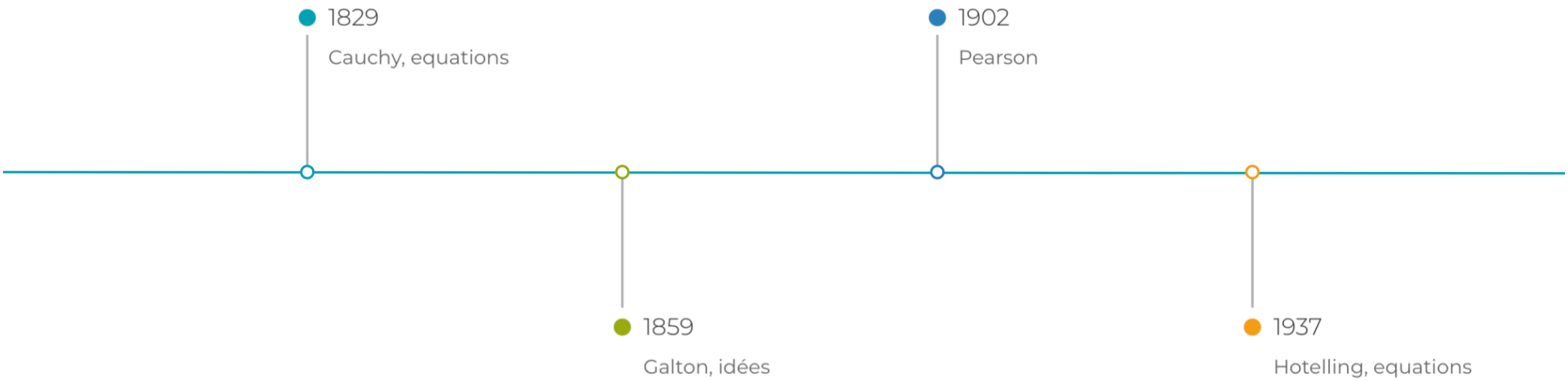
- Méthodes supervisées : une observation = caractéristiques / variables (p.ex : l'expression de gène) + une variable de réponse (p.ex. : survie).
- L'objectif est alors de prédire la réponse à l'aide des variables (par exemple, quels gènes prédisent le mieux ou sont associés à la survie des patients).
- Pour le moment, on explore (avec l'ACP et la CCA)

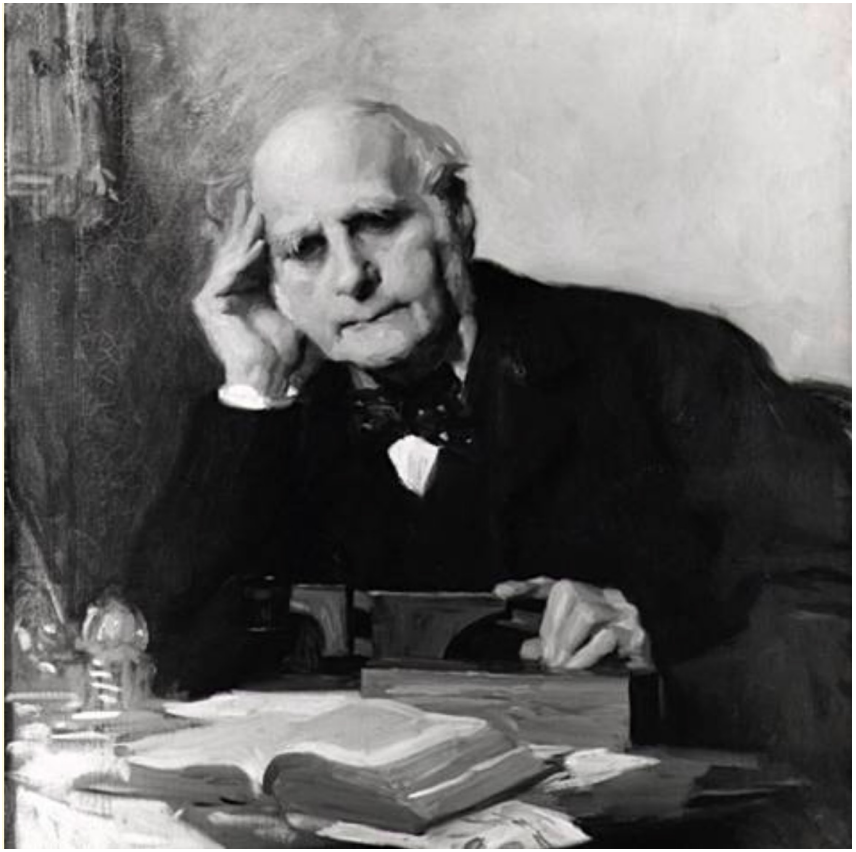
En image (merci Laura)



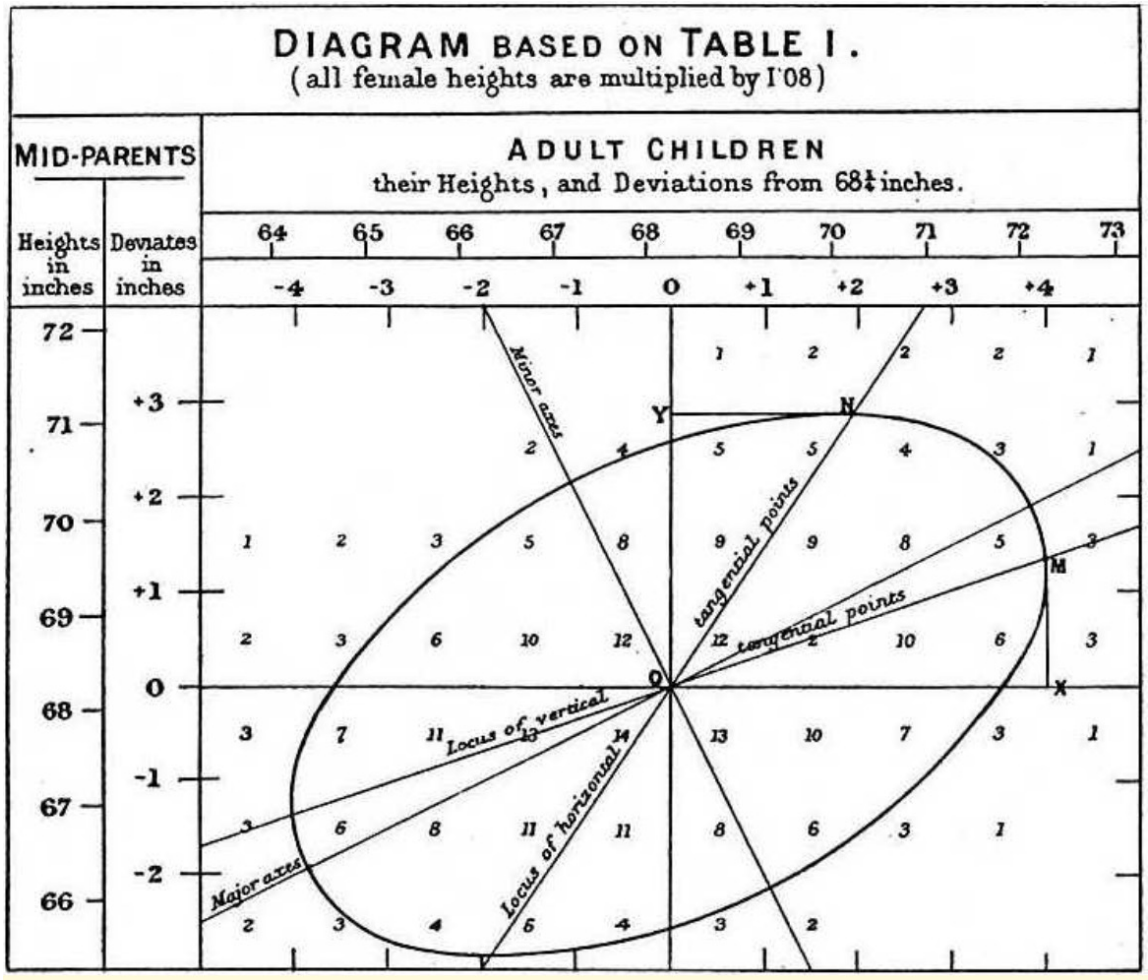


Un peu d'histoire

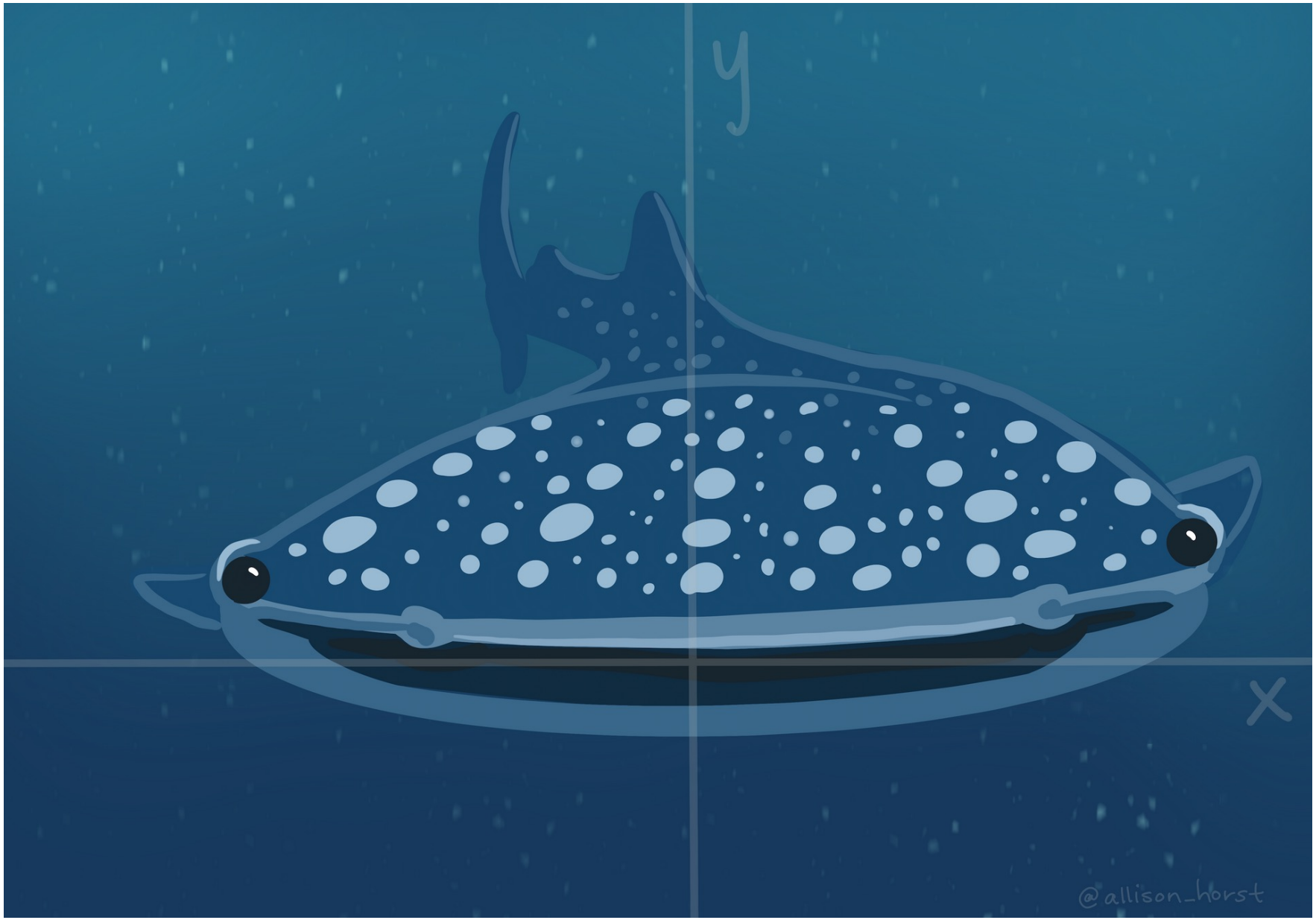




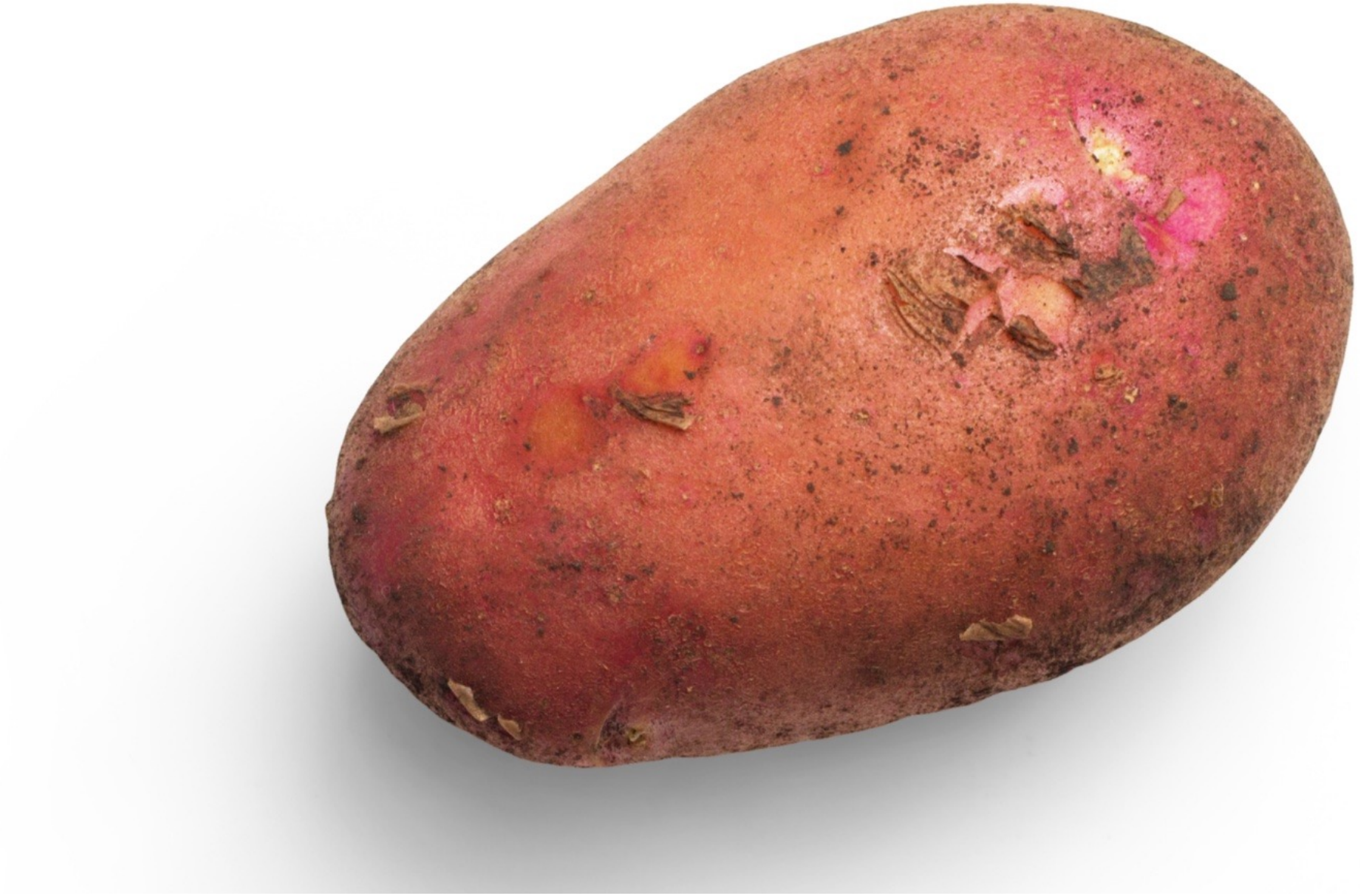
Sir Francis Galton



from Galton 1859

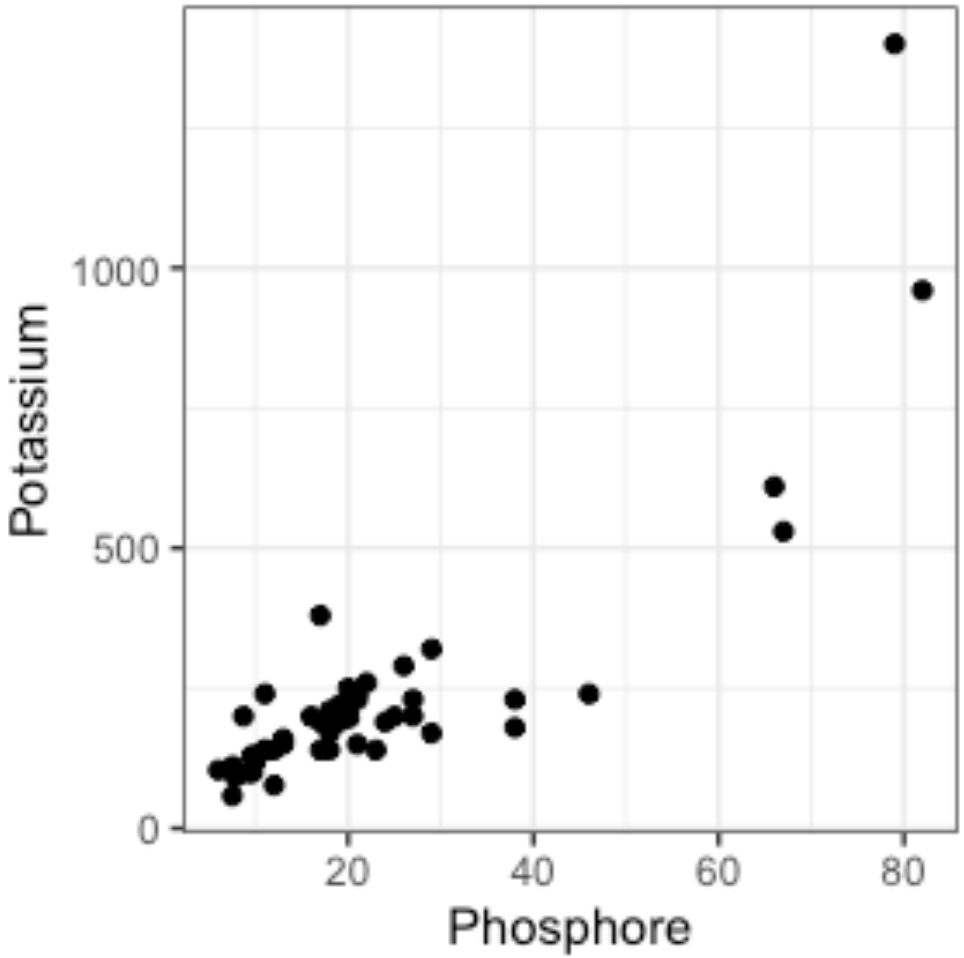

y
x
@allison_horst


y
x
@allison_horst

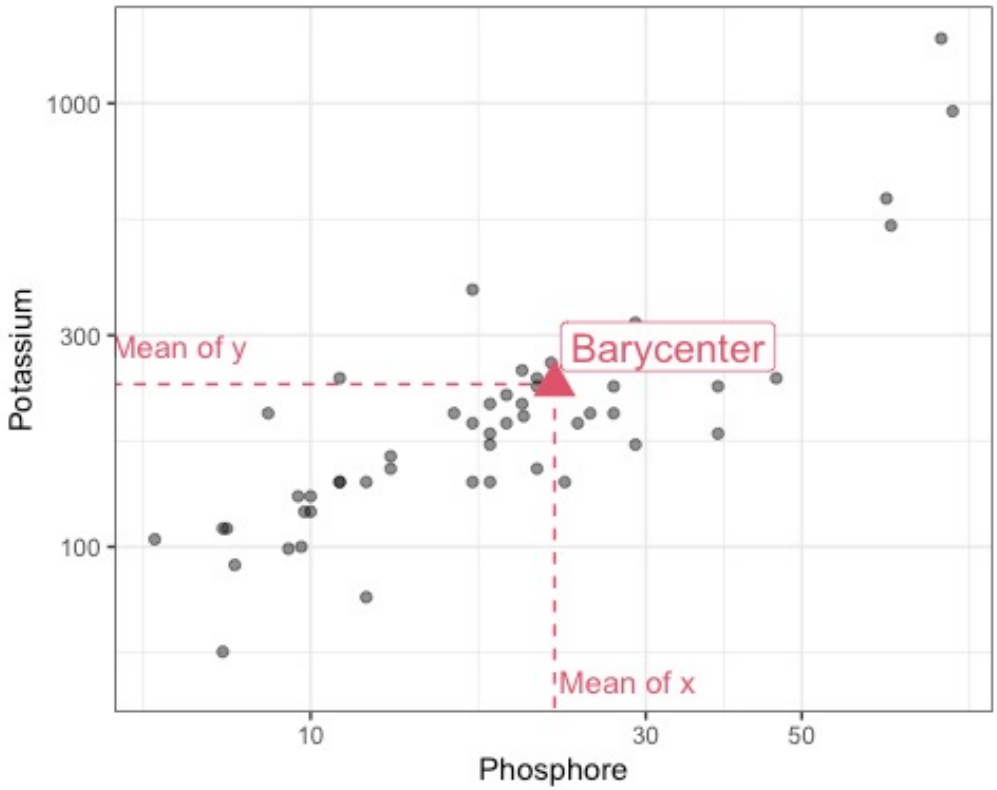


Covariance et corrélation

Sur des fruits



Barycentre (rappel)



Covariance (rappel)

Un point est-il proche ou loin du barycentre ? Rectangle !

$$(x_i - \bar{x}) \times (y_i - \bar{y})$$

La covariance est (presque) l'aire moyenne :

$$\mathrm{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \times (y_i - \bar{y})$$

Coefficient de corrélation

La covariance varie entre $-\infty$ et $+\infty$.

La corrélation est, par définition, une mesure de lien linéaire entre -1 et +1:

- -1 est une relation linéaire négative parfaite,
- 0 correspond à un lien nul,
- +1 est une relation linéaire positive parfaite.

Corrélation de Pearson

$$\mathrm{cor}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

… en bref…

$$\mathrm{cor}(x, y) = \frac{\mathrm{cov}(x, y)}{\mathrm{sd}(x)\mathrm{sd}(y)}$$

“Exercices”

Covariance :

```
cov(fruits$Potassium,
    fruits$Phosphore)
#> [1] 3315.292
```

Corrélation :

```
cor(fruits$Potassium,
    fruits$Phosphore)
#> [1] 0.8587945
```

Corrélation de Spearman's

Notée ρ. Sur les rangs !

- r_x les rangs de x,
- r_y les rangs de y.

$$\rho(x, y) = \mathrm{cor}\left(r_x, r_y\right)$$

Propriétés :

- Robuste aux valeurs exceptionnelles,
- Invariante par transformation monotone (p.ex. : log, square-root),
- Ne supporte pas les ex-aequos

Corrélation de Kendall

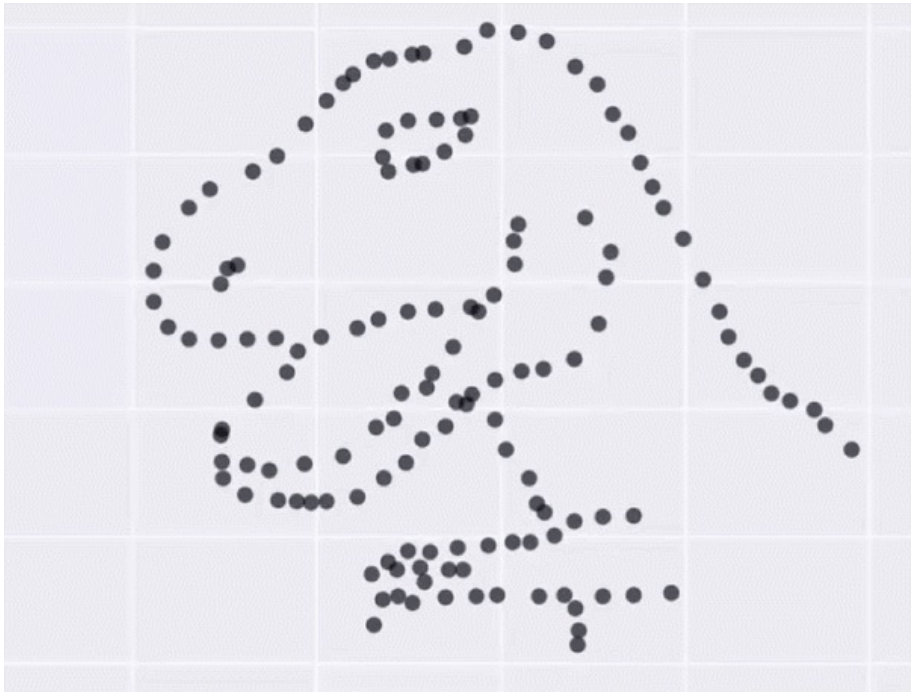
Paires de points : sur des sujets i et j.

- Paire Concordante : $\left(x_i < x_j \text{ et } y_i < y_j\right)$ OU $\left(x_i > x_j \text{ et } y_i > y_j\right)$
- Paire Discordante : $\left(x_i < x_j \text{ et } y_i > y_j\right)$ OU $\left(x_i > x_j \text{ et } y_i < y_j\right)$

$$\tau(x, y) = \frac{n_C - n_D}{n_0},$$

avec n_C le nombre de paires concordantes, n_D le nombre de paires discordantes et n_0 le nombre total.

Beware of naked numbers!



Datasaurus

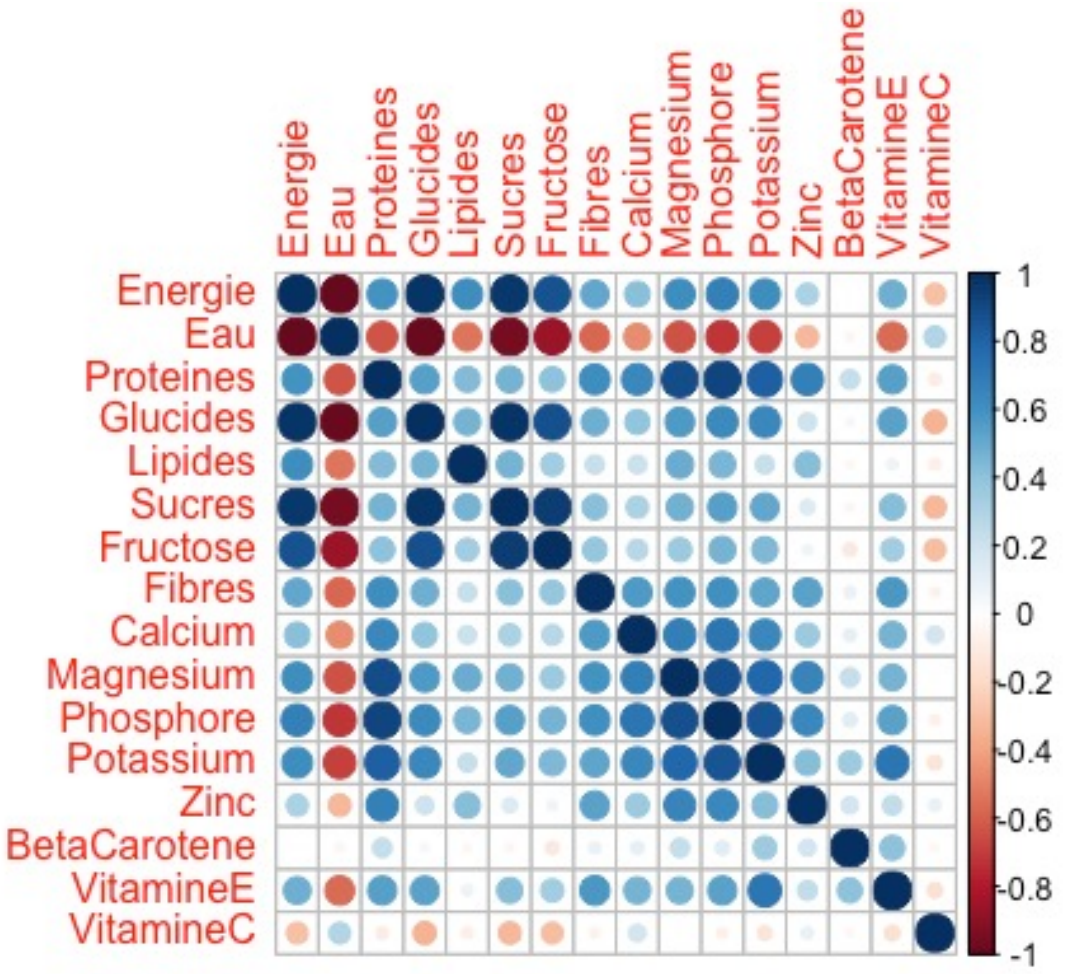
Plot the correlation

To compute all correlations:

```
cormat <- cor(fruits[, -(1:2)])
```

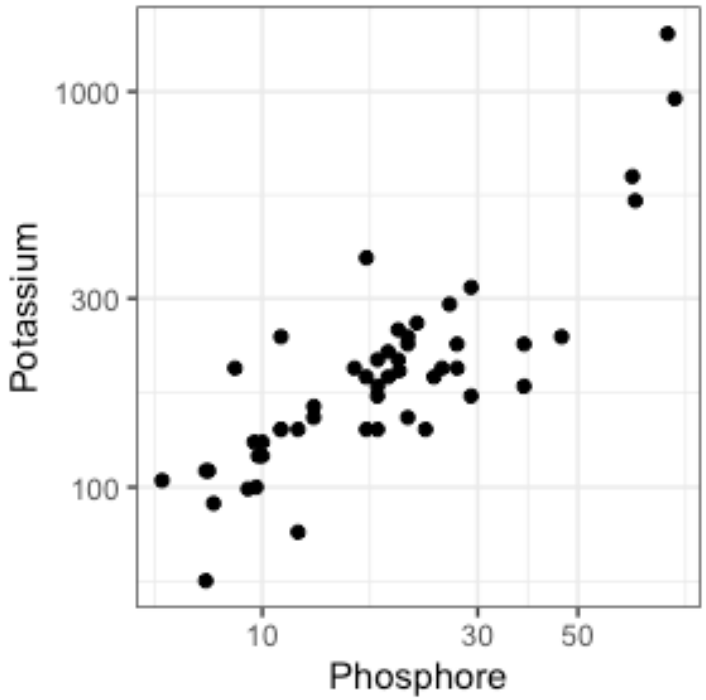
To make a “correlogram”:

```
corrplot(cormat)
```

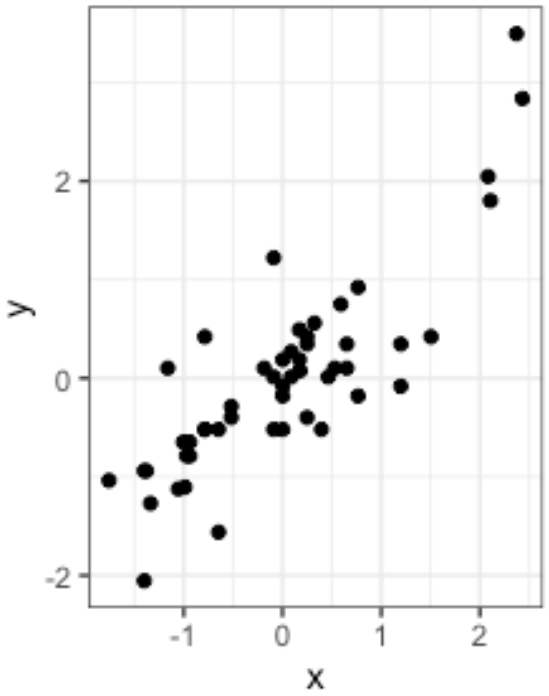


ACP en deux dimensions

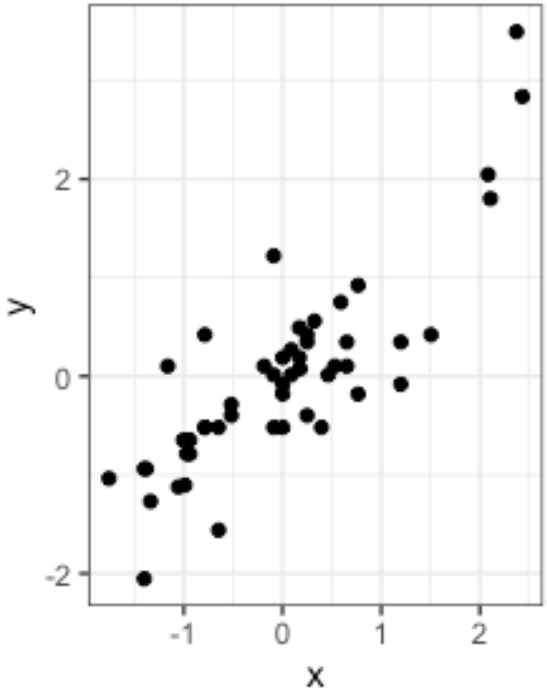
Sur les fruits



Pour **me** rendre
la vie plus facile



J’ai fait quoi ?

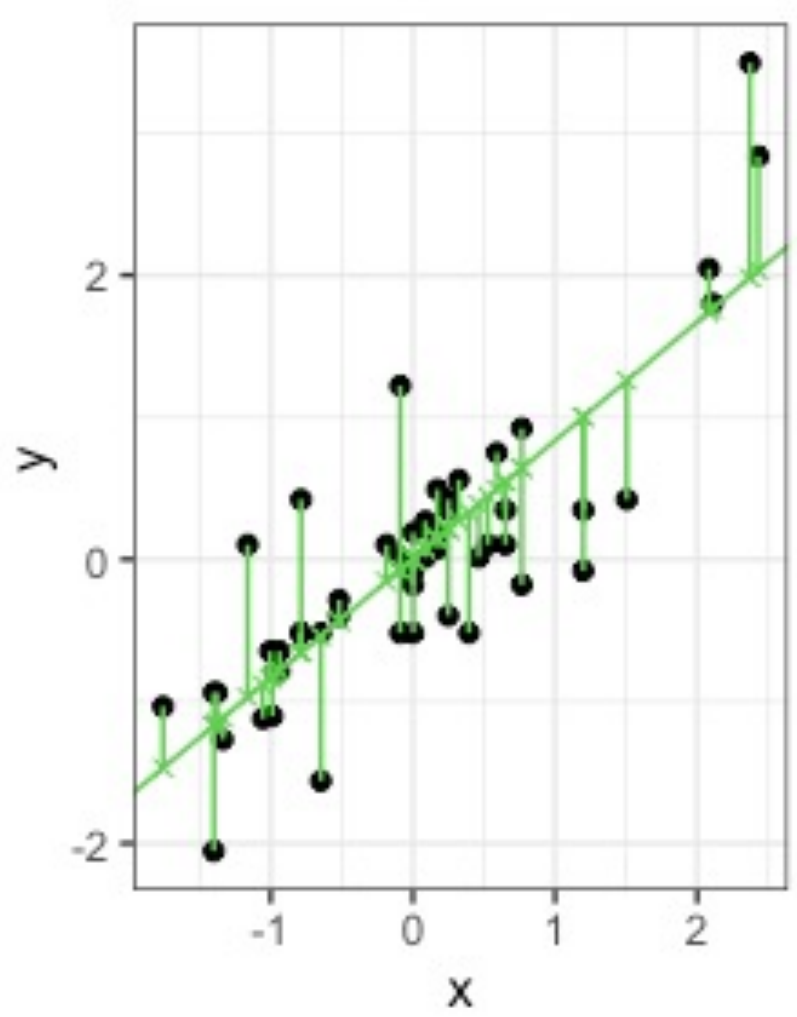


Régression linéaire

Modèle : $y = ax + b$

$a = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}$ est la pente de la droite,

$b = \bar{y} - a\bar{x}$ est l'ordonnées à l'origine.



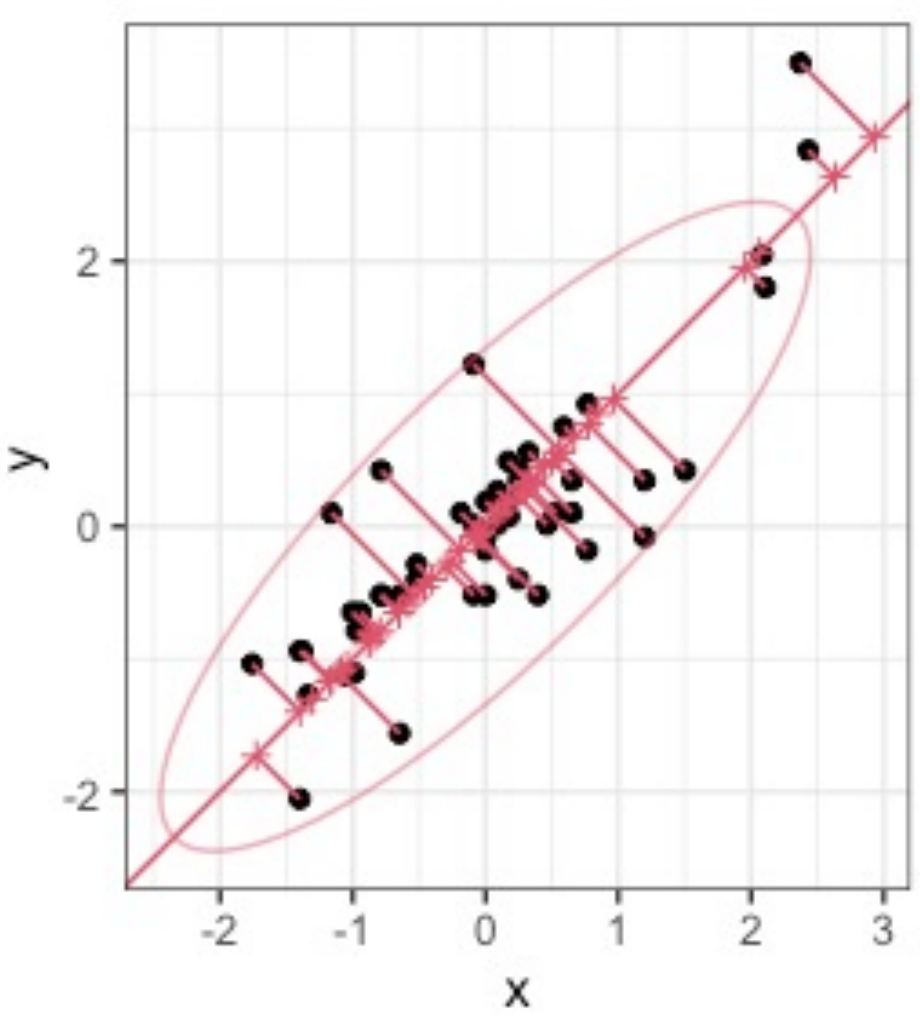
Première composante principale

$$PC_1 = a_1 x + a_2 y$$

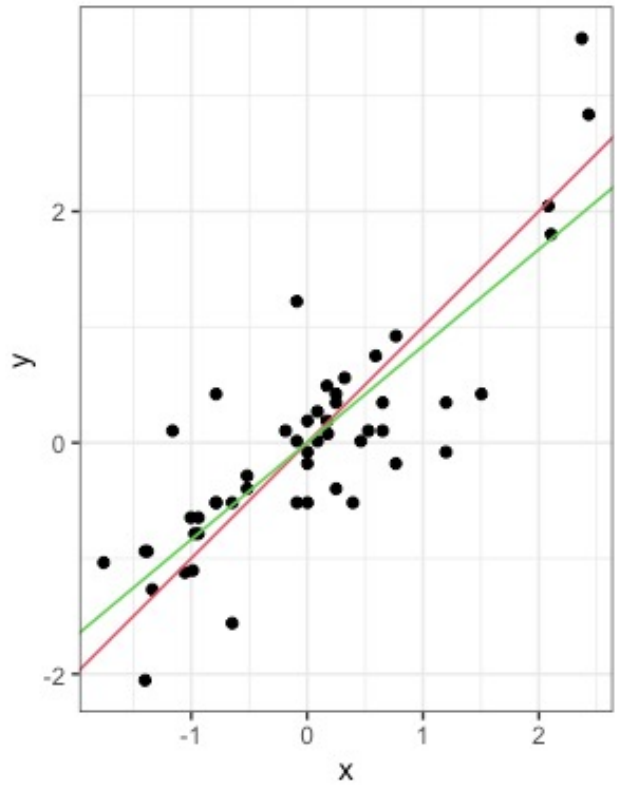
a_1 et a_2 sont les poids

PC_1 est la première composante principale

a_1 et a_2 sot calculés de telle sorte que PC_1 a la plus grande variance **ET** $a_1^2 + a_2^2 = 1$



Les deux sur le même graphe



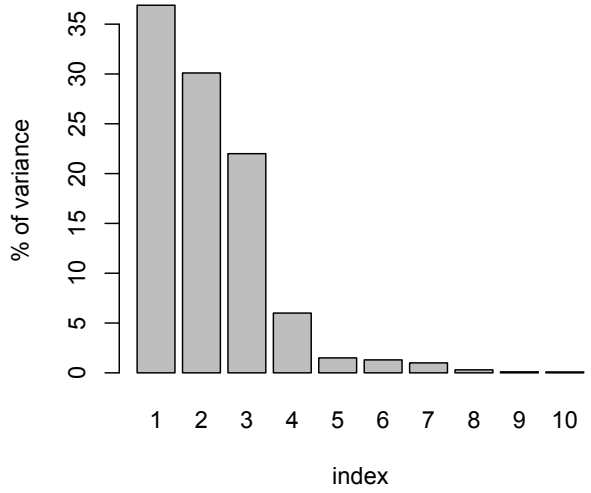
Généralités

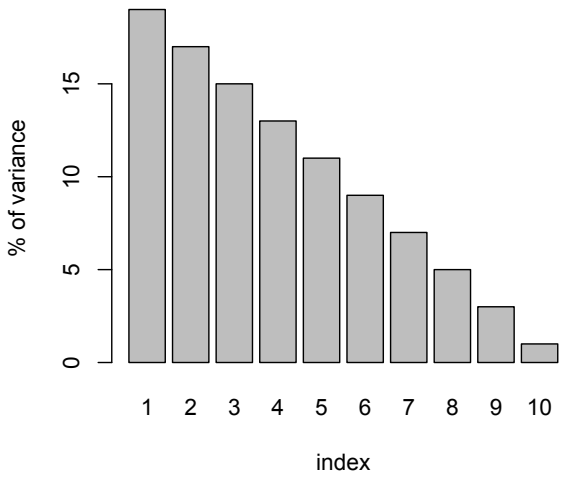
Combien de composantes principales

- Il y a au plus min(n-1,p) PCs
- Exemple : p = 200 gènes et n = 30 échantillons → au plus 29 composantes principales
- Que souhaitez-vous faire avec l'ACP ?
  - Visualisation ? Plutôt une poignée de composantes
  - Capturer un certain pourcentage de variabilité ?
  - Capturer la plus grande partie de la variabilité ?

The elbow rule

- Always useful to have a look at the plot of the fraction of total variance explained by each component.
- *Scree* graph (Cattell, 1966) : Look at the plot of the percentage of variance l_k against k and decide which value of k defines an *'elbow'* in the graph (subjective). In practice, rarely easy to choose.

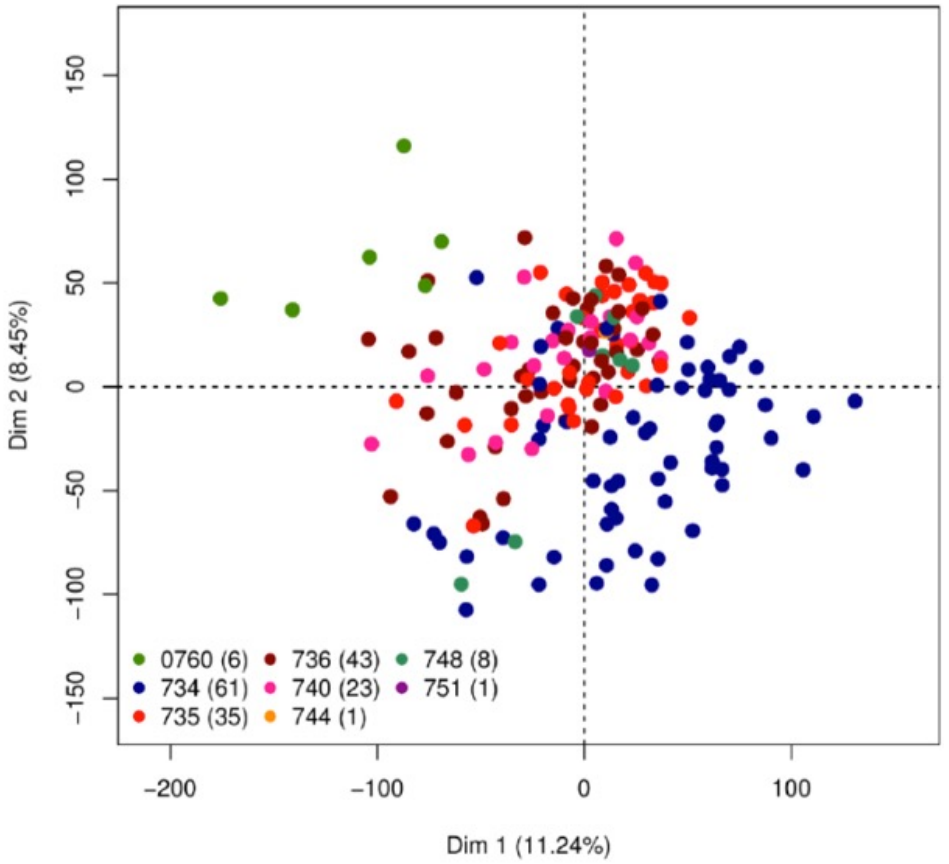




La carte des observations/patient·e·s

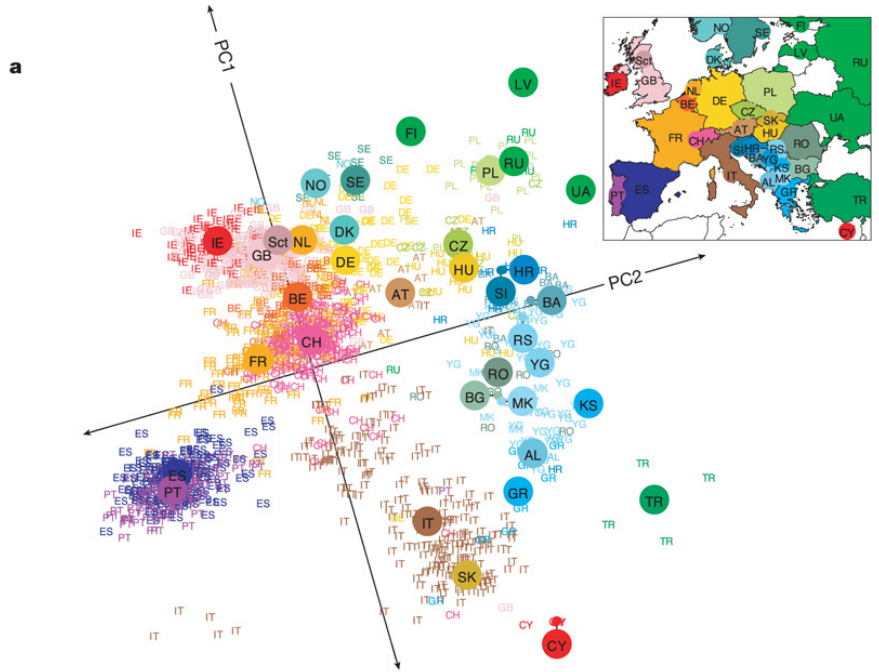
- Example: TCGA gene expression data set: Tumor samples colored by batch and projected on the first and the second PCs.

PCA allows to discover that the main source of variance is batch effect on this dataset



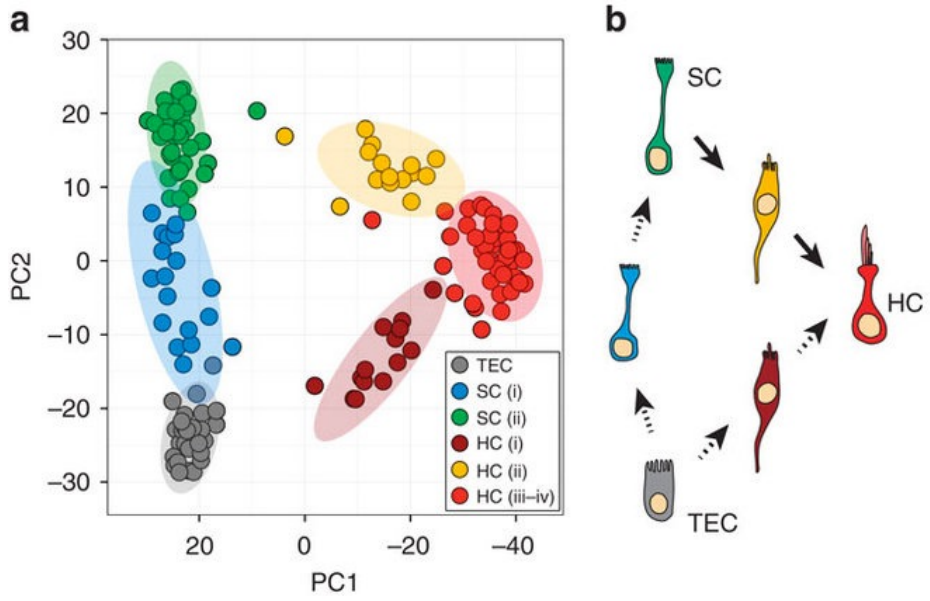
La carte des observations/individus

- Example: A dataset of genotypes measured on a SNP array: the projection of individuals on the first and the second PCs highlights population structure within Europe.

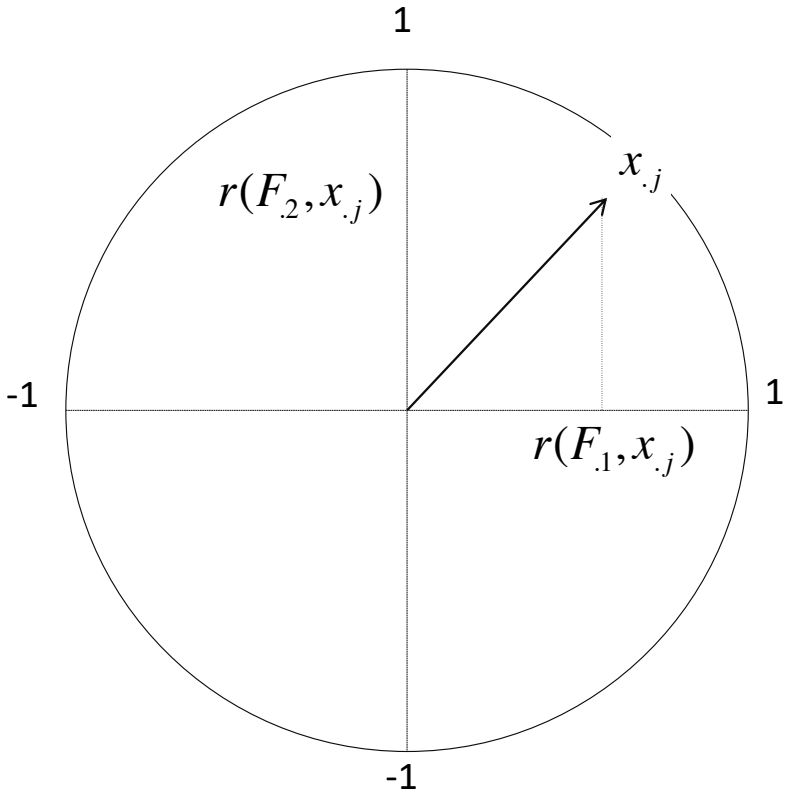


La carte des observations/cellules

- Example: RNA-Seq single cell data for single cells of the inner ear: the projection of samples on first and the second PCs reflects cell types.

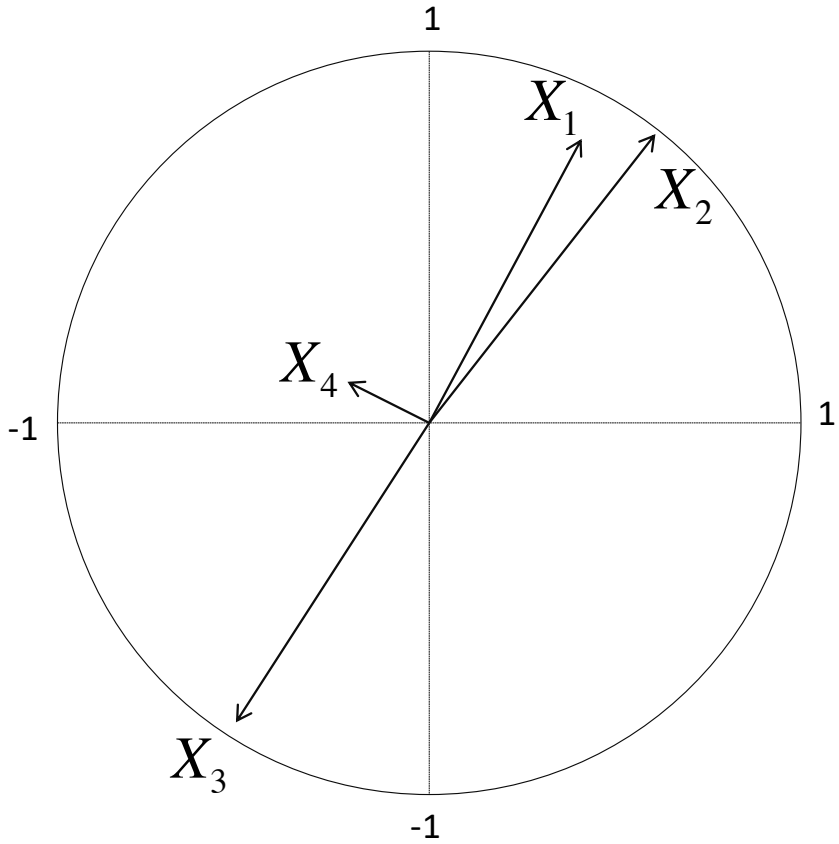


Le cercle des corrélations



Interprétation des composantes à partir des variables

- Une flèche proche du cercle indique que la variable est bien représentées par les CPs (X_1, X_2, X_3).
- Une flèche proche de 0 ne peut pas participer à l'interprétation de la variabilité capturée par les CPs (X_4).
- Deux flèches proches du cercle et proches entre elles sont très corrélées positivement (X_1, X_2).
- Deux flèches qui sont à l'opposé l'une de l'autre tout en restant proche du cercle sont très corrélées négativement (X_1, X_3 or X_2, X_3).



ACP en pratique

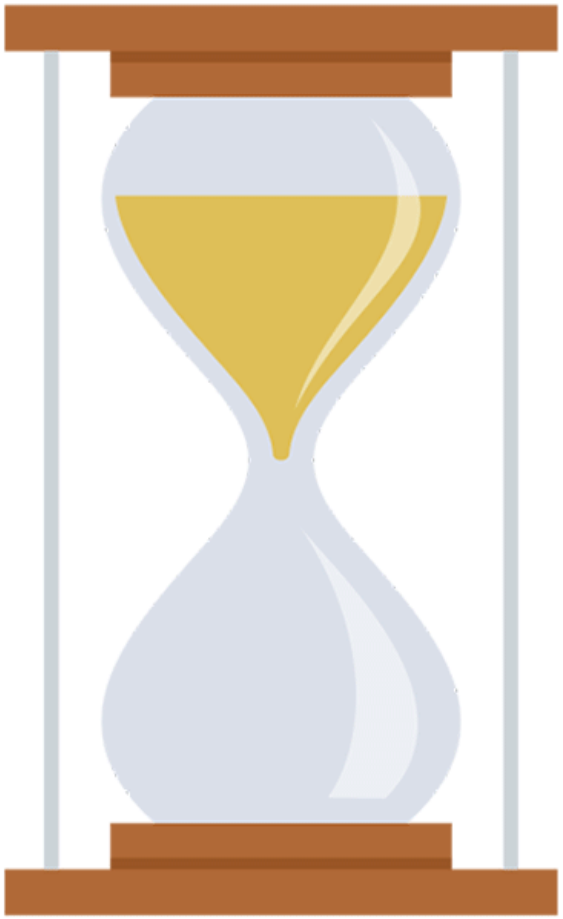
Sur les données fruits parce que c’est plus facile

Un peu de théorie

Intuition

- La première composante principale est une nouvelle variable qui est la plus corrélée possible à toutes les autres variables

$$\arg\max_{\mathbf{U}} \sum_{j=1}^{J} \mathrm{cor}(\mathbf{X}_j, \mathbf{U})^2$$



Un petit rappel : la décomposition en valeurs singulières

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\top$$

$$\begin{cases} \mathbf{X}\mathbf{X}^\top\mathbf{u} = \delta^2\mathbf{u} \\ \mathbf{X}^\top\mathbf{X}\mathbf{v} = \delta^2\mathbf{v} \end{cases} \qquad \begin{cases} \mathbf{X}^\top\mathbf{u} = \delta\mathbf{v} \\ \mathbf{X}\mathbf{v} = \delta\mathbf{u} \end{cases}$$

Méthode de la puissance itérée

$$\mathbf{v}^{(k+1)} \leftarrow \frac{1}{\|\mathbf{A}\mathbf{v}^{(k)}\|_2}\mathbf{A}\mathbf{v}^{(k)}$$

A essayer en premier lieu sur les données simulées très simples !

Comment obtenir la deuxième composante ? Par déflation

$$\widetilde{\mathbf{A}} \leftarrow \mathbf{A} - \lambda \mathbf{v}\mathbf{v}^{\top}$$

A essayer en premier lieu sur les données simulées très simples !