# Apprentissage non supervisé

# Chapitre 1 : Introduction

Allou Samé
allou.same@ifsttar.fr

# Plan


1. Introduction
   - Objectifs
   - Quelques applications
   - Terminologie
2. Statistiques descriptives
   - Tableaux de données
   - Types de variables
   - Transformations de variables
   - Statistiques descriptives monodimensionnelles
   - Statistiques descriptives bidimensionnelles
   - Cas multidimensionel
3. Mesures de proximité
   - Distances, normes
   - Dissimilarité et similarité
   - Ultramétrique
4. Réduction de la dimensionalité par ACP

# Objectifs de la classification non supervisée

## Objectifs

- Organiser les données en groupes (ou classes) homogènes
- Obtenir une représentation simplifiée d'un ensemble de données (analyse exploratoire)
- Réduire la taille des données (chaque groupe est remplacé par son représentant)
- Aider les praticiens à analyser leurs données

## Point de vue historique

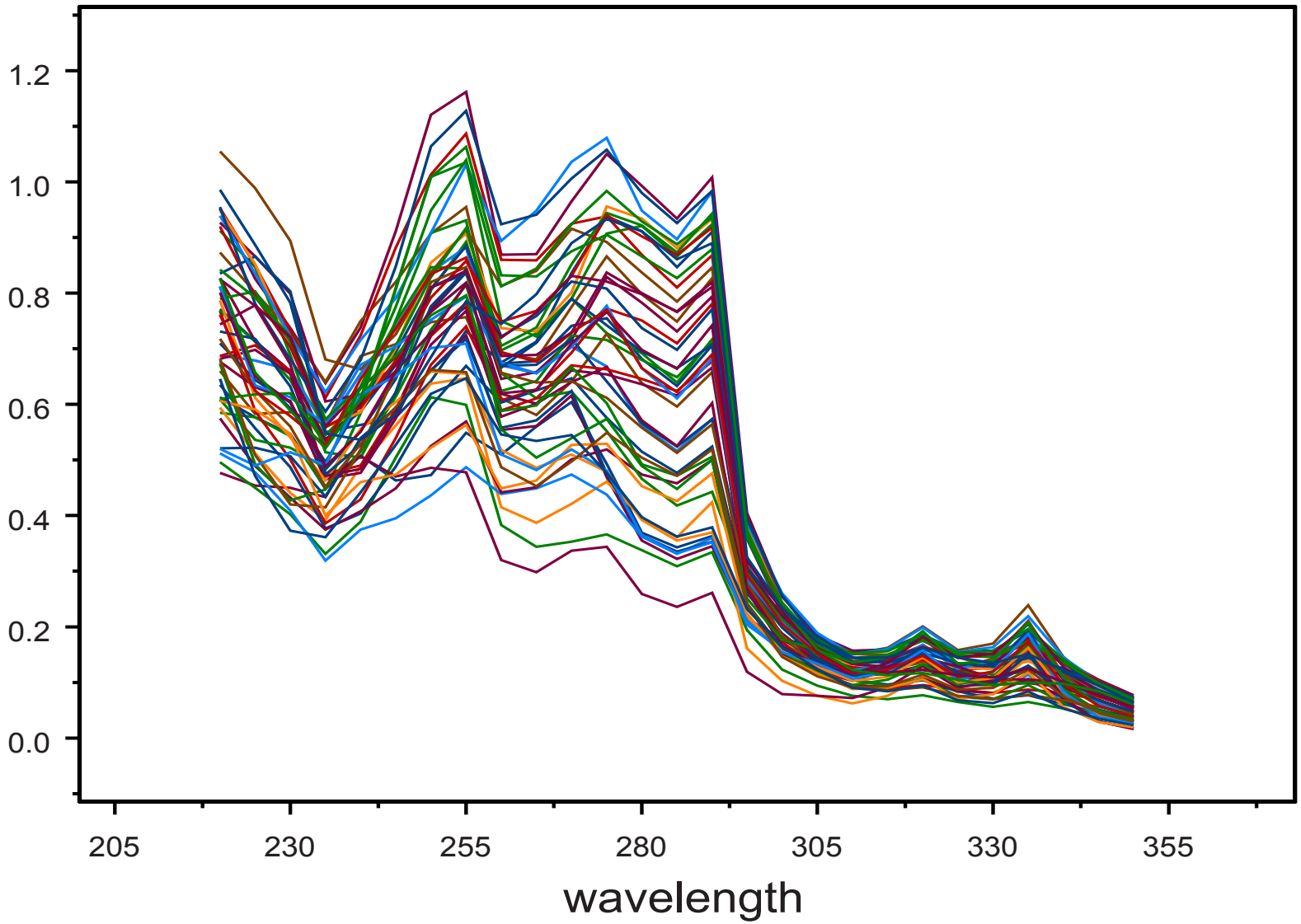
- Classification des genres naturels en trois groupes : animaux, plantes, minéraux
- Classification (nomenclature) des espèces animales et végétales (Von Carl Linné, $17^{e}$ siècle)

# Exemples d'applications

- Informatique : webmining, regroupement de pages web, d'usagers, compression de données
- Traitement d'image : quantification vectorielle, segmentation en zones homogènes
- Traitement du signal : reconnaissance de la parole et de sons...
- Neurosciences : classification des potentiels d'action ("spike sorting")
- Médecine & Bio-informatique : classification des maladies, regroupement de gènes
- Astronomie, géographie : regroupement d'étoiles, de planètes, partitionnement de régions et de villes
- Marketing : segmentation de la clientèle en classes homogènes
- Sociologie : typologie des cultures, des langues, analyse des réseaux sociaux
- Sciences sociales : identification de comportements

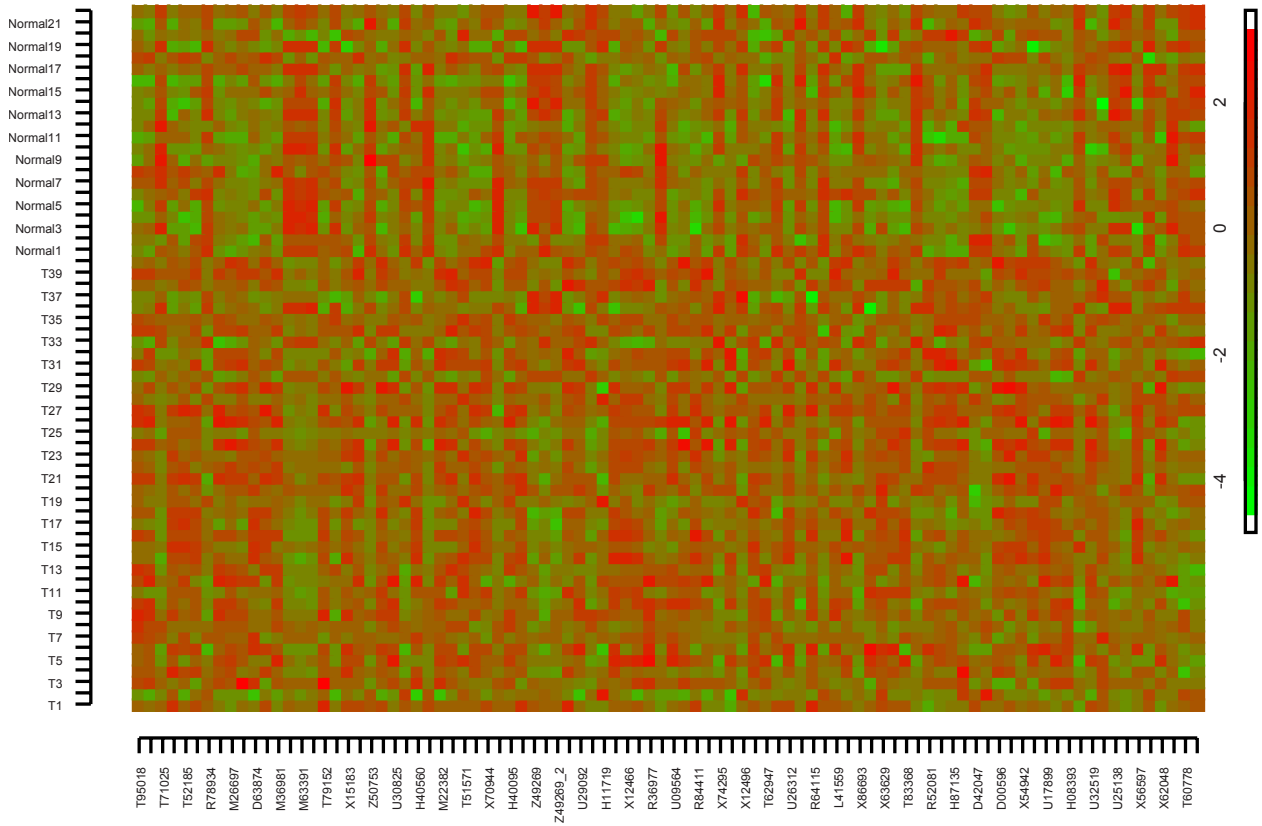
# Exemples

Données de spectrométrie de masse (Brereton, 2003)



# Exemples

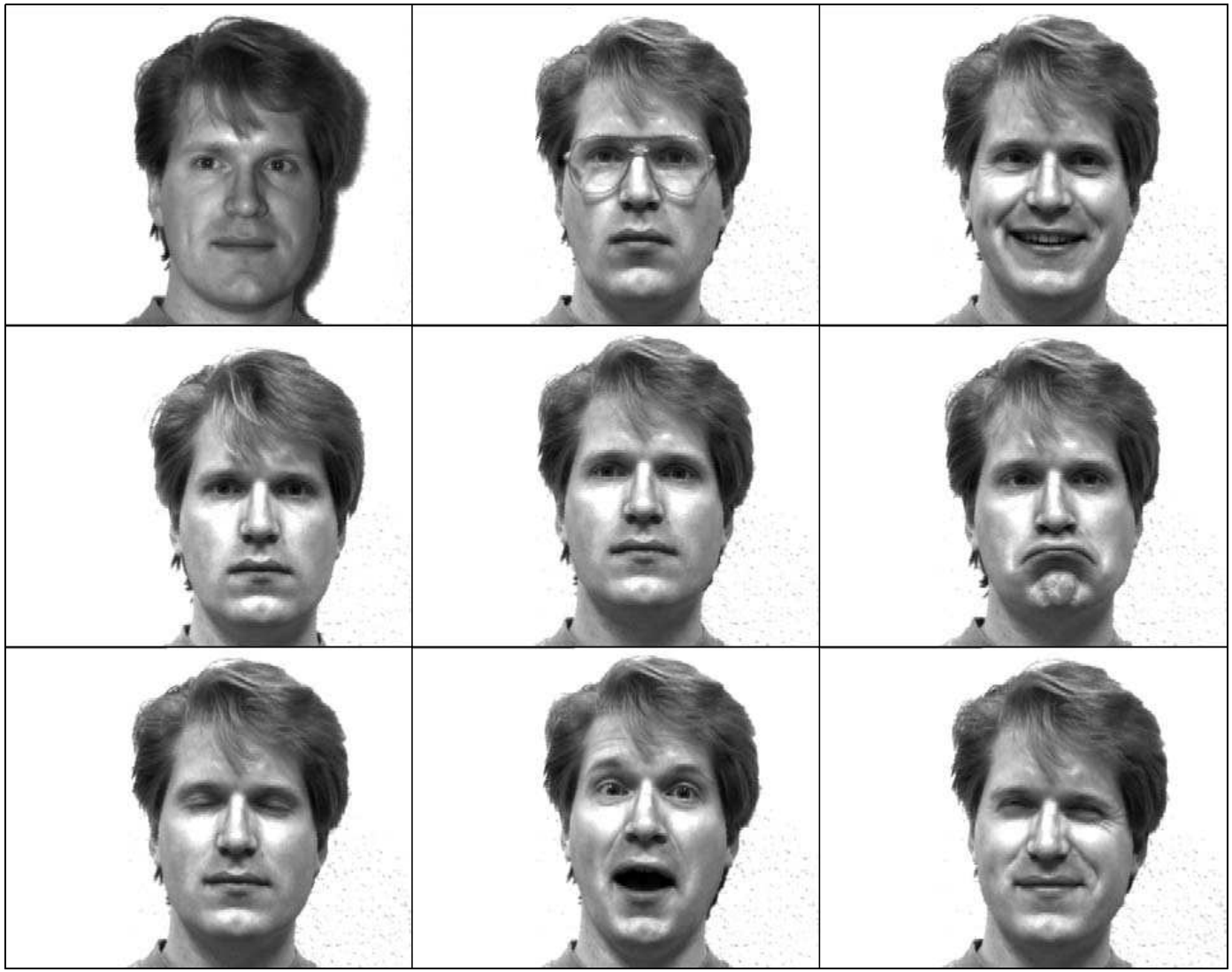
Puces ADN : données d'expression de 92 gènes pour 62 tissus (Alon et al., 1999)




# Genes = 92  # cell-lines= 62

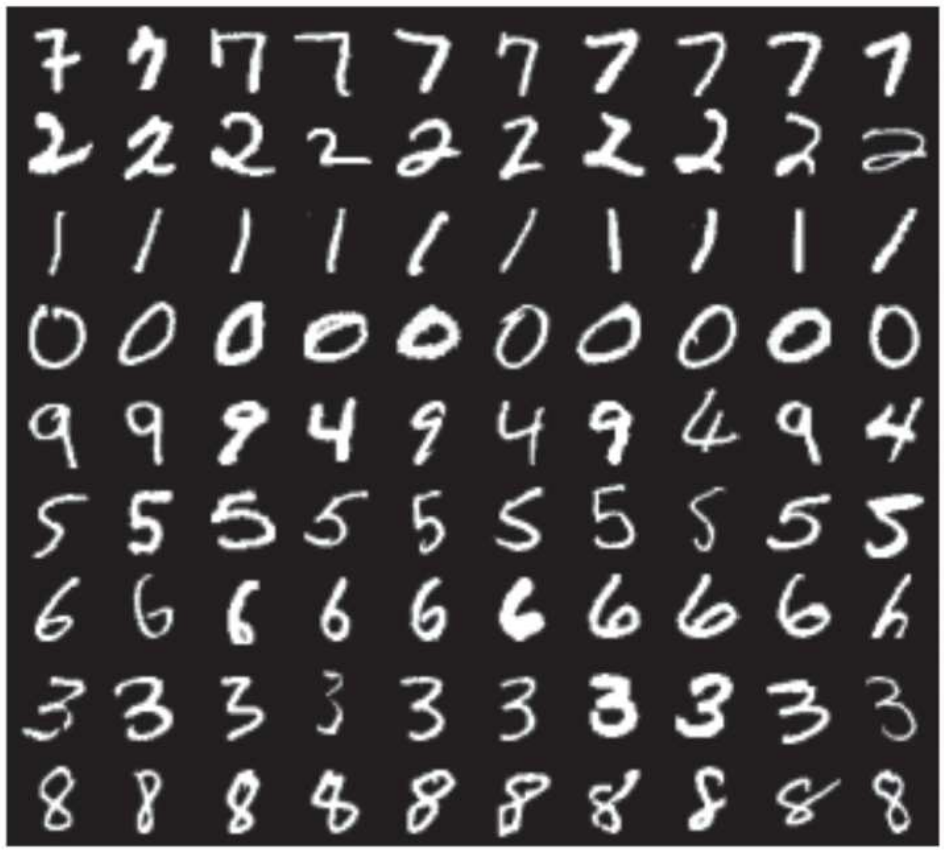
# Exemples

Reconnaissance faciale (Belhumeur et al., 1997)



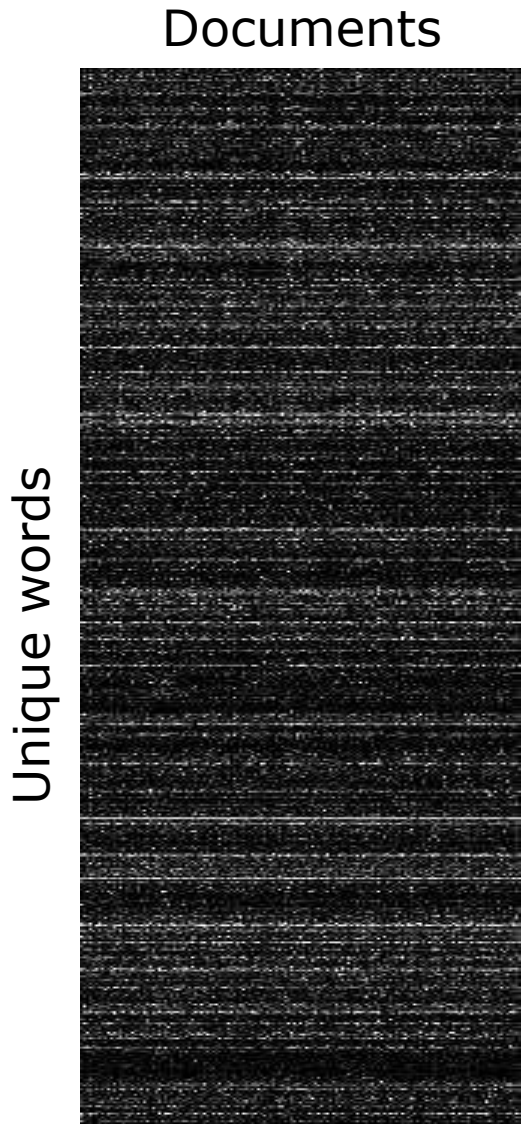
# Exemples

Chiffres manuscrits (extrait de la base MNIST)



# Exemple de données

Occurrence de mots dans des documents



# Classification automatique Vs. Classement

## Classification automatique

- Organisation des données en groupes homogènes (les groupes sont inconnus)
- Apprentissage non supervisé

## Classement

- Rangement des données dans des groupes connus à l'avance ; on parle aussi de discrimination
- Apprentissage supervisé

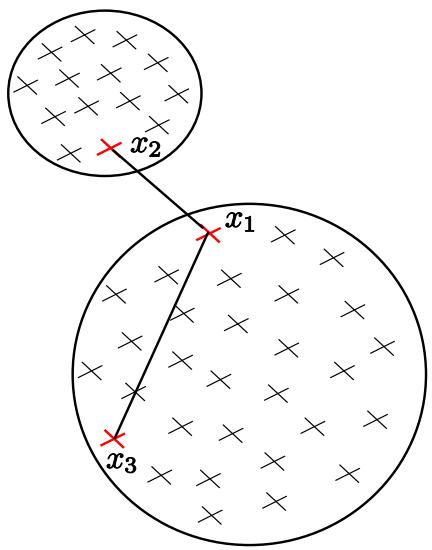
## Terminologie anglais-français

| français | anglais |
|---|---|
| classification automatique | clustering, cluster analysis |
| classement, discrimination | classification |

# Notion de classe

Plusieurs définitions ont été proposées :

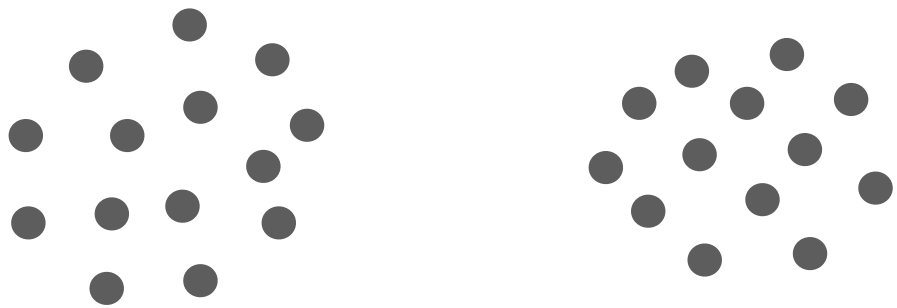
- Les classes sont des groupes d'objets homogènes : les données appartenant à une même classe se ressemblent
- Les classes sont des groupes d'objets bien séparés : les données provenant de classes différentes sont dissemblables
- Les classes sont des sous-ensembles de points de l'espace tel que la distance entre deux points d'une même classe est plus petite que la distance entre deux points de classes différentes




contre exemple

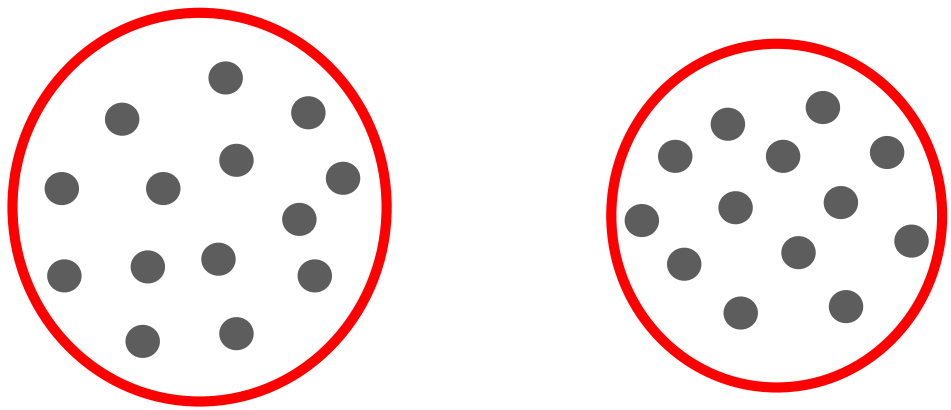
# Exemple de configurations de classes

Classes bien séparées



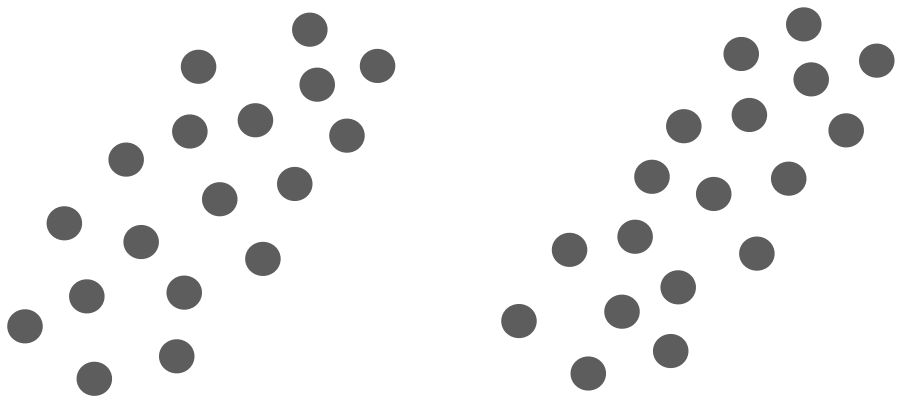
# Exemple de configuration de classes

Classes bien séparées



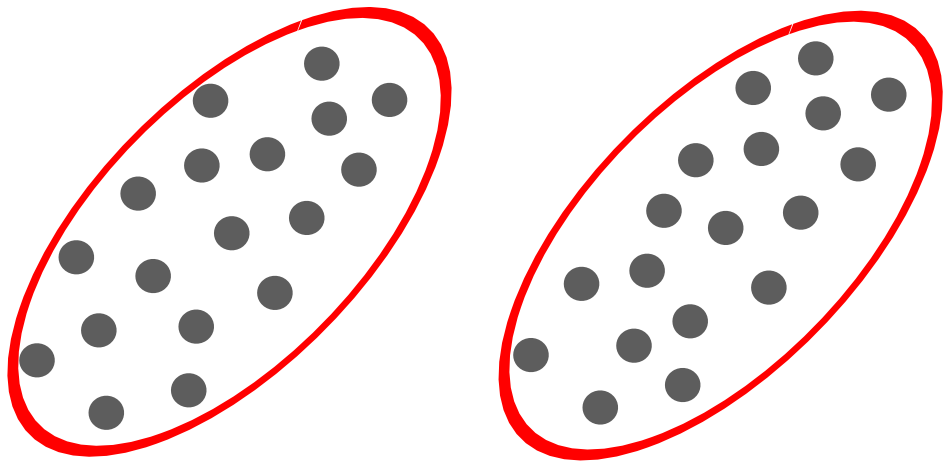
# Exemple de configuration de classes

Classes allongées



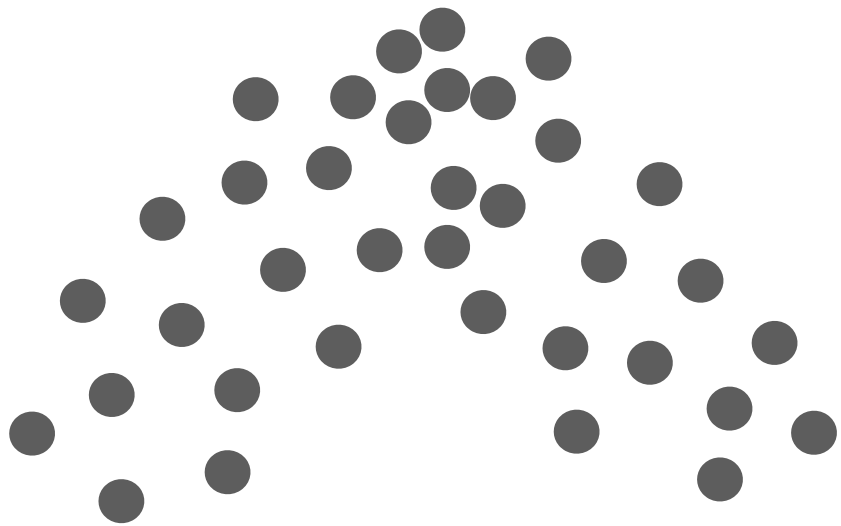
# Exemple de configuration de classes

Classes allongées



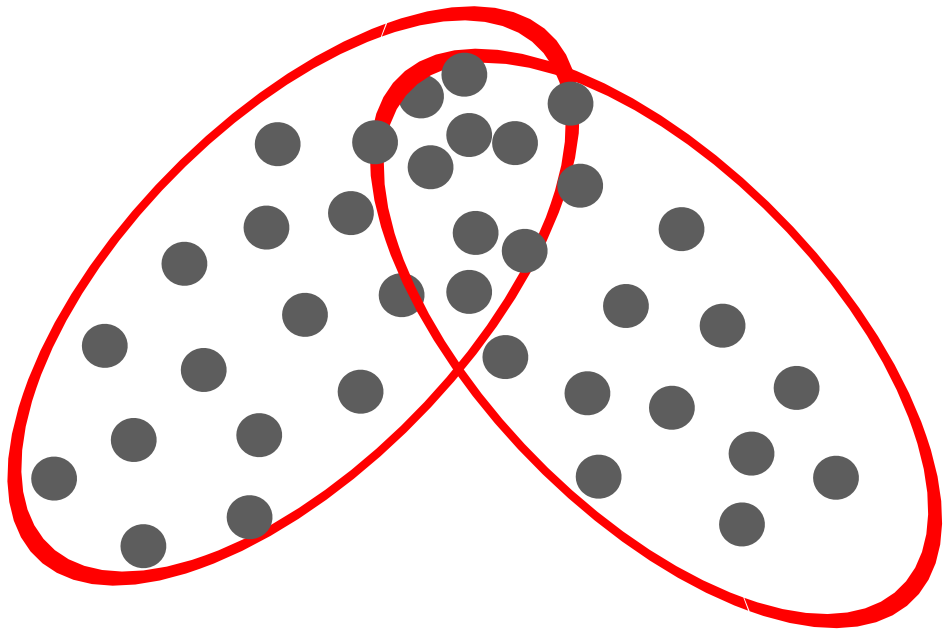
# Exemple de configuration de classes

Classes qui se chevauchent



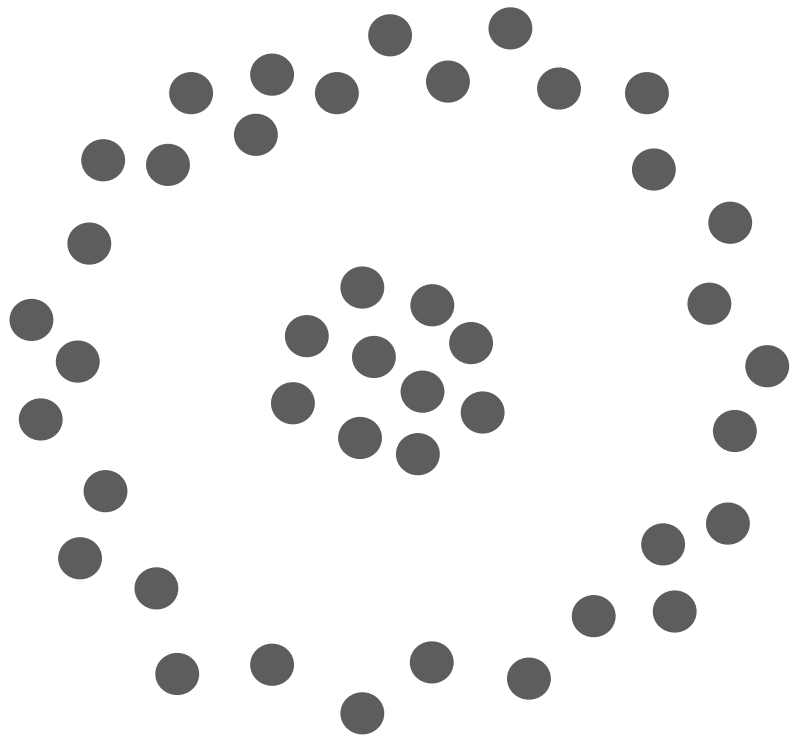
# Exemple de configuration de classes

Classes qui se chevauchent



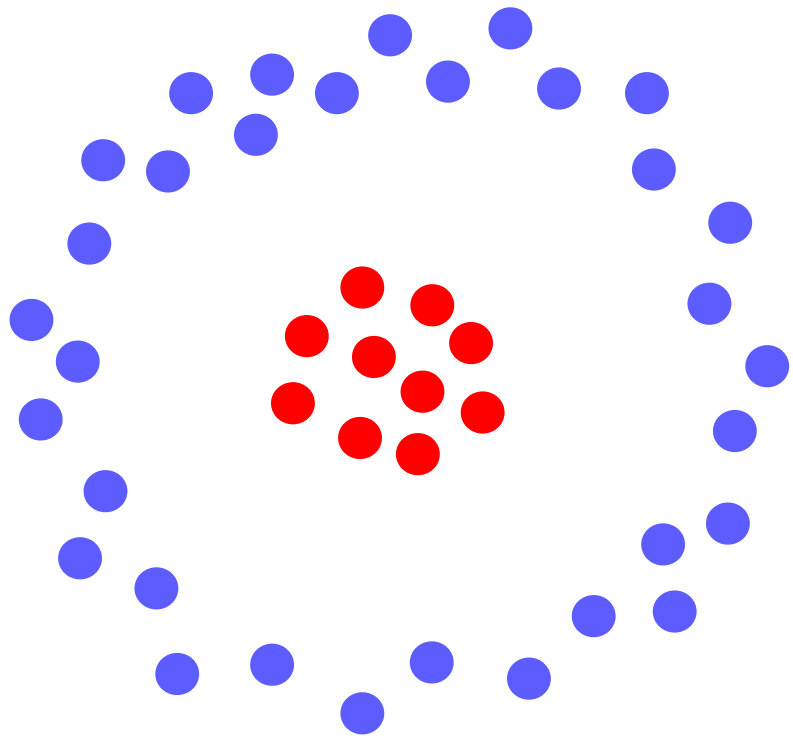
# Exemple de configuration de classes

Classes imbriquées



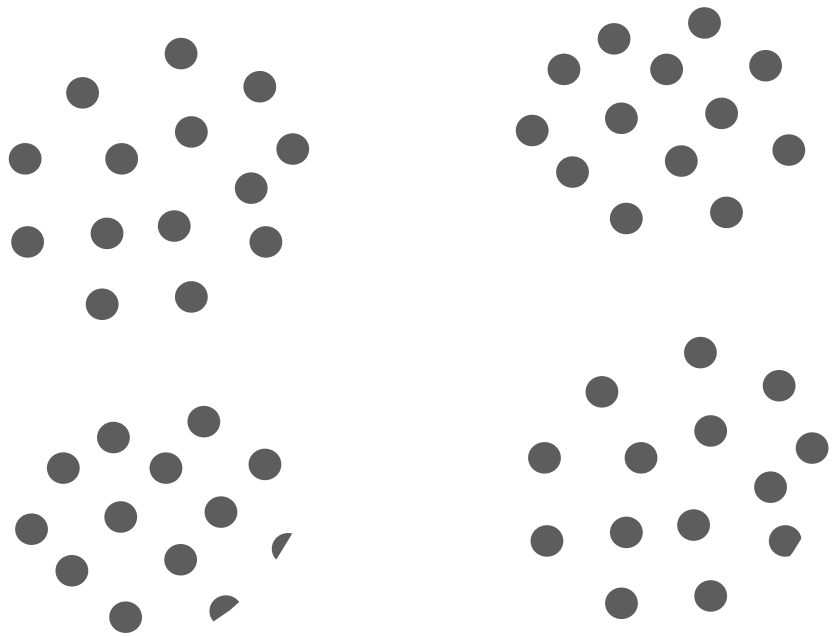
# Exemple de configuration de classes

Classes imbriquées



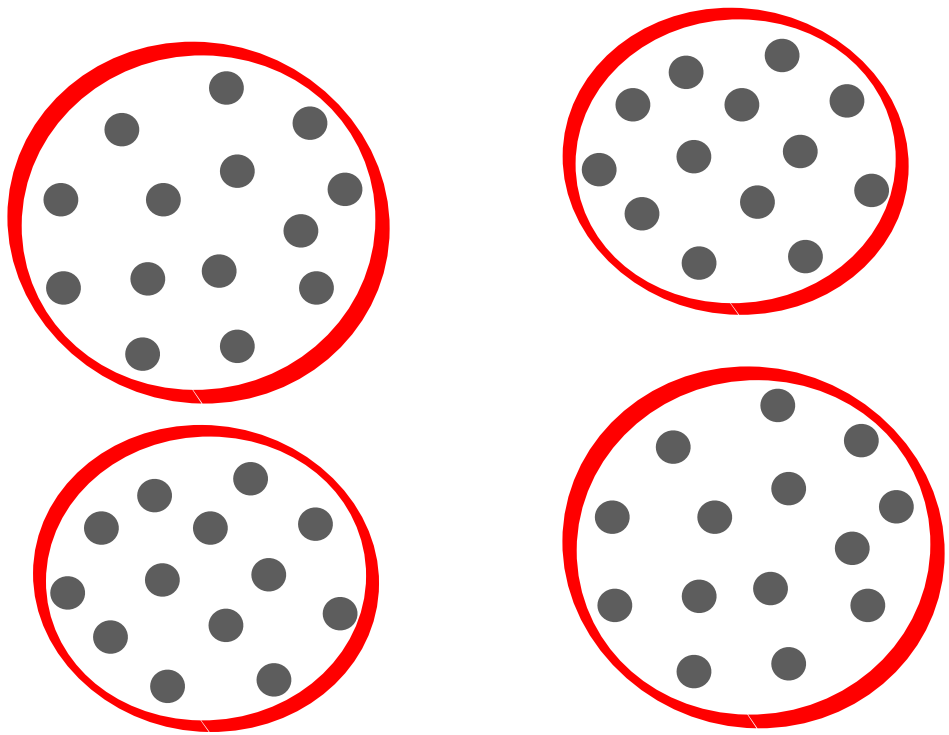
# Exemple de configuration de classes

Combien de classes ?



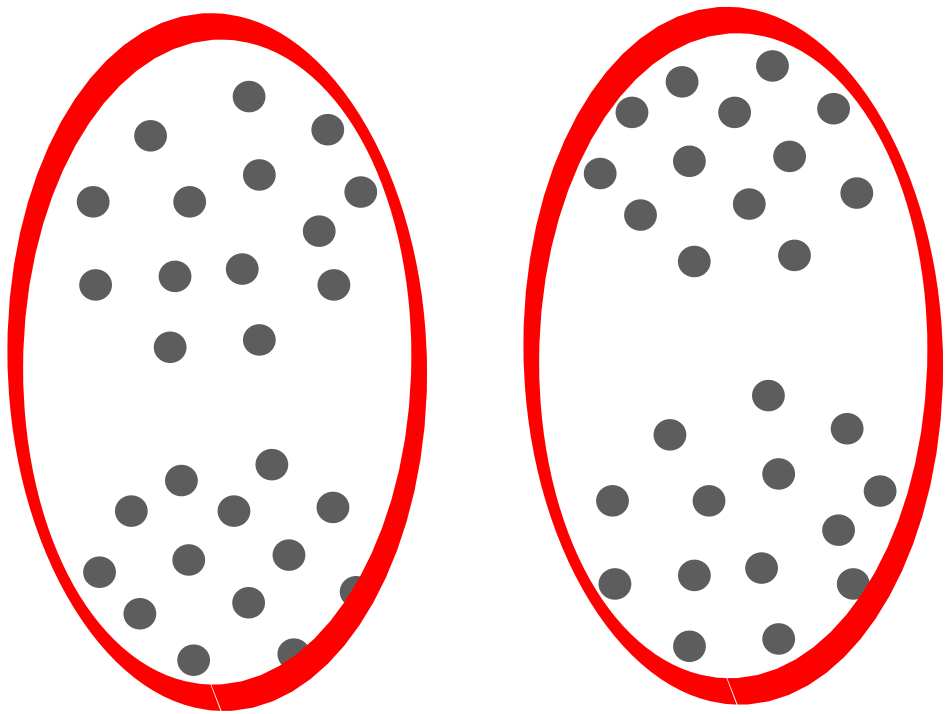
# Exemple de configuration de classes

4 classes ?



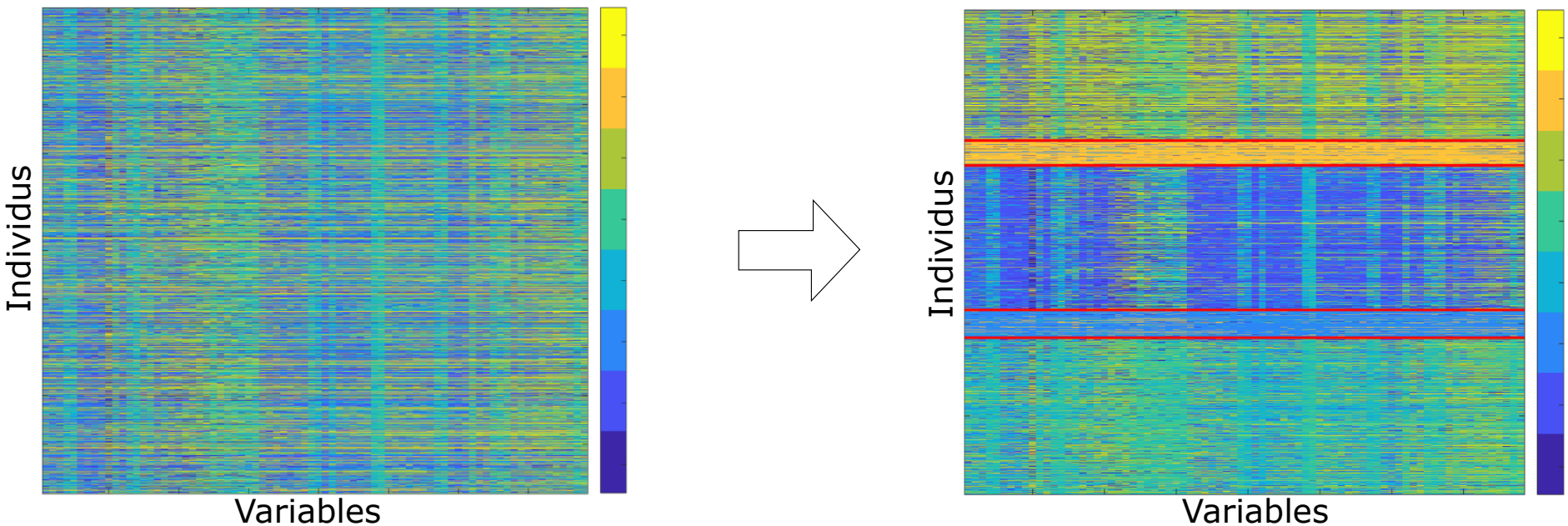
# Exemple de configuration de classes

2 classes ?



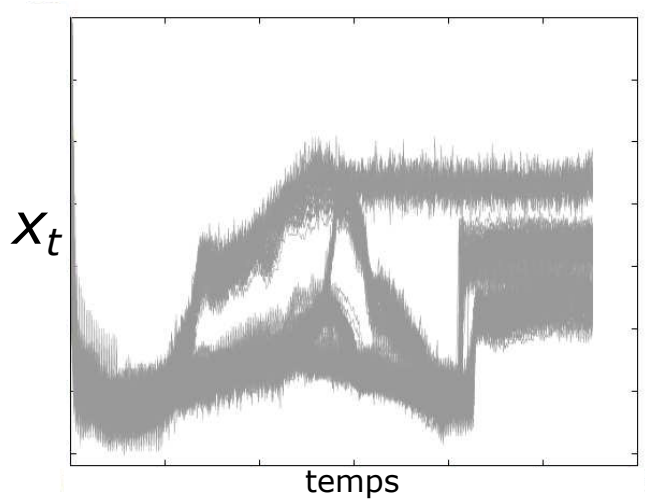
# Exemple de configuration de classes

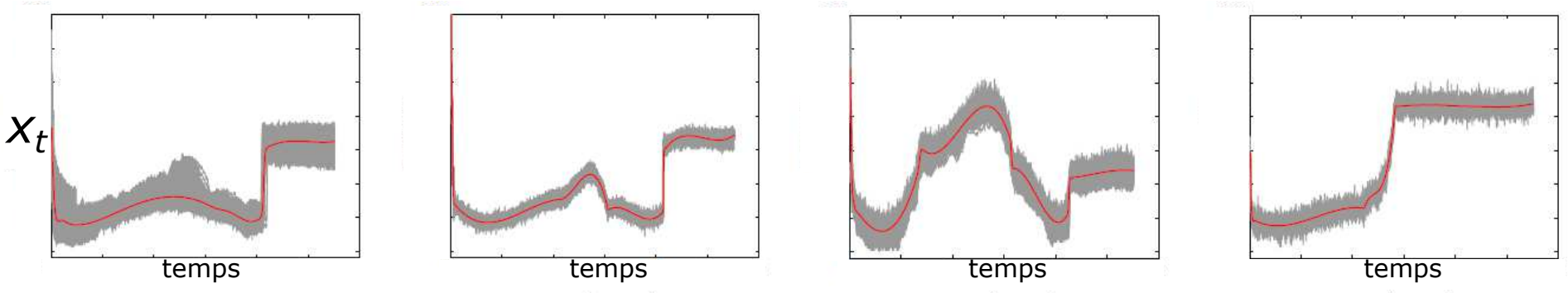
Données catégorielles



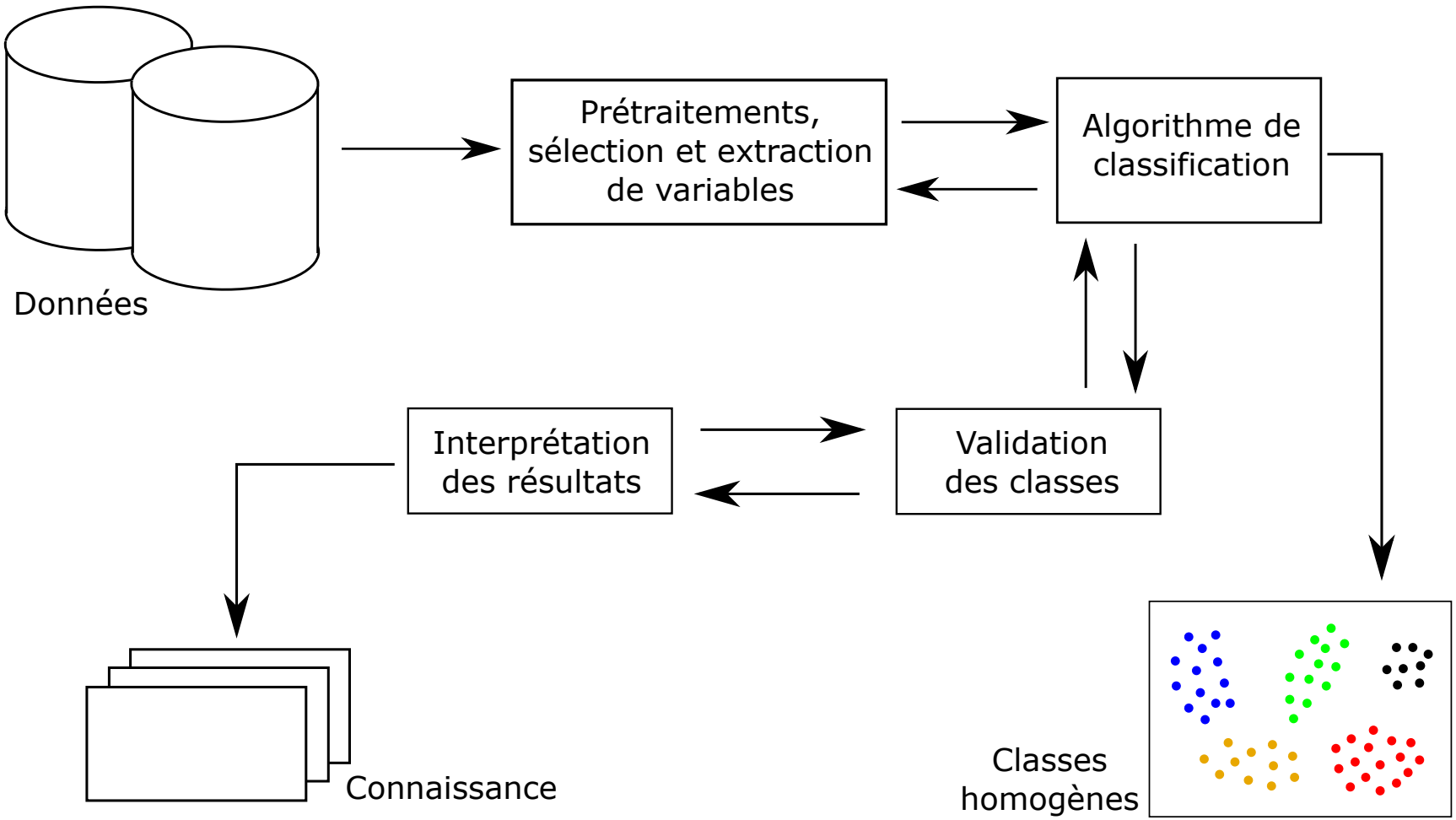
# Exemple de configuration de classes

Séries temporelles



# Différentes étapes du processus de classification automatique



# Format des données : tableau individus-variables

- Ensemble de $n$ individus décrits par $p$ variables (caractères)

$$E = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$$

avec $\mathbf{x}_i = (x_i^1, \ldots, x_i^j, \ldots, x_i^p)$

- L'ensemble $E$ peut aussi être représenté sous la forme d'un tableau $\mathbf{X}$ à $n$ lignes et $p$ colonnes

$$\mathbf{X} = \begin{pmatrix} x_1^1 & \ldots & x_1^j & \ldots & x_1^p \\ x_i^1 & \ldots & x_i^j & \ldots & x_i^p \\ x_n^1 & \ldots & x_n^j & \ldots & x_n^p \end{pmatrix}$$

# Exemple : données Iris (R. A. Fisher, 1936)

- Exemple classique étudié en statistique
- Souvent utilisé pour illustrer les méthodes de classification
- 150 iris provenant de 3 classes : Virginia, Versicolor et Setosa
- Données : longueur et largeur du sépale et du pétale

| long-sp | larg-sp | long-pt | larg-pt | espece |
|---|---|---|---|---|
| 5,1 | 3,5 | 1,4 | 0,2 | Iris-setosa |
| 4,9 | 3,0 | 1,4 | 0,2 | Iris-setosa |
| 4,7 | 3,2 | 1,3 | 0,2 | Iris-setosa |
| 4,6 | 3,1 | 1,5 | 0,2 | Iris-setosa |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| 5,5 | 2,4 | 3,7 | 1,0 | Iris-versicolor |
| 5,8 | 2,7 | 3,9 | 1,2 | Iris-versicolor |
| 6,0 | 2,7 | 5,1 | 1,6 | Iris-versicolor |
| 5,4 | 3,0 | 4,5 | 1,5 | Iris-versicolor |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| 6,3 | 3,3 | 6,0 | 2,5 | Iris-virginica |
| 5,8 | 2,7 | 5,1 | 1,9 | Iris-virginica |
| 7,1 | 3,0 | 5,9 | 2,1 | Iris-virginica |
| 6,3 | 2,9 | 5,6 | 1,8 | Iris-virginica |

# Autre format de données : tableau de proximités

## Tableau de proximités

Tableau carré de valeurs mesurant une ressemblance ou une dissemblance entre objets : distances géographiques, distances routières, durées de trajets, corrélations entre variables

Exemple : distances croisées entre des villes européennes

| | Lond | Stoc | Lisb | Madr | Par | Amst | Berl | Prag | Rome | Dubl |
|---|---|---|---|---|---|---|---|---|---|---|
| Londres | 0 | 569 | 667 | 530 | 141 | 140 | 357 | 396 | 569 | 190 |
| Stockholm | 569 | 0 | 1212 | 1043 | 617 | 446 | 325 | 423 | 787 | 648 |
| Lisbonne | 667 | 1212 | 0 | 201 | 596 | 768 | 923 | 882 | 714 | 714 |
| Madrid | 530 | 1043 | 201 | 0 | 431 | 608 | 740 | 690 | 516 | 622 |
| Paris | 141 | 617 | 596 | 431 | 0 | 177 | 340 | 337 | 436 | 320 |
| Amst | 140 | 446 | 768 | 608 | 177 | 0 | 218 | 272 | 519 | 302 |
| Berlin | 357 | 325 | 923 | 740 | 340 | 218 | 0 | 114 | 472 | 514 |
| Prague | 396 | 423 | 882 | 690 | 337 | 272 | 114 | 0 | 364 | 573 |
| Rome | 569 | 787 | 714 | 516 | 436 | 519 | 472 | 364 | 0 | 755 |
| Dublin | 190 | 648 | 714 | 622 | 320 | 302 | 514 | 573 | 755 | 0 |

# Types de variables

**Quantitative** : variable à valeurs dans $\mathbb{R}$

- **discrète** : variable à valeurs dans un sous-ensemble dénombrable de $\mathbb{R}$
  ex : age en années, nombre d'enfants dans un foyer
- **continue** : variable est à valeurs dans un intervalle de $\mathbb{R}$
  ex : taille, poids, consommation, revenu, montant facture d'électricité, diamètre d'une pièce en sortie d'usine...

**Qualitative** : variable à valeur dans un ensemble fini

- **nominale** : pas de relation d'ordre entre les modalités
  ex : sexe, situation familiale
- **ordinale** : relation d'ordre entre les modalités
  ex : réponse à un sondage ayant pour modalités : "très bon", "bon", "moyen", "mauvais", "très mauvais"

# Transformations de variables quantitatives

## Centrer-réduire

- Centrer : soustraire de chaque valeur la moyenne de la variable
- Réduire : diviser chaque valeur par l'écart-type de la variable
  Permet d'uniformiser l'échelle de grandeur des variables
- Centrer et réduire : faire les deux opérations

## Discrétiser

Transformer une variable quantitative en variable qualitative

- discrétisation définie a priori : ex. remplacer l'âge par une des valeurs 1, 2, 3, 4 suivant les intervalles : (1) 0-18 ans, (2) 19-40 ans, (3) 41-65 ans, (4) $>$ 65 ans
- Transformer des données numériques en histogramme : discrétisation en intervalles de même longueur

# Transformation d'une variable qualitative en variable binaire

- Cas d'une variable nominale : codage disjonctif complet (on remplace la variable qualitative par les indicatrices de chaque modalité)

$$\begin{array}{ccc} a & & 1\ 0\ 0 \\ b & & 0\ 1\ 0 \\ a & \Longleftrightarrow & 1\ 0\ 0 \\ c & & 0\ 0\ 1 \\ c & & 0\ 0\ 1 \end{array}$$

- Cas d'une variable ordinale : codage additif

$$\begin{array}{ccc} a & & 1\ 0\ 0 \\ b & & 1\ 1\ 0 \\ a & \Longleftrightarrow & 1\ 0\ 0 \\ c & & 1\ 1\ 1 \\ c & & 1\ 1\ 1 \end{array}$$

# Statistiques élémentaires sur une variable quanti.

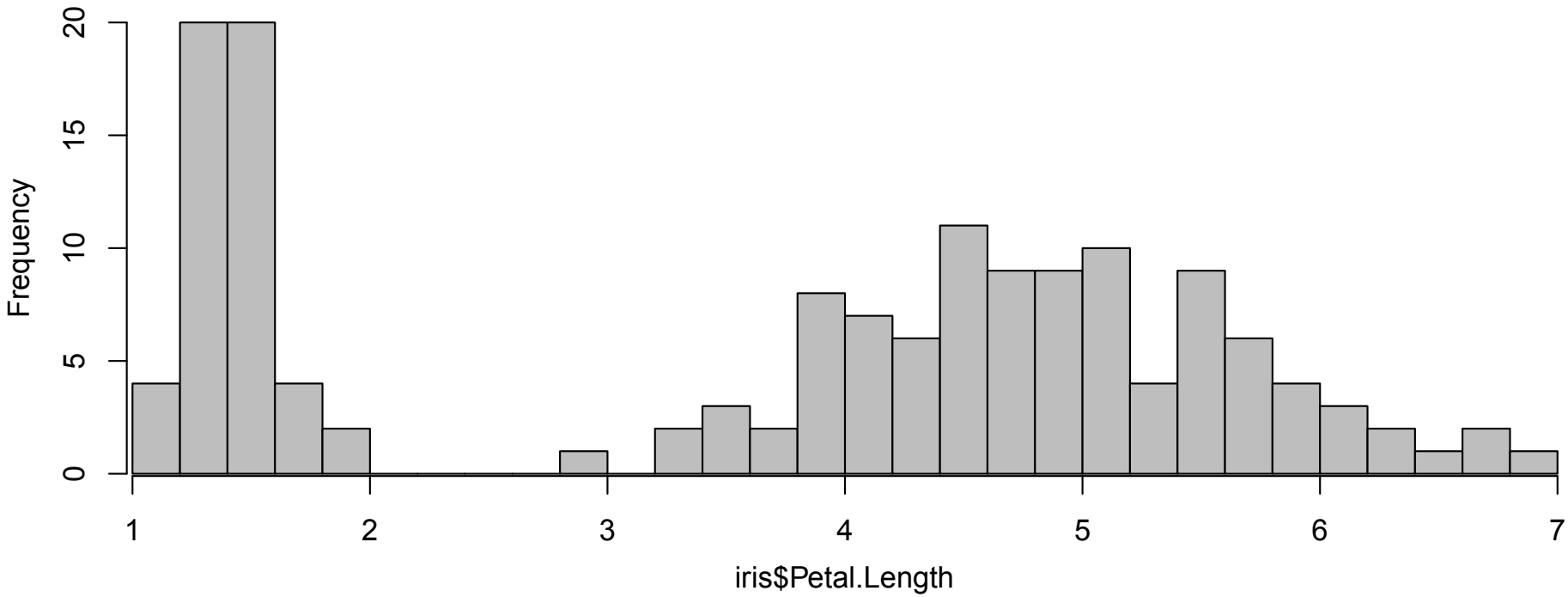
- Minimum et maximum
- Moyenne empirique : $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$
- Variance empirique : $(s_j)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^j - \overline{x}_j)^2$
- Écart-type empirique : $s_j = \sqrt{(s_j)^2}$
- Quartiles :
  - premier quartile $q_1$ : valeur qui partage l'échantillon en 25% à gauche et 75% à droite
  - deuxième quartile $q_2$ qui correspond à la médiane : valeur qui partage l'échantillon en 50% à gauche et 50% à droite
  - troisième quartile $q_3$ : valeur qui partage l'échantillon en 75% à gauche et 25% à droite

# Statistiques élémentaires sur données Iris

```
> summary(iris)
```

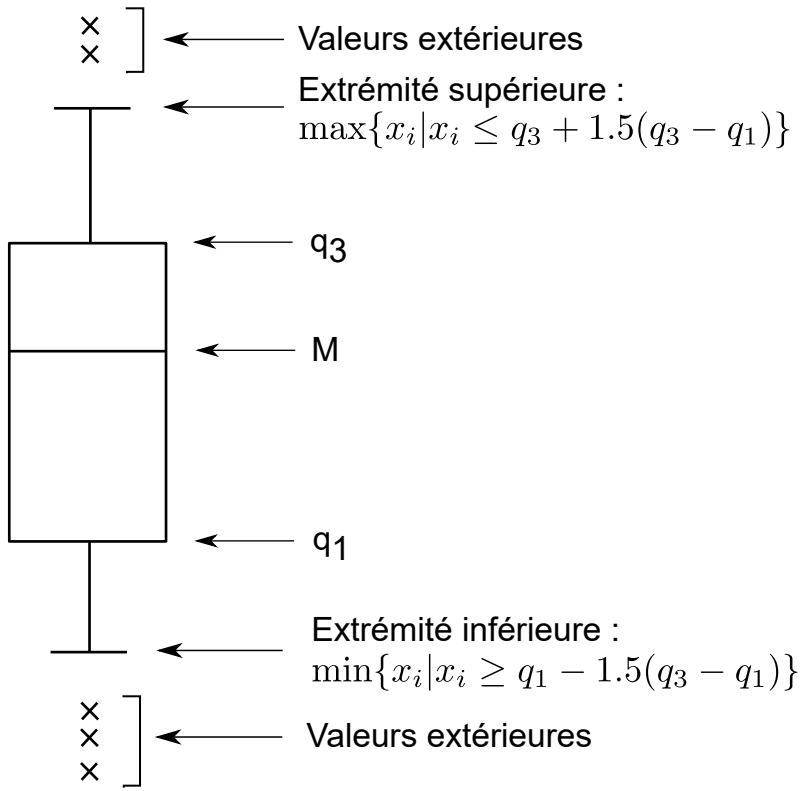
| sep_length | sep_width | pet_length | pet_width |
|---|---|---|---|
| Min. :4.300 | Min. :2.000 | Min. :1.000 | Min. :0.100 |
| 1st Qu.:5.100 | 1st Qu.:2.800 | 1st Qu.:1.600 | 1st Qu.:0.300 |
| Median :5.800 | Median :3.000 | Median :4.350 | Median :1.300 |
| Mean :5.843 | Mean :3.054 | Mean :3.759 | Mean :1.199 |
| 3rd Qu.:6.400 | 3rd Qu.:3.300 | 3rd Qu.:5.100 | 3rd Qu.:1.800 |
| Max. :7.900 | Max. :4.400 | Max. :6.900 | Max. :2.500 |

# Histogramme

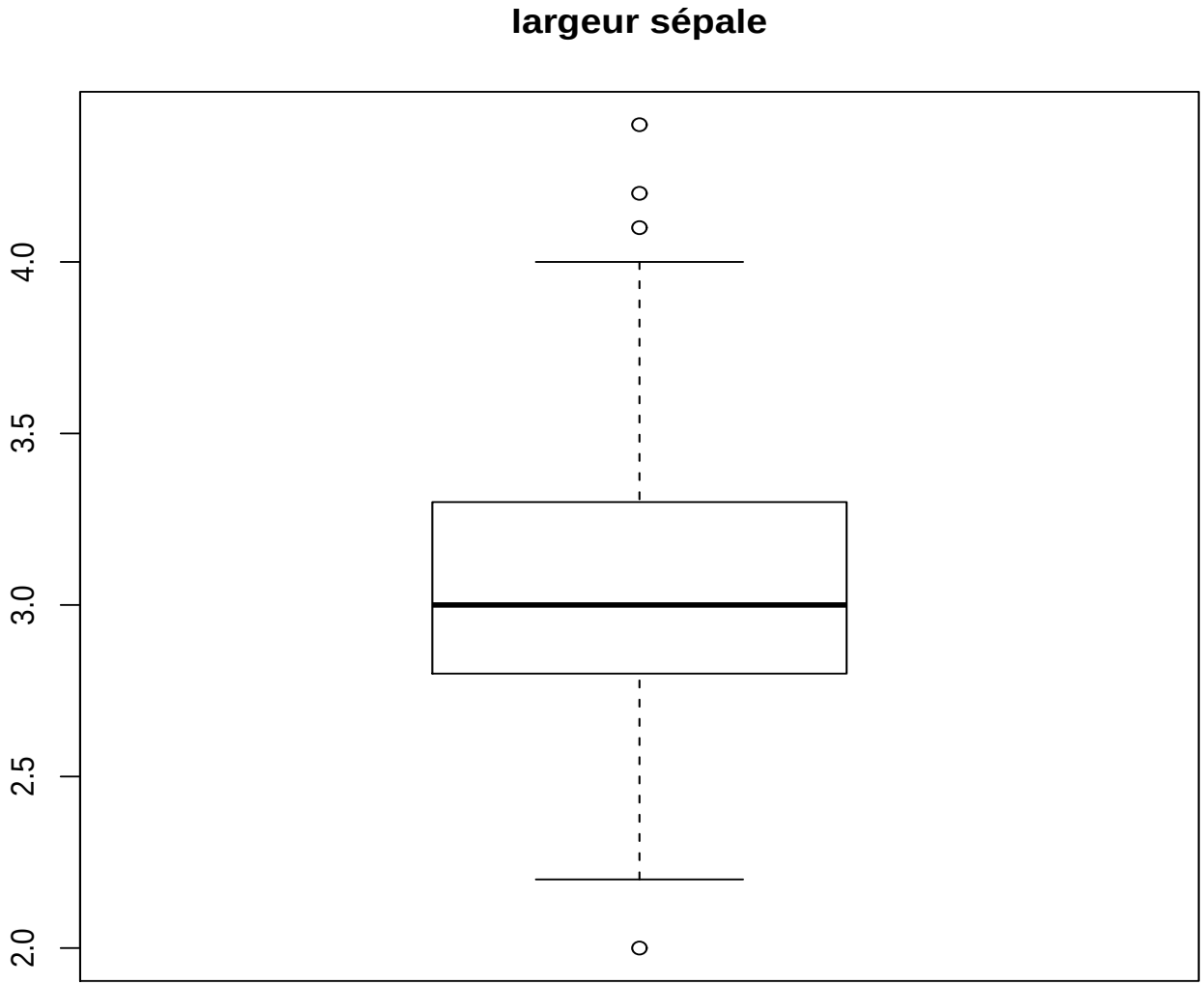


# Boîte à moustache ou boxplot

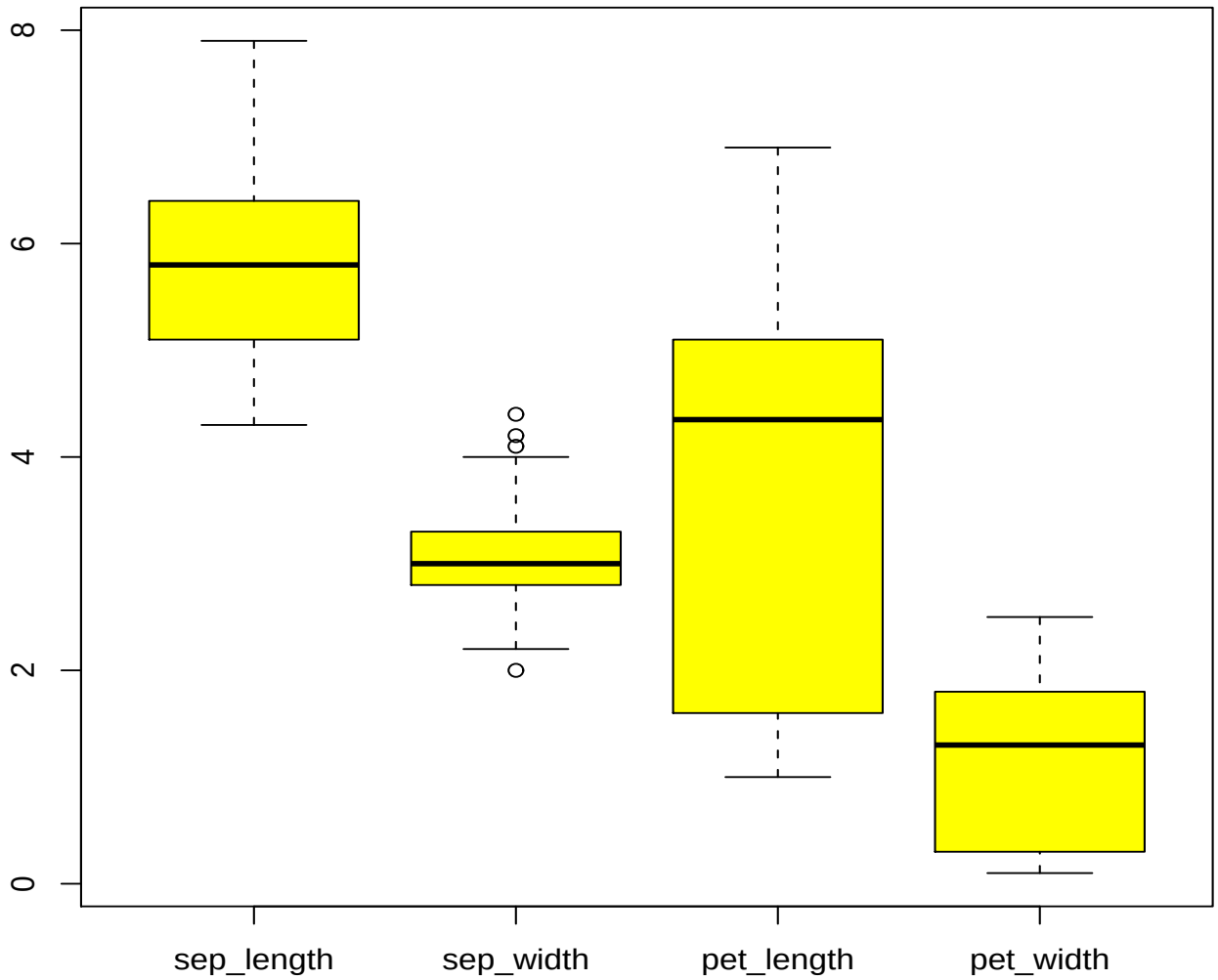
- Représente les principales caractéristiques d'une variable numérique
- Permet de repérer d'éventuelles valeurs aberrantes
- Facilite la comparaison de plusieurs distributions



# Exemple de boîte à moustaches



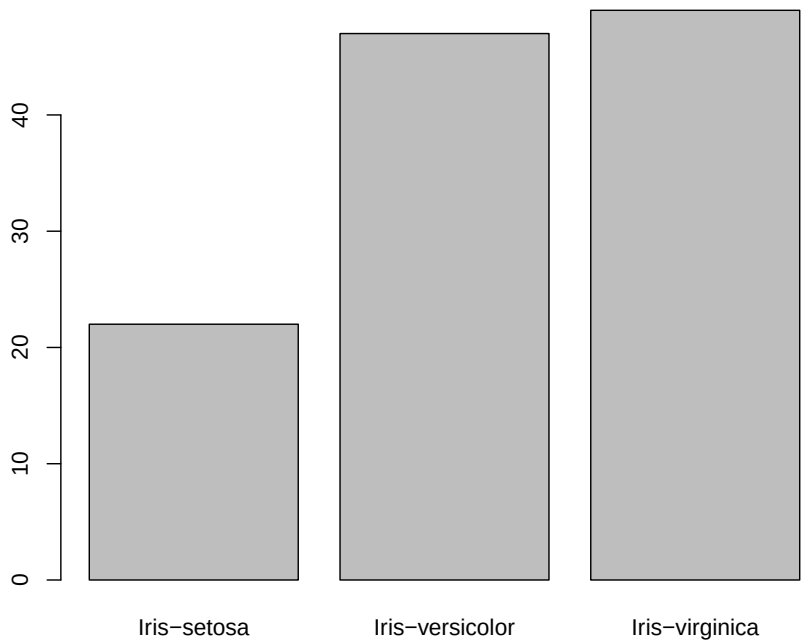
# Boîtes à moustaches (pour les 4 variables quantitatives des données Iris)



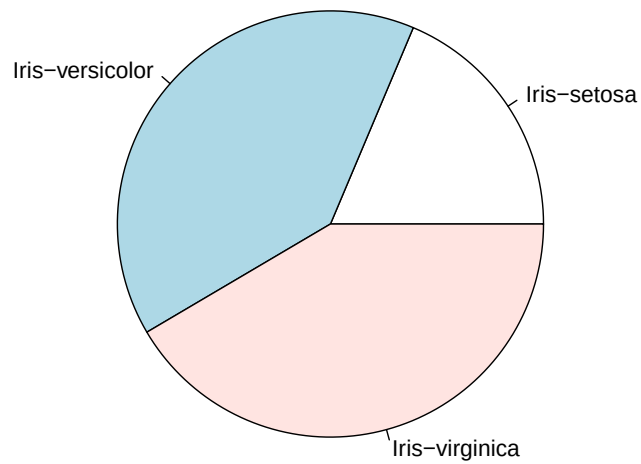
# Description d'une variable qualitative

Variable « espèce » du jeu de données « Iris » pour les longueurs de sépale supérieures à 5

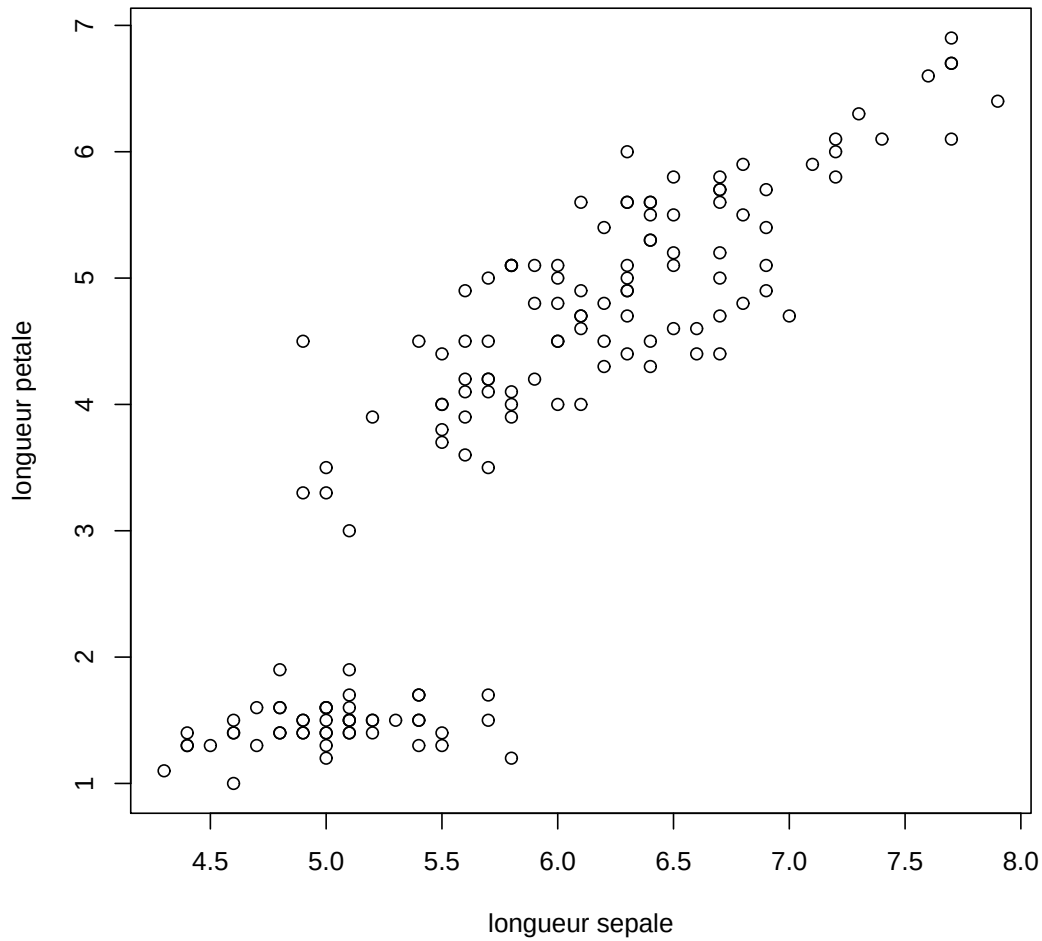
## Diagramme en barre




## Camembert

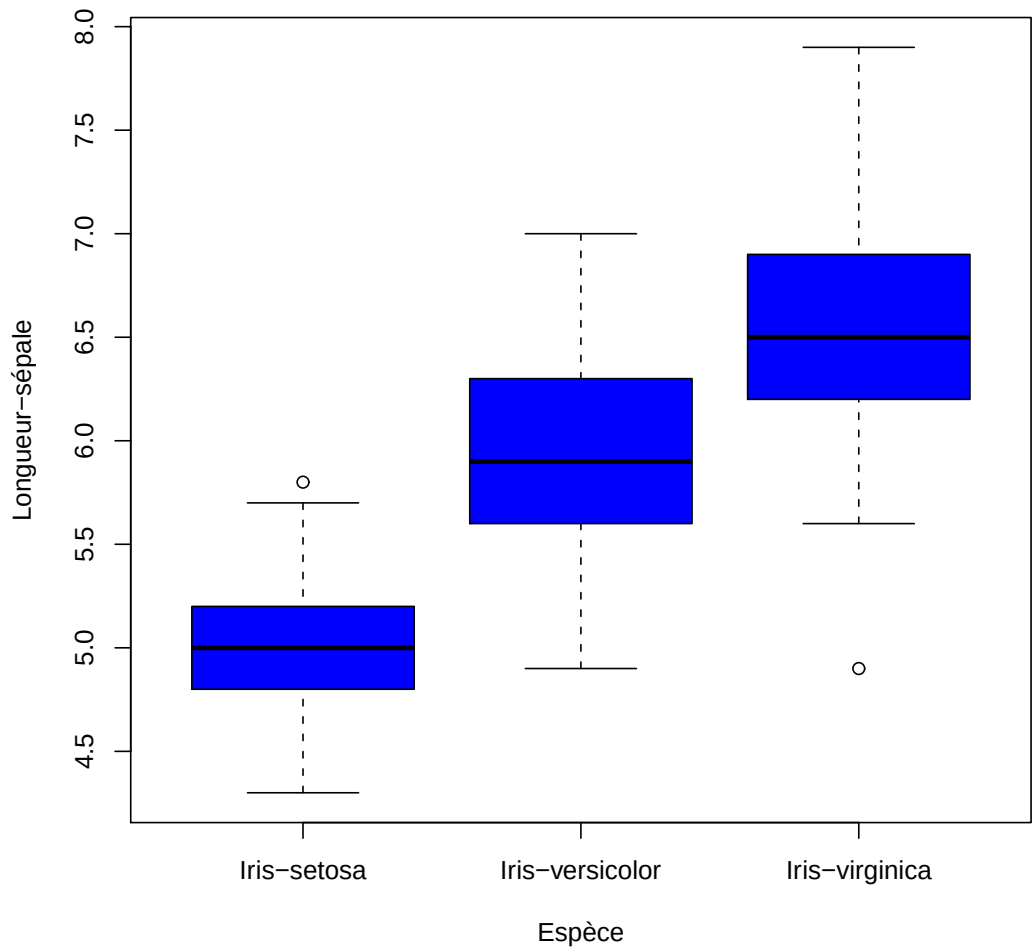


# Description bidimensionnelle de deux variables quantitatives : nuage de points



# Description bidimensionnelle : variable quanti.Vs. variable quali.

## boxplots parallèles



# Description bidimensionnelle : deux variables qualitatives

données

| $\mathbf{x}^j$ | $\mathbf{x}^{j'}$ |
|---|---|
| a | c |
| b | e |
| a | d |
| b | e |
| b | c |
| b | d |
| a | e |
| a | d |
| b | e |

tableau de contingence

| | c | d | e |
|---|---|---|---|
| a | 1 | 2 | 1 |
| b | 1 | 1 | 3 |

diagrammes en barre
lignes du tableau de contingence

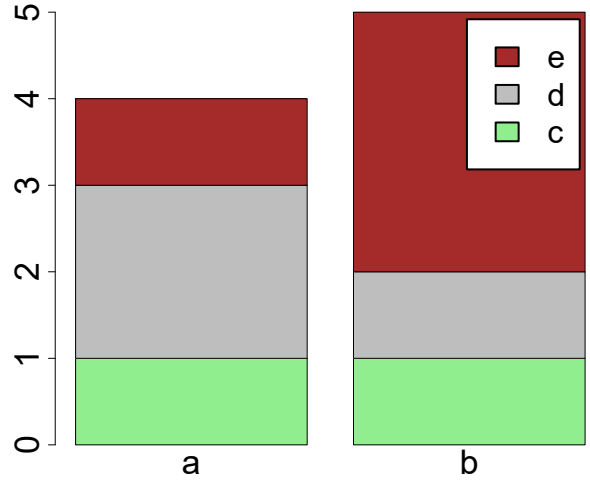
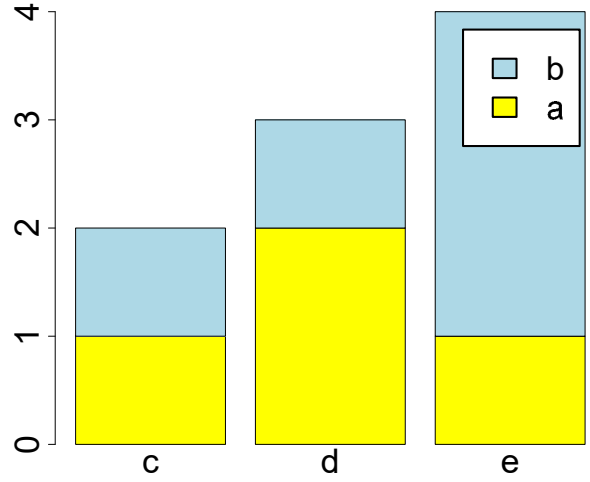



diagramme en barre
colonnes du tableau de contingence



# Degré de liaison entre deux variables quantitatives : covariance et corrélation

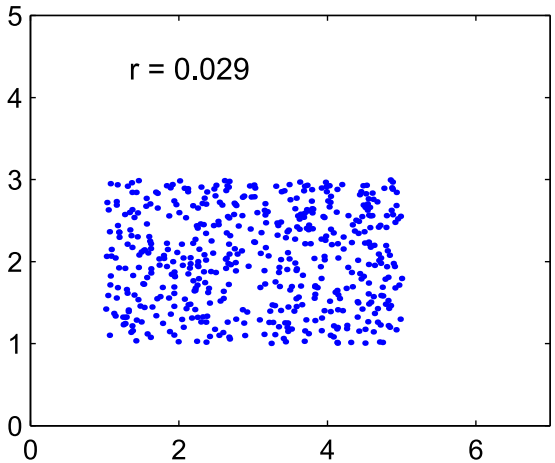
## Covariance empirique

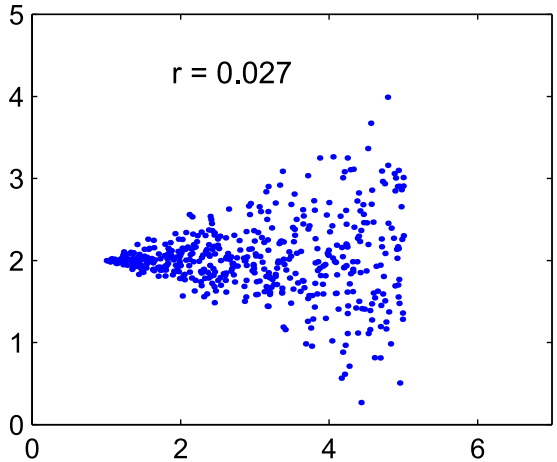
$$s_{jj'} = \frac{\sum_{i=1}^{n}(x_i^j - \overline{x}_j)(x_i^{j'} - \overline{x}_{j'})}{n}$$

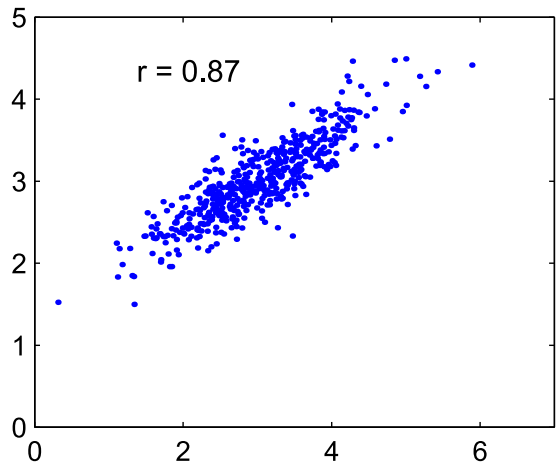
## Coefficient de corrélation empirique

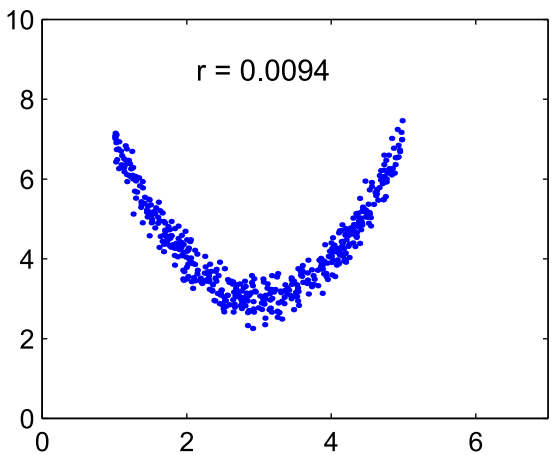
$$r_{jj'} = \frac{s_{jj'}}{s_j \, s_{j'}}$$

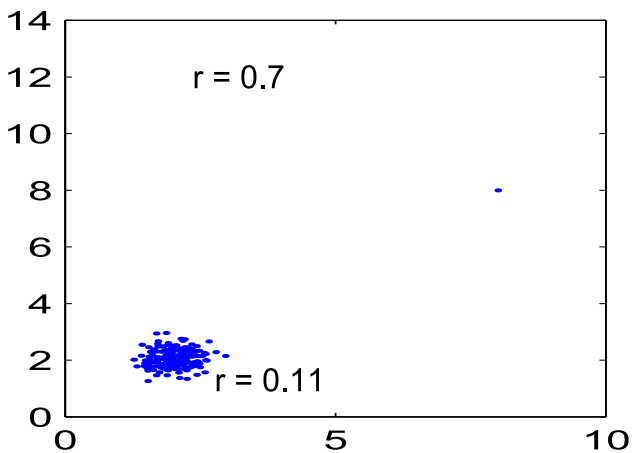
# Exemples de coefficients de corrélation

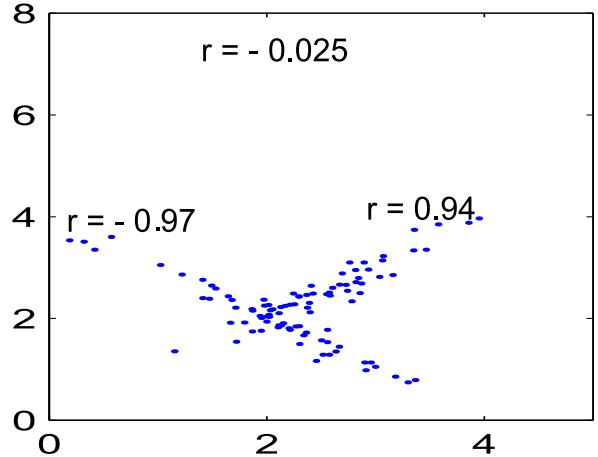


# Degré de liaison entre une variable quanti. $\mathbf{x}^j$ et une variable quali. $\mathbf{x}^{j'}$

On suppose que la variable qualitative $\mathbf{x}^{j'}$ prend $L$ modalités

On note $n_\ell$ le nombre d'occurrences de chaque modalité $(1 \leq \ell \leq L)$

Pour la variable $\mathbf{x}^j$, on note $\overline{x}_j(\ell)$ la moyenne des valeurs correspondant à la modalité $\ell$ et $\overline{x}_j$ la moyenne globale et on définit les quantités suivantes :

## Indicateur numérique : rapport variance inter-classe et variance totale

$$\rho = \frac{\sum_{\ell=1}^{L} n_\ell (\overline{x}_j(\ell) - \overline{x}_j)^2}{\sum_{i=1}^{n} (x_i^j - \overline{x}_j)^2}$$

Plus le rapport est élevé, plus la dépendance est forte

# Indicateur numérique de liaison entre deux variables qualitatives $\mathbf{x}^j$ et $\mathbf{x}^{j'}$

On suppose que la variable qualitative $\mathbf{x}^j$ prend $K$ modalités $(1 \leq k \leq K)$

On suppose que la variable qualitative $\mathbf{x}^{j'}$ prend $L$ modalités $(1 \leq \ell \leq L)$

On note $n_{k\ell}$ le nombre de co-occurrences des modalités $k$ et $\ell$

On note $n_{k\bullet} = \sum_\ell n_{k\ell}$ et $n_{\bullet\ell} = \sum_k n_{k\ell}$

## Indicateur numérique : mesure du Khi-Deux

$$D^2 = \sum_{k,\ell} \frac{\left(n_{k\ell} - \frac{n_{k\bullet} n_{\bullet\ell}}{n}\right)^2}{\frac{n_{k\bullet} n_{\bullet\ell}}{n}}$$

- L'effectif dit théorique $n^*_{k\ell} = \frac{n_{k\bullet} n_{\bullet\ell}}{n}$ traduit la situation d'indépendance entre les deux variables
- Plus $D^2$ est élevé, plus on s'éloigne de la situation d'indépendance

# Description multidimensionnelle (plus de deux variables)

## Matrice de covariance

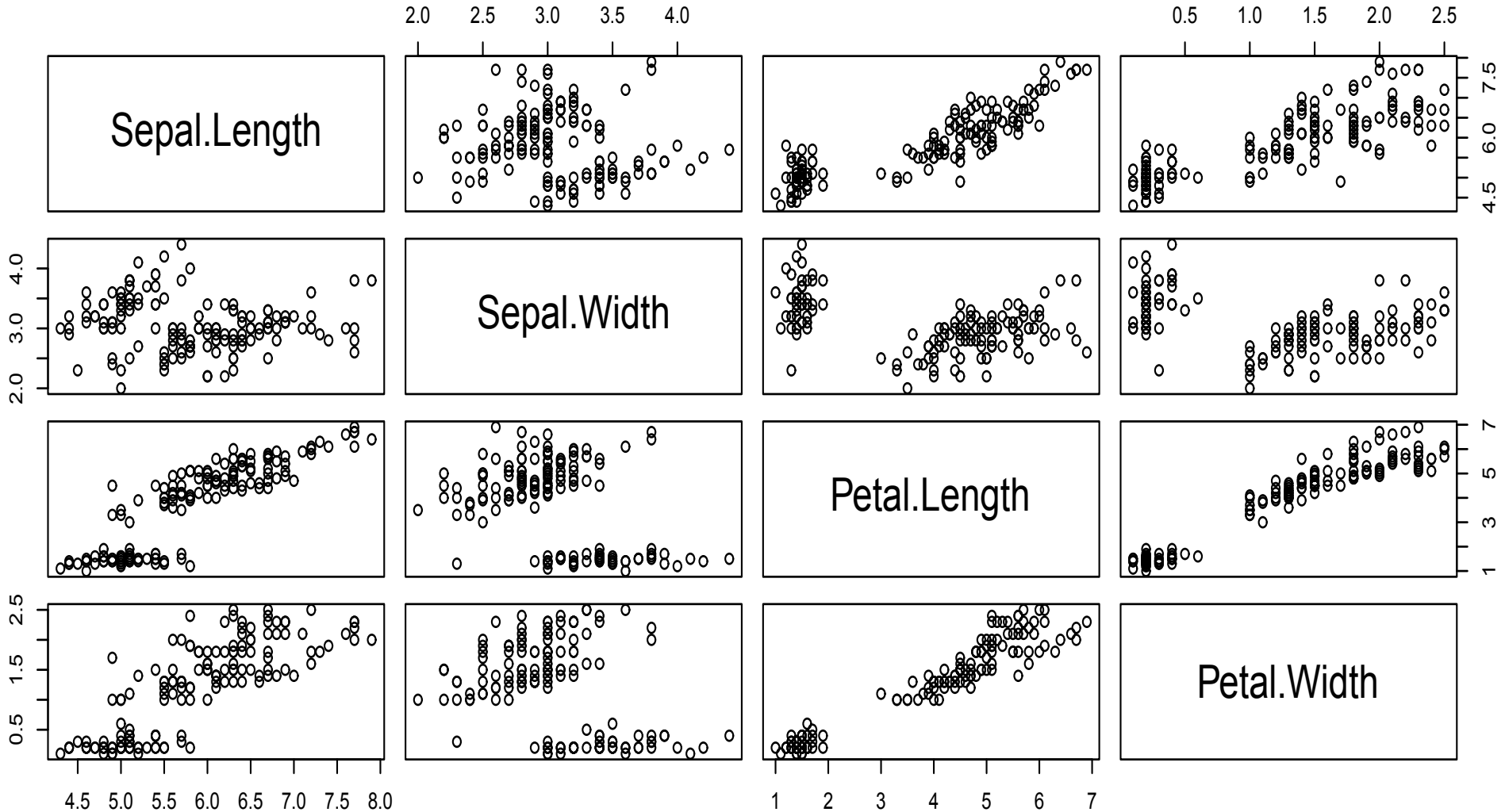
| | sep_length | sep_width | pet_length | pet_width |
|---|---|---|---|---|
| sep_length | 0.68569349 | -0.03926847 | 1.2736823 | 0.5169038 |
| sep_width | -0.03926847 | 0.18800403 | -0.3217128 | -0.1179812 |
| pet_length | 1.27368231 | -0.32171276 | 3.1131794 | 1.2963874 |
| pet_width | 0.51690379 | -0.11798121 | 1.2963874 | 0.5824143 |

## Matrice de corrélation

| | sep_length | sep_width | pet_length | pet_width |
|---|---|---|---|---|
| sep_length | 1.0000000 | -0.1093693 | 0.8717542 | 0.8179536 |
| sep_width | -0.1093693 | 1.0000000 | -0.4205161 | -0.3565441 |
| pet_length | 0.8717542 | -0.4205161 | 1.0000000 | 0.9627571 |
| pet_width | 0.8179536 | -0.3565441 | 0.9627571 | 1.0000000 |

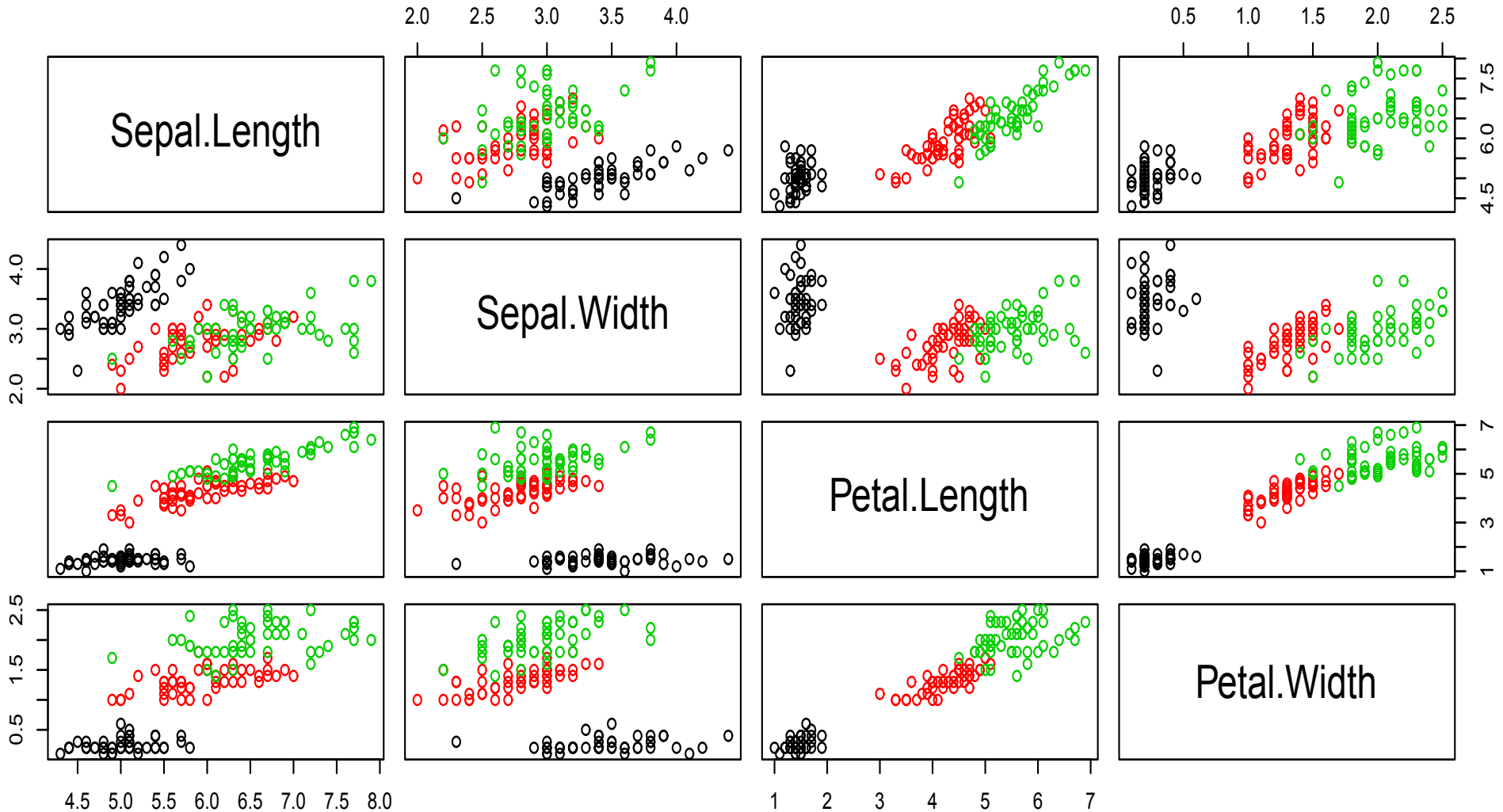
# Description multidimensionnelle

## Tableau des nuages de points



# Description multidimensionnelle

## Tableau des nuages de points



# Distance et Norme

## Distance

Une distance $d$ sur un espace métrique $E$ est une application de $E \times E \to \mathbb{R}^+$ vérifiant :

(i) $\forall \mathbf{x}, \mathbf{y} \in E, \quad d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$ (séparation)

(ii) $\forall \mathbf{x}, \mathbf{y} \in E, \quad d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symétrie)

(iii) $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in E, \quad d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ (inégalité triangulaire)

## Norme

Une norme $\| \cdot \|$ sur un $\mathbb{R}-$espace vectoriel $E$ est une application de $E \to \mathbb{R}^+$ vérifiant :

(i) $\forall \mathbf{x} \in E, \quad \| \mathbf{x} \| = 0 \Leftrightarrow \mathbf{x} = 0$

(ii) $\forall \mathbf{x} \in E, \lambda \in \mathbb{R}, \quad \| \lambda \mathbf{x} \| = |\lambda| \, \| \mathbf{x} \|$

(iii) $\forall \mathbf{x}, \mathbf{y} \in E, \quad \| \mathbf{x} + \mathbf{y} \| \leq \| \mathbf{x} \| + \| \mathbf{y} \|$

# Equivalence Distance et Norme

- A une norme $\| \cdot \|$, on peut associer la distance définie par :

$$d(\mathbf{x}, \mathbf{y}) = \| \mathbf{x} - \mathbf{y} \|$$

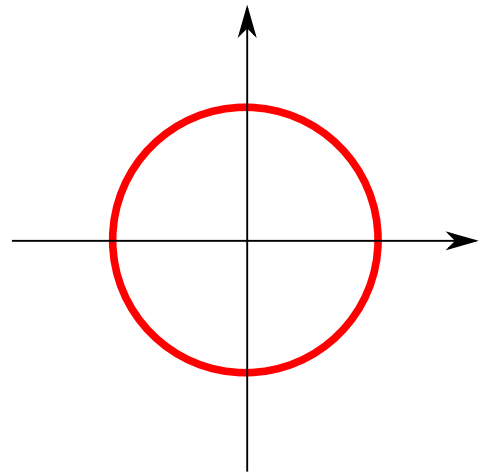
- A une distance $d$, on peut associer sous certaines conditions la norme $\| \cdot \|$ définie par

$$\| \mathbf{x} \| = d(0, \mathbf{x})$$

# Distances usuelles pour données quantitatives

## Distance euclidienne ou distance $L_2$

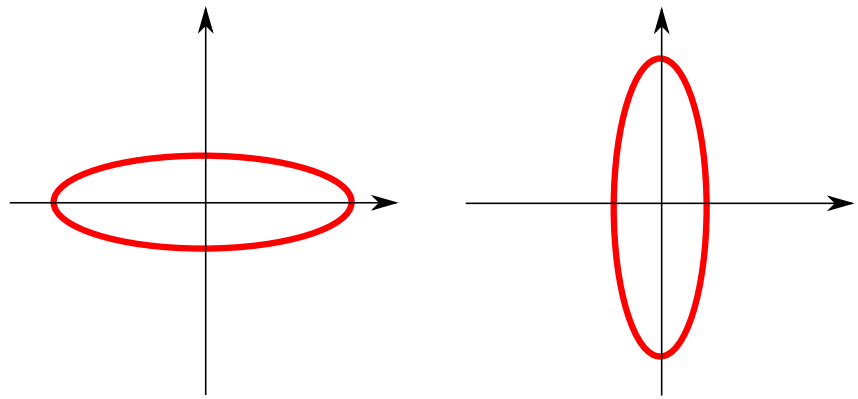
$$\begin{aligned} d^2(\mathbf{x},\mathbf{y}) &= \sum_{j=1}^{p}(x^j - y^j)^2 \\ &= (\mathbf{x}-\mathbf{y})'\mathbf{I}(\mathbf{x}-\mathbf{y}) \end{aligned}$$



avec $\mathbf{I}$ = matrice identité

## Distance euclidienne pondérée

$$\begin{aligned} d^2(\mathbf{x},\mathbf{y}) &= \sum_{j=1}^{p} w_j(x^j - y^j)^2 \\ &= (\mathbf{x}-\mathbf{y})'\mathbf{D}(\mathbf{x}-\mathbf{y}) \end{aligned}$$



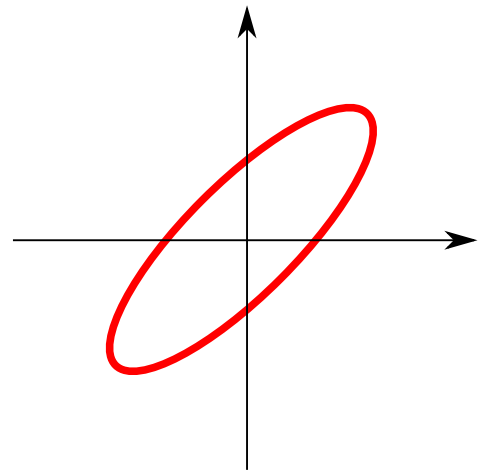
avec $\mathbf{D} = \text{diag}(w_1, \ldots, w_p)$ et $w_j > 0$

# Distances usuelles pour données quantitatives

## Distance de Mahalanobis

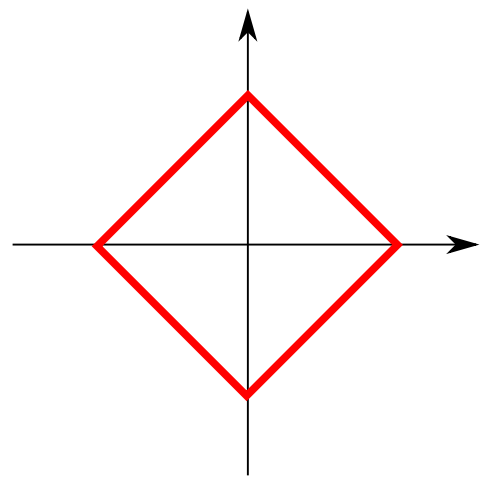
$$d^2(\mathbf{x}, \mathbf{y}) \quad = \quad (\mathbf{x} - \mathbf{y})' \mathbf{M} (\mathbf{x} - \mathbf{y})$$

avec $\mathbf{M}$ matrice symétrique définie positive



## Distance de Manhattan ou distance $L_1$

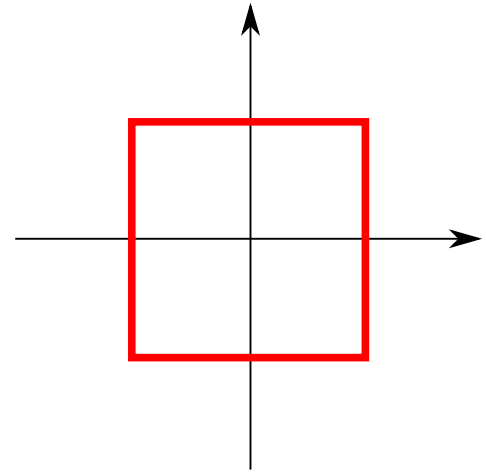
$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} |x^j - y^j|$$



# Distances usuelles pour données quantitatives

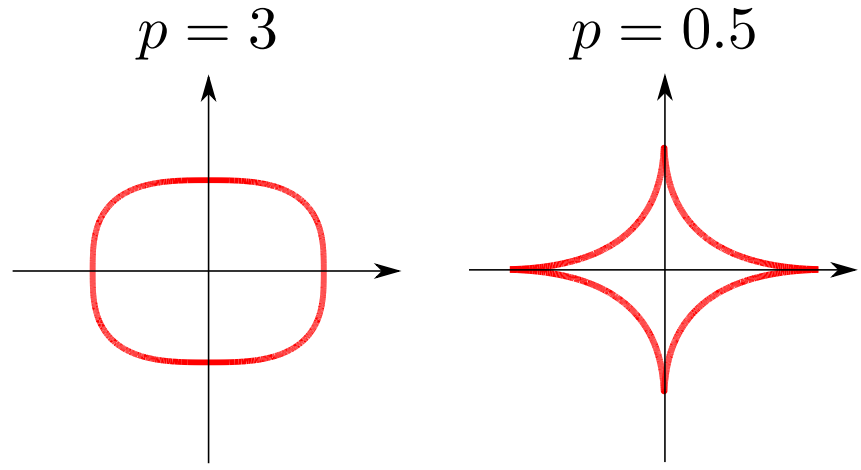
## Distance de Tchebychev ou distance $L_\infty$

$$d(\mathbf{x}, \mathbf{y}) = \max_{1 \leq j \leq p} |x^j - y^j|$$



## Distance de Minkowski ou distance $L_p$ (généralisant $L_1, L_2$)

$$d(\mathbf{x}, \mathbf{y}) = \left( \sum_{j=1}^{p} |x^j - y^j|^p \right)^{1/p}$$



# Distances pour données qualitatives

## Distance de Hamming

$$d_H(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} \delta(x^j, y^j)$$

avec

$$\delta(x^j, y^j) = \begin{cases} 1 & \text{si } x^j \neq y^j \\ 0 & \text{sinon} \end{cases}$$

## Exemple

$\mathbf{x}$ = aabbabbbacb

$\mathbf{y}$ = caccbbaaabc

$$d(\mathbf{x}, \mathbf{y}) = ?$$

# Dissimilarité et similarité

## Dissimilarité

Une mesure de dissimilarité est une application de $E \times E \to \mathbb{R}^+$ vérifiant :

(i) $\forall \mathbf{x}, \mathbf{y} \in E, \quad d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symétrie)

(ii) $\mathbf{x} = \mathbf{y} \quad \Longrightarrow \quad d(\mathbf{x}, \mathbf{y}) = 0$

## Similarité

Une mesure de similarité est une application de $E \times E \to \mathbb{R}^+$ vérifiant :

(i) $\forall \mathbf{x}, \mathbf{y} \in E, \quad d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symétrie)

(ii) $s(\mathbf{x}, \mathbf{y}) = s_{max} \quad \Longleftrightarrow \quad \mathbf{x} = \mathbf{y}$

(iii) $\forall \mathbf{x}, \mathbf{y} \in E, \quad s(\mathbf{x}, \mathbf{y}) \leq s_{max}$

# Equivalence entre dissimilarité et similarité

- Si $s$ est une mesure de similarité, alors l'application $d$ définie par

$$d(\mathbf{x}, \mathbf{y}) = s_{max} - s(\mathbf{x}, \mathbf{y})$$

  est une mesure de dissimilarité

- Si $d$ est une mesure de dissimilarité, alors l'application $s$ définie par

$$s(\mathbf{x}, \mathbf{y}) = d_{max} - d(\mathbf{x}, \mathbf{y})$$

  est une mesure de similarité

# Ultramétrique

Une ultramétrique $\delta$ sur un ensemble $E$ est une fonction de $E \times E \to \mathbb{R}^+$ vérifiant :

(i) $\forall \mathbf{x}, \mathbf{y} \in E, \quad \delta(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$

(ii) $\forall \mathbf{x}, \mathbf{y} \in E, \quad \delta(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{y}, \mathbf{x})$

(iii) $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in E, \quad \delta(\mathbf{x}, \mathbf{z}) \leq \max(\delta(\mathbf{x}, \mathbf{y}), \delta(\mathbf{y}, \mathbf{z}))$
(inégalité ultramétrique)

## Propriétés de l'ultramétrique

- L'inégalité (iii) entraine l'inégalité triangulaire
- On peut donc vérifier qu'une ultramétrique est une distance

L'ultramétrique joue un rôle fondamental en classification (on verra dans la suite qu'il y a un lien direct entre ultramétrique et hiérarchie)

# Exemples

$$\mathbf{X} = \begin{pmatrix} 1 & 4 \\ 1 & 2 \\ 4 & 3 \\ 5 & 4 \\ 5 & 1 \end{pmatrix}$$

Matrice de dissimilarités (distance euclidienne)

| | $\mathbf{x}_1$ | $\mathbf{x}_2$ | $\mathbf{x}_3$ | $\mathbf{x}_4$ | $\mathbf{x}_5$ |
|---|---|---|---|---|---|
| $\mathbf{x}_1$ | 0 | 2 | 3.16 | 4 | 5 |
| $\mathbf{x}_2$ | 2 | 0 | 3.16 | 4.47 | 4.12 |
| $\mathbf{x}_3$ | 3.16 | 3.16 | 0 | 1.41 | 2.24 |
| $\mathbf{x}_4$ | 4 | 4.47 | 1.41 | 0 | 3 |
| $\mathbf{x}_5$ | 5 | 4.12 | 2.24 | 3 | 0 |

Matrice de similarités

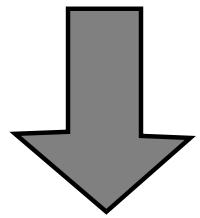
$d_{max} = 5$ et $s(\mathbf{x}_i, \mathbf{x}_{i'}) = d_{max} - d(\mathbf{x}, \mathbf{x}_{i'})$

| | $\mathbf{x}_1$ | $\mathbf{x}_2$ | $\mathbf{x}_3$ | $\mathbf{x}_4$ | $\mathbf{x}_5$ |
|---|---|---|---|---|---|
| $\mathbf{x}_1$ | 5 | 3 | 1.84 | 1 | 0 |
| $\mathbf{x}_2$ | 3 | 5 | 1.84 | 0.53 | 0.88 |
| $\mathbf{x}_3$ | 1.84 | 1.84 | 5 | 3.59 | 2.76 |
| $\mathbf{x}_4$ | 1 | 0.53 | 3.59 | 5 | 2 |
| $\mathbf{x}_5$ | 0 | 0.88 | 2.76 | 2 | 5 |

# Réduction de la dimensionalité

$$\mathbf{x}_i \in \mathbb{R}^p$$

$$(4.0,\ -1.6,\ 3.5,\ -2.8,\ 4.7,\ -1.5,\ 3.5,\ 2.5,\ 6.3,\ -2.6,\ -3.3)$$



$$\mathbf{y}_i \in \mathbb{R}^q$$

$$(0.6,\ -1.4,\ 2.2)$$

## Objectifs

- Compression de données
- Visualisation en 2D ou 3D (au dela de 3 variables il devient compliqué de représenter le nuage de points)
- Extraction de caractéristiques pertinentes

# Quelques méthodes de réduction de la dimensionnalité

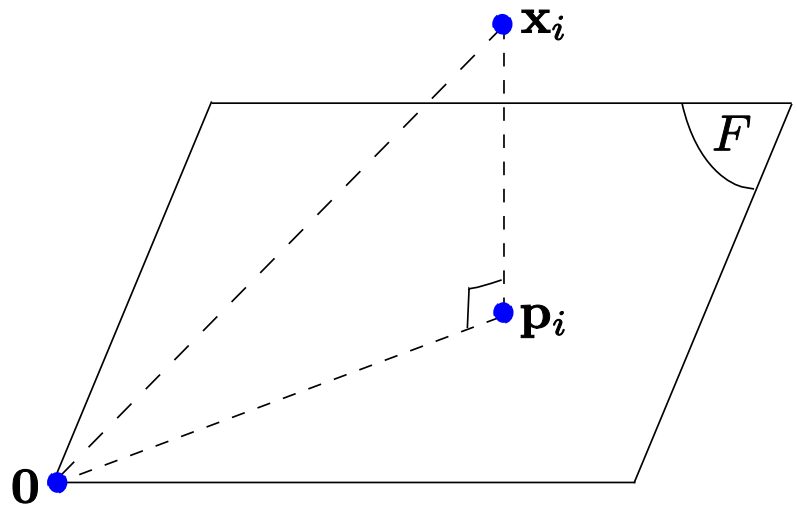
- Méthodes linéaires
  - Analyse en composantes principales (ACP ou PCA)
  - Analyse en facteurs communs ou *factor analysis*
  - Analyse en composantes indépendantes (ICA)
- Méthodes non linéaires
  - Cartes auto organisatrices (réseaux de neurones autoassociateurs)
  - Multidimensional scaling (MDS)
  - Locally linear embedding (LLE)
  - Auto-encodeurs : réseau de neurones à architecture potentiellement profonde

# Analyse en Composantes Principales

- L'ACP trouve un sous-espace de dimension $q < p$ qui passe au plus proche des données (quantitatives)
- Les nouvelles données $\mathbf{y}_i$ sont les coordonnées de la projection de $\mathbf{x}_i$ dans le sous-espace
- l'ACP permet de passer du tableau de $n$ individus décrits par $p$ variables à un tableau de $n$ individus décrits par $q$ variables avec $q < p$
- Ces nouvelles variables sont appelées composantes principales
- Différentes formulations :
  - minimisation de l'erreur de reconstruction (Pearson, 1901)
  - maximisation de la variance (Hotelling, 1933)
  - modèle probabiliste à variables latentes (Tipping, Bishop, 1999)

# Problème posé par l'ACP

Trouver un sous espace $F$ de $\mathbb{R}^p$ tel que l'écart entre les individus et leur projection soit minimal




Cela se traduit par la minimisation par rapport aux $\mathbf{p}_i$ du critère

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{p}_i\|^2$$

# Procédure

- On commence par trouver un premier axe par minimisation de l'écart entre les individus et leur projection sur cet axe
- On cherche un second axe qui, parmi tous les axes $\perp$ au premier, minimise l'inertie relativement à cet axe
- Ainsi de suite...

On peut montrer que l'ensemble des axes principaux est obtenu par diagonalisation de la matrice de covariance $\mathbf{S}$ ou la matrice de corrélation $\mathbf{R}$ si les variables ont été centrées et réduites.

$$\mathbf{S} = \begin{bmatrix} s_1^2 & \dots & s_{1j} & \dots & s_{1p} \\ \vdots & \ddots & & & \vdots \\ s_{j1} & & s_j^2 & & s_{jp} \\ \vdots & & & \ddots & \vdots \\ s_{p1} & \dots & s_{pj} & \dots & s_p^2 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} 1 & \dots & r_{1j} & \dots & r_{1p} \\ \vdots & \ddots & & & \vdots \\ r_{j1} & & 1 & & r_{jp} \\ \vdots & & & \ddots & \vdots \\ r_{p1} & \dots & r_{pj} & \dots & 1 \end{bmatrix}$$

# Détermination des axes et des composantes principaux

## Axes factoriels ou axes principaux

Obtenus par diagonalisation de la matrice $\mathbf{S}$ ou $\mathbf{R}$

- Soient $\lambda_1, \ldots, \lambda_p$ les valeurs propres ordonnées par ordre décroissant et $\mathbf{u}_1, \ldots, \mathbf{u}_p$ les vecteurs propres correspondants
- Le premier axe factoriel est celui dont la direction est $\mathbf{u}_1$, le second axe factoriel est celui de direction $\mathbf{u}_2$ et ainsi de suite

## Composantes principales

Les composantes principales sont les coordonnées des points $\mathbf{x}_i$ sur les différents axes factoriels :

1. La 1re composante principale $\mathbf{c}^1 = (c_i^1, \ldots, c_n^1)$ contient les coordonnées des projections sur l'axe factoriel $\mathbf{u}_1$
2. La 2e composante principale $\mathbf{c}^2 = (c_i^2, \ldots, c_n^2)$ contient les coordonnées des projections sur l'axe $\mathbf{u}_2$

# Détermination des composantes principales (2/2)

## Remarque

- les composantes principales (les nouvelles variables) sont des combinaisons linéaires des variables initiales
- elles sont non corrélées deux à deux puisque les axes associés à ces variables sont orthogonaux
- on peut vérifier que la variance d'une composante principale $\mathbf{c}^{\alpha}$ est égale à la valeur propre $\lambda_{\alpha}$

# Choix du nombre d'axes principaux à retenir

On calcule un critère de qualité de représentation pour les différents sous espaces :

- pour le sous-espace engendré par l'axe principal $\mathbf{u}_1$ :

$$\text{qualité}(\mathbf{u}_1) = \frac{\lambda_1}{\sum_{i=1}^{p} \lambda_\alpha}$$

- pour le sous espace engendré par les axes $\mathbf{u}_1$ et $\mathbf{u}_2$ :

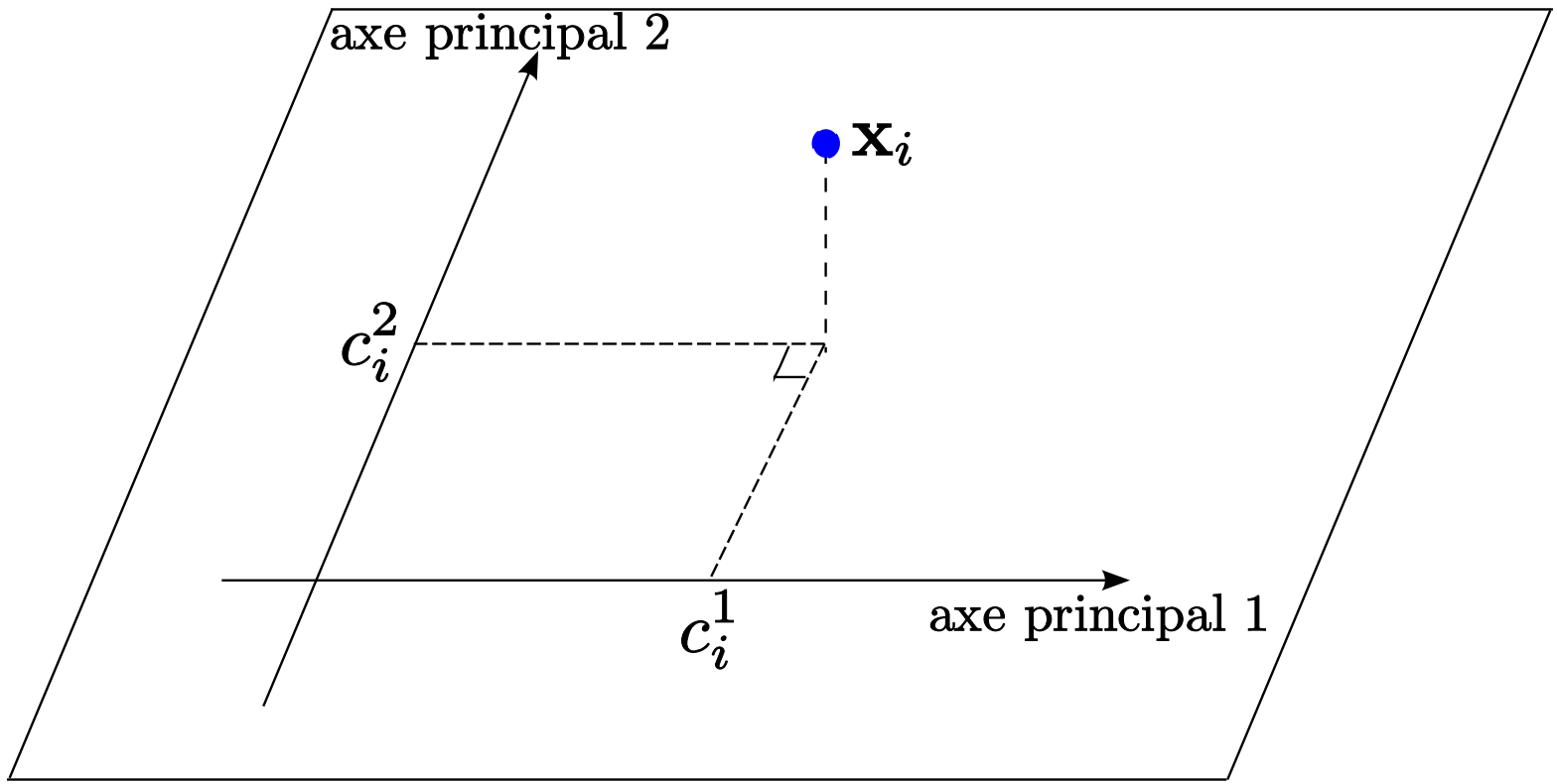
$$\text{qualité}(\mathbf{u}_1, \mathbf{u}_2) = \frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{p} \lambda_\alpha}$$

- et ainsi de suite...

On choisit le nombre d'axes $k$ à partir duquel le critère est supérieur à un certain seuil : ex. 90% de la variance

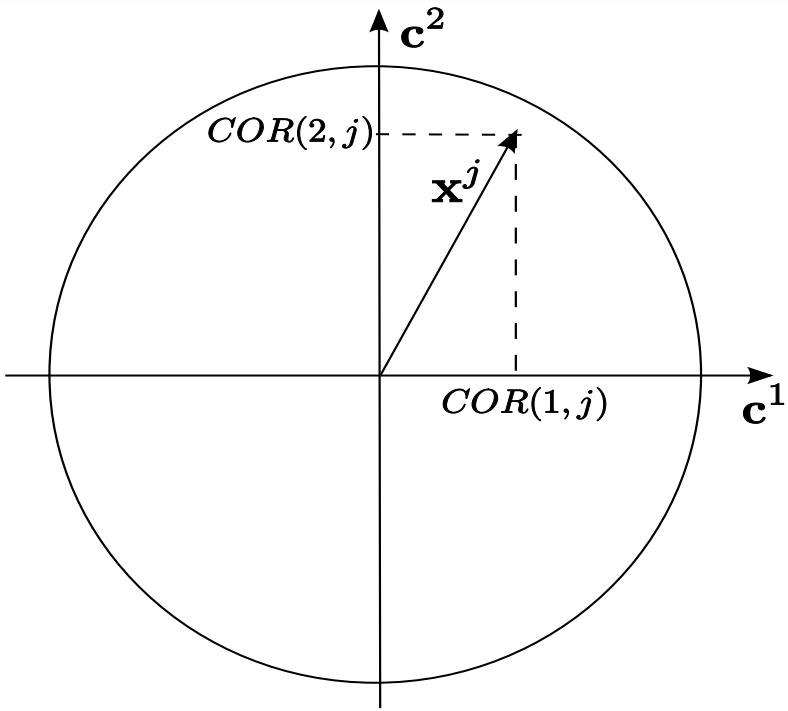
# Représentation des individus

La projection du nuage de points initial dans le premier plan factoriel est obtenue grâce aux composantes principales



# Représentation des variables : cercle des corrélations

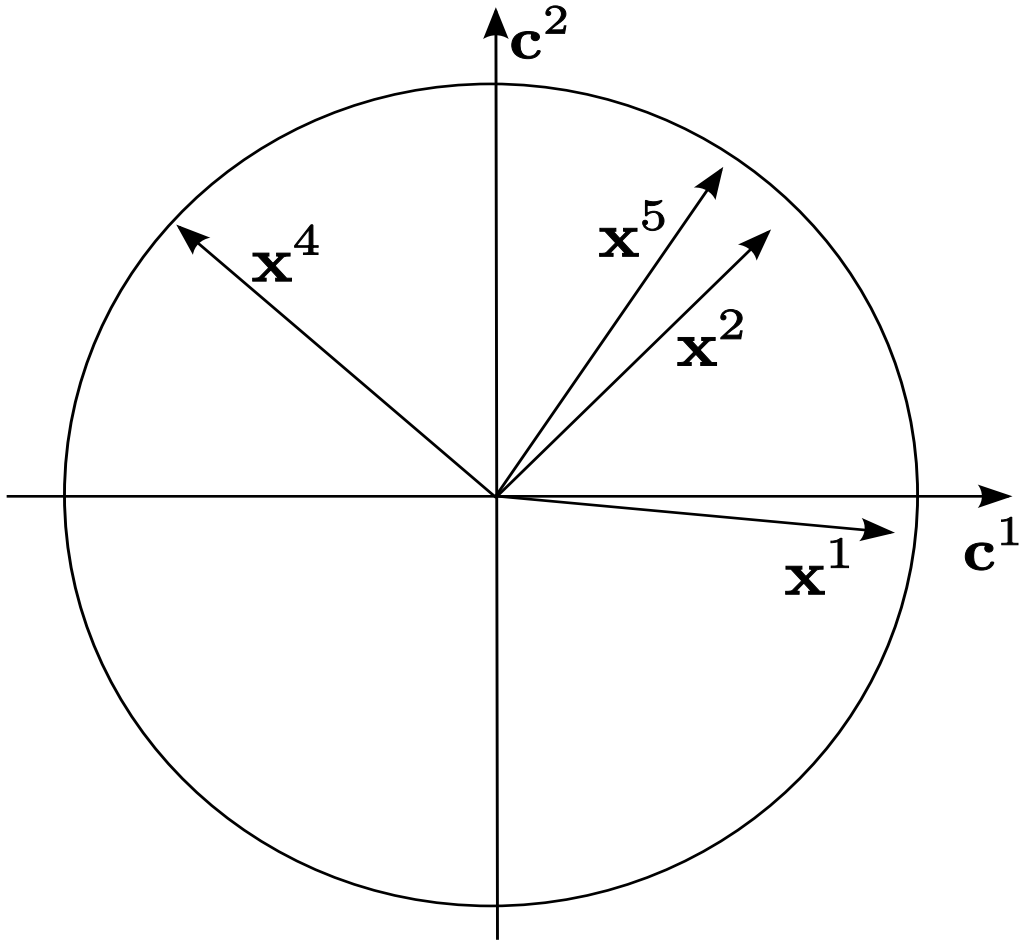
- Permet de voir les liens entre les variables initiales et aussi les liens entre composantes principales et variables initiales
- Facilite l'interprétation des axes principaux




$$COR(\alpha, j) = cor\left(\mathbf{x}^j, \mathbf{c}^\alpha\right)$$

# Exemple de représentation des variables

Quels liens existent entre les variables représentées ci-dessous ?
Existe t-il des liens entre les variables initiales et les composantes principales ?



# Exemple

Données : 40 étudiants décrits par 7 variables (tests QI, taille, poids, taille du cerveau déterminée par IRM)

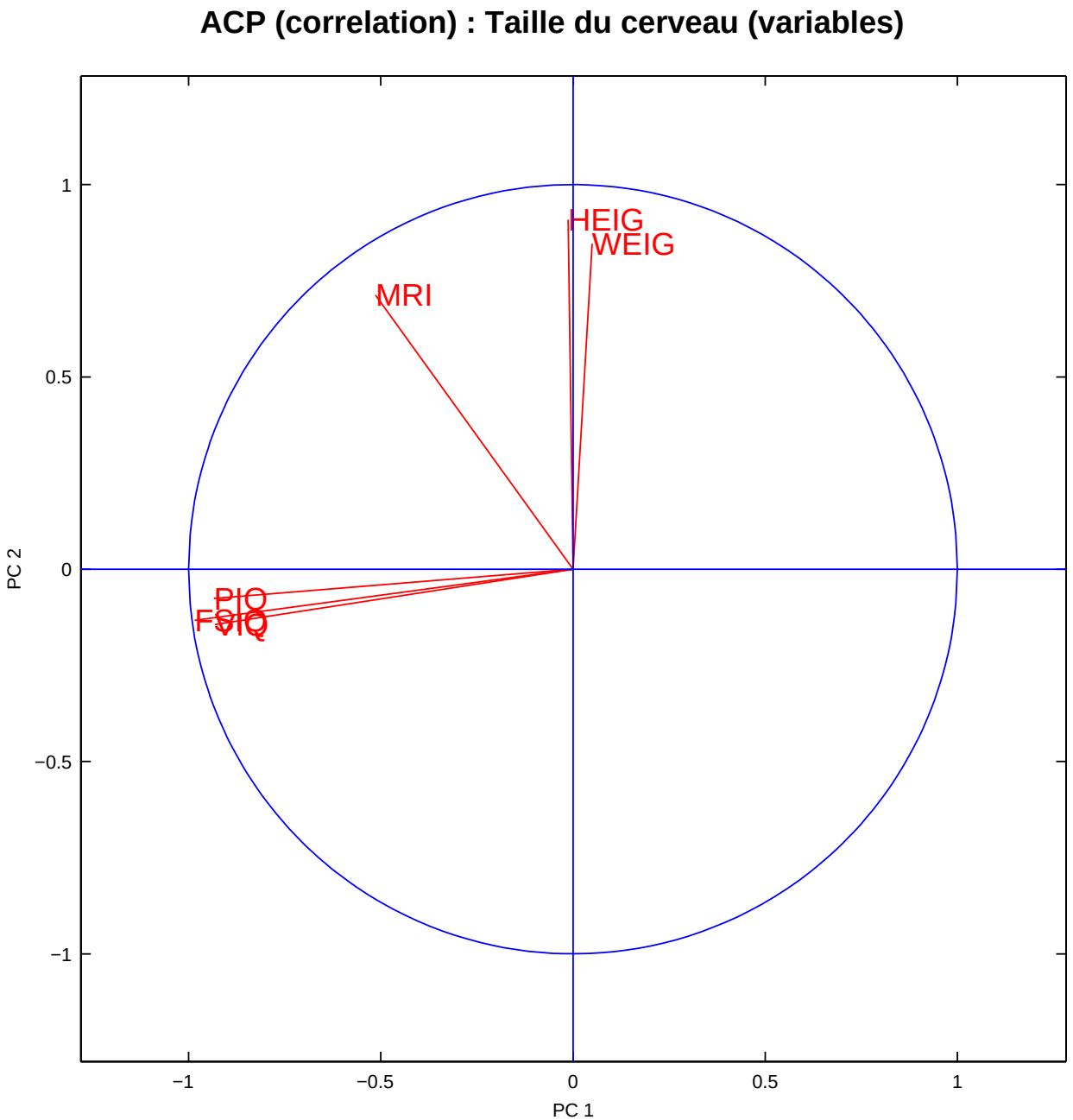
## Corrélations

| | FSIQ | VIQ | PIQ | WEIG | HEIG | MRI |
|---|---|---|---|---|---|---|
| FSIQ | 1.00 | 0.95 | 0.93 | −0.13 | −0.10 | 0.36 |
| VIQ | 0.95 | 1.00 | 0.78 | −0.16 | −0.08 | 0.34 |
| PIQ | 0.93 | 0.78 | 1.00 | −0.05 | −0.09 | 0.39 |
| WEIG | −0.13 | −0.16 | −0.05 | 1.00 | 0.63 | 0.43 |
| HEIG | −0.10 | −0.08 | −0.09 | 0.63 | 1.00 | 0.60 |
| MRI | 0.36 | 0.34 | 0.39 | 0.43 | 0.60 | 1.00 |

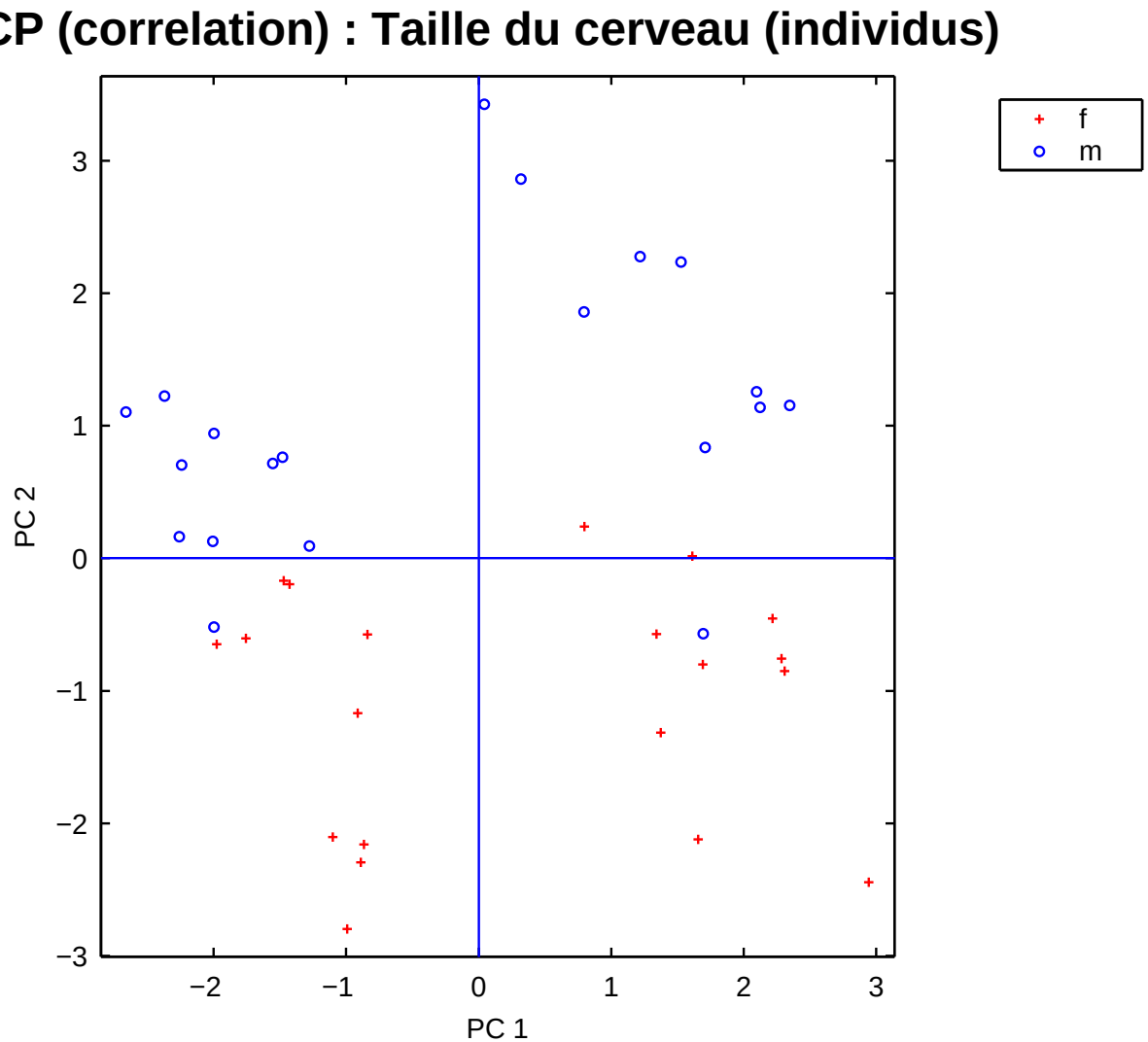
## ACP : valeurs propres

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Variance | 2.97 | 2.09 | 0.453 | 0.287 | 0.189 | 0.0026 |
| Pourc. de variance | 49.57 | 34.90 | 7.549 | 4.790 | 3.146 | 0.0432 |
| Pourcentage cumulé | 49.57 | 84.47 | 92.021 | 96.810 | 99.957 | 100.0000 |

# Exemple


ACP (correlation) : Taille du cerveau (variables)
HEIG
WEIG
MRI
PIO
FSIQ
VIQ
PC 1
PC 2
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1

# Exemple


ACP (correlation) : Taille du cerveau (individus)
f
m
PC 2
PC 1
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3