

Statistique

Benjamin Bobbia

ISAE



Point lexicque

La statistique est une discipline qui concerne les affaires de l'État. Il convient de distinguer :

- **les statistiques** : valeurs des informations recueillies,
- **la statistique** : ensemble des techniques utilisées pour l'étude méthodique des faits sociaux.

Dans un cadre statistique, il est commun d'appeler :

- **un individu** un élément individuel considéré dans l'étude,
- **une population** l'ensemble des individus considérés,
- **un échantillon** une partie de la population étudiée.

Ces définitions se généralisent aux méthodes utilisées pour l'étude d'un ensemble d'observations, appelé **jeu de données**, que les mathématiques permettent de placer dans un cadre formel. Il s'agit de l'**objet de ce cours**.

Statistique mathématique

Domaine des mathématiques dédié à l'étude d'un phénomène par la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation.

Ces définitions se généralisent aux méthodes utilisées pour l'étude d'un ensemble d'observations, appelé **jeu de données**, que les mathématiques permettent de placer dans un cadre formel. Il s'agit de l'**objet de ce cours**.

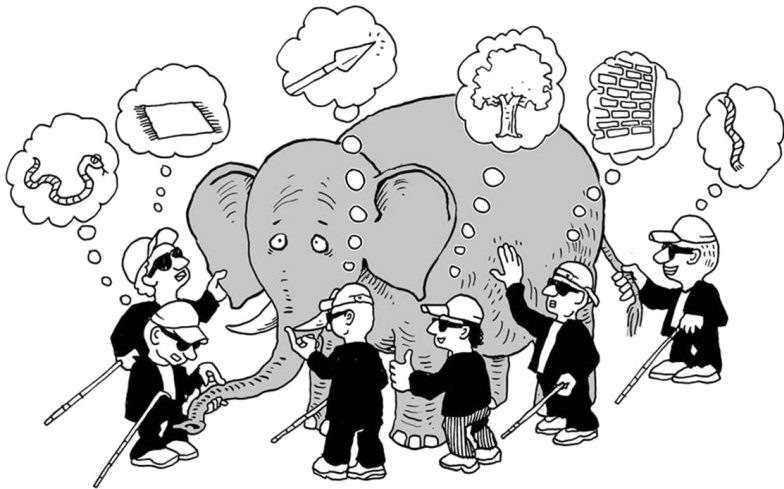
Statistique mathématique

Domaine des mathématiques dédié à l'étude d'un phénomène par la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation.

Cette définition ne donne pas de finalité à cette étude :

- si l'objectif est de **décrire** ou de **résumer** l'information contenue dans les données, nous parlerons de **statistique exploratoire**,
- si l'objectif est de **prédire** ou de **généraliser** à partir des données, elles devront être placées dans un modèle mathématique (généralement probabiliste) et nous parlerons de **statistique inférentielle**.

Statistique exploratoire



1.1 Notions élémentaires

Définitions

Variable et observations

Une **variable** x est une entité à valeurs dans un espace $\mathcal{X}$ qui peut être observée. Dans la suite, un jeu de données issu de n observations de x est appelé un **échantillon** et noté $x_1, \dots, x_n \in \mathcal{X}$.

Une variable x à valeurs dans un espace $\mathcal{X}$ est appelée :

- **quantitative** si $\mathcal{X}$ est un sous-espace de $\mathbb{R}^p$ (température, position spatiale, ...),
- **qualitative** ou **catégorielle** si $\mathcal{X}$ est un ensemble fini (catégorie socio-professionnelle, ...).

Lorsque $\mathcal{X} = \mathbb{R}$, la variable x est dite **réelle**.

Tendance centrale

Considérons une variable réelle x et des observations $x_1, \dots, x_n \in \mathbb{R}$.

Objectif : résumer les données numériques de l'échantillon par un ou plusieurs nombres (mais pas trop).

- **Moyenne** : somme des observations divisée par leur nombre,
- **Médiane** : valeur qui permet de couper les observations en 2 parties de tailles égales,
- **Mode** : valeur la plus fréquente dans l'échantillon,
- **Quartiles** : 3 valeurs qui permettent de couper les observations en 4 parties de tailles égales,
- **Déciles** : 9 valeurs qui permettent de couper les observations en 10 parties de tailles égales,
- **Percentiles** : 99 valeurs qui permettent de couper les observations en 100 parties de tailles égales,
- **Quantiles** : valeurs qui généralisent les indicateurs précédents.

Moyenne et médiane

Soient x et y des variables réelles et des échantillons $x_1, \dots, x_n \in \mathbb{R}$ et $y_1, \dots, y_n \in \mathbb{R}$ associés.

La moyenne est **linéaire** en l'échantillon,

$$\forall a \in \mathbb{R}, \overline{ax} = \frac{1}{n} \sum_{k=1}^n ax_k = a\overline{x} \quad \text{et} \quad \overline{x+y} = \frac{1}{n} \sum_{k=1}^n x_k + y_k = \overline{x} + \overline{y}.$$

La médiane est **homogène** mais **pas additive** (sauf cas particuliers),

$$\forall a \in \mathbb{R}, \text{Med}(ax) = a\text{Med}(x) \quad \text{et} \quad \text{Med}(x+y) \neq \text{Med}(x) + \text{Med}(y).$$

$$\begin{aligned} x_1 &= 0, x_2 = 1, x_3 = 2 : \text{Med}(x) = 1, \\ y_1 &= 2, y_2 = 0, y_3 = 0 : \text{Med}(y) = 0, \\ \text{Med}(x+y) &= 2 \text{ et } \text{Med}(x) + \text{Med}(y) = 1. \end{aligned}$$

Moyenne et médiane

La moyenne $\bar{x}$ minimise les **écarts au carré**,

$$\forall t \in \mathbb{R}, \sum_{k=1}^n (x_k - \bar{x})^2 \leq \sum_{k=1}^n (x_k - t)^2.$$

La médiane $\text{Med}(x)$ minimise les **écarts en valeur absolue**,

$$\forall t \in \mathbb{R}, \sum_{k=1}^n |x_k - \text{Med}(x)| \leq \sum_{k=1}^n |x_k - t|.$$

La médiane est plus **robuste** par rapport aux valeurs extrêmes.

Moyenne et médiane

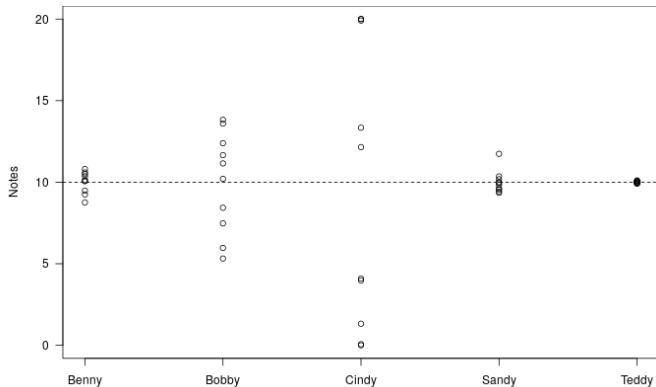
Considérons une entreprise de 12 personnes :

- 8 ouvriers (1201 €)
- 1 chef d'atelier (2000 €)
- 1 directeur technique (5000 €)
- 1 directeur des RH (8000 €)
- 1 directeur général (10000 €)

Salaire moyen : 2884 € – Salaire médian : 1201 €

Autres exemples : gains au Loto, ...

Capacité à résumer



Cette figure représente les notes obtenues par 5 étudiants dans 10 matières différentes. Ces étudiants ont tous une moyenne égale à 10.

Donneriez-vous la même appréciation à chacun d'entre eux ? Est-ce que la moyenne est un « bon » résumé de leurs résultats ?

Dispersion

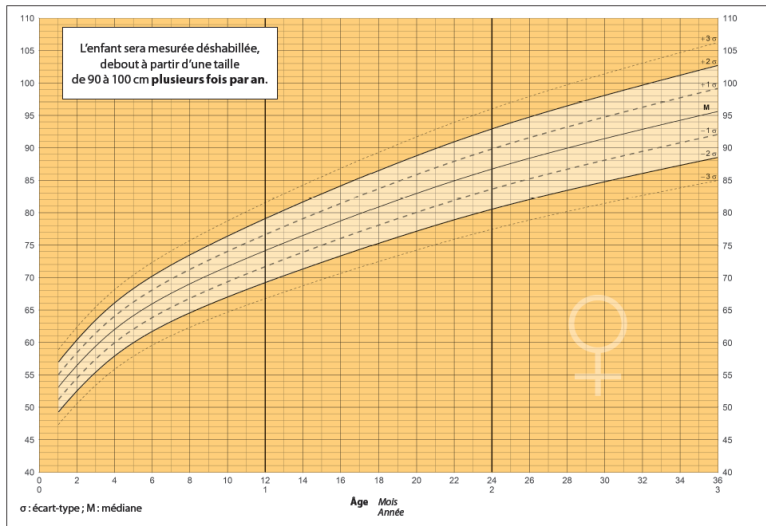
Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.

- **Variance** : $\sigma^2(x) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$.
- **Écart-type** : $\sigma(x) = \sqrt{\sigma^2(x)}$, cet indicateur s'exprime dans les mêmes unités que la moyenne $\bar{x}$. Ainsi, ces quantités peuvent être additionnées pour considérer la dispersion à l'aide d'**intervalles**,

$$\forall a \in \mathbb{R}_+, [\bar{x} - a\sigma(x), \bar{x} + a\sigma(x)].$$

- **Étendue** : $\max_k x_k - \min_k x_k$.
- **Écart interquartile** : distance entre le premier et le troisième quartile.

Exemple d'intervalles



Courbes de croissance AFPA – CRESS/INSERM - CompuGroup Medical, 2018 [enfants nés à plus de 2 500 g et suivis par des médecins sur le territoire métropolitain].

DE 1 MOIS À 3 ANS (CM)

TAILLE DES FILLES

Variance

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.

$$\text{Variance : } \sigma^2(x) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2.$$

La variance est **quadratique**,

$$\forall a \in \mathbb{R}, \sigma^2(ax) = \frac{1}{n} \sum_{k=1}^n (ax_k - a\bar{x})^2 = a^2 \sigma^2(x).$$

La variance est **invariante par translation**,

$$\forall b \in \mathbb{R}, \sigma^2(x + b) = \frac{1}{n} \sum_{k=1}^n ((x_k + b) - (\bar{x} + b))^2 = \sigma^2(x).$$

Variance

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.

$$\text{Variance : } \sigma^2(x) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2.$$

La variance est **nulle** si et seulement si l'échantillon est **constant**,

$$\sigma^2(x) = 0 \iff (x_k - \bar{x})^2 = 0, \forall k \iff x_k = \bar{x}, \forall k.$$

La variance peut être utilisée comme une **mesure de la capacité à résumer** l'information contenue dans le jeu de données réelles par la moyenne. Autrement dit, la variance quantifie la **quantité d'information non expliquée** par la moyenne.

Centrer et réduire

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.

- **Centrer** : retrancher la moyenne.
- **Réduire** : diviser par l'écart-type.

Version centrée-réduite / z-transformation

Les observations $z_1, \dots, z_n$ centrées et réduites sont données par

$$\forall i \in \{1, \dots, n\}, z_i = \frac{x_i - \bar{x}}{\sigma(x)}.$$

Par définition, $\bar{z} = 0$ et $\sigma^2(z) = 1$.

Ces opérations permettent d'exprimer les données dans une échelle neutre en les débarrassant de leurs unités physiques. Une fois centrées et réduites, les observations s'expriment comme un **nombre d'écart-types par rapport à la moyenne**.

Covariance

Soit un couple (x, y) de variables réelles.

Observations : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

La **covariance** entre les observations de x et y est définie par

$$\sigma(x, y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}).$$

Covariance

Soit un couple (x, y) de variables réelles.

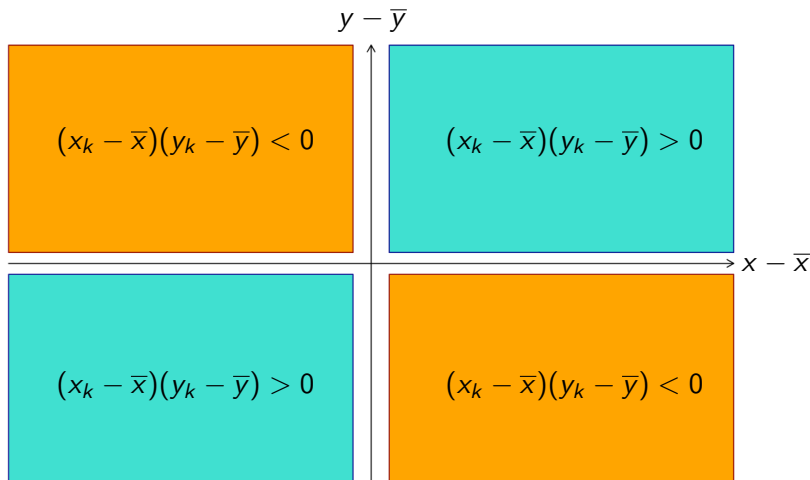
Observations : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

La **covariance** entre les observations de x et y est définie par

$$\sigma(x, y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}).$$

- La covariance $\sigma(x, y)$ est **positive** si les x_k et les y_k ont tendance à être simultanément du même côté de leurs moyennes respectives.
 $\Rightarrow$ **Relation croissante** (e.g. température et nombre de glaces vendues)
- La covariance $\sigma(x, y)$ est **négative** si les x_k et les y_k ont tendance à être simultanément du côté opposé de leurs moyennes respectives.
 $\Rightarrow$ **Relation décroissante** (e.g. température et consommation de chauffage)

Covariance (signe)



Propriétés de la covariance

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Covariance : $\sigma(x, y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$.

La covariance est ...

- **symétrique** : $\sigma(x, y) = \sigma(y, x)$,
- **bilinéaire** :

$$\forall a \in \mathbb{R}, \sigma(ax, y) = a\sigma(x, y) \quad \text{et} \quad \sigma(x + x', y) = \sigma(x, y) + \sigma(x', y),$$

- **définie positive** : $\sigma(x, x) = \sigma^2(x) \geq 0$.

La covariance est donc un ???

Propriétés de la covariance

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Covariance : $\sigma(x, y) = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$.

La covariance est ...

- **symétrique** : $\sigma(x, y) = \sigma(y, x)$,
- **bilinéaire** :

$$\forall a \in \mathbb{R}, \sigma(ax, y) = a\sigma(x, y) \quad \text{et} \quad \sigma(x + x', y) = \sigma(x, y) + \sigma(x', y),$$

- **définie positive** : $\sigma(x, x) = \sigma^2(x) \geq 0$.

La covariance est donc un **produit scalaire** (et l'écart-type est la norme associée).

Cauchy-Schwarz

$$|\sigma(x, y)| \leq \sigma(x)\sigma(y)$$

Coefficient de corrélation linéaire de Pearson

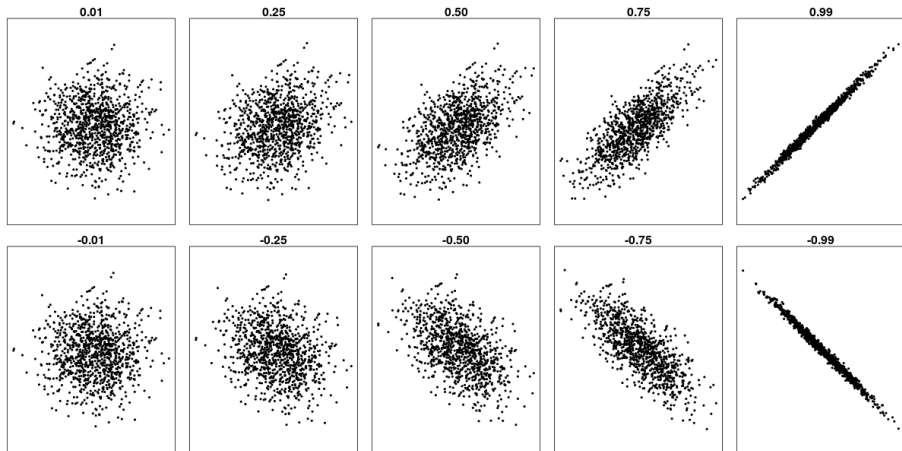
Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Le coefficient de corrélation **linéaire** de Pearson entre les observations des x et y est défini par

$$\rho(x, y) = \frac{\sigma(x, y)}{\sigma(x)\sigma(y)} \in [-1, 1].$$

- Par linéarité de la covariance, $\rho(x, y)$ est également la covariance des versions centrées-réduites de x et y .
- Le signe de $\rho(x, y)$ s'interprète comme celui de $\sigma(x, y)$.
- Si $|\rho(x, y)| = 1$, alors il y a égalité dans Cauchy-Schwarz et donc colinéarité, *i.e.* **les observations sont distribuées sur une droite.**

Coefficient de corrélation linéaire (exemple)



Régression linéaire simple

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Objectif : trouver l'équation de la droite $y = ax + b$ qui passe « au plus près » des points observés.

Critère des moindres carrés

$$\forall a, b \in \mathbb{R}, \gamma(a, b) = \frac{1}{n} \sum_{k=1}^n (y_k - ax_k - b)^2$$

La **droite de régression** est donnée par l'équation $y = \hat{a}x + \hat{b}$ telle que $(\hat{a}, \hat{b})$ minimise le critère des moindres carrés.

Les écarts au modèle linéaire $\varepsilon_k = y_k - \hat{a}x_k - \hat{b}$ sont appelés les **résidus**.

Régression linéaire simple

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Objectif : trouver l'équation de la droite $y = ax + b$ qui passe « au plus près » des points observés.

Point d'annulation du gradient :

$$\begin{aligned}\nabla \gamma(a, b) = 0 &\iff \begin{cases} \frac{1}{n} \sum_{k=1}^n x_k (y_k - ax_k - b) = 0 \\ \frac{1}{n} \sum_{k=1}^n (y_k - ax_k - b) = 0 \end{cases} \\ &\iff \begin{cases} a\sigma^2(x) = \sigma(x, y) \\ b = \bar{y} - a\bar{x} \end{cases}\end{aligned}$$

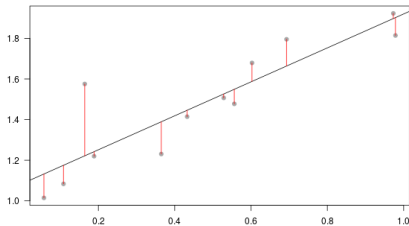
Régression linéaire simple

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Objectif : trouver l'équation de la droite $y = ax + b$ qui passe « au plus près » des points observés.

Si $\sigma^2(x) > 0$, la **droite de régression** est donnée par l'équation $y = \hat{a}x + \hat{b}$ où

$$\hat{a} = \frac{\sigma(x, y)}{\sigma^2(x)} = \frac{\rho(x, y)\sigma(y)}{\sigma(x)} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$



$$\bar{x} = 0.471 \quad \sigma^2(x) = 0.090$$

$$\bar{y} = 1.478 \quad \sigma^2(y) = 0.080$$

$$\rho(x, y) = 0.882$$

$$\Rightarrow y = 0.832x + 1.086$$

Régression linéaire : Qualité de l'ajustement

Comment peut-on mesurer la qualité de cette approximation linéaire ?

Régression linéaire : Qualité de l'ajustement

Comment peut-on mesurer la qualité de cette approximation linéaire ?

On peut décomposer la variance des $(y_i)_{i=1}^n$:

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SCT} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SCR} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SCE}.$$

Régression linéaire : Qualité de l'ajustement

Comment peut-on mesurer la qualité de cette approximation linéaire ?

On peut décomposer la variance des $(y_i)_{i=1}^n$:

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SCT} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SCR} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SCE}.$$

On regarde la proportion de variance expliquée par la régression :

$$R^2 = \frac{SCE}{SCT}.$$

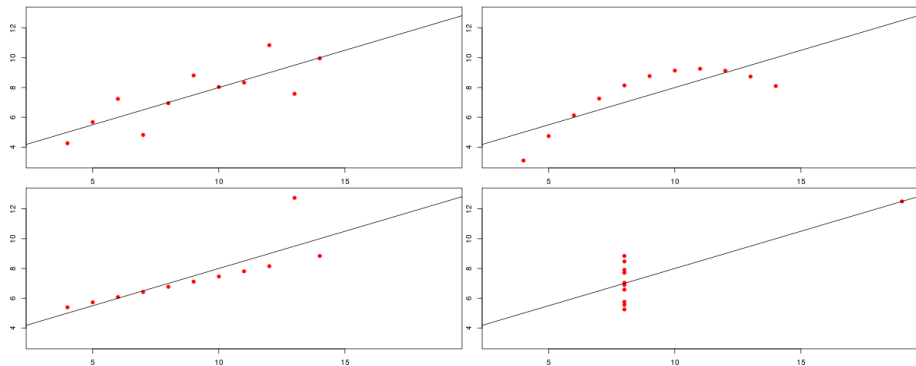
R^2 est appelé coefficient de détermination :

- R^2 proche de 1 $\implies$ Régression pertinente.
- R^2 proche de 0 $\implies$ Régression non pertinente.

Coefficient de corrélation linéaire (contre-exemple)

... make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.

F. J. Anscombe, 1973



$$\bar{x} = 9.0 \quad \sigma^2(x) = 10.0 \quad \bar{y} = 7.5 \quad \sigma^2(y) = 3.75 \quad \rho(x, y) = 0.8165$$

Coefficient de corrélation de Spearman

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Objectif : étudier l'existence d'un lien entre les observations de x et de y
sans faire l'hypothèse de linéarité.

Rangs

Soit $k \in \{1, \dots, n\}$, rx_k (resp. ry_k) désigne le rang de x_k (resp. y_k) dans la séquence des observations triées par ordre croissant. En cas d'égalité, le rang est donné par le rang moyen des observations égales.

Exemple : si $x_1 = 17$, $x_2 = 8$, $x_3 = 19$ et $x_4 = 81$, alors

$$rx_1 = 2, rx_2 = 1, rx_3 = 3, rx_4 = 4.$$

Exemple : si $x_1 = 17$, $x_2 = 8$, $x_3 = 17$ et $x_4 = 81$, alors

$$rx_1 = 2.5, rx_2 = 1, rx_3 = 2.5, rx_4 = 4.$$

Coefficient de corrélation de Spearman

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Coefficient de corrélation de Spearman

$$\rho_S(x, y) = \rho(rx, ry) = \frac{\sigma(rx, ry)}{\sigma(rx)\sigma(ry)}$$

Si **tous les rangs sont distincts**, alors

$$\rho_S(x, y) = 1 - \frac{6}{n(n^2 - 1)} \sum_{k=1}^n (rx_k - ry_k)^2.$$

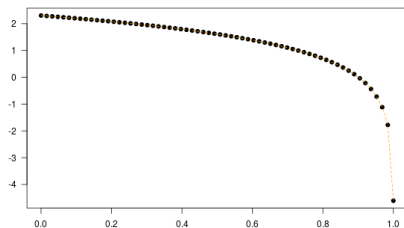
En effet, il est facile de voir dans ce cas que $\overline{rx} = \overline{ry} = (n+1)/2$ et que $\sigma^2(rx) = \sigma^2(ry) = (n^2 - 1)/12$. Le résultat découle du calcul suivant,

$$\frac{1}{n} \sum_{k=1}^n (rx_k - ry_k)^2 = \frac{n^2 - 1}{6} - 2\sigma(rx, ry).$$

Coefficient de corrélation de Spearman

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

- Par construction, $\rho_S(x, y) \in [-1, 1]$.
- Si $|\rho_S(x, y)|$ est proche de 1, le lien entre les observations de x et y est donné par une **fonction monotone**.
- Le **signe** de $\rho_S(x, y)$ donne le sens de monotonie : croissante si $\rho_S(x, y) > 0$ et décroissante si $\rho_S(x, y) < 0$.



$$\rho(x, y) = -0.816$$

$$\rho_S(x, y) = -1$$

Coefficient de corrélation de Kendall

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Objectif : étudier l'existence d'un lien entre les observations de x et de y sans faire l'hypothèse de linéarité.

Concordance

Soit $k_1, k_2 \in \{1, \dots, n\}$, les couples d'observations (x_{k_1}, y_{k_1}) et (x_{k_2}, y_{k_2}) sont dits **concordants** si

$$x_{k_1} < x_{k_2} \text{ et } y_{k_1} < y_{k_2} \quad \textbf{ou} \quad x_{k_1} > x_{k_2} \text{ et } y_{k_1} > y_{k_2},$$

ou **discordants** si

$$x_{k_1} < x_{k_2} \text{ et } y_{k_1} > y_{k_2} \quad \textbf{ou} \quad x_{k_1} > x_{k_2} \text{ et } y_{k_1} < y_{k_2}.$$

Si $x_{k_1} = x_{k_2}$ ou $y_{k_1} = y_{k_2}$, le couple n'est ni concordant, ni discordant.

Coefficient de corrélation de Kendall

Échantillon : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$.

Coefficient de corrélation de Kendall

$$\rho_K(x, y) = \frac{2(R_+ - R_-)}{n(n-1)}$$

où R_+ (resp. R_-) est le nombre de couples concordants (resp. discordants).

- Un simple argument combinatoire donne $\rho_K(x, y) \in [-1, 1]$.
- Une valeur $|\rho_K(x, y)|$ proche de 1 suggère un lien **monotone** entre les observations de x et de y .
- L'interprétation est similaire à celle du coefficient de corrélation de Spearman.

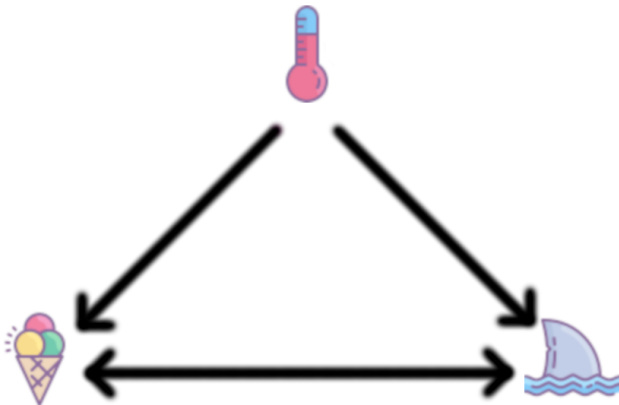
Cum hoc sed non propter hoc (Avec ceci mais pas à cause de ceci)

« Corrélation n'est pas causalité »

Quelques exemples :

- Le fait de dormir avec ses chaussures est fortement corrélé avec celui de se réveiller avec la « gueule de bois ».
→ Dormir avec des chaussures donne-t-il la « gueule de bois » ? Un autre facteur est-il impliqué ?
- La fréquence des attaques de requins est fortement corrélée avec la vente de glaces sur la plage.
→ Est-on plus appétissant pour les requins en ayant mangé de la glace ?
- Les personnes qui meurent ont très souvent vu un médecin dans les jours qui ont précédé.
→ Est-ce dangereux de rencontrer un médecin ?

Corrélation n'est pas causalité



Corrélation partielle

Échantillon : $(x_1, y_1, z_1), \dots, (x_n, y_n, z_n) \in \mathbb{R}^3$.

La **corrélation partielle** entre x et y **conditionnellement** à z est la corrélation linéaire entre les résidus ε^x et ε^y des régressions linéaires sur (x, z) et (y, z) respectivement.

La corrélation partielle permet d'évaluer la corrélation entre les observations de deux variables après avoir contrôlé l'effet perturbateur d'une ou de plusieurs autres variables.

Corrélation partielle

Échantillon : $(x_1, y_1, z_1), \dots, (x_n, y_n, z_n) \in \mathbb{R}^3$.

La **corrélation partielle** entre x et y **conditionnellement** à z est la corrélation linéaire entre les résidus ε^x et ε^y des régressions linéaires sur (x, z) et (y, z) respectivement.

Corrélation linéaire

	Temp	Glace	Requin
Temp	1.00	0.95	0.93
Glace	.	1.00	0.88
Requin	.	.	1.00

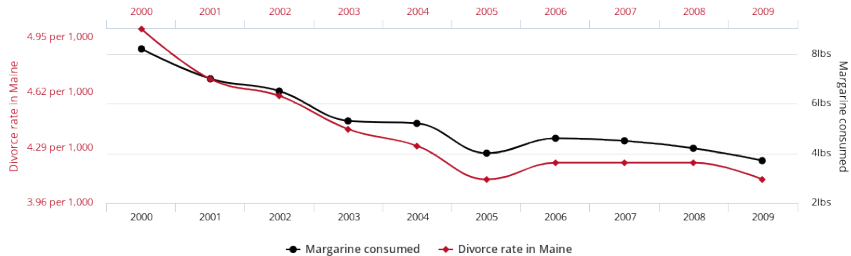
Corrélation partielle

	Temp	Glace	Requin
Temp	1.00	0.74	0.66
Glace	.	1.00	0.04
Requin	.	.	1.00

Spurious correlation

Divorce rate in Maine correlates with Per capita consumption of margarine

Correlation: 99.26% ($r=0.992558$)



Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

tylervigen.com

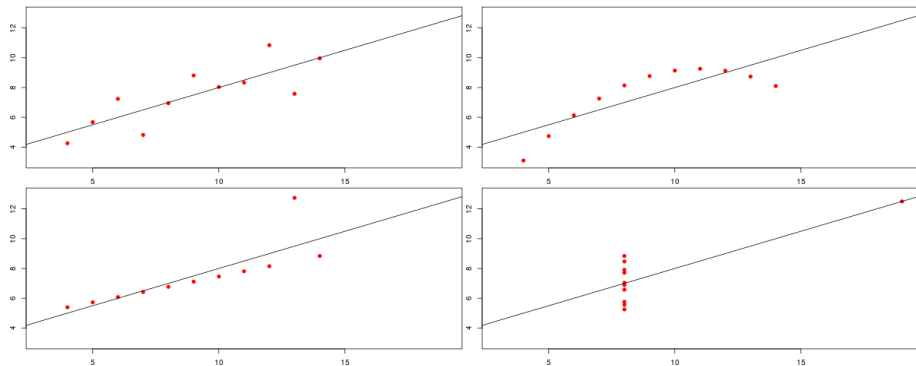
<http://www.tylervigen.com/spurious-correlations>

1.2 Représentations graphiques

Pourquoi visualiser ?

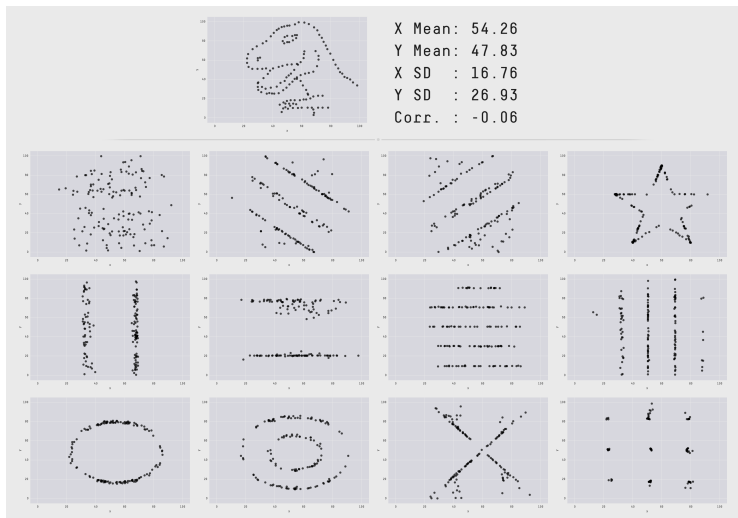
... make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.

F. J. Anscombe, 1973



$$\bar{x} = 9.0 \quad \sigma^2(x) = 10.0 \quad \bar{y} = 7.5 \quad \sigma^2(y) = 3.75 \quad \rho(x, y) = 0.8165$$

Datasaurus (Alberto Cairo)



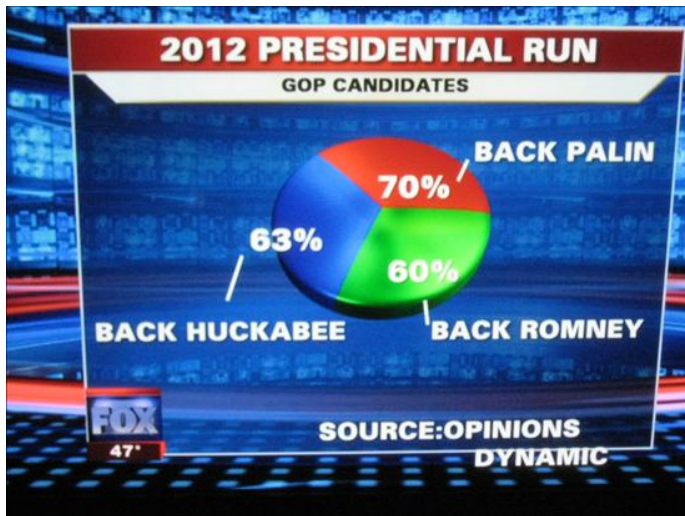
<https://www.autodeskresearch.com/publications/samestats>

The Visual Display of Quantitative Information (E. Tufte, 1983)

Tufte a popularisé plusieurs bonnes pratiques graphiques :

- Montrer les données.
- Inciter celui ou celle qui regarde à penser.
- Éviter de distordre ce que les données ont à dire.
- Présenter beaucoup de données sur une petite surface.
- Révéler les données à des niveaux différents : d'un aperçu global à des structures plus fines.
- Servir un objectif clair et raisonnable.
- Être étroitement intégré à une description statistique du jeu de données.

Mauvaise visualisation I

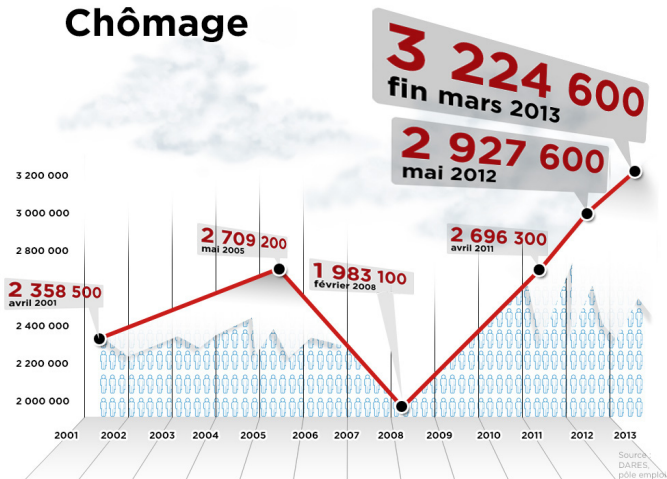


Mauvaise visualisation II



Mauvaise visualisation III

Chômage



Données de type « effectif »

A	30
B	15
C	30
D	20
E	25



Diagramme en bâtons (pile)

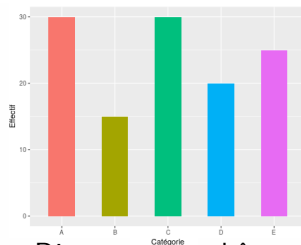
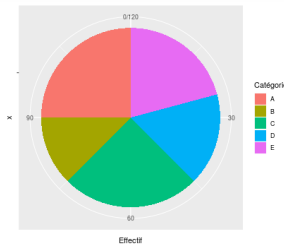
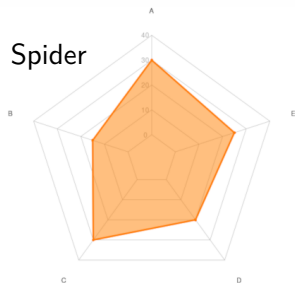


Diagramme en bâtons

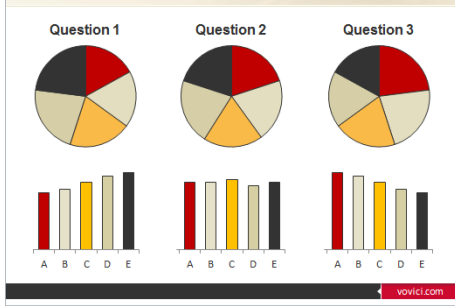
Radar / Spider



Camembert (Pie)

Les problèmes du camembert

Bar Charts are Easier to Interpret than Pie Charts



Lisibilité difficile : les diagrammes en bâtons sont souvent préférables

Attention aux représentations en 3D.
Est-ce évident que A et C ont la même valeur ?



Données de type « effectif » (2D)

	X	Y
A	30	25
B	15	15
C	30	20
D	20	30
E	25	35

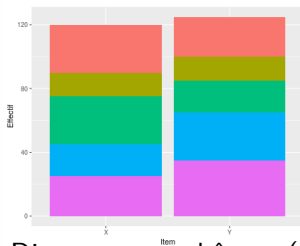


Diagramme en bâtons (pile)

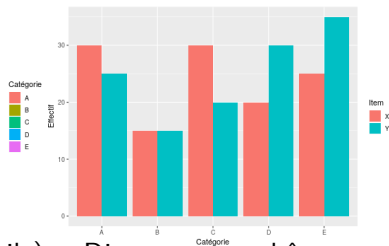
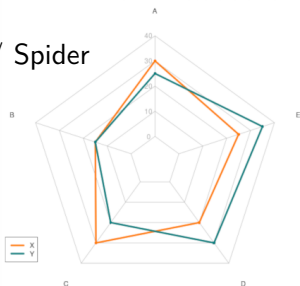


Diagramme en bâtons

Radar / Spider

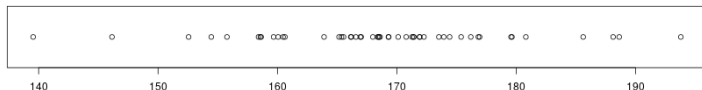
Tore
(Donut)

Données réelles

Taille de 50 individus en centimètres :

166.9	159.7	174.4	188.1	166.2	158.6	139.5	193.8	167.0	168.5
171.3	179.6	165.2	171.4	169.3	176.8	168.0	158.6	160.1	168.6
155.8	170.1	158.4	173.5	172.3	146.1	170.8	176.2	185.6	165.4
171.4	160.6	168.5	171.9	169.3	171.9	166.6	163.9	154.5	179.6
176.9	180.8	175.4	166.1	165.5	188.6	168.4	173.9	152.6	160.5

Nuage de points
(strip chart)

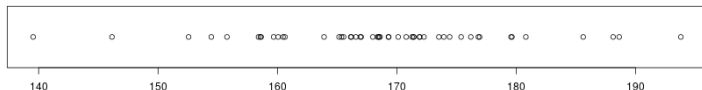


Données réelles

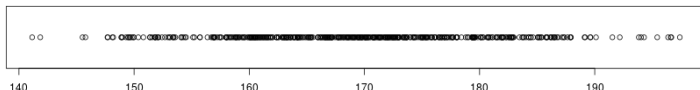
Taille de 50 individus en centimètres :

166.9	159.7	174.4	188.1	166.2	158.6	139.5	193.8	167.0	168.5
171.3	179.6	165.2	171.4	169.3	176.8	168.0	158.6	160.1	168.6
155.8	170.1	158.4	173.5	172.3	146.1	170.8	176.2	185.6	165.4
171.4	160.6	168.5	171.9	169.3	171.9	166.6	163.9	154.5	179.6
176.9	180.8	175.4	166.1	165.5	188.6	168.4	173.9	152.6	160.5

Nuage de points
(strip chart)



Avec 500 observations, la lisibilité devient délicate ...



Données réelles

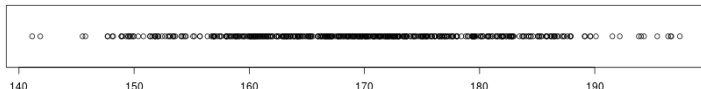
Taille de 50 individus en centimètres :

166.9	159.7	174.4	188.1	166.2	158.6	139.5	193.8	167.0	168.5
171.3	179.6	165.2	171.4	169.3	176.8	168.0	158.6	160.1	168.6
155.8	170.1	158.4	173.5	172.3	146.1	170.8	176.2	185.6	165.4
171.4	160.6	168.5	171.9	169.3	171.9	166.6	163.9	154.5	179.6
176.9	180.8	175.4	166.1	165.5	188.6	168.4	173.9	152.6	160.5

Nuage de points
(strip chart)



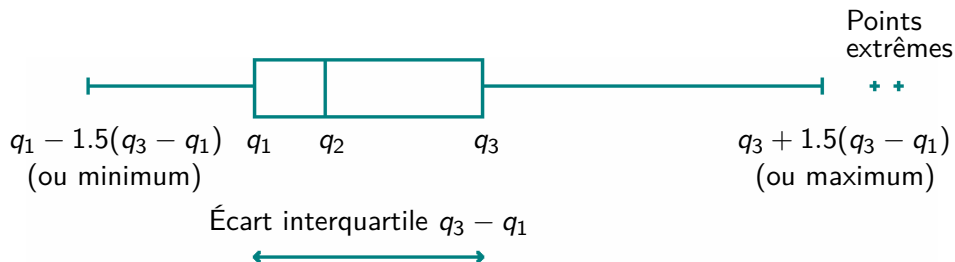
Avec 500 observations, la lisibilité devient délicate ...



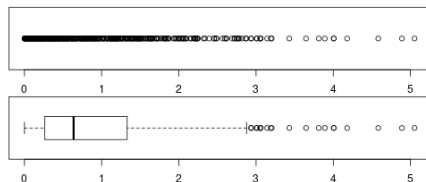
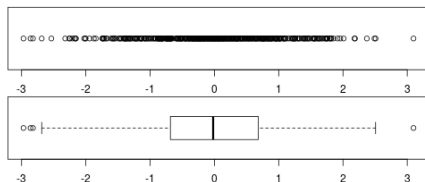
... et il est nécessaire de **résumer** l'information affichée.



Boxplot

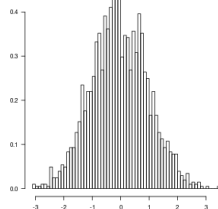
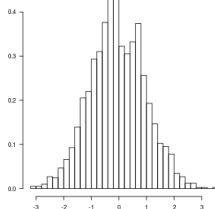
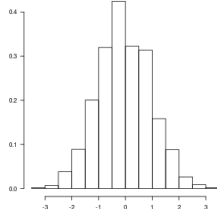
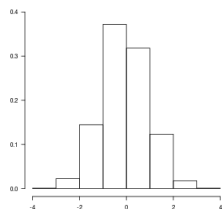


Notations : premier quartile q_1 , médiane q_2 et troisième quartile q_3 .



Histogramme et noyau

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.



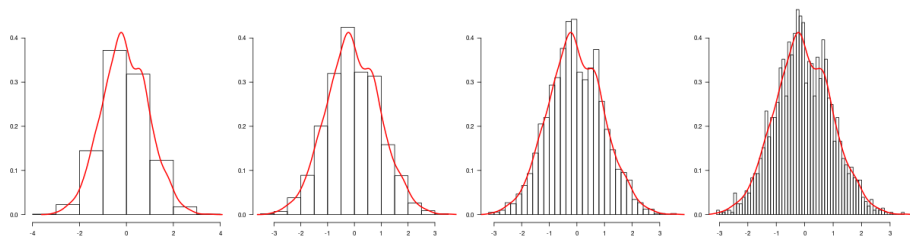
Pour un **histogramme** construit sur des blocs de taille $h > 0$, la valeur prise sur le segment $[b - h/2, b + h/2[$ vaut

$$\frac{\text{Nombre de points dans } [b - h/2, b + h/2[}{\text{Nombre de points total}} = \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{|b - x_k| \leq h/2}$$

$\Rightarrow$ La surface de l'histogramme vaut 1 (**fonction de densité**)

Histogramme et noyau

Échantillon réel : $x_1, \dots, x_n \in \mathbb{R}$.



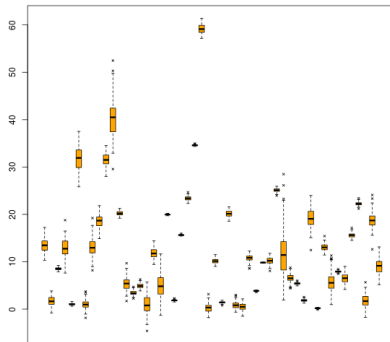
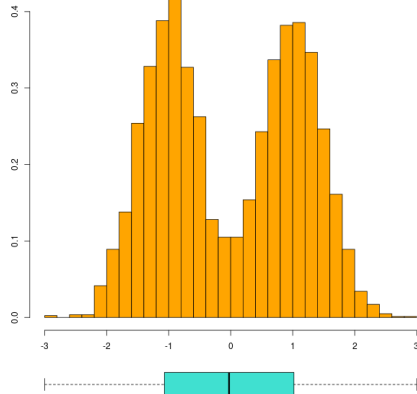
Un **noyau** $K : \mathbb{R} \rightarrow [0, \infty[$ est une fonction paire telle que $\int_{\mathbb{R}} K(t)dt = 1$.
 L'**estimateur par noyau** de la distribution des observations est donné par

$$\hat{f}_K(t) = \frac{1}{n} \sum_{k=1}^n K(t - x_k).$$

$\Rightarrow$ La surface sous la courbe vaut 1 (**fonction de densité**)

Boxplot versus Histogramme

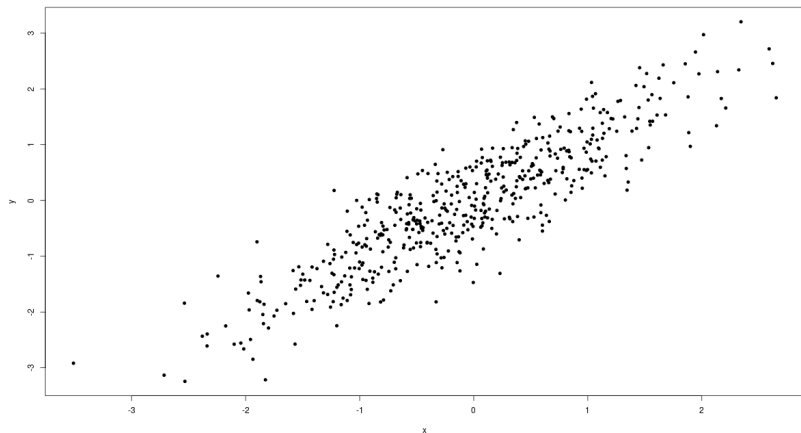
Un histogramme permet de capter des distributions multimodales ...



... mais des boxplots permettent de rendre lisibles de nombreuses variables sur un même graphique.

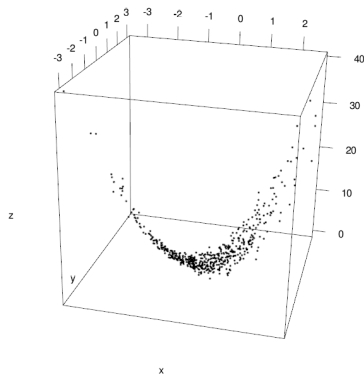
Données quantitatives (2D)

Dans le cas d'observations bidimensionnelles $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$, il est possible de tracer un **nuage de points** (ou **scatter plot**).



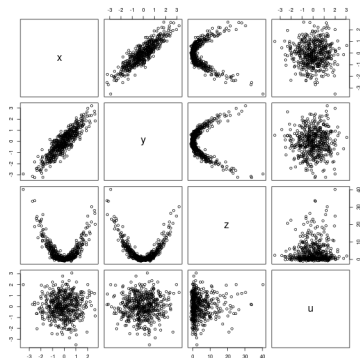
Données quantitatives (3D)

Pour des observations tridimensionnelles $(x_1, y_1, z_1), \dots, (x_n, y_n, z_n) \in \mathbb{R}^3$, il est encore possible de considérer un **nuage de points** en utilisant des outils de visualisation en 3D (e.g. le package `rgl` avec R).



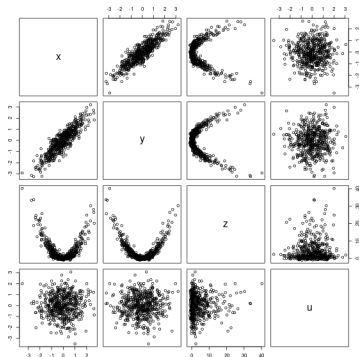
Données quantitatives multidimensionnelles

À partir de la dimension 3, il devient difficile de proposer une représentation graphique simple des données. Une possibilité consiste à tracer les nuages de points associés à chaque paire de variables observées.



Données quantitatives multidimensionnelles

À partir de la dimension 3, il devient difficile de proposer une représentation graphique simple des données. Une possibilité consiste à tracer les nuages de points associés à chaque paire de variables observées.

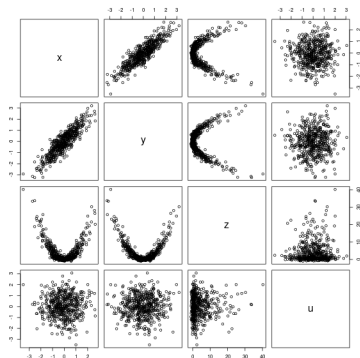


Acceptable pour une première exploration grossière mais plusieurs inconvénients :

- lisibilité difficile pour un nombre important de variables,
- étude limitée à des paires de variables,
- projections uniquement sur des plans parallèles aux axes,
- . . .

Données quantitatives multidimensionnelles

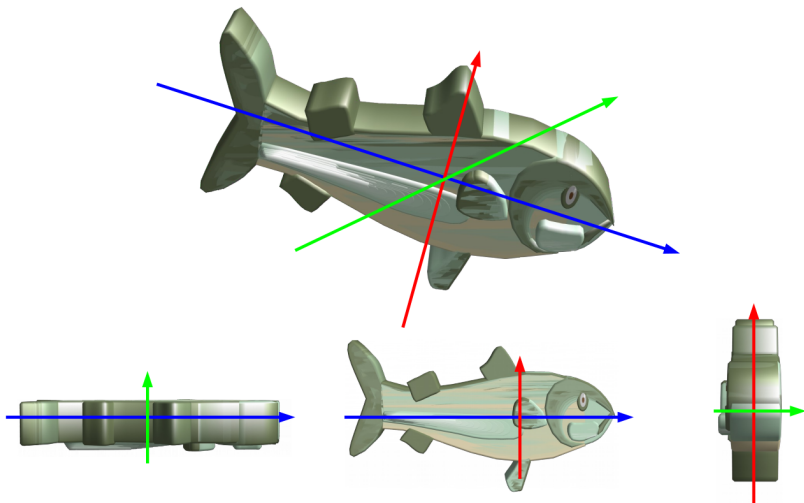
À partir de la dimension 3, il devient difficile de proposer une représentation graphique simple des données. Une possibilité consiste à tracer les nuages de points associés à chaque paire de variables observées.



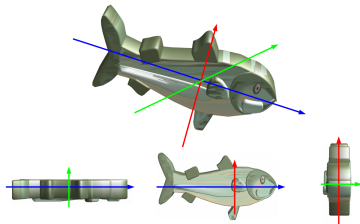
Nous avons besoin d'une méthode systématique et plus souple.

1.3 Analyse en composantes principales (ACP)

Motivation : un problème de projection



Motivation : un problème de projection



Parmi les projections en 2D, toutes ne permettent pas de reconnaître aussi facilement l'objet initial, *i.e.* elles ne contiennent pas toutes la même **quantité d'information** et n'ont pas la même **capacité à résumer**.

La projection du milieu apparaît comme la plus fidèle à l'original. Elle correspond au plan où l'objet initial s'étale le plus, *i.e.* admet la **plus grande variabilité**.

L'information apportée par la 3ème dimension est **minimale** et sa perte n'est pas préjudiciable.

Quelques notations

Dans cette section, nous considérons un jeu de données quantitatives issu de n observations de p variables $x^1, \dots, x^p$ à valeurs réelles,

$$x_1, \dots, x_n \in \mathbb{R}^p$$

avec, pour tout $k \in \{1, \dots, n\}$,

$$x_k = \begin{pmatrix} x_k^1 \\ \vdots \\ x_k^p \end{pmatrix} \in \mathbb{R}^p.$$

Autrement dit, pour tout $k \in \{1, \dots, n\}$ et $\ell \in \{1, \dots, p\}$, $x_k^\ell \in \mathbb{R}$ est la valeur prise par la **variable** x^ℓ sur le k -ème individu.

Quelques notations

Échantillon : $x_1, \dots, x_n \in \mathbb{R}^p$

La **matrice des données (centrées)** de taille $n \times p$ est

$$X = \begin{pmatrix} x_1^1 - \bar{x}^1 & \dots & x_1^p - \bar{x}^p \\ \vdots & \vdots & \vdots \\ x_n^1 - \bar{x}^1 & \dots & x_n^p - \bar{x}^p \end{pmatrix}$$

où $\bar{x}^1, \dots, \bar{x}^p \in \mathbb{R}$ sont les moyennes des observations des variables $x^1, \dots, x^p$ respectivement,

$$\forall \ell \in \{1, \dots, p\}, \bar{x}^\ell = \frac{1}{n} \sum_{k=1}^n x_k^\ell.$$

Pour tout $k \in \{1, \dots, n\}$ et $\ell \in \{1, \dots, p\}$, la **k -ème ligne** de la matrice X contient les observations faites sur le **k -ème individu** et la **ℓ -ème colonne** de la matrice X contient les observations centrées de la **variable x^ℓ** .

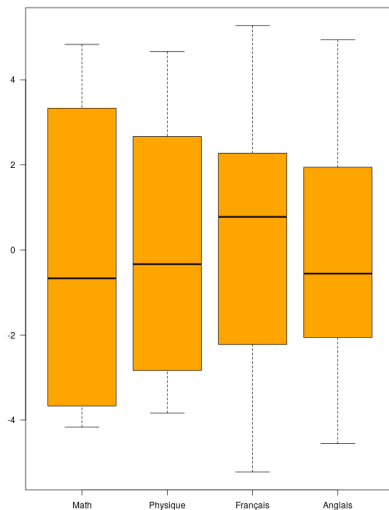
Exemple-jouet : observations

Considérons les notes obtenues par $n = 9$ étudiants dans $p = 4$ matières :

	Math	Physique	Français	Anglais
Benny	6.0	6.0	5.0	5.5
Bobby	8.0	8.0	8.0	8.0
Brandy	6.0	7.0	11.0	9.5
Coby	14.5	14.5	15.5	15.0
Daisy	14.0	14.0	12.0	12.5
Emily	11.0	10.0	5.5	7.0
Judy	5.5	7.0	14.0	11.5
Marty	13.0	12.5	8.5	9.5
Sandy	9.0	9.5	12.5	12.0
Moyenne	9.67	9.83	10.22	10.06

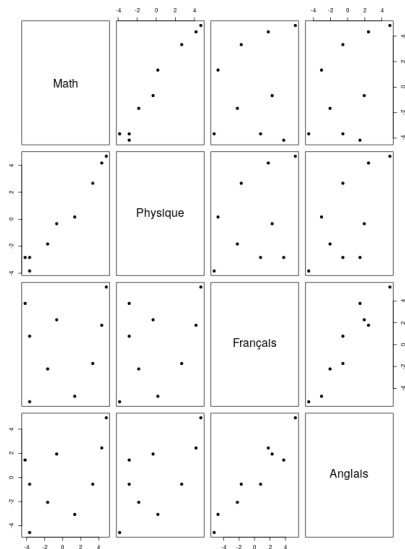
Exemple-jouet : données et visualisation élémentaire

$$X = \begin{pmatrix} -3.67 & -3.83 & -5.22 & -4.56 \\ -1.67 & -1.83 & -2.22 & -2.06 \\ -3.67 & -2.83 & 0.78 & -0.56 \\ 4.83 & 4.67 & 5.28 & 4.94 \\ 4.33 & 4.17 & 1.78 & 2.44 \\ 1.33 & 0.17 & -4.72 & -3.06 \\ -4.17 & -2.83 & 3.78 & 1.44 \\ 3.33 & 2.67 & -1.72 & -0.56 \\ -0.67 & -0.33 & 2.28 & 1.94 \end{pmatrix}$$

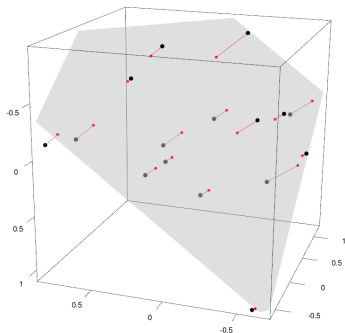


Exemple-jouet : données et visualisation élémentaire

$$X = \begin{pmatrix} -3.67 & -3.83 & -5.22 & -4.56 \\ -1.67 & -1.83 & -2.22 & -2.06 \\ -3.67 & -2.83 & 0.78 & -0.56 \\ 4.83 & 4.67 & 5.28 & 4.94 \\ 4.33 & 4.17 & 1.78 & 2.44 \\ 1.33 & 0.17 & -4.72 & -3.06 \\ -4.17 & -2.83 & 3.78 & 1.44 \\ 3.33 & 2.67 & -1.72 & -0.56 \\ -0.67 & -0.33 & 2.28 & 1.94 \end{pmatrix}$$



Principes de l'ACP



Déterminer les espaces de dimension inférieure à l'espace initial sur lesquels la projection du nuage de points initial est **la moins déformée possible**, *i.e.* celle qui conserve **le plus d'information au sens de la variabilité**.

Du point de vue des variables, cela correspond à chercher les **combinaisons linéaires** qui préservent au mieux la **structure de corrélation** entre les valeurs des données initiales.

Mesure de la variabilité

Échantillon : $x_1, \dots, x_n \in \mathbb{R}^p$

Une façon simple de généraliser la variance dans un cadre multidimensionnel consiste à définir l'**inertie (standard)** comme la somme des variances,

$$I(x) = \sum_{\ell=1}^p \sigma^2(x^\ell) = \frac{1}{n} \sum_{k=1}^n \sum_{\ell=1}^p (x_k^\ell - \bar{x}^\ell)^2.$$

Cette quantité s'écrit également $I(x) = \frac{1}{n} \sum_{k=1}^n \|x_k - \bar{x}\|^2$ où

$$\bar{x} = \begin{pmatrix} \bar{x}^1 \\ \vdots \\ \bar{x}^p \end{pmatrix} \in \mathbb{R}^p \quad \text{et} \quad \forall v = \begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix} \in \mathbb{R}^p, \quad \|v\|^2 = v^\top v = \sum_{\ell=1}^p v_\ell^2.$$

Projection des données

Soit $v \in \mathbb{R}^p$ tel que $\|v\|^2 = 1$.

Pour tout $k \in \{1, \dots, n\}$, la **projection orthogonale** du vecteur observé $x_k - \bar{x}$ sur la **droite engendrée par** v est donnée par

$$\langle x_k - \bar{x}, v \rangle v$$

avec le **produit scalaire** $\langle u, v \rangle = u^\top v = \sum_{\ell=1}^p u_\ell v_\ell$ pour tout $u, v \in \mathbb{R}^p$.

Autrement dit, le **résumé des observations le long de** v est donné par les coordonnées des n vecteurs initiaux

$$\langle x_1 - \bar{x}, v \rangle, \dots, \langle x_n - \bar{x}, v \rangle \in \mathbb{R}.$$

Projection des données

Soit $v \in \mathbb{R}^p$ tel que $\|v\|^2 = 1$.

Observations projetées sur $\mathbb{R}v$: $\langle x_1 - \bar{x}, v \rangle, \dots, \langle x_n - \bar{x}, v \rangle \in \mathbb{R}$.

Les données projetées sont **centrées** par construction,

$$\frac{1}{n} \sum_{k=1}^n \langle x_k - \bar{x}, v \rangle = \left\langle \frac{1}{n} \sum_{k=1}^n x_k - \bar{x}, v \right\rangle = \langle \bar{x} - \bar{x}, v \rangle = 0.$$

L'inertie des données projetées (ici, il s'agit simplement de la variance) vaut

$$\begin{aligned} \frac{1}{n} \sum_{k=1}^n \langle x_k - \bar{x}, v \rangle^2 &= \frac{1}{n} \sum_{k=1}^n v^\top (x_k - \bar{x})(x_k - \bar{x})^\top v \\ &= v^\top \left(\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(x_k - \bar{x})^\top \right) v = v^\top \Sigma v. \end{aligned}$$

Matrice de covariance

La matrice $p \times p$ qui apparaît dans le calcul de l'inertie précédent,

$$\Sigma = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(x_k - \bar{x})^\top,$$

s'appelle la **matrice de covariance** car, pour tout $\ell_1, \ell_2 \in \{1, \dots, p\}$,

$$\Sigma_{\ell_1 \ell_2} = \frac{1}{n} \sum_{k=1}^n \left(x_k^{\ell_1} - \bar{x}^{\ell_1} \right) \left(x_k^{\ell_2} - \bar{x}^{\ell_2} \right) = \sigma(x^{\ell_1}, x^{\ell_2}).$$

Cette matrice s'écrit également $\Sigma = X^\top W X$ avec $W = n^{-1} Id_n$ et vérifie

- Σ est **symétrique** : $\Sigma_{\ell_1 \ell_2} = \sigma(x^{\ell_1}, x^{\ell_2}) = \sigma(x^{\ell_2}, x^{\ell_1}) = \Sigma_{\ell_2 \ell_1}$,
- Σ est **positive** : $\forall u \in \mathbb{R}^p$, $u^\top \Sigma u = \|Xu\|^2/n \geq 0$.

Première direction principale

Déterminer la droite sur laquelle la projection du nuage de points initial conserve **le plus d'information au sens de la variabilité** revient donc à chercher le vecteur unitaire $v \in \mathbb{R}^p$ tel que $v^T \Sigma v \in \mathbb{R}$ soit maximal.

Puisque Σ est symétrique et positive, le vecteur v solution de ce problème est donné par ???

Première direction principale

Déterminer la droite sur laquelle la projection du nuage de points initial conserve **le plus d'information au sens de la variabilité** revient donc à chercher le vecteur unitaire $v \in \mathbb{R}^p$ tel que $v^\top \Sigma v \in \mathbb{R}$ soit maximal.

Puisque Σ est symétrique et positive, le vecteur v solution de ce problème est donné par le **vecteur propre associé à la plus grande valeur propre de Σ** .

En effet, le théorème spectral assure qu'il existe une **matrice orthogonale** V (i.e. $V^{-1} = V^\top$) de taille $p \times p$ dont les colonnes $v^1, \dots, v^p \in \mathbb{R}^p$ forment une **base orthonormale** de $\mathbb{R}^p$ et correspondent aux **vecteurs propres** de Σ associés aux **valeurs propres** respectives

$$\lambda_1 \geq \dots \geq \lambda_p \geq 0.$$

Le vecteur $v^1 \in \mathbb{R}^p$ est la **première direction principale** et l'inertie maximale le long d'une droite est donnée par $v^{1\top} \Sigma v^1 = \lambda_1$.

Part d'inertie expliquée

Le problème unidimensionnel précédent se généralise immédiatement à la recherche d'un espace de dimension $d \leq p$ sur lequel la projection du nuage de points initial conserve **le plus d'information au sens de la variabilité**.

La solution est fournie par l'**espace E_d engendré par les d vecteurs propres orthonormaux $v^1, \dots, v^d \in \mathbb{R}^p$** qui sont donnés par les d premières colonnes de la matrice V (*i.e.* les d premières directions principales).

L'inertie des données projetées dans E_d vaut

$$\sum_{\ell=1}^d v^{\ell \top} \Sigma v^{\ell} = \sum_{\ell=1}^d \lambda_{\ell} \left(\leq \sum_{\ell=1}^p \lambda_{\ell} = I(x) \right).$$

La **part d'inertie expliquée** est la quantité

$$\frac{1}{I(x)} \sum_{\ell=1}^d \lambda_{\ell}.$$

Composantes principales

Les directions principales $v^1, \dots, v^p \in \mathbb{R}^p$ forment une **base orthonormale** de $\mathbb{R}^p$ dans laquelle les données initiales peuvent être représentées. Cette représentation correspond à des **combinaisons linéaires** des variables initiales qui préservent au mieux la **structure de corrélation**.

Les coordonnées des données initiales dans la base des directions principales sont fournies par la matrice de taille $n \times p$ définie par

$$C = XV.$$

La matrice C est la **matrice des composantes principales**.

Composantes principales

Les colonnes $c^1, \dots, c^p \in \mathbb{R}^n$ de C s'appellent les **composantes principales** et doivent être considérées comme p variables « virtuelles » obtenues par combinaisons linéaires des variables initiales,

$$\forall \ell \in \{1, \dots, p\}, \quad c^\ell = \begin{pmatrix} c_1^\ell \\ \vdots \\ c_n^\ell \end{pmatrix} = Xv^\ell \in \mathbb{R}^n.$$

Les composantes principales sont **centrées** par construction et leur matrice de covariance est donnée par

$$C^\top WC = V^\top X^\top W X V = V^\top \Sigma V = \Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_p \end{pmatrix}.$$

Composantes principales

Les composantes principales sont des variables **décorrélées** et **ordonnées par la quantité d'information**,

$$\forall \ell, \ell' \in \{1, \dots, p\}, \sigma(c^\ell, c^{\ell'}) = \begin{cases} \lambda_\ell & \text{si } \ell = \ell', \\ 0 & \text{sinon.} \end{cases}$$

Pour représenter les données initiales dans un espace de dimension $d \leq p$, nous utilisons donc les coordonnées fournies par les d premières composantes principales $c^1, \dots, c^d$.

Le cas particulier $d = 2$ correspond au **plan principal** et permet de donner une représentation graphique des données de dimension p (correspondant à la part d'inertie expliquée par les deux premières directions principales).

Exemple-jouet : plan principal

Matrice de covariance $\Sigma = X^T W X$:

	Math	Physique	Français	Anglais
Math	11.39	9.92	2.66	4.82
Physique	9.92	8.94	4.12	5.48
Français	2.66	4.12	12.06	9.29
Anglais	4.82	5.48	9.29	7.91

Inertie du nuage de points initial : $I(x) = \text{Tr}(\Sigma) = 40.31$

Matrice des directions principales :

$$V = \begin{matrix} & \begin{matrix} v^1 & v^2 & v^3 & v^4 \end{matrix} \\ \begin{pmatrix} -0.515 & 0.569 & 0.185 & 0.614 \\ -0.508 & 0.371 & -0.450 & -0.634 \\ -0.492 & -0.658 & -0.460 & 0.335 \\ -0.484 & -0.325 & 0.742 & -0.329 \end{pmatrix} \end{matrix}$$

Exemple-jouet : plan principal

Matrice des composantes principales $C = XV$:

	c^1	c^2	c^3	c^4
Benny	8.61	1.41	0.07	-0.07
Bobby	3.88	0.50	0.01	0.07
Brandy	3.21	-3.47	-0.17	-0.01
Coby	-9.85	-0.60	0.04	0.15
Daisy	-6.41	2.05	-0.08	-0.19
Emily	3.03	4.92	0.08	0.14
Judy	1.03	-6.38	-0.16	0.03
Marty	-1.95	4.20	-0.20	-0.04
Sandy	-1.55	-2.63	0.42	-0.07

Valeurs propres de Σ : $\lambda_1 = 28.23$, $\lambda_2 = 12.03$, $\lambda_3 = 0.03$ et $\lambda_4 = 0.01$

Exemple-jouet : plan principal

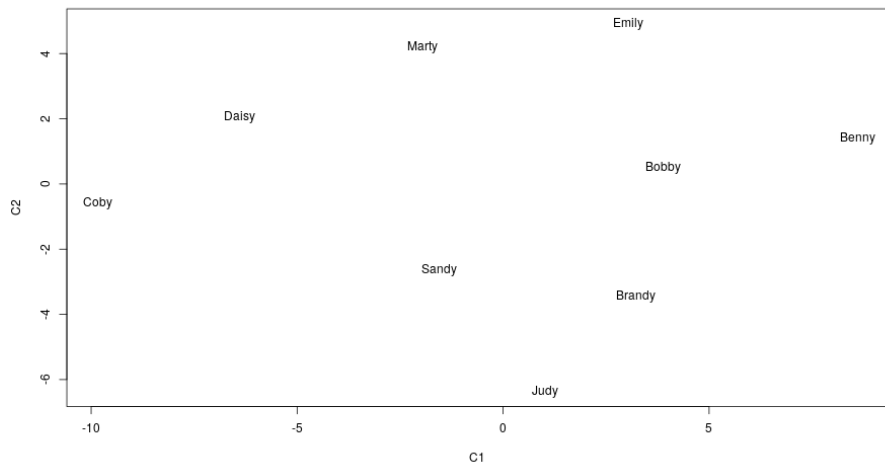
Matrice des composantes principales $C = XV$:

	c^1	c^2	c^3	c^4
Benny	8.61	1.41	0.07	-0.07
Bobby	3.88	0.50	0.01	0.07
Brandy	3.21	-3.47	-0.17	-0.01
Coby	-9.85	-0.60	0.04	0.15
Daisy	-6.41	2.05	-0.08	-0.19
Emily	3.03	4.92	0.08	0.14
Judy	1.03	-6.38	-0.16	0.03
Marty	-1.95	4.20	-0.20	-0.04
Sandy	-1.55	-2.63	0.42	-0.07

Valeurs propres de Σ : $\lambda_1 = 28.23$, $\lambda_2 = 12.03$, $\lambda_3 = 0.03$ et $\lambda_4 = 0.01$

Part d'inertie expliquée par le **plan principal** : $(\lambda_1 + \lambda_2)/I(x) = 99.89\%$

Exemple-jouet : plan principal



Représentation des variables

Pour discuter du lien entre les variables initiales et les composantes principales, nous considérons les **coefficients de corrélation linéaire** :

$$\forall \ell, \ell' \in \{1, \dots, p\}, \quad \rho(x^\ell, c^{\ell'}) = \frac{\sigma(x^\ell, c^{\ell'})}{\sigma(x^\ell)\sqrt{\lambda_{\ell'}}} = \frac{\sqrt{\lambda_{\ell'}}}{\sigma(x^\ell)} v_{\ell}^{\ell'}$$

car $X = CV^\top$ par orthogonalité, i.e. $x^\ell = \sum_{j=1}^p v_{\ell}^j c^j$.

Pour $\ell \in \{1, \dots, p\}$, le point

$$\left(\rho(x^\ell, c^1), \rho(x^\ell, c^2) \right) = \left(\frac{\sqrt{\lambda_1}}{\sigma(x^\ell)} v_{\ell}^1, \frac{\sqrt{\lambda_2}}{\sigma(x^\ell)} v_{\ell}^2 \right)$$

permet de **représenter graphiquement le lien entre x^ℓ et les deux premières composantes principales**.

Cercle des corrélations

Par construction, pour tout $\ell \in \{1, \dots, p\}$, nous avons

$$\sum_{\ell'=1}^p \rho(x^\ell, c^{\ell'})^2 = \sum_{\ell'=1}^p \frac{\lambda_{\ell'}}{\sigma^2(x^\ell)} (v_{\ell'}^\ell)^2 = \frac{1}{\sigma^2(x^\ell)} (V \Lambda V^\top)_{\ell\ell} = \frac{\Sigma_{\ell\ell}}{\sigma^2(x^\ell)} = 1.$$

Ainsi,

$$\rho(x^\ell, c^1)^2 + \rho(x^\ell, c^2)^2 \leq 1$$

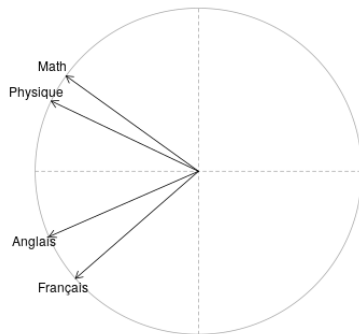
et le point $(\rho(x^\ell, c^1), \rho(x^\ell, c^2))$ est **dans le disque de rayon unité**. La proximité au cercle traduit la **qualité de la représentation** et la direction indique les **liens avec les deux premières composantes**.

La représentation de toutes les variables initiales sur un même graphique s'appelle le **cercle des corrélations**.

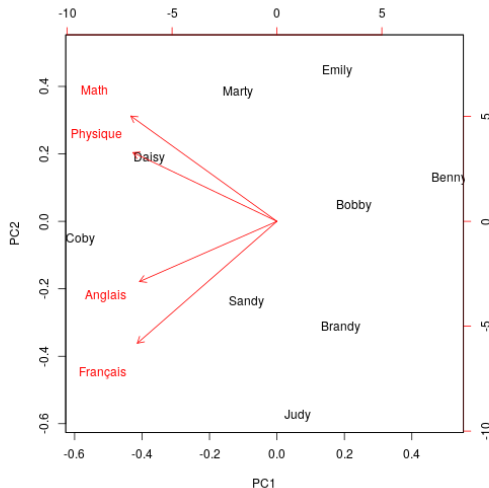
Exemple-jouet : cercle des corrélations

Coefficients de corrélation linéaire :

	c^1	c^2
Math	-0.81	0.58
Physique	-0.90	0.43
Français	-0.75	-0.66
Anglais	-0.91	-0.40



Exemple-jouet : représentation biplot



Représentation simultanée des individus et des variables

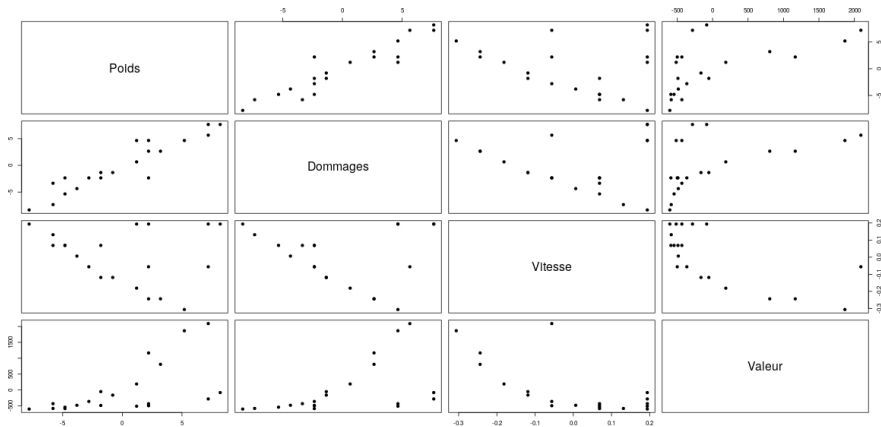
Exemple (un peu plus sérieux mais pas trop)

Considérons les $p = 4$ caractéristiques des $n = 20$ arcs et arbalètes sans enchantement du jeu « *The Elder Scrolls V : Skyrim* » (avec extensions) :

	Poids	Dommages	Vitesse	Valeur
Arc long	5	6	1.0000	30
Arc de chasse	7	7	0.9375	50
Arc nordique du Héros	7	11	0.8750	200
Arc nordique antique	8	12	0.8750	45
Arc impérial	8	9	0.8750	90
Arc de Parjure	11	12	0.8750	145
Arc orque	9	10	0.8126	150
Arc falmer	15	12	0.7500	135
Arc dwemer	10	12	0.7500	270
Arc elfique	12	13	0.6875	470
Arc nordique	11	13	0.6875	580
Arc de verre	14	15	0.6250	820
Arc d'ébonite	16	17	0.5625	1440
Arc de stalhrim	15	17	0.5625	1800
Arc daédra	18	19	0.5000	2500
Arc d'os de dragon	20	20	0.7500	2725
Arbalète	14	19	1.0000	120
Arbalète améliorée	15	19	1.0000	200
Arbalète dwemer	20	22	1.0000	350
Arbalète dwemer améliorée	21	22	1.0000	550

Exemple (un peu plus sérieux mais pas trop)

Comme dans l'exemple-jouet, nous définissons la matrice X des données centrées et une étude exploratoire grossière suggère une structure de corrélation entre les variables.



Exemple (un peu plus sérieux mais pas trop)

En diagonalisant la matrice de covariance $\Sigma = X^\top W X$ où $W = n^{-1} Id_n$, nous obtenons la matrice V des directions principales, la matrice des composantes principales $C = X V$ et les valeurs propres

$$\lambda_1 = 643016.54, \lambda_2 = 28.14, \lambda_3 = 1.37 \text{ et } \lambda_4 = 0.01.$$

L'inertie du nuage de points initial vaut $I(x) = \text{Tr}(\Sigma) = 643046.1$.

La part d'inertie expliquée par le plan principal est égale à

$$\frac{\lambda_1 + \lambda_2}{I(x)} = 99.99\%.$$

Exemple (un peu plus sérieux mais pas trop)

En diagonalisant la matrice de covariance $\Sigma = X^T W X$ où $W = n^{-1} Id_n$, nous obtenons la matrice V des directions principales, la matrice des composantes principales $C = X V$ et les valeurs propres

$$\lambda_1 = 643016.54, \lambda_2 = 28.14, \lambda_3 = 1.37 \text{ et } \lambda_4 = 0.01.$$

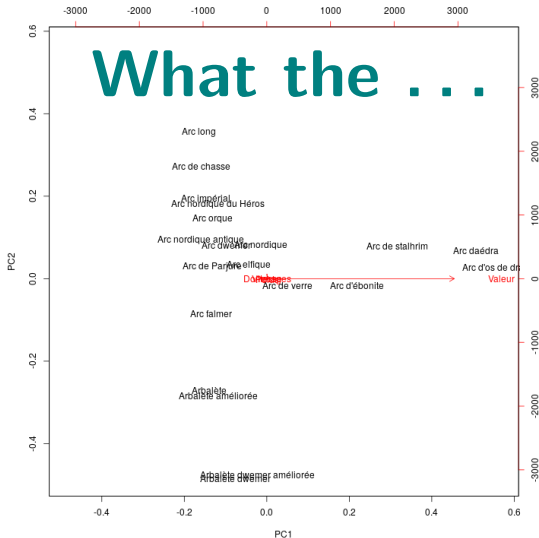
L'inertie du nuage de points initial vaut $I(x) = \text{Tr}(\Sigma) = 643046.1$.

La part d'inertie expliquée par le plan principal est égale à

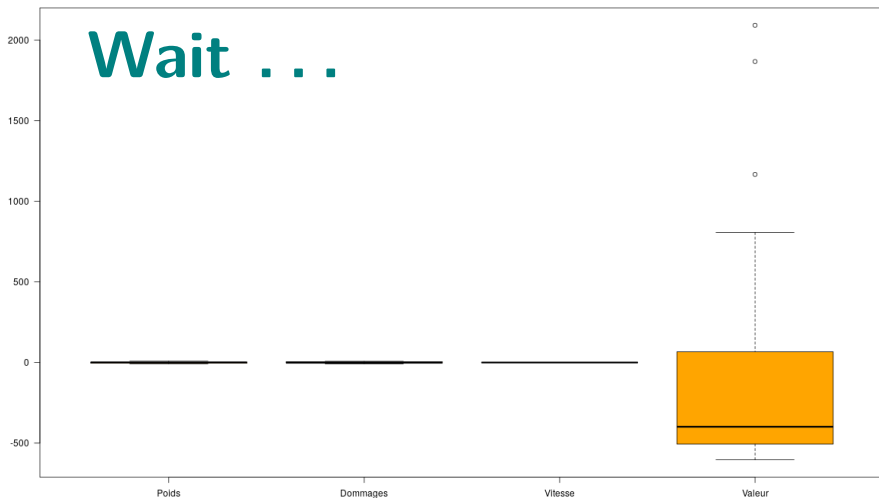
$$\frac{\lambda_1 + \lambda_2}{I(x)} = 99.99\%.$$

Oh yeah ! Go, biplot !

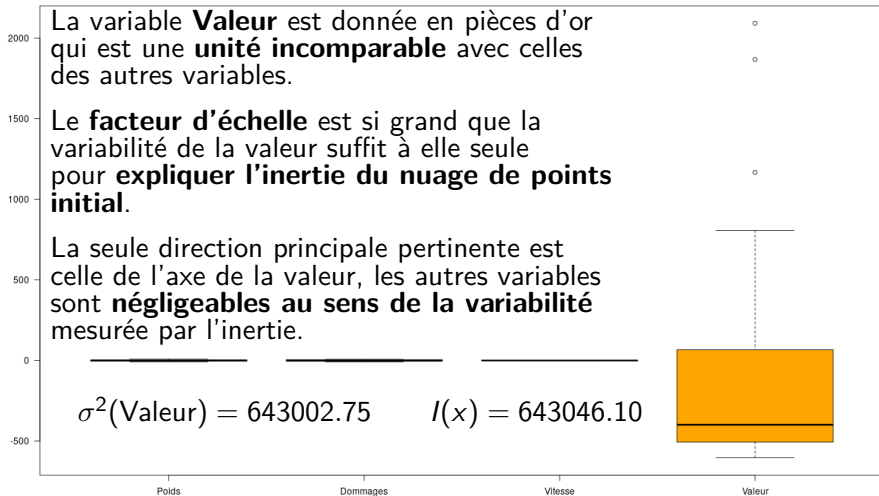
Exemple (un peu plus sérieux mais pas trop)



Exemple (un peu plus sérieux mais pas trop)



Exemple (un peu plus sérieux mais pas trop)



ACP et données réduites

La variance quantifie la variabilité des observations mais **sa valeur dépend de l'unité** utilisée. Un **changement d'échelle** modifie cette mesure,

$$\forall a \in \mathbb{R}, \sigma^2(ax) = a^2 \sigma^2(x).$$

L'inertie comme **somme des variances** attribue donc une importance à chaque variable qui **dépend de l'unité physique**.

Avantages de la réduction :

- exprime les données dans une **échelle neutre**.
- évite qu'une variable concentre toute la variabilité.

Inconvénients de la réduction :

- l'information de l'unité de mesure est perdue.
- un bruit se retrouve avec une variance apparente égale à celle d'une variable informative.

ACP et données réduites

La variance quantifie la variabilité des observations mais **sa valeur dépend de l'unité** utilisée. Un **changement d'échelle** modifie cette mesure,

$$\forall a \in \mathbb{R}, \sigma^2(ax) = a^2 \sigma^2(x).$$

L'inertie comme **somme des variances** attribue donc une importance à chaque variable qui **dépend de l'unité physique**.

Faire l'ACP des données réduites revient à **diagonaliser la matrice de corrélation**,

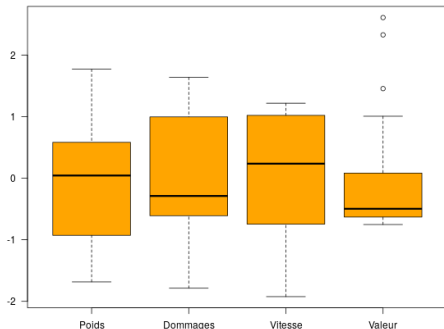
$$\begin{pmatrix} 1.0 & \rho(x^1, x^2) & \dots & \rho(x^1, x^p) \\ \rho(x^1, x^2) & 1.0 & \dots & \vdots \\ \vdots & \vdots & \dots & \rho(x^{p-1}, x^p) \\ \rho(x^1, x^p) & \rho(x^2, x^p) & \dots & 1.0 \end{pmatrix}$$

Exemple (un peu plus sérieux mais pas trop) avec réduction

La version centrée-réduite des $p = 4$ variables observées sur les $n = 20$ arcs et arbalètes conduit à considérer la matrice $\tilde{X}$ de taille $n \times p$ définie par

$$\forall k \in \{1, \dots, n\}, \forall \ell \in \{1, \dots, p\}, \tilde{X}_{k\ell} = \tilde{x}_k^\ell = \frac{x_k^\ell - \bar{x}^\ell}{\sigma(x^\ell)}.$$

Poids	Domages	Vitesse	Valeur
-1.68	-1.79	1.22	-0.75
-1.25	-1.57	0.82	-0.73
-1.25	-0.72	0.43	-0.54
-1.04	-0.50	0.43	-0.73
-1.04	-1.15	0.43	-0.68
-0.39	-0.50	0.43	-0.61
-0.82	-0.93	0.04	-0.60
0.47	-0.50	-0.35	-0.62
-0.60	-0.50	-0.35	-0.45
-0.17	-0.29	-0.75	-0.20
-0.39	-0.29	-0.75	-0.07
0.26	0.14	-1.14	0.23
0.69	0.57	-1.53	1.01
0.47	0.57	-1.53	1.45
1.12	1.00	-1.92	2.33
1.55	1.21	-0.35	2.61
0.26	1.00	1.22	-0.64
0.47	1.00	1.22	-0.54
1.55	1.64	1.22	-0.35
1.77	1.64	1.22	-0.10



Exemple (un peu plus sérieux mais pas trop) avec réduction

Matrice de corrélation $\tilde{\Sigma} = \tilde{X}^\top W \tilde{X}$ avec $W = n^{-1} Id_n$:

	Poids	Dommages	Vitesse	Valeur
Poids	1.00	0.93	-0.21	0.60
Dommages	0.93	1.00	-0.09	0.52
Vitesse	-0.21	-0.09	1.00	-0.67
Valeur	0.60	0.52	-0.67	1.00

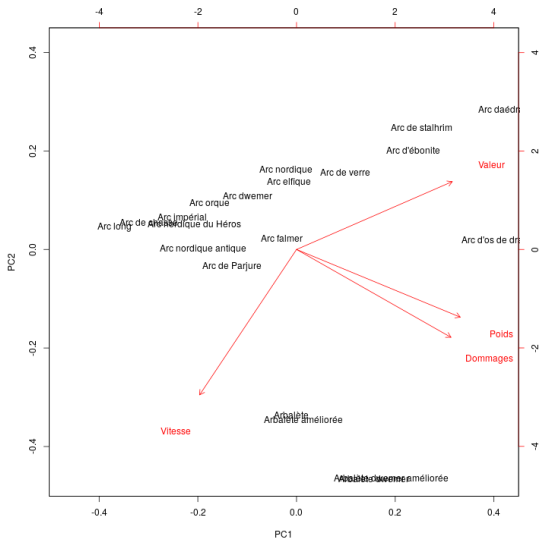
La diagonalisation de $\tilde{\Sigma}$ mène à la matrice $\tilde{V}$ des directions principales et aux valeurs propres

$$\tilde{\lambda}_1 = 2.57, \tilde{\lambda}_2 = 1.17, \tilde{\lambda}_3 = 0.20 \text{ et } \tilde{\lambda}_4 = 0.06.$$

Inertie du nuage de points initial : $I(\tilde{x}) = \text{Tr}(\tilde{\Sigma}) = 4$

Part d'inertie expliquée par le plan principal : $\frac{\tilde{\lambda}_1 + \tilde{\lambda}_2}{I(\tilde{x})} = 93.45\%$

Exemple (un peu plus sérieux mais pas trop) avec réduction



ACP et données réduites (point de vue dual)

La matrice des données réduites s'écrit également $\tilde{X} = XM^{1/2}$ où $M^{1/2}$ est la matrice **symétrique** et **positive** définie par

$$M^{1/2} = \begin{pmatrix} \frac{1}{\sigma(x^1)} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma(x^2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{\sigma(x^p)} \end{pmatrix}.$$

Dans le cas unidimensionnel de la recherche d'une direction $\tilde{v} \in \mathbb{R}^p$ de plus grande variabilité, nous avons vu qu'il fallait rendre maximale l'inertie des données projetées sur la droite $\mathbb{R}\tilde{v}$,

$$\tilde{v}^\top \tilde{\Sigma} \tilde{v} = \langle \tilde{\Sigma} \tilde{v}, \tilde{v} \rangle.$$

ACP et données réduites (point de vue dual)

Inertie des données projetées : $\tilde{v}^\top \tilde{\Sigma} \tilde{v} = \langle \tilde{\Sigma} \tilde{v}, \tilde{v} \rangle$.

En posant $\tilde{v} = M^{1/2} v$, cette inertie s'écrit

$$\begin{aligned}\tilde{v}^\top \tilde{\Sigma} \tilde{v} &= \tilde{v}^\top \tilde{X}^\top W \tilde{X} \tilde{v} \\ &= \left(M^{1/2} v\right)^\top \left(X M^{1/2}\right)^\top W \left(X M^{1/2}\right) \left(M^{1/2} v\right) \\ &= v^\top M (X^\top W X) M v = v^\top M \Sigma M v = \langle \Sigma M v, v \rangle_M\end{aligned}$$

avec, pour tout $u, v \in \mathbb{R}^p$, $\langle u, v \rangle_M = u^\top M v$.

Le **produit scalaire** $\langle \cdot, \cdot \rangle_M$ modifie la géométrie de $\mathbb{R}^p$ en donnant une **importance différente à chaque coordonnée** (i.e. à chaque variable) en fonction des variances observées. Ainsi, l'ACP calculée sur les données réduites n'est rien d'autre que l'**ACP des données initiales calculée avec une structure euclidienne** induite par la matrice M .

Exemple (plus sérieux)

Nous disposons de $p = 5$ variables morphologiques mesurées sur $n = 20$ individus contenant 10 hommes et 10 femmes.

TEpaule : tour d'épaules (cm)

TPoitrine : tour de poitrine (cm)

TTaille : tour de taille (cm)

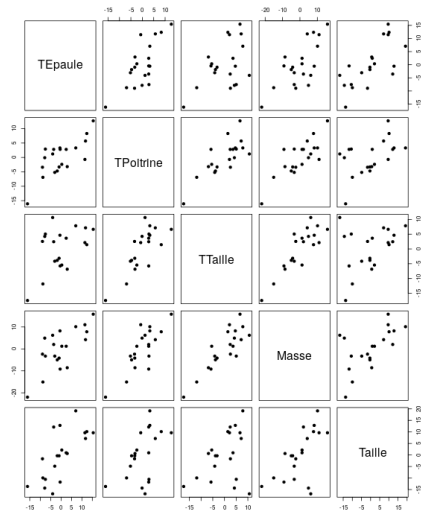
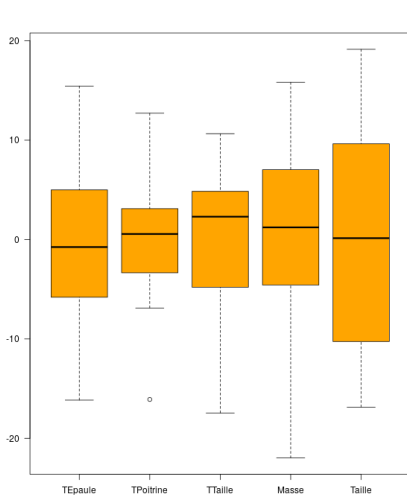
Masse : masse (kg)

Taille : taille (cm)

	TEpaule	TPoitrine	TTaille	Masse	Taille
H1	106.2	89.5	71.5	65.6	174.0
H2	110.5	97.0	79.0	71.8	175.3
H3	115.1	97.5	83.2	80.7	193.5
H4	104.5	97.0	77.8	72.6	186.5
H5	107.5	97.5	80.0	78.8	187.2
H6	119.8	99.9	82.5	74.8	181.5
H7	123.5	106.9	82.0	86.4	184.0
H8	120.4	102.5	76.8	78.4	184.5
H9	111.0	91.0	68.5	62.0	175.0
H10	119.5	93.5	77.5	81.6	184.0
F1	105.0	89.0	71.2	67.3	169.5
F2	100.2	94.1	79.6	75.5	160.0
F3	99.1	90.8	77.9	68.2	172.7
F4	107.6	97.0	69.6	61.4	162.6
F5	104.0	95.4	86.0	76.8	157.5
F6	108.4	91.8	69.9	71.8	176.5
F7	99.3	87.3	63.5	55.5	164.4
F8	91.9	78.1	57.9	48.6	160.7
F9	107.1	90.9	72.2	66.4	174.0
F10	100.5	97.1	80.4	67.3	163.8

Exemple (plus sérieux)

X : matrice des données **centrées**



Exemple (plus sérieux)

Matrice de covariance $\Sigma = X^\top W X$ avec $W = n^{-1} Id_n$:

	TEpaule	TPoitrine	TTaille	Masse	Taille
TEpaule	65.21	35.85	26.68	52.55	58.13
TPoitrine	35.85	35.64	32.20	43.42	30.78
TTaille	26.68	32.20	48.24	53.76	26.31
Masse	52.55	43.42	53.76	81.42	56.55
Taille	58.13	30.78	26.31	56.55	103.84

La diagonalisation de Σ mène à la matrice V des directions principales et aux valeurs propres

$$\lambda_1 = 242.87, \lambda_2 = 57.17, \lambda_3 = 22.31, \lambda_4 = 8.18 \text{ et } \lambda_5 = 3.81.$$

Inertie du nuage de points initial : $I(x) = 334.3$

Part d'inertie expliquée par le plan principal : $\frac{\lambda_1 + \lambda_2}{I(x)} = 89.74\%$

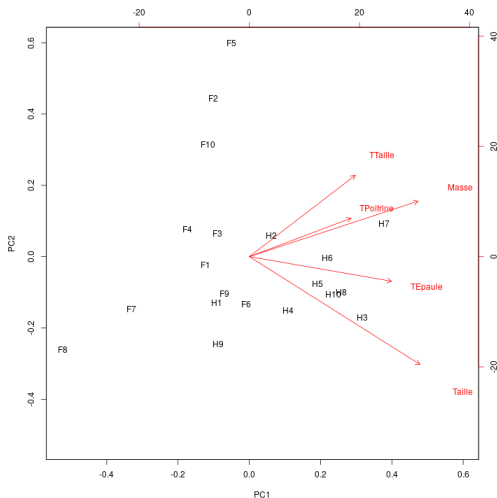
Exemple (plus sérieux)

PC1 : « Gabarit »

Sépare les grands (valeurs élevées pour les 5 variables) à droite et les petits à gauche.

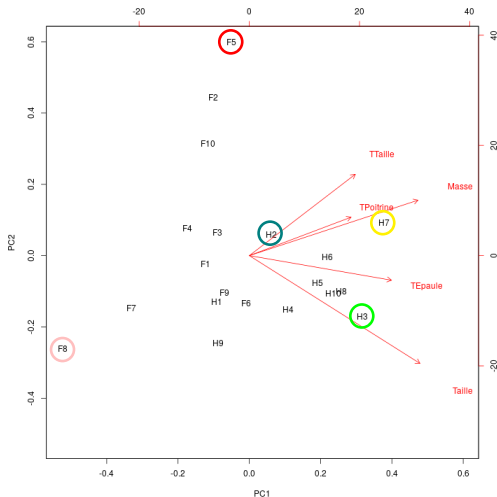
PC2 : « Embonpoint »

Sépare les variables liées à la taille et à la carrure (en bas) et celles liées à la masse et aux tours de taille et de poitrine (en haut).



Exemple (plus sérieux)

	TEpaule	TPoitrine	TTaille	Masse	Taille
H1	106.2	89.5	71.5	65.6	174.0
H2	110.5	97.0	79.0	71.8	175.3
H3	115.1	97.5	83.2	80.7	193.5
H4	104.5	97.0	77.8	72.6	186.5
H5	107.5	97.5	80.0	78.8	187.2
H6	119.8	99.9	82.5	74.8	181.5
H7	123.5	106.9	82.0	86.4	184.0
H8	120.4	102.5	76.8	78.4	184.5
H9	111.0	91.0	68.5	62.0	175.0
H10	119.5	93.5	77.5	81.6	184.0
F1	105.0	89.0	71.2	67.3	169.5
F2	100.2	94.1	79.6	75.5	160.0
F3	99.1	90.8	77.9	68.2	172.7
F4	107.6	97.0	69.6	61.4	162.6
F5	104.0	95.4	86.0	76.8	157.5
F6	108.4	91.8	69.9	71.8	176.5
F7	99.3	87.3	63.5	55.5	164.4
F8	91.9	78.1	57.9	48.6	160.7
F9	107.1	90.9	72.2	66.4	174.0
F10	100.5	97.1	80.4	67.3	163.8

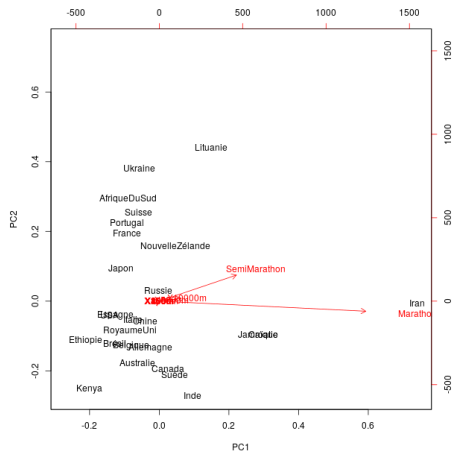
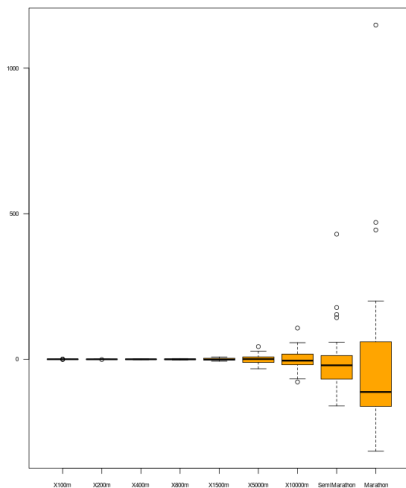


Exemple (un dernier)

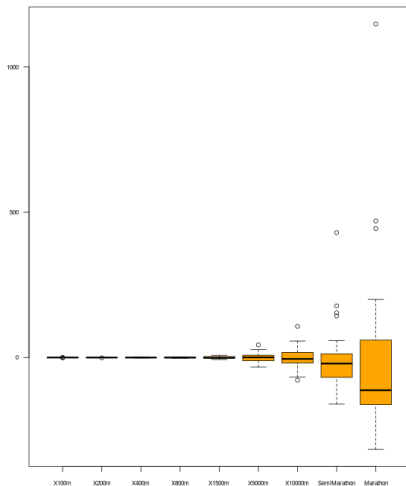
Voici $n = 26$ records nationaux (en sec.) de $p = 9$ épreuves d'athlétisme :

	X100m	X200m	X400m	X800m	X1500m	X5000m	X10000m	SemiMarathon	Marathon
Australie	9.93	20.06	44.38	104.40	211.96	775.76	1649.73	3602	7671
Belgique	10.02	20.19	44.78	103.86	214.13	769.71	1612.30	3605	7640
Brésil	10.00	19.89	44.29	101.77	213.25	799.43	1648.12	3573	7565
RoyaumeUni	9.87	19.87	44.36	101.73	209.67	780.41	1638.14	3609	7633
Canada	9.84	20.17	44.44	103.68	211.71	793.96	1656.01	3650	7809
Chine	10.17	20.54	45.25	106.44	216.49	805.14	1670.00	3635	7695
Croatie	10.25	20.76	45.64	104.07	213.30	817.76	1704.32	3827	8225
Ethiopie	10.50	21.08	45.89	106.08	211.13	757.35	1577.53	3535	7439
France	9.99	20.16	44.46	103.15	208.98	778.83	1642.78	3658	7596
Allemagne	10.06	20.20	44.33	103.65	211.58	774.70	1641.53	3634	7727
Inde	10.30	20.73	45.48	105.77	218.00	809.70	1682.89	3672	7920
Iran	10.29	21.11	46.37	104.74	218.80	833.40	1762.65	4103	8903
Italie	10.01	19.72	45.19	103.17	212.78	785.59	1636.50	3620	7642
Jamaïque	9.58	19.19	44.49	105.21	219.19	813.10	1712.44	3816	8199
Japon	10.00	20.03	44.78	106.18	217.42	793.20	1655.09	3625	7576
Kenya	10.26	20.43	44.18	102.01	206.34	759.74	1587.85	3513	7467
Lituanie	10.33	20.88	45.73	106.64	220.90	797.90	1651.50	3851	7955
NouvelleZélande	10.11	20.42	46.09	104.30	212.17	790.19	1661.95	3732	7815
Portugal	9.86	20.01	46.11	104.91	210.07	782.86	1632.47	3665	7596
Russie	10.10	20.23	44.60	102.47	212.28	791.99	1673.12	3675	7747
AfriqueDuSud	10.06	20.11	44.59	102.69	213.56	794.16	1649.94	3678	7593
Espagne	10.14	20.59	44.96	103.83	208.95	782.54	1634.44	3592	7562
Suède	10.18	20.30	44.56	105.54	216.49	797.59	1675.74	3655	7838
Suisse	10.16	20.41	44.99	102.55	211.75	787.54	1673.16	3686	7643
Ukraine	10.07	20.00	45.11	105.08	210.33	790.78	1679.80	3711	7635
USA	9.69	19.32	43.18	102.60	209.30	776.27	1633.98	3583	7538

Exemple (un dernier)



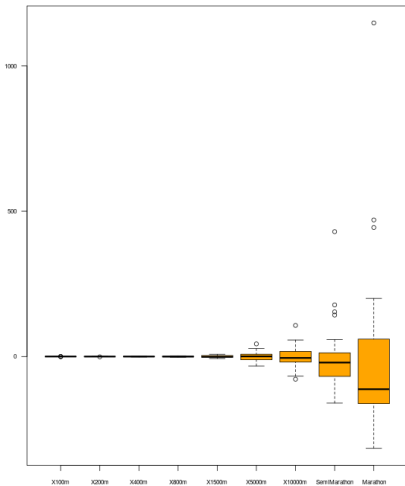
Exemple (un dernier)



Les épreuves de longue durée capturent presque toute la variabilité du jeu de données.

Qu'allons-nous perdre en réduisant les variances ?

Exemple (un dernier)

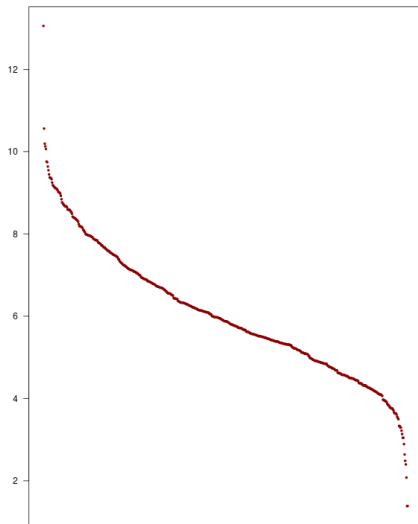
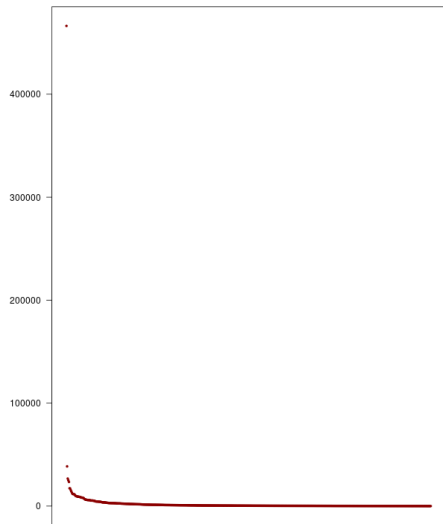


Les épreuves de longue durée capturent presque toute la variabilité du jeu de données.

Qu'allons-nous perdre en réduisant les variances ?

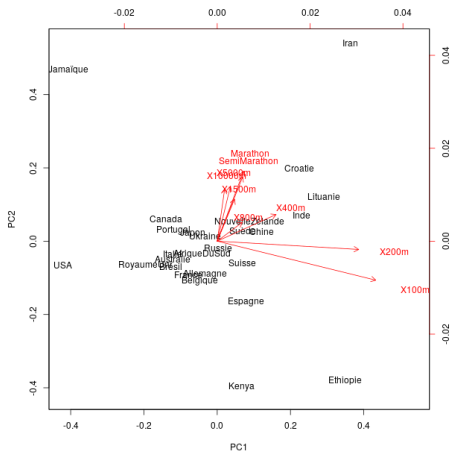
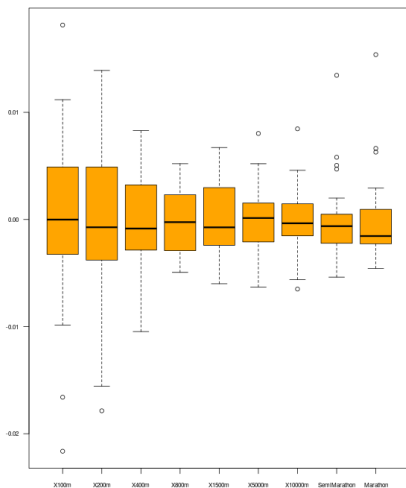
Les observations suggèrent une **structure d'échelle exponentielle** entre les différentes variables. Une réduction exprimerait les observations dans une même échelle neutre mais ferait également **disparaître cette structure entre les variables**. Il faut considérer une **transformation globale** plutôt que des transformations par variable.

Transformation logarithmique

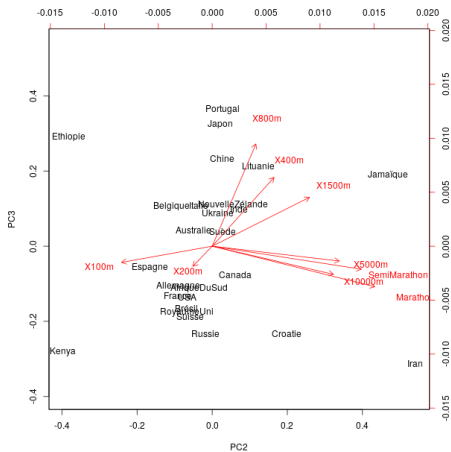
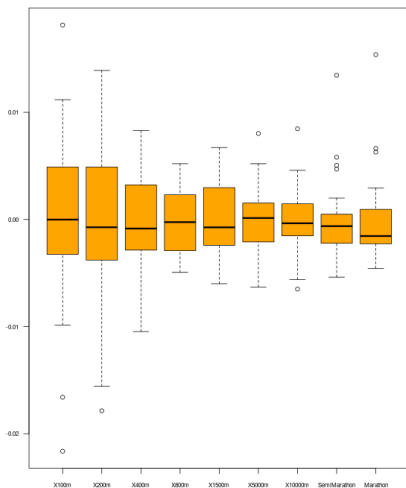


Données : population des 589 communes de Haute-Garonne en 2014, *INSEE*.

Exemple (un dernier) avec transformation logarithmique



Exemple (un dernier) avec transformation logarithmique



1.4 Analyse factorielle discriminante (AFD)

Motivation : visualiser des groupes

En 1962, le biologiste russe Lubischew a publié une étude de $n = 74$ co-léoptères issus de 3 espèces notées A, B et C. Pour chaque insecte, $p = 6$ variables morphologiques ont été mesurées.

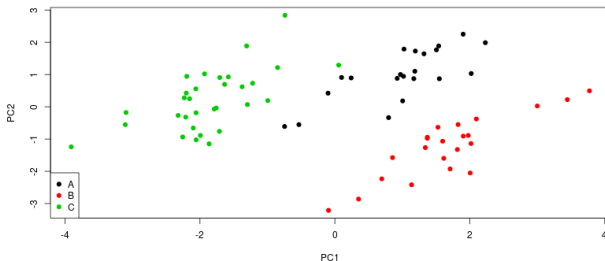
	V1	V2	V3	V4	V5	V6	Species
[1]	191	131	53	150	15	104	A
[2]	185	134	50	147	13	105	A
[3]	200	137	52	144	14	102	A
[4]	173	127	50	144	16	97	A
...							
[22]	158	141	58	145	8	107	B
[23]	146	119	51	140	11	111	B
[24]	151	130	51	140	11	113	B
[25]	122	113	45	131	10	102	B
...							
[44]	186	107	49	120	14	84	C
[45]	211	122	49	123	16	95	C
[46]	201	114	47	130	14	74	C
[47]	242	131	54	131	16	90	C
...							

Objectif : déterminer des combinaisons de variables qui permettent de **discriminer** au mieux les 3 espèces et donner une **représentation graphique** de cette discrimination.

Motivation : visualiser des groupes

Première idée naïve

Faire une ACP sur les observations (réduites) des 6 variables quantitatives et faire apparaître la variable catégorielle Species sur le résultat.



Est-ce la « meilleure » façon de faire ? La variable catégorielle est utilisée uniquement **a posteriori**, pouvons-nous utiliser sa connaissance de manière plus pertinente ?

Apprentissage supervisé

Le cadre statistique que nous considérons ici est celui de l'**apprentissage supervisé** dans lequel nous disposons de n observations de :

- p variables réelles $x^1, \dots, x^p$,
- 1 variable catégorielle t à valeurs dans $\{\tau_1, \dots, \tau_q\}$.

Autrement dit, les données sont de la forme suivante :

$$(x_1, t_1), \dots, (x_n, t_n) \in \mathbb{R}^p \times \{\tau_1, \dots, \tau_q\}.$$

Objectif

Étudier ou prédire la **modalité** de t en fonction des variables $x^1, \dots, x^p$.

Décomposition de la matrice de covariance

Matrice de covariance : $\Sigma = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(x_k - \bar{x})^\top$

Les observations de la variable t induisent une **partition** naturelle des observations en q groupes $G_1, \dots, G_q$ (supposés non vides),

$$\forall m \in \{1, \dots, q\}, G_m = \{k \in \{1, \dots, n\} \text{ tels que } t_k = \tau_m\}.$$

Pour chaque $m \in \{1, \dots, q\}$, nous pouvons définir le vecteur $\bar{x}_m \in \mathbb{R}^p$ des moyennes des observations du groupe G_m ,

$$\bar{x}_m = \begin{pmatrix} \bar{x}_m^1 \\ \vdots \\ \bar{x}_m^p \end{pmatrix} \text{ avec } \forall \ell \in \{1, \dots, p\}, \bar{x}_m^\ell = \frac{1}{n_m} \sum_{k \in G_m} x_k^\ell$$

où n_m est la taille du groupe G_m .

Décomposition de la matrice de covariance

Matrice de covariance : $\Sigma = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(x_k - \bar{x})^\top$

Faisons apparaître ces moyennes par groupe dans l'expression de la matrice de covariance,

$$\begin{aligned}\Sigma &= \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(x_k - \bar{x})^\top \\ &= \frac{1}{n} \sum_{m=1}^q \sum_{k \in G_m} ((x_k - \bar{x}_m) + (\bar{x}_m - \bar{x})) ((x_k - \bar{x}_m) + (\bar{x}_m - \bar{x}))^\top \\ &= \frac{1}{n} \sum_{m=1}^q \sum_{k \in G_m} (x_k - \bar{x}_m)(x_k - \bar{x}_m)^\top + \frac{1}{n} \sum_{m=1}^q \sum_{k \in G_m} (\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})^\top\end{aligned}$$

car $\sum_{k \in G_m} (x_k - \bar{x}_m) = 0$ par définition.

Décomposition de la matrice de covariance

Matrice de covariance : $\Sigma = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(x_k - \bar{x})^\top$

Ainsi,

$$\Sigma = \Sigma_w + \Sigma_b$$

où nous avons défini :

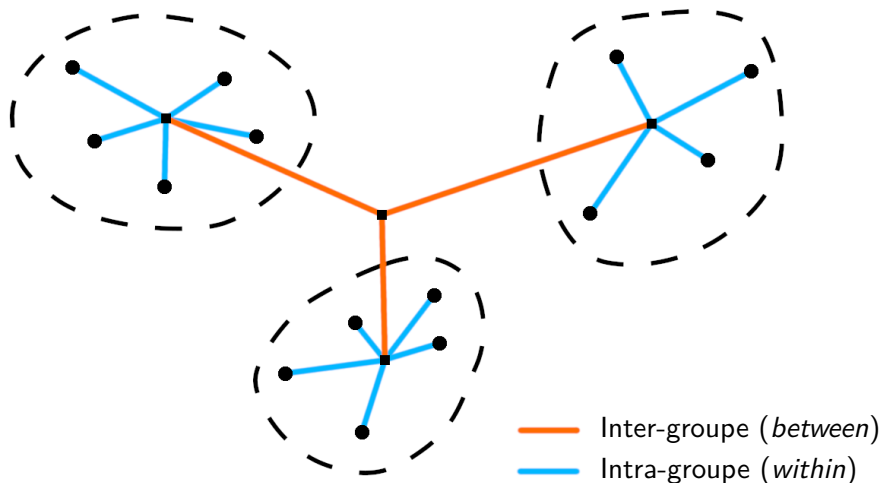
- la **matrice de covariance intra-groupe** (*within*)

$$\Sigma_w = \frac{1}{n} \sum_{m=1}^q n_m \times \frac{1}{n_m} \sum_{k \in G_m} (x_k - \bar{x}_m)(x_k - \bar{x}_m)^\top,$$

- la **matrice de covariance inter-groupe** (*between*)

$$\Sigma_b = \frac{1}{n} \sum_{m=1}^q n_m (\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})^\top.$$

Décomposition de la matrice de covariance (visuel)



Principe de l'AFD (cas unidimensionnel)

Nous avons vu dans la section consacrée à l'ACP que l'inertie des données projetées sur la droite engendrée par un vecteur $v \in \mathbb{R}^p$ unitaire est égale à

$$v^T \Sigma v = \underbrace{v^T \Sigma_w v}_{\text{Inertie intra-groupe}} + \underbrace{v^T \Sigma_b v}_{\text{Inertie inter-groupe}}$$

où la décomposition de l'inertie découle de celle de la matrice de covariance.

L'inertie intra-groupe quantifie la **variabilité à l'intérieur des groupes** et l'inertie inter-groupe celle **entre les groupes**.

Principe de l'AFD (cas unidimensionnel)

Nous avons vu dans la section consacrée à l'ACP que l'inertie des données projetées sur la droite engendrée par un vecteur $v \in \mathbb{R}^p$ unitaire est égale à

$$v^T \Sigma v = \underbrace{v^T \Sigma_w v}_{\text{Inertie intra-groupe}} + \underbrace{v^T \Sigma_b v}_{\text{Inertie inter-groupe}}$$

où la décomposition de l'inertie découle de celle de la matrice de covariance.

L'inertie intra-groupe quantifie la **variabilité à l'intérieur des groupes** et l'inertie inter-groupe celle **entre les groupes**.

Des **groupes homogènes et bien séparés** correspondent donc à une **petite inertie intra-groupe** et une **grande inertie inter-groupe**.

Principe de l'AFD (cas unidimensionnel)

Pour trouver une direction qui permette de **discriminer** au mieux les groupes d'observations $G_1, \dots, G_q$, il faut chercher à rendre simultanément **l'inertie inter-groupe maximale** et **l'inertie intra-groupe minimale**.

Pour cela, l'analyse factorielle discriminante consiste à déterminer un vecteur unitaire $v \in \mathbb{R}^p$ qui maximise le rapport de l'inertie inter-groupe sur l'inertie totale,

$$\frac{v^T \Sigma_b v}{v^T \Sigma v}.$$

Remarque : nous supposons dans la suite que la matrice de covariance Σ est inversible.

Encore un peu d'algèbre linéaire (ne dites pas non...)

Si $v \in \mathbb{R}^p$ est un vecteur propre de $\Sigma^{-1}\Sigma_b$ associé à la valeur propre λ , alors nous avons

$$\frac{v^\top \Sigma_b v}{v^\top \Sigma v} = \lambda.$$

Les valeurs propres de $\Sigma^{-1}\Sigma_b = \Sigma^{-1/2}(\Sigma^{-1/2}\Sigma_b)$ coïncident avec celles de $\Sigma^{-1/2}\Sigma_b\Sigma^{-1/2}$ qui est symétrique et positive puisque Σ_b est une matrice de covariance. Par conséquent, la matrice $\Sigma^{-1}\Sigma_b$ est diagonalisable et ses valeurs propres sont positives.

Par construction, Σ_b est engendrée par q vecteurs **centrés**, son rang vaut donc au plus $q - 1$. Il en va de même pour $\Sigma^{-1}\Sigma_b$ et les $q - 1$ valeurs propres (potentiellement) non triviales sont notées

$$\lambda_1 \geq \dots \geq \lambda_{q-1} \geq 0.$$

Encore un peu d'algèbre linéaire (ne dites pas non...)

Si $v \in \mathbb{R}^p$ est un vecteur propre de $\Sigma^{-1}\Sigma_b$ associé à la valeur propre λ , alors nous avons

$$\frac{v^T \Sigma_b v}{v^T \Sigma v} = \lambda.$$

La solution du problème de l'AFD unidimensionnel est donc donnée par le vecteur propre unitaire v^1 de la matrice $\Sigma^{-1}\Sigma_b$ associé à la plus grande valeur propre λ_1 .

Principe de l'AFD (cas général)

Nous nous intéressons maintenant à la recherche d'un espace E_d engendré par d vecteurs $v^1, \dots, v^d \in \mathbb{R}^p$ libres qui permette de discriminer au mieux les groupes d'observations $G_1, \dots, G_q$.

L'inertie des données projetées dans E_d s'exprime à l'aide d'un **déterminant** (admis) et le principe de l'AFD consiste à maximiser le rapport

$$\frac{\det(V^\top \Sigma_b V)}{\det(V^\top \Sigma V)}$$

où V est la matrice de taille $p \times d$ dont les colonnes sont données par les vecteurs $v^1, \dots, v^d$.

Principe de l'AFD (cas général)

Nous nous intéressons maintenant à la recherche d'un espace E_d engendré par d vecteurs $v^1, \dots, v^d \in \mathbb{R}^p$ libres qui permette de discriminer au mieux les groupes d'observations $G_1, \dots, G_q$.

Il est encore possible de montrer que la solution est donnée par les vecteurs propres $v^1, \dots, v^d$ de la matrice $\Sigma^{-1}\Sigma_b$ associés au d plus grandes valeurs propres $\lambda_1 \geq \dots \geq \lambda_d$.

Attention

La **dimension maximale est limitée** par le nombre de modalités de la variable catégorielle t . En effet, le rang de la matrice $\Sigma^{-1}\Sigma_b$ est au plus $q - 1$ et il ne peut pas y avoir plus de directions pour discriminer les q groupes.

En particulier, pour une variable binaire ($q = 2$), il n'y a qu'une seule direction discriminante **quel que soit le nombre p de variables explicatives**.

L'AFD est une variante d'ACP

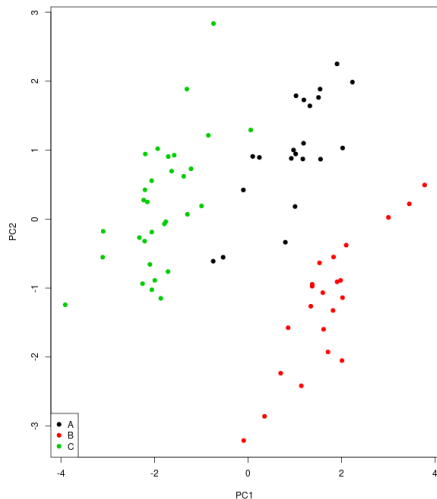
À un **reparamétrage** près, l'AFD peut se comprendre comme une variante de l'**ACP des vecteurs moyens** $\bar{x}_1, \dots, \bar{x}_m \in \mathbb{R}^p$.

Bien que ce point de vue soit souvent utilisé dans la littérature pour son apparente simplicité, la formulation du problème en ces termes demande **plus de technicité**. Il faut en particulier prendre le point de vue dual évoqué dans le cas de l'ACP sur les données réduites (avec $M = \Sigma^{-1}$ pour l'AFD).

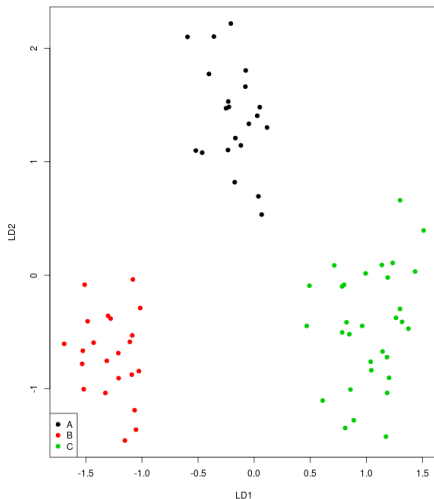
L'avantage principal de ce parallèle est de proposer une **représentation graphique** des données dans le plan engendré par les deux premières directions discriminantes, *i.e.* le plan qui permet de discriminer au mieux les groupes d'observations $G_1, \dots, G_q$.

Exemple des b b tes

ACP



AFD



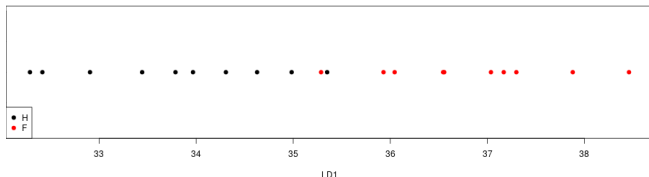
Exemple binaire

Reprenons les données morphologiques constituées de 10 hommes et 10 femmes. Les moyennes par groupe sont données dans le tableau suivant,

	TEpaule	TPoitrine	TTaille	Masse	Taille
Homme	113.80	97.23	77.88	75.27	182.55
Femme	102.31	91.15	72.82	65.88	166.17

La variable catégorielle est binaire, il n'y a donc au plus que 1 **seule direction discriminante**.

	LD1
TEpaule	0.119
TPoitrine	-0.022
TTaille	0.107
Masse	-0.112
Taille	0.139



Selon ces coefficients, la variable qui joue le rôle le plus faible dans la discrimination entre les hommes et les femmes est le tour de poitrine ...

AFD « décisionnelle »

L'AFD peut être utilisée dans un cadre d'apprentissage supervisé pour **prédire la modalité d'un nouvel individu** à partir des observations des variables quantitatives.

Une méthode simple pour affecter un nouvel individu à une classe donnée consiste à **utiliser le vecteur moyen dont il est le plus proche**. Cette approche souffre cependant de conservatisme et il existe des **règles de classification** plus complexes (plus proches voisins adaptatifs, ...).

Une faiblesse de l'AFD pour répondre à la prédiction d'une modalité apparaît lorsque le nombre de modalités est **faible** et celui des variables quantitatives est **élevé**. Par construction, il y aura peu de directions discriminantes et la méthode se retrouvera limitée. D'autres méthodes existent pour répondre à cette question (**arbres de décision, régression logistique, réseaux de neurones, ...**) et feront l'objet de cours à venir.

1.5 Classification

Problème de la classification

Objectif : regrouper des objets $x_1, \dots, x_n \in \mathcal{E}$ qui se « ressemblent ».



Problème de la classification

Objectif : regrouper des objets $x_1, \dots, x_n \in \mathcal{E}$ qui se « ressemblent ».

Le problème de la classification est **moins bien posé** que celui de l'apprentissage supervisé car **les classes ne sont pas connues a priori**.

Plusieurs questions se posent :

- que savons-nous de l'**espace** $\mathcal{E}$?
- existe-t-il une « **bonne** » **classification** ?
- connaissons-nous le **nombre de classes** a priori ?
- comment mesurons-nous la « **ressemblance** » ?
- pouvons-nous définir une notion de **similitude entre les objets** ?
- pouvons-nous définir une notion de **similitude entre des groupes d'objets** ?
- ...

Problème de la classification

Objectif : regrouper des objets $x_1, \dots, x_n \in \mathcal{E}$ qui se « ressemblent ».

Le problème de la classification est **moins bien posé** que celui de l'apprentissage supervisé car **les classes ne sont pas connues a priori**.

Il existe un (très) grand nombre de méthodes pour aborder ce problème de la classification. Certaines de ces méthodes feront l'objet d'autres cours et nous nous concentrons ici sur deux approches « classiques » :

- **Agrégation autour de centres mobiles** (*a.k.a.* K -means)
→ nombre de classes **connu** a priori
- **Classification ascendante hiérarchique**
→ nombre de classes **inconnu** a priori

Agrégation autour de centres mobiles

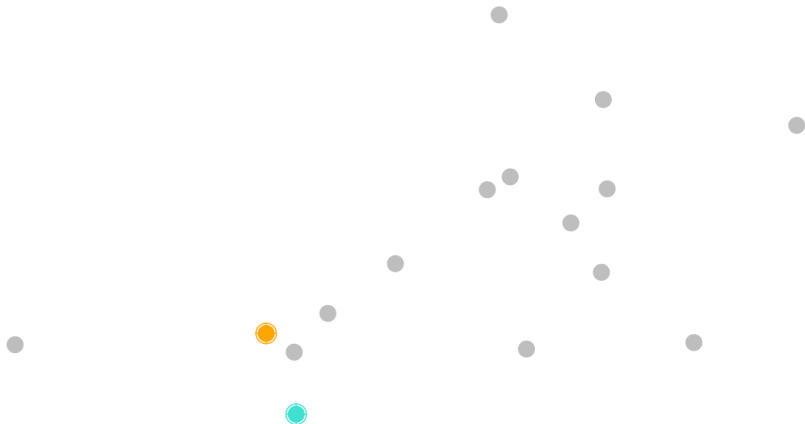
Prérequis : déterminer le **nombre K de classes** soit par une connaissance a priori du phénomène étudié, soit par une autre méthode (nous en reparlerons plus tard).

Algorithme

- ❶ Initialiser K centres distincts (tirages aléatoires ou choix imposés)
- ❷ Répéter les étapes suivantes :
 - Affecter chaque objet au centre le plus proche
 - Recalculer les centres de chaque groupe
- ❸ Terminer lorsque les objets ne changent plus de groupe entre 2 itérations successives

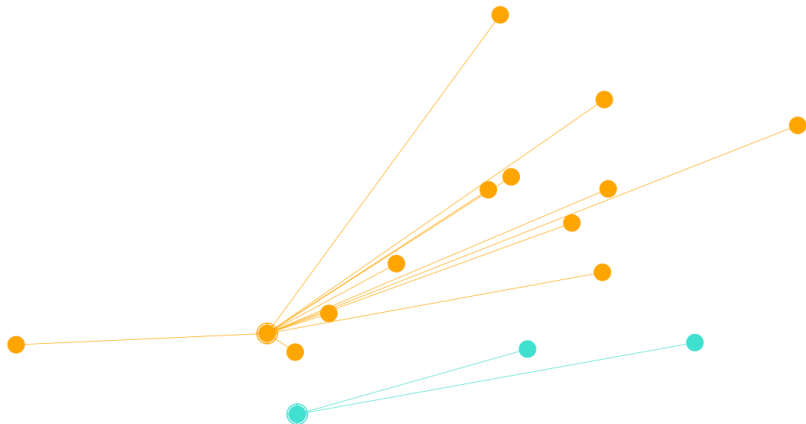
À l'issue de cet algorithme, nous obtenons une classification des données en K groupes.

Illustration de l'algorithme ($K = 2$)



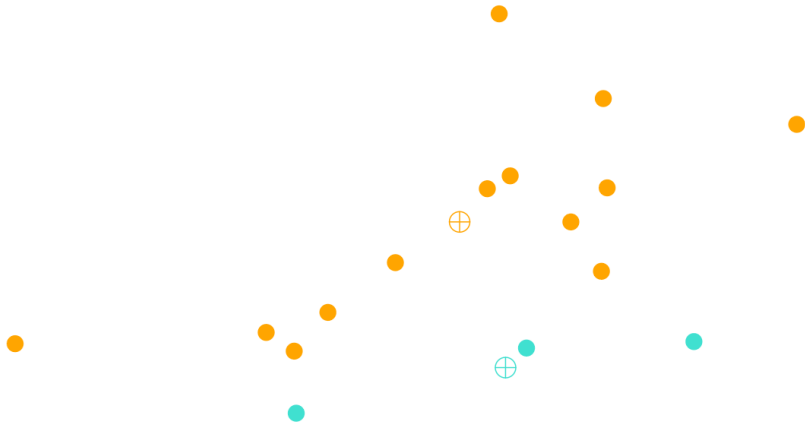
Initialisation de 2 centres aléatoires

Illustration de l'algorithme ($K = 2$)



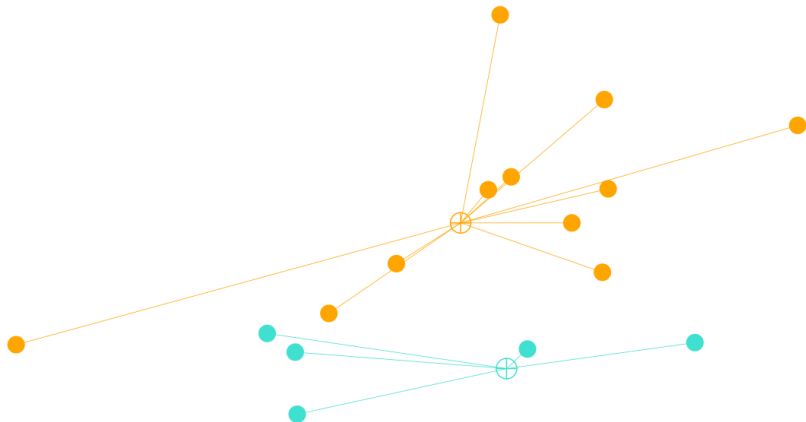
Affectation de chaque point au centre le plus proche

Illustration de l'algorithme ($K = 2$)



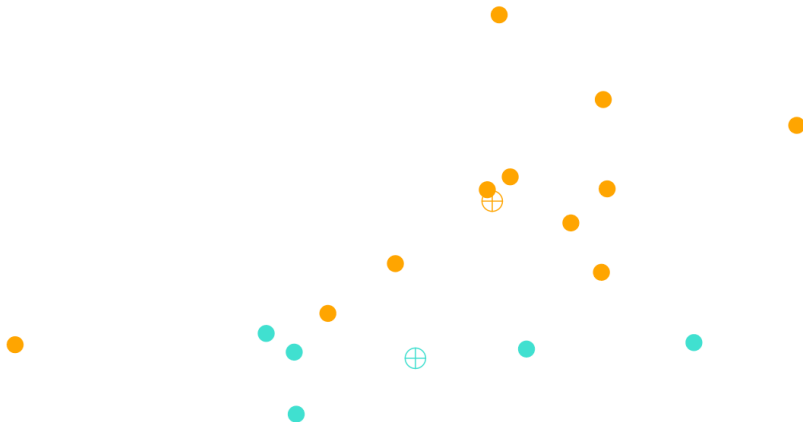
Mise à jour des centres

Illustration de l'algorithme ($K = 2$)



Affectation de chaque point au centre le plus proche (on itère ...)

Illustration de l'algorithme ($K = 2$)



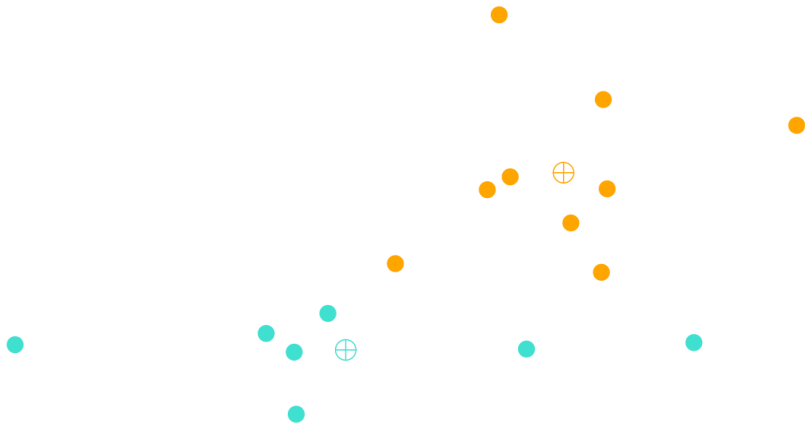
Mise à jour des centres (on itère ...)

Illustration de l'algorithme ($K = 2$)



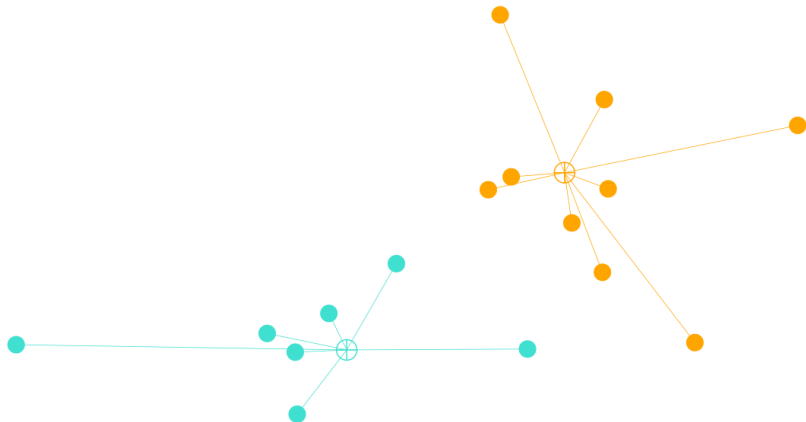
Affectation de chaque point au centre le plus proche (on itère ...)

Illustration de l'algorithme ($K = 2$)



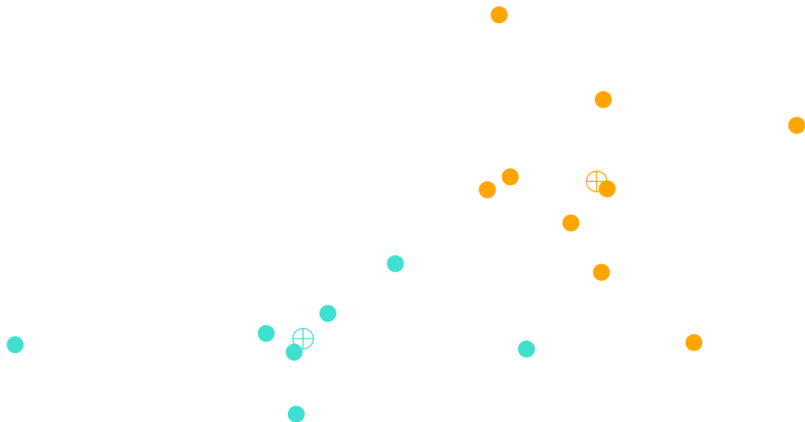
Mise à jour des centres (on itère ...)

Illustration de l'algorithme ($K = 2$)



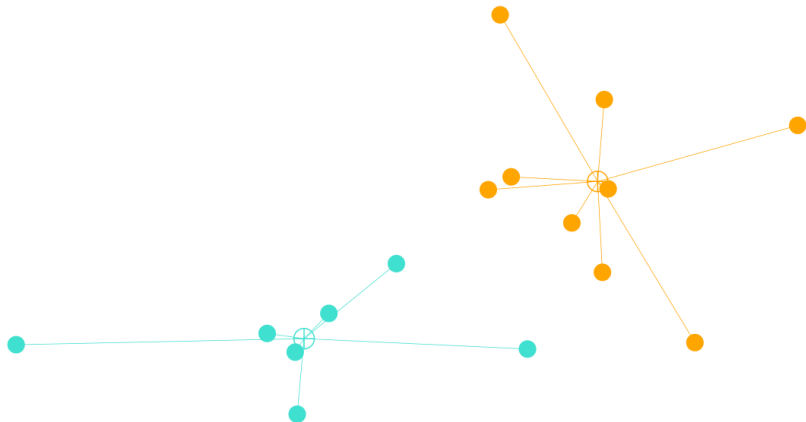
Affectation de chaque point au centre le plus proche (on itère ...)

Illustration de l'algorithme ($K = 2$)



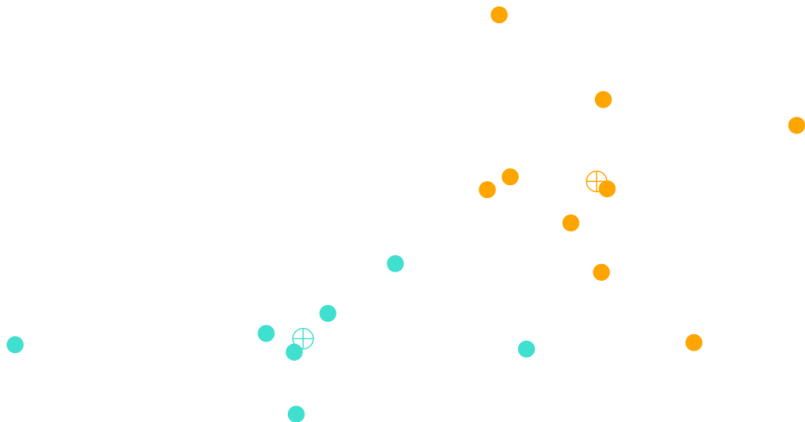
Mise à jour des centres (on itère ...)

Illustration de l'algorithme ($K = 2$)



Affectation de chaque point au centre le plus proche, rien ne change !

Illustration de l'algorithme ($K = 2$)



Classification finale des données en 2 groupes

Cas des centres explicites

Un cas particulier est celui où les **deux conditions suivantes** sont réunies :

- la **similitude** entre deux objets se mesure à l'aide d'une fonction $s : \mathcal{E}^2 \rightarrow \mathbb{R}$ (e.g. distance, variance, corrélation, ...),
- les centres $c_1, \dots, c_K$ des groupes respectifs $G_1, \dots, G_K$ peuvent être calculés **explicitement** tels que

$$\forall m \in \{1, \dots, K\}, c_m \text{ minimise } c \in \mathcal{E} \mapsto \frac{1}{n_m} \sum_{k \in G_m} s(x_k, c)$$

où n_m désigne la taille du groupe G_m .

Dans ce cadre, l'algorithme précédent est une **méthode de minimisation d'un critère de variabilité intra-groupe**.

$$(c_1, \dots, c_K) \in \mathcal{E}^K \mapsto \frac{1}{K} \sum_{m=1}^K \frac{1}{n_m} \sum_{k \in G_m} s(x_k, c_m).$$

Cas des centres explicites

Un cas particulier est celui où les **deux conditions suivantes** sont réunies :

- la **similitude** entre deux objets se mesure à l'aide d'une fonction $s : \mathcal{E}^2 \rightarrow \mathbb{R}$ (e.g. distance, variance, corrélation, ...),
- les centres $c_1, \dots, c_K$ des groupes respectifs $G_1, \dots, G_K$ peuvent être calculés **explicitement** tels que

$$\forall m \in \{1, \dots, K\}, c_m \text{ minimise } c \in \mathcal{E} \mapsto \frac{1}{n_m} \sum_{k \in G_m} s(x_k, c)$$

où n_m désigne la taille du groupe G_m .

Dans ce cadre, l'algorithme précédent est une **méthode de minimisation d'un critère de variabilité intra-groupe**.

Ce critère admet généralement des **minima locaux**. Comme dans le cas général, la solution trouvée peut **ne pas être optimale** et dépendre **fortement** de l'initialisation.

Initialisation de l'algorithme



Exemple de l'influence du **tirage aléatoire** des centres initiaux.

Initialisation de l'algorithme

Lorsque les centres initiaux sont tirés **au hasard**, des exécutions successives de l'algorithme peuvent conduire à des **classification différentes**. Pour minimiser l'impact de cette initialisation aléatoire, nous pouvons :

- relancer la procédure plusieurs fois et affecter les objets à une classe selon un principe de **vote majoritaire**,
- imposer un choix **non aléatoire** des centres initiaux (nous en reparlerons bientôt),
- renforcer la procédure en **imposant** à certains objets d'être toujours dans le **même groupe**.

Agrégation autour de centres mobiles (variantes)

L'algorithme des centres mobiles est simple à mettre en œuvre en pratique et il en existe plusieurs variantes :

- les centres peuvent être recalculés **après chaque affectation** d'un objet à un groupe, il s'agit des **nuées dynamiques**. Cette variante se stabilise plus rapidement mais accroît le **risque d'une solution sous optimale**.
- lorsque nous ne disposons pas des objets eux-mêmes mais seulement de la **matrice de similitude** S (i.e. les mesures $S_{kk'}$ des similitudes entre toutes les paires d'objets $(x_k, x_{k'}) \in \mathcal{E}^2$), alors les centres doivent être définis comme les **objets les plus « centraux »** des groupes selon un **critère de variabilité intra-groupe approché**,

$$\forall m \in \{1, \dots, K\}, c_m = x_{k_m} \text{ où } k_m \text{ minimise } k' \in G_m \mapsto \frac{1}{n_m} \sum_{k \in G_m} S_{kk'}.$$

• ...

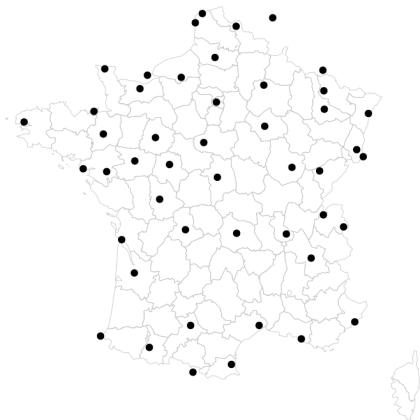
Exemple géographique (distances IGN)

Pour $n = 47$ villes de France ou frontalières, nous mesurons la « similitude » entre deux villes avec la distance IGN. Les données brutes correspondent donc à la matrice de similitude suivante,

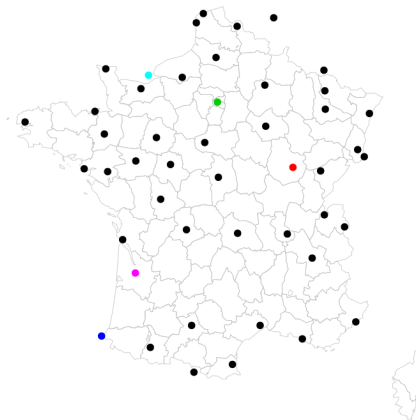
	Amiens	Andorre	Angers	Bâle	LaBaule	Besançon	Bordeaux	Boulogne	Bourges	Brest	Bruxelles	Caen	...
Amiens	0	1020	440	560	590	560	730	120	380	610	210	240	...
Andorre	1020	0	760	1130	830	970	430	1020	680	1130	1200	950	...
Angers	440	760	0	770	160	620	340	480	260	380	600	220	...
Bâle	560	1130	770	0	940	160	840	690	500	1090	560	800	...
LaBaule	590	830	160	940	0	770	400	550	430	270	760	350	...
Besançon	560	970	620	160	770	0	700	610	350	960	550	640	...
Bordeaux	730	430	340	840	400	700	0	830	400	620	890	580	...
Boulogne	120	1020	480	690	550	610	830	0	480	690	260	300	...
Bourges	380	680	260	500	430	350	400	480	0	630	550	360	...
Brest	610	1130	380	1090	270	960	620	690	630	0	910	370	...
Bruxelles	210	1200	600	560	760	550	890	260	550	910	0	450	...
Caen	240	950	220	800	350	640	580	300	360	370	450	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Selon les variantes envisagées, nous utiliserons uniquement ces distances ou les données GPS de chaque agglomération.

Exemple géographique ($K = 5$)

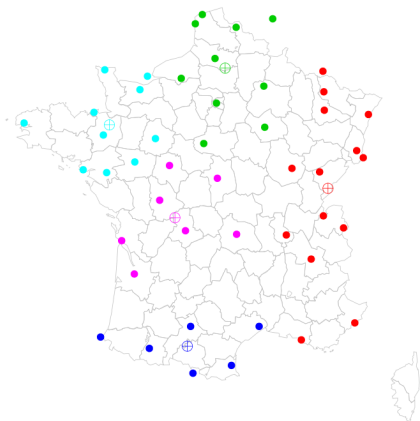


Positions des 47 villes

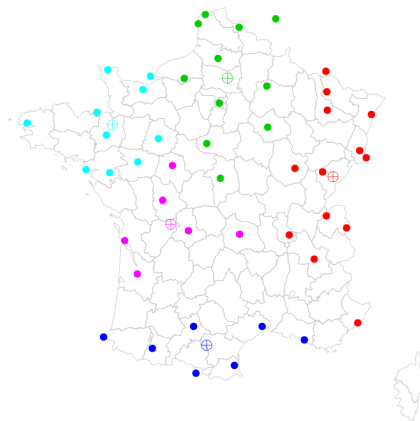


Initialisation de 5 centres aléatoires
(Dijon, Paris, Hendaye, Le Havre, Bordeaux)

Exemple géographique ($K = 5$)

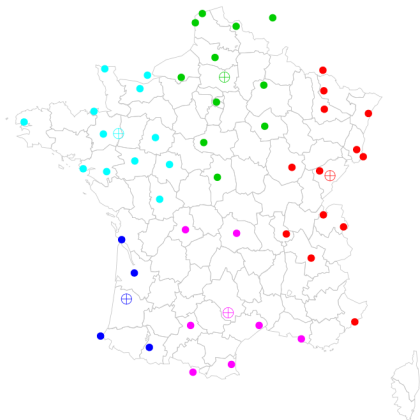


Centres mobiles classiques
(distance L^2 , 6 itérations)

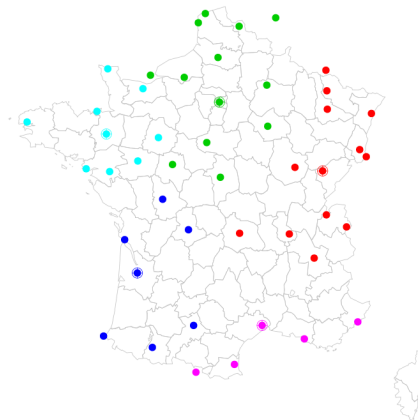


Centres mobiles classiques
(distance L^1 , 4 itérations)

Exemple géographique ($K = 5$)



Nuées dynamiques
(distance L^2 , 3 itérations)



Matrice de similitude uniquement
(5 itérations)

Classification ascendante hiérarchique (CAH)

Prérequis : choisir un **critère d'agglomération** pour donner un sens à la notion de **similitude entre des groupes d'objets**.

Algorithme

- 1 Initialiser n groupes « singletons » contenant chacun un objet
- 2 Répéter : regrouper les deux groupes les plus proches au sens du critère d'agglomération
- 3 Terminer lorsque il n'y a plus qu'un seul groupe contenant les n objets

À l'issue de cet algorithme, nous obtenons un diagramme appelé **dendrogramme** qui décrit les agglomérations effectuées.

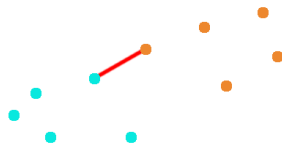
Critères d'agglomération (*linkage*)

Objectif : pour deux ensembles $A, B \subset \{1, \dots, n\}$ **disjoints**, nous voulons donner un sens à la **similitude entre les groupes d'objets**

$$\{x_k, k \in A\} \quad \text{et} \quad \{x_k, k \in B\}.$$

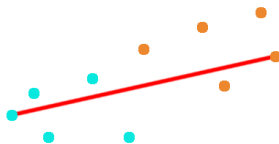
Si nous ne disposons que de la matrice S des similitudes $S_{kk'} = s(x_k, x_{k'})$ entre les objets x_k et $x_{k'}$ pour tout $k, k' \in \{1, \dots, n\}$, alors nous pouvons définir les critères suivants :

Minimum (*single*)



$$\min_{k \in A, k' \in B} S_{kk'}$$

Maximum (*complete*)



$$\max_{k \in A, k' \in B} S_{kk'}$$

Moyen (*average*)



$$\frac{1}{n_A n_B} \sum_{k \in A, k' \in B} S_{kk'}$$

Critères d'agglomération (*linkage*)

Objectif : pour deux ensembles $A, B \subset \{1, \dots, n\}$ **disjoints**, nous voulons donner un sens à la **similitude entre les groupes d'objets**

$$\{x_k, k \in A\} \quad \text{et} \quad \{x_k, k \in B\}.$$

Si il est possible de manipuler la **fonction de similitude** s et de **calculer explicitement** les « objets moyens » $c_A, c_B \in \mathcal{E}$ des deux groupes d'objets, alors la similitude entre ces centres peut être utilisée,

$$s(c_A, c_B).$$

Ce critère ne tient pas compte des tailles n_A et n_B des groupes, ce qui rend son interprétation difficile. Cette similitude peut être renormalisée de façon à correspondre à la **perte d'inertie inter-groupe** associée au regroupement de A et B . Cette méthode est très utilisée en pratique et s'appelle le **critère de Ward**,

$$\frac{n_A n_B}{n_A + n_B} s(c_A, c_B).$$

Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Nous disposons des similitudes entre $n = 5$ objets (données brutes ou calcul avec une fonction s).

	I1	I2	I3	I4	I5
I1	.				
I2	3.61	.			
I3	5.10	2.24	.		
I4	10.34	8.12	7.28	.	
I5	3.00	2.00	4.12	8.60	.

Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Les deux objets les plus proches sont I2 et I5.

	I1	I2	I3	I4	I5
I1	.				
I2	3.61	.			
I3	5.10	2.24	.		
I4	10.34	8.12	7.28	.	
I5	3.00	2.00	4.12	8.60	.

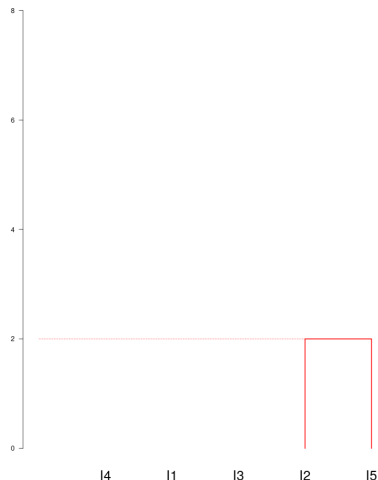


Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Nous regroupons les objets I2 et I5 dans un **nœud** N1 et nous recalculons les similitudes avec ce nouveau groupe à l'aide du critère d'agglomération.

	I1	I3	I4	N1
I1	.			
I3	5.10	.		
I4	10.34	7.28	.	
N1	3.00	2.24	8.12	.

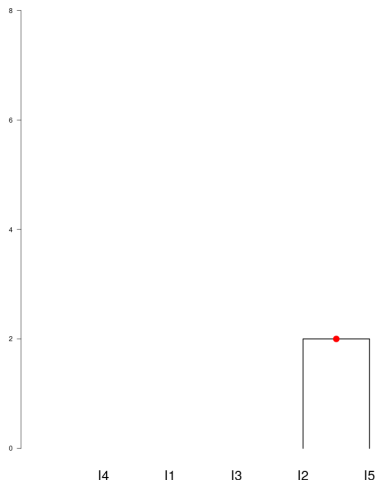


Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

La plus faible similitude est maintenant celle entre I3 et N1.

	I1	I3	I4	N1
I1	.			
I3	5.10	.		
I4	10.34	7.28	.	
N1	3.00	2.24	8.12	.

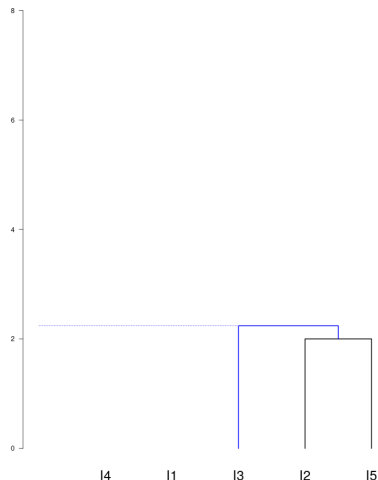


Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Le groupe singleton $\{I3\}$ et le groupe N1 sont regroupés dans un nœud N2 et les similitudes sont mises à jour.

	I1	I4	N2
I1	.		
I4	10.34	.	
N2	3.00	7.28	.

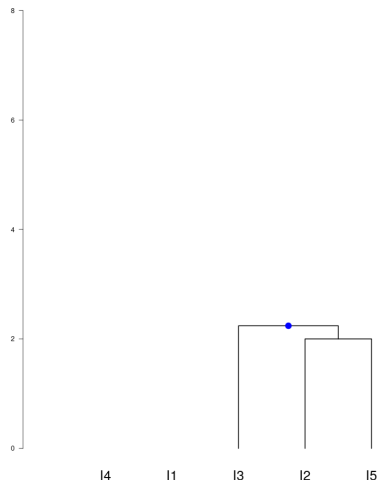


Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

La plus petite similitude est observée pour I1 et N2.

	I1	I4	N2
I1	.		
I4	10.34	.	
N2	3.00	7.28	.

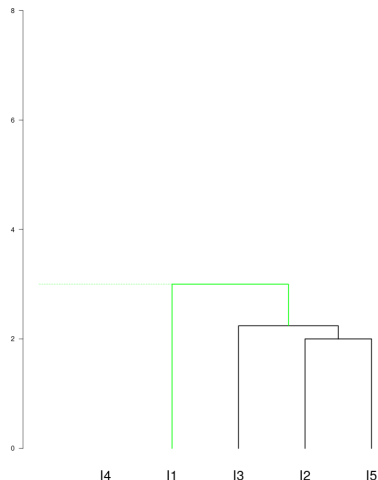


Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

La dernière similitude est celle entre l'objet I4 et le groupe N3 formé par les objets I1, I2, I3 et I5.

	I4	N3
I4	.	
N3	7.28	.

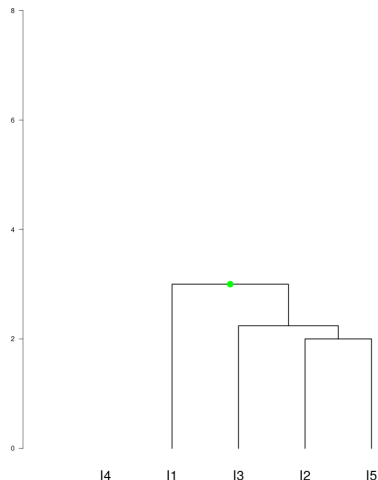


Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Pour terminer l'algorithme, il suffit de regrouper tous les objets dans un dernier nœud N4 à la hauteur 7.28.

	I4	N3
I4	.	
N3	7.28	.

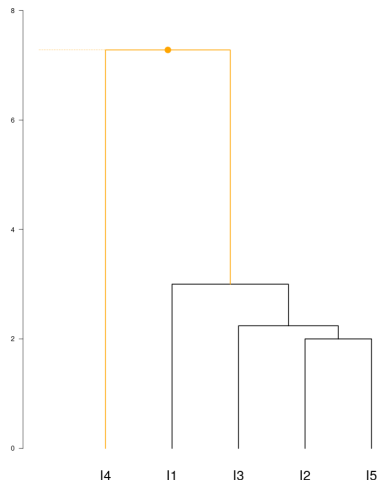
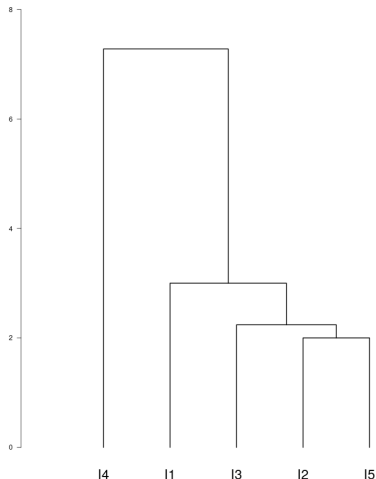


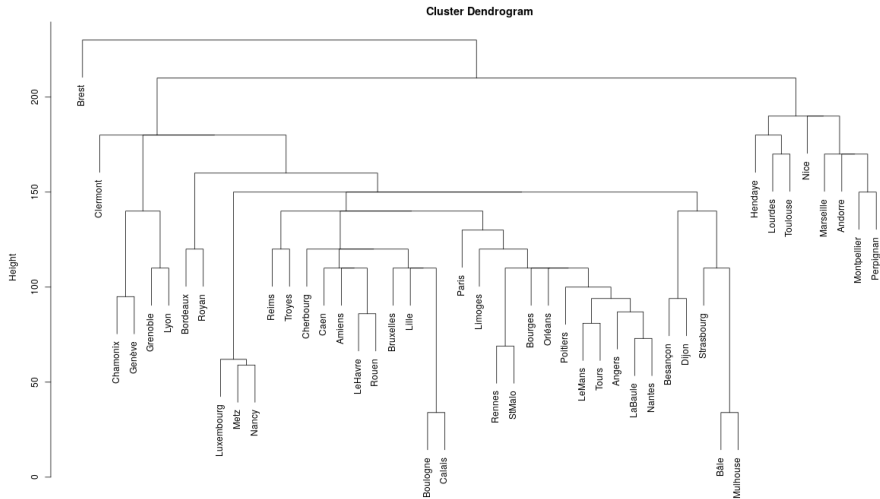
Illustration de l'algorithme

Critère d'agglomération : minimum (*single linkage*)

Le **dendrogramme** ainsi obtenu rend compte des différentes étapes de regroupement ainsi que des hauteurs des **sauts de similitudes** effectués.

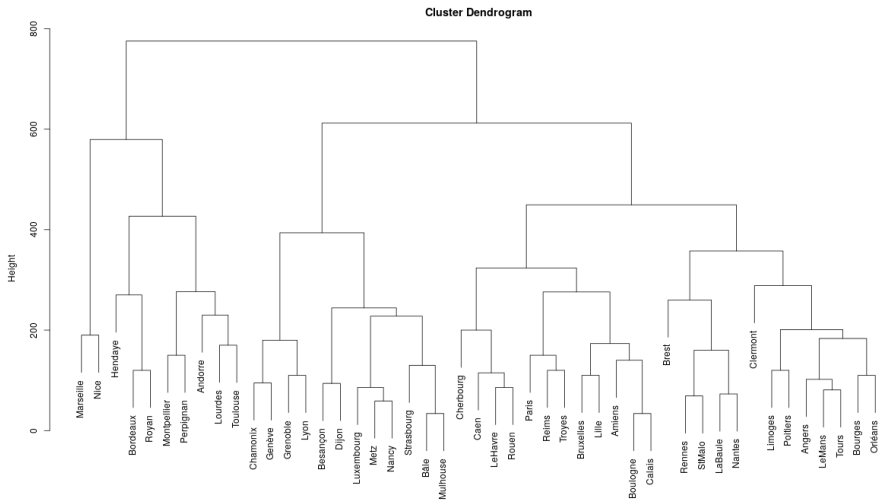


Exemple géographique (distances IGN)



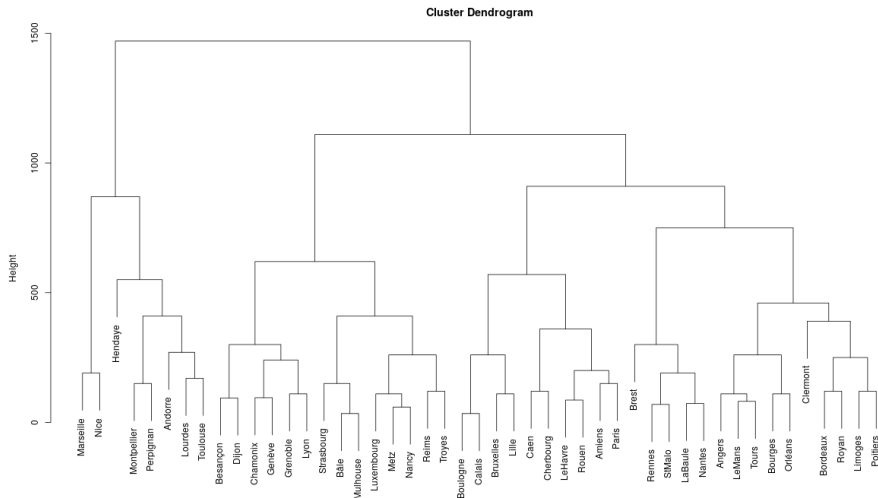
Critère d'agglomération minimum (*single linkage*)

Exemple géographique (distances IGN)



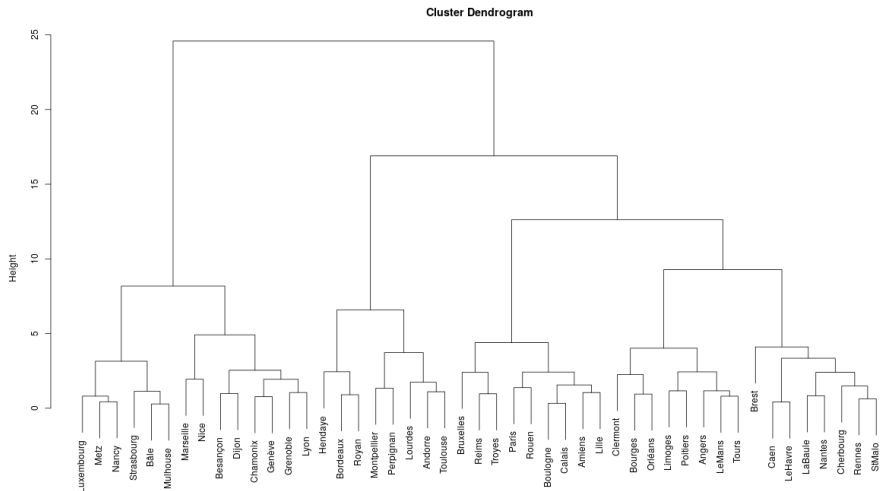
Critère d'agglomération moyen (*average linkage*)

Exemple géographique (distances IGN)



Critère d'agglomération maximum (*complete linkage*)

Exemple géographique (distances GPS)



Critère d'agglomération de Ward

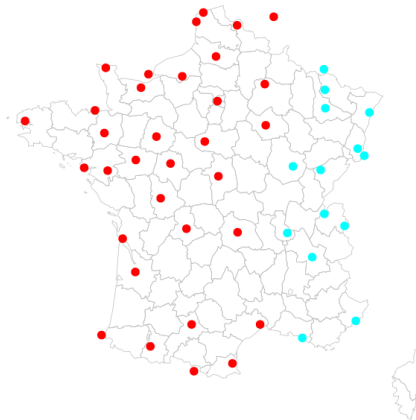
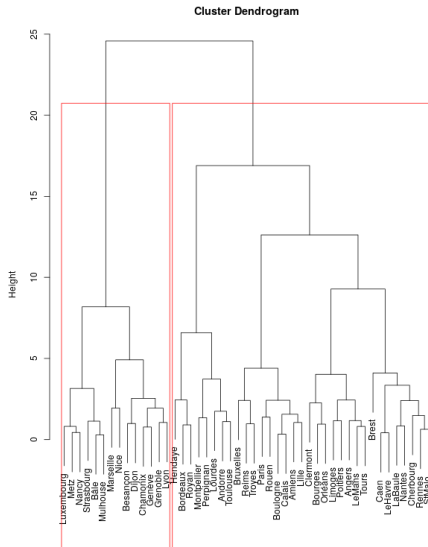
Utilisation d'un dendrogramme

Afin de déduire une classification des objets initiaux à partir d'un dendrogramme, il faut se donner une **hauteur de coupe**. Les groupes objets donnés par les « branches » obtenues forment les classes.

Plus le dendrogramme est coupé haut, plus la classification est grossière, *i.e.* peu de classes voire même une seule contenant tous les objets.

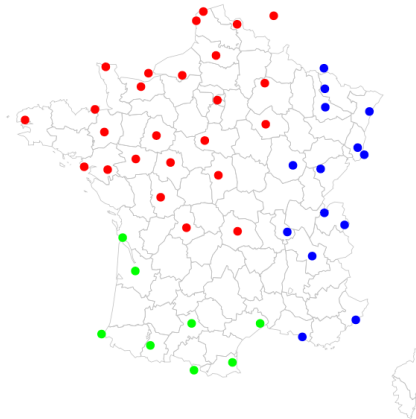
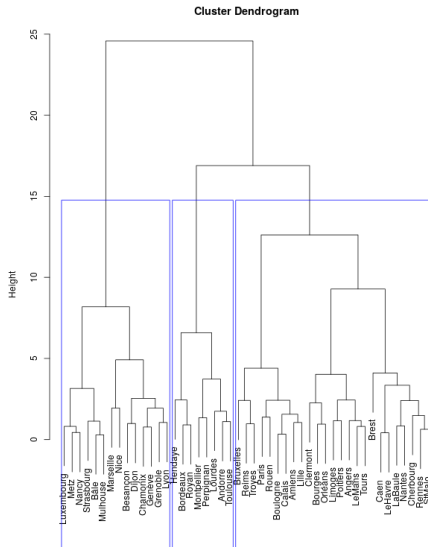
Une hauteur de coupe est pertinente si elle se trouve entre deux nœuds séparés par une hauteur relativement « grande ». Avec le critère de Ward, cela s'interprète comme une **part d'inertie inter-groupe expliquée** similaire à ce que nous avons manipulé dans le cadre de l'ACP.

Exemple géographique (distances GPS, critère de Ward)



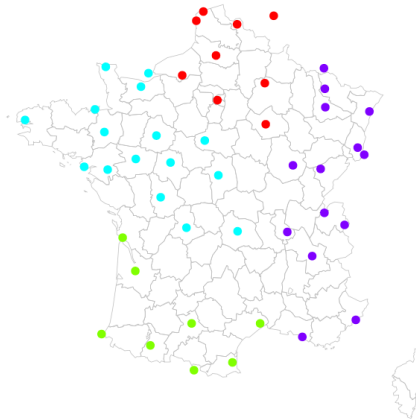
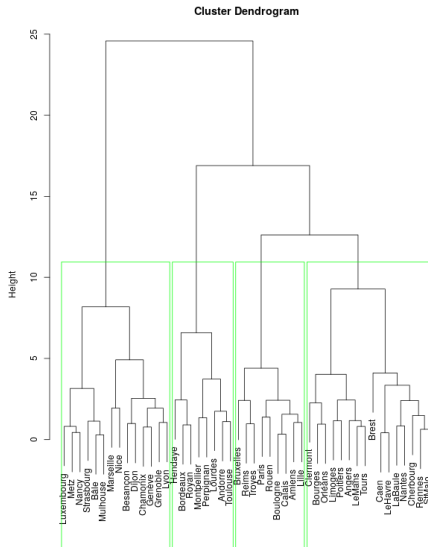
2 classes

Exemple géographique (distances GPS, critère de Ward)



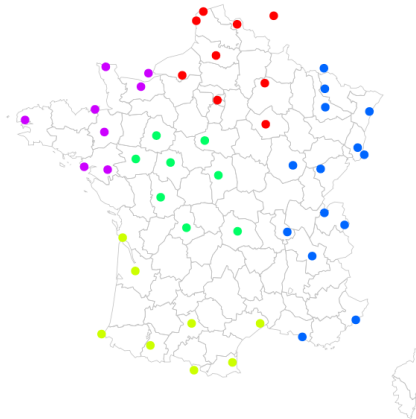
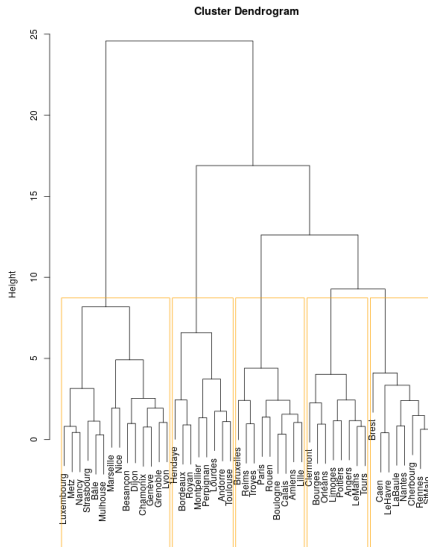
3 classes

Exemple géographique (distances GPS, critère de Ward)



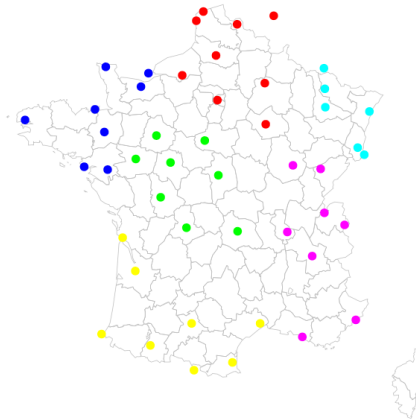
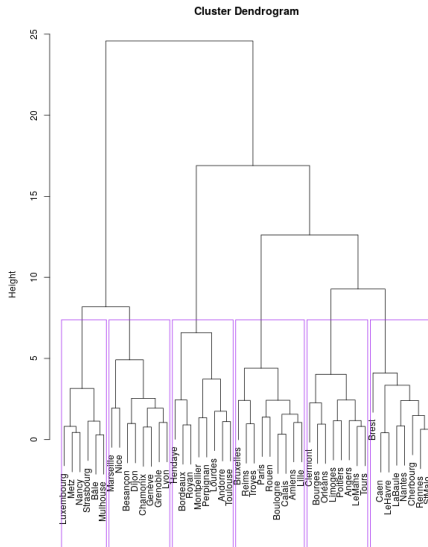
4 classes

Exemple géographique (distances GPS, critère de Ward)



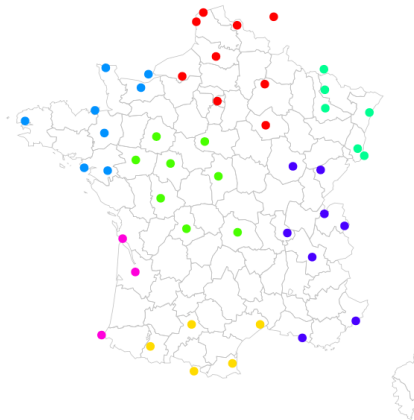
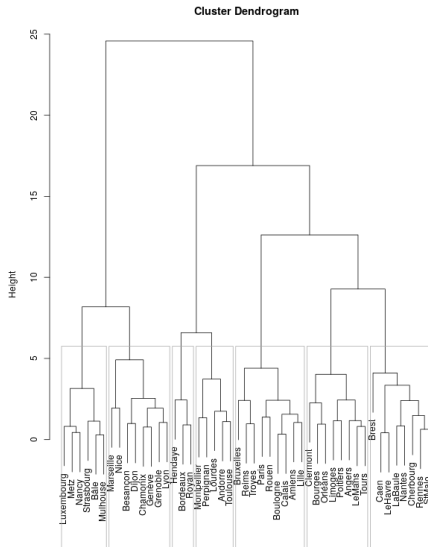
5 classes

Exemple géographique (distances GPS, critère de Ward)



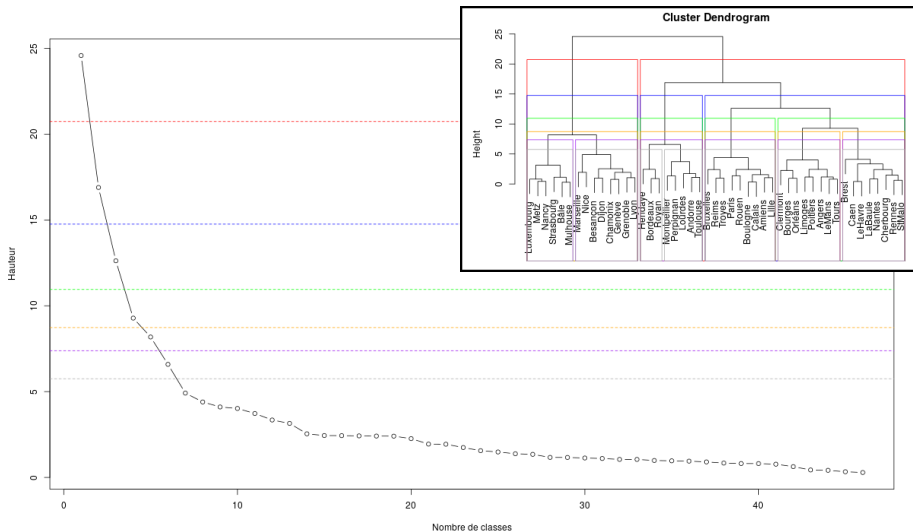
6 classes

Exemple géographique (distances GPS, critère de Ward)



7 classes

Exemple géographique (distances GPS, critère de Ward)



Centres mobiles versus CAH

Centres mobiles

Avantages :

- Classification **robuste** car minimum d'un critère
- Algorithme simple à mettre en œuvre et rapide

Inconvénients :

- Demande de connaître le nombre K de classes

CAH

Avantages :

- Ne nécessite pas de connaître le nombre K de classes

Inconvénients :

- Classification **sensible aux données** car la topologie du dendrogramme dépend fortement des premiers regroupements
- Algorithme lourd en temps de calcul

Stabilisation

Pour tirer parti des avantages des deux méthodes, il est possible de les enchaîner en utilisant le résultat de l'une comme initialisation de la suivante.

- **Étape optionnelle** : si le nombre n d'objets à classer est trop important pour envisager une CAH, nous pouvons appliquer les centres mobiles avec un nombre $K_0 \ll n$ de classes **grand** et nous restreindre à ces K_0 groupes comme des objets élémentaires dans les étapes suivantes.
- **Étape CAH** : le dendrogramme permet de **déterminer un nombre de classes** K_1 « pertinent » d'après les données (critère d'inertie, ...) et de fournir une première classification **peu robuste**.
- **Étape Centres Mobiles** : pour **stabiliser** la classification obtenue à l'étape précédente, nous utilisons une agrégation autour de K_1 centres mobiles avec la classification de la CAH comme initialisation.