Classification non-supervisée

Thomas A Rieutord

Classification non-supervisée

Thomas Rieutord

Chercheur au CNRM (Météo-France, CNRS – UMR 3589)
Groupe de Météorologie pour l'Assimilation et la Prévision (GMAP)
Equipe RECYF (prévision d'ensemble, prévisibilité...)

Décembre 2021

Outline

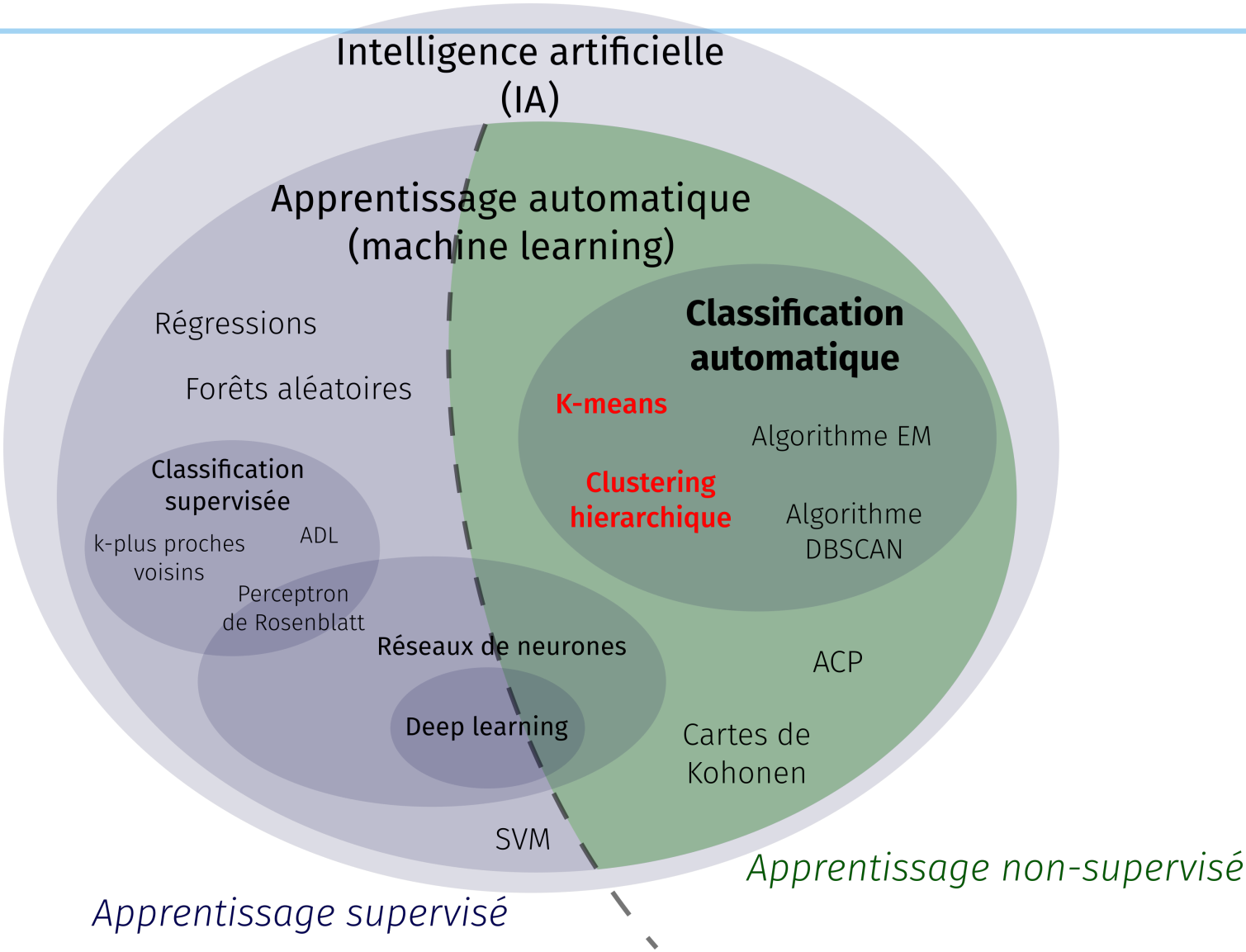
Outline

De quoi parle-t-on ?

Classification automatique : attribuer une classe à des individus, définis par des caractéristiques, à l'aide d'un algorithme.

Individu : un élément de notre échantillon. Ils sont représentés par un vecteur.
Caractéristique : quantité commune à tous les individus. Elles sont les composantes des vecteurs qui forment les individus.
Classe : groupement d'individus dont les caractéristiques sont proches.
Algorithme : suite d'opérations explicites permettant de résoudre un problème (ici : regrouper les individus en classes).

De quoi parle-t-on ?



La classification automatique dans le paysage de l'intelligence artificielle et du machine learning.

Régression vs classification

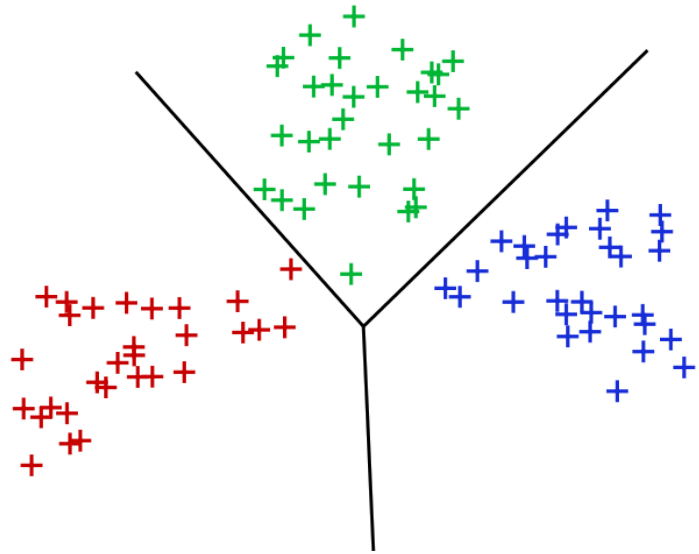
En statistiques, on cherche à expliquer une variable de sortie Y en fonction de variables d'entrée $X_1, ..., X_p$. Expliquer cette relation revient à trouver f telle que

$$Y = f(X_1, ..., X_p)$$

- Lorsque Y est une variable **continue**, on parle de **régression**.
- Lorsque Y est une varaible **discrète**, on parle de **classification**.

Supervisé vs non-supervisé

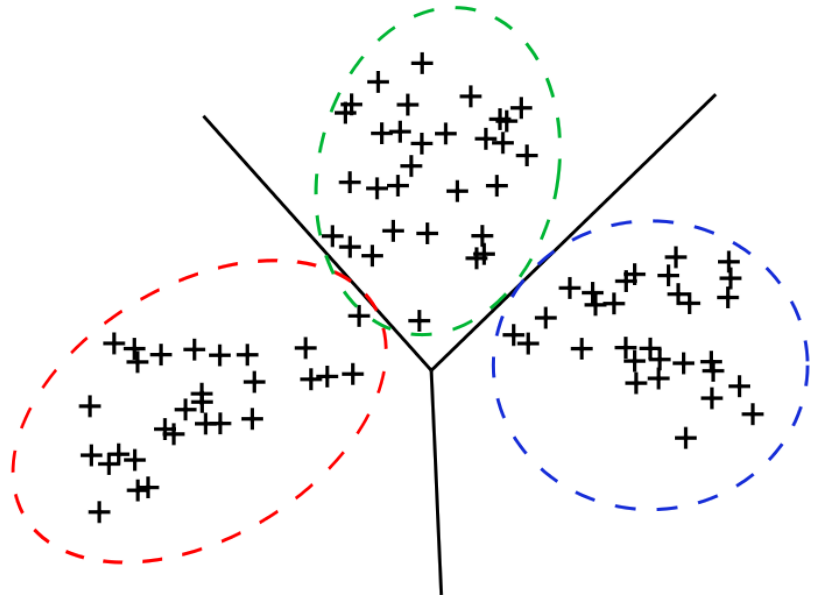
Supervisé



Trouver les meilleures frontières à partir de

- Position des points
- Couleur des points (référence)

Non-supervisé

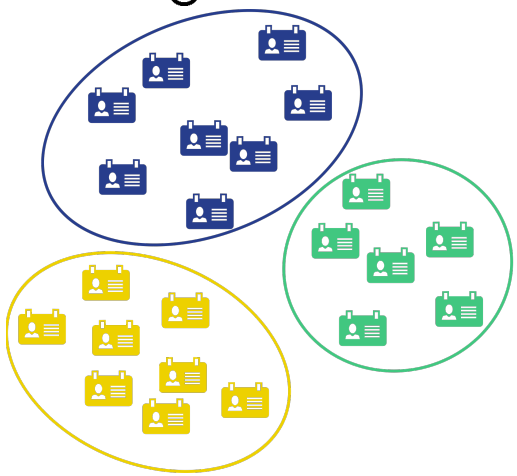


Trouver les zones de grandes densité dans

- Position des points SEULEMENT
- Pas de référence

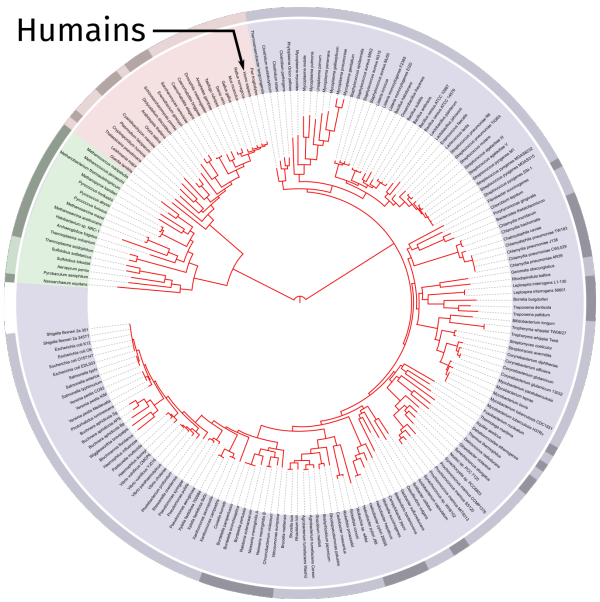
Pourquoi faire des classifications non-supervisées ?

- Profilage



Individus ? personnes physiques.
Caractéristiques ? données socio-économiques (âge, études...) + autres, éventuellement (historique conso, préférences).

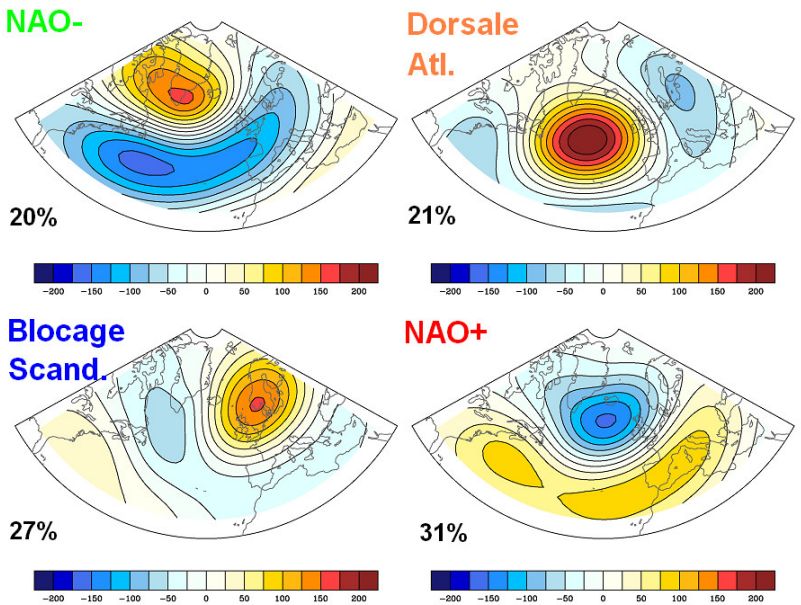
- Taxonomie (classification du vivant)



Individus ? espèces vivantes.
Caractéristiques ? gènes.

Pourquoi faire des classifications non-supervisées ?

- Régimes de temps



© Météo-France

Individus ? carte mensuelle de Z500.
Caractéristiques ? 10 premières composantes principales.

⇒ identification de schémas récurrents qui aident à prévoir le temps à moyenne échéance

De quoi parle-t-on ?

Classification automatique : attribuer une classe à des individus, définis par des caractéristiques, à l'aide d'un algorithme.

Individu (ou observation): un vecteur $x^i = (x_1^i, ..., x_p^i)$.
Caractéristique (ou prédicteur) : composante $j \in [\![1, p]\!]$ des vecteurs x^i.
Algorithme : retourne une attribution $C(i) = k$.
Classe : ensemble d'objets $C_k = \{i, C(i) = k\}$.
On dispose de N individus ayant tous p caractéristiques :

$$X = \begin{pmatrix} x_1^1 & \dots & x_p^1 \\ \vdots & & \vdots \\ x_1^N & \dots & x_p^N \end{pmatrix} \in \mathbb{R}^{N \times p}$$

Quelques notations

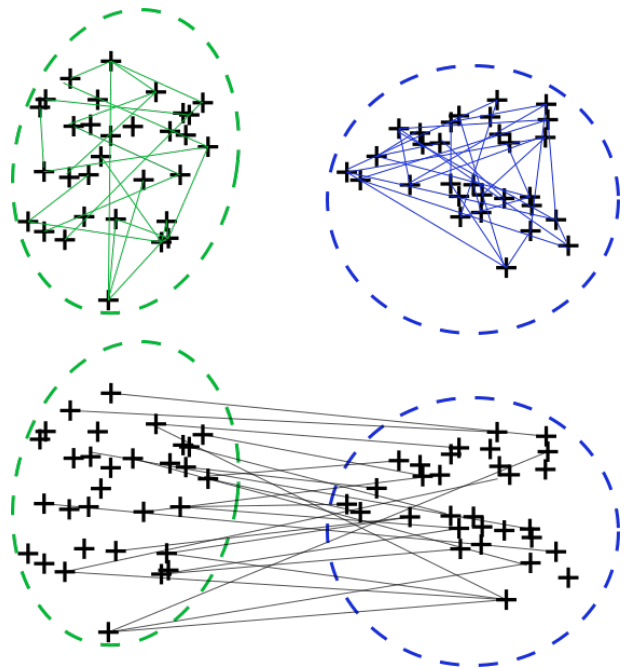
On cherche à grouper N individus en K groupes.
$\Longrightarrow$ trouver une fonction d'attribution C (i.e $C(i) = k$) qui minimise une fonction coût $W(C)$.

- Dispersion "within cluster" :

$$W(C) = \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x^i, x^{i'})$$

- Dispersion "between cluster":

$$B(C) = \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d(x^i, x^{i'})$$



- Dispersion "totale": $T = \sum_{i=1}^{N} \sum_{i'=1}^{N} d(x^i, x^{i'}) = W(C) + B(C)$

où $d(x^i, x^{i'})$ mesure la **dissimilarité entre** x^i **et** $x^{i'}$.

Outline

Dissimilarité entre caractéristiques

- **Variable quantitative**

Exemple: $X_j \in [0, +\infty[$ ou $X_j \in [-1, 1]$...
Toute distance convient. Par exemple, si l est une fonction strictement croissante, on peut prendre

$$d_j(x_j^i, x_j^{i'}) = l(|x_j^i - x_j^{i'}|)$$

La plus répandue est l'écart au carré :

$$d_j(x_j^i, x_j^{i'}) = (x_j^i - x_j^{i'})^2$$

Dissimilarité entre caractéristiques

- **Variable ordinale** (catégories ordonnées)

Exemple: $X_j \in$ {très mauvais,mauvais,bien,très bien}...
Les valeurs prises par X_j sont numérotées de 1 à M. On remplace chaque élément de la colonne j par

$$\frac{k - 1/2}{M}$$

où $k \in [\![1, M]\!]$ est le rang de X_j parmis les valeurs prises par X_j.
On traite ensuite cette variable tranformée comme une variable quantitative[1].

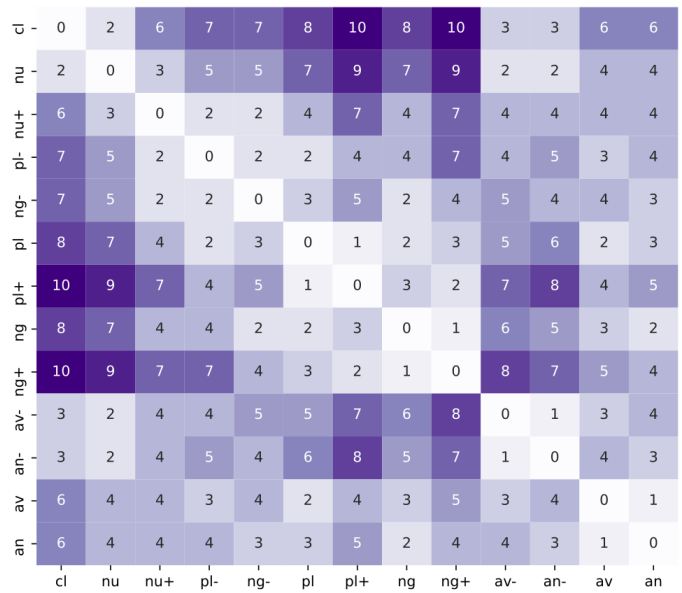
[1] NB: Rigoureusement elle ne l'est pas : la variable tranformée prend ses valeurs dans $\{\frac{1}{2M}, ..., 1 - \frac{1}{2M}\}$ qui tend vers $[0, 1]$ quand $M \to +\infty$

Dissimilarité entre caractéristiques

- **Variable catégorielle** (catégories non-ordonnées)

Exemple: temps sensible à Météo-France $X_j \in \mathcal{E}$ avec $\mathcal{E}$ ={pictogrammes pluie, soleil etc.}...

Il faut définir une matrice de dissimilarité $(L_{rr'})_{r,r' \in \mathcal{E}}$ telle que $d_j(x_j^i = r, x_j^{i'} = r') = L_{rr'}$



Le plus souvent, on prend $L_{rr'} = 0$ si $r \neq r'$, 1 sinon.

Dissimilarité entre individus

On dispose de $X = \begin{pmatrix} x_1^1 & \dots & x_p^1 \\ \vdots & & \vdots \\ x_1^N & \dots & x_p^N \end{pmatrix}$

et de $d_j(x_j^i, x_j^{i'})$ pour chaque caractérique $j \in [\![1, p]\!]$.
La dissimilarité entre individus est alors donnée par

$$d_{ii'} = d(x^i, x^{i'}) = \sum_{j=1}^{p} w_j d_j(x_j^i, x_j^{i'})$$

- Choix des w_j ?
- Problème des unités ?
- Problème des données manquantes ?

Dissimilarité entre individus

Choix des w_j ?

$$\overline{D} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d(x^i, x^{i'}) = \sum_{j=1}^{p} w_j \overline{d_j}$$

avec $\overline{d_j} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} d_j(x_j^i, x_j^{i'})$

- La contribution de la j-ème caractéristique à la dissimilarité totale est $w_j \overline{d_j}$.
- Prendre $w_j = 1/\overline{d_j}$ égalise les contributions...

Peut être contre-productif si certaines caractéristiques "séparent mieux" que d'autres.

Dissimilarité entre individus

Choix des w_j ?

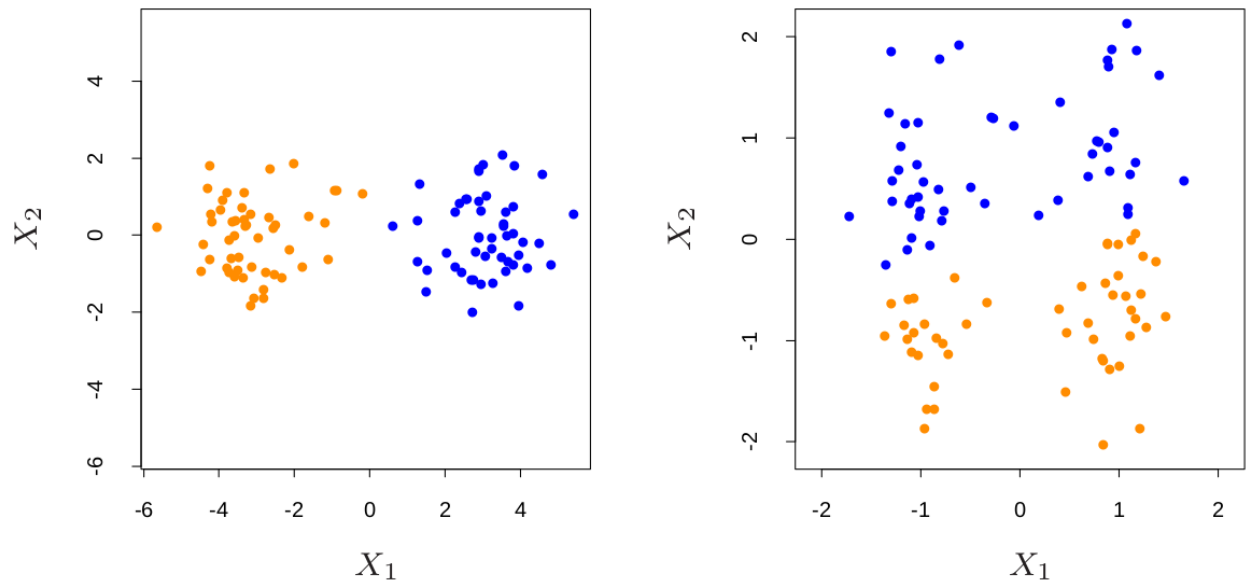


FIGURE 14.5. *Simulated data: on the left, K-means clustering (with $K=2$) has been applied to the raw data. The two colors indicate the cluster memberships. On the right, the features were first standardized before clustering. This is equivalent to using feature weights $1/[2 \cdot \mathrm{var}(X_j)]$. The standardization has obscured the two well-separated groups. Note that each plot uses the same units in the horizontal and vertical axes.*

©figure de Hastie, Tibshirani, Friedman (2001).

Dissimilarité entre individus

Problème des unités ?

Si X_1 est une P_{mer} en hectopascal (~ 1000) et X_2 est une surcôte en mètres (~ 0.1), et qu'on calcule la dissimilarité directement : $d(x^i, x^{i'}) = \underbrace{(x_1^i - x_1^{i'})^2}_{\sim 100} + \underbrace{(x_2^i - x_2^{i'})^2}_{\sim 0.01}$

X_1 est prépondérant dans la dissimilarité à cause de son unité...

- Les caractéristiques doivent être adimensionnées dans le calcul de la dissimilarité.

$\Rightarrow$ Normalisation...

- "uniforme" : $\tilde{X}_j = \dfrac{X_j - \min(X_j)}{\max(X_j) - \min(X_j)}$
- "normale" : $\tilde{X}_j = \dfrac{X_j - \overline{X_j}}{\sigma(X_j)}$
- autre...

- Ce choix a une grande influence sur les résultats.

$\Rightarrow$ Nécessite d'être fait par un utilisateur expert.

Dissimilarité entre individus

Problème des données manquantes ?

En pratique, il arrive que certaines caratéristiques ne soient pas disponibles pour certains individus (panne capteur, erreur, valeur aberrante...).

Soient deux individus x^i et $x^{i'}$.

- Si x^i et $x^{i'}$ ont des caractéristiques $J_{commun} \subset [\![1, p]\!]$ en commun

$\rightarrow$ on omet les caractéristiques manquantes

$$d_{ii'} = \sum_{j \in J_{commun}} w'_j d_j(x^i, x^{i'}) \text{ avec } \sum_{j \in J_{commun}} w'_j = \sum_{j=1}^{p} w_j$$

- Si x^i et $x^{i'}$ n'ont aucune caractéristique en commun

$\rightarrow$ $d_{ii'}$ ne peut pas être calculée..

  - retirer un des deux individus (e.g. celui avec le plus de données manquantes)
  - compléter les données manquantes (*data imputation*)

Dissimilarité entre individus

Problème des données manquantes ?

Prenons par exemple $X = \begin{pmatrix} \times & x_2^1 & x_3^1 & x_4^1 & x_5^1 \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 & x_5^2 \\ \times & \times & x_3^3 & x_4^3 & x_5^3 \\ x_1^4 & x_2^4 & \times & \times & \times \end{pmatrix}$

où "$\times$" est une donnée manquante.

- $d_{12} = \sum_{j=\mathbf{2}}^{5} w'_j d_j(x^1, x^2)$ avec $\sum_{j=\mathbf{2}}^{5} w'_j = \sum_{j=1}^{5} w_j$
- $d_{34}? \rightarrow$ retirer x^4 de X (individu avec le plus de données manquantes)

Résumé sur la dissimilarité

- Pour chaque caractéristique, prendre une dissimilarité adaptée à sa nature.
- Adimensionner les caractérisques (normalisation ou choix des w_j).
- Définir une stratégie pour les données manquantes.

Définir la dissimilarité est une question qui nécessite une **expertise** sur le problème à résoudre.

Outline

Rappels

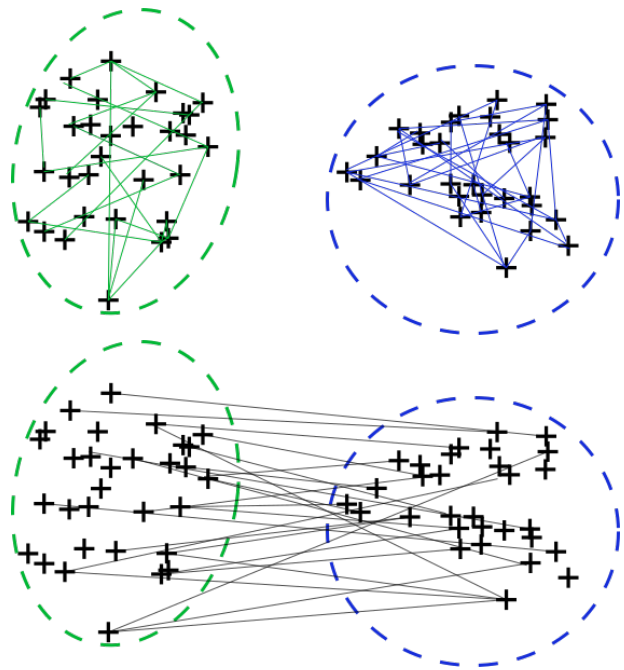
On cherche à grouper N individus en K groupes.
$\Longrightarrow$ trouver une fonction d'attribution C (i.e $C(i) = k$) qui minimise une fonction coût $W(C)$.

- Dispersion "within cluster" :

$$W(C) = \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x^i, x^{i'})$$

- Dispersion "between cluster":

$$B(C) = \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')\neq k} d(x^i, x^{i'})$$



- Dispersion "totale": $T = \sum_{i=1}^{N} \sum_{i'=1}^{N} d(x^i, x^{i'}) = W(C) + B(C)$

où $d(x^i, x^{i'})$ mesure la **dissimilarité entre** x^i **et** $x^{i'}$.

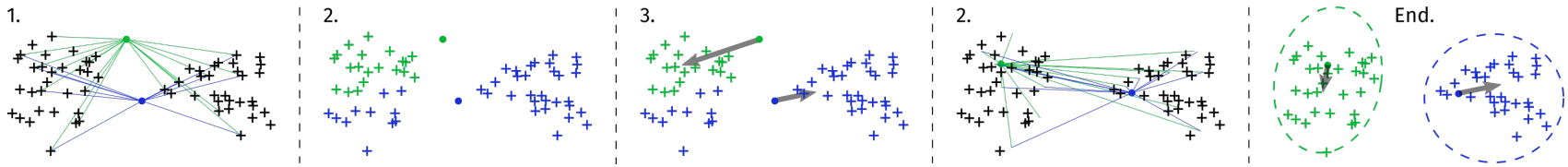
L'algorithme des K-means

- Le nombre de groupes K est donné.
- Toutes les caractéristiques sont quantitatives.
- La dissimilarité est la distance euclidienne : $d(x^i, x^{i'}) = \sum_{j=1}^{p} (x_j^i - x_j^{i'})^2$

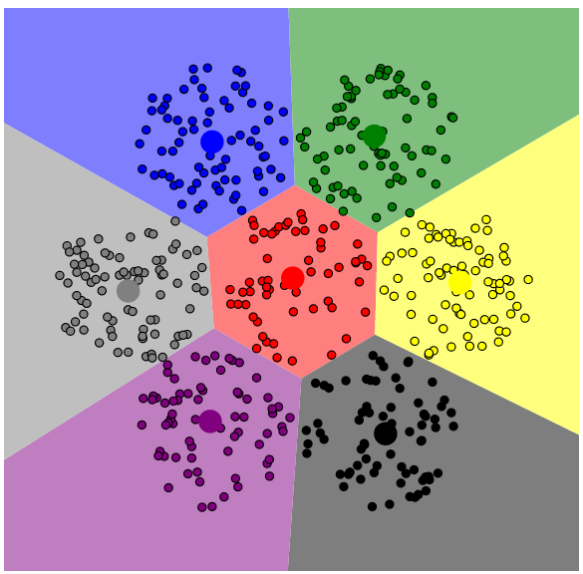
Algorithme des K-means

1. Prendre K individus $m_1, ..., m_K$ comme centroïdes.
2. Attribuer les individus selon $C(i) = \arg\min_k \{d(x^i, m_k)\}$
3. Redéfinir les centroïdes par $m_k = \frac{\sum_{i=1}^{N} x^i \mathbf{1}_{C(i)=k}}{\sum_{i=1}^{N} \mathbf{1}_{C(i)=k}}$
4. Répéter les étapes 2. et 3. tant que $W(C)$ décroit.

Convergence des K-means



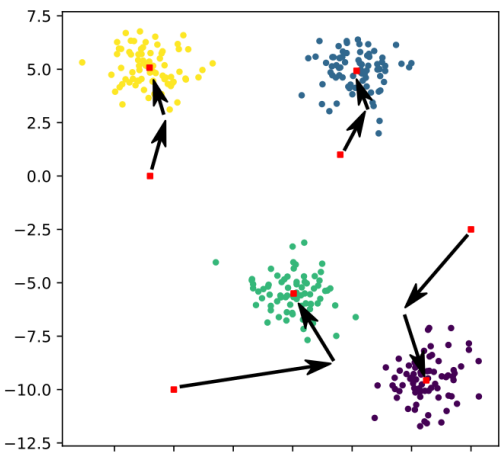
- Assez rapide (quelques dizaines d'itérations suffisent)
- Vers un minimum local de $W(C)$.



Les frontières sont toujours des hyperplans (donc des droites en 2D). Elles forment ce qu'on appelle un **diagramme de Voronoï**.

Initialisation des K-means

- Influence beaucoup le résultat final
- Exemple d'initialisation:
  - aléatoire, répétée plusieurs fois.
  - points extrêmes (stéréotypes caricaturaux)
  - aléatoire guidé (ex: "K-means++")



Exemple d'initialisation

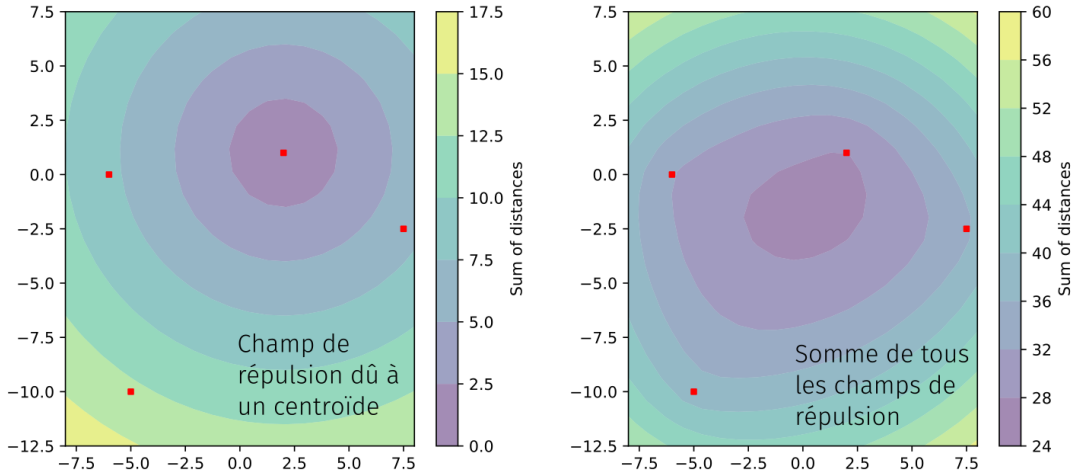
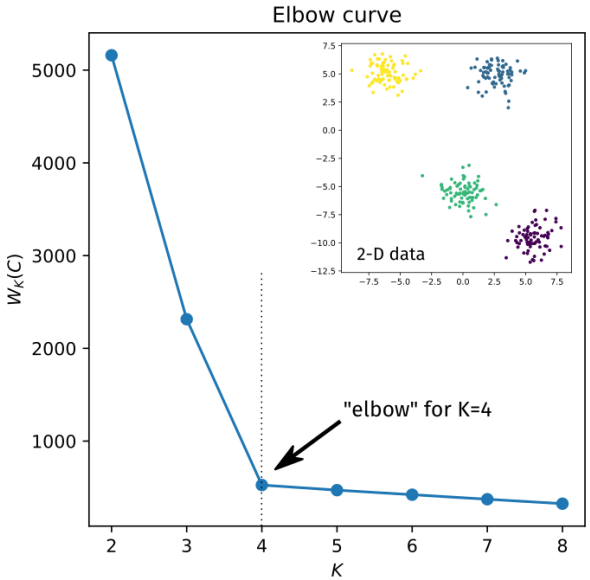


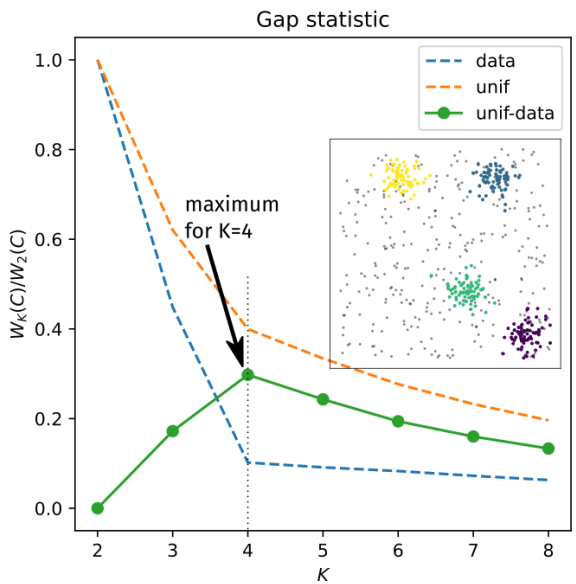
Illustration de l'initialisation K-means++ avec 4 centroïdes

Choix de K

- Peut être imposée par le problème (e.g. 26 lettres dans l'alphabet)
- Estimé automatiquement depuis les données
  - Répéter l'algorithme pour K prenant des valeurs raisonnables et chercher un "coude" dans la courbe de $W_K(C)$.
  - Comparer avec des données suivant une loi uniforme ("gap statistics", Tibshirani et al., 2001).
  - Optimiser un score de classification

Pas de validation croisée possible!





Choix de K : méthode de la silhouette

Soit i un individu affecté à un groupe k. On définit:

- $a(i) = \frac{1}{|C_k|-1} \sum_{i' \in C_k, i' \neq i} d(x^i, x^{i'})$: dissimilarité moyenne entre i et les membres de son groupe.
- $\Delta(i,l) = \frac{1}{|C_l|} \sum_{i' \in C_l} d(x^i, x^{i'})$: dissimilarité moyenne entre i et les membres du groupe l.
- $b(i) = \min_{l \neq k} \Delta(i,l)$: dissimilarité moyenne entre i et les membres du groupe voisin.
- $s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$: **silhouette** du point i. $s(i) \in [-1, 1]$

Quand $s(i) \simeq 1$, $a(i) \ll b(i)$, donc i est bien classé.
Quand $s(i) \simeq -1$, $a(i) \gg b(i)$, donc i est mal classé.

$\Rightarrow$ la silhouette est un score d'évaluation de la classification.

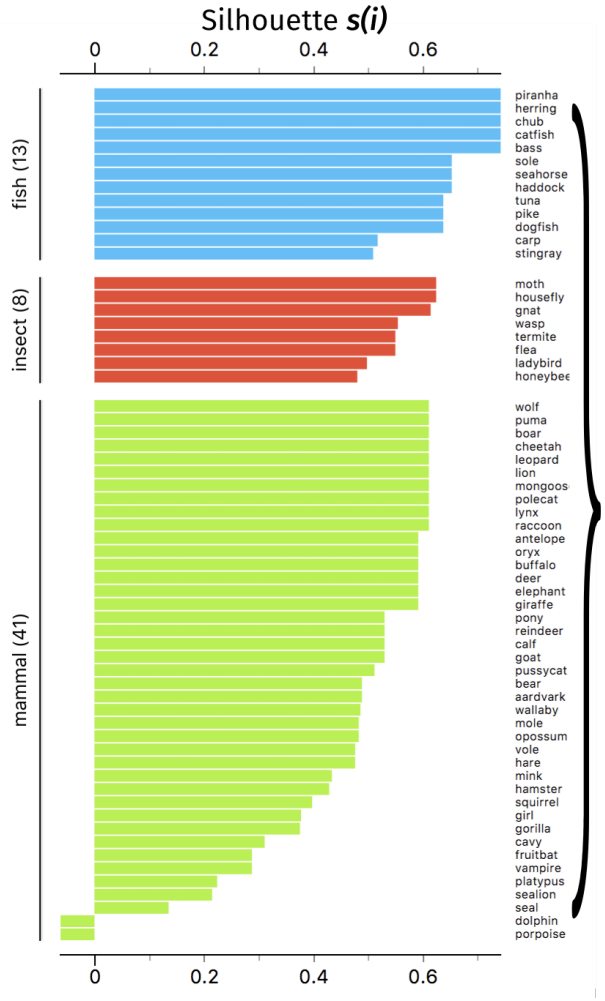
Choix de K : méthode de la silhouette

En pratique, on calcule $s(i)$ pour tous les points i et on trace les $s(i)$ par ordre décroissant sur la verticale.

Les points sont regroupés par cluster.

On voit que les clusters bleu et rouge n'ont que des fortes silhouettes. En revanche, le cluster vert contient des points à silhouette faible.

$\Rightarrow$ Passer de $K = 3$ à $K = 4$?



Variantes : K-medians

K-medians

- Toutes les caractéristiques sont quantitatives.
- La dissimilarité est la distance de Manhattan : $d(x^i, x^{i'}) = \sum_{j=1}^{p} |x_j^i - x_j^{i'}|$

Algorithme des K-medians

1. Prendre K individus $m_1, ..., m_K$ comme centroïdes.
2. Attribuer les individus selon $C(i) = \arg\min_k \{d(x^i, m_k)\}$
3. Redéfinir les centroïdes par $m_k = \text{median}\{x^i, C(i) = k\}$
4. Répéter les étapes 2. et 3. tant que $W(C)$ décroit.

Variantes : K-medoids

K-medoids (ou PAM : Partition Around Medoids)

- Caractéristiques sont quelconques.
- La dissimilarité est quelconque.
- Les centroïdes sont des individus de l'échantillon.

Algorithme des K-medoids

1. Prendre K individus $m_1, ..., m_K$ parmi $x^1, ..., x^N$ comme centroïdes.
2. Attribuer les individus selon $C(i) = \arg\min_k \{d(x^i, m_k)\}$
3. Redéfinir les centroïdes par $m_k = x^{i_k^*}$ où $i_k^* = \arg\min_{i, C(i)=k} \sum_{C(i')=k} d(x^i, x^{i'})$
4. Répéter les étapes 2. et 3. tant que $W(C)$ décroit.

Résumé sur les K-means

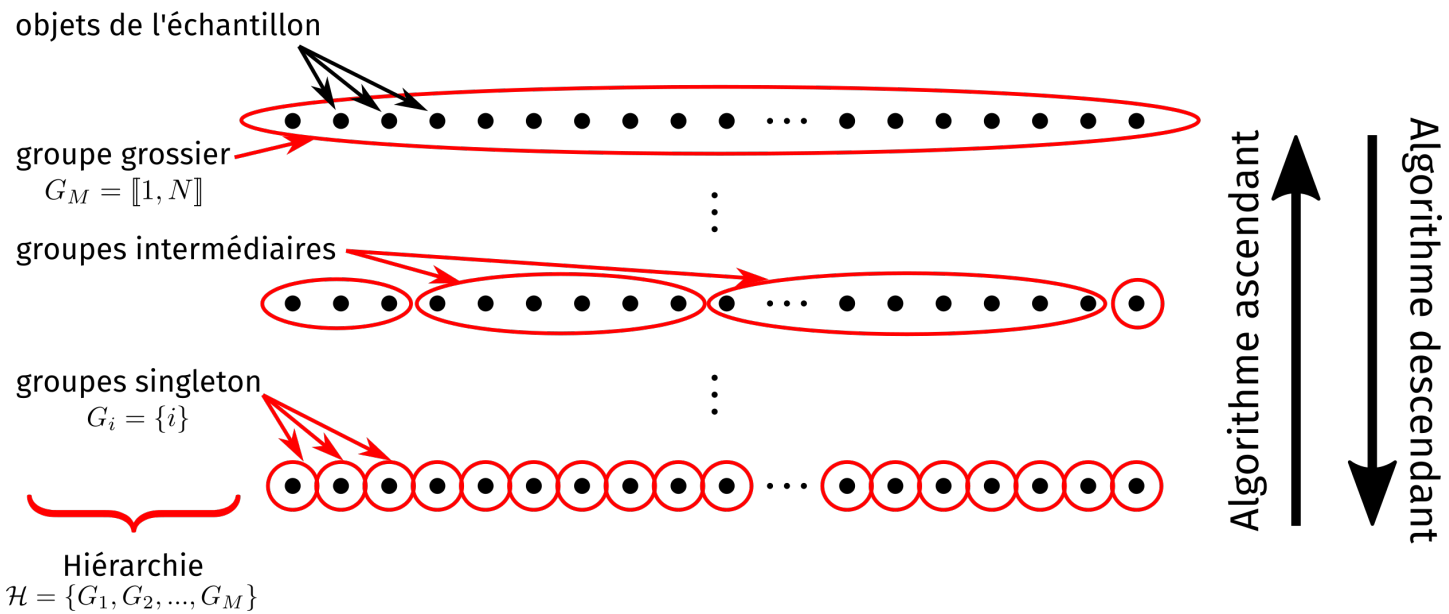
- Il s'agit d'un algorithme très populaire et assez vieux (1967).
- Nombre de groupes K donné a priori.
- **Converge** rapidement, mais vers un minimum local de $W(C)$.
- L'**initialisation** est donc importante, on conseille de répéter l'algorithme plusieurs fois avec différents points de départ.
- Plusieurs astuces existent pour **choisir** K ("coude" dans courbe de $W(C)$, gap statistic...)
- Des **variantes** de l'algorithme permettent de l'adapter à des contraintes particulières (choix de la dissimilarité...)

Outline

Classification hiérarchique ascendante

Une classification hiérarchique donne une **hiérarchie** de groupes (i.e. des groupes imbriqués les uns dans les autres).

- Groupe : ensemble d'indices $G \subset [\![1, N]\!]$.
- Hierarchie : ensemble de groupes $\mathcal{H} \in \mathcal{P}([\![1, N]\!])$ tel que
  - $[\![1, N]\!] \in H$
  - Si G et H sont dans $\mathcal{H}$, $G \cap H \neq \emptyset \Rightarrow (G \subset H$ ou $H \subset G)$



Classification hiérarchique ascendante

Algorithme ascendant (ou agglomératif):

- On commence par considérer tous les individus comme des groupes (singletons) puis on réunit les 2 groupes les plus proches en un nouveau groupe qui remplace les 2 précédents. On réunit à nouveau les 2 groupes les plus proches, et ainsi de suite jusqu'à agglomérer tous les points.

Algorithme descendant (ou divisif):

- On commence par considérer le groupe contenant tous les individus et on le divise de façon à maximiser la dissimilarité intra-groupes. Le groupe avec la plus grande variance est ensuite divisé de la même façon et ainsi de suite jusqu'aux singletons.
- Les algorithmes divisifs sont assez rares

Distance cophénétique

Dans un algorithme ascendant, on commence par considérer tous les points comme des groupes puis on réunit les 2 groupes les plus proches en un groupe père, et ainsi de suite.

$\Rightarrow$ Par construction, la distance entre les groupes réunis est croissante : on parle de **distance cophénétique**.

$\Rightarrow$ On peut représenter les réunions successives par une structure en arbre : on parle de **dendrogramme**.

Distance cophénétique

Clustering hiérarchique nécessite une distance entre groupes (linkage). On attend de cette distance que

- Les classes qui en résultent soient bonnes :
  - Compacité : classes de petit diamètre
  - Proximité : point d'un classe + proche que ceux de la classe voisine
- Une augmentation du nombre de points lui soit bénéfique.
- Elle ne soit pas affectée par une transformation monotone des distances : utiliser $d_{ii'}$ ou $f(d_{ii'})$ ne doit rien changer si f est monotone.

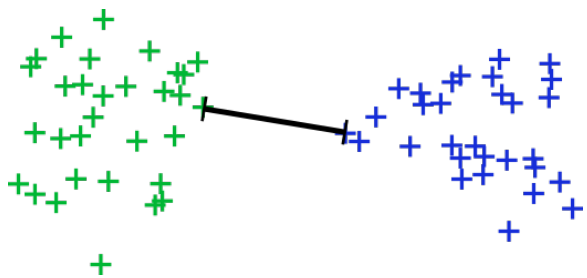
Soient G et H deux groupes de $[\![1, N]\!]$ et $|G|$, $|H|$ leurs cardinaux respectifs.

Distance cophénétique

Saut minimal[1]**:** $D(G,H) = \min_{i \in G, i' \in H} d_{ii'}$

- Tendance à faire des chaines de points, les clusters résultants seront étirés, de large diamètre.

$\Rightarrow$ Détection de bandes/strates...



Avantages

- Points d'un même cluster sont proches entre eux
- Invariant par transformation monotone

Inconvénients

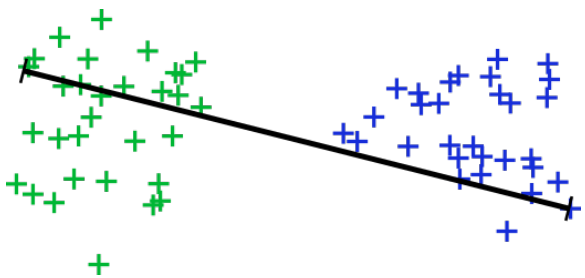
- Clusters peu compacts
- Tend vers 0 quand $N \to \infty$
- Pas lié à une stat sur la population

[1] En anglais: *simple linkage*

Distance cophénétique

Saut maximal[1]**:** $D(G,H) = \max_{i \in G, i' \in H} d_{ii'}$

- Opposé du saut minimal : crée des groupes compacts de petits diamètres mais dont les points au bord peuvent être plus proche d'un autre cluster que d'un point du même cluster.



$\Rightarrow$ Détection de paquets bien isolés

Avantages

- Cluster compacts
- Invariant par transformation monotone

Inconvénients

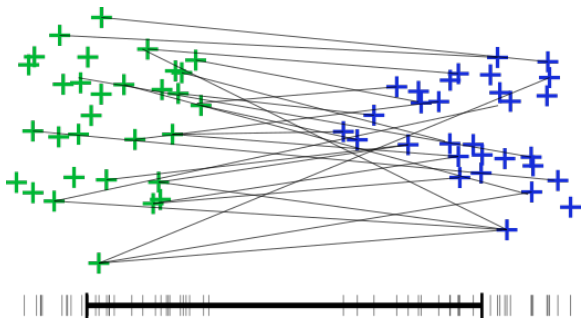
- Points de clusters différents peuvent être + proche que points d'un même cluster.
- Tend vers $+\infty$ quand $N \to \infty$
- Pas lié à une stat sur la population

[1] En anglais: *complete linkage*

Distance cophénétique

Distance moyenne[1]: $D(G,H) = \frac{1}{|G||H|} \sum_{i \in G, i' \in H} d_{ii'}$

- Compromis entre saut minimal et maximal. Elle n'est pas invariante par transformation monotone mais elle est consistante avec la formulation du clustering par mélange de modèle et converge quand $N \to \infty$.



Avantages

- Compromis compacité/proximité
- Quand $N \to \infty$, tend vers $\mathbb{E}\left[d(X,X')|X \in G, X' \in H\right]$

Inconvénients

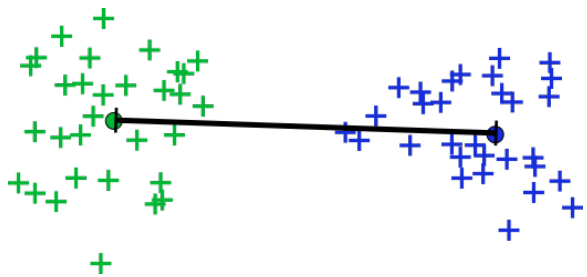
- Non invariant par transformation monotone

[1] En anglais: *group average*

Distance cophénétique

Distance à centroïdes: $D(G,H) = d(\overline{x_G}, \overline{x_H})$
où $\overline{x_G} = \frac{1}{|G|}\sum_{i\in G} x^i$

- Résume un cluster par son point moyen, ce qui a du sens pour les clusters avec des formes assez régulières et de taille équilavente mais qui peut être risqué autrement.



Avantages

- Invariant par transformation monotone
- Stable quand $N \to \infty$

Inconvénients

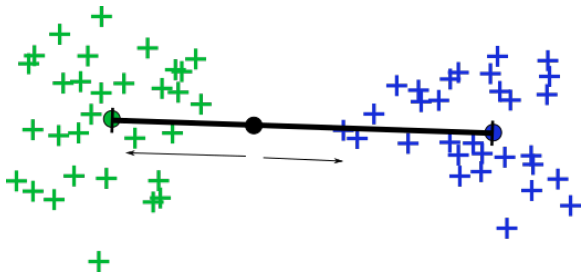
- Ne tient pas compte de la forme des clusters.
- Petits clusters comptent autant que les gros.

Distance cophénétique

Distance de Ward: $D(G,H) = \dfrac{w_G w_H}{w_G + w_H} d(\overline{x_G}, \overline{x_H})$

où $\overline{x_G} = \frac{1}{|G|}\sum_{i \in G} x^i$ et $w_G = \frac{|G|}{N}$

- Distance qui maximise l'augmentation d'inertie inter-classe[1]. Dans le cas de la distance euclidienne, c'est équivalent à minimiser $W(C)$, ce qui en fait un choix judicieux.



Avantages

- Invariant par transformation monotone
- Stable quand $N \to \infty$
- Maximise l'augmentation d'inertie inter-classe.

Inconvénients

- Ne tient pas compte de la forme des clusters.
- Inertie n'a de sens que pour la distance euclidienne.

[1] Inertie inter-classe : $I_e = \sum_{k=1}^{K} \frac{|C_k|}{N} d(\overline{x_{C_k}}, \overline{X})^2$ où $\overline{X} = \sum_i x^i / N$

Distance cophénétique : bilan

Nous avons vu 5 distances possibles entre deux groupes G et H :

- Saut minimal: $D(G,H) = \min_{i \in G, i' \in H} d_{ii'}$
- Saut Maximal: $D(G,H) = \max_{i \in G, i' \in H} d_{ii'}$
- Distance moyenne: $D(G,H) = \frac{1}{|G||H|} \sum_{i \in G, i' \in H} d_{ii'}$
- Distance à centroïdes: $D(G,H) = d(\overline{x_G}, \overline{x_H})$
- Distance de Ward: $D(G,H) = \frac{w_G w_H}{w_G + w_H} d(\overline{x_G}, \overline{x_H})$

où $\overline{x_G} = \frac{1}{|G|} \sum_{i \in G} x^i$ et $w_G = \frac{|G|}{N}$.
Mais il en existe beaucoup d'autres...

Choix ? Arbitrage avantages/inconvénients par un expert du problème à résoudre.

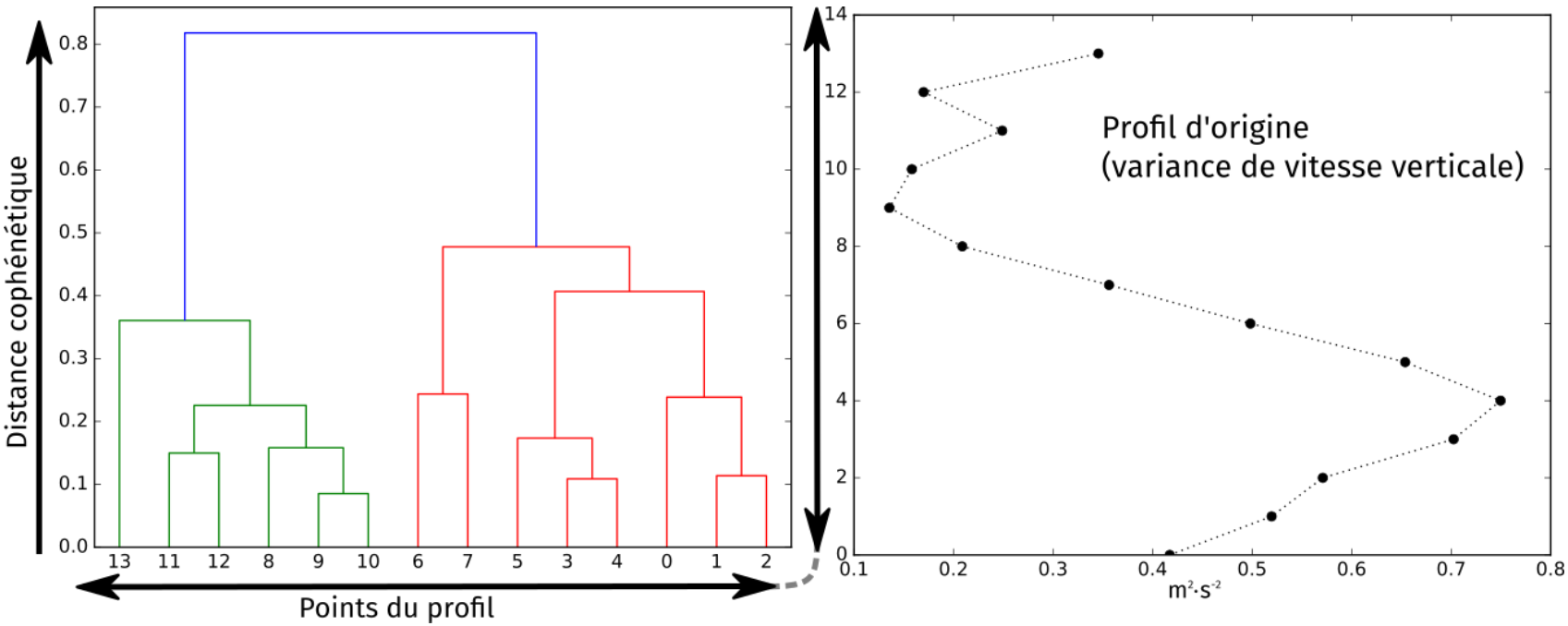
Dendrogramme

Dans un algorithme ascendant, on commence par considérer tous les individus comme des groupes puis on réunit les 2 groupes les plus proches en un groupe père, et ainsi de suite.

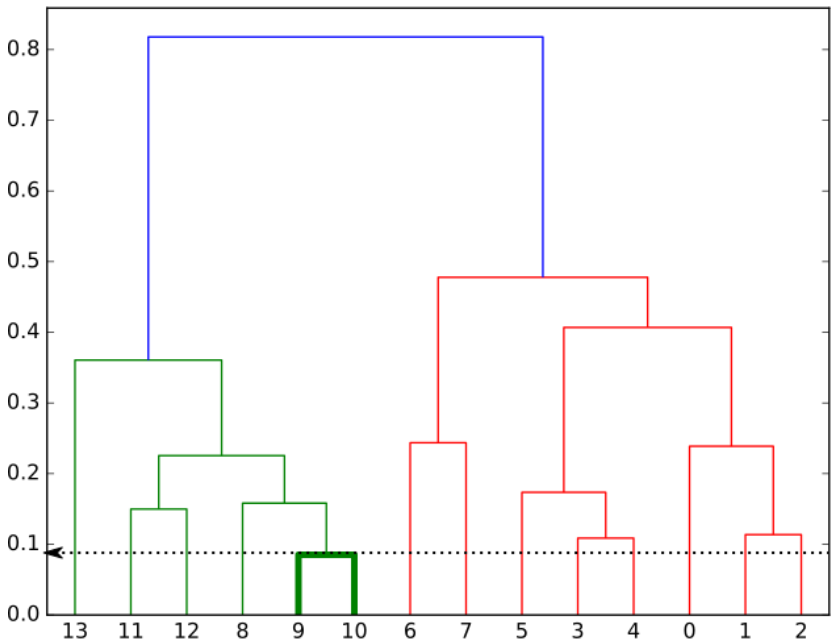
$\Rightarrow$ Par construction, la distance entre les groupes réunis est croissante : on parle de **distance cophénétique**.

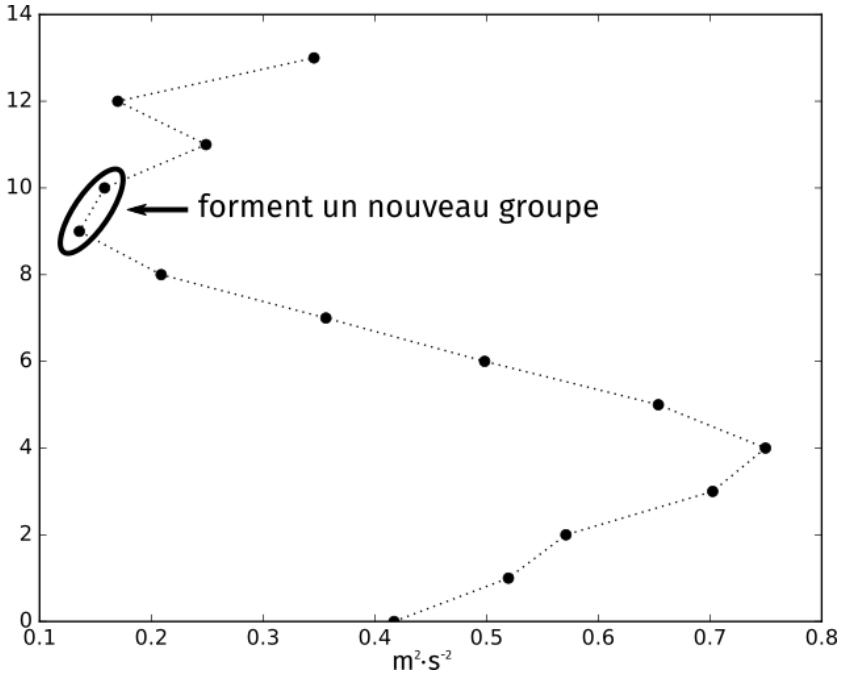
$\Rightarrow$ On peut représenter les réunions successives par une structure en arbre : on parle de **dendrogramme**.

Dendrogramme

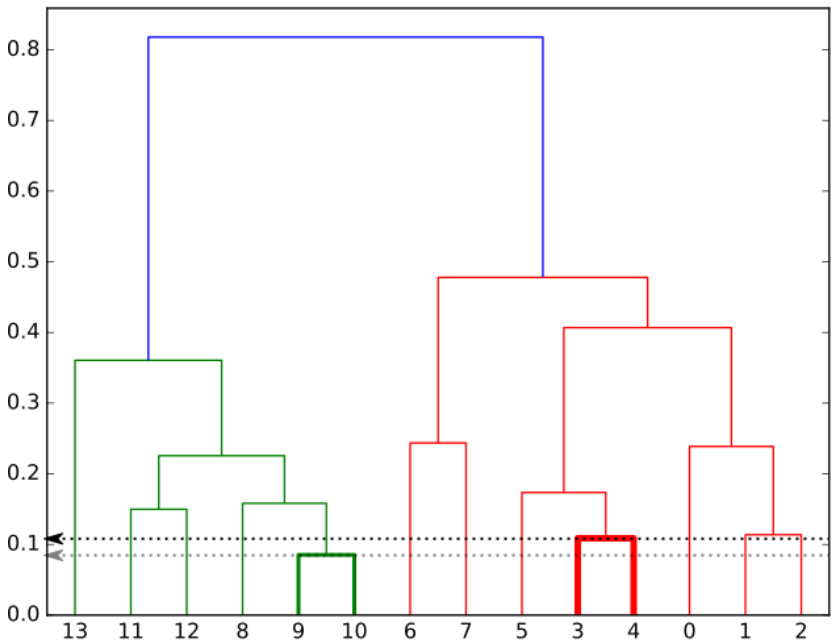


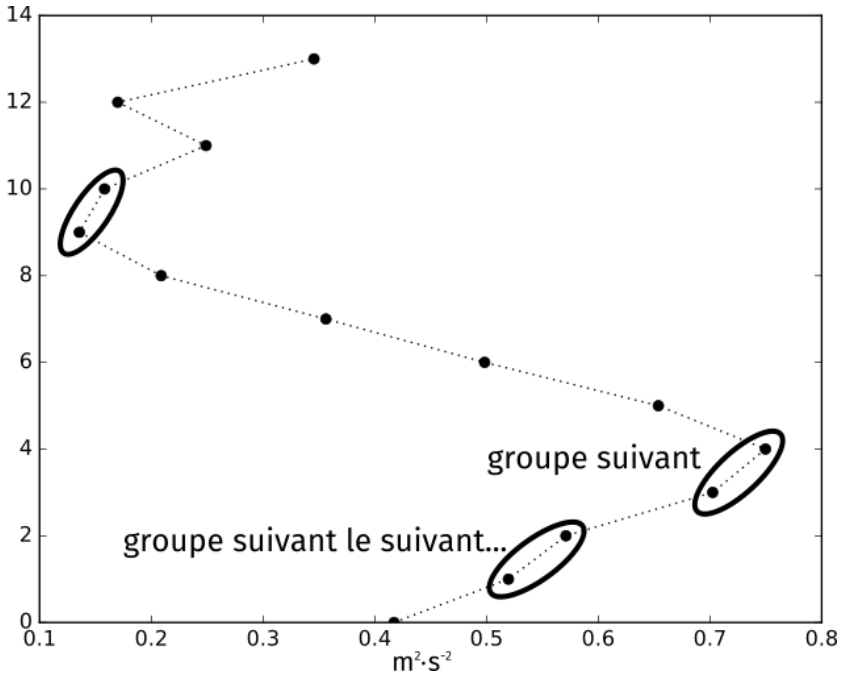
Dendrogramme



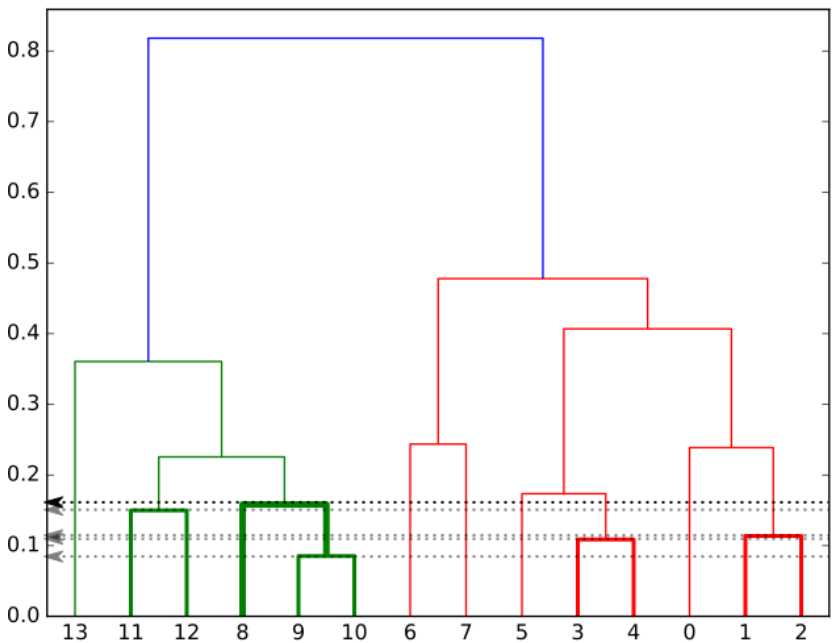


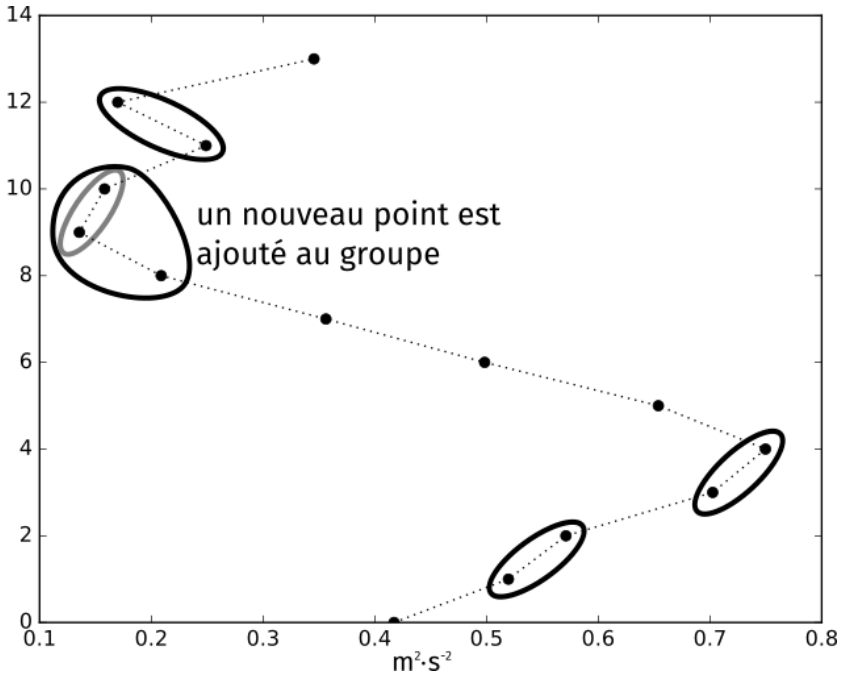
Dendrogramme



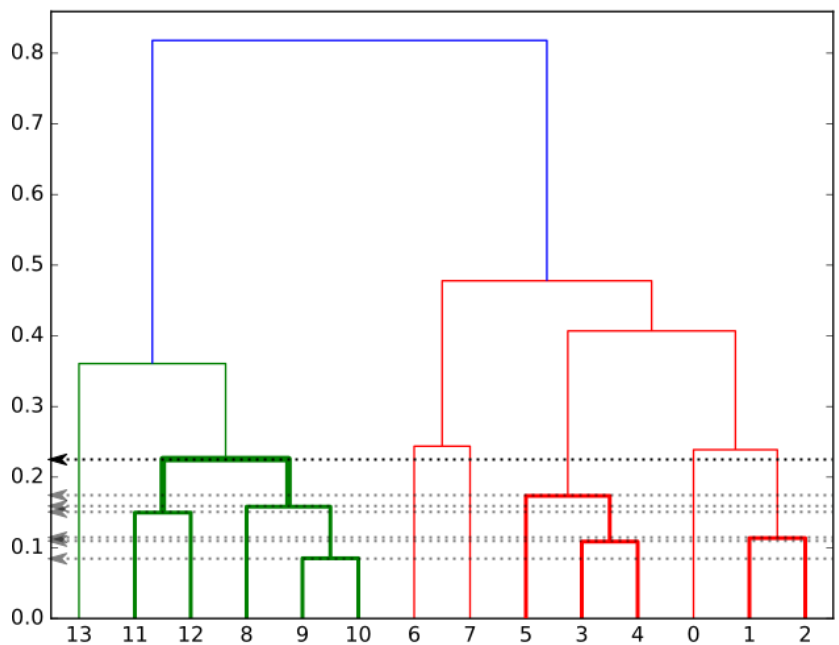


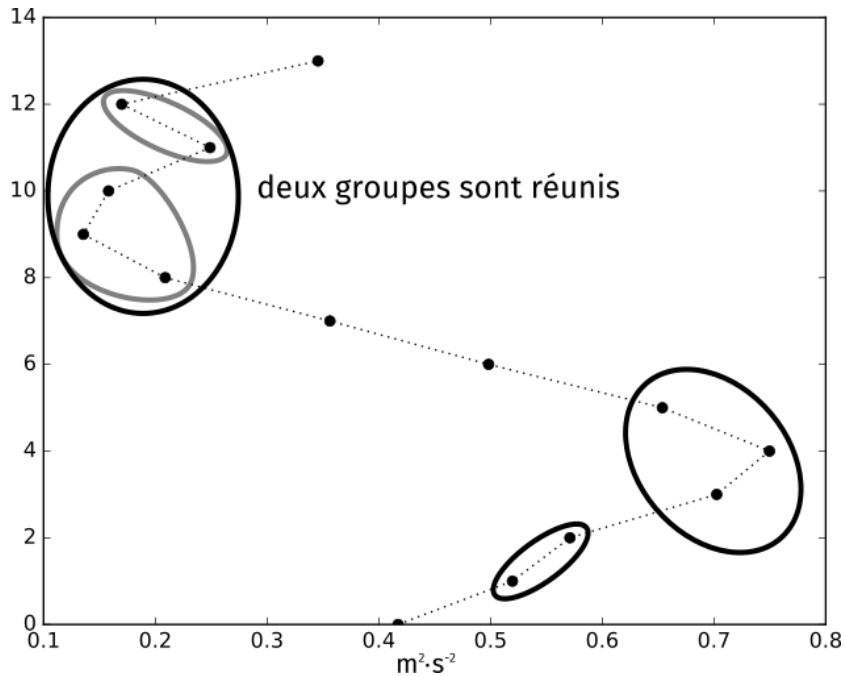
Dendrogramme



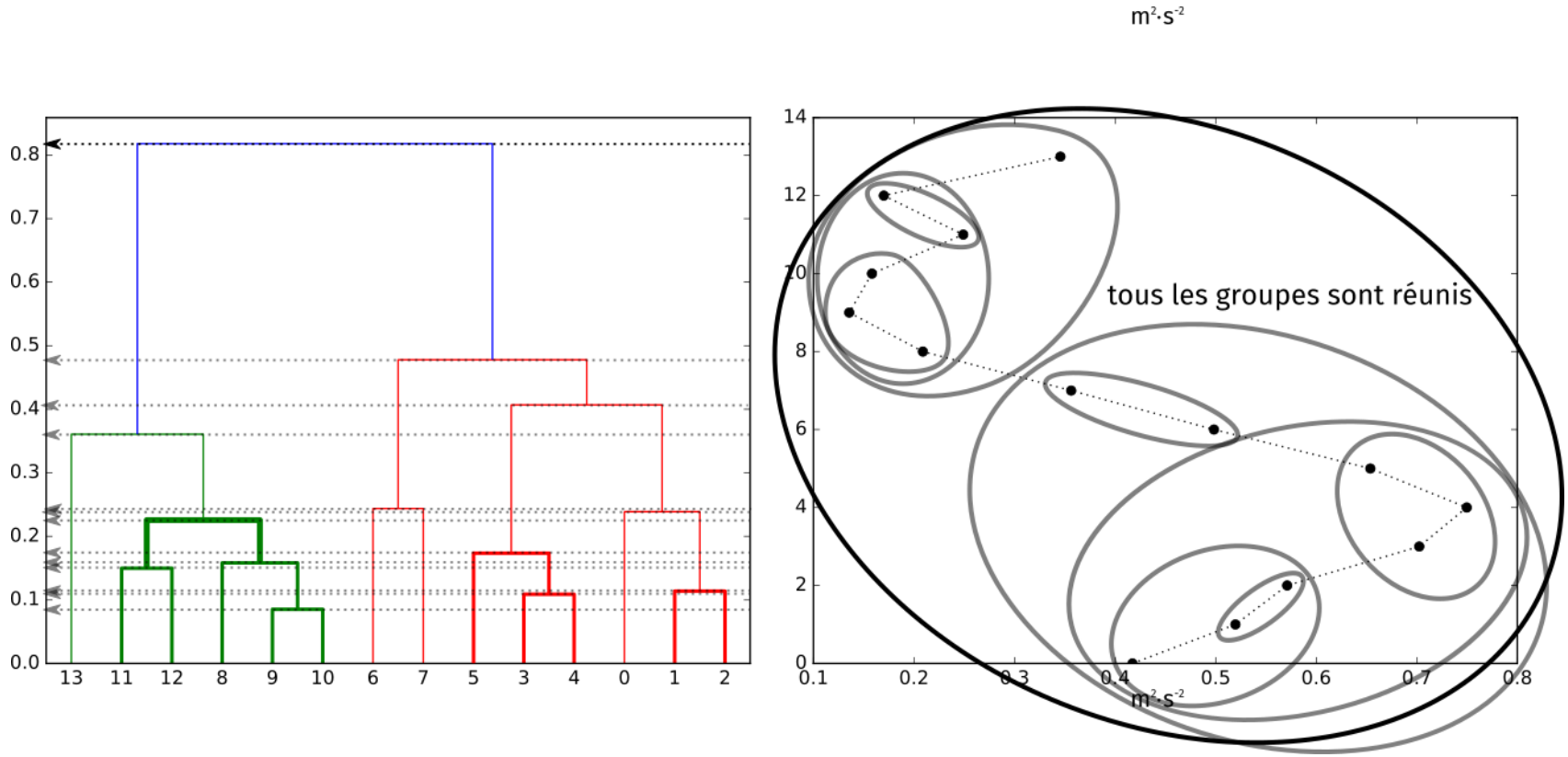


Dendrogramme





Dendrogramme



Dendrogramme

Avantages

- Résumé visuel et complet des données
- Permet d'identifier des groupes à plusieurs échelles
- Les groupes sont imbriqués (cohérence entre les échelles)
- Une fois le nombre de groupes fixés, les groupes sont déjà calculés.

Inconvénients

- Dépend du choix de distance cophénétique.
- C'est le résultat d'un algorithme et non les données brutes.
- Donne toujours une structure hiérarchique, même lorsqu'il n'y en a pas.
- De petites modifications des données peuvent produire un dendrogramme assez différent.

Choix de K ?

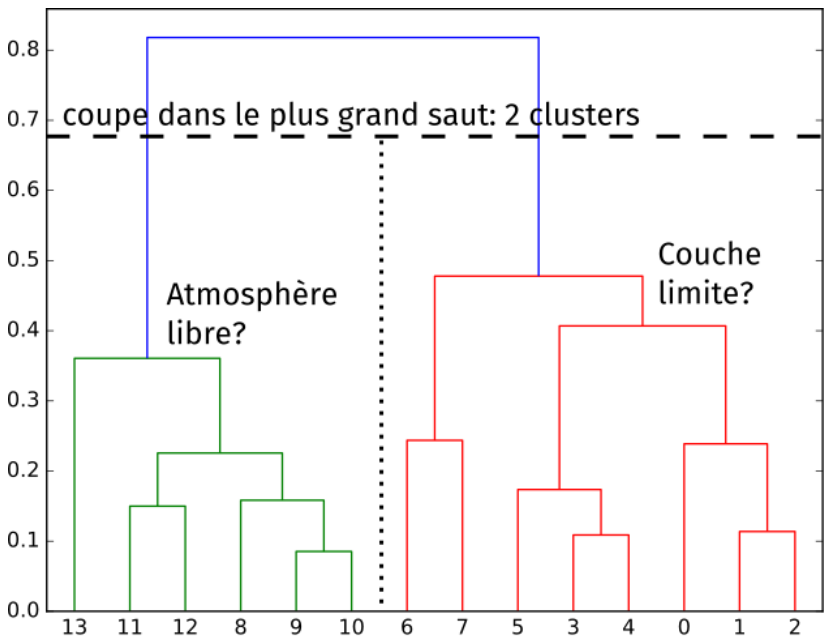
Le clustering hiérarchique ne donne pas une partition mais une hiérarchie, visualisable avec un dendrogramme. Pour obtenir une partition, il faut choisir un niveau dans la hiérachie.

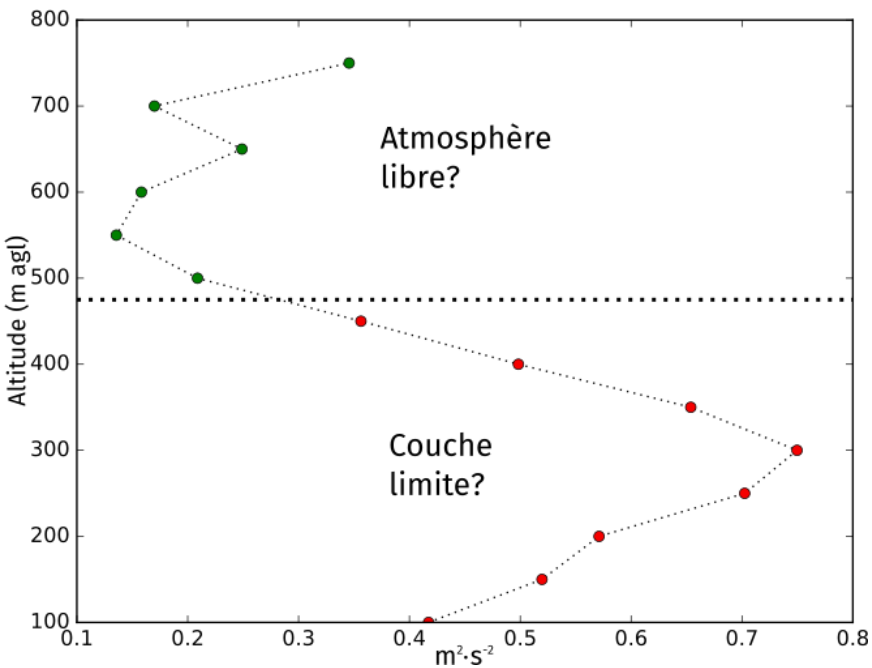
$\Rightarrow$ **Où couper dans le dendrogramme ?**

- Chercher les liens qui suivent de grands sauts
- De façon automatique : calculer les coefficients d'inconsistance

Choix de K : saut le plus grand

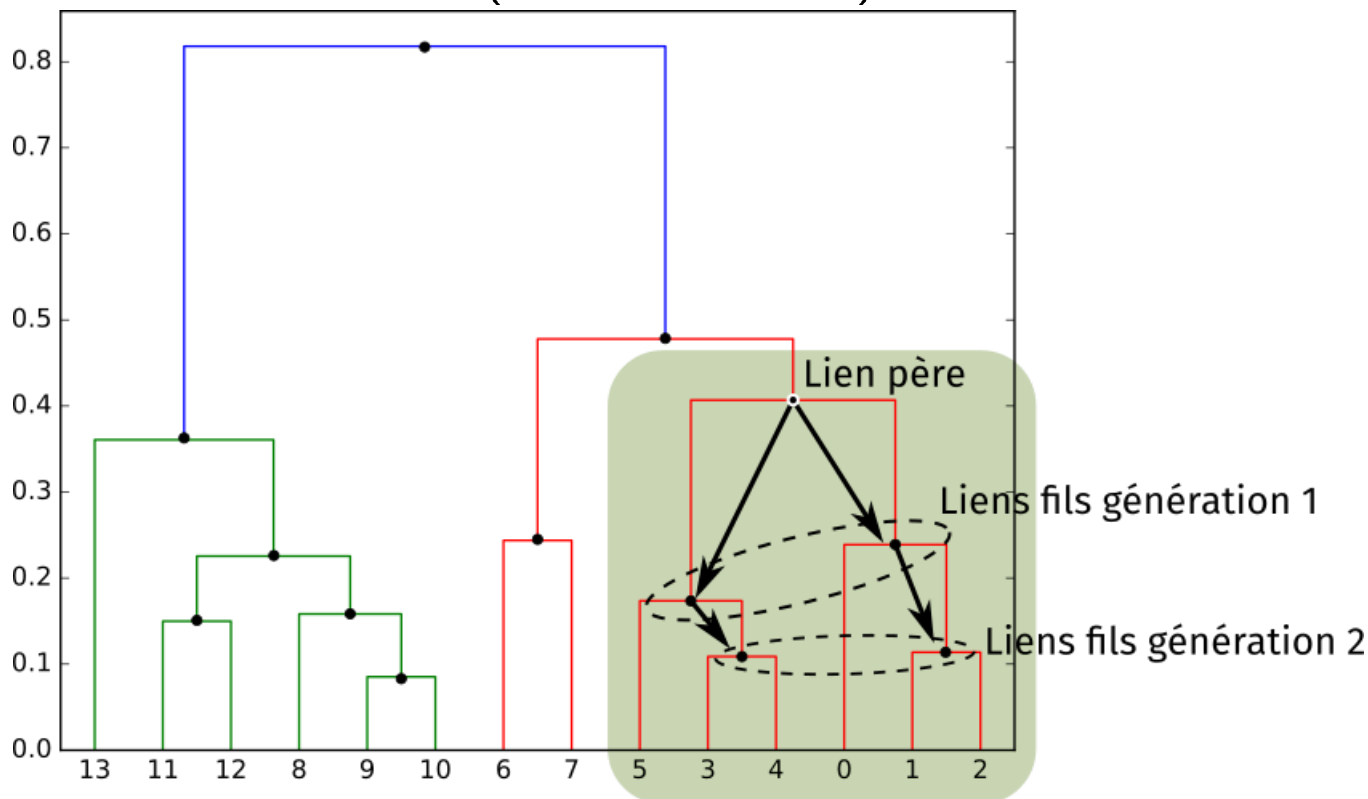
Détecter le saut le plus grand (par examen visuel ou en dérivant la distance cophénétique suivant les générations).





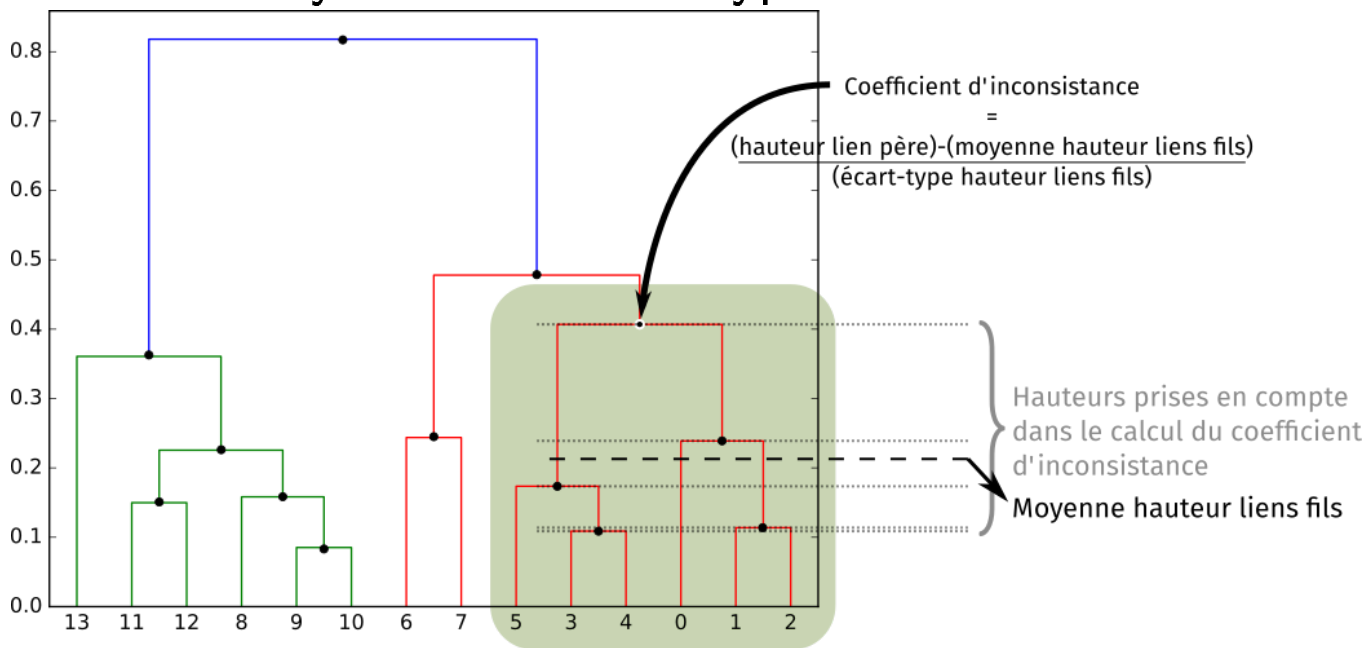
Choix de K : coefficient d'inconsistance

Pour chaque lien, on identifie ses liens fils sur d générations (souvent $d = 2$).



Choix de K : coefficient d'inconsistance

On calcule la moyenne et l'écart type de la hauteur de ces liens.



L'écart normalisé avec la hauteur du lien père est le coefficient d'inconsistance.

On coupe le lien avec le **plus fort coefficient d'inconsistance**.

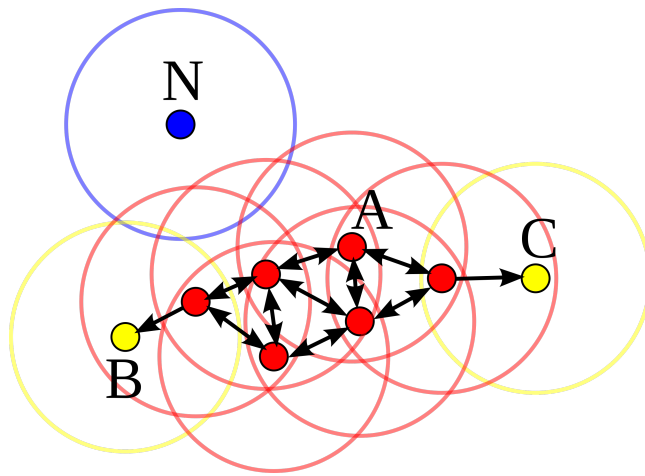
Outline

L'algorithme DBSCAN

Density-Based Spatial Clustering of Applications with Noise
(Ester et al., 1996)

Pseudo-algorithme:

1. Pour tout point x^i, trouver son ϵ-voisinage.
2. Les points avec plus de m voisins appelé *points cœur* (A). Les points cœur connecté entre eux sont mis dans le même cluster.
3. Les points non-cœur sont soit *au bord* (B,C) s'ils sont voisins d'un point cœur, soit *aberrant* (N) sinon.



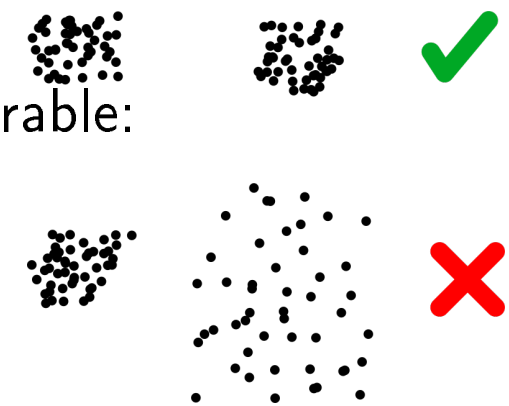
L'algorithme DBSCAN

Avantages

- Trouve automatiquement le nombre "naturel" de classes.
- Les groupes peuvent être de forme quelconque (circulaire, entrelacés...).
- Résilient aux données aberrantes et peut même les identifier.

Inconvénients

- Choix des paramètres ϵ et m.
- Les points au bord relié à plus d'un groupe peuvent changer d'affectation suivant l'ordre des points.
- Les groupes doivent être de densité comparable:



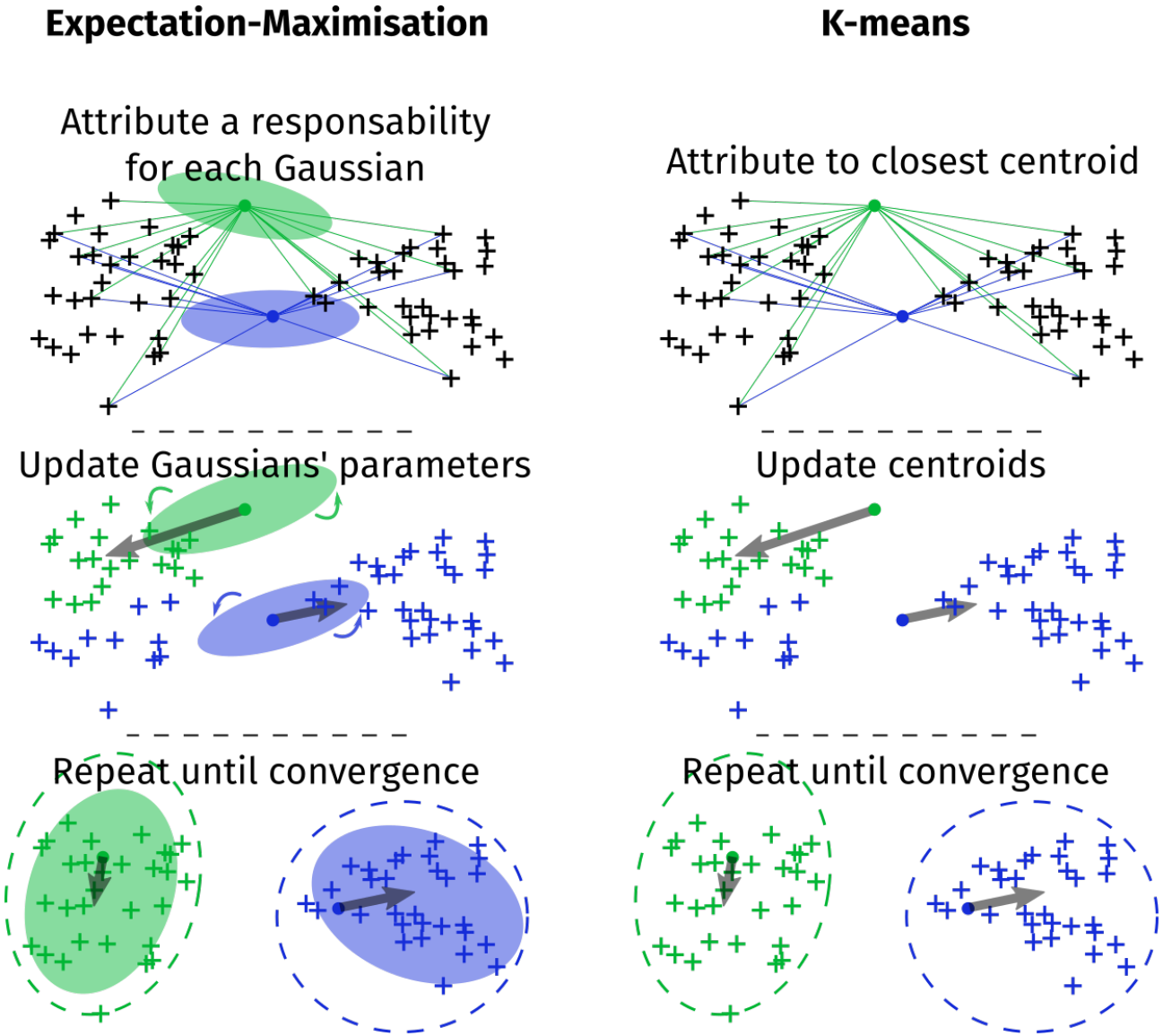
L'algorithme Expectation-Maximisation

- Le nombre de groupes K est donné.
- On suppose que chaque groupe est distribué selon une loi gaussienne $\mathcal{N}(\mu_k, \Sigma_k)$.
- On cherche alors à estimer les paramètres μ_k et Σ_k ainsi que la responsabilité de chaque gaussienne.

Algorithme EM

1. Initialiser $(\hat{\mu}_k, \hat{\sigma}_k, \hat{\pi}_k)_{k \in [\![1,K]\!]}$ (aléatoirement).
2. Expectation: attribuer une responsabilité à chaque gaussienne pour la position des points
3. Maximisation: recaluler les paramètres qui maximisent la vraisemblance
4. Répéter les étapes 2. et 3. tant que $W(C)$ décroit.

L'algorithme EM: lien avec les K-means ?



Indice de Davies-Bouldin

Se base sur

- les points moyens de chaque groupe $\mu_k = \frac{1}{|I_k|} \sum_{i \in I_k} x^i$
- la distance moyenne entre un point et le centre de son groupe $\bar{\delta}_k = \frac{1}{|I_k|} \sum_{i \in I_k} d(x^i, \mu_k)$

$$S_{DB} = \frac{1}{K} \sum_{k=1}^{K} \max_{k' \neq k} \left(\frac{\bar{\delta}_k + \bar{\delta}_{k'}}{d(\mu_k, \mu_{k'})} \right)$$

Interprétation:

- S_{DB} petit: bonne classification
- S_{DB} grand: mauvaise classification
- $S_{DB} = 0$: cas dégénéré où les points d'un groupes sont confondus avec son centre

Indice de Dunn

Se base sur

- les points moyens de chaque groupe $\mu_k = \frac{1}{|I_k|} \sum_{i \in I_k} x^i$
- les diamètres de chaque groupe $\Delta_k = \max_{i,i' \in I_k} d(x^i, x^{i'})$

$$S_D = \frac{\min\limits_{1 \leqslant k < k' \leqslant K} d(\mu_k, \mu_{k'})}{\max\limits_{1 \leqslant k \leqslant K} \Delta_k}$$

Interprétation:

- S_D petit: mauvaise classification
- S_D grand: bonne classification
- $S_D = 0$: lorsqu'au moins deux groupes ont le même centre (clusters circulaires concentriques)

Indice de Calinski-Harabasz

Se base sur

- la variance inter-groupes $B = \sum_{k=1}^{K} |I_k| \|\mu_k - \mu\|$
- les variances intra-groupe $W_k = \frac{1}{|I_k|} \sum_{i \in I_k} \|x^i - \mu_k\|$

$$S_{CH} = \frac{(N-K)B}{(K-1)\sum_{k=1}^{K} W_k}$$

Interprétation:

- S_{CH} petit: mauvaise classification
- S_{CH} grand: bonne classification
- $S_{CH} = 0$: cas dégénéré où tous les points sont confondus
- Croît linéairement avec N: ordre de grandeur très variable suivant données

Bibliographie

Sources:

- Friedman, Hastie, & Tibshirani (2001). *The elements of statistical learning.* Springer Edition. 2nd Edition. Chapitre 14.
- Saporta (2006). *Probabilités, analyse des données et statistique.*
- Jain, Murty, & Flynn (1999). *Data clustering: a review.* ACM computing surveys (CSUR).
- Rousseeuw (1987). *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of computational and applied mathematics.
- Desgraupes (2017). *Clustering indices.* Comprehensive R Archive Network.

Références:

- (Ester et al., 1996) : Ester, Kriegel, Sander & Xu (1996, August). *A density-based algorithm for discovering clusters in large spatial databases with noise.*
- (Tibshirani, 2001) Tibshirani, Walther, & Hastie (2001). *Estimating the number of clusters in a data set via the gap statistic.* Journal of the Royal Statistical Society: Series B (Statistical Methodology).

Merci pour votre attention
Questions ?

INP-ENM & Météo-France

thomas.rieutord@meteo.fr

www.enm.meteo.fr

www.umr-cnrm.fr