Partie 3 - Classification non-supervisée (*clustering*)

Anne-Sophie Charest

École interdisciplinaire outils et méthodes
Cheminement 2 - Science des données

Apprentissage statistique (*machine learning*)

28-30 août 2024

Plan

Objectif et méthodes

On veut grouper des observations en un certain nombre de groupes homogènes, sans connaitre ce nombre de groupes à l'avance, ni l'appartenance des individus aux groupes.

Plusieurs méthodes :

- Classification hiérarchique
- Classification par regroupements
- Classification basée sur des modèles
- Classification basée sur une densité

Aujourd'hui : k-moyennes, et classification hiérarchique.

Des exemples

- Bail, C. A. (2008). The configuration of symbolic boundaries against immigrants in Europe. American Sociological Review, 73(1), 37-59. ici
- Discovering diverse mechanisms of migration : The Mexico–US Stream 1970– 2000. Population and Development Review. ici

Ces deux liens viennent d'un plan de cours pour Machine Learning for Social Sciences par Jorge Cimentada, trouvé ici.

Méthode des k-moyennes

Idée générale

Soit un ensemble d'observations $x_1, \ldots, x_n$.
Soit $C_1, \ldots, C_K$ les ensembles des index des observations dans chacun des groupes.
Par ex. $C_3 = \{2,5\} \implies$ observations 2 et 5 dans le groupe 3

On veut une partition des observations, c'est-à-dire

- $C_1 \cup C_2 \cup \ldots \cup C_K = \{1,2,\ldots,n\}$
- $C_k \cap C_{k'} = \emptyset$ pour tout $k \neq k'$

Idée générale

Soit un ensemble d'observations $x_1, \ldots, x_n$.
Soit $C_1, \ldots, C_K$ les ensembles des index des observations dans chacun des groupes.
Par ex. $C_3 = \{2,5\} \implies$ observations 2 et 5 dans le groupe 3

On veut une partition des observations, c'est-à-dire

- $C_1 \cup C_2 \cup \ldots \cup C_K = \{1,2,\ldots,n\}$
- $C_k \cap C_{k'} = \emptyset$ pour tout $k \neq k'$

On va essayer de minimiser la variation à l'intérieur de chaque groupe.

Mesure de la variabilité intra-groupe :

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Détails techniques

Mesure de la variabilité intra-groupe :

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

On veut donc choisir les $C_1, \ldots, C_K$ qui minimisent

$$\sum_k \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Détails techniques

Mesure de la variabilité intra-groupe :

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

On veut donc choisir les $C_1, \ldots, C_K$ qui minimisent

$$\sum_k \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

C'est un problème d'optimisation très difficile.
On peut toutefois approximer la solution assez facilement.

Un algorithme (celui de Lloyd)

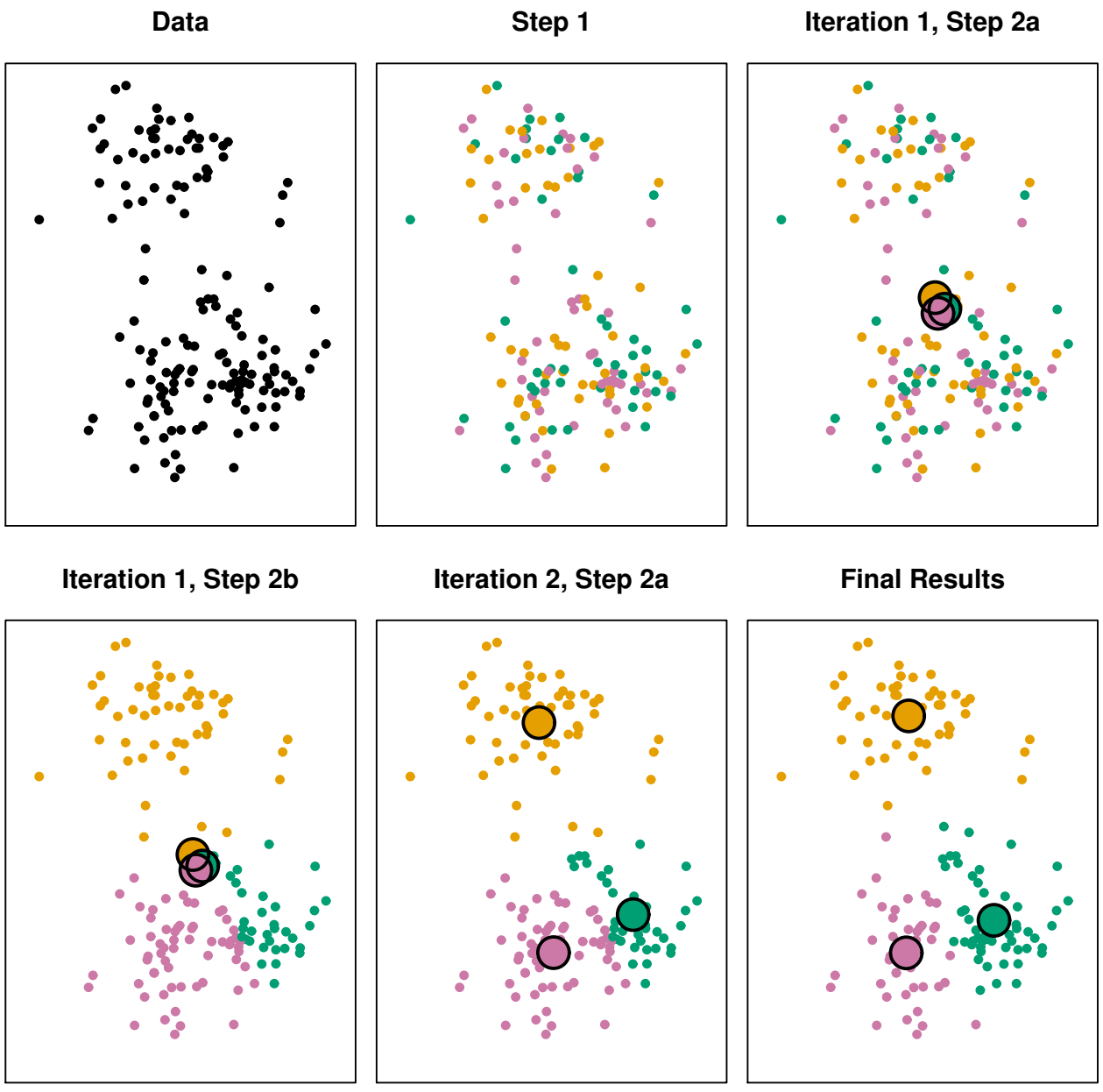
1. Assigner aléatoirement chaque observation à un des k groupes.
2. Calculer la moyenne des observations dans chacun des groupes.
3. Assigner chaque observation au groupe duquel elle est le plus proche.
4. Répéter les étapes 2 et 3 jusqu'à convergence.

Un algorithme (celui de Lloyd)

1. Assigner aléatoirement chaque observation à un des k groupes.
2. Calculer la moyenne des observations dans chacun des groupes.
3. Assigner chaque observation au groupe duquel elle est le plus proche.
4. Répéter les étapes 2 et 3 jusqu'à convergence.

Note : À l'étape 1, on peut aussi simplement choisir au hasard k observations pour être les centres des k groupes. On ira directement à l'étape 3 pour la première étape.

Exemple


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Convergence

On peut prouver assez facilement que l'algorithme convergera nécessairement vers un minimum local.

Pour s'assurer que ce minimum local est un minimum global, on répète habituellement l'algorithme avec plusieurs valeurs de départ différentes. (par exemple avec l'option `nstart` dans la fonction `kmeans` de R)

Exemple

Choix du nombre de groupes

Il faut décider dès le départ du nombre de groupes désiré pour utiliser l'algorithme k-moyennes.

On fait habituellement la classification avec plusieurs valeurs différentes de k.

On doit ensuite choisir la valeur de k.
Il existe plusieurs méthodes pour ce faire.

Hypothèses implicites

En minimisant la somme des carrés des distances à l'intérieur des groupes, la méthode des k-moyennes fait implicitement les hypothèses suivantes :

- Les variables suivent des lois normales
- Les variances de toutes les variables sont similaires
- Les groupes sont de mêmes tailles

On peut l'appliquer dans d'autres cas, mais les résultats risquent de ne pas être aussi bons.
Voir ces quelques exemples :
https://www.r-bloggers.com/2015/01/k-means-clustering-is-not-a-free-lunch/.

Autres caractéristiques

- Sensible aux valeurs extrêmes
- Uniquement pour des variables continues
- Dépend de la standardisation des variables

Variante : K-modes

Partitionne les observations en k groupes pour que la somme des distances entre les observations et le mode de leur groupe soit minimisée.

Utile pour des données catégoriques.

R : fonction `kmodes` de la librairie `klaR`

Variante : K-medoids

On ajoute la contrainte que le centre d'un groupe soit une des observations de ce groupe.

Utile par exemple si je veux regrouper des personnes autour d'un centre qui est aussi une personne.

Peut se faire à partir d'une matrice de dissimilarités.

R : fonction `pam` de la librairie `cluster`.

Choix de k

Aperçu

Il existe une panoplie de mesures pour

- évaluer la qualité d'une classification non supervisée
- comparer deux classifications entre elles

On peut évidemment utiliser ces méthodes pour choisir le nombre de groupes.

La librairie `NbClust` implémente 30 mesures pour choisir le nombre de classes. Elle peut être utilisée avec plusieurs méthodes de classification, mais seulement celles implémentées dans la librairie.

D'autres librairies implémentent certains autres critères, par exemple

- `cclust`
- `clusterSim`
- `clv`
- `clValid`

Ultimement, ça devient presque plus un art qu'une science...

Mesures de la qualité d'une classification

On distingue en général deux types de mesures :

- Mesures internes
  e.g. Index de Dunn, Connectivité, Largeur de silhouette,...
- Mesures de stabilité
  Stabilité si on enlève une des variables (voir clValid), si on modifie le jeu de données utilisé,...

Index de Dunn

On maximise l'index suivant :

$$D = \frac{min_{i \neq j}\, d(C_i, C_j)}{max_k\, d'(C_k)}$$

Le numérateur donne la distance minimale entre deux groupes. Le dénominateur donne la distance intra-groupe maximale (diamètre) de tous les groupes.

L'index de Dunn cherche donc à créer des groupes denses et bien séparés.

Connectivité

On veut minimiser l'index suivant :

$$Conn(C) = \sum_{i=1}^{n} \sum_{j=1}^{L} X_{i,nn_i(j)}$$

où $nn_i(j)$ donne l'index du j-ième plus proche voisin de i et

$$X_{i,nn_i(j)} = \begin{cases} 1/j, & \text{si } i \text{ et } nn_i(j) \text{ ne sont pas dans le même groupe} \\ 0, & \text{sinon} \end{cases}$$

La connectivité mesure donc si les points proches sont dans le même groupe.

La valeur de L doit être choisie. (10 par défaut dans la fonction `clValid`)

Largeur de silhouette

La silhouette de l'observation i mesure la confiance dans le choix du groupe pour l'observation i :

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)}$$

où
a_i est la distance moyenne entre l'observation i et les autres observations de son groupe

b_i est la distance moyenne entre l'observation i et les observations du groupe le plus proche de i (parmi ceux auxquels il n'appartient pas)

On souhaite maximiser la silhouette moyenne des observations.
On peut aussi regarder la distribution des silhouettes à l'intérieur des groupes pour trouver des groupes clairement séparés.

Classification hiérarchique

Aperçu

On n'obtiendra pas qu'une seule classification des individus, mais une série de classifications imbriquées.

2 stratégies :

- Algorithme ascendant
- Algorithme descendant

Dans les deux cas, on obtient une série de classifications hiérarchiques contenant de 1 à n groupes.

Algorithme descendant

Procédure :

- Au départ toutes les observations sont dans le même groupe.
- On divise itérativement le groupe le moins homogène en deux groupes.
- À la fin, chaque observation est son propre groupe.

Caractéristiques :

- Demande beaucoup de temps de calcul si on veut obtenir la hiérarchie totale.
- Moins utilisé que les algorithmes ascendants.

Un exemple : fonction `DIANA` de la librairie `cluster` dans R.

Algorithme ascendant

Procédure :

- Au départ chaque observation est son propre groupe.
- On regroupe itérativement les deux groupes les plus proches.
- À la fin, toutes les observations sont dans le même groupe.

Plusieurs méthodes de classification hiérarchique ascendante existent et sont employées en pratique. (détails à suivre)

Elles sont implémentés dans la fonction `hclust` de R.

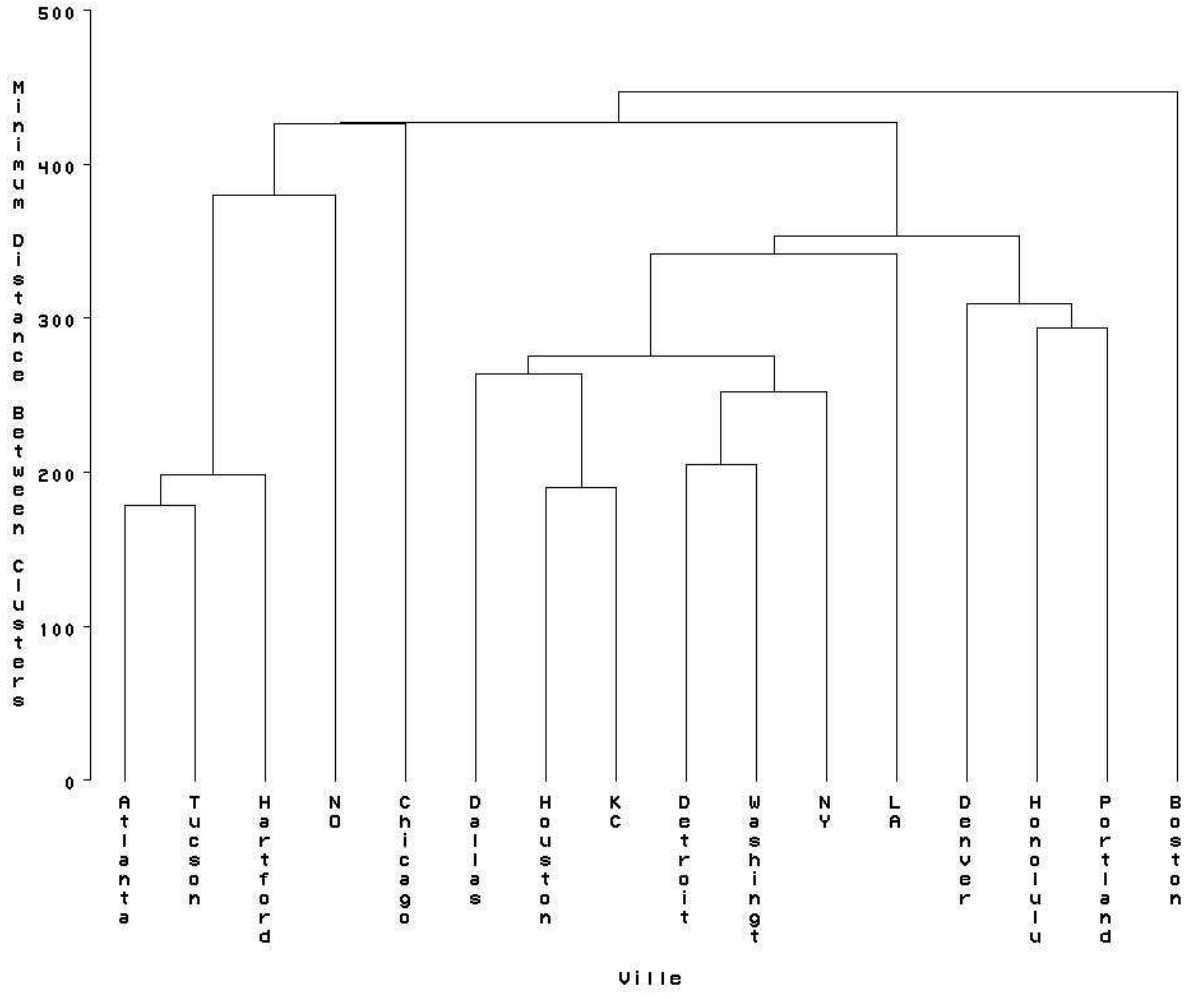
Un exemple - données

Données (Hartigan (1975)) :
Sept variables ont été mesurées dans 16 villes américaines.

- Meurtres, viols, vols
- Agressions, cambriolages, larcins
- Vols de voitures

Toutes les variables sont des taux par 100 000 habitants et pour l'année 1970.

Exemple - Dendrogramme



Selon le nombre de groupes qu'on désire, on n'a qu'à couper le dendrogramme à cet endroit.

Algorithme ascendant - différentes variantes

Il existe plusieurs variantes d'algorithmes de classification hiérarchique ascendants. Ils se distinguent par les choix de :

1. Mesure de distance ou de dissimilarité entre deux observations
2. Mesure de distance ou de dissimilarité entre deux groupes

On a déjà parlé des différentes mesures de distance ou de dissimilarité entre deux observations. Voyons maintenant comment mesurer la distance ou dissimilarité entre deux groupes.

Mesure de distance entre deux observations

On appelle d une mesure de distance entre les objets si, pour tous $i, j, k \in \{1, \ldots, n\}$, on a

- $d(i,i) = 0$;
- $d(i,j) = d(j,i)$;
- $d(i,k) \leq d(i,j) + d(j,k)$.

Exemple : Distance de Minkowski

$$\mathscr{L}_q = ||x_i - x_j||_q = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^q \right)^{1/q}.$$

- Si q = 1 : distance de Manhattan
- Si q = 2 : distance euclidienne

Attention !

La distance $\mathscr{L}_q$ n'est pas invariante au changement d'échelle.

Exemple :

Objet	poids (g)	taille (cm)
1	10	7
2	20	2
3	30	10

On trouve

$$d_{12} = 11.2, \quad d_{13} = 20.2, \quad d_{23} = 12.8.$$

Mais si la taille est exprimée en mm, on trouve

$$d_{12} = 51.0, \quad d_{13} = 36.1, \quad d_{23} = 80.6.$$

Alors l'objet 1 est-il plus près de l'objet 2 ou de l'objet 3 ?

Autres exemples de distance

- La distance de Mahanalobis (empirique)

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)}$$

- La distance standardisée par variable :

$$d(x_i, x_j) = \sum_{k=1}^{p} \frac{x_{ik} - x_{jk}}{s_k},$$

où $s_k =$ écart-type de la variable k.
Il s'agit du cas particulier de la distance de Mahanalobis si les variables ne sont pas corrélées (ou qu'on ignore la corrélation entre les variables).

Note : Dans les deux cas, la distance entre x_i et x_j dépend des autres observations du jeu de données.

Indice de similarité

On appelle s un indice de similarité entre les objets si, pour tous $i, j, k \in \{1, \ldots, n\}$, on a

- $s(i, j) \geq 0$;
- $s(i, j) = s(j, i)$;
- $s(i, j)$ augmente lorsque i et j se ressemblent davantage.

Généralement, $s(i, j) \leq 1$ pour tous $1 \leq i, j \leq n$.

Inversement, on parle de dissimilarité si de plus grandes valeurs indiquent des observations plus différentes.

Conversions

Une distance peut se transformer en similarité en posant

$$s_{ij} = \frac{1}{1 + d_{ij}},$$

mais l'inverse n'est pas vrai, à cause de l'inégalité du triangle.

On peut également transformer une dissimilarité en similarité, ou inversement, à l'aide d'une transformation inversement monotone.

Usage de similarités

Certains indices de similarité sont particulièrement utiles pour des observations avec plusieurs variables dichotomiques.

	variables					
	1	2	3	4	5	6
i	0	1	1	0	1	1
j	1	0	1	0	0	1

Ces indices se basent sur le nombre d'attributs concordants et le nombre d'attributs discordants.

Indices les plus communs

On compte d'abord les paires concordantes et discordantes.

		j		
		1	0	Total
i	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$a+b+c+d$

- Indice de Jaccard : $\dfrac{a}{a+b+c}$
- Indice de Russell ou de Rao : $\dfrac{a}{a+b+c+d}$
- Indice de Sokal ou de Michener : $\dfrac{a+d}{a+b+c+d}$

Distance entre deux groupes

On veut réunir les observations qui sont les moins distantes (ou les plus semblables).

Une fois qu'on aura commencé à créer des groupes d'observations, il faudra mesurer la distance entre ceux-ci.

On doit donc définir

$$d(A,B) = \text{distance entre deux groupes A et B,}$$

où $A, B \subset \{1, \ldots, n\}$ et $A \cap B = \emptyset$.

Comment faire ?

Méthode du plus proche voisin (single)

Dans cette méthode, dite “single linkage” en anglais, on pose

$$d(A,B) = \min\left\{d_{ij} : i \in A, j \in B\right\}.$$

- Parfois aussi appelée la technique du plus proche voisin (*nearest-neighbor*).
- Permet de retrouver des groupes à forme très irrégulière.
- Tend à créer un grand groupe en ajoutant une à une des observations situées proches de la dernière observation ajoutée (*chaining effect*).
- Tend à créer des groupes avec de grands diamètres (distance maximal entre 2 observations du groupe).

Note : Si on travaille plutôt avec des similitudes, alors

$$s(A,B) = \max\left\{s_{ij} : i \in A, j \in B\right\}.$$

Méthode du voisin le plus distant (complete)

Dans cette méthode, dite “complete linkage” en anglais, on pose

$$d(A,B) = \max\left\{d_{ij} : i \in A, j \in B\right\}.$$

- Tend à former des groupes compacts de tailles égales.
- Tend à créer des groupes pour lesquels certaines observations sont plus proches d’observations dans d’autres groupes que de certaines observations de leur propre groupe.

Note : Si on travaille plutôt avec des similitudes, alors

$$s(A,B) = \min\left\{s_{ij} : i \in A, j \in B\right\}.$$

Méthode de la moyenne

La distance entre deux groupes est donnée par

$$d(A,B) = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d(x_i, x_j).$$

- Sorte de compromis entre les deux premières méthodes.
- Tend à former des groupes de mêmes variances.
- Contrairement aux deux premières méthodes, son résultat n'est pas invariant sous transformation monotone des distances.

Méthode du centroïde

La distance entre deux groupes est donnée par

$$d(A,B) = d(\bar{x}_A, \bar{x}_B),$$

où

$$\bar{x}_A = \frac{1}{n_A} \sum_{i \in A} x_i, \quad \bar{x}_B = \frac{1}{n_B} \sum_{j \in B} x_j.$$

- N'est pas non plus invariant sous transformation monotone des distances.
- Peut mener à un phénomène d'inversion, soit deux groupes qui fusionnent à une hauteur inférieure à celle de ces deux groupes.

Méthode de Ward (1963) [1]

On ne définit pas ici explicitement de mesure de distance entre deux groupes. À chaque étape, on fusionne deux groupes choisis de façon à minimiser la somme des variances intra-groupe.

- Tente d'obtenir des groupes compacts et sphériques.
- Optimal si les groupes sont approximativement de loi normale multivariée de même matrice de variance-covariance.
- Quelques versions ; voir l'aide de `clust` et les références qui y sont données.

1. Joe H. Ward (1963)., Hierarchical grouping to optimize an objective function., *Journal of the American Statistical Association*, 58, 236–244.

Notes

- On a vu les méthodes les plus populaires, mais il en existe plusieurs autres.
- À l'aide du dendogramme, on peut obtenir un groupe de taille n, puis 2 groupes, puis 3 groupes, et ainsi de suite. On peut utiliser les mêmes critères que pour k-moyennes pour choisir le nombre de groupes.

Autres méthodes

Très variées !

Encore une fois, il en existe plein, incluant certaines qui associent chaque observation à plus d'un seul groupe, ou en partie à plusieurs groupes.

Voici une liste de certaines fonctions disponibles dans R, avec entre crochets la librairie dans laquelle les trouver :

- kmeans [stats]
- pam, clara and fanny [cluster]
- dbscan [fpc]
- Mclust [mclust]
- HCPC [FactoMineR]
- hkmeans [factoextra]

Note

”Some of the figures in this presentation are taken from ”An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors : G. James, D. Witten, T. Hastie and R. Tibshirani ”