Classification hiérarchique

Cathy Maugis-Rabusseau

4modIA / INSA Toulouse & ENSEEIHT

2023-2024

Plan

- 1 Hiérarchie indicée et CAH
- 2 Mesures d'agrégation entre classes
- 3 Coupure du dendrogramme
- 4 Applications
- Conclusion

Introduction

Données : On observe n individus décrits par p variables

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \text{ avec } x_i = (x_{i1}, \dots, x_{ip}) \in \mathcal{X}$$

- Objectif : Hiérarchiser les données c'est à dire obtenir une suite de partitions emboîtées des données.
- Notation : on note d la dissimilarité choisie entre les individus

Hiérarchie

Définition : Hiérarchie

Une **hiérarchie** ${\mathcal H}$ est un ensemble de parties de ${\mathbf X}$ satisfaisant:

- $\mathbf{2} \mathbf{X} \in \mathcal{H}$

Exemple:

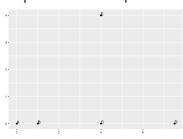
 $\{\{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{3,4\}, \{1,2,3,4\}\}$ est une hiérarchie de $\{1,2,3,4\}$.

Définition: Hiérarchie indicée

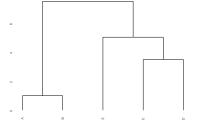
Une **hiérarchie indicée** est un couple (\mathcal{H}, h) où \mathcal{H} est une hiérarchie et $h: \mathcal{H} \to \mathbb{R}^+$ satisfait :

- \lozenge $\forall A, B \in \mathcal{H}, A \neq B, A \subset B \Rightarrow h(A) \leq h(B)$

Représentation par dendrogramme



- $\mathcal{H} = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{C, D\}, \{A, B\}, \{C, D, E\}, \{A, B, C, D, E\}\}$
- $h(\{x\}) = 0, \forall x \in \{A, B, C, D, E\}$
- $h(\{A, B\}) = 1$
- $h({C, D}) = 3.5$
- $h({C, D, E}) = 5.04$
- $h({A, B, C, D, E}) = 7.52$



La représentation du dendrogramme n'est pas unique : si \mathbf{X} est un ensemble de n points, il existe 2^{n-1} possibilités pour ordonner les feuilles de l'arbre.

Construction d'une hiérarchie indicée

• 1ère stratégie : on part du bas du dendrogramme (les singletons) et on agrège deux à deux les parties les plus proches jusqu'à obtenir qu'une seule classe ⇒ Classification Ascendante Hiérarchique (CAH)

Question centrale : Comment choisir les classes à agréger ?

• 2ème stratégie : on part du haut du dendrogramme en procédant par divisions successives de \underline{x} jusqu'à obtenir des classes réduites à des singletons

⇒ Classification Descendante Hiérarchique (CDH)

Question centrale : Comment choisir la classe à diviser à chaque étape ?

Algorithme général de CAH

• Initialisation : on part de la partition en singletons

$$\mathcal{P}_n = \{\{x_1\}, \dots, \{x_n\}\}$$

- Étapes agrégatives :
 - on part de la partition précédente $\mathcal{P}_K = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$ en K classes
 - ▶ on agrège les deux classes C_k et $C_{k'}$ qui minimisent une **mesure** d'agrégation $D(C_k, C_{k'})$: $C_{k \cup k'} = C_k \cup C_{k'}$
 - on obtient ainsi une partition en K-1 classes
- On recommence l'étape d'agrégation jusqu'à obtenir une partition en une seule classe

Les choix à faire

- Choix d'une dissimilarité d entre les points
- Choix d'une **mesure d'agrégation** *D* entre classes
- Construction d'un dendrogramme
- Critère pour la coupure du dendrogramme pour en déduire une classification des données

Plan

- Hiérarchie indicée et CAH
- 2 Mesures d'agrégation entre classes
- 3 Coupure du dendrogramme
- 4 Applications
- Conclusion

Lien simple (Single linkage)

$$D(\mathcal{C}_k, \mathcal{C}_{k'}) = \min_{i \in \mathcal{C}_k, \ell \in \mathcal{C}_{k'}} d(x_i, x_\ell)$$

- Arbre couvrant minimal
- Classes avec des diamètres très différents
- Effet de chaînage : tendance à l'agrégation plutôt qu'à la création de nouvelles classes
- Sensibilité aux individus bruités

Lien complet (Complete linkage)

$$D(\mathcal{C}_k, \mathcal{C}_{k'}) = \max_{i \in \mathcal{C}_k, \ell \in \mathcal{C}_{k'}} d(x_i, x_\ell)$$

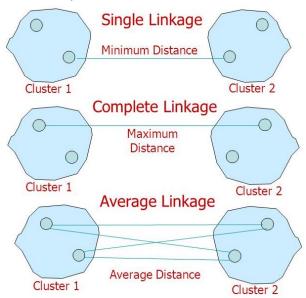
- Crée des classes compactes (contrôle du diamètre) : cette fusion engendre l'accroissement le plus faible des diamètres
- Pas de contrôle de la séparation: classes arbitrairement proches
- Sensibilité aux individus bruités

Lien moyen (Average linkage)

$$D(C_k, C_{k'}) = \frac{1}{|C_k||C_{k'}|} \sum_{i \in C_k \ell \in C_{k'}} d(x_i, x_\ell)$$

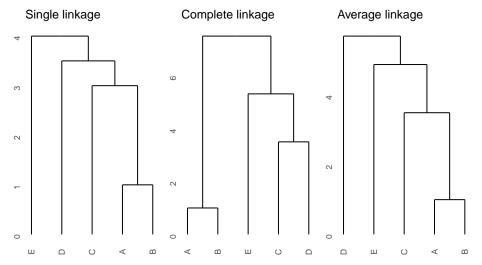
- Compromis entre les deux liens précédents : bon équilibre entre séparation des classes et diamètre des classes
- Tendance à produire des classes de variance proche

Single / Complete / Average



Exemple jouet

 \bullet d = distance euclidienne usuelle



Mesures d'agrégation de Ward

$$D(\mathcal{C}_k, \mathcal{C}_{k'}) = \frac{|\mathcal{C}_k||\mathcal{C}_{k'}|}{|\mathcal{C}_k| + |\mathcal{C}_{k'}|} d(m_k, m_{k'})^2$$

où m_k (resp. $m_{k'}$) centre de gravité de \mathcal{C}_k (resp. $\mathcal{C}_{k'}$) et d est une distance euclidienne.

- Tendance à construire des classes ayant des effectifs égaux pour un niveau de hiérarchie donné
- Favorise les classes sphériques

Méthode de Ward

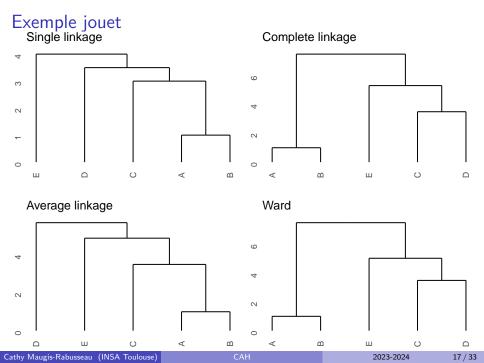
Proposition

Soit $\mathcal{P}_K = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$ une partition des données et soit $k \neq k'$. Si l'on rassemble les deux classes \mathcal{C}_k et $\mathcal{C}_{k'}$ en une classe notée $\mathcal{C}_{k \cup k'}$ alors l'inertie interclasse diminue (l'inertie intraclasse augmente) de :

$$\frac{|\mathcal{C}_k||\mathcal{C}_{k'}|}{|\mathcal{C}_k|+|\mathcal{C}_{k'}|}d(m_k,m_{k'})^2.$$

- m_k (resp. $m_{k'}$) centre de gravité de C_k (resp. $C_{k'}$)
- d distance euclidienne

Méthode de Ward : Elle consiste à choisir à chaque étape les deux classes dont le regroupement implique une augmentation minimale de l'inertie intraclasse.



Formule de Lance et Williams

Cette formule permet de mettre à jour les distances pour l'agrégation

$$D(\mathcal{C}_{u}, \mathcal{C}_{k \cup k'}) = \alpha_{1}D(\mathcal{C}_{u}, \mathcal{C}_{k}) + \alpha_{2}D(\mathcal{C}_{u}, \mathcal{C}_{k'}) + \alpha_{3}D(\mathcal{C}_{k}, \mathcal{C}_{k'}) + \alpha_{4}|D(\mathcal{C}_{u}, \mathcal{C}_{k}) - D(\mathcal{C}_{u}, \mathcal{C}_{k'})|$$

Lien	α_1	$lpha_{2}$	$lpha_3$	$\alpha_{ extsf{4}}$
simple	0.5	0.5	0	-0.5
complet	0.5	0.5	0	0.5
moyen	$rac{ \mathcal{C}_k }{ \mathcal{C}_{k'} + \mathcal{C}_k }$	$\frac{ \mathcal{C}_{k'} }{ \mathcal{C}_{k'} + \mathcal{C}_k }$	0	0
Ward	$\frac{ \mathcal{C}_u + \mathcal{C}_k }{ \mathcal{C}_k + \mathcal{C}_{k'} + \mathcal{C}_u }$	$\frac{ \mathcal{C}_u + \mathcal{C}_{k'} }{ \mathcal{C}_k + \mathcal{C}_{k'} + \mathcal{C}_u }$	$-\frac{ \mathcal{C}_u }{ \mathcal{C}_k + \mathcal{C}_{k'} + \mathcal{C}_u }$	0

Indicer la hiérarchie

- En général, $\forall A, B \in \mathcal{H}, \ h(A \cup B) = D(A, B)$
- Si (H, h) ainsi définie ne vérifie pas les propriétés d'une hiérarchie indicée, on peut utiliser la relation suivante:

$$\forall A, B \in \mathcal{H}, \ h(A \cup B) = \max[D(A, B), h(A), h(B)]$$

• Lien entre hiérarchie indicée et distance ultramétrique

Plan

- Hiérarchie indicée et CAH
- 2 Mesures d'agrégation entre classes
- 3 Coupure du dendrogramme
- 4 Applications
- Conclusion

Comment faire?

- Le choix du niveau de coupure du dendrogramme détermine le nombre de classes et ces classes sont alors uniques
- On peut définir la coupure du dendrogramme en déterminant à l'avance le nombre de classes dans lesquelles on désire répartir l'ensemble des données
- Le choix du niveau de coupure peut être facilité par l'examen des indices croissants de niveau de l'arbre hiérarchique
- ullet On peut aussi faire ce choix en utilisant les indices tels que R^2 , CH, Silhouette, Gap Statistic, . . .

Quelques critères

- Critères fondés sur les inerties
 - R-Square :

$$K \mapsto RSQ(K) = 1 - \frac{I_{intra}(\mathcal{P}_K)}{I_{totale}} = \frac{I_{inter}(\mathcal{P}_K)}{I_{totale}}$$

On retient l'endroit où la courbe $K \mapsto RSQ(K)$ forme un coude.

Semi-Partial R-Square :

$$K \mapsto SPRSQ(K) = \frac{I_{inter}(\mathcal{P}_K) - I_{inter}(\mathcal{P}_{K-1})}{I_{totale}}$$

On retient l'endroit où on a la plus forte réduction du SPRSQ.

Pseudo-F (Calinski-Harabasz) :

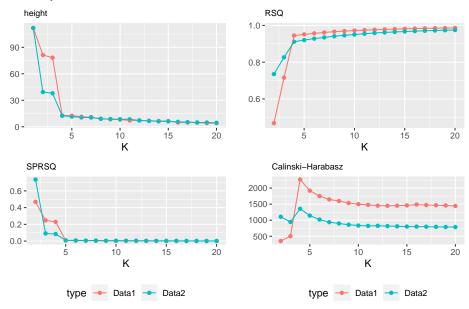
$$K \mapsto PseudoF(K) = \frac{I_{inter}(\mathcal{P}_K)/(K-1)}{I_{intra}(\mathcal{P}_K)/(n-K)}$$

On cherche un pic sur cette courbe

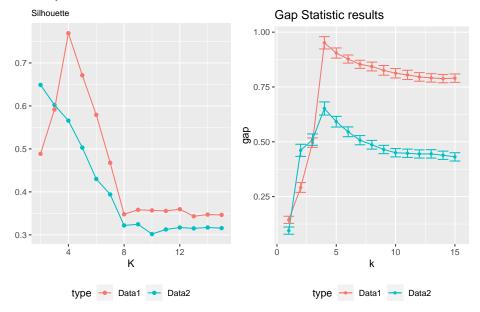
- Critère Silhouette
- Le Gap Statistique

Cathy Maugis-Rabusseau (INSA Toulouse)

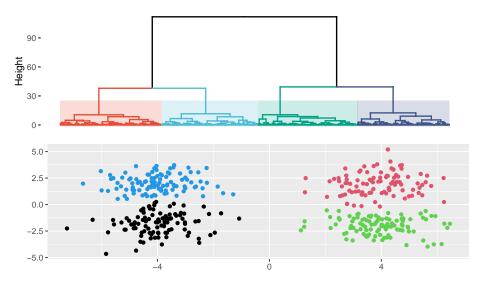
Exemple des données simulées



Exemple des données simulées



Exemple des données simulées



Plan

- Hiérarchie indicée et CAH
- 2 Mesures d'agrégation entre classes
- 3 Coupure du dendrogramme
- 4 Applications
- Conclusion

Quelques commandes



- hc=hclust(d,method=)
 - ★ d: tableau de distances comme produit par dist()
 - ★ method : agrégation "ward.D2", "single", "complete", "average", . . .
- plot(hc,hang=,...) ou ggdendrogram(hc,...) ou fviz_dend()
 pour tracer le dendrogramme
- ► cutree(hc,k=..) pour obtenir la classification en k classes



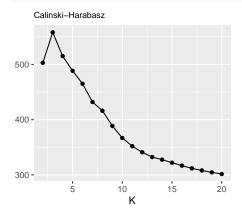
Avec Python

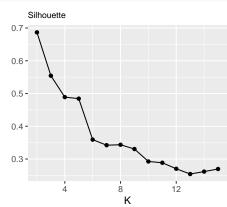
scipy.cluster.hierarchy [scipy]

- linkage(.) : method='single','complete','average','ward',...
- dendrogram(.) : pour tracer le dendrogramme
- fcluster(.) : pour obtenir un clustering à partir du dendrogramme

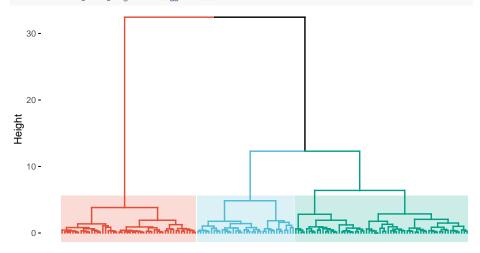
Exemple des iris

dx<-dist(iris[,-5],method="euclidian")
hward<-hclust(dx,method="ward.D2")</pre>



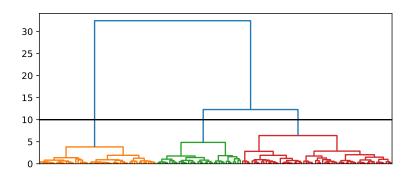


Exemple des iris



Exemple des iris avec

```
pyiris=r.auxiris
from matplotlib import pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage
Z=linkage(pyiris,method='ward')
dendrogram(Z,no_labels=True,color_threshold=10);
plt.axhline(y=10, c='k')
```



Plan

- Hiérarchie indicée et CAH
- 2 Mesures d'agrégation entre classes
- 3 Coupure du dendrogramme
- 4 Applications
- Conclusion

Avantages et inconvénients CAH

Avantages :

- ▶ Méthode flexible pour le niveau de finesse de la classification
- Prise en compte facile de distances et d'indices de similarité de n'importe quel type
- Rapidité d'execution et reproductible
- Inconvénients :
 - Choix de la coupure de l'arbre
 - La partition obtenue à une étape dépend de celle à l'étape précédente