Chapitre 5-4) Classification Ascendante Hiérarchique (CAH)

Maxime El Masri

3 MIC / INSA Toulouse

2023-2024

Partie 4

Introduction

- Données : On observe n individus décrits par p variables

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

  L'ensemble des individus sera également désigné par la matrice $\mathbf{X}$.

- 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:

1. $\forall\ i = 1 \dots n,\ \{x_i\} \in \mathcal{H}$
2. $\mathbf{X} \in \mathcal{H}$
3. $\forall A, B \in \mathcal{H},\ A \cap B \in \{\emptyset, A, B\}$ (autrement dit $A \cap B = \emptyset$ ou $A \subset B$ ou $B \subset A$).

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

Remarque : $\{\{1,2\},\{3,4\}\}$ est une partition de $\{1,2,3,4\}$. ($\rightarrow$ méthodes par partitionnements de type kmeans, DBSCAN...)

Hiérarchie indicée

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 :

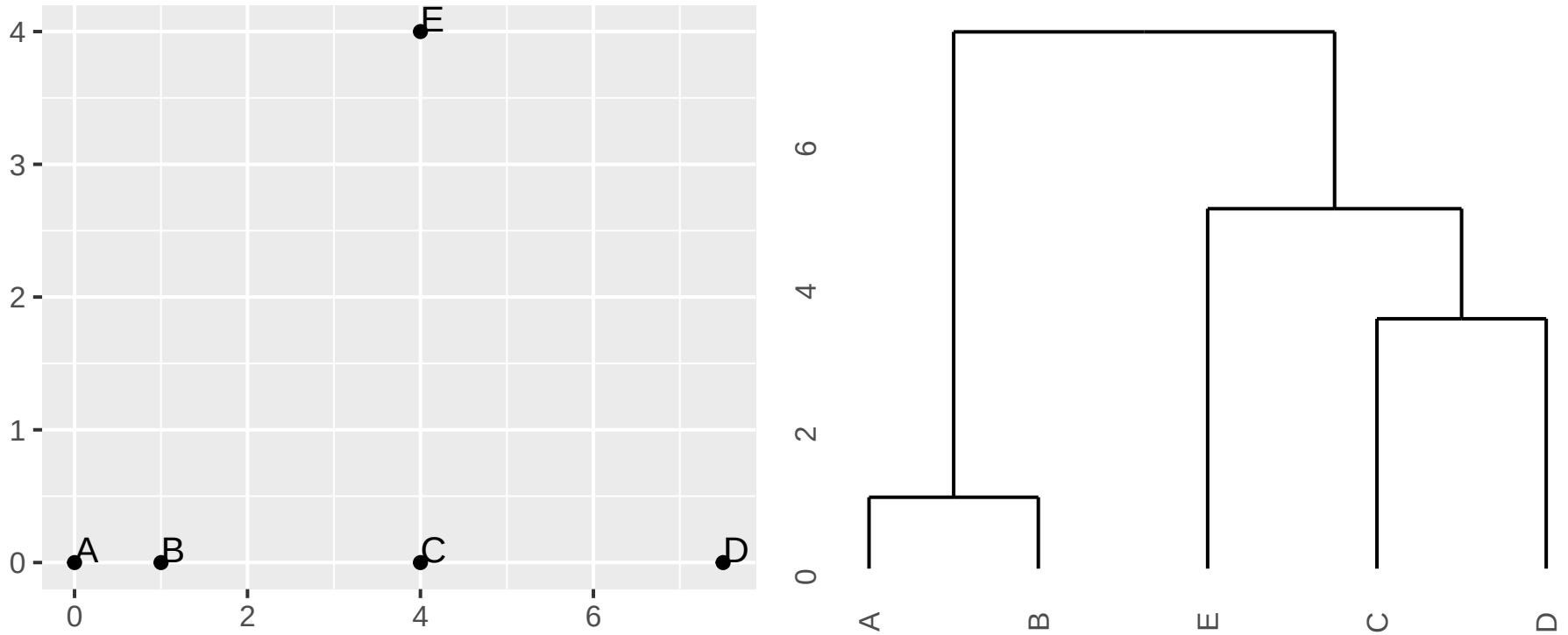
1. $\forall A \in \mathcal{H},\ h(A) = 0 \Leftrightarrow A$ est un singleton
2. $\forall A, B \in \mathcal{H},\ A \neq B,\ A \subset B \Rightarrow h(A) \leq h(B)$

Exemple : $\mathcal{H} = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{C, D\}, \{A, B\}, \{C, D, E\}, \{A, B, C, D, E\}\}$

- $h(\{x\}) = 0,\ \forall x \in \{A, B, C, D, E\}$
- $h(\{A, B\}) = 1,\ h(\{C, D\}) = 3.5$
- $h(\{C, D, E\}) = 5$
- $h(\{A, B, C, D, E\}) = 7.5$

$\longrightarrow$ Représentation graphique d'une hiérarchie indicée : le **dendrogramme**.

Représentation par dendrogramme



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. $\Rightarrow$ **Classification Ascendante Hiérarchique (CAH)**

Question : Comment choisir les classes à agréger ?

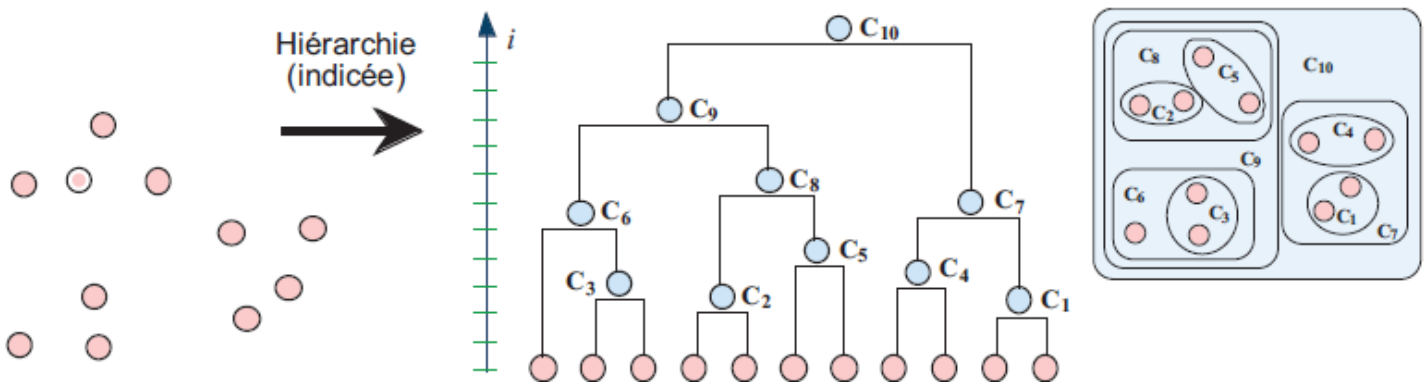
- 2ème stratégie : on part du haut du dendrogramme en procédant par divisions successives de $\mathbf{X}$ jusqu'à obtenir des classes réduites à des singletons
  $\Rightarrow$ **Classification Descendante Hiérarchique (CDH)**

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

Algorithme général de CAH

- Initialisation : partition en n singletons $\mathcal{P}_n = \{\{x_1\}, \ldots, \{x_n\}\}$
- Étapes agrégatives :
  - on part de la partition précédente $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ en K classes
  - on agrège les deux classes $\mathcal{C}_k$ et $\mathcal{C}_{k'}$ qui minimisent une **mesure d'agrégation** $D(\mathcal{C}_k, \mathcal{C}_{k'})$
  - on obtient ainsi une partition en $K - 1$ classes
- On recommence l'étape d'agrégation jusqu'à obtenir une partition en une seule classe

(Bisson 2001)



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
- Choix d'un **critère pour la coupure du dendrogramme** pour en déduire une classification des données

Partie 4

Mesures de lien

Soit d une dissimilarité sur $\mathcal{X}$

- Lien simple (*Single linkage*) ou saut minimum :

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

  - Classes avec des diamètres très différents, aux formes irrégulières

- Lien complet (*Complete linkage*) ou saut maximum :

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

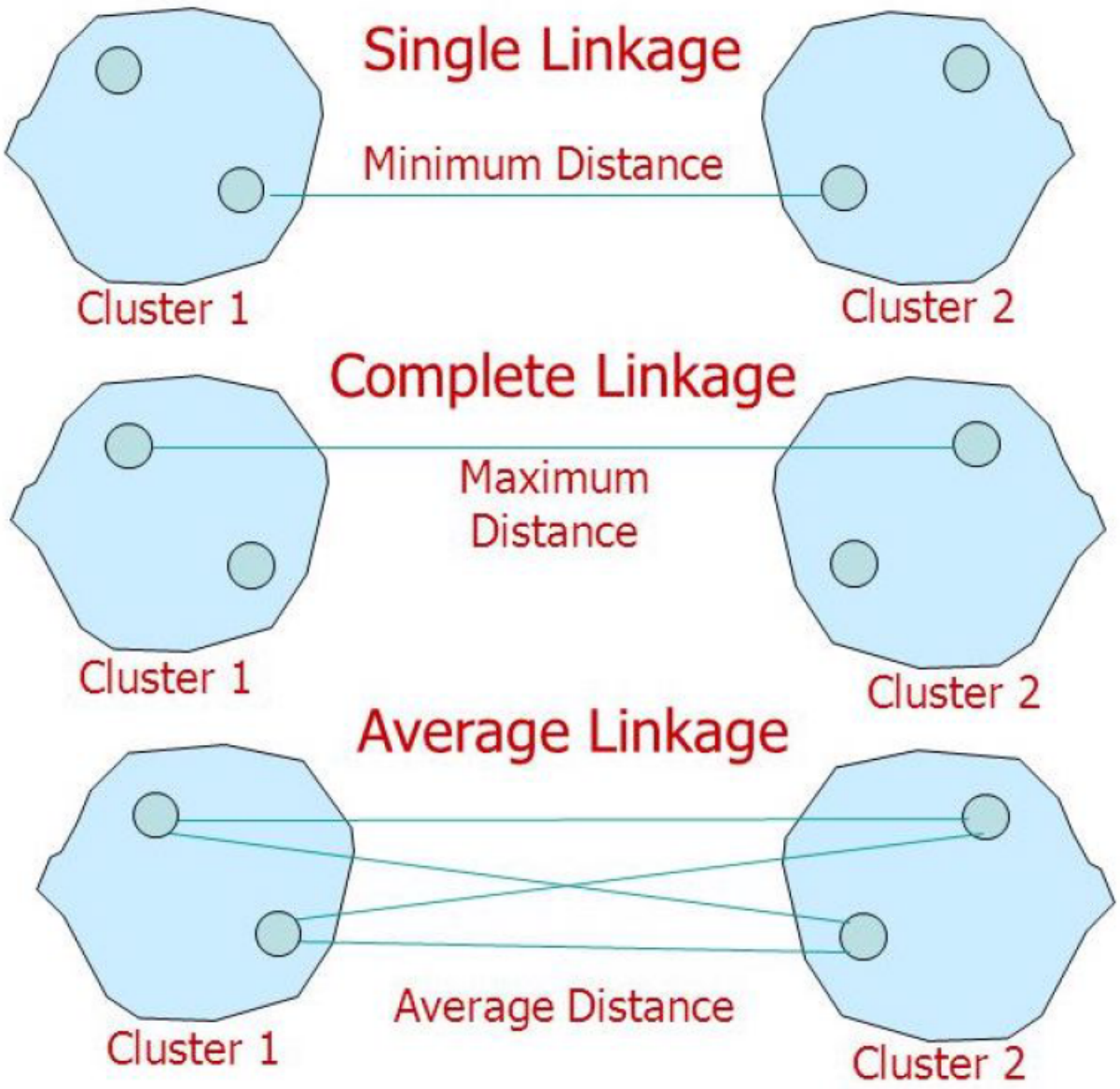
  - Classes compactes, de tailles similaires et rapprochées
  - Sensible aux données aberrantes

- Lien moyen (*Average linkage*) ou saut moyen :

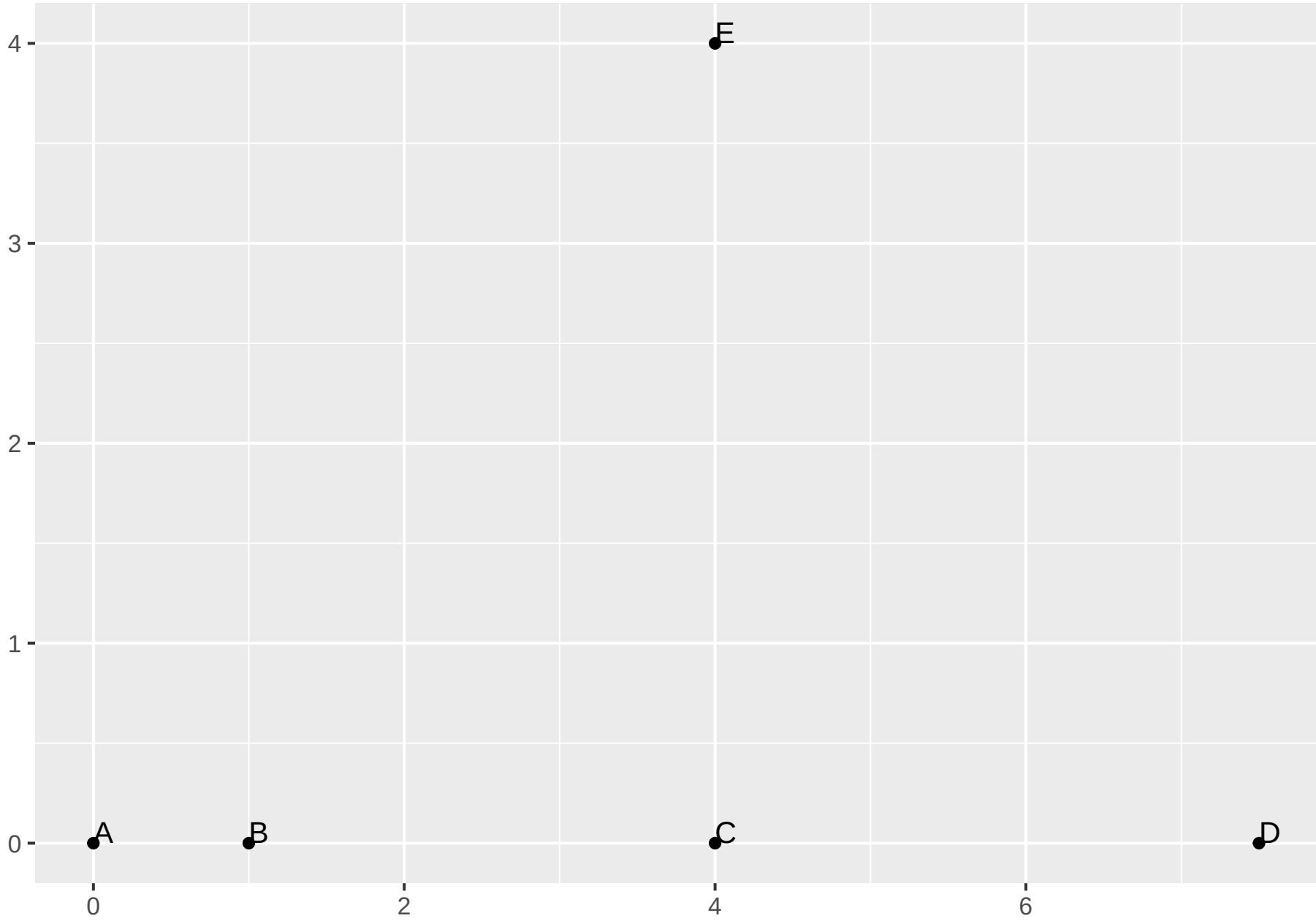
$$D(\mathcal{C}_k, \mathcal{C}_{k'}) = \frac{1}{|\mathcal{C}_k||\mathcal{C}_{k'}|} \sum_{i \in \mathcal{C}_k} \sum_{\ell \in \mathcal{C}_{k'}} d(x_i, x_\ell)$$

  - Classes de variance proche
  - Plus robuste aux données aberrantes

Mesures de lien



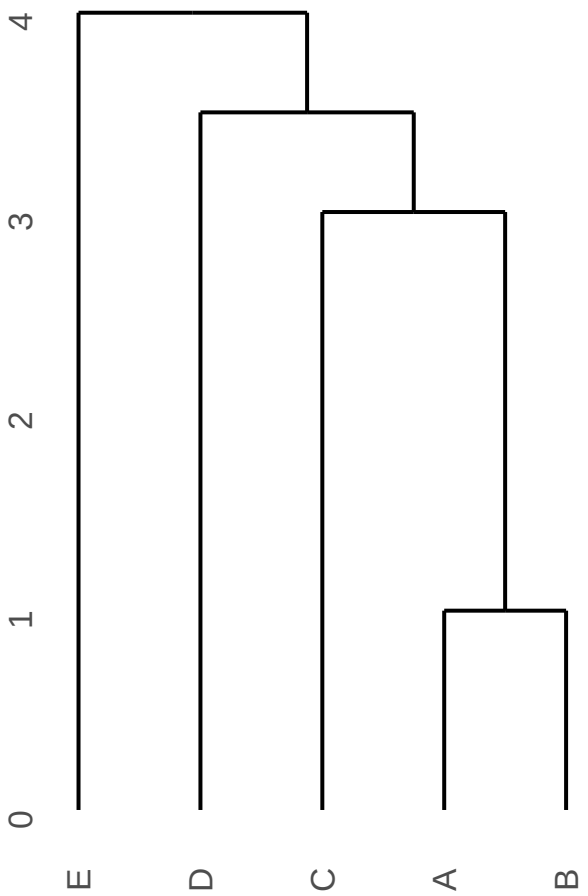
Exemple jouet



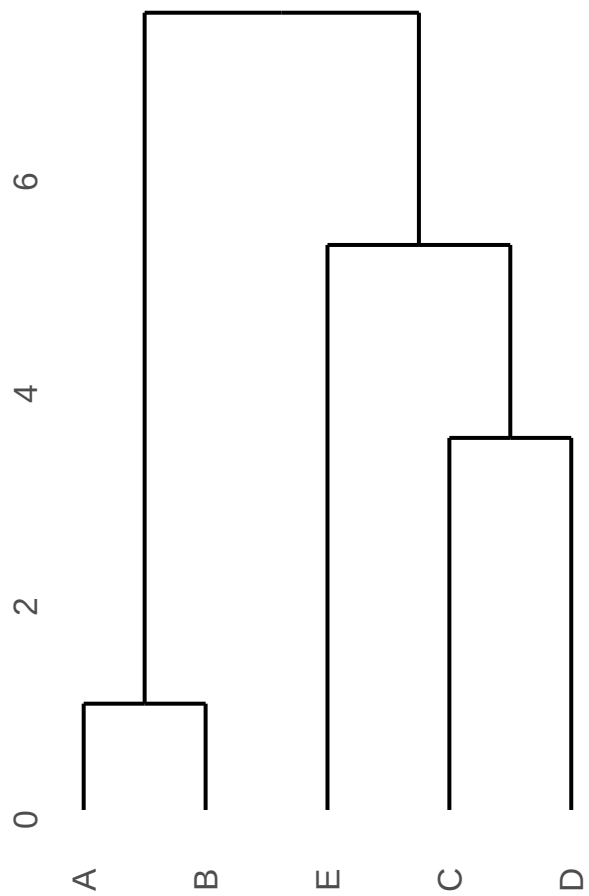
Exemple jouet

- d : distance euclidienne usuelle

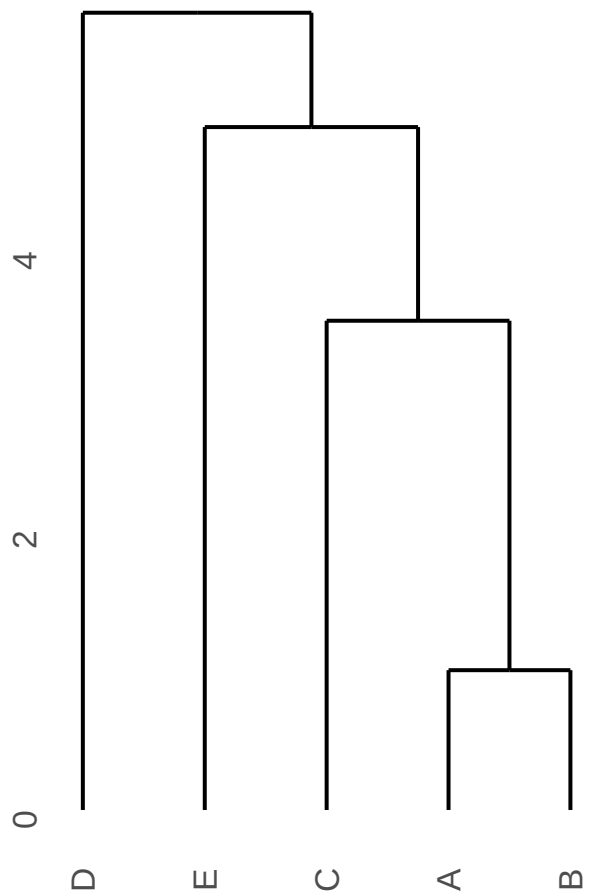
Lien simple



Lien Complet



Lien moyen



Méthode de Ward

Propriété

Soit $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ 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 intraclasse augmente (l'inertie interclasse diminue) de :

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

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

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

Mesure d'agrégation de Ward

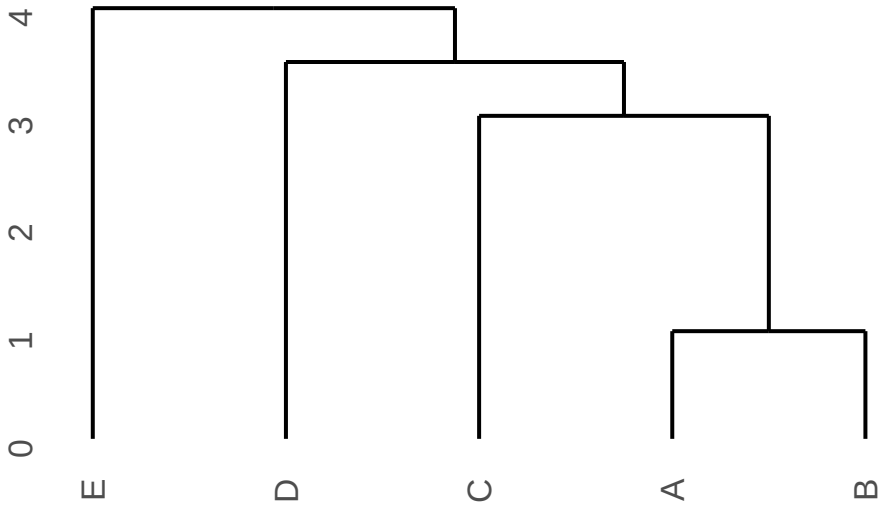
$$D_W(\mathcal{C}_k, \mathcal{C}_{k'}) = \frac{|\mathcal{C}_k||\mathcal{C}_{k'}|}{|\mathcal{C}_k| + |\mathcal{C}_{k'}|} d(m_k, m_{k'})^2$$

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

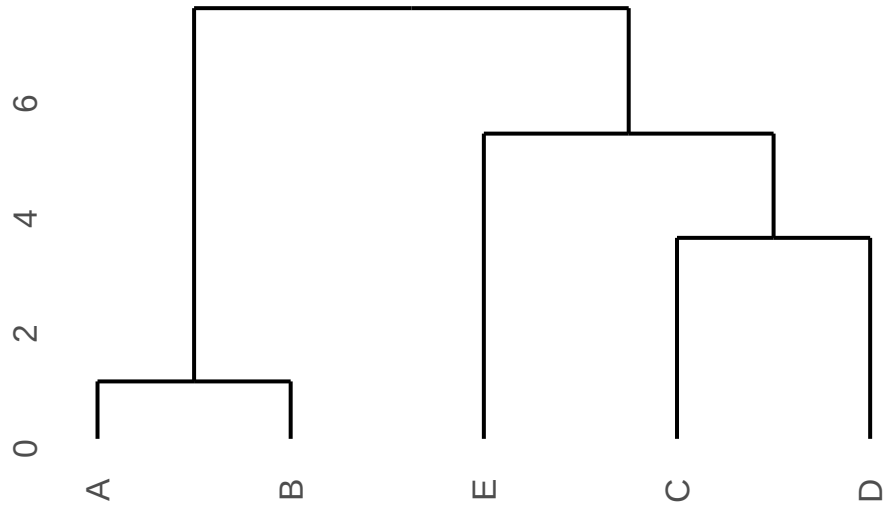
- Tendance à construire des classes ayant des effectifs très proches
- Méthode optimale lorsque les classes sont gaussiennes
- Méthode utilisée par défaut dans R notamment.

Exemple jouet

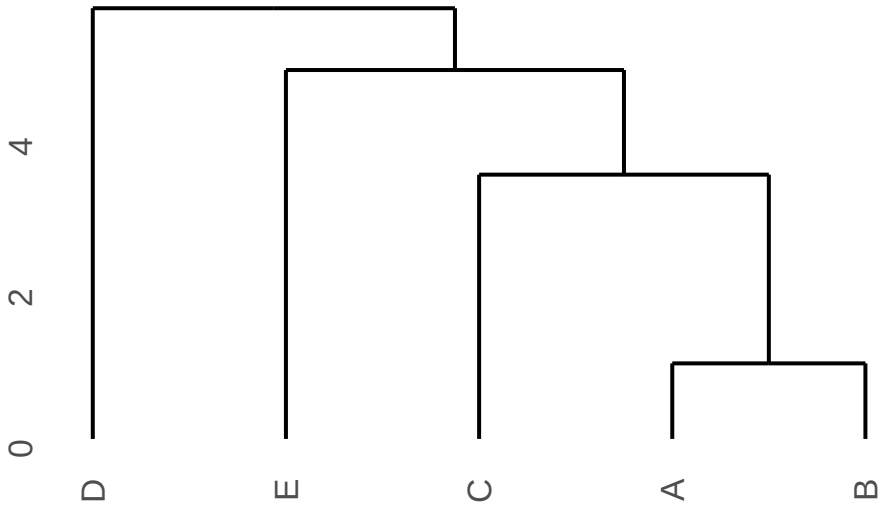
Single linkage



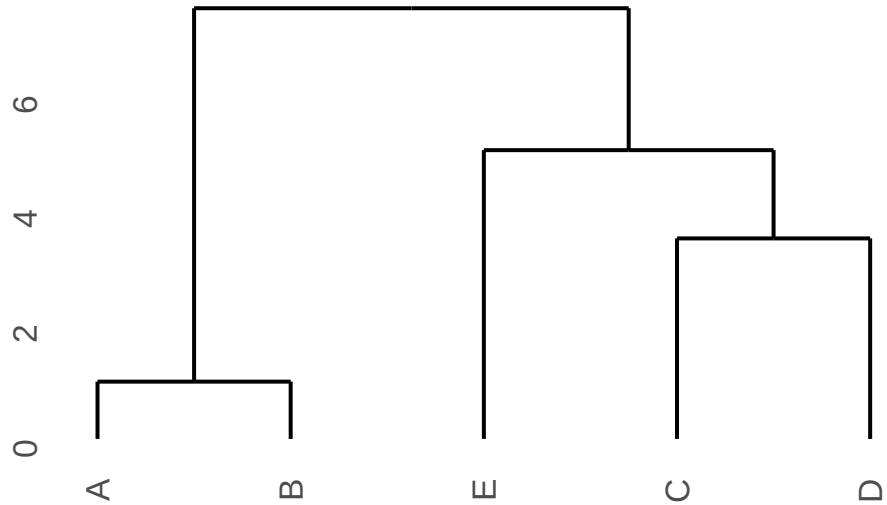
Complete linkage



Average linkage



Ward



Formule de Lance et Williams

Permet de mettre à jour les distances pour l'agrégation :

$$\begin{aligned} D(\mathcal{C}_u, \mathcal{C}_{k\cup k'}) &= \alpha_1 D(\mathcal{C}_u, \mathcal{C}_k) + \alpha_2 D(\mathcal{C}_u, \mathcal{C}_{k'}) + \alpha_3 D(\mathcal{C}_k, \mathcal{C}_{k'}) \\ &\quad + \alpha_4 |D(\mathcal{C}_u, \mathcal{C}_k) - D(\mathcal{C}_u, \mathcal{C}_{k'})| \end{aligned}$$

Lien	α_1	α_2	α_3	α_4
simple	0.5	0.5	0	-0.5
complet	0.5	0.5	0	0.5
moyen	$\frac{\lvert\mathcal{C}_k\rvert}{\lvert\mathcal{C}_{k'}\rvert+\lvert\mathcal{C}_k\rvert}$	$\frac{\lvert\mathcal{C}_{k'}\rvert}{\lvert\mathcal{C}_{k'}\rvert+\lvert\mathcal{C}_k\rvert}$	0	0
Ward	$\frac{\lvert\mathcal{C}_u\rvert+\lvert\mathcal{C}_k\rvert}{\lvert\mathcal{C}_k\rvert+\lvert\mathcal{C}_{k'}\rvert+\lvert\mathcal{C}_u\rvert}$	$\frac{\lvert\mathcal{C}_u\rvert+\lvert\mathcal{C}_{k'}\rvert}{\lvert\mathcal{C}_k\rvert+\lvert\mathcal{C}_{k'}\rvert+\lvert\mathcal{C}_u\rvert}$	$-\frac{\lvert\mathcal{C}_u\rvert}{\lvert\mathcal{C}_k\rvert+\lvert\mathcal{C}_{k'}\rvert+\lvert\mathcal{C}_u\rvert}$	0

Indicer la hiérarchie

- En général, $\forall A, B \in \mathcal{H},\ h(A \cup B) = D(A, B)$ (vrai pour les mesures des liens simple et complet, et pour la mesure de Ward)

- Si (H, h) ainsi définie ne vérifie pas les propriétés d'une hiérarchie indicée, on peut utiliser la relation suivante:

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

(c'est le cas si D est la mesure du lien moyen)

Algorithme CAH de Ward

- Initialisation : Calculer les mesures de Ward entre les n singletons $\{x_1\}, \ldots, \{x_n\}$:

$$D_W(\{x_i\}, \{x_j\}) = \frac{1}{2} d(x_i, x_j)^2$$

- Étapes agrégatives :
  - on part de la partition précédente $\mathcal{P}_K = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ en K classes
  - on agrège les deux classes $\mathcal{C}_k$ et $\mathcal{C}_{k'}$ qui minimisent la mesure de Ward $D_W(\mathcal{C}_k, \mathcal{C}_{k'})$ pour une partition en $K-1$ classes
  - on calcule la mesure de Ward entre $\mathcal{C}_k \cup \mathcal{C}_{k'}$ et les autres $\mathcal{C}_j$ avec la formule de Lance et Williams :

$$D_W(\mathcal{C}_j, \mathcal{C}_{k \cup k'}) = \sum_{i \in \{k, k'\}} \frac{|\mathcal{C}_j| + |\mathcal{C}_i|}{|\mathcal{C}_j| + |\mathcal{C}_k| + |\mathcal{C}_{k'}|} D_W(\mathcal{C}_j, \mathcal{C}_i) - \frac{|\mathcal{C}_j|}{|\mathcal{C}_j| + |\mathcal{C}_k| + |\mathcal{C}_{k'}|} D_W(\mathcal{C}_k, \mathcal{C}_{k'})$$

- On recommence l'étape d'agrégation jusqu'à obtenir une partition en une seule classe

Partie 4

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
- On peut couper le dendrogramme à une hauteur considérée comme suffisamment réduite (= faible inertie intra-classe pour la mesure de Ward)
- On peut aussi faire ce choix en utilisant les indices tels que R^2, CH, Silhouette ...

Rappels

- 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 $K \mapsto \mathrm{RSQ}(K)$ forme un coude.
  - Semi-Partial R-Square :

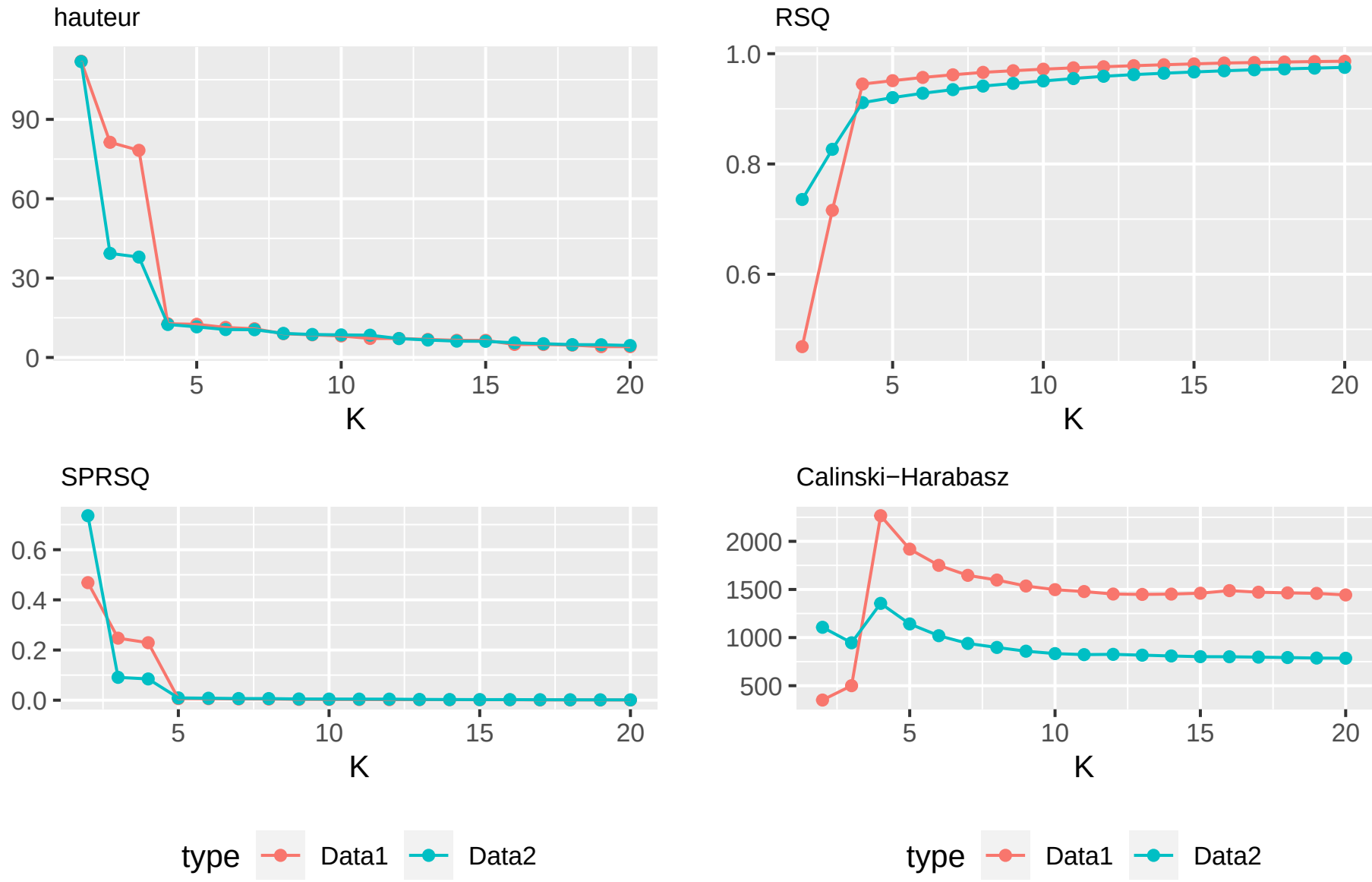
$$K \mapsto SPRSQ(K) = \frac{I_{inter}(\mathcal{P}_K) - I_{inter}(\mathcal{P}_{K-1})}{I_{totale}}$$

    On retient l'endroit où on a la plus forte réduction du SPRSQ.
  - CH (Calinski-Harabasz) :

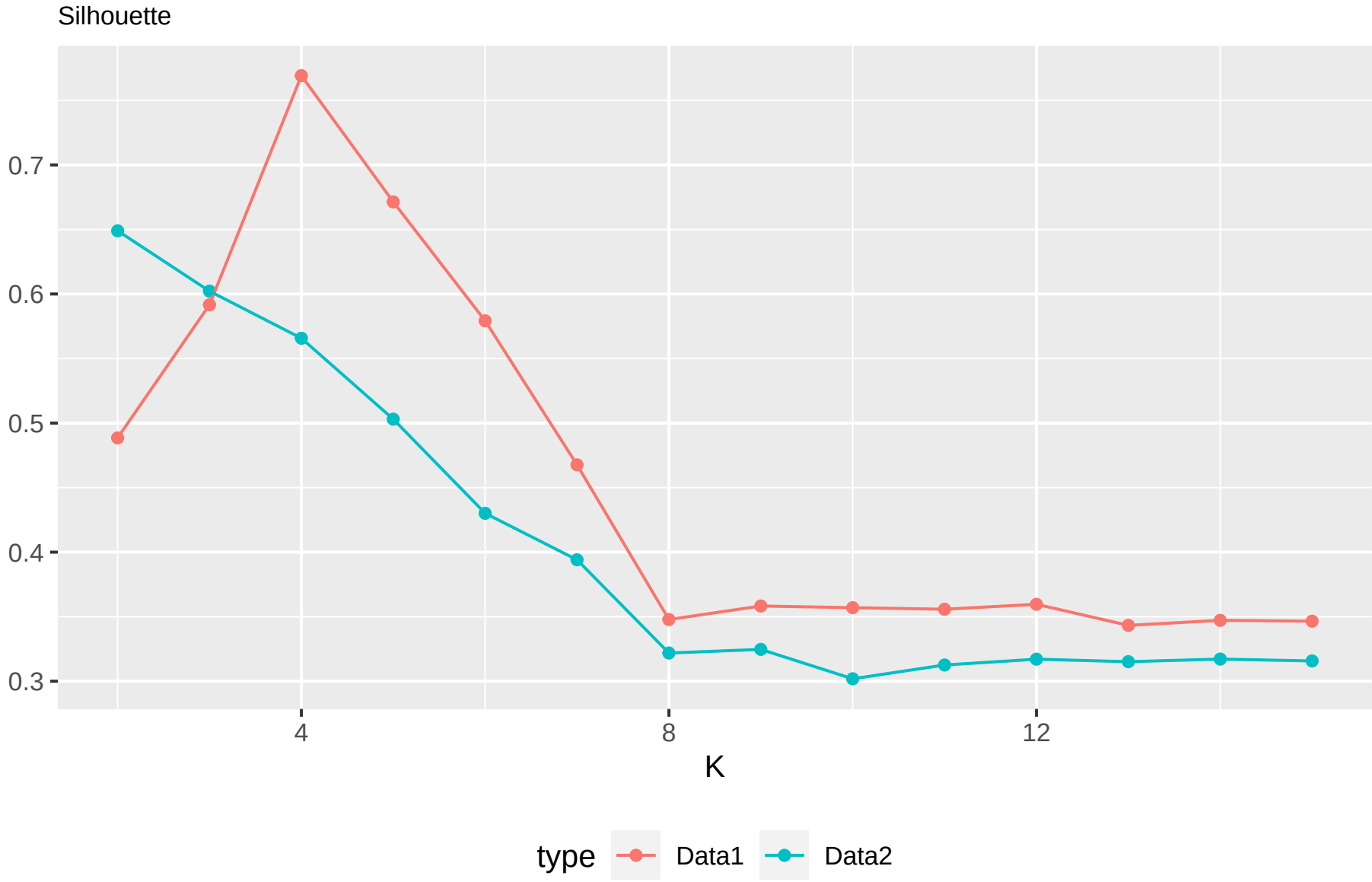
$$K \mapsto CH(K) = \frac{I_{inter}(\mathcal{P}_K)/(K-1)}{I_{intra}(\mathcal{P}_K)/(n-K)}$$

    On cherche un pic sur cette courbe
- Critère Silhouette

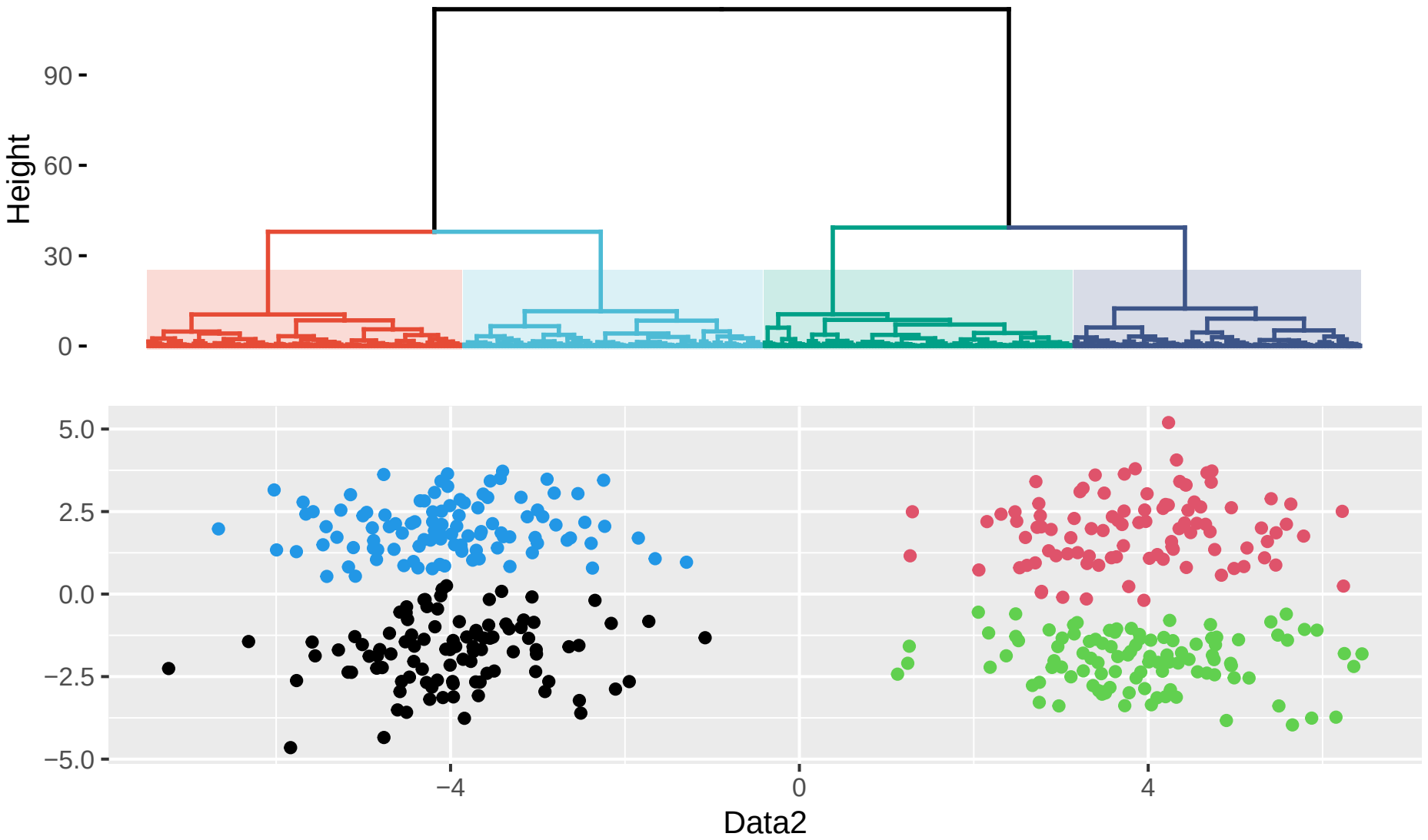
Exemple des données simulées



Exemple des données simulées



Exemple des données simulées



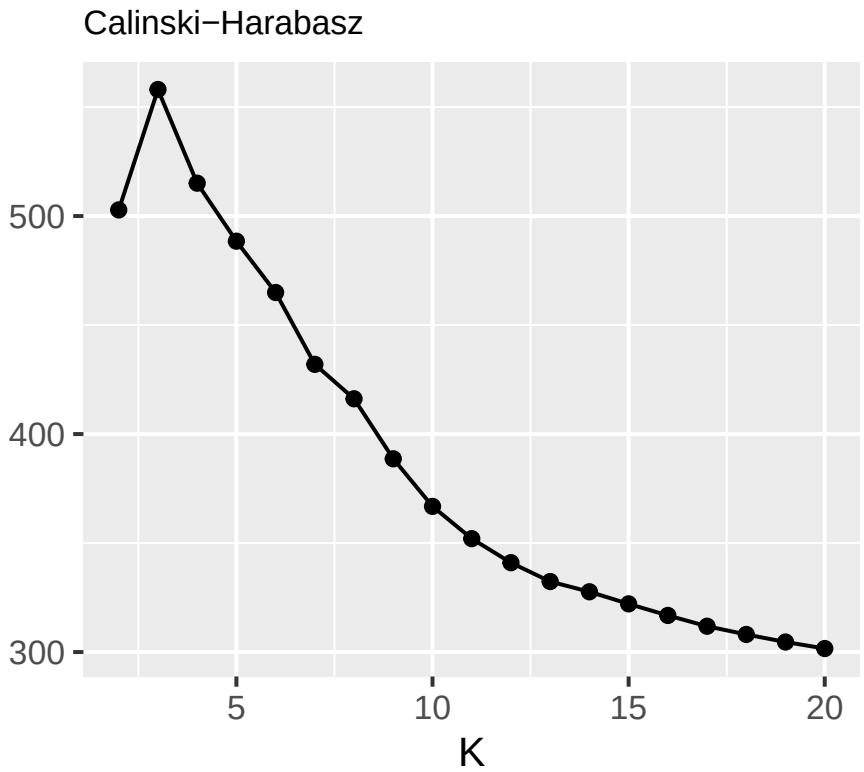
Partie 4

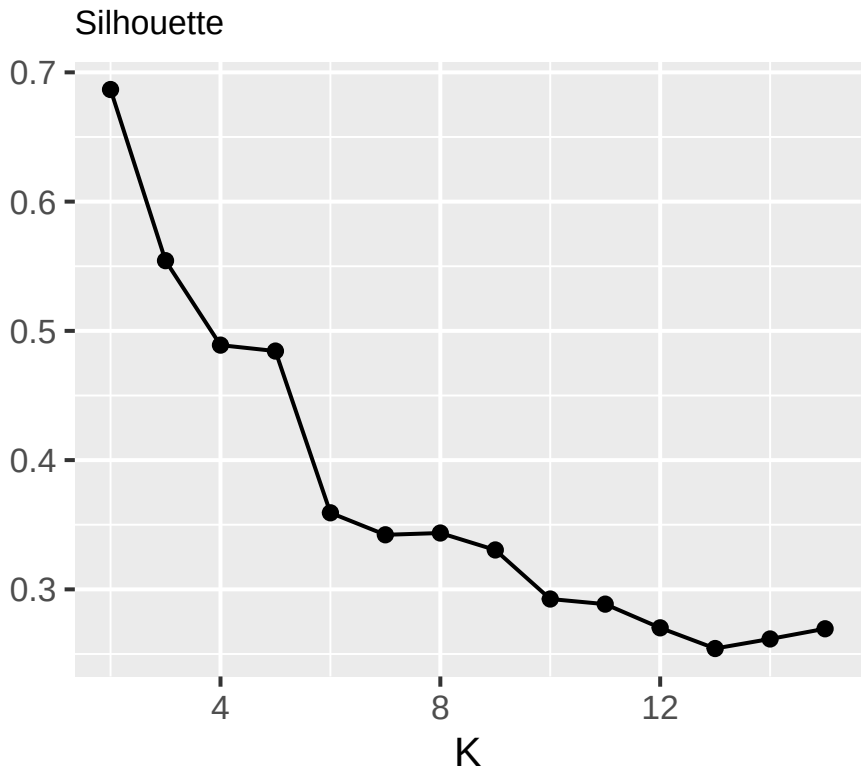
Commandes R

- `hc=hclust(d,method=)`
  - *d* : tableau de distances (produit par exemple par `dist()` ou `daisy()`)
  - *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.

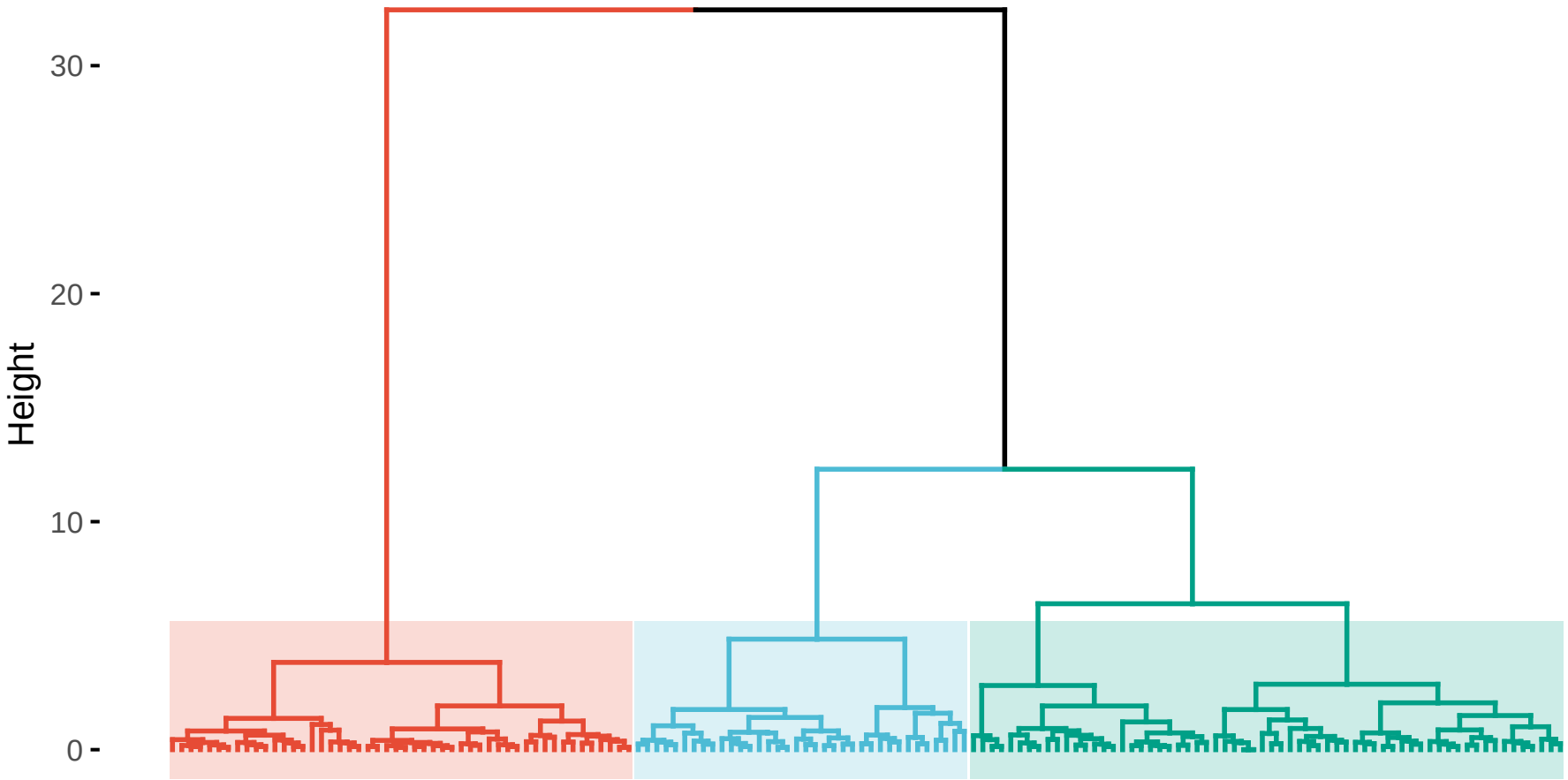
Exemple des iris

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





Exemple des iris



Avantages et inconvénients CAH

- Avantages :
  - Méthode flexible pour le niveau de finesse de la classification (pas nécessaire de fixer le nombre de classes à l'avance)
  - Prise en compte facile de distances et d'indices de similarité de n'importe quel type
- Limites :
  - Choix de la coupure de l'arbre
  - Classes très différentes selon la mesure d'agrégation choisie
  - Coûteux en calcul et mémoire pour de grands jeux de données.