Classifications nos supervisées

UFR Environnement - Departement PCMI

PLAN

- CHI- Généralités
- CHII - Analyse en Composante Principale (ACP)
- CHIII - Classification Ascendante Hiérarchique (CAH)
- CHIV- K-means clustering

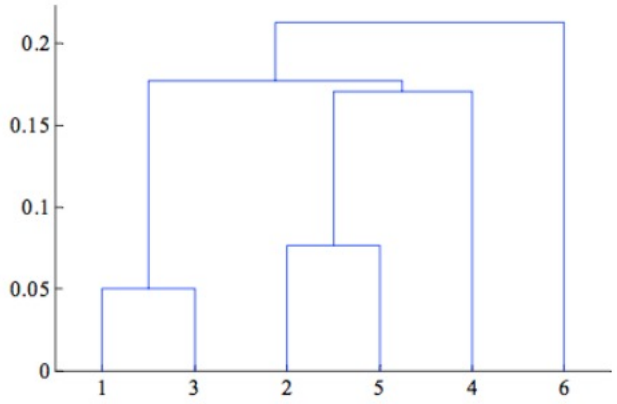
CHIII- Classification Ascendante Hiérarchique (CAH)

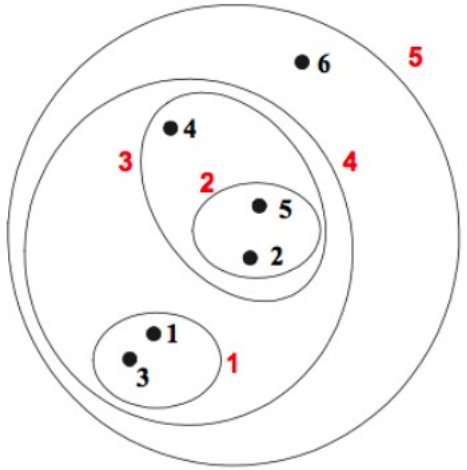
CHIII- Classicfication Assendante Hierarchique (CAH)

Introduction

Le clustering hiérarchique est, en fait, exactement ce qu'il semble être.

- Il produit un ensemble de grappes imbriquées organisées de manière hiérarchique.
- Le résultat peut être visualisé sous la forme d'un "dendrogramme", un diagramme en arbre qui enregistre les séquences de fusions ou de scissions :





CHIII- Classicfication Assendante Hierarchique (CAH)

I-1 Algorithme du Clustering

L'algorithme de base est simple : commencez par placer chaque point dans son propre groupe et calculez une matrice de proximité D(i,j) entre chaque paire de points i et j.

- Répéter ensuite :
    1. Fusionner les deux clusters les plus proches.
    2. Mettre à jour la matrice de proximité.
- Jusqu'à ce qu'il ne reste plus qu'une seule grappe.

L'opération clé est le calcul de la proximité de deux grappes. Les différents algorithmes se distinguent par des approches différentes pour définir la distance entre les grappes.

CHIII- Classicfication Assendante Hierarchique (CAH)

I-2 Type de Clustering hierarchique

Deux principaux types de regroupement hiérarchique

- Agglomératif :
  1. Commencer avec les points en tant que cluster individuelles.
  2. A chaque étape, fusionner la paire de grappes la plus proche jusqu'à ce qu'il ne reste plus qu'une seule grappe (ou k cluster).
- Partitionnement :
  1. Commencer par une seule grappe globale
  2. À chaque étape, diviser une grappe jusqu'à ce que chaque grappe contienne un point (ou qu'il y ait k clusters).

CHIII- Classicfication Assendante Hierarchique (CAH)

I-3 Mise en œuvre sur un exemple simple

Effectuons une CAH manuellement par liens complets et traçons le dendrogramme correspondant.

Étape 1 : Nous repérons dans la matrice de distance la paire qui a l'indice de dissimilarité le plus petit et effectuons un premier regroupement. A_E = groupe I à la distance 5,1. La matrice de distances est simplifiée par rapport à ce groupe I en considérant la règle utilisée (ici, liens complets, donc on garde la plus grande distance entre toutes les paires possibles lorsqu'il y a des groupes).

- Distance entre B et I = B_A = 15
- Distance entre C et I = C_E = 7,28
- Distance entre D et I = D_E = 7,81
- Distance entre F et I = F_A = 10,39

CHIII- Classicfication Assendante Hierarchique (CAH)

I-2 Mise en œuvre sur un exemple simple (suite)

Matrice de distance recalculée :

	I	B	C	D
B	15.00			
C	7.28	10.86		
D	7.81	13.04	5.48	
F	10.39	7.42	5.57	9.64

Étape 2 : Nous répétons le processus. C_D = groupe II à la distance la plus petite de 5,48.

•Distance entre I et II = I_D = 7,81

•Distance entre B et II = B_D = 13,04

•Distance entre F et II = F_D = 9,64

CHIII- Classification Assendante Hierarchique (CAH)

I-2 Mise en œuvre sur un exemple simple (suite)

Matrice de distance recalculée :

	I	B	II
B	15.00		
II	7.81	13.04	
F	10.39	7.42	9.64

Étape 3 : B_F = groupe III à la distance 7,42

- Distance entre I et III = I_B = 15
- Distance entre II et III = II_B = 13,04
- Matrice de distance recalculée :

	I	III
III	15.00	
II	7.81	13.04

CHIII- Classicfication Assendante Hierarchique (CAH)

I-2 Mise en œuvre sur un exemple simple (suite)

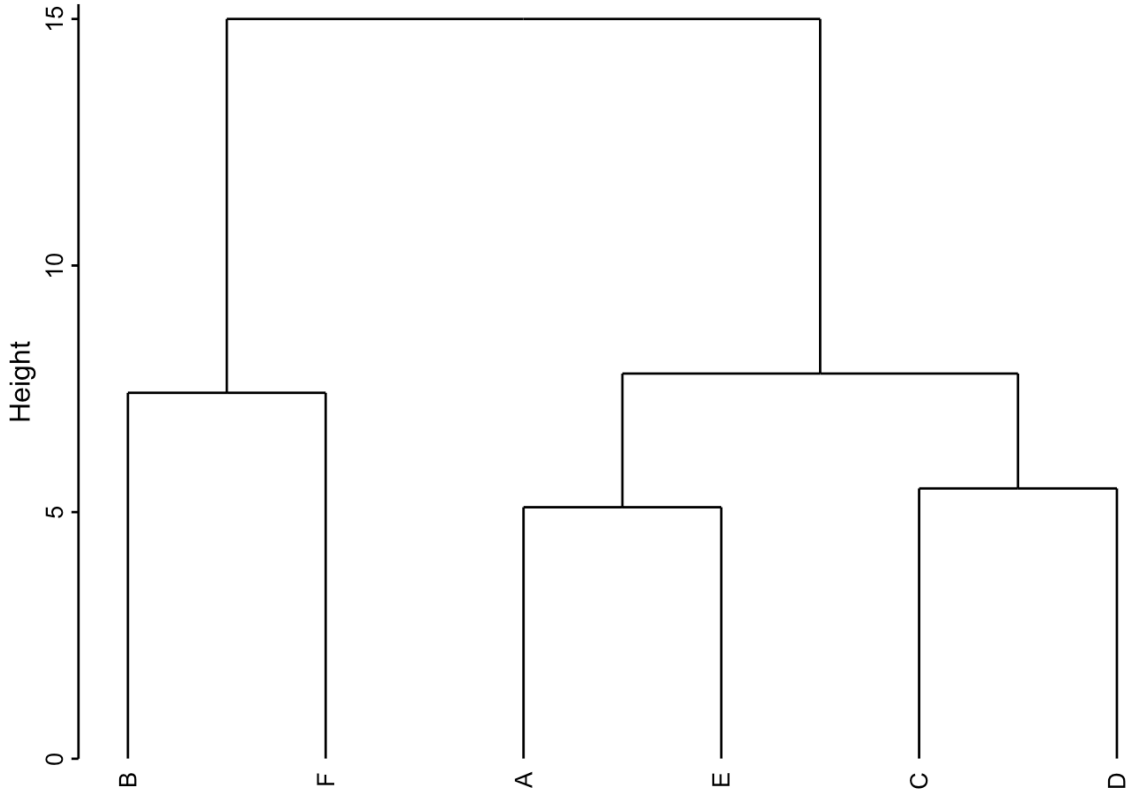
Étape 4 : I_II = groupe IV à la distance 7,81

•Distance entre III et IV = III_I = 15

Matrice de distance recalculée :

	III
IV	15

le dendrogramme résultant :

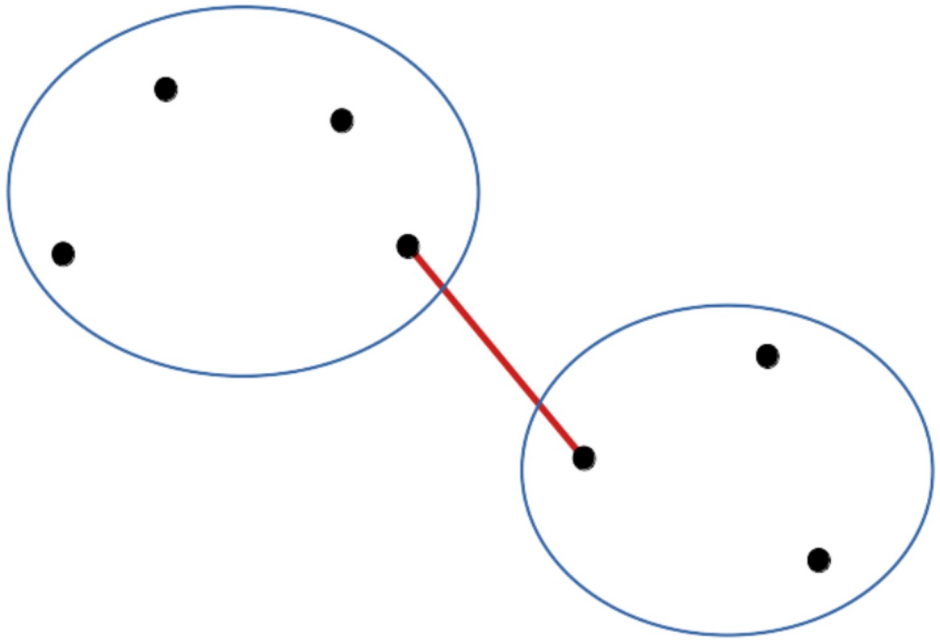


CHIII- Classicfication Assendante Hierarchique (CAH)

II- Fonctions de liaison

Tant que l'on compare des individus isolés entre eux, il n'y a pas d'ambiguïté. Par contre, dès que nous comparons un groupe avec un individu isolé, ou deux groupes entre eux, nous avons plusieurs stratégies possibles pour calculer leurs distances

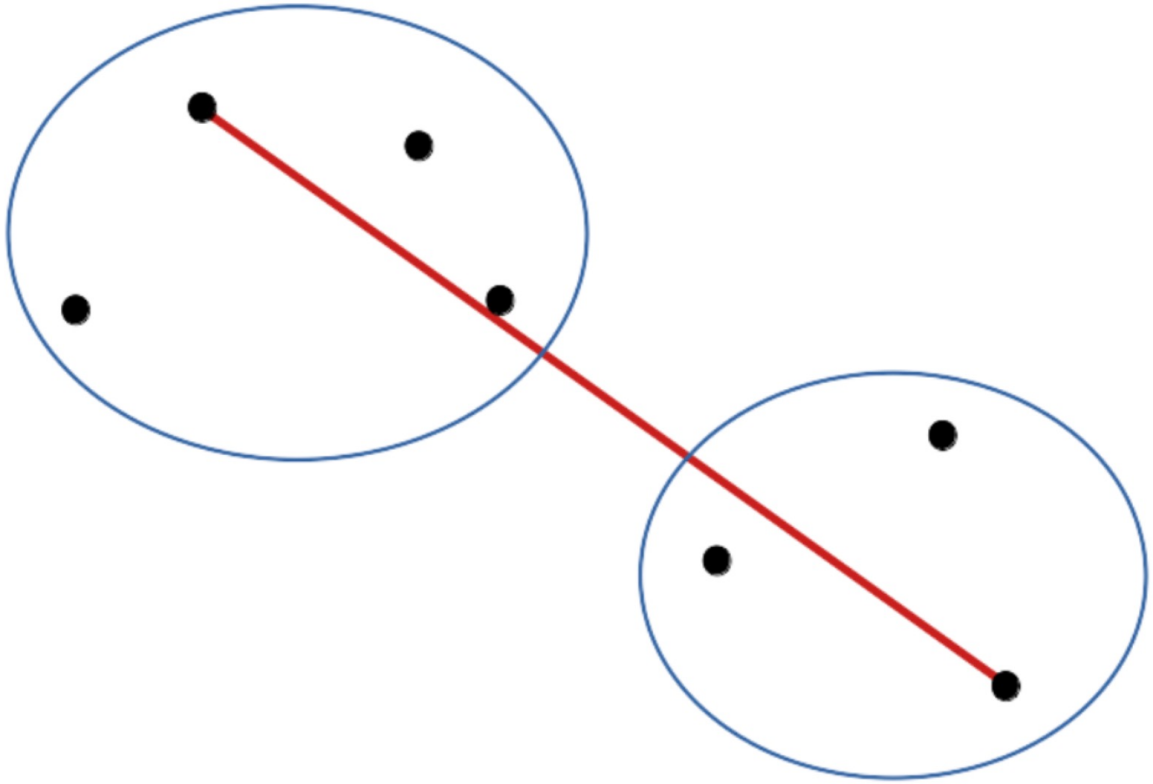
- **Liens simples** : la distance entre les plus proches voisins au sein des groupes est utilisée



CHII- Analyse en Composante Principale (ACP)

II- Fonctions de liaison (suite)

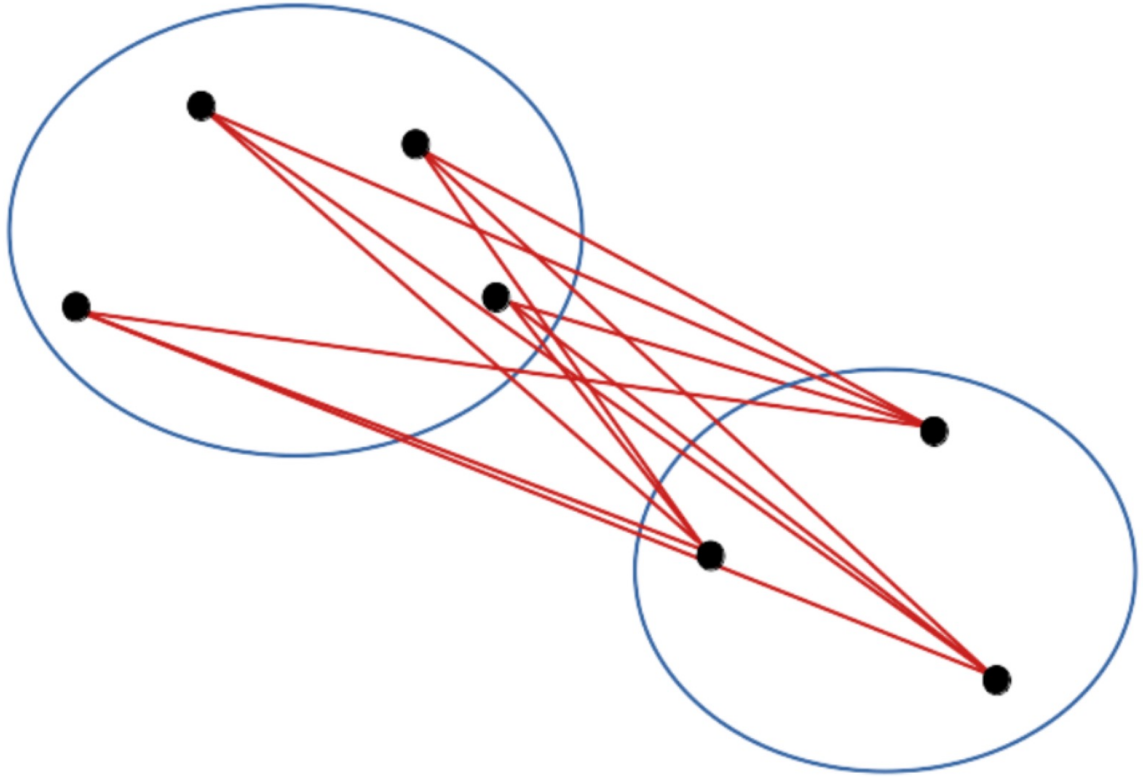
- **Liens complet** : méthode utilisée par défaut si non précisée : la distance entre les plus lointains voisins est considérée. C'est le premier dendrogramme que nous allons réaliser.



CHIII- Classicfication Assendante Hierarchique (CAH)

II- Typologie de méthodes agglomératives (suite)

- **Liens moyen** encore appelés méthode UPGMA : moyenne des liens entre toutes les paires possibles intergroupes.



CHIII- Classicfication Assendante Hierarchique (CAH)

II- Fonctions de liaison (suite)

- **Liens médians**, **centroïdes** moins utilisées. Elles ont l'inconvénient de produire parfois des inversions dans le dendrogramme, c'est-à-dire qu'un regroupement plus avant se fait parfois à une hauteur plus basse sur l'axe des ordonnées, ce qui rend le dendrogramme peu lisible et beaucoup moins esthétique. D.
- **Méthode de Ward D2** : Considérant le partitionnement de la variance totale du nuage de points (on parle aussi de l'inertie du nuage de points) entre variance interclasse et variance intraclasse, la méthode vise à maximiser la variance interclasse et minimiser la variance intraclasse, ce qui revient d'ailleurs au même. Cette technique fonctionne souvent très bien pour obtenir des groupes bien individualisés.

CHIII- Classicfication Assendante Hierarchique (CAH)

III- Fonctions de liaison (suite

- **Liens médians**, **centroïdes** moins utilisées. Elles ont l'inconvénient de produire parfois des inversions dans le dendrogramme, c'est-à-dire qu'un regroupement plus avant se fait parfois à une hauteur plus basse sur l'axe des ordonnées, ce qui rend le dendrogramme peu lisible et beaucoup moins esthétique. D.
- **Méthode de Ward D2** : Considérant le partitionnement de la variance totale du nuage de points (on parle aussi de l'inertie du nuage de points) entre variance interclasse et variance intraclasse, la méthode vise à maximiser la variance interclasse et minimiser la variance intraclasse, ce qui revient d'ailleurs au même. Cette technique fonctionne souvent très bien pour obtenir des groupes bien individualisés.

CHIII- Classification Assendante Hierarchique (CAH)

III- Fonctions de liaison (suite)

Chacune a ses avantages et ses inconvénients :

- Min : plus sensible au bruit et aux valeurs aberrantes, mais laisse intactes les grandes grappes évidentes.
- Max : plus robuste au bruit et aux valeurs aberrantes, mais tend à briser les grands groupes.
- Moyenne et centroïde : une sorte de compromis entre les deux.