

Matière :ADD

Niveau :Master1

CLASSIFICATION ASCENDANTE HIÉRARCHIQUE (CAH)

Cours 4

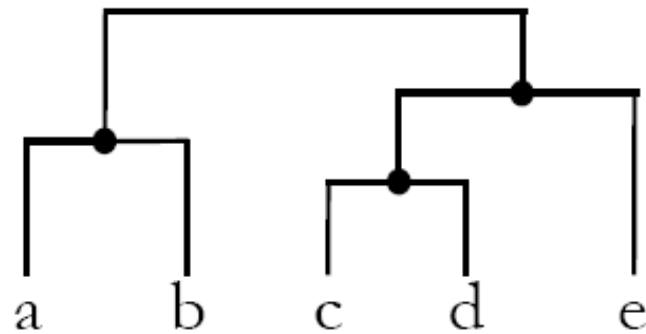


Introduction (CAH)

- Elle consiste à regrouper progressivement les individus dans un groupe
- il faut d'abord mettre les individus les plus proches ensemble
- Opérer des regroupements en classes homogènes d'un ensemble d'individus.
- L'état de rapprochement ou d'éloignement entre les individus est mesuré souvent par le biais de la distance euclidienne

Introduction (CAH)

- L'idée de l'algorithme de classification ascendante hiérarchique (CAH) est de créer, à chaque étape, une partition en regroupant les deux éléments les plus proches.
- . Le terme "élément" désigne aussi bien un individu qu'un groupe d'individus.



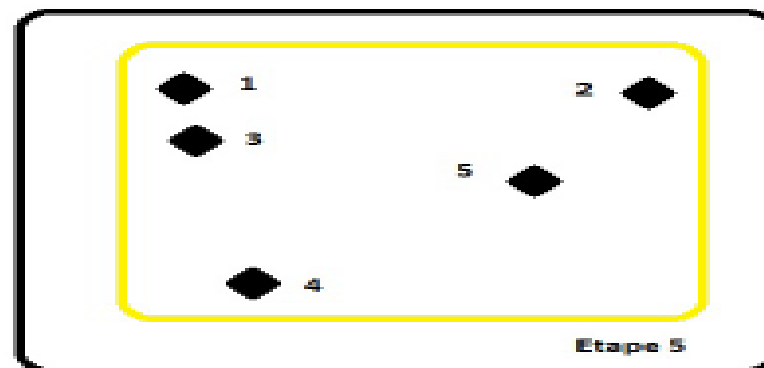
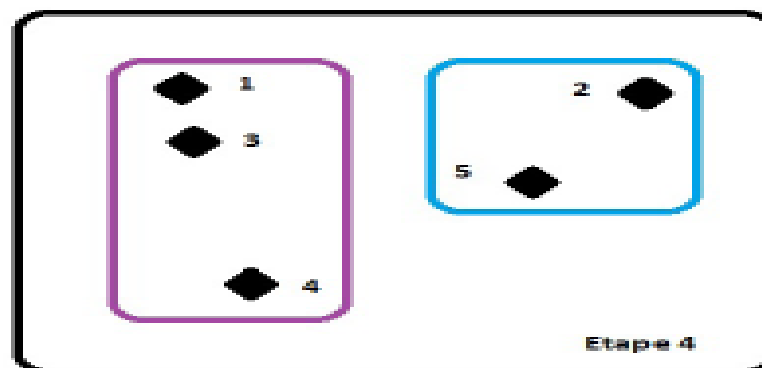
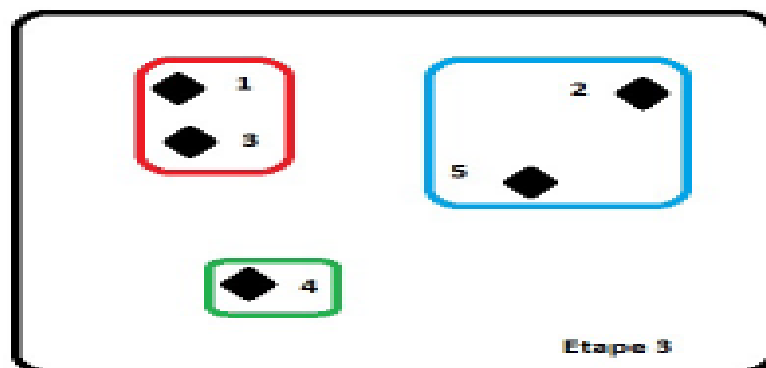
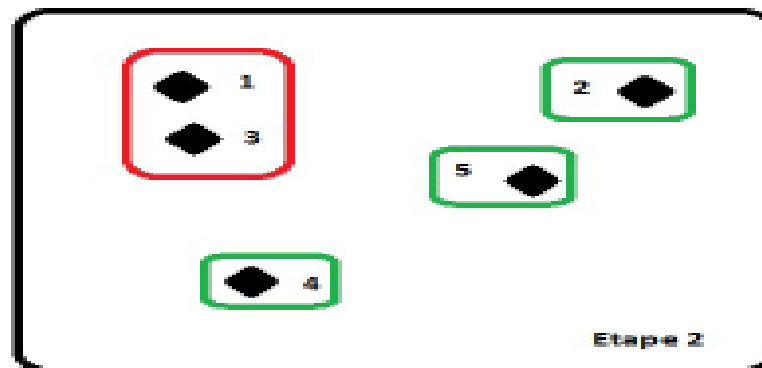
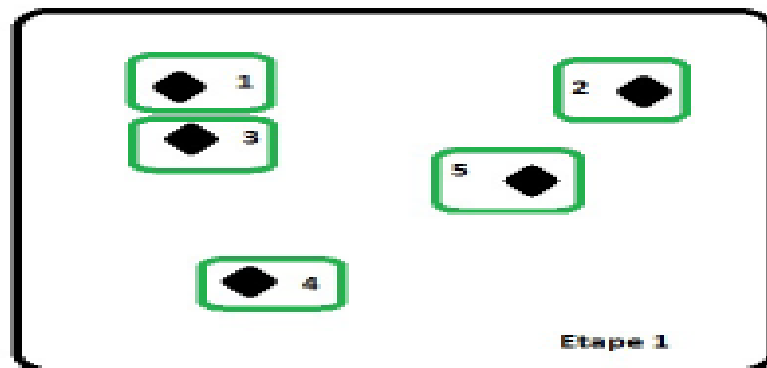
**Arbre de classification
ou dendrogramme**



Description de l'algorithme

- On choisit un écart. On construit le tableau des écarts pour la partition initiale des n individus
- On parcourt le tableau des écarts pour identifier le couple d'individus ayant l'écart le plus petit. Le regroupement de ces deux individus forme un groupe A . On a donc une partition de $n-1$ éléments : A et les $n-2$ individus restants.
- On calcule le tableau des écarts entre les $n-1$ éléments obtenus à l'étape précédente et on regroupe les deux éléments ayant l'écart le plus petit, On a donc une partition de $n-2$ éléments
- On itère la procédure précédente jusqu'à ce qu'il ne reste qu'un seul élément.

Exemple graphique



Distance

- On peut aborder le problème de la ressemblance entre individus par le biais de la notion de distance.

- **Exemple** : distance euclidienne :

On appelle distance euclidienne entre x et y la distance : CAH

6


$$d(x, y) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}.$$

Tableau des distances

- Soit d une distance. On appelle tableau des distances associées aux individus (w_1, \dots, w_n) le tableau :

$\mathbf{D} =$

	w_1	w_2	\dots	w_{n-1}	w_n
w_1	0	$d_{1,2}$	\dots	$d_{1,n-1}$	$d_{1,n}$
w_2	$d_{2,1}$	0	\dots	\dots	\dots
\dots	\dots	\dots	\dots	\dots	\dots
w_{n-1}	$d_{n-1,1}$	\dots	\dots	0	$d_{n-1,n}$
w_n	$d_{n,1}$	\dots	\dots	$d_{n,n-1}$	0



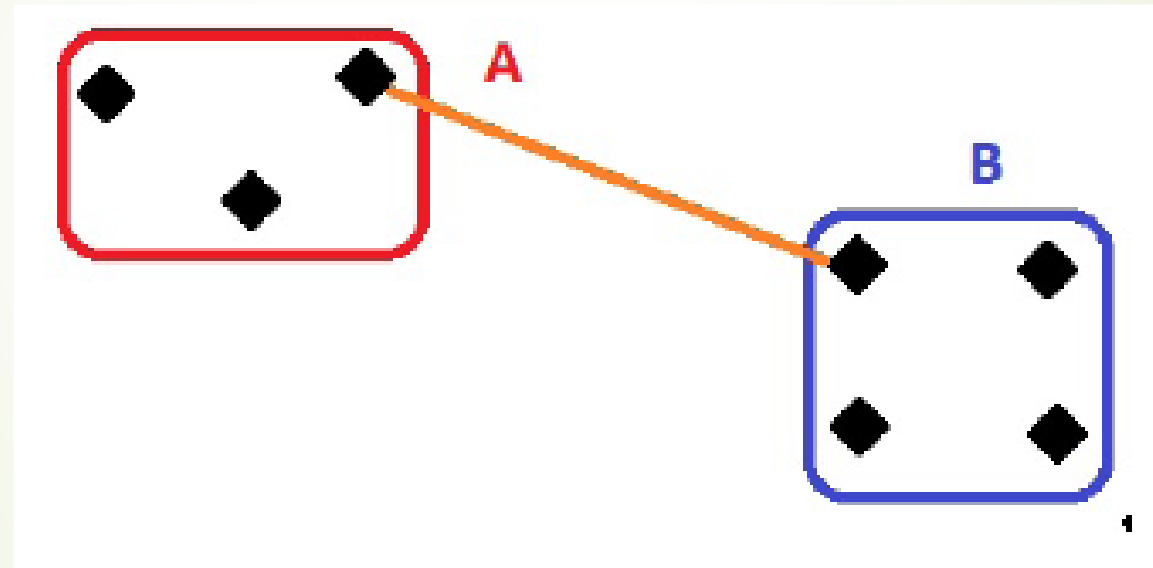
Écarts (ressemblance entre groupes d'individus)

- On appelle écart toute application définie à partir d'une distance et évaluant la ressemblance entre deux groupes d'individus.
- Plus l'écart entre deux éléments est petit, plus ils se ressemblent.

Écart simple / Méthode du plus proche voisin

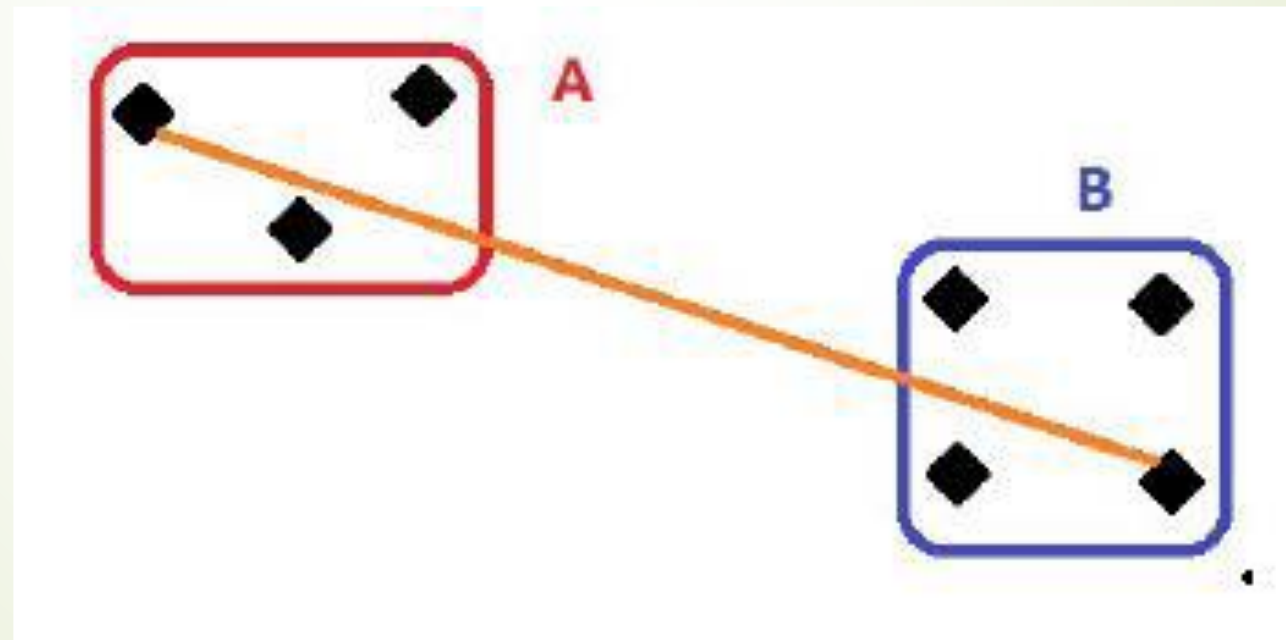
- L'écart entre deux groupes A et B est caractérisé par la distance la plus faible (petite) entre un point de A et un point de B:

CAH
9



Écart complet /Méthode du voisin le plus éloigné

- L'écart entre deux groupes A et B est caractérisé par la distance la plus forte (grande) entre un point de A et un point de B :

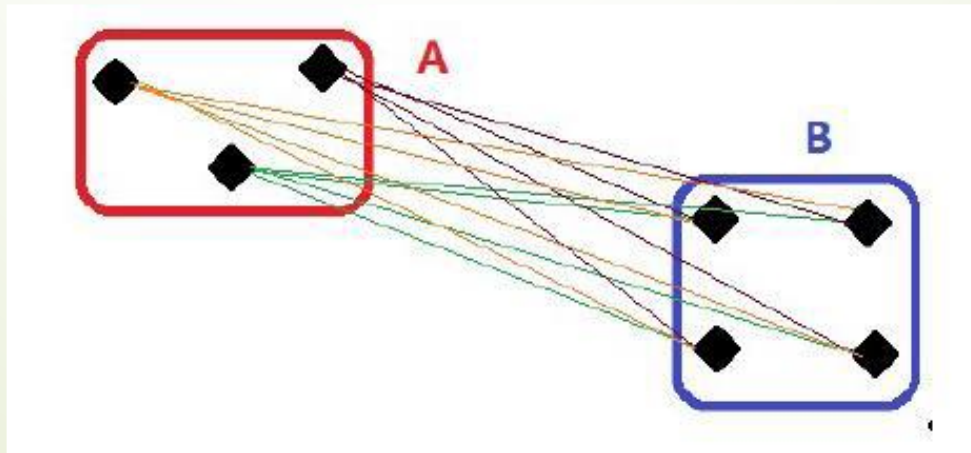


Écart moyen /Méthode de la distance moyenne

- L'écart entre deux groupes A et B est caractérisé par la distance moyenne entre les points de A et B :

$$e(A, B) = \frac{1}{n_A n_B} \sum_{\omega \in A} \sum_{\omega_* \in B} d(\omega, \omega_*),$$

où n_A est le nombre d'individus dans A, et n_B le nombre d'individus dans B.



Écart de Ward

- Soit d la distance euclidienne. La méthode de Ward considère l'écart :

$$e(A, B) = \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B),$$

où g_A est le centre de gravité de A, et g_B celui de B

- Cette méthode prend en compte à la fois la dispersion à l'intérieur d'un groupe et la dispersion entre les groupes. Elle est utilisée par défaut dans la plupart des programmes informatiques

Tableau des écarts

- Soit e un écart défini par une des méthodes précédentes. On appelle tableau des écarts associé aux groupes d'individus (A_1, \dots, A_n) le tableau :


$$\mathbf{E} =$$

	A_1	A_2	\dots	A_{n-1}	A_n
A_1	0	$e_{1,2}$	\dots	$e_{1,n-1}$	$e_{1,n}$
A_2	$e_{2,1}$	0	\dots	\dots	\dots
\dots	\dots	\dots	\dots	\dots	\dots
A_{n-1}	$e_{n-1,1}$	\dots	\dots	0	$e_{n-1,n}$
A_n	$e_{n,1}$	\dots	\dots	$e_{n,n-1}$	0



Dendrogramme

- Les partitions faites à chaque étape de l'algorithme de la CAH peuvent se visualiser via un arbre appelé dendrogramme.
- Sur un axe apparait les individus à regrouper et sur l'autre axe sont indiqués les écarts correspondants aux différents niveaux de regroupement.
- Cela se fait graphiquement par le biais de branches et de noeuds.
- Une partition naturelle se fait en coupant l'arbre au niveau du plus grand saut de noeuds.



Exemple (en présentiel)

Nous considérons ici 8 points :A,B,C,D,E,F,G,H

