

Analyse des données avancée

Chapitre 13 : La classification ascendante hiérarchique



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

L'algorithme

Le critère de Ward

L'arbre et sa coupure

Les limites de la méthode

Ce qu'il faut retenir

L'algorithme

Classification ascendante hiérarchique

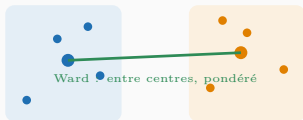
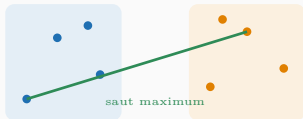
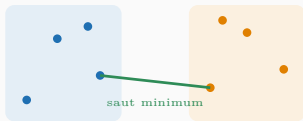
1. Au départ, **chaque individu est une classe** : n classes.
2. À chaque étape, **fusionner les deux classes les plus proches**.
3. S'arrêter quand il n'en reste qu'une.

Ce que l'algorithme produit

Non pas une partition, mais **une suite emboîtée de partitions** — de n classes à une seule.

C'est ce qui fait la force de la méthode : **on n'a pas à décider du nombre de classes avant de calculer**.
On regarde l'arbre, puis on coupe.

Les critères d'agrégation



Trois façons de mesurer la distance entre deux *groupes*, et trois arbres différents.

Saut minimum : le lien le plus court. Il produit des classes en chaîne, longues et étirées — un défaut connu sous le nom d'*effet de chaîne*.

Saut maximum : le lien le plus long. Il produit des classes compactes, mais coupe les groupes allongés.

Ward : on agrège les deux classes dont la fusion *coûte le moins* d'inertie interclasse.

$$\Delta = \frac{n_a n_b}{n_a + n_b} d^2(G_a, G_b)$$

C'est le seul critère cohérent avec l'ACP : même notion d'inertie, même géométrie. C'est pourquoi on enchaîne presque toujours ACP puis CAH de Ward.

Le critère de Ward

Pourquoi c'est presque toujours celui qu'on retient

Le coût d'une fusion

$$\Delta(a, b) = \frac{n_a n_b}{n_a + n_b} d^2(G_a, G_b)$$

On fusionne les deux classes dont la réunion **fait perdre le moins d'inertie interclasse**.

Deux termes, deux effets

Le facteur $d^2(G_a, G_b)$ dit que les classes doivent être **proches**.

Le facteur $\frac{n_a n_b}{n_a + n_b}$ dit qu'il est plus coûteux de fusionner deux **grosses** classes. **Ward tend donc à produire des classes de tailles comparables** — ce qui est en général souhaitable en segmentation : une classe de deux clients ne sert à rien.

Même monnaie, même géométrie

Ward raisonne en **inertie** : exactement la quantité que l'ACP a maximisée, et exactement celle que l'on décomposera pour juger la partition.

Enchaîner ACP puis CAH de Ward, ce n'est donc pas juxtaposer deux méthodes : **c'est poursuivre le même calcul**. C'est pourquoi cette séquence est devenue la pratique standard des études de marché.

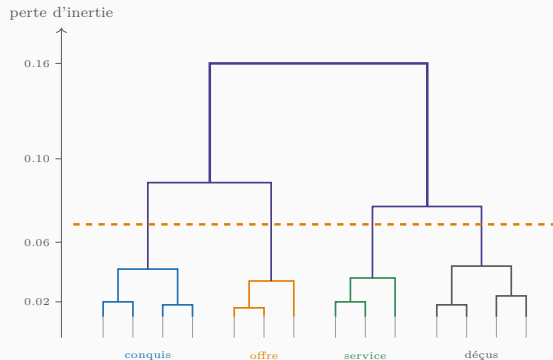
La décomposition de l'inertie d'une partition

$$I_{\text{totale}} = I_{\text{intraclasse}} + I_{\text{interclasse}}$$

Une bonne partition maximise l'**inter** — les classes bien séparées — et minimise l'**intra** — les classes homogènes. **Les deux objectifs n'en font qu'un.**

L'arbre et sa coupe

Le dendrogramme



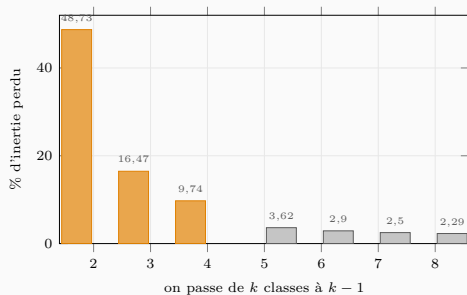
L'arbre se lit de bas en haut, la décision se prend de haut en bas.

En bas : 300 classes d'un individu. **En haut** : une seule classe.

La hauteur d'une branche est la perte d'inertie que coûte la fusion. Une longue branche signale que l'on vient de réunir deux groupes qui n'avaient rien à faire ensemble.

On coupe donc juste en dessous d'une longue branche — et la coupure donne le nombre de classes, qu'aucune formule n'impose.

Où couper, en chiffres



La règle : couper avant le grand saut.

Passer de 5 à 4 classes ne coûte que 3,62 % de l'inertie. Passer de 4 à 3 en coûte 9,74 % — presque le triple.

On s'arrête donc à quatre classes. Les fusions suivantes détruisent une information que l'on tient à garder.

Et l'on vérifie toujours que le nombre retenu s'interprète. Une partition statistiquement bonne mais que l'on ne sait pas nommer ne sert à rien — surtout en marketing, où la segmentation doit être communicable.

Quatre classes, sur nos trois cents clients

Classe	Effectif	F_1 moyen	F_2 moyen
1	61	+2,50	-0,45
2	82	+1,06	+1,31
3	59	+0,06	-1,61
4	98	-2,49	+0,16

La qualité de la partition

L'inertie interclasse représente **74,93 %** de l'inertie du nuage projeté.

Trois quarts de la dispersion des clients s'expliquent par leur appartenance à l'une des quatre classes.

Le quart restant est la variabilité à l'intérieur des classes — et il n'y a aucune raison de vouloir l'annuler : des clients d'un même segment ne sont pas identiques.

Les limites de la méthode

Ce que la CAH ne sait pas faire

1. **Une fusion est définitive.** Deux individus réunis à l'étape 3 le restent jusqu'au bout, même si l'étape 200 montre qu'ils n'auraient pas dû l'être.
2. **Le calcul est lourd** : il faut la matrice des $n(n - 1)/2$ distances. À 100 000 clients, c'est impraticable.

Le premier défaut a un remède

Il s'appelle la **consolidation** : reprendre la partition obtenue et laisser un algorithme la corriger, individu par individu.

C'est exactement ce que fait la seconde famille d'algorithmes. **Les deux méthodes sont complémentaires, non concurrentes** — et le chapitre 14 le montre sur nos données.

Un arbre existe toujours

L'algorithme produira un dendrogramme sur n'importe quelles données, **y compris sur trois cents points tirés au hasard.**

Il n'existe aucun test pour savoir si les classes sont réelles. Une CAH ne découvre pas des groupes : elle en construit.

Comment se prémunir

Trois vérifications, dans cet ordre. **Le saut d'inertie est-il net?** Un arbre sans coupure évidente signale un nuage sans structure. **Les classes s'interprètent-elles? Se retrouve-t-on avec une autre méthode?**

Sur nos données, le saut passe de 3,62 à 9,74 %, les quatre classes se nomment sans peine, et le chapitre 14 confirmera la partition. **Trois signaux concordants.**

Ce qu'il faut retenir

Six points

1. La CAH part de n classes et fusionne à chaque étape les deux plus proches.
2. Elle produit **une hiérarchie**, non une partition : on choisit après coup.
3. **Ward** minimise la perte d'inertie interclasse — même monnaie que l'ACP.
4. On coupe **avant le grand saut** : ici de 3,62 à 9,74 %, donc quatre classes.
5. Notre partition retient 74,93 % de l'inertie en interclasse.
6. Deux limites : les fusions sont **définitives**, et le calcul est **lourd** sur gros fichiers.

La suite

Une seconde famille d'algorithmes corrige les deux défauts — au prix d'une exigence nouvelle : il faut lui dire combien de classes on veut.

Puisque la CAH vient de répondre à cette question, les deux méthodes s'enchaînent naturellement.