TP6 : Classification Ascendante Hiérarchique

A. L. N'Guessan, V. Roca & W. Heyse

Le 02 Octobre 2025

Contents

1 Rappels

La **Classification Ascendante Hiérarchique** (CAH) est une méthode exploratoire utilisée pour regrouper des individus ou des objets similaires en classes homogènes, tout en permettant une représentation hiérarchique des regroupements sous forme d'un **dendrogramme**.
L'objectif principal est de simplifier et de structurer des données complexes en groupes cohérents, tout en conservant une trace de l'ordre dans lequel les regroupements ont été effectués. Contrairement aux méthodes de partitionnement (comme k-means), la CAH ne nécessite pas de fixer a priori le nombre de classes.

2 Fonctionnement Mathématique de la CAH

2.1 Matrice de Distance

Le point de départ de la CAH est la construction d'une matrice de distance (ou de similarité), qui quantifie les écarts entre les individus.

- **Distance Euclidienne** (classique pour des variables quantitatives) :
  Pour deux individus i et j, la distance est donnée par :

$$d(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

  où x_{ik} et x_{jk} représentent les valeurs des p variables pour les individus i et j.

Il existe aussi d'autres mesures, comme la distance de Manhattan, le coefficient de corrélation ou des mesures adaptées aux variables qualitatives (indice de Jaccard, distance du khi2, etc.).

2.2 Critères d'aggrégation

Une fois les distances calculées, les individus les plus proches sont regroupés en clusters selon un critère de liaison. Les critères les plus courants incluent :

- **Méthode de la liaison simple** :
  La distance entre deux clusters est définie par la plus petite distance entre leurs éléments :

$$d(C_1, C_2) = \min_{i \in C_1, j \in C_2} d(i, j)$$

- **Méthode de la liaison complète** :
  La distance entre deux clusters est définie par la plus grande distance entre leurs éléments :

$$d(C_1, C_2) = \max_{i \in C_1, j \in C_2} d(i, j)$$

- **Méthode de la moyenne de groupe** :
  La distance est la moyenne des distances entre tous les éléments des deux clusters :

$$d(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{i \in C_1} \sum_{j \in C_2} d(i, j)$$

- **Méthode de Ward** (optimisation de l'inertie) :
  La fusion minimise la perte d'inertie inter-classes, en cherchant à minimiser l'augmentation de l'inertie totale :

$$\Delta I = \frac{|C_1| \cdot |C_2|}{|C_1| + |C_2|} \cdot \|\mathbf{g}_{C_1} - \mathbf{g}_{C_2}\|^2$$

  où $\mathbf{g}_{C_1}$ et $\mathbf{g}_{C_2}$ sont les barycentres des clusters C_1 et C_2.

2.3 Construction de l'Arbre Hiérarchique

L'arbre hiérarchique (ou dendrogramme) est construit en répétant les étapes suivantes jusqu'à ce que tous les individus soient regroupés dans un seul cluster :
1. Identifier les deux clusters les plus proches (selon le critère de liaison choisi).
2. Fusionner ces clusters.
3. Mettre à jour la matrice de distance.

Le dendrogramme est une représentation visuelle qui montre les regroupements successifs. Les branches illustrent les étapes de fusion, et leur hauteur indique les distances entre les clusters.

2.4 Élagage de l'Arbre

L'élagage consiste à choisir un niveau de coupure dans le dendrogramme pour définir le nombre final de clusters. Ce choix peut être guidé par :
- L'observation des **ruptures dans les hauteurs des branches** du dendrogramme. (Criètre du coude) - Des indices statistiques comme :
- **Indice de silhouette** : Mesure la cohésion interne et la séparation externe des clusters.
- **Critère de Calinski-Harabasz** : Évalue le rapport entre la variance inter-clusters et intra-clusters.

2.5 Qualité des Partitions

La qualité d'une partition obtenue peut être évaluée par :
- **Inertie inter-classes** : Part de la variance expliquée par les clusters. Une inertie inter-classes élevée indique des clusters bien séparés.

$$I_{\text{inter}} = \sum_{k=1}^{K} |C_k| \cdot \|\mathbf{g}_{C_k} - \mathbf{g}_T\|^2$$

où $\mathbf{g}_T$ est le barycentre total des données.
- **Indice de Dunn** : Rapport entre la distance minimale inter-cluster et la taille maximale intra-cluster.

2.6 Interprétation des Résultats

2.7 Lecture du Dendrogramme

Les **hauteurs des branches** indiquent les distances (ou dissimilarités) entre les clusters. Une branche plus haute signifie que les clusters fusionnés sont plus dissemblables.
L'élagage du dendrogramme fournit une partition des individus. Le nombre de clusters peut être déterminé en coupant l'arbre à une hauteur appropriée.

2.8 Interprétation des Clusters

Chaque cluster regroupe des individus similaires selon les variables initiales. Les caractéristiques communes des membres du cluster peuvent être identifiées en étudiant les moyennes ou distributions des variables dans chaque cluster.

3 Exercice 1

On propose d'analyser les caractéristiques de certains dignes représantants de la culture fromagère Française disponible dans le fichier `fromage.csv`.

1. Chargez le jeu de données `fromage.csv` et proposez une analyse descriptive des données.
2. Réslisez une ACP des données et représentez le premier plan factoriel.
3. Semble-t-il y avoir des groupes d'individus distincts des autres ?
4. A l'aide de la fonction `dist`, construire une matrice de distance euclidienne entre les individus à partir des données.
5. Utilisez la fonction `hclust` afin de réaliser une CAH, vous utiliserez le critère d'aggrégation de Ward.
6. Représentez le dendrogramme associé à l'arbre.
7. Utilisez le fonction `cutree` afin d'élaguer l'arbre et d'obtenir une partition des donnez. Faites varier le nombre de classe entre 1 et 10 et déterminez pour chaque partition l'intertie inter-classe. Représentez l'inertie inter-classe en fonction du nombre de groupe et déterminez le nombre de groupe à retenir.
8. Reprenez la représentation dans le premier plan factoriel de l'ACP et colorez les individus selon leur groupe.
9. Interprétez les groupes trouvés.
10. Utilisez à présent le critère d'aggrégation du lien simple et recommencez l'analyse. Quelles différences observez-vous ?

4 Exercice 2

Reprenez les données du TP5.

1. Reproduisez l'ACM du TP5 avec la fonction `MCA`.
2. Utilisez la fonction `HPCP` du package `FactoMineR` afin de réaliser une CAH sur les coordonées des individus projetés. Choisissez le nombre de groupes en utlisant l'inertie.
3. Représentez les groupes à l'aide de la fonction `fviz_cluster`.
4. En vous aidant de votre interprétation des axes au TP précédant et des valeurs données dans `desc.var`, `desc.axes` et `desc.ind`. Interprétez les clusters.