Chapitre 9
Classification non supervisée

Lise BELLANGER

Lab-STICC UMR 6285, CNRS, Université Bretagne Sud, Vannes

Ce chapitre s'inspire de plusieurs chapitres de l'ouvrage de Bellanger & Tomassone (2014). Les lecteurs intéressés trouveront aussi de bons compléments dans les ouvrages suivants : Lebart *et al.*, (2000), Nakache & Confais (2005), Saporta (2011).

La constitution de classes* peut être l'objectif principal de l'analyse d'un corpus de données. La **classification non supervisée** consiste à regrouper les observations* d'une population* en un nombre limité de classes (encore appelé groupes) qui :

- ne sont pas prédéfinies mais déterminées au cours de l'opération ;
- regroupent les individus ayant des caractéristiques similaires et séparent les observations ayant des caractéristiques différentes.

Le terme classification non supervisée sous-entend le fait que le *tableau de données n'est pas structuré*. Dans la mesure où les techniques employées pour constituer des classes sont très souvent algorithmiques, on trouve aussi le terme de **classification automatique** ou clustering en anglais. Le but d'une classification est de construire une partition ou un système de classes, emboîtées ou pas. Cette construction doit être bâtie sur l'ensemble des critères jugés pertinents par l'utilisateur. On peut obtenir les classes d'observations soit à partir de l'ensemble des variables* qui les décrivent, on obtient une **classification polythétique**, soit en prenant les variables une par une pour séparer les observations ; on obtient alors une **classification monothétique**. Nous nous concentrerons dans cette partie sur les méthodes de classification polythétique adaptées à la classification d'observations. Après avoir donné quelques idées générales communes à l'ensemble des techniques de classification (§1), nous décrirons les deux principales catégories de méthodes : (i) par partitionnement (§2) et (ii) hiérarchiques (§3). Enfin, nous ferons un rapide bilan et évoquerons des méthodes plus récentes.

1 Idées générales

1.1 Concepts courants en classification

Quelles que soient les techniques utilisées pour regrouper des observations en classes, on souhaite toujours que celles qui sont dans une même classe soient plus voisines que celles qui sont dans des classes séparées. Une technique de classification repose donc sur deux idées complémentaires : cohésion interne des classes et isolement entre classes. A ces deux idées s'ajoute celle de hiérarchie possible entre classes. Enfin, classifier implique qu'il existe une discontinuité ; s'il n'y en a pas, tout découpage risque d'être, au moins en partie, artificiel. C'est la raison pour laquelle certaines techniques peuvent permettre un certain recouvrement des classes. Ces concepts sont schématisés sur la Figure 1.

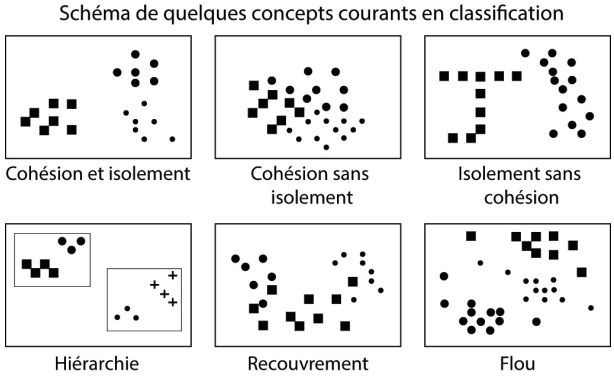


FIG. 1 : Concepts courants en classification. Figure tirée de Bellanger & Tomassone (2014)

Une « bonne » classification produira des classes avec une petite dissimilarité à l'intérieur de chaque classe et une grande dissimilarité entre les classes. Sa qualité dépend donc de la mesure de **dissimilarité** et de la méthode de classification choisie. Cette mesure permettra de quantifier la cohésion interne et l'isolement entre classes. Pour classifier des individus, la première étape consiste donc à choisir une mesure de dissimilarité entre individus, le plus souvent la distance euclidienne* ; mais il existe d'autres mesures de dissimilarité, plus adaptées à certains types de données spécifiques !

1.2 Mesures de dissimilarité

Nous venons d'employer le terme **dissimilarité** dans son acception courante. Il faut maintenant lui donner une définition plus précise. On définit d'abord une distance d sur un ensemble $\boldsymbol{E} = \{\boldsymbol{i}, i = 1, ..., n\}$ de n objets comme une *application* de $\boldsymbol{E} \times \boldsymbol{E}$ dans $\mathbb{R}^+$ telle que :

$$\begin{aligned} &d(\boldsymbol{i}, \boldsymbol{j}) \geq 0 \text{ et } d(\boldsymbol{i}, \boldsymbol{j}) = 0 \Leftrightarrow \boldsymbol{i} = \boldsymbol{j} \\ &d(\boldsymbol{i}, \boldsymbol{j}) = d(\boldsymbol{j}, \boldsymbol{i}) \\ &d(\boldsymbol{i}, \boldsymbol{j}) \leq d(\boldsymbol{i}, \boldsymbol{k}) + d(\boldsymbol{j}, \boldsymbol{k}) \text{ (inégalité triangulaire)} \end{aligned}$$

La mesure ainsi définie n'est une distance euclidienne que si elle est engendrée par un **produit scalaire** [1]. Il existe bien d'autres types de distance pour les variables continues* : la distance de Manhattan* par exemple est la somme des valeurs absolues des différences entre les coordonnées x et y [2] ; elle est beaucoup moins sensible aux grandes différences, elle peut donc être particulièrement intéressante dans certains cas.

Si l'inégalité triangulaire* n'est pas vérifiée, on a une dissimilarité ou **dissemblance**. On parle de **similarité** ou **ressemblance** quand on a une application s telle que :

$$\begin{aligned} &s(\boldsymbol{i}, \boldsymbol{j}) \geq 0 \\ &s(\boldsymbol{i}, \boldsymbol{j}) = s(\boldsymbol{j}, \boldsymbol{i}) \\ &s(\boldsymbol{i}, \boldsymbol{i}) = s(\boldsymbol{j}, \boldsymbol{j}) \geq s(\boldsymbol{i}, \boldsymbol{j}) \ \forall \boldsymbol{i}, \boldsymbol{j} \end{aligned}$$

[1] On appelle *produit scalaire* une opération qui permet de mesurer à la fois la longueur des vecteurs et l'angle qu'ils forment entre eux. Autrement dit, il fournit un nombre qui indique dans quelle mesure deux vecteurs "pointent dans la même direction". Ce nombre permet de calculer la longueur d'un vecteur et la distance entre deux points. Pour deux vecteurs u et v, le produit scalaire s'écrit

$$\langle u, v \rangle = \|u\| \, \|v\| \cos(\theta),$$

où $\|u\|$ est la longueur de u, $\|v\|$ celle de v, et θ l'angle entre eux. La distance euclidienne entre deux points x et y peut alors s'exprimer comme

$$d(x, y) = \|x - y\| = \sqrt{\langle x - y, \, x - y \rangle}.$$

Une distance est dite *euclidienne* lorsqu'elle peut être calculée de cette manière.

[2] La distance dite de *Manhattan* tire son nom du plan orthogonal des rues de Manhattan et correspond à la distance parcourue le long de chemins rectilinéaires, comme si l'on se déplaçait uniquement sur des rues perpendiculaires (Everitt et al., 2001). Autrement dit, pour aller d'un point à un autre, on additionne les déplacements horizontaux et verticaux. Pour deux points (x_i, y_i) et (x_j, y_j) sur un plan, la distance de Manhattan s'écrit

$$D_{ij} = |x_i - x_j| + |y_i - y_j|.$$

Plus généralement, dans un espace à n dimensions, pour deux points $X = (x_1, \ldots, x_n)$ et $Y = (y_1, \ldots, y_n)$, elle correspond à la somme de toutes les différences absolues entre les coordonnées :

$$D(X, Y) = \sum_{k=1}^{n} |x_k - y_k|.$$

Quand les variables sont toutes quantitatives, on utilise généralement la distance euclidienne calculée sur les variables centrées réduites (*cf.* 4 et *cf.* 7, section 3.3.2). C'est le cas quand, parallèlement à une ACP (Analyse en Composantes Principales*, *cf.* 6) sur variables centrées réduites, on veut utiliser des distances entre les observations. Par contre, avec une AFC (Analyse Factorielle des Correspondances*, *cf.* 6), il est naturel de prendre une distance utilisant la métrique du χ^2* (*cf.* 2.2.8). Quand les objets sont décrits par des variables qualitatives de nature binaire, il existe un grand nombre de mesures de similarité (voir par exemple Nakache & Confais (2005)).

Une fois fixé le choix d'une mesure de dissimilarité entre observations, il suffit de calculer cette mesure pour les $\frac{n(n-1)}{2}$ couples pour obtenir une matrice symétrique $\mathbf{D}_{n\times n}$. C'est donc à partir de cette dernière que s'effectuera la classification. Il faut remarquer que si la matrice $\mathbf{X}_{n\times p}$ existe elle permet de calculer $\mathbf{D}_{n\times n}$ qui peut très bien être la donnée de base de l'algorithme de classification.

1.3 Caractère combinatoire de la classification

Le nombre de partitions d'un ensemble à n objets, appelé n-ième **nombre de Bell**, devient très rapidement astronomique même avec l'utilisation d'un ordinateur. Il vaut :

$$B_n = \frac{1}{e}\sum_{i=1}^{\infty}\frac{i^n}{i!} = \sum_{k=1}^{n} B_{n,k}$$

où $B_{n,k}$ est le nombre de partitions en k classes de n objets aussi appelé nombre de Stirling.

Par exemple, pour $n = 13, B_{13} = 27644437$; ce qui signifie qu'il y a plus de 27 millions de partitions possibles pour un ensemble de 13 observations ! Il est donc matériellement impossible d'obtenir une réponse exacte à une question d'apparence aussi anodine que *"quelle est la meilleure classification de n objets en K classes ?"*. La réponse est de nature purement algorithmique ; ce qui signifie que, sauf dans de très rares cas, nous devons donc nous satisfaire de solutions approchées.

2 Classification par partitionnement : méthode k-means

2.1 Inertie inter-classe et inertie intra-classe

L'**inertie** [3] d'un nuage de n observations se définit comme la moyenne des carrés des distances de ces observations à leur centre de gravité ou à leur point moyen noté g. Si $d(\boldsymbol{i}, \boldsymbol{g})$ est la distance de l'observation $\boldsymbol{i}$ au centre de gravité $\boldsymbol{g}$), nous pouvons écrire (en supposant que chaque observation a le même poids $\frac{1}{n}$) que l'inertie totale vaut :

$$I_T = \frac{1}{n}\sum_{i=1}^{n} d^2(\boldsymbol{i}, \boldsymbol{g})) \text{ où } \boldsymbol{g} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \text{ et } \boldsymbol{i} = \boldsymbol{x}_i \; i = 1, .., n$$

Pour un ensemble d'observations données, l'inertie totale I_T est constante.

[3] L'inertie mesure la dispersion d'un nuage autour de son centre de gravité g.

Supposons que nous ayons K classes, la $k^{\text{ième}}$ classe C_k de centre de gravité $\boldsymbol{g}_k$ est formée de n_k $(k = 1, \cdots, K)$ observations, son inertie vaut :

$$I_k = \frac{1}{n_k} \sum_{\boldsymbol{i} \in C_k} d^2(\boldsymbol{i}, \boldsymbol{g}_k) \text{ où } \boldsymbol{g}_k = \frac{1}{n_k} \sum_{\boldsymbol{i} \in C_k} \boldsymbol{x}_i$$

Elle est d'autant plus homogène que ses éléments sont proches de son centre, son inertie est alors faible.

Une mesure globale de l'homogénéité des classes appelée **inertie intra-classe** (d'où le symbole W comme *within*), est donc la moyenne des inerties des K classes, chacune pondérée par son importance relative :

$$I_W = \sum_{k=1}^{K} \frac{n_k}{n} I_k = \frac{1}{n} \sum_{k=1}^{K} \sum_{\boldsymbol{i} \in C_k} d^2(\boldsymbol{i}, \boldsymbol{g}_k)$$

L'inertie du nuage des K centres de gravité, appelée **inertie inter-classe** (d'où le symbole B comme *between*), est :

$$I_B = \sum_{k=1}^{K} \frac{n_k}{n} d^2(\boldsymbol{g}_k, \boldsymbol{g})$$

Le théorème de Huygens [4] nous permet d'écrire :

$$I_T = I_B + I_W$$

Il est alors naturel de choisir comme critère de classification celui qui rend l'inertie intra-classe I_W minimale ou l'inertie inter-classe I_B maximale. On peut choisir par exemple de déterminer le nombre de classes K qui maximise le critère global de qualité de la classification :

$$C_K^2 = \frac{I_B}{I_T}$$

2.2 Regroupement d'observations autour de centres mobiles : méthode k-means

On commence par fixer le nombre K de classes, puis on passe par les étapes suivantes :

1. Choisir les K centres initiaux $\boldsymbol{g}_k^{(0)}$, par exemple par tirage pseudo-aléatoire, qui induisent une première partition $C_k^{(0)} (k = 1, \cdots, K)$; si l'observation $\boldsymbol{i}$ est plus proche de $\boldsymbol{g}_k^{(0)}$ que de tous les autres centres alors $\boldsymbol{i} \in C_k^{(0)}$.

[4] Le théorème de Huygens, issue de l'étude de l'énergie cynétique en physique, permet de calculer l'inertie totale du nuage de points (c'est-à-dire sa dispersion autour du centre de gravité g) en faisant la somme de l'inertie intra-classe et de l'inertie inter-classe. Plus l'inertie totale est grande et plus le nuage est dispersé par rapport à son centre de gravité et à l'inverse, plus elle tend vers 0 et plus les points sont rapprochés voir confondus (IG = 0). Une illustration du théorème est proposée dans le chap. 10.

2. Affecter chaque observation à la classe dont elle est le plus proche du centre. Il en résulte une **partition** en K classes.

3. Recalculer les centres $\boldsymbol{g}_k^{(1)}$ de chaque classe en effectuant une moyenne des observations affectées à cette classe à l'étape précédente : chaque classe est alors représentée par son centre qui, en général, n'est pas un point de l'échantillon. Les centres de gravité $\boldsymbol{g}_k^{(1)}$ induisent une nouvelle partition $C_k^{(1)}$.

4. Itérer les étapes 2 et 3 jusqu'à ce que le critère de classification reste stable, c'est-à-dire quand le contenu des classes n'est plus modifié.

Le processus de classification est simple, mais il souffre d'un défaut sérieux : si un minimum de I_W est toujours atteint, ce n'est, peut-être, qu'un minimum local. Le choix de la première partition $C_k^{(0)}$ est donc important : on peut choisir le contenu des classes au hasard, ou prendre les points les plus éloignés en faisant d'abord une ACP, ou prendre les observations en fonction de connaissances *a priori*.

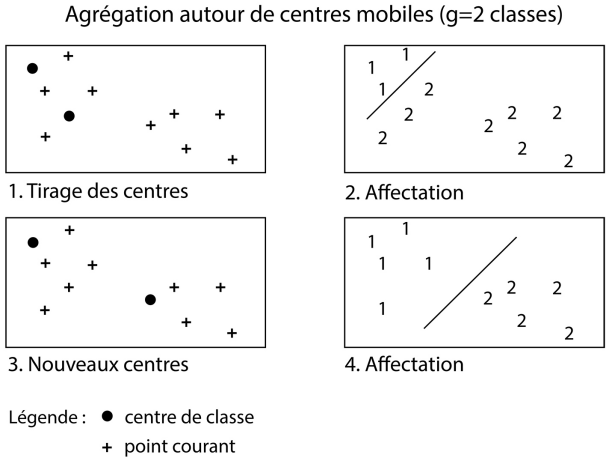


FIG. 2 : Etapes d'une classification par centres mobiles. Figure tirée de Bellanger & Tomassone (2014)

Il existe de nombreuses méthodes de type ***k*-means** [5] (*cf.* par exemple Hartigan & Wong, 1979). Elles diffèrent suivant la sélection des k objets initiaux, le critère de classification, la prise en compte de données catégorielles ou mixtes. La plus connue est la méthode des **centres mobiles** (Forgy, 1965).

Les méthodes de type k-means ont de nombreux avantages : simplicité, compréhension, applicables à des données de grande taille. Cependant, ces méthodes ont le gros

[5] Les méthodes dites *k-means* ou *k-moyennes* regroupent des techniques visant à partitionner un ensemble de points en k groupes (ou clusters) de telle sorte que les points d'un même groupe soient les plus proches possibles les uns des autres. Plus précisément, l'algorithme cherche à minimiser la *variance intra-classe*, c'est-à-dire la somme des distances au carré entre chaque point et le centre de son groupe. Chaque groupe est représenté par un centre, appelé *centroïde*, qui correspond à la moyenne des points qui lui appartiennent.

désavantage d'être sensibles aux outliers* et de fournir des classes **convexes** [6]. Ainsi, dans le cas où les données ne sont pas parfaitement structurées et où la présence d'outliers n'est pas rare, comme cela peut être le cas pour les jeux de données en archéologie, cette méthode peut ne pas être adaptée.

Il existe de nombreuses variations des méthodes k-means pour lesquelles chaque classe est représentée par un objet autre que la moyenne ; citons k-medians ou k-medoids. Pour plus de détails, voir Nakache & Confais (2005) ou Kaufman & Rousseeuw (1990).

2.3 Nombre de classes K à retenir : le graphique des silhouettes

L'**indice silhouette** [7] S_i se définit pour chaque observation $\boldsymbol{i}$ dont la classe obtenue est notée C_k. Il est défini par :

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

où :

- $a_i = \frac{1}{n_k - 1} \sum_{\boldsymbol{j} \in C_k, \boldsymbol{j} \neq \boldsymbol{i}} d(\boldsymbol{i}, \boldsymbol{j})$ est la distance moyenne entre l'observation $\boldsymbol{i}$ et toutes les observations de sa classe C_k d'effectif n_k.
- $b_i = \frac{1}{n_{k'}} \sum_{\boldsymbol{i}' \in C_{k'}} d(\boldsymbol{i}, \boldsymbol{i}')$ est la distance moyenne entre l'observation $\boldsymbol{i}$ et les observations du groupe le plus voisin $C_{k'}$. La classe $C_{k'}$ ainsi trouvée est la deuxième meilleure classe de l'observation $\boldsymbol{i}$.

L'indice silhouette S_i prend des valeurs comprises entre -1 et $+1$. Son interprétation est simple :

$S_i = 1 \Rightarrow$ l'observation $\boldsymbol{i}$ est bien classée,
$S_i = 0 \Rightarrow$ l'observation $\boldsymbol{i}$ se situe entre 2 classes,
$S_i = -1 \Rightarrow$ l'observation $\boldsymbol{i}$ est mal classée.

Le *graphique des silhouettes* est obtenu en rangeant les S_i par valeur décroissante dans chaque classe.

Pour étudier la qualité d'une partition obtenue ; mais aussi pour déterminer le nombre de classes à retenir, on peut moyenner groupe par groupe l'indice silhouette pour les K classes retenues :

[6] Une classe est dite *convexe* si, pour tous les couples de points qu'elle contient, le segment qui les relie reste entièrement à l'intérieur de la classe, sans vides ni trous. Dans ce cas, le barycentre g (centre moyen) représente bien le centre de la classe. Si la classe n'est pas convexe, le barycentre peut se trouver dans une zone vide, donnant une impression trompeuse de la répartition réelle des points. Autrement dit, le barycentre reflète fidèlement le centre de la classe uniquement si celle-ci est convexe.

[7] L'indice silhouette est une mesure qui permet d'évaluer la qualité d'un regroupement (clustering). Il indique, pour chaque observation, à quel point elle est bien placée dans sa classe. Pour chaque observation, la valeur S reflète à la fois sa cohésion (proximité avec les autres points de sa propre classe) et sa séparation (distance à la classe la plus proche).

$$S_{sil}(K) = \frac{1}{n}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in C_k} S_i$$

La *moyenne des silhouettes* $S_{sil}(K)$ permet de comparer la qualité de différentes partitions obtenue en faisant varier K. On retiendra la partition correspondant à la valeur la plus grande.

3 Classification ascendante hiérarchique (CAH)

Les méthodes de **classification hiérarchique** conduisent à une chaîne de partitions emboîtées que l'on peut représenter par un **arbre de classification** appelé **dendrogramme** : elles ont comme point de départ une matrice $\mathbf{D}_{n\times n}$ (des distances ou plus généralement des dissimilarités entre observations) obtenue soit par un calcul intermédiaire à partir des données de base contenues dans $\mathbf{X}_{n\times p}$; soit directement disponible. La plus connue est la **classification ascendante hiérarchique**, notée CAH (*cf.* 2.2.8).

3.1 Le principe

Calculer $\mathbf{D}_{n\times n}$ matrice des dissimilarités entre individus. Chaque individu constitue une classe.

1. Détecter les deux classes les plus proches et les agréger pour n'en former qu'une.
2. Calculer le nouveau tableau des dissimilarités entre classes (entre une classe et un individu) à l'aide d'une mesure de dissimilarité adaptée appelée stratégie d'**agrégation**.
3. Itérer les étapes 1 et 2 jusqu'à ce que tous les individus forment une seule classe.

Le dendrogramme (Figure 3) permet alors de visualiser graphiquement la hiérarchie des partitions imbriquées qui sont autant de solutions potentielles (*cf.* 2.2.8, figure 4).

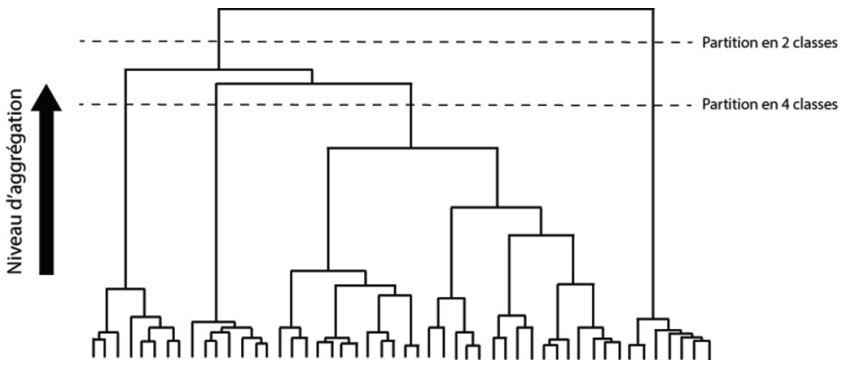


FIG. 3 : Exemple de dendrogramme associée à une CAH avec coupure visuelle en 2 et 4 classes

3.2 Exemple de la stratégie d'agrégation du lien simple

A partir d'une matrice des dissimilarités $\mathbf{D}$, cette stratégie consiste à :

1. Chercher la plus petite distance $d(\boldsymbol{i}_1, \boldsymbol{i}_2)$ qui fournit les deux observations les plus voisines $\{\boldsymbol{i}_1, \boldsymbol{i}_2\}$ qui sont choisies pour former un premier groupe $G^{(1)}(\boldsymbol{i}_1, \boldsymbol{i}_2)$; il y a donc $n-2$ observations encore isolées.
2. Chercher la seconde plus petite distance entre les n observations ; deux possibilités se présentent :
   - soit ce sont deux observations différentes des premières : elles forment alors un nouveau groupe $G^{(2)}(\boldsymbol{i}_3, \boldsymbol{i}_4)$; il y a donc $n-4$ observations encore isolées.
   - soit l'une des deux observations est l'une des deux premières, $\boldsymbol{i}_1$ ou $\boldsymbol{i}_2$: dans ce cas la nouvelle observation $\boldsymbol{i}_3$ est agrégée au premier groupe pour former un nouveau groupe à trois observations $G^{(2)}(\boldsymbol{i}_1, \boldsymbol{i}_2, \boldsymbol{i}_3)$; il y a donc $n-3$ observations encore isolées, soit :

$$d_{\min}\left(G^{(1)}(\boldsymbol{i}_1, \boldsymbol{i}_2), \boldsymbol{i}_3\right) = \min\left(d(\boldsymbol{i}_1, \boldsymbol{i}_3); d(\boldsymbol{i}_2, \boldsymbol{i}_3)\right)$$

3. Recommencer l'opération pour les observations isolées. Successivement, les observations sont ainsi regroupées jusqu'à ne former plus qu'un seul groupe.

Cette stratégie, appelée agrégation par **lien simple**, est donc élémentaire. Elle repose, dans la seconde étape, sur la définition de la distance entre une observation et un groupe ; ici le choix s'est porté sur la distance minimale entre une observation non encore agrégée et une des deux observations qui le sont déjà. Les autres stratégies diffèrent de celle-ci par ce choix.

Considérons l'exemple de cinq observations $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}, \boldsymbol{e}\}$ tiré de Bellanger & Tomassone (2014). Les dissimilarités (ici ce ne sont pas des distances) sont données par la matrice symétrique $\mathbf{D}_{5\times5}$ suivante :

$$\mathbf{D} = \begin{array}{c} \\ \boldsymbol{a} \\ \boldsymbol{b} \\ \boldsymbol{c} \\ \boldsymbol{d} \\ \boldsymbol{e} \end{array} \begin{array}{c} \begin{array}{ccccc} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{c} & \boldsymbol{d} & \boldsymbol{e} \end{array} \\ \begin{bmatrix} 0 & 6 & 14 & 6 & 8 \\ 6 & 0 & 8 & 8 & 2 \\ 14 & 8 & 0 & 4 & 12 \\ 6 & 8 & 4 & 0 & 1 \\ 8 & 2 & 12 & 1 & 0 \end{bmatrix} \end{array}$$

Au niveau 1 (Figure 4), les observations $\boldsymbol{d}$ et $\boldsymbol{e}$ sont d'abord regroupées. Puis au niveau 2, $\boldsymbol{b}$ voisine de $\boldsymbol{e}$ est agrégée ; ensuite au niveau 3, $\boldsymbol{c}$ voisine de $\boldsymbol{d}$; enfin au niveau 4, $\boldsymbol{a}$ à égale distance de $\boldsymbol{b}$ et de $\boldsymbol{d}$.

Cet exemple illustre l'**effet de chaînage** caractéristique de la stratégie : les observations ont tendance à se regrouper de proche en proche s'il n'y a pas de séparation nette entre observations.

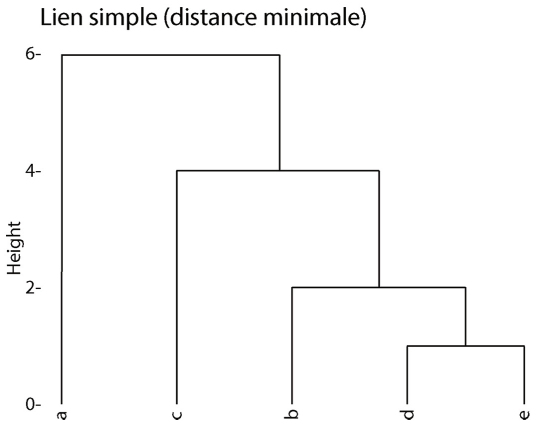


FIG. 4 : Dendrogramme obtenu par lien simple, pour la matrice des dissimilarités $\mathbf{D}$, ($n = 5$)

3.3 Autres stratégies d'agrégation

Nous avons noté que la définition de la distance entre une observation et un groupe pouvait jouer un rôle crucial dans la stratégie d'agrégation. Nous avons pris comme distance la plus petite, mais nous aurions aussi pu choisir la plus grande ou la moyenne des distances aux observations déjà regroupées.

Reprenons l'exemple précédent des cinq observations $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}, \boldsymbol{e}\}$. Une fois choisis $\boldsymbol{d}$ et $\boldsymbol{e}$ pour former le premier groupe, si nous choisissons la stratégie de la distance maximale, à savoir :

$$d_{\max}\left(G^{(1)}\left(\boldsymbol{i}_1, \boldsymbol{i}_2\right), \boldsymbol{i}_3\right) = \max\left(d\left(\boldsymbol{i}_1, \boldsymbol{i}_3\right); d\left(\boldsymbol{i}_2, \boldsymbol{i}_3\right)\right)\Big)$$

nous aurons :

$$\begin{aligned} d_{\max}\left(\boldsymbol{a}, G^{(1)}(\boldsymbol{d}, \boldsymbol{e})\right) &= \max(6; 8) = 8 \\ d_{\max}\left(\boldsymbol{b}, G^{(1)}(\boldsymbol{d}, \boldsymbol{e})\right) &= \max(8; 2) = 8 \\ d_{\max}\left(\boldsymbol{c}, G^{(1)}(\boldsymbol{d}, \boldsymbol{e})\right) &= \max(4; 12) = 12 \end{aligned}$$

toutes ces quantités sont supérieures à $d(\boldsymbol{a}, \boldsymbol{b}) = 6$, donc le nouveau groupe est $G^{(2)}(\boldsymbol{a}, \boldsymbol{b})$. Cette stratégie s'appelle celle du **saut maximal** ou **lien complet**.

Si nous choisissons la stratégie de la distance moyenne, nous aurons comme distance des trois autres observations à $G^{(1)}(\boldsymbol{d}, \boldsymbol{e})$:

$$\begin{aligned} d\left(\boldsymbol{a}, G^{(1)}(\boldsymbol{d}, \boldsymbol{e})\right) &= (6 + 8)/2 = 7 \\ d\left(\boldsymbol{b}, G^{(1)}(\boldsymbol{d}, \boldsymbol{e})\right) &= (8 + 2)/2 = 5 \\ d\left(\boldsymbol{c}, G^{(1)}(\boldsymbol{d}, \boldsymbol{e})\right) &= (4 + 12)/2 = 8 \end{aligned}$$

Enfin, comme on peut noter que lorsque l'on remplace deux classes A et B par leur réunion, la diminution de l'inertie interclasse (et donc l'augmentation de l'inertie in-

TAB. 1 : Paramétrisation des stratégies d'agrégation n_k représente l'effectif de la classe C_k. (Everitt *et al.*, 2001).

$\alpha_K{}^{(*)}$	β	γ	Nom de la stratégie / En anglais	Remarques
$\frac{1}{2}$	0	$-\frac{1}{2}$	Voisin le plus proche / Single linkage	Classes formées de singleton. Effet de chaînage. Pas de prise en compte de la str ucture des classes. Utilise distance ou dissimilarité.
$\frac{1}{2}$	0	$\frac{1}{2}$	Voisin le plus éloigné / Complete linkage	Trend à former des classes compactes. Pas de prise en compte de la structure des classes. Utilise distance ou dissimilarité.
$\frac{n_K}{n_K+n_L}$	0	0	Moyenne du groupe / Average linkage	Tend à regrouper les classes de faible variance. Prise en compte de la structure des classes. Utilise distance ou dissimilarité.
1/2	0	0	Moyenne non pondérée / Mc Quitty linkage	Prise en compte de la structure des classes. Utilise distance ou dissimilarité.
$\frac{n_K}{n_K+n_L}$	$\frac{-n_K n_L}{(n_K+n_L)^2}$	0	Centroïde / Centroid linkage	Effet prépondérant de la classe qui possède l'effectif maximal. Inversions possibles. Utilise distance.
$\frac{1}{2}$	$-\frac{1}{4}$	0	Médiane / Median linkage	Inversions possibles. Utilise distance.
$\frac{n_K+n_M}{n_K+n_L+n_M}$	$\frac{-n_M}{n_K+n_L+n_M}$	0	Variance minimale / Ward's method	Sensible aux aberrants et aux outliers. Tend à produire des classes de même taille. Utilise distance.
$^{(*)}$: pour α_L échanger les indices K et L				

traclasse) est égale à $\frac{n_A n_B}{n(n_A n_B)} d^2\left(\boldsymbol{g}_{A,}, \boldsymbol{g}_B\right)$, une autre stratégie possible, la stratégie de **Ward** consiste à choisir à chaque étape le regroupement de classes tel que l'augmentation de l'inertie intraclasse soit minimale.

La formule de Lance et Williams (Lance & Williams, 1967) permet d'unifier plusieurs stratégies d'agrégation en CAH. Si, à une étape déterminée k de l'agrégation, nous avons deux groupes $G^{(\cdot)}(K)$ et $G^{(\cdot)}(L)$, respectivement formés de K et L observations, déjà fusionnés en $G^{(\cdot)}(K+L)$ et un groupe $G^{(\cdot)}(M)$ formé de M observations, nous n'avons besoin que de trois distances (ou dissimilarités) $d(K,L)$, $d(K,M)$ et $d(L,M)$ pour calculer :

$$d\left(G^{(\cdot)}(K+L), G^{(\cdot)}(M)\right) = \alpha_K d(K,M) + \alpha_L d(L,M) + \beta d(K,L) + \gamma |d(K,M) - d(L,M)|$$

où les coefficients $\{\alpha_K, \alpha_L, \beta, \gamma\}$ dépendent de la stratégie d'agrégation utilisée et satisfont $\alpha_K + \alpha_L = 1, \beta < 1, |\gamma| < 1$ pour ne pas créer d'inversions dans le dendrogramme. Le Tableau 1 présente les plus plus courantes.

Enfin, le regroupement hiérarchique peut aussi se faire de manière descendante. Dans ce cas, on parlera de **classification descendante hiérarchique**. Partant de toutes les observations qui forment au départ un seul groupe :

1. On commence par choisir l'observation qui est la plus éloignée de toutes les autres, elle constitue le noyau d'un nouveau groupe,
2. On affecte à ce nouveau groupe toutes les observations qui sont plus proches de lui que du groupe initial, on a donc deux groupes,
3. On recommence l'opération à l'intérieur de chacun des groupes jusqu'à ce que l'on obtienne une seule observation par groupe.

Cette procédure algorithmique est donc à l'opposé de celle de la CAH présentée précédemment. Il existe des techniques descendantes bien adaptées aux données binaires $\{0, 1\}$, aux variables dites dichotomiques. Dans cette situation, une façon élémentaire pour diviser les observations consiste à séparer les données selon les valeurs d'une des variables. La variable retenue pour réaliser la division est choisie de manière à ce que les sous-classes obtenues soient aussi homogènes que possible. Le but de cette méthode est donc de trouver LA variable qui est le meilleur prédicteur des différences au sein du corpus d'observations étudié. Le résultat de cette classification monothétique aboutit à une hiérarchie de classes où, à chaque étape, une classe est définie par la valeur d'une des variables binaires. La technique n'utilise donc pas des distances ou des ressemblances mais les valeurs des variables l'une après l'autre au travers d'une mesure d'association entre variables. La question fondamentale est celle du choix des variables dans cette procédure pas à pas. Pour plus de détails, on pourra consulter Kaufman et Rousseeuw (1990), Rousseeuw (1986).

4 Bilan

Les méthodes de classification non supervisée permettent de regrouper des objets qui ont des caractéristiques semblables quand l'affectation à des classes n'est pas connue *a priori*. Il existe de nombreux algorithmes ! Nous n'avons présenté ici que les plus classiques (k-means et CAH) adaptés à la classification d'observations. Nous renvoyons le lecteur intéressé aux nombreux ouvrages sur le sujet. Il est important de noter que l'identification du « bon » nombre de classes donc de la meilleure partition est un problème de recherche « ouvert » en classification non supervisée de type hiérarchique ou k-means, c'est-à-dire non résolu, qui donne lieu à de nombreux travaux. L'indice silhouette (§2.3) fait partie des indicateurs pouvant aider à décider du "bon" nombre de classes à conserver. Parmi les méthodes de classification non supervisées non abordées ici :

- Les approches probabilistes basées sur les **modèles de mélange** [8]

  (Fraley & Raftery, 2002) fournissent un cadre formel pour résoudre des problèmes pratiques qui se posent en classification, tels que la détermination du nombre de classes, et permettent d'estimer l'incertitude associée à la classification. Les limites de ces méthodes résident principalement dans le choix de la loi de probabilité des composantes du mélange,

- Les méthodes de classification de variables permettant de construire des groupes de variables qui rassemblent les variables qui portent les mêmes informations (redondantes, corrélées), et qui dissocient les variables qui expriment des informations complémentaires (Nakache & Confais, 2005 ; Chavent *et al.*, 2012).

[8] Les *modèles de mélange* sont des méthodes de clustering qui supposent que les données proviennent d'un mélange de plusieurs *sous-populations* sous-jacentes, chacune correspondant à une classe et représentée par un modèle statistique (par exemple une distribution normale). Chaque point a une certaine probabilité d'appartenir à une sous-population (et donc à la classe correspondante). Ainsi, cette méthode permet à la fois de regrouper les points en classes et de *classer* chaque point de manière probabiliste selon son appartenance aux différentes classes. Contrairement aux méthodes comme k-means, qui attribuent chaque point à une seule classe de façon stricte, les modèles de mélange tiennent compte de l'incertitude dans l'appartenance des points.

Cependant, il faut avoir conscience qu'aucune méthode de classification non supervisée n'est entièrement satisfaisante et qu'il existe peu de règles absolues pour les choisir. Voici quelques éléments de réflexion :

- Le choix de la mesure de distance ou de dissimilarité est crucial et doit faire l'objet d'un examen préalable minutieux. La distance euclidienne n'est pas toujours adaptée !
- A moins qu'il n'existe une séparation nette entre les observations, il ne faut pas s'attendre à avoir des résultats parfaitement clairs,
- La présence de données suspectes rend les résultats très instables ; la mise en œuvre de différentes techniques de classification peuvent aider à les identifier,
- La comparaison des résultats obtenus par plusieurs stratégies de classification est souvent utile,
- La possibilité d'associer à la classification une représentation graphique qui s'apparente à un plan ou une carte[9] à l'aide de méthodes factorielles doit toujours être exploitée.

[9] On parle ici d'une représentation graphique dans un plan factoriel (*cf.* 6), comparable à une "carte" et qui permet de visualiser les classes issues de la classification statistique ainsi que leurs relations. Ce type de représentation est fréquemment utilisé dans les études de provenance (*cf.* 2.1.4, 2.1.5, 2.1.6) afin de visualiser les distances statistiques entre des groupes de compositions élémentaires.