

# STA211 : Méthodes descriptives pour le pré-traitement des données (classification)

Vincent Audigier, Ndèye Niang-Kéita

01 mars, 2019

- 1 Introduction
- 2 Méthodes de partitionnement
  - 2.1 Agrégation autour des centres mobiles
    - 2.1.1 Principe
    - 2.1.2 Propriétés
    - 2.1.3 Choix du nombre de classes
    - 2.1.4 Formes fortes
  - 2.2 Autres méthodes de partitionnement
- 3 Méthodes hiérarchiques
  - 3.1 Distance entre individus ou groupes
  - 3.2 Algorithme de classification ascendante hiérarchique
    - 3.2.1 Construction de la suite de partitions imbriquées
    - 3.2.2 Construction de l'arbre hiérarchique
  - 3.3 Choix du nombre de classes
  - 3.4 Propriétés
    - 3.4.1 Choix de la mesure de dissimilarité
    - 3.4.2 Choix du critère d'agrégation
- 4 Méthodes mixtes
- 5 Description des classes
  - 5.1 Variables quantitatives
  - 5.2 Variables qualitatives
- 6 Données qualitatives
- 7 Complémentarité de l'analyse factorielle et des méthodes de classification
- Références

## 1 Introduction

Les méthodes d'analyse factorielle permettent de mettre en évidence les ressemblances entre individus ainsi que les liaisons entre variables, et ainsi d'identifier des groupes (ou *classes*) d'individus ou de variables similaires. Néanmoins, ces groupes restent assez subjectifs dans le sens où les méthodes d'analyse factorielle ne permettent pas d'affecter de façon automatique et claire un objet (individu ou variable) à un groupe.

La classification a pour objectif de former une (ou plusieurs) partition(s) d'objets, i.e. de définir  $K$  groupes d'objets ( $C_1, \dots, C_K$ ) sans recouvrements ni intersections, de sorte que chaque objet soit affecté à un groupe et à un seul. On parle également de *segmentation*, de *clustering*, ou de *classification non-supervisée*. Tout comme les méthodes d'analyse factorielle, ces méthodes s'appliquent en général sur un tableau de

données rectangulaire  $X = (x_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$  croisant des individus en lignes et des variables en colonnes (bien

que cela ne soit pas toujours nécessaire comme nous le verrons plus tard). Nous considérerons dans un premier temps que toutes les variables sont quantitatives, mais ces méthodes peuvent aussi être étendues à des données qualitatives ou mixtes.

La classification est généralement effectuée sur les individus, bien qu'il soit également possible de classifier les variables (ce sujet sera abordé dans un prochain cours). Ces groupes d'individus ne sont pas connus *a priori*, mais sont en quelque sorte "dictés" par les données. Sur les données *German Credit* par exemple, il ne s'agira pas de retrouver les bons et les mauvais payeurs, mais plutôt d'identifier clairement les profils de clients d'un point de vue multidimensionnel. Par exemple, en se basant sur les variables relatives au type de crédit, il pourrait s'agir d'identifier les profils de clients du point de vue du type de crédit demandé. Les groupes ne sont alors pas connus à l'avance.

La classification fournit une méthodologie différente et complémentaire à l'analyse factorielle pour analyser les données multivariées. Ainsi, afin de rester dans un cadre similaire à celui de l'analyse factorielle, on présente ici des méthodes de classification géométriques, basées sur des distances. Parmi elles, on distingue les méthodes de partitionnement et les méthodes hiérarchiques. Notons, qu'il existe cependant d'autres méthodes basées sur des modèles, mais non abordées ici.

## 2 Méthodes de partitionnement

Les méthodes de partitionnement consistent à définir une partition de l'ensemble des individus en  $K$  (défini à l'avance) groupes. La plus classique d'entre elles est la méthode d'agrégation autour des centres mobiles, aussi appelée méthode des  $k$ -means.

### 2.1 Agrégation autour des centres mobiles

On se place ici dans un espace Euclidien, c'est-à-dire un espace (ici  $\mathbb{R}^p$ ) muni d'un produit scalaire définissant une distance entre les éléments de cet espace (ici les individus). Cette distance (Euclidienne) va permettre de formaliser la notion de ressemblance entre individus. Parmi les distances Euclidiennes, on retrouve les deux distances déjà évoquées en ACP, i.e.

- la distance usuelle :  $d(i, i') = \sqrt{\sum_{j=1}^p (x_{ij} - x_{i'j})^2}$
- la distance usuelle pondérée par l'inverse de la variance :  $d(i, i') = \sqrt{\sum_{j=1}^p \frac{(x_{ij} - x_{i'j})^2}{s_{x_j}^2}}$

correspondant respectivement à la distance entre deux individus  $i$  et  $i'$  dans le cas non-normé et normé.

Citons aussi la distance de Mahalanobis définie par  $d(i, i') = \sqrt{(x_{i1}, \dots, x_{ip}) V^{-1} (x_{i'1}, \dots, x_{i'p})^\top}$  où

$V^{-1}$  désigne l'inverse de la matrice de variance-covariance des  $p$  variables. A l'image de la distance usuelle pondérée par l'inverse de la variance des variables, qui diminue l'effet de la variance dans le calcul des distances, la pondération par l'inverse de la matrice de variance-covariance permet de diminuer l'influence de la covariance entre les variables.

A partir de la définition d'une distance entre les individus, il va être possible de définir l'inertie d'un ensemble d'individus et, de ce fait, la qualité d'une partition. En effet, on cherchera à minimiser l'inertie au sein des groupes constitués (*inertie intra-classe*), de façon à avoir des classes les plus homogènes possibles, traduisant une classification de qualité. L'inertie intra-classe s'oppose à l'*inertie inter-classes*, correspondant à l'inertie

entre les barycentres de chaque groupe. La relation de Huygens relie l'inertie intra-classe à l'inertie inter-classes selon la relation

$$\text{Inertie Totale} = \text{Inertie inter-classes} + \text{Inertie intra-classe}$$

$$\text{autrement dit selon } \frac{1}{n} \sum_{i=1}^n d^2(i, G) = \sum_{k=1}^K \frac{n_k}{n} d^2(G_k, G) + \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^n d^2(i, G_k)$$

où  $K$  désigne le nombre de classes et  $k$  l'indice d'une de ces classes,  $n_k$  désigne l'effectif de la classe  $C_k$ ,  $G$  est le barycentre du nuage de points et  $G_k$  le barycentre du sous nuage défini par la classe  $C_k$ .

Dans la mesure où l'inertie totale du nuage de points est constante quelque soit le partitionnement en  $K$  classes considéré, cette relation nous permet de dire qu'il est équivalent de rechercher la partition  $(C_1, \dots, C_K)$  qui minimise l'inertie intra-groupe, ou celle qui maximise l'inertie inter-groupes. Le problème de la recherche de la meilleure partition se résume à la recherche de la partition  $C_1, \dots, C_K$  minimisant

$$\sum_{k=1}^K \sum_{i=1}^n d^2(i, G_k)$$

Néanmoins, il n'est pas possible de résoudre ce problème globalement, dans la mesure où le nombre de partitions possibles en  $K$  classes de  $n$  éléments est trop grand pour permettre une recherche exhaustive (cf Figure 2.1).

$n$	$k$	1	2	3	4	5	6	7	8	9	10	11	$P_n$
1	1	1											1
2	1		1										2
3	1		3	1									5
4	1		7	6	1								15
5	1		15	25	10	1							52
6	1		31	90	65	15	1						203
7	1		63	301	350	140	21	1					877
8	1		127	966	1 701	1 050	266	28	1				4 140
9	1		255	3 025	7 770	6 951	2 646	462	36	1			21 147
10	1		511	9 330	34 105	42 525	22 827	5 880	750	45	1		115 975
11	1		1 023	28 501	145 750	246 730	179 487	63 987	11 880	1 155	55	1	678 970
12	1		2 047	86 526	611 501	1 379 400	1 323 652	627 396	159 027	22 275	1 705	66	4 213 597

Figure 2.1: Nombre de partitions possibles pour un nombre  $n$  d'individus et un nombre  $k$  de classes. Source Saporta (2006)

L'algorithme d'agrégation autour des centres mobiles fournit une solution localement optimale à ce problème.

## 2.1.1 Principe

L'algorithme d'agrégation autour de centres mobiles (aussi appelé algorithme des  $k - means$ ) proposé par Forgy (1965) est le suivant

### Algorithme d'agrégation autour de centre mobiles

1. Initialisation : tirer au hasard, ou sélectionner pour des raisons extérieures à la méthode,  $K$  points dans l'espace des individus, en général  $K$  individus de l'ensemble, appelés *centres* ou *noyaux*
2. Répéter :

- Allouer chaque individu au centre (c'est-à-dire à la classe) le plus proche au sens de la distance choisie
- Calculer le centre de gravité de chaque classe, qui devient le nouveau noyau.
- S'arrêter si le critère de variance intra-classe ne diminue plus (ou de façon équivalente, si le critère de variance inter-classes ne croît plus)

L'étape d'allocation consiste à réaffecter tous les individus à une classe et donc de mettre à jour la partition courante. On notera qu'il est possible qu'une classe (ou plusieurs) se vide lors de cette étape. Dans ce cas, on pourra éventuellement retirer aléatoirement un noyau complémentaire ou simplement poursuivre avec une classification à  $K - 1$  classes. L'algorithme s'arrête quand les classes sont stables. Il est aussi possible de l'arrêter après un nombre d'itérations défini à l'avance.

Une illustration de l'algorithme dans un cas bivarié est proposée en Figure 2.2 pour un choix de  $K = 3$ .

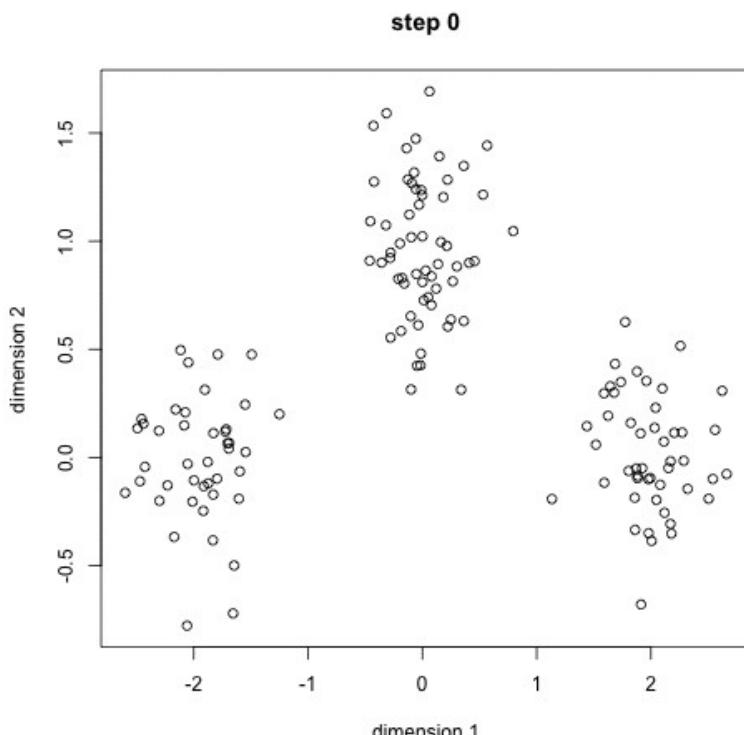


Figure 2.2: Agrégation autour des centres mobiles. Source <https://www.kdnuggets.com/2018/06/5-clustering-algorithms-data-scientists-need-know.html> (<https://www.kdnuggets.com/2018/06/5-clustering-algorithms-data-scientists-need-know.html>)

## 2.1.2 Propriétés

A chaque étape de l'algorithme, l'inertie intra-classe diminue. En effet, le barycentre d'une classe est le point qui minimise la somme des écarts au carré avec les individus de cette classe. Par conséquent, l'étape de mise à jour des centres des classes par le barycentre ne peut que diminuer l'inertie intra-classe. Par ailleurs, une fois cette mise à jour effectuée, la réallocation des individus au centre le plus proche ne peut que diminuer l'inertie intra-classe. Cette propriété assure que pour une configuration initiale donnée, on obtiendra une partition d'inertie intra-classe minimale. Néanmoins, pour deux configurations initiales différentes, rien n'assure que l'on obtienne les mêmes partitions à l'issue de l'algorithme. Par conséquent, ce minimum d'inertie intra-classe n'est qu'un minimum *local*. En pratique, on pourra exécuter l'algorithme pour plusieurs initialisations et retenir la partition minimisant la variance intra-classe.

Cet algorithme a une complexité linéaire avec le nombre d'individus, ce qui est faible. Ainsi, il pourra être facilement appliqué sur des jeux de données comportant un grand nombre d'individus. En pratique, la convergence est atteinte en quelques dizaines d'itérations maximum.

Enfin, il faut noter qu'en plus du choix de la distance utilisée, cet algorithme nécessite de spécifier le nombre de classes de façon a priori.

### 2.1.3 Choix du nombre de classes

Parfois, le nombre de classes à considérer est assez évident, comme sur les données en Figure 2.2, une visualisation des données suffisant à le déterminer. Néanmoins, l'utilisateur n'a généralement aucune idée a priori du nombre de classes le plus pertinent pour ses données. On pourrait penser qu'il suffirait alors d'exécuter plusieurs fois l'algorithme pour des nombres de classes différents, puis retenir la partition dont l'inertie intra-classe est minimale. Ceci n'est pas possible, car l'inertie intra-classe entre deux partitions à nombre de classes différent n'est pas comparable. En effet, plus le nombre de classes est grand, plus l'inertie intra-classe est petite. Dans le cas extrême où  $K = 1$ , l'inertie intra-classe atteint la valeur de l'inertie totale du nuage, tandis qu'elle vaut 0 si  $K = n$ . Dès lors, un choix du nombre de classes basé sur l'inertie intra doit être relativisé au nombre de classes considéré. Pour cela, une stratégie classique consiste à représenter l'inertie intra-classe en fonction du nombre de classes, puis de visualiser le coude où l'adjonction d'une classe ne correspond à rien dans la structuration des données.

### 2.1.4 Formes fortes

Etant donné que l'algorithme d'agrégation autour des centres mobiles est sensible à son initialisation, il peut être intéressant d'analyser les intersections des partitions obtenues, appelées *formes fortes*. En effet, si pour deux initialisations de l'algorithme (ou plus) un regroupement d'individus est toujours au sein d'une même classe, alors ceci indique que regroupement est stable, traduisant une véritable structuration des données. Les regroupements d'individus tantôt dans une classe, tantôt dans une autre, correspondent quant à eux aux *formes faibles*. Il est important de noter que le label (i.e. le numéro) d'une classe est purement arbitraire et que par conséquent, pour deux exécutions de l'algorithme, les partitions peuvent être identiques alors que les numéros des classes sont permutés. Dans pareil cas, on n'observera pas de formes faibles, mais uniquement des formes fortes.

La figure 2.3 illustre l'analyse des formes fortes pour un jeu de données à deux dimensions pour un nombre de classes  $K = 3$  et pour deux initialisations différentes.

		2ème exécution		
		C1	C2	C3
1ère exécution	C1	30	0	72
	C2	0	99	1
	C3	98	0	0

On observe les coïncidences entre les classes. C<sub>3</sub> de la 1<sup>ère</sup> exécution correspond au C<sub>1</sub> de la 2<sup>nde</sup>, etc.

Les zones d'indécisions (en gris) correspondent à des zones frontières entre les classes. « Formes faibles ».

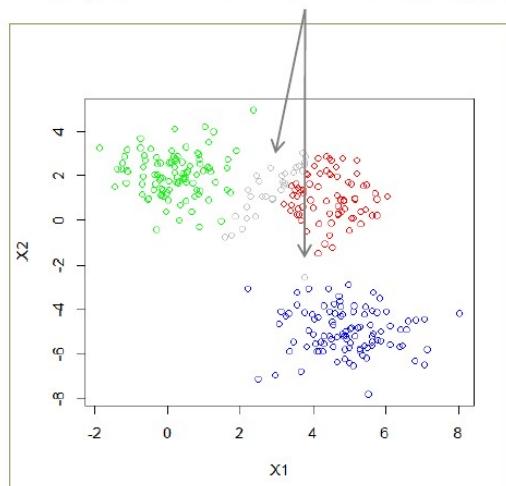


Figure 2.3: Illustration de l'analyse des formes fortes. Source [http://eric.univ-lyon2.fr/~ricco/cours/slides/classif\\_centres\\_mobiles.pdf](http://eric.univ-lyon2.fr/~ricco/cours/slides/classif_centres_mobiles.pdf) ([http://eric.univ-lyon2.fr/~ricco/cours/slides/classif\\_centres\\_mobiles.pdf](http://eric.univ-lyon2.fr/~ricco/cours/slides/classif_centres_mobiles.pdf))

Il apparaît clairement que ce type d'analyse est rapidement limitée quand le nombre d'initialisations est élevé.

## 2.2 Autres méthodes de partitionnement

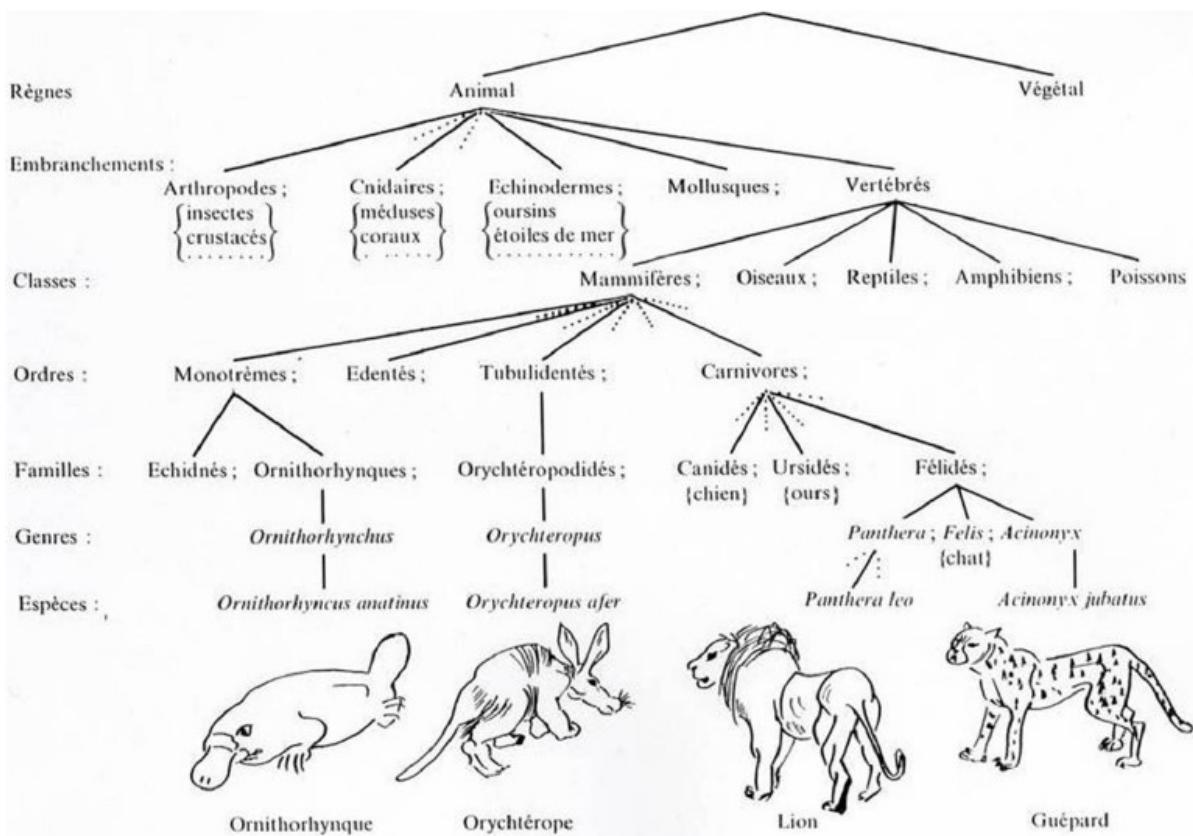
Des variantes de cet algorithme ont été proposées dans la littérature. On peut notamment citer l'algorithme des *k – means* de MacQueen (1967) qui consiste à réaffecter les centres dès qu'un individu change de classe, alors que précédemment l'étape de mise à jour des centres n'avait lieu qu'une fois tous les individus réalloués. Cette variante accélère la convergence, mais rend l'algorithme dépendant à l'ordre dans lequel les individus sont réalloués.

Une autre variante est celle des *nuées dynamiques*, proposée par Diday (1973). Elle consiste à utiliser un mode de représentation d'un groupe d'individus différent de celui du barycentre, par exemple en considérant un sous-ensemble de points (les plus centraux), un plan factoriel, une distribution de probabilité,... cela permet de corriger l'influence d'éventuelles valeurs extrêmes sur le calcul du barycentre.

Enfin, citons la méthode *isodata*, qui est une autre méthode de partitionnement proposée par Ball and Hall (1965). Cette méthode se caractérise par un raffinement de la technique de réallocation dans la méthode des centres mobiles avec une gestion dynamique du nombre de classes. Elle permet notamment d'empêcher la formation de classes d'effectifs trop faibles ou de diamètres trop grands.

## 3 Méthodes hiérarchiques

Les méthodes hiérarchiques constituent l'autre grande famille des méthodes de classification géométriques. Elles visent à construire non pas une partition, mais une suite de partitions imbriquées les unes dans les autres. Cette suite de partitions imbriquées se représentent sous la forme d'un arbre, appelé *arbre hiérarchique* ou *dendrogramme*. L'arbre hiérachique du règne animal constitue un bel exemple de ce type de représentation (cf Figure 3.1). Par exemple, on retrouve au sein des animaux les différents embranchements (Arthropodes, Cnidaires, Echinodermes, Mollusques, Vertébrés), au sein de l'embranchement des vertébrés on retrouve différentes classes (Mammifères, Oiseaux, Reptiles, Amphibiens, Poissons), etc.. Chaque niveau de l'arbre (embranchements, classes, ...) définit ainsi une partition des animaux.



[CL. Nat.J, TIB n°2, § I.1]

Figure 3.1: Arbre hiérarchique du règne animal.

Cette suite de partitions peut se construire de façon ascendante, ce qui signifie que l'on commence par construire la partition à  $K = n$  classes, puis on fusionne les deux classes les plus proches pour obtenir une partition à  $K = n - 1$  classes, etc, jusqu'à obtenir la partition à une seule classe ; soit de façon divisive, ce qui signifie que l'on commence par construire la partition à  $K = 1$  classe, puis on divise cette classe en deux selon un critère, pour obtenir une partition à  $K = 2$  classes, etc, jusqu'à obtenir une partition à  $K = n$  classes. Nous présentons ici uniquement les méthodes ascendantes qui constituent les approches les plus couramment utilisées.

## 3.1 Distance entre individus ou groupes

La méthode de classification ascendante hiérarchique (CAH) nécessite de définir dans un premier temps une distance, ou plus simplement une *mesure de dissimilarité* entre individus. Rappelons qu'une mesure de dissimilarité  $d$ , dans notre contexte, est une fonction de l'espace des individus  $\mathbb{R}^p \times \mathbb{R}^p$  à valeurs dans l'ensemble des réels positifs vérifiant

- $d(i, i') = 0$  si et seulement si  $i = i'$
- $d(i, i') = d(i', i)$

pour tous les couples d'individus  $(i, i')$ . Par rapport à une distance, une mesure de dissimilarité ne vérifie pas nécessairement l'inégalité triangulaire

- $d(i, i') \leq d(i, i'') + d(i'', i')$ .

En ce sens, cette méthode offre potentiellement plus de liberté que la méthode des  $k - means$  nécessitant l'emploi d'une distance Euclidienne.

Cette distance va permettre d'identifier les deux individus les plus proches dans le jeu de données et donc de

définir la partition à  $K = n - 1$  classes. Se pose ensuite la question de la distance entre un groupe d'individus et un individu, puis celle de la distance entre deux groupes. Ceci revient à définir la façon de fusionner des regroupements d'éléments (un élément désignant un individu ou un groupe d'individus), on parle de *critère d'agrégation*. Plusieurs choix sont possibles, par exemple, pour deux classes  $C_1$  et  $C_2$  on peut considérer

- le *saut minimum*  $d(C_1, C_2) = \min_{a \in C_1, b \in C_2} d(a, b)$
- le *saut maximum*  $d(C_1, C_2) = \max_{a \in C_1, b \in C_2} d(a, b)$
- le *saut moyen*  $d(C_1, C_2) = \frac{1}{n_1 \times n_2} \sum_{a \in C_1} \sum_{b \in C_2} d(a, b)$  où  $n_k$  désigne le cardinal de la classe  $k$
- la *distance de Ward*  $d(C_1, C_2) = \sqrt{\frac{n_1 \times n_2}{n_1 + n_2} d^2(G_{C_1}, G_{C_2})}$

Dès lors, il est possible de définir l'algorithme de CAH, définissant la suite de partitions emboîtées, ainsi que sa représentation sous la forme d'un dendrogramme.

## 3.2 Algorithme de classification ascendante hiérarchique

### 3.2.1 Construction de la suite de partitions imbriquées

L'algorithme pour construire la suite de partitions emboîtées est le suivant :

#### Algorithme de classification ascendante hiérarchique

1. Initialisation :

- construire la partition à  $K$  classes où chaque classe contient un unique individu
- calculer la matrice des distances entre individus deux à deux (e.g. distance euclidienne classique)

2. Répéter :

- regrouper les deux classes les plus proches au sens de la distance entre classes choisie (e.g. distance de Ward)
- mettre à jour le tableau des distances en remplaçant les deux classes regroupées par la nouvelle et en calculant sa distance avec chacune des autres classes
- s'arrêter quand il ne reste qu'une seule classe

Cet algorithme est illustré en Figure 3.2 (graphique de gauche) sur un petit jeu de données comportant 6 éléments de  $\mathbb{R}^2$  à classer. La distance entre individus est la distance euclidienne usuelle et le critère d'agrégation est celui du saut moyen.

### 3.2.2 Construction de l'arbre hiérarchique

L'arbre hiérarchique associé à une suite partition s'obtient de manière ascendante. Les éléments terminaux de l'arbre sont appelés *feuilles*, il y a donc autant de feuilles que d'individus. Deux feuilles sont reliées par des *branches* qui se rejoignent au niveau d'un *noeud*. On commence donc par relier les deux feuilles

correspondant aux deux individus les plus proches. Le niveau auquel ce noeud est positionné verticalement dans l'arbre est déterminé à partir de la valeur de la distance entre ces individus (en général, il s'agit exactement de la valeur de cette distance). Par la suite, quand on regroupera un groupe d'individus avec un autre individu, ou un autre groupe, le noeud sera positionné au niveau de la valeur de la distance entre regroupements.

En ``coupant'' l'arbre par une ligne horizontale, on obtient une partition, d'autant plus fine que la section est proche des éléments terminaux. La Figure 3.2 (graphique de droite) illustre la construction de l'arbre hiérarchique pour le petit jeu de données précédent.

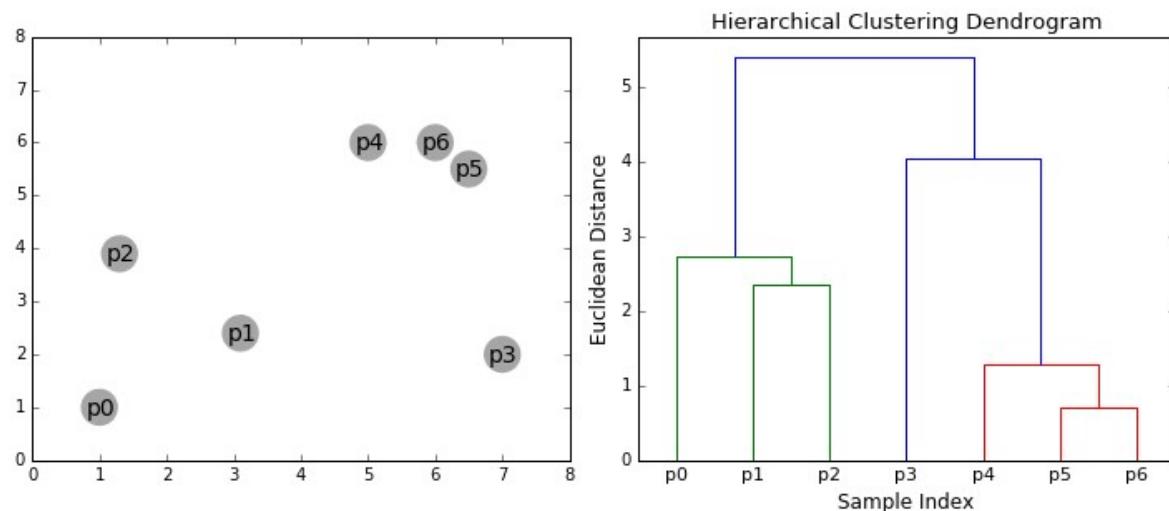


Figure 3.2: Classification ascendante hiérarchique. Source <https://www.kdnuggets.com/2018/06/5-clustering-algorithms-data-scientists-need-know.html> (<https://www.kdnuggets.com/2018/06/5-clustering-algorithms-data-scientists-need-know.html>)

Remarquons que la hiérarchie construite ne présente pas d'inversion pour les critères d'agrégation précédemment énoncés (Section 3.1), i.e. que si un noeud a été construit par agrégation de deux noeuds  $a$  et  $b$  à un certain niveau, alors les agrégations ultérieures ne peuvent se faire qu'à un niveau supérieur ou égal (Lebart, Morineau, and Piron (2006), pp. 282).

### 3.3 Choix du nombre de classes

L'arbre hiérarchique fournit une représentation riche de la structure des individus, dans le sens où il permet de visualiser les ressemblances entre individus et entre groupes d'individus via une succession de partitions emboîtées. Néanmoins, il est souvent nécessaire en pratique d'établir une partition unique des individus. De ce point de vue, le choix du nombre de classes, i.e. du niveau de coupure de l'arbre, s'impose.

La façon classique de procéder est d'analyser les écarts entre les indices d'agrégation, autrement dit la longueur des branches de l'arbre. Des fortes différences entre deux niveaux d'agrégation successifs indiqueront une modification importante de la structure des données lorsqu'on a procédé au regroupement et suggéreront donc de pas fusionner ces groupes (cf Figure 3.3).

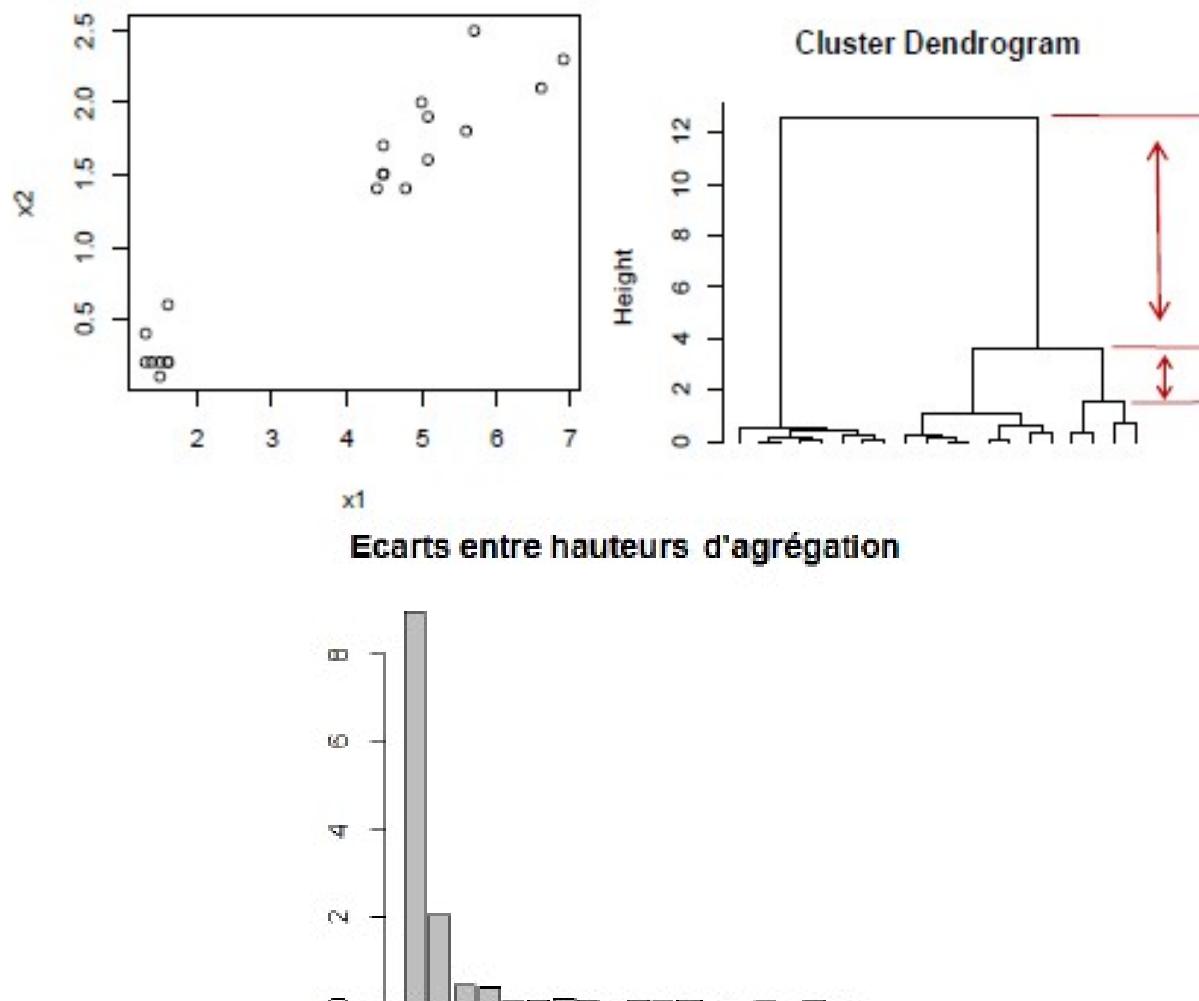


Figure 3.3: Choix du nombre de classes : exemple sur un jeu à deux variables  $x_1$   $x_2$ . De gauche à droite : données, dendrogramme, diagramme en barres des écarts d'inertie

Tout comme pour les méthodes de partitionnement, cette façon de choisir le nombre de classes repose sur un critère statistique qui donne une indication sur le nombre à choisir. Le nombre d'individus sera également un élément important pour choisir le nombre de classes : avoir un nombre de classe proche du nombre d'individus conduirait à avoir des classes de petits effectifs. Aussi l'interprétation qui sera faite des classes peut aussi permettre d'ajuster ce nombre : si une partition plus fine est plus facile à caractériser, alors on pourra avoir intérêt à choisir un nombre de classes plus élevé que celui indiqué par le dendrogramme.

Le choix du nombre de classes reste un problème ouvert. Il existe d'autres façons de choisir le nombre de classes, en particulier en utilisant des critères de validation interne. Ces critères sont basés sur la définition de mesures propres aux classes, comme la distance entre les observations et le centre de la classe à laquelle elles sont rattachées. Ces indices reposent sur les propriétés suivantes d'une ``bonne classification''

- des individus d'une même classe partagent les mêmes propriétés (compacité).
- des individus appartenant à des classes différentes aient peu de propriétés en commun (séparabilité).

Pour évaluer le respect de ces deux notions, différentes mesures basées sur les distances entre les observations et les centres de classe ont été définies pour quantifier l'adéquation entre une partition et l'idée que l'on se fait d'une bonne classification.

Parmi les plus connus, on peut citer l'indice silhouette ou l'indice de Davies-Bouldin, mais il en existe bien d'autres. On se reportera à Charrad et al. (2014) pour une description plus exhaustive de ces indices et leur détermination à l'aide du package R *NbClust*.

## 3.4 Propriétés

D'un point de vue algorithmique, la CAH est très complexe du fait du nombre de distances à calculer entre éléments (de l'ordre de  $n^3$ ). L'algorithme ne peut donc pas être mis en oeuvre sur des jeux de données comportant beaucoup d'individus. Néanmoins, des méthodes plus efficaces ont été développées abaissant à une complexité de l'ordre de  $n^2$  (méthode des voisins réciproques McQuitty (1966)). Aussi, la méthode ne nécessite pas le choix a priori du `` bon'' nombre de classes, au contraire, elle permet de donner une idée de celui-ci. Toutefois elle nécessite le choix d'un critère d'agrégation en plus du choix d'une distance entre individus. De ces choix dépendent également les propriétés de la méthode, nous y revenons par la suite.

### 3.4.1 Choix de la mesure de dissimilarité

Parmi les choix classiques de distances (ou simplement dissimilarités) entre individus, on retrouve les distances Euclidiennes déjà évoquées dans la méthode d'agrégation autour des centres mobiles. Cependant, d'autres mesures peuvent être privilégiées. Par exemple, la distance cosinus

$$d(i, i') = 1 - \cos(i, i') = 1 - \frac{\sum_{j=1}^p (x_{ij} \times x_{i'j})}{\sqrt{\sum_{j=1}^p x_{ij}^2} \times \sqrt{\sum_{j=1}^p x_{i'j}^2}}$$

Celle-ci est notamment populaire en text mining : dans ce cas les individus sont des textes, décrits par les fréquences absolues de certains mots, définissant les variables. La distance cosinus permet alors de comparer des textes de longueurs différentes. En effet, si un texte est long, il aurait tendance à être considéré comme éloigné des autres selon la distance Euclidienne car ses fréquences de mots (i.e. ses coordonnées dans l'espace) seraient bien plus grandes. Autrement dit, la norme du vecteur associé à cet individu serait élevée. La distance cosinus remédie à cela en considérant l'angle entre les vecteurs définis par les individus, sans tenir compte de leur norme.

Le choix de la distance est important et dépend des données analysées. Il convient donc de réfléchir à son choix avec soin. Notons que quand la classification vient en complément d'une analyse factorielle, ce choix s'impose de lui-même, car il a déjà été réfléchi auparavant.

### 3.4.2 Choix du critère d'agrégation

Les critères d'agrégation définissent quel individu sera rattaché au classes préexistantes. En fonction de la forme des classes, certains critères pourront être plus adaptés que d'autres. Le saut minimal a notamment la particularité de détecter des classes allongées, irrégulières, voire sinueuses (cf Figure 3.4). En effet, ce critère tend à rattacher systématiquement l'individu le plus proche du bord d'une classe existante. En procédant de proche en proche, on peut ainsi détecter des classes bien distinctes les unes des autres dans l'espace, même si elles sont proches dans l'espace. A contrario, si deux classes bien distinctes sont reliées l'une à l'autre par une succession de points, alors les classes ne seront pas identifiées. C'est ce qu'on appelle *l'effet de chaîne*. Le saut maximum lui n'aura pas cette tendance et tendra à produire des classes de diamètres égaux. Il est peu utilisé notamment parce qu'il est très sensible aux valeurs aberrantes. Le saut moyen est un intermédiaire entre les deux.

Le critère de Ward est certainement celui qui pourrait paraître le plus naturel, car il consiste à rechercher la partition d'inertie inter-classe la plus forte, et en ce sens, correspond à ce que l'on attend d'une bonne partition. Néanmoins, la notion d'inertie nécessite de se placer dans un espace Euclidien, ce qui n'est pas toujours possible dès lors que les données ne se présentent plus sous la forme d'un tableau individus ×

variables. En effet, en toute rigueur, l'algorithme de CAH nécessite simplement de disposer en entrée d'un tableau de distances individus  $\times$  individus. On peut obtenir ce tableau à partir d'un tableau individus  $\times$  variables, mais ce n'est pas une nécessité. Or, la distance de Ward est une fonction de la distance des barycentres et cette notion de barycentre n'a pas de sens si on ne dispose que les distances entre individus deux à deux.

La distance de Ward a tendance à conduire à l'obtention de classes sphériques et de mêmes effectifs, elle sera donc limitée sur des classes allongées. Par ailleurs du fait qu'elle est fondée sur l'inertie, elle sera naturellement sensible aux valeurs aberrantes. Elle est la méthode la plus utilisée, et est bien adaptée au cadre posé lors d'une analyse factorielle.



Figure 3.4: Quelques exemples de classes aux formes particulières. Source : <https://www.kdnuggets.com/2018/06/5-clustering-algorithms-data-scientists-need-know.html> (<https://www.kdnuggets.com/2018/06/5-clustering-algorithms-data-scientists-need-know.html>)

## 4 Méthodes mixtes

Les méthodes d'agrégation autour des centres mobiles et de classification ascendante hiérarchique possèdent des propriétés différentes, dès lors il devient intéressant d'utiliser les deux conjointement pour bénéficier des avantages de chacune, on parle de *méthodes de classification mixtes*.

La méthode de classification ascendante hiérarchique fournit un moyen assez simple d'avoir une idée du nombre de classes, tandis que la méthode d'agrégation autour des centres mobiles est peu complexe algorithmiquement (contrairement à la CAH) et permet d'obtenir une partition optimale en termes d'inertie. Notons que la CAH par méthode de Ward ne permet généralement pas d'obtenir une partition telle l'inertie intra-classe soit la plus faible possible. En effet, la CAH permet à chaque étape de trouver la partition à  $K - 1$  classes telle que l'inertie intra soit la plus faible, mais ceci en partant de la partition à  $K$  classes. Il s'agit là d'une contrainte forte, empêchant que des individus précédemment dans une même classe soit répartis dans des classes différentes. Dès lors, la partition obtenue à  $K - 1$  classes n'est pas nécessairement la meilleure en termes d'inertie.

L'idée de la classification mixte est alors de commencer par effectuer une agrégation autour des centres mobiles pour un grand nombre de classes (disons 50 pour fixer les idées). On réduit ainsi le nombre d'éléments à classer passant d'un nombre  $n$  d'individus potentiellement grand à un nombre modéré de classes. Ainsi, on peut appliquer une CAH sur les groupes obtenus et définir une classification à  $K$  classes. Le nombre de classe étant désormais fixé, on exécute l'algorithme d'agrégation autour des centres mobiles pour le nombre de classes choisi en initialisant l'algorithme via les centres des classes établies par la CAH. De cette façon, on obtient une partition à  $K$  classes d'inertie intra-classe plus faible que celle donnée par la CAH. Cette dernière étape s'appelle *la consolidation*. On pourra noter que cette étape de consolidation fait perdre la hiérarchie précédemment établie.

# 5 Description des classes

Dans une optique exploratoire, il sera indispensable d'effectuer une description des classes obtenues à partir des variables. Notons qu'il peut s'agir de variables utilisées pour la classification (variables actives), ou de variables supplémentaires. L'objectif est ici de définir les variables caractérisantes de chacune des classes à partir de tests statistiques en distinguant le cas des variables quantitatives de celui des variables qualitatives. Ceci s'effectuera essentiellement en comparant les moyennes ou les pourcentages à l'intérieur des classes, à ceux obtenus sur l'ensemble des individus. Nous en donnons ici les éléments essentiels qu'il pourra être utile de compléter en consultant Lebart, Morineau, and Piron (2006), pp.291-294.

## 5.1 Variables quantitatives

Pour une variable  $X$  continue, on note  $\bar{X}$  sa moyenne,  $\bar{X}_k$  sa moyenne dans la classe  $k$ ,  $S_X^2$  sa variance empirique,  $S_{X_k}^2$  sa variance empirique dans la classe  $k$ . La valeur-test pour la classe  $k$  est

$$t_k = \frac{\bar{X}_k - \bar{X}}{S_{X_k}^2} \text{ avec } S_{X_k}^2 = \frac{n - n_k}{n - 1} \frac{S_X^2}{n_k}$$

On effectue ensuite un classement des variables actives en fonction de la valeur absolue des valeurs-test, permettant d'identifier les variables les plus caractérisantes d'une classe. Dans le cas de variables supplémentaires, la distribution de la statistique  $t_k$  est connue sous l'hypothèse que l'appartenance d'un individu à la classe est sans lien avec la variable considérée. Il s'agit d'une gaussienne standard. Ainsi une valeur supérieure à 2 sera considérée comme statistiquement significative et traduira un lien entre la variable quantitative et l'appartenance à la classe. Dans le cas d'une variable active, cette interprétation statistique n'a plus de sens car la variable a nécessairement servi à déterminer la classification. Néanmoins, elle fournira un moyen pratique de hiérarchiser les variables.

## 5.2 Variables qualitatives

Pour une variable  $X$  qualitative à  $Q$  modalités, on notera  $n_q$  le nombre d'individus à présenter la modalité  $q$  dans la population et  $n_{kq}$  ce nombre dans la classe  $K$ . Pour évaluer le ``degré de présence'' de la modalité  $q$  d'une variable qualitative dans une classe  $k$  d'individus, on compare  $\frac{n_{kq}}{n_k}$  et  $\frac{n_q}{n}$ . Les différences les plus significatives selon la valeur test

$$\frac{n_{kq} - n_k \frac{n_q}{n}}{\sqrt{n_k \frac{n-n_k}{n-1} \frac{n_q}{n} \left(1 - \frac{n_q}{n}\right)}}$$

permettent d'isoler les modalités les plus caractéristiques d'une classe. Cette statistique suit une loi normale standard pour un nombre d'individus suffisamment élevé, permettant de se référer à la valeur seuil 2 : une valeur supérieure indiquera que la modalité caractérise la classe, tandis qu'une valeur inférieure ne sera pas considérée comme statistiquement significative.

# 6 Données qualitatives

Une façon simple d'appliquer les méthodes de classification non-supervisées à des données qualitatives consiste à appliquer les approches classiquement utilisées pour des données quantitatives sur l'ensemble de toutes les composantes de l'ACM plutôt que sur les données brutes. En effet, d'une part les composantes permettent de transformer les variables d'origine en variables quantitatives, et d'autre part, en utilisant toutes les composantes pour effectuer la classification, on ne perd aucune information par rapport aux données brutes, car l'intégralité de l'inertie du nuage est portée par les axes utilisés pour la classification.

Une autre façon de procéder consiste à utiliser une distance adaptée à la nature des données dans les algorithmes de classification, en particulier, on pourra utiliser la distance du  $\chi^2$  qui est une distance Euclidienne (permettant donc d'appliquer l'agrégation autour des centres mobiles ou la CAH) :

$$d^2(i, i') = \frac{n}{p} \sum_{q=1}^Q \frac{(z_{iq} - z_{i'q})^2}{n_q}$$

où  $Z = (z_{iq})_{\substack{1 \leq i \leq n \\ 1 \leq q \leq Q}}$  est le tableau disjonctif codant pour le tableau de données qualitatives.

Notons que la distance du  $\chi^2$  affecte un poids important aux modalités rares et ne fait pas de distinction entre la présence simultanée d'un caractère sur deux individus, ou l'absence simultanée de ce caractère. Ce choix ne sera pas pertinent en fonction des applications. Par exemple, en phytosociologie on s'intéresse aux associations végétales, c'est-à-dire aux espèces présentes dans les mêmes milieux. On dira alors que deux espèces sont proches si elles sont présentes simultanément dans les mêmes stations (une station étant un site représentatif d'un milieu), l'absence simultanée des deux espèces n'ayant pas une réelle valeur. On définira alors la distance entre deux individus (espèces) selon la distance de Jaccard qui ne tient compte que de la présence simultanée des espèces dans une même station

$$1 - \frac{n_{++}}{n_{++} + n_{+-} + n_{-+}}$$

où  $n_{++}$  est le nombre de stations où les deux espèces sont présentes,  $n_{+-}$  est le nombre de stations où seule l'espèce  $i$  est présente et  $n_{-+}$  est le nombre de stations où seule l'espèce  $i'$  est présente. D'autres exemples de distances (Euclidienne) sont donnés en bas de la section.

Huang (1997) a proposé d'une version modifiée des  $k$ -means dédiées aux variables qualitatives et appelée *algorithme des k-modes*. Il a la spécificité de considérer comme centre d'une classe le vecteur des modalités les plus fréquentes au sein de la classe (noté  $M_k$ ).

## Algorithme des k-modes

1. Initialisation : Tirer au hasard  $k$  centres

2. Répéter :

- Allouer un individu au groupe dont il est le plus proche au sens de la distance du  $\chi^2$  ou  $d(i, i') = \sum_{q=1}^Q (z_{iq} - z_{i'q})^2$  (distance du  $\chi^2$  sans pondération selon l'effectif). Mettre à jour le mode du groupe après chaque allocation
- Re-tester la distance entre individus selon les différents modes : si un individu est tel que son mode le plus proche n'est pas celui de son groupe, alors le réallouer à l'autre groupe et mettre à jour les modes des deux groupes.

3. S'arrêter à stabilisation

Cet algorithme minimise localement

$$\sum_{k=1}^K \sum_{i \in C_k} d(i, M_k)$$

mais est plutôt instable du fait du choix particulier du centre d'une classe qui peut être très influencé par l'ajout d'un unique individu.

Quelques exemples de distances Euclidiennes

<i>Sokal, Michener</i>	$\frac{n_{++} + n_{--}}{n_{++} + n_{-+} + n_{+-}}$
<i>Russel - Rao</i>	$\frac{n_{++}}{n_{++} + n_{-+} + n_{+-}}$
<i>Ochiai</i>	$\frac{n_{++}}{\sqrt{(n_{++} + n_{-+})(n_{+-} + n_{--})}}$
<i>OchiaiII</i>	$\frac{n_{++} \times n_{--}}{\sqrt{(n_{++} + n_{--})(n_{++} + n_{+-})(n_{--} + n_{-+})(n_{--} + n_{+-})}}$
<i>Dice</i>	$\frac{n_{++}}{2n_{++} + n_{-+} + n_{+-}}$
<i>Rogers - Tanimoto</i>	$\frac{n_{++} + n_{--}}{n_{++} + 2(n_{-+} + n_{+-}) + n_{--}}$
<i>Kulzinsky</i>	$\frac{n_{++}}{n_{-+} + n_{+-}}$

## 7 Complémentarité de l'analyse factorielle et des méthodes de classification

Les méthodes d'analyse factorielle ont plusieurs limites auxquelles les méthodes de classification peuvent apporter des solutions.

Tout d'abord, il est généralement difficile d'interpréter les axes au-delà du plan principal. La raison est que les axes sont orthogonaux, par conséquent le plan 3-4 ne porte que de l'information qui n'est pas présente sur le plan principal (idem pour les plans suivants). Par ailleurs, la visualisation des premiers axes ne reflète pas l'intégralité de la structure des données car une partie de l'inertie est aussi portée par les derniers axes qui ne sont pas analysés. La classification, elle, est effectuée sur l'intégralité des données. Par conséquent, le positionnement des classes sur les plans factoriels (soit par un simple coloriage des individus, soit par l'ajout d'une variable qualitative supplémentaire indiquant la classe de chaque individu) va permettre de visualiser l'information portée par l'ensemble des axes et donc d'enrichir l'interprétation et corriger les déformations liées à l'opération de projection. Si deux points sur un plan sont proches et dans la même classe, alors la proximité entre ces individus est alors globalement observée sur les autres axes, tandis que si ils sont proches mais de classes différentes, l'information portée par le plan n'est pas suffisante pour résumer les ressemblances entre ces deux individus.

Aussi, ces méthodes d'analyse factorielle sont sensibles aux individus aberrants, ces individus influençant grandement la construction des axes. Au contraire, les méthodes de classification (hiérarchiques en particulier) sont peu sensibles localement aux points marginaux isolés : ces individus ne seront pas agrégés avec les autres dans la partie basse du dendrogramme.

Enfin, les graphiques deviennent rapidement inextricables en présence d'un grand nombre d'individus. La représentation des classes sur les plans va permettre de simplifier la description des ressemblances et oppositions entre individus par une donnée plus synthétique. En effet, les axes fournissent une information continue des différences entre individus (via leurs coordonnées sur les axes). En représentant par exemple le centre de gravité de chaque classe on obtient une information discontinue, plus simple, qui permet par ailleurs d'alléger les représentations graphiques.

## Références

- Ball, G., and D. Hall. 1965. "ISODATA: A Novel Method of Data Analysis and Pattern Classification." Menlo Park: Stanford Research Institute.
- Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set." *Journal of Statistical Software* 61 (6): 1–36. <http://www.jstatsoft.org/v61/i06/> (<http://www.jstatsoft.org/v61/i06/>).
- Diday, E. 1973. "The Dynamic Clusters Method in Nonhierarchical Clustering." *International Journal of Computer & Information Sciences* 2 (1): 61–88.
- Forgy, E. 1965. "Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications." *Biometrics* 21: 768–80.
- Huang, Zhixue. 1997. "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining." In *In Research Issues on Data Mining and Knowledge Discovery*, 1–8.
- Husson, F. 2016a. "Cours d'Analyse En Composantes Principales." [https://husson.github.io!\[\]\(858b895d2d81c819fd5bd63547df3684\_img.jpg\)/img/AnaDo\\_ACP\\_cours\\_slides.pdf](https://husson.github.io/img/AnaDo_ACP_cours_slides.pdf) ([https://husson.github.io/img/AnaDo\\_ACP\\_cours\\_slides.pdf](https://husson.github.io/img/AnaDo_ACP_cours_slides.pdf)).
- . 2016b. "Cours d'Analyse Factorielle Multiple (Cours Complet)." <https://www.youtube.com/watch?v=wCTaFaVKGAM> (<https://www.youtube.com/watch?v=wCTaFaVKGAM>).
- Lebart, L., A. Morineau, and M. Piron. 2006. *Statistique Exploratoire Multidimensionnelle: Visualisations et Inférences En Fouille de Données*. Sciences Sup. Mathématiques. Dunod.
- MacQueen, J. B. 1967. "Some Methods for Classification and Analysis of Multivariate Observations." In *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, edited by L. M. Le Cam and J. Neyman, 1:281–97. University of California Press.
- McQuitty, Louis L. 1966. "Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data." *Educational and Psychological Measurement* 26 (4): 825–31.
- Saporta, G. 2006. *Probabilités, Analyse Des Données et Statistique*. Editions Technip.
- Tufféry, S. 2007. *Data Mining et Statistique décisionnelle: L'intelligence Des Données*. Editions Technip.
- . 2015. *Modélisation Prédictive et Apprentissage Statistique Avec R*: Éditions Technip.
- Wikistat. 2016. "Statistique Descriptive Bimensionnelle — Wikistat." <http://wikistat.fr/pdf/st-l-des-bi.pdf> (<http://wikistat.fr/pdf/st-l-des-bi.pdf>).