

STA211 : Méthodes non-supervisées

Vincent Audigier, Ndeye Niang-Keita

15 juin, 2019

- 1 Introduction
- 2 Classification de variables
 - 2.1 Objectifs et contexte
 - 2.2 Classification hiérarchique
 - 2.2.1 Ascendante
 - 2.2.2 Descendante
 - 2.3 Classification par partitionnement direct
 - 2.4 Conclusion
- 3 Cartes de Kohonen
 - 3.1 Principe général
 - 3.2 Exemple
 - 3.3 Variantes
 - 3.3.1 Fonction de voisinage
 - 3.3.2 Version batch
 - 3.3.3 Données non-numériques
 - 3.4 En pratique
 - 3.4.1 Qualité de représentation
 - 3.4.2 Classification par CAH
 - 3.4.3 Choix de la grille
 - 3.4.4 Nombre d'itérations
 - 3.4.5 Packages R
 - 3.5 Conclusion
- 4 Règles d'association
 - 4.1 Quelques définitions et concepts de base
 - 4.1.1 Représentation et lecture des données
 - 4.1.2 Notions de support et de confiance
 - 4.2 Recherche de règles d'association
 - 4.2.1 Algorithme Apriori
 - 4.2.2 Quelques algorithmes alternatifs
 - 4.2.3 Indices de pertinence des règles
 - 4.3 Exemple
 - 4.3.1 Supermarché
 - 4.3.2 German Credit
 - 4.4 Conclusion
- 5 Analyse multi-blocs
 - 5.1 Objectifs de l'analyse multi-blocs
 - 5.2 Méthodes
 - 5.2.1 Notations
 - 5.2.2 Double ACP
 - 5.2.3 STATIS/STATIS duale
 - 5.2.4 AFM

- 5.3 Conclusion
- Références

1 Introduction

Parmi les méthodes non-supervisées, on retrouve essentiellement les méthodes de classification, les méthodes de recherche de règles d'association et les méthodes d'analyse factorielle (cf introduction (<https://par.moodle.lecnam.net/mod/resource/view.php?id=90955>)).

Les méthodes de classification consistent à identifier des groupes d'individus (ou de variables) homogènes. Nous nous focaliserons sur les méthodes basées sur des distances qui sont essentiellement basées sur des approches géométriques mais des méthodes basées sur des modèles existent également. On distingue les méthodes de partitionnement et les méthodes hiérarchiques. Les méthodes de partitionnement consistent à définir une partition de l'ensemble des individus (ou des variables) en K (défini à l'avance) groupes telle que la dispersion des individus (au sens de la distance préalablement définie) au sein (resp. à l'extérieur) des groupes soit la plus faible (resp. forte) possible. Elles regroupent les méthodes des centres mobiles et ses variantes (k-means, nuées dynamiques) et les cartes de Kohonen pour les méthodes les plus connues. Les méthodes hiérarchiques visent quant à elle à construire des classes imbriquées les unes dans les autres. On distingue les méthodes ascendantes, qui partent d'une partition très fine puis qui fusionnent les classes les plus voisines (au sens de la distance choisie) de proche en proche, des méthodes descendantes, qui partent, elles, d'une classe regroupant l'ensemble des individus et qui la partitionnent de façon itérative de façon à obtenir des classes plus fines. A l'exception des cartes de Kohonen, ces méthodes sont supposées connues du lecteur pour ce qui concerne la classification des individus. Ainsi, nous nous focaliserons sur les cartes de Kohonen, et présenterons la classification des variables qui, bien que similaire dans le principe à celle des individus, revêt des particularités qu'il convient de préciser.

Les méthodes de recherche de règles d'association, globalement moins répandues que les méthodes précédentes, sont très utilisées dans le domaine du marketing. Elles consistent à comparer, pour certains produits, les probabilités d'achats simultanés aux probabilités obtenues dans le cas où il n'y aurait pas de lien entre les achats des produits. Ceci permet d'identifier des associations importantes quand le volume des données ne permet pas une étude exhaustive de l'ensemble des associations.

Enfin, les méthodes d'analyse factorielle (<https://par.moodle.lecnam.net/mod/resource/view.php?id=93508>) sont des approches géométriques qui consistent à résumer les données sous la forme de graphiques de dimension deux. La visualisation de ces graphiques synthétiques permet alors d'identifier les ressemblances entre les individus, les relations entre les variables et de résumer des groupes d'individus qui se ressemblent par quelques variables caractéristiques. Les méthodes d'analyse factorielle sont complémentaires aux précédentes. En particulier, elles peuvent être employées comme méthode de pré-traitement pour se ramener à des variables quantitatives lorsque celles-ci sont qualitatives ou mixtes, pour réduire le nombre de variables ou encore obtenir une meilleure partition lors de l'application d'une méthode de classification des individus (cf *tandem clustering*). Les méthodes d'analyse factorielle les plus classiques (ACP, ACM) sont supposées connues. Ainsi, nous évoquerons uniquement les méthodes multi-blocs.

2 Classification de variables

2.1 Objectifs et contexte

Quand les variables sont nombreuses, il est parfois très utile de les regrouper en groupes homogènes. Ceci

peut servir à pour analyser la multicolinéarité entre les variables, ou dans une optique de pré-traitement à réduire leur nombre en leur substituant un petit nombre de variables latentes synthétiques, représentatives de chacune des classes. Comme pour la classification des individus, on distingue 3 stratégies de partitionnement : les méthodes hiérarchiques ascendantes, descendantes, et les méthodes de partitionnement direct.

2.2 Classification hiérarchique

2.2.1 Ascendante

Lors d'une classification ascendante hiérarchique de variables, les groupements se font par agglomération progressive des variables deux à deux. On réunit les variables les plus proches au sens d'une distance à déterminer (ou simplement d'une dissimilarité). Il faut aussi définir un critère d'agrégation afin de pouvoir par la suite définir la distance d'une variable ou d'une classe à un groupe déjà établi. Notons que le nombre d'opérations à effectuer pour parvenir à une classification par cette méthode évolue en p^3 où p désigne le nombre de variables à classer. Cela signifie que sur un nombre de variables important, cette méthode ne peut pas être utilisée. Cependant des algorithmes permettent d'abaisser la complexité en p^2 (voir e.g. G. Saporta (2006)).

La classification selon cette procédure repose donc sur le choix de l'indice de dissimilarité entre variables et du critère d'agrégation de deux groupes de variables. Les critères d'agrégation utilisés pour la classification de variables sont les mêmes que ceux utilisés pour la classification ascendante hiérarchique des individus, à savoir saut minimal, saut maximal, saut moyen ou méthode de Ward dans le cas où la dissimilarité est une distance euclidienne. On pourra par exemple consulter G. Saporta (2006) pour plus de précisions. Nous précisons ici simplement les indices de similarité (ou dissimilarité) les plus courants.

Lorsque les variables sont de nature quantitative, le coefficient de corrélation linéaire est l'indice naturel de similarité utilisé (l'indice de dissimilarité associé s'avère être par ailleurs une distance euclidienne sur les variables).

Lorsque les individus sont décrits par un ensemble de variables qualitatives, les indices les plus classiques sont le χ^2 (qui est une distance euclidienne) ou le V de Cramer. Il faut toutefois noter que la nature précise des liaisons entre les variables d'un même groupe sera généralement difficile à apprécier. Il pourra être préférable dans ce cas de se tourner vers la classification des modalités des variables, i.e., de considérer le tableau disjonctif associé et d'effectuer la classification sur les variables indicatrices de ce tableau.

Le cas spécifique de variables binaires (fréquent lors de l'étude d'associations), on trouve dans Fichet and le Calve (1984) plusieurs indices de similarité définis tous à partir du tableau de contingence des paires de variables ($X_j, X_{j'}$) :

- Concordances simples (Sokal, Michener)

$$\frac{N_{11} + N_{00}}{N_{11} + N_{01} + N_{10}}$$

- Jaccard

$$\frac{N_{11}}{N_{11} + N_{01} + N_{10} + N_{00}}$$

- Russel-Rao

$$\frac{N_{11}}{N_{11} + N_{01} + N_{10}}$$

- Ochiai

$$\frac{N_{11}}{\sqrt{(N_{11} + N_{01})(N_{10} + N_{00})}}$$

- Ochiai II

$$\frac{N_{11} \times N_{00}}{\sqrt{(N_{11} + N_{00})(N_{11} + N_{10})(N_{00} + N_{01})(N_{00} + N_{10})}}$$

- Dice

$$\frac{N_{11}}{2N_{11} + N_{01} + N_{10}}$$

- Rogers-Tanimoto

$$\frac{N_{11} + N_{00}}{N_{11} + 2(N_{01} + N_{10}) + N_{00}}$$

- Kulzinsky

$$\frac{N_{11}}{N_{01} + N_{10}}$$

où

- N_{11} désigne le nombre de caractéristiques communes aux individus i , ni i' ;
- N_{10} désigne le nombre de caractéristiques possédées par i et pas par i' ;
- N_{01} désigne le nombre de caractéristiques possédées par i' et pas par i ;
- N_{00} désigne le nombre de caractéristiques que ne possèdent ni i , ni i' ;

Une connaissance a priori des données pourrait guider le choix de l'une ou l'autre des mesures. Chacune de ces mesures de similarité possède ses propres propriétés qui influencent les résultats de la classification. On notera que les mesures de dissimilarité associées sont toutes euclidiennes (Fichet and le Calve (1984)).

Une fois les différentes partitions à K classes, $K - 1$ classes, ... établies, on détermine le niveau de coupure de l'arbre, à partir de l'évolution du gain d'inertie inter-classes pour la méthode de Ward.

2.2.2 Descendante

Les méthodes de classification descendante procèdent par dichotomies successives à la construction d'un arbre hiérarchique descendant dont les segments terminaux constituent une partition des éléments à classer. La partition obtenue est telle que les éléments d'une même classe sont les plus ressemblants possible et deux éléments appartenant à des classes différentes sont les moins ressemblants possible. Un des intérêts majeurs des méthodes descendantes est que l'interprétation des classes obtenues est bien plus aisée car une classe est décrite par les règles de divisions successives.

Une des techniques de classification de variables couramment utilisée est la méthode VARCLUS de SAS Sarle (1990). Elle consiste à :

- réaliser une analyse en composantes principales des variables. Les deux premières composantes factorielles associées aux deux plus grandes valeurs propres si la seconde est supérieure à 1 sont retenues.
- affecter chaque variable à la composante principale avec laquelle elle est le plus corrélée.
- répéter les étapes précédentes sur les groupes obtenus tant que la seconde valeur propre est supérieure à 1.

2.3 Classification par partitionnement direct

Parmi les méthodes de partitionnement direct, la méthode la plus connue est la méthode CLV proposée par Vigneau and Qannari (2003) destinée aux variables de nature quantitative. C'est une approche similaire aux K-moyennes, consistant à rechercher une partition en K classes des variables en maximisant un critère exprimant la colinéarité entre les variables d'une classe :

$$n \sum_{k=1}^K \delta_{k,j} cov^2(v_j, u_k)$$

sous la contrainte $u_k u_{k'} = 1$ où $\delta_{k,j} = 1$ si la variable j est dans la classe k et 0 sinon, et $cov^2(v_j, u_k)$ est le carré de la covariance entre la variable v_j et la variable latente u_k représentant la classe k .

La solution du critère peut être obtenue en utilisant un algorithme de partitionnement alternant deux étapes : l'étape d'affectation des variables aux différents classes et l'étape d'estimation des nouveaux centroïdes des classes.

Après initialisation des classes, la variable latente u_k pour chaque classe est la première composante principale de l'ACP la matrice dont les variables sont restreintes aux variables de la classe k . Cette composante représente le nouveau centroïde de la classe k . Puis, dans un processus itératif, une variable v_j est affectée à la classe qui maximise le carré de sa covariance avec la variable latente u_k . Ici les groupes de variables constitués sont corrélés positivement ou négativement. Il existe une autre version de la méthode qui tient compte du sens de la liaison de façon à ne pas regrouper des variables opposées.

Cette méthode a été généralisée par Chavent et al. (2012) aux variables exclusivement qualitatives ou mixtes en utilisant les composantes principales de l'ACM (pour des données qualitatives) ou de l'AFDM (pour des données mixtes). La méthode proposée porte le nom ClustOfVar.

On notera que des versions ascendantes de ces approches existent également.

2.4 Conclusion

La classification de variables qualitatives vise à regrouper les variables en classes homogènes : les variables dans un même groupe sont fortement liées entre elles, les variables dans des classes différentes sont faiblement liées. Comme pour la classification des individus, on peut procéder par des méthodes de partitionnement hiérarchiques ou directes.

Il est fréquent de vouloir effectuer à la fois de la classification des individus et des variables sur un même jeu de données. Dès lors, on peut se demander si certains groupes d'individus ne seraient pas caractérisés par un seul des groupes de variables plutôt que par l'intégralité de celles-ci. Ceci est fréquent dans certains domaines comme la bioinformatique, le textmining, le webmining ou encore le marketing : des clients peuvent avoir des habitudes de consommation complètement différentes par rapport à certains produits, par exemple, une personne ayant un nourrisson achètera fréquemment des couches et du lait maternel, tandis

qu'une personne sans enfant n'en achètera pas. Ces clients sont donc complètement opposés par rapport à ces produits (variables). Pour autant s'ils sont tous les deux amateurs de cuisine asiatique ils achèteront fréquemment des pâtes de riz et de la sauce soja. En effectuant uniquement de la classification des consommateurs, ces deux individus ont peu de chances d'être dans une même classe, alors qu'ils ont pourtant certaines habitudes de consommation communes. Pour prendre en compte cet aspect, on peut alors procéder à du *biclustering* (aussi appelé *co-clustering* ou *classification croisée*). Le principe est de constituer simultanément des groupes d'individus et de variables tels que chaque groupe d'individus ne soit construit qu'à partir des ressemblances vis-à-vis de certaines variables d'un même groupe. Il existe différentes façons d'effectuer cette classification (voir Charrad and Ben Ahmed (2011)).

3 Cartes de Kohonen

Les cartes de Kohonen (aussi appelées cartes auto-organisatrices ou Self-Organizing Maps en anglais) sont une autre méthode de classification non-supervisées. Elles permettent de former des groupes d'individus homogènes selon une méthode assez similaire à la méthode des k -moyennes. La spécificité des cartes de Kohonen est que les centres des classes (appelés prototypes ou vecteurs référents) sont ``encouragés'' à vivre dans un sous-espace de l'espace des individus. Ainsi, les cartes de Kohonen peuvent être vues comme une méthode des k -moyennes contraintes. L'intérêt par rapport aux k -moyennes est que cette approche permet d'effectuer de la classification non-supervisée en grande dimension.

3.1 Principe général

Comme pour les k -moyennes, il existe deux versions de l'algorithme des cartes de Kohonen : une version *on-line* dans laquelle les vecteurs référents sont mis à jour à chaque prise en compte d'un nouvel individu (tiré au hasard sans remise), et une version *batch*, où les vecteurs référents sont mis à jour uniquement après avoir passé en revue l'intégralité des individus (de façon déterministe en regardant le premier, puis le second,...). Nous présentons dans cette section la version on-line qui correspond à la version originelle de l'algorithme SOM.

On considère n individus dans un espace V de dimension p (inclu dans \mathbb{R}^p). K vecteurs référents $(w_k)_{1 \leq k \leq K}$ sont d'abord disposés sur un plan de cet espace. A chacun de ces vecteurs référents dans V , on associe un représentant dans une grille A , i.e. un couple de coordonnées $s_k \in \mathcal{S}_1 \times \mathcal{S}_2$ avec $\mathcal{S}_1 = \{1, 2, \dots, s_1^{\max}\}$ $\mathcal{S}_2 = \{1, 2, \dots, s_2^{\max}\}$. Cette grille est généralement rectangulaire mais d'autres types, telle que les grilles hexagonales, peuvent être considérés. Pour chaque individu x_i pris au hasard dans V , on cherche dans un premier temps le vecteur référent w_k le plus proche au sens d'une certaine distance dans \mathbb{R}^p (généralement euclidienne). On déplace alors tous les vecteurs référents les plus proches de w_k en direction de x_i selon un pas Δw_k (voir Figure 3.1). La notion de voisinage entre vecteurs référents dépend d'une distance dans l'espace des coordonnées de la grille. Deux vecteurs référents sont voisins, si leurs couples de coordonnées dans la grille sont proches (et non pas s'ils sont proches dans l'espace des individus V). Toute la spécificité de la méthode repose sur ce dernier point : les vecteurs référents sont ainsi "attirés" par les données, mais contraints à rester rattachés aux autres via une grille en deux dimensions.

Plus précisément, les vecteurs référents sont déplacés selon

$$w_k \leftarrow w_k + \underbrace{\alpha(x_i - w_k)}_{\Delta w_k}$$

où α est un scalaire appelé *learning rate*. Si α est petit (proche de 0), les voisins sont peu déplacés, et si α est

grand (proche de 1), alors le déplacement est important. Par ailleurs, pour définir à partir de quelle distance deux vecteurs référents ne sont plus voisins, la version la plus simple de l'algorithme SOM consiste à utiliser la distance Euclidienne et à se donner un seuil r déterminant si un vecteur référent est voisin ou non.

Lors de la mise en oeuvre de l'algorithme, les valeurs de α et r sont diminuées progressivement de telle sorte qu'à la fin de l'algorithme, $\alpha = 0$ ce qui implique que les vecteurs référents ne sont plus déplacés, et $r = 1$, ce qui signifie que le voisinage est réduit à un seul vecteur référent. Plus précisément, on se donne d'abord une valeur minimale et maximale pour r et pour α . Généralement, α et r évoluent selon un pas régulier entre ces bornes. Le nombre de pas pour α correspond au nombre d'individus tandis que le nombre de pas pour r est généralement de l'ordre de plusieurs milliers. Pour un r fixé, on parcourt les différentes valeurs de α et pour chacune d'entre elle on tire un individu unique sans remise et on met à jour les vecteurs référents. Une fois tous les individus tirés (et donc toutes les valeurs de α considérées) on passe à la valeur suivante de r et ainsi de suite.

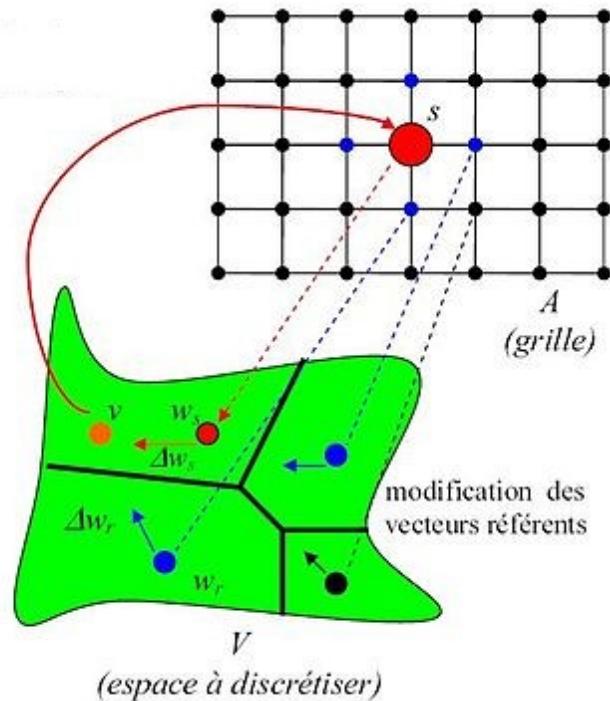


Figure 3.1: Schéma de l'algorithme des cartes de Kohonen. Source : Wikipédia

Remarque : les cartes de Kohonen peuvent également vues comme un réseau de neurones où les neurones sont les éléments de la grille.

3.2 Exemple

On applique la méthode sur les données iris de Fisher. Ce jeu de données contient les mesures en centimètres des longueurs et largeurs des pétales et sépales pour 150 fleurs appartenant à 3 espèces d'iris (setosa, versicolor et virginica). Les quatre premières variables (continues) sont utilisées pour construire la carte.

On utilise une carte rectangulaire à 5×5 noeuds. L'algorithme est itéré 750 fois. Les données sont réparties en 25 classes de la façon suivante :

Table 3.1: Effectifs des classes. Seuls les effectifs des classes non vides sont reportés

| Indice de classe | 5 | 6 | 7 | 10 | 11 | 12 | 15 | 16 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
|------------------|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|
| | | | | | | | | | | | | | | | | |

| effectif | 48 | 4 | 1 | 1 | 6 | 1 | 1 | 3 | 5 | 5 | 2 | 23 | 5 | 10 | 15 | 20 |
|----------|----|---|---|---|---|---|---|---|---|---|---|----|---|----|----|----|
|----------|----|---|---|---|---|---|---|---|---|---|---|----|---|----|----|----|

On voit que certains prototypes n'ont aucun individu rattaché (1, 2, 3, 4, 8, 9, 13, 14 et 17), tandis que d'autres ne sont rattachés qu'à un seul individu (7, 10, 12 et 15). Ceci n'est pas surprenant, car le choix du nombre de noeuds a été choisi a priori, et n'est pas problématique dans un premier objectif de visualisation des structures de groupes.

La carte obtenue est présentée en Figure 3.2.

Profil des noeuds

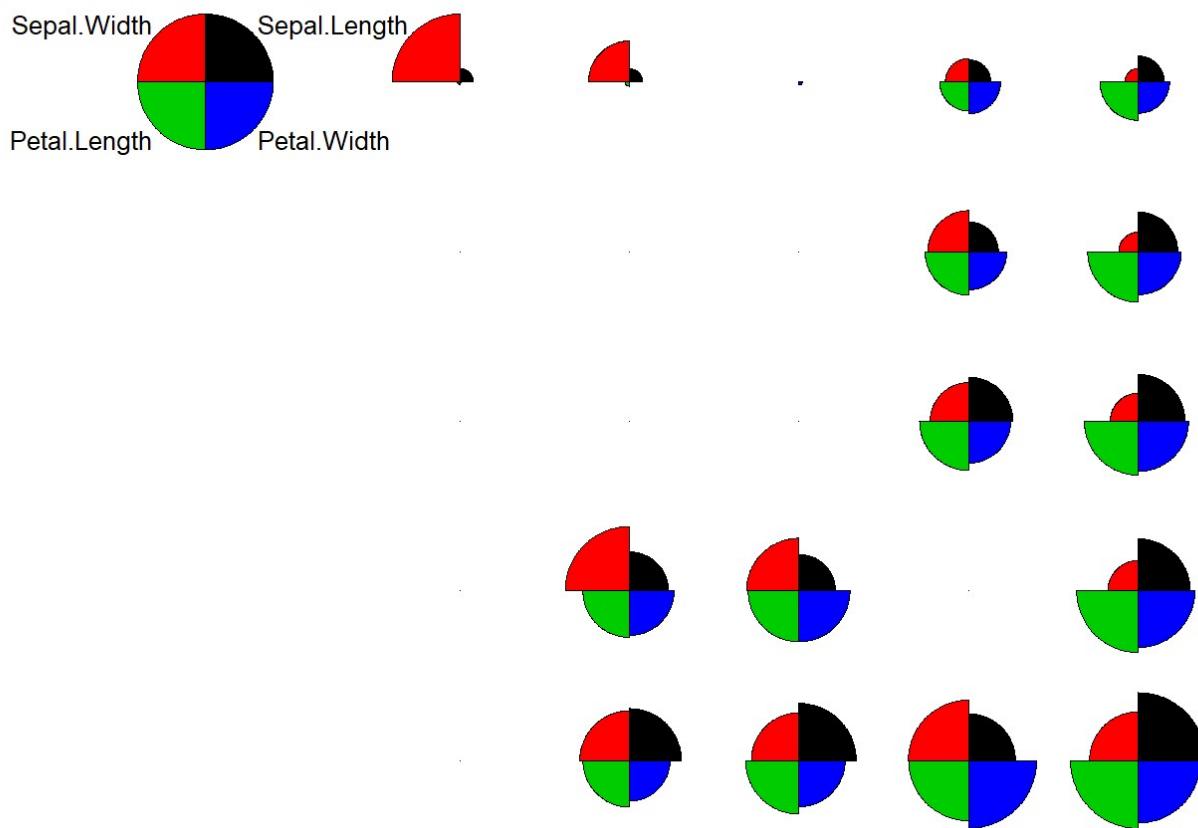


Figure 3.2: Carte de Kohonen pour les données iris.

Dans cette figure, chaque radar correspond à une classe (les groupes vides ne sont pas représentés). Chaque cadran du radar correspond à une variable. Plus la valeur moyenne d'une variable est élevée dans un groupe (relativement à la moyenne globale), plus le cadran correspondant est visible. Par exemple, dans la classe 5 (représentée en haut à gauche), la largeur du sépale est particulièrement élevée. Ce type de représentation est une façon de visualiser la carte. Il est clair qu'elle est rapidement limitée quand le nombre de variables est grand. Dans ce cas, on pourra par exemple procéder préalablement à une analyse de la variance afin d'identifier les variables discriminantes, puis à ne représenter que ces variables de façon à alléger les graphiques.

On peut également regarder les distances entre prototypes (Figure 3.3), ce qui permet d'apprécier les proximités entre classes.

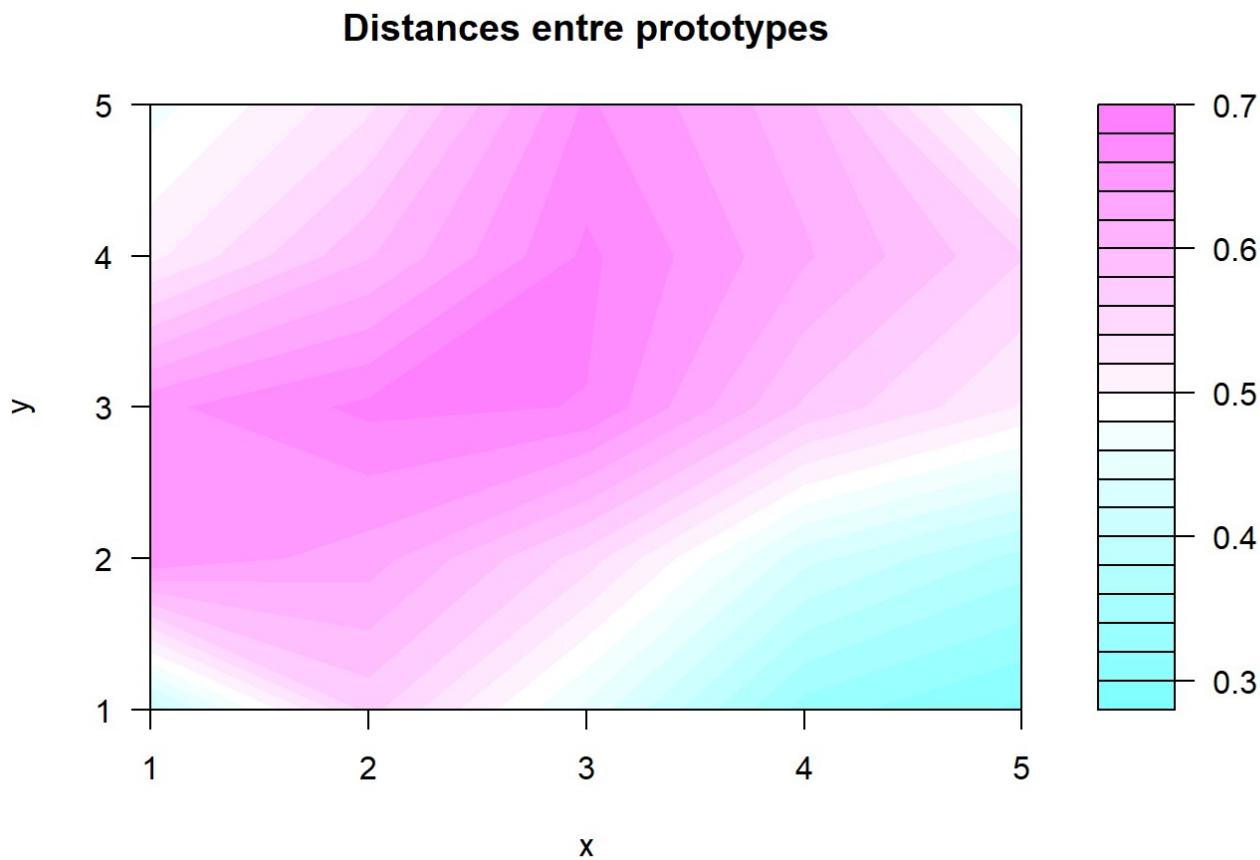


Figure 3.3: Distance entre prototypes : distance moyenne entre un prototype et ses voisins. Les coordonnées d'un point correspondent au couple de coordonnées associé dans la grille. Les valeurs intermédiaires sont obtenues par lissage.

Sur les données iris, on repère qu'il y des écarts importants sur la diagonale. Ceci traduit par exemple que les fleurs affectées au groupe de coordonnées (3,5) (groupe 15) sont plutôt différentes de celles du groupe de coordonnées (4,5) (groupe 20), bien que les deux groupes soient adjacents sur la grille. En revanche, les groupes de coordonnées (4,1), (5,1) et (5,2) (groupes 16, 21 et 22) sont eux assez proches.

NB : la numérotation des groupes est ici effectuée à partir du coin inférieur gauche, puis en remontant colonne par colonne de la façon suivante :

Table 3.2: Schéma de la numérotation des clusters sur la carte

| | | | | |
|---|----|----|----|----|
| 5 | 10 | 15 | 20 | 25 |
| 4 | 9 | 14 | 19 | 24 |
| 3 | 8 | 13 | 18 | 23 |
| 2 | 7 | 12 | 17 | 22 |
| 1 | 6 | 11 | 16 | 21 |

3.3 Variantes

3.3.1 Fonction de voisinage

La méthode précédente est la version la plus simple des cartes de Kohonen. Une méthode plus sophistiquée (et très utilisée) consiste à déplacer les vecteurs référents selon

$$w_k \leftarrow w_k + \alpha h(\| s_k - s_l \|)(x_i - w_k)$$

où s_l désigne le couple de coordonnées associé au vecteur référent le plus proche de x_i et h une *fonction de voisinage* donnant plus de poids aux vecteurs référents dont les représentants sur la grille sont les plus proches de s_l . La fonction h est un noyau, c'est-à-dire une fonction symétrique, positive, d'intégrale égale à 1. Les choix les plus classiques pour h sont le noyau Gaussien, ou le noyau uniforme (cf Figure 3.4) dont les expressions sont celles des fonctions de densité des lois du même nom. Ces noyaux sont paramétrables, ce qui permet de régler la taille du voisinage (e.g. en modifiant la variance pour le noyau Gaussien).

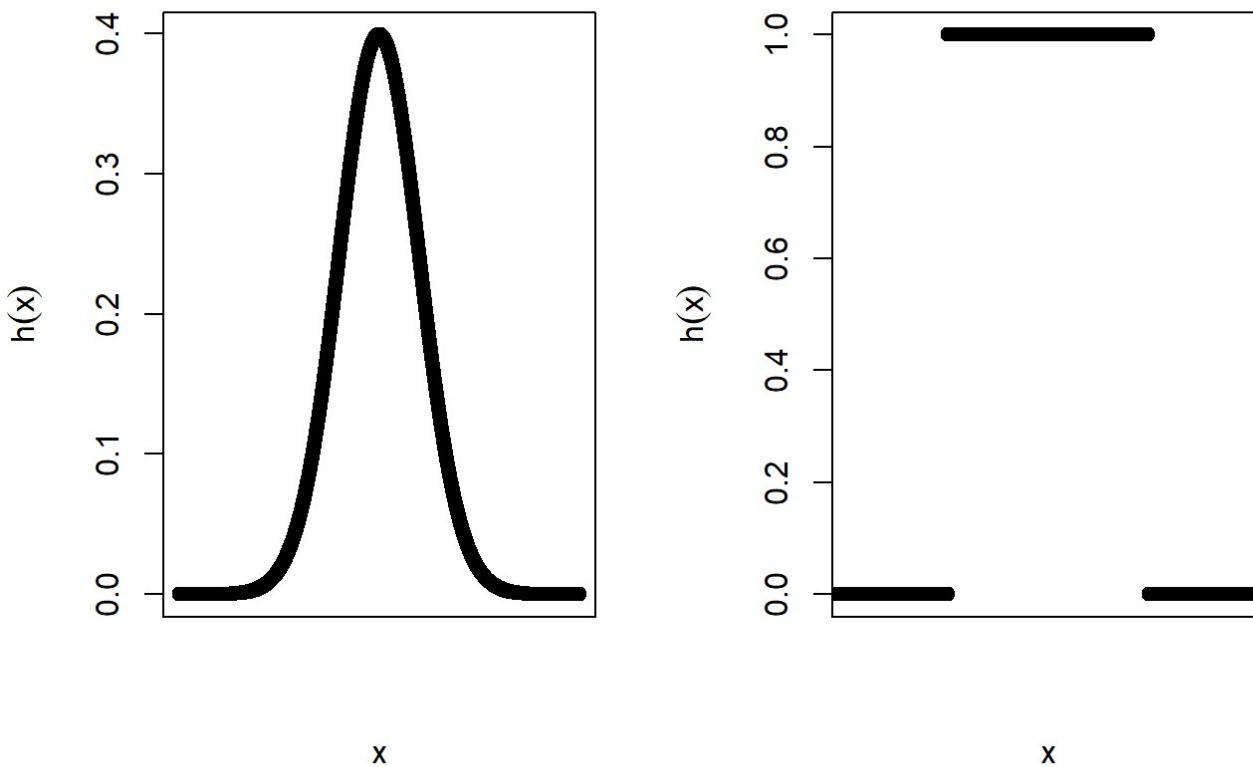


Figure 3.4: Noyau Gaussien et noyau uniforme

On notera que le noyau uniforme correspond à la version simple de l'algorithme présentée précédemment.

3.3.2 Version batch

La version présentée dans la section 3.1 est une version on-line où les vecteurs référents sont mis à jour à chaque nouvel individu. La version batch consiste elle à mettre à jour les vecteurs référents identifiant dans un premier temps tous les individus rattachés à un vecteur référent, puis à mettre à jour les vecteurs référents selon le barycentre des individus qui lui sont rattachés. Les poids des individus sont définis selon la fonction de voisinage h (Hastie, Tibshirani, and Friedman (2009)).

La version batch permet d'accélérer l'algorithme, tout en donnant des résultats similaires. Elle est néanmoins très sensible à l'initialisation. Contrairement à la version on-line, elle est déterministe, ce qui implique que pour des initialisations identiques l'algorithme donnera toujours la même carte.

3.3.3 Données non-numériques

Des méthodes ont également été proposées pour le cas où les données ne sont pas toutes numériques. L'algorithme de Kohonen a été étendu d'abord au cas de données qualitatives en particulier binaires dans le cadre de l'algorithme Binbatch (Lebbah, Badran, and Thiria (2000)) par exemple et à travers les variantes de l'algorithme Kohonen Multiple Correspondence Analysis (KMCA) (Cottrell, Ibbou, and Letrémy (2004)). Ensuite, pour prendre en compte les données mixtes, Lebbah et al. (2005) et Chen and Marques (2005) ont proposé simultanément une version mixte quasi-identique de l'algorithme SOM. Les auteurs utilisent une extension de la distance euclidienne classique combinant une partie continue regroupant l'ensemble des variables quantitatives et une partie constituée de variables qualitatives mises sous forme disjonctive au préalable. Dans l'algorithme NCSOM proposé par Chen and Marques (2005), les auteurs combinent simplement la distance euclidienne classique et la distance de Hamming pour données binaires. Lebbah et al. (2005) propose d'utiliser dans l'algorithme MTM (Mixed Topological Map) un paramètre d'ajustement de l'influence de la partie quantitative des données par rapport à la partie qualitative dans l'évaluation de la similarité entre les observations.

3.4 En pratique

3.4.1 Qualité de représentation

Les résultats de l'algorithme SOM dépendent des paramètres d'initialisation pouvant engendrer ainsi une instabilité. Il est habituel d'effectuer plusieurs initialisations puis de retenir comme carte finale la meilleure au sens d'un indice interne de qualité tel que l'erreur de quantification, topologique ou l'erreur de distorsion.

Definition 3.1 (Erreur de quantification)

$$\frac{1}{n} \sum_{i=1}^n \| w_{k(x_i)} - x_i \|^2$$

où $k(x_i)$ désigne la classe de x_i . Cette erreur est la mesure classique d'homogénéité utilisée en classification correspondant au coût à remplacer les données par le représentant de la classe auxquelles elles sont rattachées.

Definition 3.2 (Erreur topologique) L'erreur topologique correspond à la proportion des données pour lesquelles les 2 référents les plus proches ne correspondent pas à des unités adjacentes sur la carte.

Definition 3.3 (Mesure de distortion)

$$distorsion = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K h(\| s_k - s_{k(x_i)} \|) \| w_k - x_i \|^2$$

Cette erreur est celle qui est minimisée (au moins localement) par l'algorithme de Kohonen. Elle combine un critère de qualité du clustering (i.e. l'homogénéité des classes, mesurée par l'erreur de quantification) et d'organisation correcte (c'est-à-dire que les distances entre neurones reflètent celles des individus associés, ce qui est mesuré par l'erreur topologique)

3.4.2 Classification par CAH

L'auto-organisation par les cartes de Kohonen, ne permet pas de résoudre directement un problème de classification, elle permet simplement d'affecter une observation à un sous-ensemble d'une partition, représenté par un référent. Il peut être utile de se servir de cette quantification vectorielle comme point de départ d'un clustering.

Puisque chaque sous-ensemble de données est associé à un référent, le problème de classification se résume à celui de l'étiquetage de chaque vecteur référent à l'une des classes. Il faut donc introduire une seconde étape consistant à étiqueter tous les vecteurs référents. On peut par exemple procéder par classification ascendante hiérarchique.

3.4.3 Choix de la grille

Pour ce qui est du choix du nombre de noeuds sur la grille, celui-ci varie généralement entre une douzaine et une centaine. Il se détermine par essais-erreurs en regardant le résultat pour un nombre de noeuds fixé et en ajustant au besoin, mais dans une optique de classification, celui-ci n'est pas aussi essentiel que dans la méthode des k -moyennes car la carte ne sera qu'une étape préalable à la CAH.

Par ailleurs, puisque l'algorithme SOM est une version contrainte des k -moyennes, il est important de vérifier que ces contraintes sont raisonnables pour les données considérées. Pour cela, on peut comparer l'erreur de quantification à celle obtenue en procédant par K-moyennes (Hastie, Tibshirani, and Friedman (2009)). On s'attend nécessairement à observer une erreur de quantification plus grande pour l'algorithme SOM.

Néanmoins, un écart trop important serait suspect et amènerait à éventuellement modifier la forme de la grille ou la fonction de voisinage.

3.4.4 Nombre d'itérations

Le nombre d'itérations de l'algorithme permet d'assurer sa convergence. Il est généralement choisi en regardant l'évolution de la distorsion en fonction du nombre d'itérations effectuées.

3.4.5 Packages R

Il existe différents packages R permettant de mettre de construire des cartes topologiques. On pourra notamment utiliser *SOMbrero*, *som*, ou *kohonen*. On donne ici quelques commandes utiles pour analyser le jeu de données iris via le package *SOMBREO*.

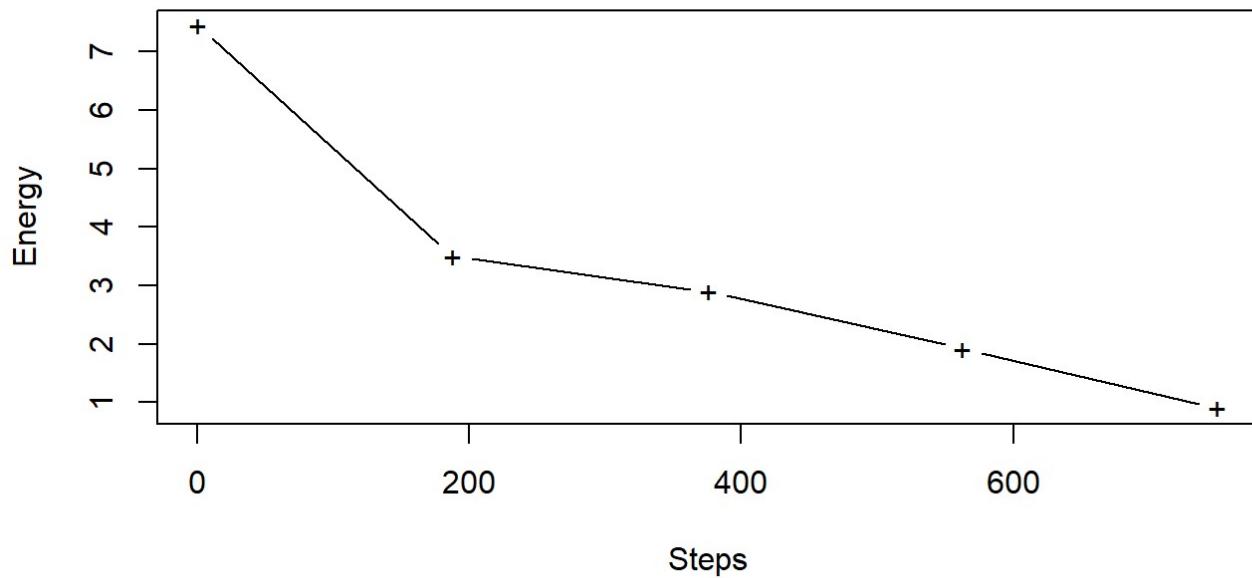
Pour construire une carte de kohonen, on pourra utiliser

```
library(SOMBREO)
set.seed(255)
iris.som <- trainSOM(x.data=iris[,1:4], verbose=TRUE, nb.save=10)
```

Pour vérifier que le nombre d'itérations est suffisant, on représente la décroissance de la distorsion en fonction du nombre d'itérations.

```
plot(iris.som, what="energy")
```

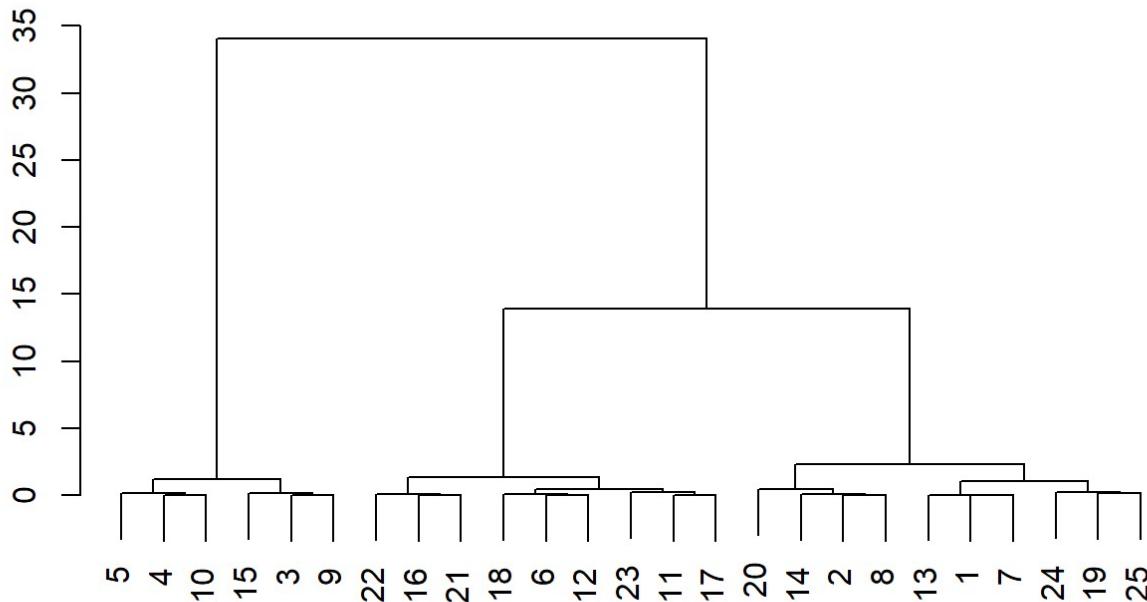
Energy evolution



Pour construire la classification ascendante hiérarchique selon la méthode de Ward à partir de la carte, on utilisera

```
plot(superClass(iris.som))
```

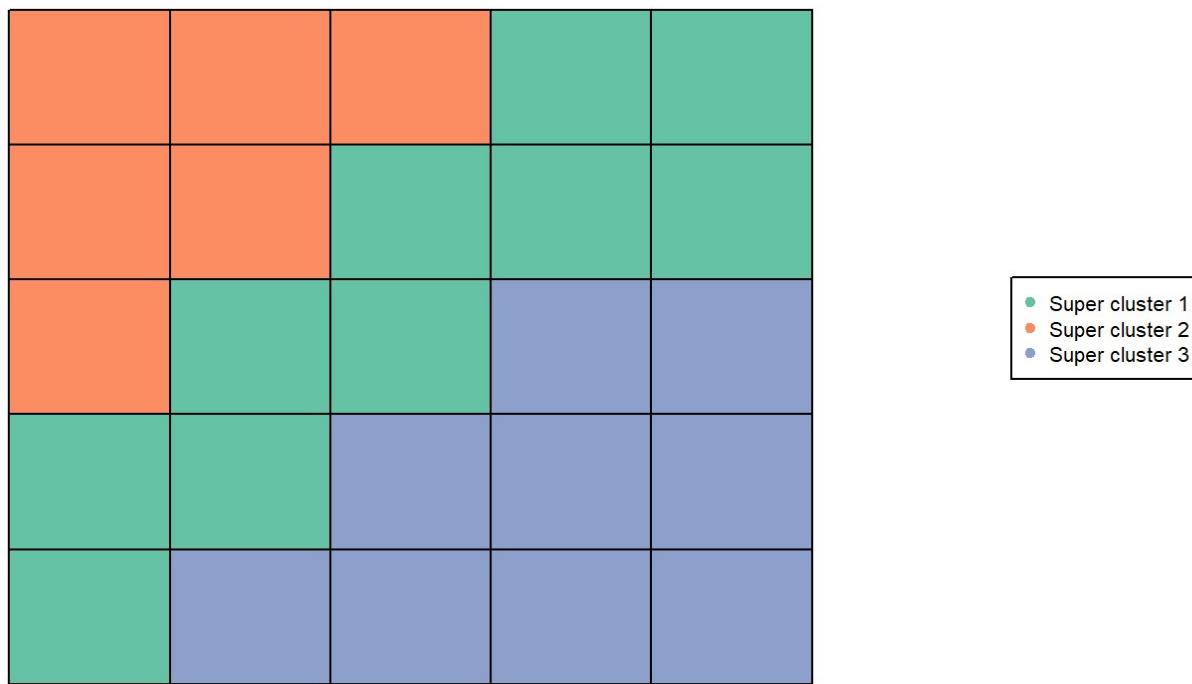
Super-clusters dendrogram



En coupant l'arbre à 3 classes (ce qui correspond par ailleurs au nombre d'espèces d'iris du jeu de données),

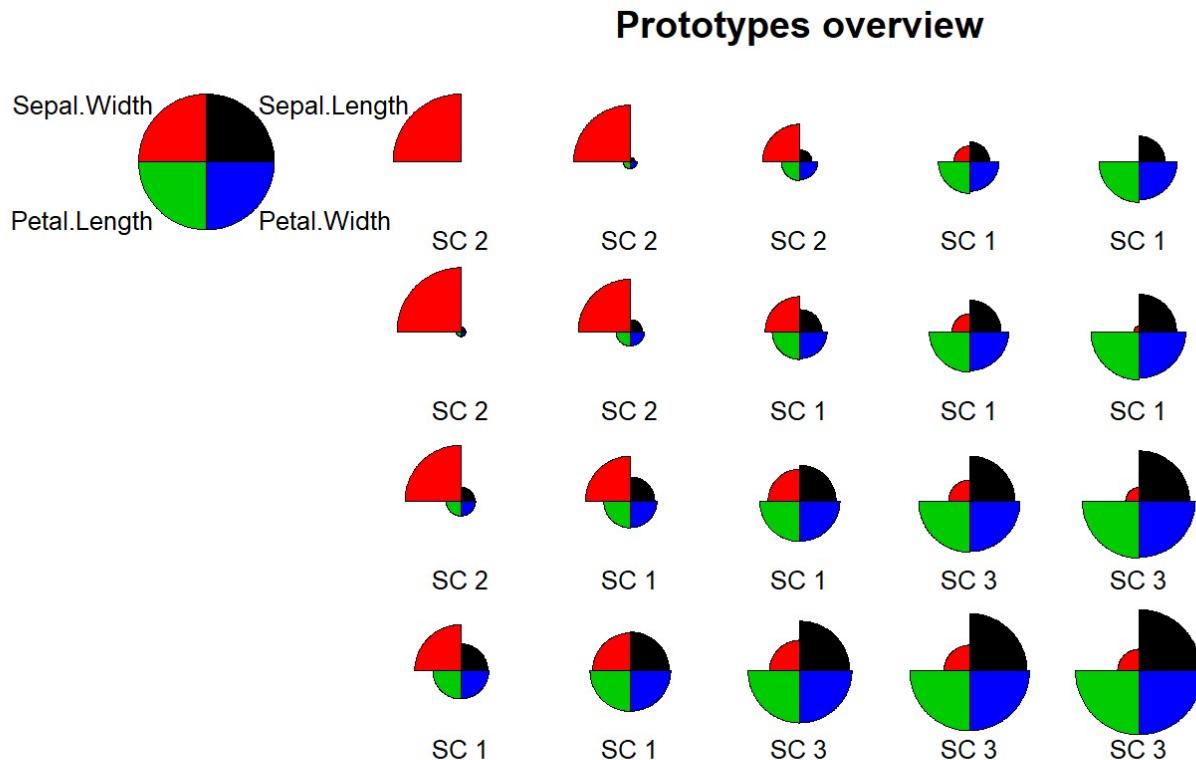
on obtient la classification suivante

```
my.sc <- superClass(iris.som, k=3)
plot(my.sc, type="grid", plot.legend=TRUE)
```



La description des classes peut notamment s'appuyer sur le regroupement des 'radars' par classe

```
plot(my.sc, type="radar", key.loc=c(-0.5,5), mar=c(0,10,2,0))
```





Les outils graphiques pour analyser les cartes sont riches, nous en avons fourni que les principaux. Pour aller plus loin on pourra par exemple consulter les vignettes (<https://cran.r-project.org/web/packages/SOMbrero/vignettes/doc-numericSOM.html>) du package *SOMbrero* ou le tutoriel (https://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Kohonen_SOM_R.pdf) de R. Rakotomalala utilisant le package *kohonen*.

3.5 Conclusion

Les cartes topologiques peuvent être considérées comme une version régularisée des K-moyennes. Il est aussi possible de les présenter comme des réseaux de neurones particulier où chaque noeud de la grille correspond à un neurone. Les cartes ont la propriété de conservation topologique, de sorte que la proximité des neurones sur la carte traduit une proximité dans l'espace initial des données. On montre qu'elles sont plus adaptées à la découverte de classes à structure non-sphérique et à la présence d'observations aberrantes. Elles offrent de plus, des possibilités de visualisation directe (sans passer par un plan factoriel d'ACP par exemple) utiles pour l'interprétation des résultats.

Les cartes constituent également une méthode de classification automatique, mais elles permettent surtout d'effectuer un prétraitement avant d'effectuer une CAH. En effet, contrairement à la CAH, la complexité de l'algorithme SOM est de l'ordre du nombre de données ce qui permet de l'utiliser sur des données où le nombre d'individus ou de variables est grand.

4 Règles d'association

Les méthodes de recherche de règles d'association ont été proposées pour découvrir quels produits étaient achetés conjointement et à quelle date dans les bases de données des ventes de supermarché (Agrawal, Imieliński, and Swami (1993)). Elles permettent d'extraire des règles de type: "lorsqu'un client achète du pain et du beurre alors 9 fois sur 10 il achète en même temps du lait". En effet, la technologie des codes-barres permet l'acquisition rapide et automatique des données relatives aux ventes. Trouver des ensembles d'articles achetés simultanément permet alors de mener des actions marketing pertinentes. Cependant, même si initialement la méthode de recherche de règles d'association a été développée dans un objectif marketing, elle peut être appliquée dans d'autres domaines si la structure des données s'y prête. Elle peut trouver des applications notamment en "webmining" dans l'analyse de la consultation des pages web sur un site Internet, ou encore en "text mining" dans la découverte de cooccurrences de termes dans les documents.

Cette section s'appuiera sur la thèse de Marie Plasse (Plasse (2006)) à laquelle le lecteur pourra se référer au besoin pour des approfondissements.

4.1 Quelques définitions et concepts de base

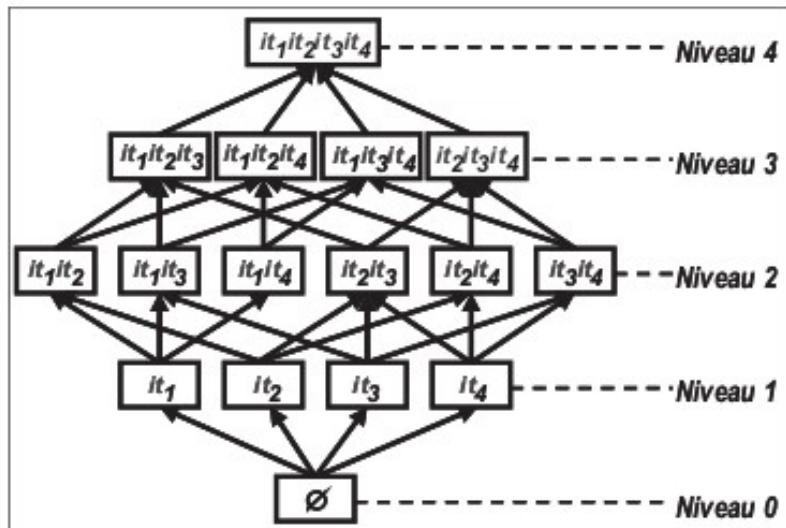
On définit l'ensemble $I = \{it_1, it_2, \dots, it_i, \dots, it_p\}$ dont les éléments sont appelés items. I peut par exemple correspondre à l'ensemble des produits d'un supermarché. On appelle alors transaction, un sous-ensemble non-vide de I . Dans l'exemple précédent, une transaction correspond au contenu d'un caddie. On note $T = \{tr_1, tr_2, \dots, tr_j, \dots, tr_n\}$ l'ensemble à n éléments des transactions.

Une règle d'association est une implication de la forme $A \rightarrow B$ où $A \subset I, B \subset I$ et $A \cap B = \emptyset$. Une règle

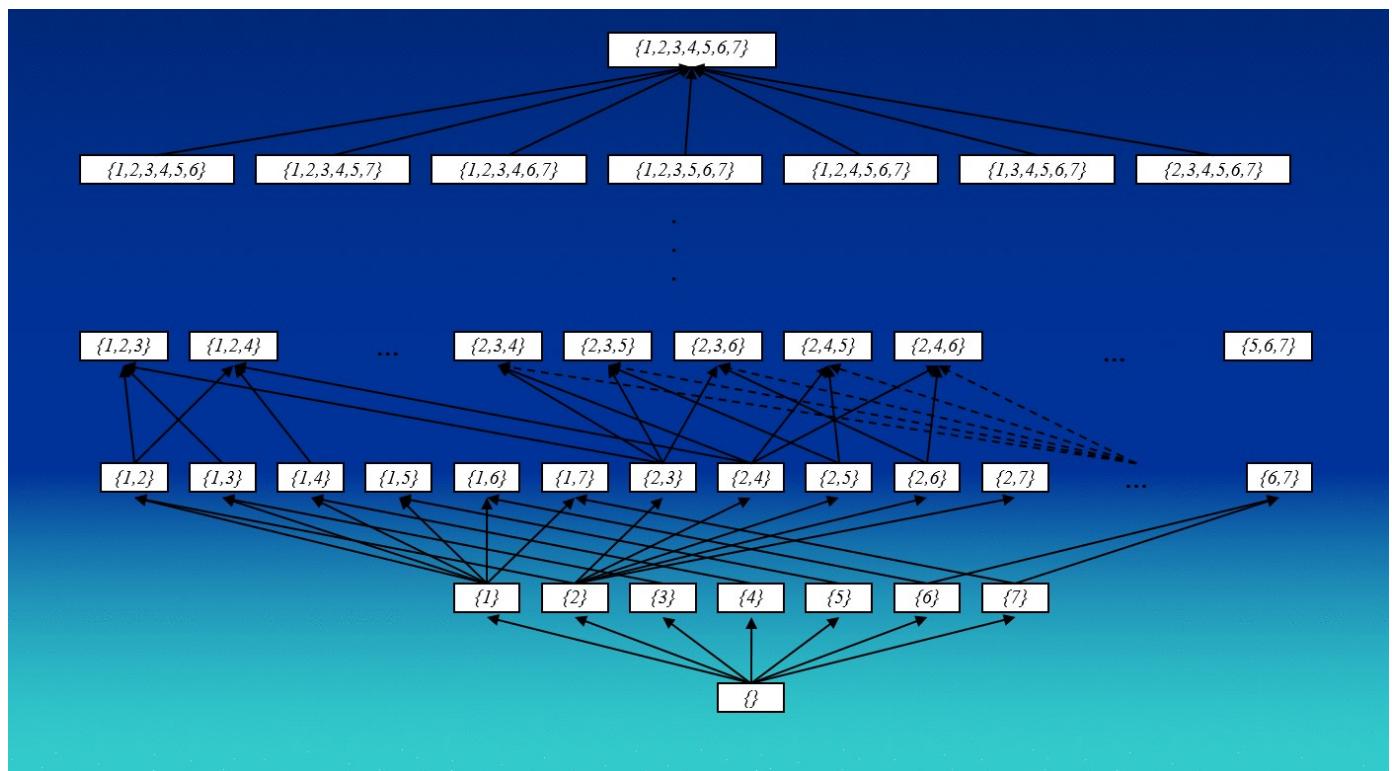
comporte donc une partie prémissé (ou antécédent) composée d'un ensemble d'items A et une partie conclusion (ou conséquent) composée d'un ensemble d'items B disjoint de A . Les ensembles d'items sont appelés itemsets.

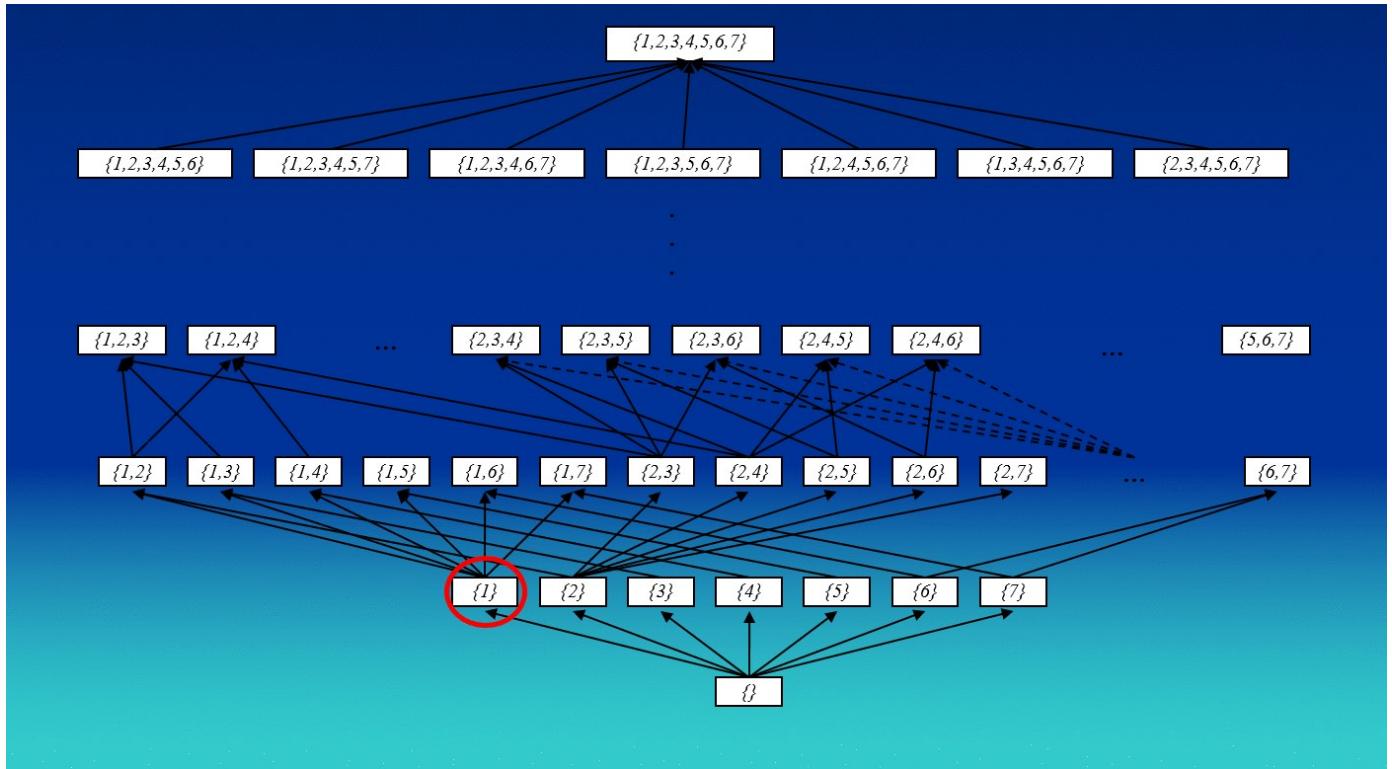
4.1.1 Représentation et lecture des données

Tous les itemsets des transactions peuvent être représentés par un treillis d'items formé de l'ensemble des parties de l'ensemble I muni de l'opération d'inclusion. La figure ci-dessous montre un exemple de treillis associé à un ensemble de 4 items.



Le treillis peut alors être lu en largeur ou en profondeur. Un parcours en largeur du treillis débute avec la lecture des itemsets de taille 1 puis continue avec la lecture des itemsets de taille 2 et ainsi de suite. Par contre, un parcours en profondeur s'intéresse d'abord à tous les itemsets commençant par le même sous-ensemble. Une fois arrivé à la profondeur maximale, la lecture revient à la racine et recommence avec un nouveau sous-ensemble (voir Figure animée ??).





(#fig:figtrellis,) Mode de lecture du treillis : en largeur en haut, en profondeur en bas

Par ailleurs, les données peuvent être représentées sous la forme d'une matrice de données binaires de type présence absence où les transactions sont en ligne et les items en colonne. Une telle matrice peut être lue de plusieurs façons (cf Figure 4.1):

| | | Lecture verticale | | | | |
|---------------------|---------------------|-------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| | | <i>Items</i> | <i>it₁</i> | <i>it₂</i> | <i>it₃</i> | <i>it₄</i> |
| Lecture horizontale | <i>Transactions</i> | | | | | |
| | tr1 | | 1 | 1 | 1 | 0 |
| | tr2 | | 1 | 1 | 0 | 0 |
| | tr3 | | 0 | 1 | 0 | 1 |
| | tr4 | | 0 | 0 | 1 | 1 |

Transaction = 200 : *fit₁, it₂*

Couverture (it₂) = {tr1, tr2, tr3}

Figure 4.1: Représentation verticale et horizontale

- Lorsque la lecture est basée sur une représentation horizontale des données, on considère les transactions et à chacune d'elles correspond une liste d'items associés.
- Lorsque la lecture utilise une représentation verticale des données, on considère les items et la base de données est vue comme un ensemble d'items; à chacun d'eux correspond une liste de transactions associées. On appelle *couverture* d'un item l'ensemble des transactions qui le contiennent.

Les algorithmes de recherche de règles d'association diffèrent selon la manière dont ils parcourent le treillis et lisent (horizontalement ou verticalement) les données.

4.1.2 Notions de support et de confiance

Une règle d'association est attachée à deux informations indiquant sa pertinence : son support *s* et sa confiance *c*. Du point de vue ensembliste, le support correspond à la proportion de transactions contenant

les items de $A \cup B$ (le nombre de transactions contenant à la fois tous les items de A et tous les items de B) par rapport à l'ensemble des transactions :

$$s = supp(A \rightarrow B) = \frac{card(t \in T / A \cup B \subseteq t)}{card(T)}$$

La confiance indique la proportion de transactions contenant les items de $A \cup B$ par rapport aux transactions contenant les items de A

$$c = conf(A \rightarrow B) = \frac{card(t \in T / A \cup B \subseteq t)}{card(t \in T / A \subseteq t)} = \frac{supp(A \rightarrow B)}{supp(A)}$$

4.2 Recherche de règles d'association

La recherche de règles d'association se décompose en deux sous problèmes. Le premier est la recherche des itemsets fréquents dont le support est supérieur ou égal à un seuil minimum, noté $minsup$. A partir de ces sous ensembles fréquents, le second problème est l'extraction des règles d'association dont la confiance est supérieure ou égale à un second seuil minimum, noté $minconf$. Les deux seuils, $minsup$ et $minconf$, sont spécifiés au préalable par l'utilisateur.

Dans cette section, nous présentons des algorithmes permettant de résoudre ces deux problèmes.

4.2.1 Algorithme Apriori

4.2.1.1 Recherche des itemsets fréquents

L'algorithme fondateur pour la recherche de règles d'association, Apriori (Agrawal, Srikant, and others (1994)), repose sur la propriété selon laquelle *tout sous-ensemble d'un ensemble fréquent est nécessairement fréquent*.

Cette propriété permet en effet de réduire considérablement le nombre d'itemsets candidats générés, c'est à dire les itemsets potentiellement fréquents imposant un calcul du support. Les ensembles candidats sont générés *a priori* - d'où le nom de l'algorithme - et leur support est calculé à l'étape suivante.

L'algorithme Apriori utilise une représentation horizontale des données et un parcours en largeur du treillis des itemsets. Le principe est de rechercher de manière itérative les ensembles fréquents de cardinal k à partir des ensembles fréquents de cardinal $k - 1$, déterminés lors de l'itération précédente. Ainsi, à la lecture k , l'algorithme calcule le support des itemsets de taille k . Pour cela, le tableau des transactions est parcouru transaction par transaction et un compteur est tenu à jour pour chacun des itemsets de taille k . A la fin de ce parcours, seuls sont conservés les itemsets dont le support est supérieur ou égal à $minsup$. A partir de ces k -itemsets fréquents, l'algorithme génère les ensembles candidats de taille $k + 1$ dont il calculera le support à la lecture $k + 1$. Lorsqu'il n'y a plus d'ensembles candidats, la procédure s'arrête (cf illustration ci-dessous)

Algorithme Apriori : etape 1

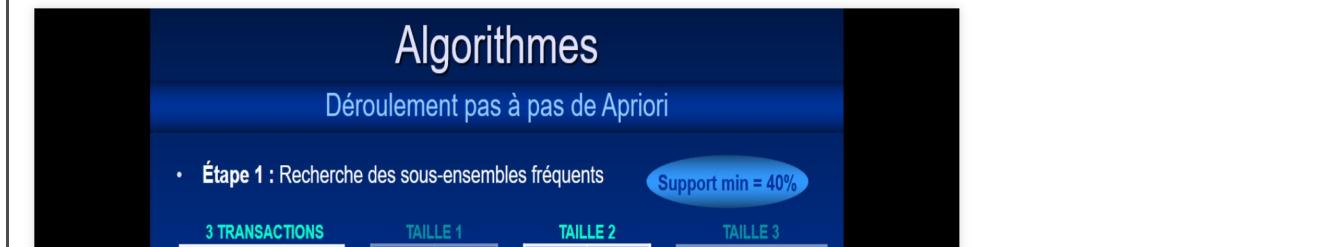


4.2.1.2 Extraction des règles

A partir des itemsets fréquents, l'étape suivante est l'extraction des règles d'association auxquelles est accordée une confiance au moins égale au seuil minimum $minconf$ également fixé *a priori*. La méthode proposée par Agrawal, Srikant, and others (1994) repose également sur la propriété selon laquelle tout sous-ensemble d'un ensemble fréquent est nécessairement fréquent.

Pour un itemset fréquent donné, l'algorithme génère toutes les règles ayant un conséquent de taille 1. Il calcule les confiances de ces règles et les compare à $minconf$ pour sélectionner les règles à extraire. Ensuite, il génère les conséquents de taille 2 en utilisant seulement les conséquents de taille 1 des règles sélectionnées à l'étape précédente. Cette procédure continue jusqu'à ce qu'il n'y ait plus de conséquent à générer pour l'itemset. La procédure est illustrée ci-dessous.

Algorithme Apriori : etape 2



En limitant le nombre d'ensembles candidats, l'algorithme Apriori économise des calculs inutiles pour le support, point déterminant dans le temps de calcul global. De plus, la base de données n'a pas besoin de tenir en mémoire car il est possible de lire le fichier transaction par transaction.

4.2.2 Quelques algorithmes alternatifs

L'algorithme Apriori a ensuite connu plusieurs extensions de ses fonctionnalités et performances. La recherche des itemsets fréquents est le point crucial de la recherche de règles d'association. De nombreux algorithmes ont, en effet, été proposés pour répondre à cette problématique. Quelques algorithmes de recherche d'itemsets fréquents proposés après Apriori sont présentés ci-après. Ces algorithmes visent principalement à limiter le nombre d'accès aux données et à accélérer les temps de calcul. Ils sont regroupés selon leur façon de représenter les données.

4.2.2.1 Représentation horizontale des données

4.2.2.1.1 Algorithme Sampling

L'algorithme Sampling (Toivonen and others (1996)) parcourt le treillis des itemsets en largeur (tout comme Apriori). Le principe de cet algorithme est de tirer un échantillon aléatoire de transactions permettant de chercher les itemsets fréquents. Cet échantillon étant de taille plus modeste que le nombre de transactions initial, il faut alors définir un seuil $minsup' < minsup$ par rapport auquel sont déterminés les itemsets fréquents. La vérification des résultats et le calcul exact du support des itemsets fréquents (par rapport au seuil initial $minsup$) sont ensuite réalisés avec le reste des transactions de la base.

4.2.2.1.2 Algorithme DIC (Dynamic Itemset Counting)

Comme Apriori et Sampling, l'algorithme Dynamic Itemset Counting (DIC) parcourt le treillis des items en largeur. Il procède par niveaux dans le treillis des itemsets : dès qu'un itemset de niveau k s'avère fréquent, on examine tous les itemsets de taille $k + 1$ générés par celui-ci (Brin et al. (1997)).

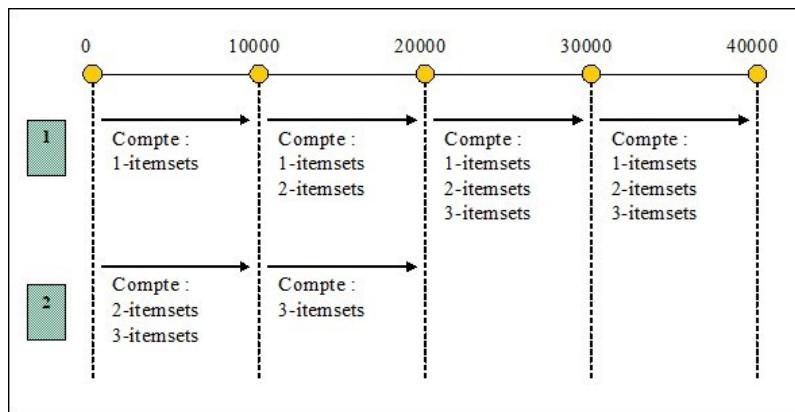


Figure 4.2: Schéma de fonctionnement de l'algorithme DIC

L'algorithme génère d'abord tous les itemsets de taille 1. Le support de ces itemsets candidats est calculé sur le premier intervalle de la base de données. A la fin du premier intervalle, les itemsets de taille 2 sont générés à partir des itemsets fréquents de taille 1. Le support de ces itemsets candidats de taille 2 commence alors à être calculé dès le second intervalle. Le procédé se répète jusqu'à l'obtention de tous les supports exacts de chaque itemset fréquent. Cet algorithme permet de réduire le nombre d'itemsets candidats et donc le nombre de calculs du support. Le nombre de lectures de la base est inférieur ou égal à celui d'Apriori, par contre il est nettement supérieur à celui de Sampling.

4.2.2.2 Représentation verticale des données

Cette représentation a pour avantage de permettre un calcul rapide du support d'un itemset donné en effectuant une simple opération d'intersection entre les couvertures des items concernés. Ainsi les algorithmes suivants sont plus rapides que Apriori.

4.2.2.2.1 Algorithme Partition

Comme Apriori, l'algorithme Partition (Savasere, Omiecinski, and Navathe (1995)) parcourt le treillis des items en largeur. Cet algorithme divise la base en parties disjointes de telle sorte que chacune, ainsi que les itemsets candidats qui seront générés, tiennent en mémoire. Ensuite, l'algorithme recherche les itemsets fréquents de chaque partie et détermine ainsi des itemsets fréquents locaux. Ces derniers sont alors réunis pour former l'ensemble des itemsets candidats globaux. Enfin, le support exact de tous les itemsets candidats globaux est calculé dans une seconde lecture de la base de données et les itemsets fréquents globaux sont ainsi obtenus.

Le partitionnement de la base soulève cependant plusieurs problèmes. Il est nécessaire de trouver un compromis entre la taille et le nombre des partitions. Elles doivent tenir en mémoire sans pour autant être trop nombreuses afin de garantir le principe de l'algorithme selon lequel "un itemset fréquent sur l'ensemble de la base doit être fréquent dans au moins une de ses partitions". Par ailleurs, dans l'idéal, chaque partition doit être représentative de la base entière pour éviter le risque de ne pas détecter un itemset fréquent.

4.2.2.2.2 Algorithme Eclat

L'algorithme Eclat "Equivalence CLASse Transformation" (Zaki (2000)) ne requiert que de faibles besoins en mémoire. Il parcourt le treillis des items en profondeur et le décompose en sous parties qui sont résolues de manière indépendante en mémoire. Cette décomposition du treillis peut se faire en fonction du préfixe des ensembles candidats.

Eclat parcourt les k -itemsets ayant pour préfixe commun l'itemset a de taille $k - 1$. Il calcule la liste des transactions associées à chacun de ces itemsets. Il détermine ainsi les k -itemsets fréquents commençant par a et les place dans une nouvelle base de données $D[a]$. Ensuite, il s'agit d'une procédure récursive :

l'algorithme recherche les itemsets fréquents commençant par b (avec a inclus dans b).

Le fait de parcourir en profondeur d'abord le treillis des ensembles conduit à calculer le support d'itemsets supplémentaires par rapport à Apriori qui ne calcule que les supports nécessaires. Mais Eclat parvient quand même à être plus rapide qu'Apriori, grâce à la représentation verticale des données. Enfin, Eclat réalise une meilleure gestion de la mémoire que l'algorithme Partition qui utilise également une lecture verticale des données.

4.2.2.3 Bilan

En termes de résultats, les nombreux algorithmes pour la recherche des ensembles fréquents aboutissent tous aux mêmes ensembles fréquents puisque la recherche des ensembles fréquents est déterministe, tout comme la recherche de règles d'association. Cependant, tous ne sont pas équivalents en termes de performances sur les temps de calcul et la gestion de la mémoire. L'efficacité d'un algorithme dépend essentiellement de trois facteurs : la façon dont les itemsets candidats sont générés, les structures de données adoptées et l'implémentation de l'algorithme. Au vu de ces critères, Eclat est l'algorithme le plus rapide.

4.2.3 Indices de pertinence des règles

L'approche support-confiance a l'avantage d'être simple. En effet, le support et la confiance sont des concepts facilement compréhensibles et ils ont un sens concret. Cependant cette approche montre plusieurs faiblesses.

Bien souvent, l'approche support-confiance précédemment décrite conduit à l'obtention de règles en trop grand nombre. Par conséquent, il est impossible de les faire valider par un expert. Dès lors, il est utile de les trier par ordre décroissant de leur intérêt au sens d'un indice de pertinence, tel que le lift (Brin et al., 1997), pour citer un des plus connus.

4.2.3.1 Faiblesse de l'approche support-confiance

Le choix de la valeur de *minsup* est difficile et est laissé à l'utilisateur. Un seuil élevé pour le support risque de disqualifier certaines règles très intéressantes, ayant un support faible mais une confiance très élevée. Par exemple, l'item "caviar" étant rare, il ne sera pas retenu alors qu'il peut cacher une règle du type "caviar → vodka" valable dans 85% des cas. D'un point de vue marketing, il est dommage de passer à côté d'une telle règle car le caviar est un produit de luxe. Sachant que l'achat de l'un entraîne de manière quasi sûre l'achat de l'autre, il pourrait être pertinent, par exemple, de faire une promotion sur la vodka lors de l'achat de caviar.

A l'inverse, fixer un seuil très bas pour le support peut entraîner l'extraction de règles sans intérêt du type "caviar → lait". En effet, le lait est un produit très commun qui figure dans un nombre élevé de transactions. Il n'est donc pas surprenant de le trouver dans le caddie de tous les clients qui achètent du caviar. Dans d'une application sur données avec des événements statistiquement rares, le support minimum doit être obligatoirement très faible. Par conséquent, les règles du type caviar → lait sont très nombreuses et doivent être sanctionnées.

Dans la règle "caviar → lait", le support mesure la fréquence globale des exemples de la règle ; il est donc faible étant donné la faible fréquence d'achat du produit caviar. La confiance est proche de 100% puisqu'elle s'intéresse à la répartition des achats de caviar entre les achats de lait et les "non achats" de lait. La répartition des "non achats" de caviar n'est pas du tout prise en compte alors qu'elle permettrait de discriminer ce type de règles. Ainsi l'approche support-confiance ne suffit pas pour discriminer les deux règles.

4.2.3.2 Indices de pertinence

La seule utilisation des seuils *minsup* et *minconf* présente donc quelques limites qui peuvent être levées grâce à l'utilisation d'autres critères pour mesurer l'intérêt statistique des règles. Ainsi, de nombreux indices de pertinence ont été présentés dans la littérature. Parmi eux, le lift (Brin et al., 1997) est un des plus connus et utilisés. Il est très facilement interprétable. C'est le rapport de la probabilité de trouver ensemble les items de l'antécédent et du conséquent sur la probabilité de les trouver ensemble alors qu'ils sont indépendants. Ainsi, les règles ayant un lift plus petit ou autour de 1 ne sont pas jugées intéressantes. Par contre, un lift égal à 2 montre que le nombre d'exemples de la règle $A \rightarrow C$ est deux fois plus grand que celui attendu sous l'indépendance.

Il est donc possible de classer les règles par ordre décroissant d'un indice de pertinence, tel que le lift, et de faire analyser par l'expert seulement les plus intéressantes au sens de cet indice. Cependant l'indice de pertinence optimal n'existe pas et le nombre élevé de tels indices est une difficulté supplémentaire pour l'utilisateur qui doit choisir le plus approprié à ses besoins. Parmi les travaux évaluant les indices de pertinence, nous pouvons citer ceux de Lallich et Teytaud (2004) qui proposent des critères d'évaluation. Aussi, Tan et al. (2002) ont réalisé une étude comparative d'une vingtaine d'indices symétriques, alors que Vaillant et al. (2004) s'intéressent aux indices dissymétriques.

Lenca et al. (2004) remarquent que certains indices sont équivalents, c'est à dire qu'ils classent les règles dans le même ordre. Par exemple, l'indice de Sebag-Schoenauer est une transformation monotone croissante de la confiance.

Afin d'évaluer les indices de pertinence, Lenca et al. (2004) définissent huit critères formels tels que :

1. le traitement non symétrique de A et de C : l'indice ne doit pas avoir la même valeur pour les règles $A \rightarrow C$ et $C \rightarrow A$
2. la décroissance avec le nombre d'occurrences de C dans la base de données (ce critère pénalise les règles du type caviar → lait).

Cependant, aucun indice de pertinence ne satisfait tous ces critères. L'utilisateur est le seul juge de l'importance accordée au respect de chaque critère.

Sur la base de ces critères, les auteurs ont effectué une étude comparative des différents indices et aboutissent à une typologie en trois classes (voir Figure 4.3)

| <i>Mesures</i> | <i>Définition</i> |
|---------------------------------------|---|
| CLASSE 1 | |
| Piatetksy-Shapiro | $nP(A)(P(C A)-P(C)) = nP(A)P(C)/\text{lift-} I)$ (avec n le nombre de transactions) |
| Indice de qualité de Cohen | $2 \frac{P(AC) - P(A)P(C)}{P(A) + P(C) - 2P(A)P(C)}$ |
| Gain Informationnel | $\log \frac{P(AC)}{P(A)P(C)} = \log(\text{lift})$ |
| Confiance centrée | $P(C A) - P(C)$ |
| Lift | $\frac{P(AC)}{P(A)P(C)}$ |
| Coefficient de corrélation de Pearson | $\frac{P(AC) - P(A)P(C)}{\sqrt{P(A)P(C)P(\bar{A})P(\bar{C})}}$ |
| Indice d'implication | $\sqrt{n} \frac{P(A\bar{C}) - P(A)P(\bar{C})}{\sqrt{P(A)P(\bar{C})}}$ |
| Indice probabiliste discriminant | $P[N(0,1) > \text{indice d'implication } CR/B]$ (La notation CR/B signifie que l'indice d'implication est préalablement centré et réduit sur une base d'exemples.) |
| CLASSE 2 | |
| Support | $P(AC)$ |
| Confiance | $P(C A)$ |
| Surprise | $\frac{P(AC) - P(A\bar{C})}{P(C)} = 2 \frac{P(A)}{P(C)}(\text{confiance} - 0.5)$ |
| Laplace | $\frac{P(C A) + n / P(A)}{1 + (2n) / P(A)}$ |
| Sebag et Schoenauer | $\frac{P(AC)}{P(A\bar{C})} = \frac{\text{confiance}}{1 - \text{confiance}}$ |
| Taux d'exemples et de contre exemples | $1 - \frac{P(A\bar{C})}{P(AC)} = 1 - \frac{1}{\text{Sebag et Schoenauer}}$ |
| CLASSE 3 | |
| Multiplicateur de cotes | $\frac{P(AC)P(\bar{C})}{P(A\bar{C})P(C)} = \text{lift . conviction}$ |
| Loevinger | $\frac{P(C A) - P(C)}{P(\bar{C})} = \frac{1}{P(\bar{C})} \text{confiance centrée} = 1 - \frac{1}{\text{conviction}}$ |
| Zhang | $\frac{P(AC) - P(A)P(C)}{\max \{P(AC)P(\bar{C}), P(C)(P(A) - P(AC))\}}$ |
| Conviction | $\frac{P(A)P(\bar{C})}{P(AC)}$ |

Figure 4.3: Typologie des indices de pertinences.

où chacun de ces indices est une fonction des huit éléments suivants



| | | | |
|------------------|---------------|---------------------|--------------|
| A | $P(AC)$ | $P(A\bar{C})$ | $P(A)$ |
| \bar{A} | $P(\bar{A}C)$ | $P(\bar{A}\bar{C})$ | $P(\bar{A})$ |
| Profils colonnes | $P(C)$ | $P(\bar{C})$ | 1 |

A titre d'exemple, considérons 100 consommateurs : 8 ont acheté du caviar, 40 ont acheté du lait et 7 ont acheté les deux en même temps. Cette règle qui n'a, en réalité, aucun intérêt va tout de même présenter des indices de pertinence élevés, comme le montre la figure 4.4

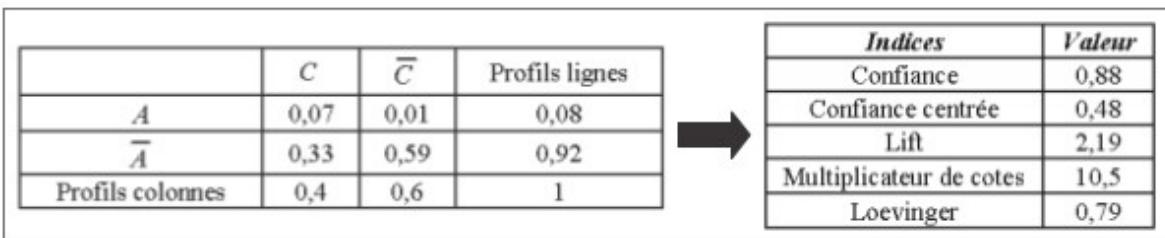


Figure 4.4: Exemple de règle sans intérêt aux indices de pertinence élevés.

Selon le lift, le nombre d'exemples de {caviar → lait} est deux fois plus grand que sous l'indépendance de {caviar} et {lait}

Il faudrait des indices de pertinence qui pénalisent mieux les règles où le conséquent est fréquent par rapport à l'antécédent. C'est le cas de l'indice accords-désaccords proposé par Kulczynski (en 1927)

$$IAD = \frac{P(AC)}{P(\bar{A}C) + P(A\bar{C})} = \frac{\text{accords positifs}}{\text{désacords}}$$

ou l'indice de Jaccard

$$Jaccard = \frac{P(AC)}{P(A) + P(C) - P(AC)} = \frac{P(AC)}{P(A \cup C)}$$

L'union représente un "ou" inclusif alors que la différence symétrique exprime un "ou" exclusif.

Les deux indices sont parfaitement équivalents dans la mesure où

$$\frac{1}{Jaccard} = \frac{1}{IAD} + 1$$

Ainsi, ils conduisent à un classement identique des règles d'association. L'indice de Jaccard présente l'intérêt d'être borné entre 0 et 1.

Sur l'exemple de la figure 4.4, ils se montrent plus sévères que les autres indices retenus tels que le lift, en effet : IAD=0.21 et Jaccard =0.17. Ce résultat n'est toutefois pas généralisable à toutes les applications.

Les meilleures performances de ces deux indices par rapport aux autres s'expliquent, en partie, par le fait que IAD et Jaccard présentent la propriété "cross-support" introduite par Xiong et al. en 2003. Les indices qui ont cette propriété permettent de discriminer les règles dont les supports de l'antécédent et du conséquent n'ont pas le même ordre de grandeur

Afin de comparer graphiquement les indices, M.Passe et al (conf egc) ont proposé une analyse graphique basée sur des courbes de niveaux et l'expression des indices en fonction des probabilités conditionnelles, notées $\lambda_A = P(C/A)$ et $\lambda_C = P(A/C)$ et de $P(C)$

| Indices | Définition |
|-------------------------------------|---|
| <i>Confiance centrée (CONFcen)</i> | $\lambda_A = CONFcen - P(C)$ |
| <i>Lift</i> | $\lambda_A = Lift \cdot P(C)$ |
| <i>Multiplicateur de cotes (MC)</i> | $\lambda_A = \frac{MC \cdot P(C)}{1 - P(C) + MC \cdot P(C)}$ |
| <i>Loevinger (LOE)</i> | $\lambda_A = LOE(1 - P(C)) + P(C)$ |
| <i>IAD</i> | $\lambda_A = \frac{I}{\frac{I}{IAD} - \frac{I}{\lambda_C} + 2}$ |
| <i>Jaccard (JAC)</i> | $\lambda_A = \frac{I}{\frac{I}{JAC} - \frac{I}{\lambda_C} + I}$ |

Ces nouvelles expressions permettent de tracer des courbes de niveaux (Figure 4.5). Chaque courbe correspond à un niveau de valeur de l'indice de pertinence. Les règles se situant sur une même courbe ne sont donc pas discriminées par l'indice puisqu'il a la même valeur.

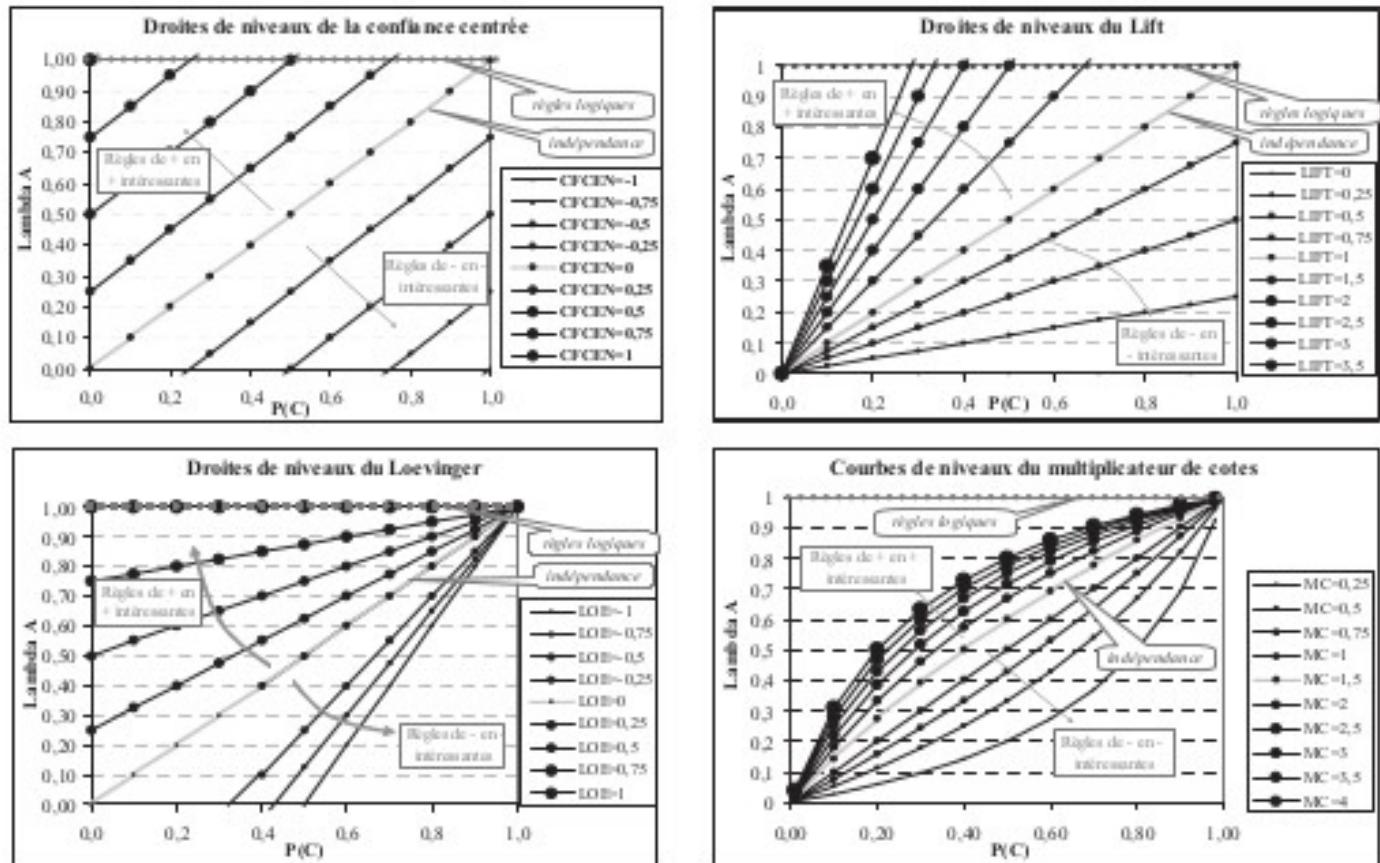


FIG. 1 – Courbes de niveaux en fonction de λ_A et de $P(C)$.

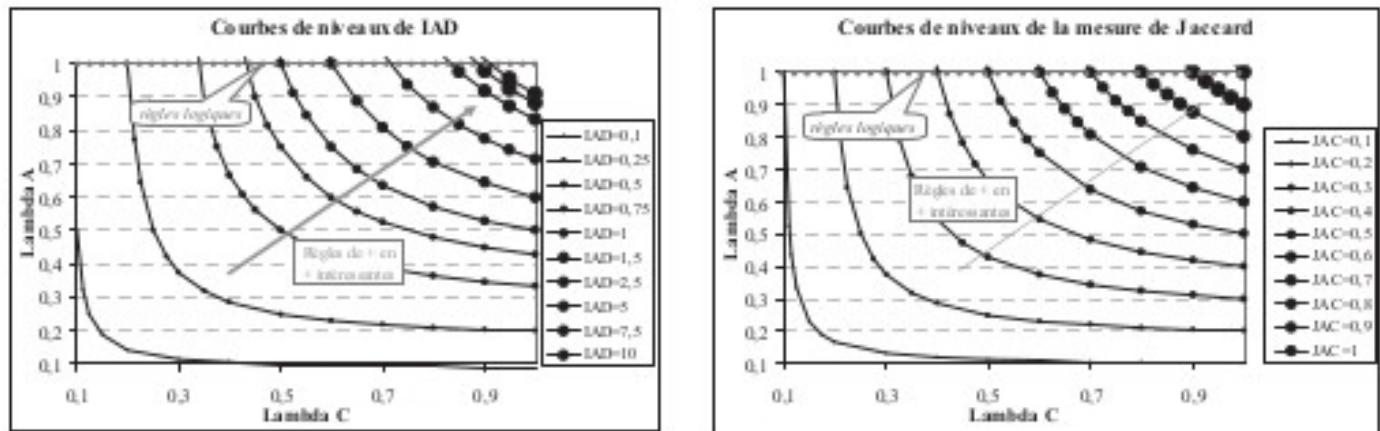


Figure 4.5: Exemples de courbes de niveau

4.3 Exemple

Le package R arules permet de rechercher des règles d'association selon les algorithmes Apriori ou Eclat. Nous proposons ici de l'utiliser afin d'illustrer la méthode de recherche d'associations sur un premier jeu de données ``jouet'' portant sur des achats de produits de consommation où les règles sont faciles à interpréter. Par la suite, nous rechercherons les règles d'association dans le jeu de données German Credit.

4.3.1 Supermarché

On commence par saisir les données et à les convertir dans un format adapté à l'application des algorithmes

de recherche de règles.

```
#charger le package arules  
library(arules)
```

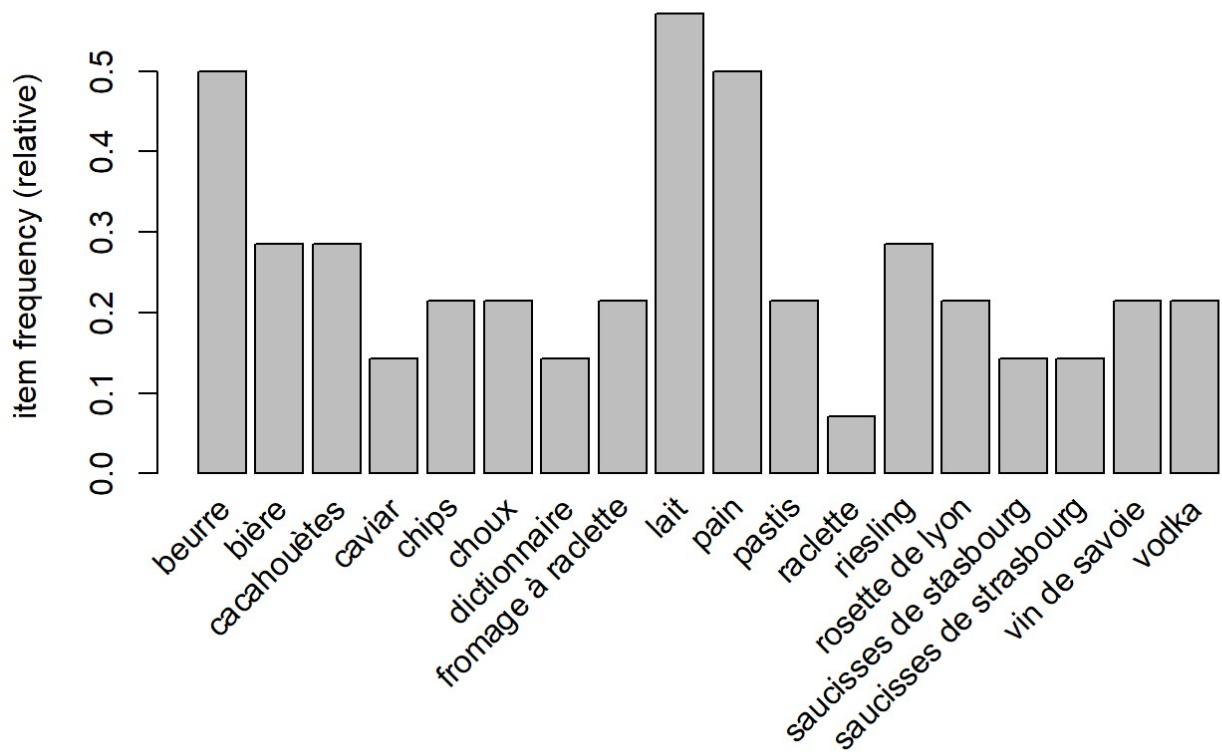
```
#saisie des données  
supermarche<-list(c("dictionnaire", "lait", "pain", "beurre"),  
                     c("caviar", "vodka", "fromage à raclette", "rosette de lyon", "vin d  
e savoie"),  
                     c("saucisses de stasbourg", "choux", "riesling", "pastis", "bière", "caca  
houètes", "chips"),  
                     c("lait", "pain", "beurre"),  
                     c("pastis", "bière", "cacahouètes", "chips"),  
                     c("lait", "pain", "beurre", "pastis", "bière", "cacahouètes", "chips"),  
                     c("lait", "pain", "beurre", "riesling", "raclette"),  
                     c("lait", "pain", "beurre", "cacahouètes"),  
                     c("saucisses de stasbourg", "choux", "riesling", "fromage à raclette", "  
rosette de lyon", "vin de savoie"),  
                     c("lait", "pain", "beurre", "caviar", "vodka"),  
                     c("bière", "vodka", "saucisses de strasbourg"),  
                     c("lait", "pain", "beurre"),  
                     c("fromage à raclette", "vin de savoie", "rosette de lyon"),  
                     c("dictionnaire", "lait", "saucisses de strasbourg", "choux", "riesling"  
))
```



```
#conversion au format transactions  
supermarche.trans <- as(supermarche, "transactions")
```

Le jeu de données comporte 14 transactions et 18 items dont la fréquence est répartie de la façon suivante :

```
#Fréquences des items  
itemFrequencyPlot(supermarche.trans)
```



Le beurre, le lait et le pain sont les items les plus fréquents dans les différentes transactions. On recherche les règles telles que le support soit supérieur à 3/18 selon l'algorithme Apriori.

```
#extraction des règles avec support supérieur à 3/18
supermarche.regles <- apriori(supermarche.trans, parameter=
  list(supp=3/18, conf=0, minlen=2, target="rules"))

inspect(supermarche.regles)
```

| ## | lhs | rhs | support | confidence | lift | cou |
|---------|----------------------|-------------------------|-----------|------------|----------|-----|
| nt | | | | | | |
| ## [1] | {fromage à raclette} | => {rosette de Lyon} | 0.2142857 | 1.000 | 4.666667 | |
| 3 | | | | | | |
| ## [2] | {rosette de Lyon} | => {fromage à raclette} | 0.2142857 | 1.000 | 4.666667 | |
| 3 | | | | | | |
| ## [3] | {fromage à raclette} | => {vin de savoie} | 0.2142857 | 1.000 | 4.666667 | |
| 3 | | | | | | |
| ## [4] | {vin de savoie} | => {fromage à raclette} | 0.2142857 | 1.000 | 4.666667 | |
| 3 | | | | | | |
| ## [5] | {rosette de Lyon} | => {vin de savoie} | 0.2142857 | 1.000 | 4.666667 | |
| 3 | | | | | | |
| ## [6] | {vin de savoie} | => {rosette de Lyon} | 0.2142857 | 1.000 | 4.666667 | |
| 3 | | | | | | |
| ## [7] | {pastis} | => {chips} | 0.2142857 | 1.000 | 4.666667 | |
| 3 | | | | | | |
| ## [8] | {chips} | => {pastis} | 0.2142857 | 1.000 | 4.666667 | |
| 3 | | | | | | |
| ## [9] | {pastis} | => {bière} | 0.2142857 | 1.000 | 3.500000 | |
| 3 | | | | | | |
| ## [10] | {bière} | => {pastis} | 0.2142857 | 0.750 | 3.500000 | |
| 3 | | | | | | |
| ## [11] | {pastis} | => {cacahuètes} | 0.2142857 | 1.000 | 3.500000 | |
| 3 | | | | | | |
| ## [12] | {cacahuètes} | => {pastis} | 0.2142857 | 0.750 | 3.500000 | |
| 3 | | | | | | |
| ## [13] | {choux} | => {riesling} | 0.2142857 | 1.000 | 3.500000 | |
| 3 | | | | | | |
| ## [14] | {riesling} | => {choux} | 0.2142857 | 0.750 | 3.500000 | |
| 3 | | | | | | |
| ## [15] | {chips} | => {bière} | 0.2142857 | 1.000 | 3.500000 | |
| 3 | | | | | | |
| ## [16] | {bière} | => {chips} | 0.2142857 | 0.750 | 3.500000 | |
| 3 | | | | | | |
| ## [17] | {chips} | => {cacahuètes} | 0.2142857 | 1.000 | 3.500000 | |
| 3 | | | | | | |
| ## [18] | {cacahuètes} | => {chips} | 0.2142857 | 0.750 | 3.500000 | |
| 3 | | | | | | |
| ## [19] | {bière} | => {cacahuètes} | 0.2142857 | 0.750 | 2.625000 | |
| 3 | | | | | | |
| ## [20] | {cacahuètes} | => {bière} | 0.2142857 | 0.750 | 2.625000 | |
| 3 | | | | | | |
| ## [21] | {beurre} | => {pain} | 0.5000000 | 1.000 | 2.000000 | |
| 7 | | | | | | |
| ## [22] | {pain} | => {beurre} | 0.5000000 | 1.000 | 2.000000 | |
| 7 | | | | | | |
| ## [23] | {beurre} | => {lait} | 0.5000000 | 1.000 | 1.750000 | |
| 7 | | | | | | |
| ## [24] | {lait} | => {beurre} | 0.5000000 | 0.875 | 1.750000 | |
| 7 | | | | | | |

```

## [25] {pain}                => {lait}           0.5000000 1.000 1.750000
7
## [26] {lait}                => {pain}           0.5000000 0.875 1.750000
7
## [27] {fromage à raclette,
##       rosette de lyon}    => {vin de savoie}  0.2142857 1.000 4.666667
3
## [28] {fromage à raclette,
##       vin de savoie}      => {rosette de Lyon} 0.2142857 1.000 4.666667
3
## [29] {rosette de Lyon,
##       vin de savoie}      => {fromage à raclette} 0.2142857 1.000 4.666667
3
## [30] {chips,
##       pastis}              => {bière}          0.2142857 1.000 3.500000
3
## [31] {bière,
##       pastis}              => {chips}          0.2142857 1.000 4.666667
3
## [32] {bière,
##       chips}               => {pastis}          0.2142857 1.000 4.666667
3
## [33] {chips,
##       pastis}              => {cacahouètes}   0.2142857 1.000 3.500000
3
## [34] {cacahouètes,
##       pastis}              => {chips}          0.2142857 1.000 4.666667
3
## [35] {cacahouètes,
##       chips}               => {pastis}          0.2142857 1.000 4.666667
3
## [36] {bière,
##       pastis}              => {cacahouètes}   0.2142857 1.000 3.500000
3
## [37] {cacahouètes,
##       pastis}              => {bière}          0.2142857 1.000 3.500000
3
## [38] {bière,
##       cacahouètes}         => {pastis}          0.2142857 1.000 4.666667
3
## [39] {bière,
##       chips}               => {cacahouètes}   0.2142857 1.000 3.500000
3
## [40] {cacahouètes,
##       chips}               => {bière}          0.2142857 1.000 3.500000
3
## [41] {bière,
##       cacahouètes}         => {chips}          0.2142857 1.000 4.666667
3
## [42] {beurre,
##       pain}                => {lait}           0.5000000 1.000 1.750000

```

```

7
## [43] {beurre,
##       lait}          => {pain}           0.5000000 1.000 2.000000
7
## [44] {lait,
##       pain}          => {beurre}         0.5000000 1.000 2.000000
7
## [45] {bière,
##       chips,
##       pastis}        => {cacahuètes}    0.2142857 1.000 3.500000
3
## [46] {cacahuètes,
##       chips,
##       pastis}        => {bière}          0.2142857 1.000 3.500000
3
## [47] {bière,
##       cacahuètes,
##       pastis}        => {chips}          0.2142857 1.000 4.666667
3
## [48] {bière,
##       cacahuètes,
##       chips}          => {pastis}         0.2142857 1.000 4.666667
3

```

Le nombre de règles obtenues est de 48. Par exemple, la règle 45 est {bière, chips, pastis} => {cacahuètes}. Son support vaut 21% ce qui signifie que parmi l'ensemble des transactions, 21% contiennent les items bière, chips, pastis, cacahuètes. Aussi, sa confiance vaut 1 ce qui signifie que tous les gens qui achètent bière, chips et pastis achètent aussi des cacahuètes. On peut remarquer que le caviar ou le dictionnaire n'apparaissent pas dans ces différentes règles. En effet, ces produits sont plutôt rares, il faudrait abaisser le support minimum pour pouvoir les observer. Au contraire, le lait est souvent présent, mais comme ce produit est très fréquent, les règles qui le contiennent ne sont pas forcément très intéressantes. Parmi ces 48 règles, on propose de s'intéresser à celles donc le lift est le plus élevé.

```

#afficher les 10 règles avec le lift le + élevé
regles.triees <- sort(supermarche.regles,by="lift")
inspect(regles.triees[1:10])

```

| ## | lhs | rhs | support | confidence | lift | cou |
|---------|--|-------------------------|-----------|------------|----------|-----|
| nt | | | | | | |
| ## [1] | {fromage à raclette} | => {rosette de Lyon} | 0.2142857 | 1 | 4.666667 | |
| 3 | | | | | | |
| ## [2] | {rosette de Lyon} | => {fromage à raclette} | 0.2142857 | 1 | 4.666667 | |
| 3 | | | | | | |
| ## [3] | {fromage à raclette} | => {vin de savoie} | 0.2142857 | 1 | 4.666667 | |
| 3 | | | | | | |
| ## [4] | {vin de savoie} | => {fromage à raclette} | 0.2142857 | 1 | 4.666667 | |
| 3 | | | | | | |
| ## [5] | {rosette de Lyon} | => {vin de savoie} | 0.2142857 | 1 | 4.666667 | |
| 3 | | | | | | |
| ## [6] | {vin de savoie} | => {rosette de Lyon} | 0.2142857 | 1 | 4.666667 | |
| 3 | | | | | | |
| ## [7] | {pastis} | => {chips} | 0.2142857 | 1 | 4.666667 | |
| 3 | | | | | | |
| ## [8] | {chips} | => {pastis} | 0.2142857 | 1 | 4.666667 | |
| 3 | | | | | | |
| ## [9] | {fromage à raclette, rosette de Lyon} | => {vin de savoie} | 0.2142857 | 1 | 4.666667 | |
| 3 | | | | | | |
| ## [10] | {fromage à raclette, vin de savoie} | => {rosette de Lyon} | 0.2142857 | 1 | 4.666667 | |
| 3 | | | | | | |

Selon ce critère, les règles {fromage à raclette} => {rosette de Lyon}, ou {rosette de Lyon} => {vin de savoie}, ou {fromage à raclette,rosette de Lyon} => {vin de savoie} font partie des plus intéressantes, à titre d'exemple. A la lecture du tableau, on voit que tous les individus achetant du fromage à raclette achètent systématiquement de la rosette de Lyon (confiance à 1), que ceux qui achètent de la rosette achètent systématiquement du vin de savoie (confiance à 1), et que donc, ceux qui achètent fromage à raclette et rosette de Lyon achètent systématiquement du vin de savoie.

Pour approfondir cet exemple, on pourra faire cet exercice (https://par.moodle.lecnam.net/pluginfile.php/236574/mod_folder/content/0/RA-%20Ex1.pdf?forcedownload=1) dont la correction est disponible ici (https://par.moodle.lecnam.net/pluginfile.php/236574/mod_folder/content/0/RA-Ex1-Corr.pdf?forcedownload=1).

4.3.2 German Credit

Nous appliquons à présent la recherche de règle sur le jeu German Credit dans lequel les variables quantitatives ont été recodées en facteur). L'illustration ci-dessous est issue d'une étude de cas de R. Rakotomalala (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Assoc_Rules_Comparison.pdf)

```
#charger le fichier de données
german.quali <- read.table(file="credit-german.txt", header=T, dec=". ", sep="\t")
```

On recherche les règles d'association selon l'algorithme Apriori, pour un support de 0.25 et une confiance de 0.75. On spécifie également que seuls les itemsets de taille comprise entre 2 et 10 sont inspectés.

```
#transformer les données attributs-variables
#en données transactionnelles
german.trans <- as(german.quali,"transactions")

#extraction des règles
german.regles <- apriori(german.trans,parameter=
                           list(supp=0.25,conf=0.75,minlen=2,maxlen=10,target="rules"))
```

On obtient alors 1928 règles dont voici un extrait

```
#afficher les 10 premières règles parmi les 1928
inspect(german.regles[1:10])
```

| ## | lhs | rhs | suppor |
|--------------|---|-------------------------------|--------|
| t confidence | lift count | | |
| ## [1] | {checking_status=0<=X<200} | => {foreign_worker=yes} | |
| 0.264 | 0.9814126 1.0191201 264 | | |
| ## [2] | {checking_status=<0} | => {foreign_worker=yes} | |
| 0.259 | 0.9452555 0.9815737 259 | | |
| ## [3] | {purpose=radio/tv} | => {num_dependents=one} | |
| 0.250 | 0.8928571 1.0566357 250 | | |
| ## [4] | {purpose=radio/tv} | => {foreign_worker=yes} | |
| 0.275 | 0.9821429 1.0198784 275 | | |
| ## [5] | {property_magnitude=real estate} | => {foreign_worker=yes} | |
| 0.262 | 0.9290780 0.9647747 262 | | |
| ## [6] | {credit_history=critical/other existing } | => {other_payment_plans=none} | |
| 0.251 | 0.8566553 1.0524021 251 | | |
| ## [7] | {credit_history=critical/other existing } | => {other_parties=none} | |
| 0.268 | 0.9146758 1.0084628 268 | | |
| ## [8] | {credit_history=critical/other existing } | => {foreign_worker=yes} | |
| 0.279 | 0.9522184 0.9888042 279 | | |
| ## [9] | {class=bad} | => {num_dependents=one} | |
| 0.254 | 0.8466667 1.0019724 254 | | |
| ## [10] | {class=bad} | => {other_parties=none} | |
| 0.272 | 0.9066667 0.9996325 272 | | |

On peut ensuite trier les règles selon le lift et s'intéresser aux règles les plus pertinentes au sens de ce critère

```
#afficher les 5 règles avec le lift le + élevé
regles.triees <- sort(german.regles,by="lift")
inspect(regles.triees[1:5])
```

| ## | lhs | rhs | support | confidence |
|--|-------|-----------|---------|------------|
| lift count | | | | |
| ## [1] {other_payment_plans=none, | | | | |
| ## existing_credits=one, | | | | |
| ## own_telephone=none} => {credit_history=existing paid} | 0.263 | 0.8218750 | | |
| 1.550708 263 | | | | |
| ## [2] {other_payment_plans=none, | | | | |
| ## housing=own, | | | | |
| ## existing_credits=one, | | | | |
| ## num_dependents=one} => {credit_history=existing paid} | 0.253 | 0.8031746 | | |
| 1.515424 253 | | | | |
| ## [3] {other_payment_plans=none, | | | | |
| ## housing=own, | | | | |
| ## existing_credits=one} => {credit_history=existing paid} | 0.287 | 0.8016760 | | |
| 1.512596 287 | | | | |
| ## [4] {other_payment_plans=none, | | | | |
| ## housing=own, | | | | |
| ## existing_credits=one, | | | | |
| ## foreign_worker=yes} => {credit_history=existing paid} | 0.271 | 0.7970588 | | |
| 1.503885 271 | | | | |
| ## [5] {other_payment_plans=none, | | | | |
| ## existing_credits=one} => {credit_history=existing paid} | 0.415 | 0.7950192 | | |
| 1.500036 415 | | | | |

Une connaissance précise des données sera nécessaire pour exploiter ces règles.

4.4 Conclusion

D'un point de vue technique, les méthodes de recherche de règles d'association reposent sur une optimisation algorithmique pour identifier toutes les règles au support et confiance élevés. Il faut pour cela limiter le nombre de lectures de la base et limiter la quantité de données stockée en mémoire. La plupart des algorithmes de recherche de règles d'association procèdent comme Apriori, l'algorithme fondateur. Ils commencent par rechercher les sous-ensembles fréquents d'items, puis ils extraient les règles d'association. Ces deux étapes supposent de fixer des seuils minimums pour le support et la confiance, paramètres qui ne sont pas toujours suffisants pour évaluer la pertinence d'une règle d'association. De nombreux indices de pertinence ont donc été proposés dans la littérature pour combler les lacunes de l'approche support - confiance. Ainsi, en aval de la recherche de règles d'association, il est possible de classer les règles par ordre décroissant de leur pertinence au sens d'un indice donné. Cependant, le choix d'un indice particulier dépend beaucoup des données analysées et de ce que cherche l'utilisateur. Il a été précisé que les données sont stockées sous un format présence/absence. Quand les variables sont qualitatives (à plusieurs modalités), il est alors nécessaire de les recoder par leurs indicatrices (c'est d'ailleurs ce qui a été appliqué lors de la mise en oeuvre sur le jeu German Credit en Section 3.2). Aussi, il peut parfois être utile de regrouper des items, de façon à diminuer le nombre de règles, typiquement, on pourra regrouper des produits similaires (e.g. des chaussettes de pointure différente) sous un item générique (e.g. chaussette). En pratique, rechercher des associations sur un grand ensemble d'événements rares conduit à une profusion de règles difficiles à interpréter de part leur nombre et leur complexité. Dans ce cas, il peut être pertinent de constituer des groupes de variables plus restreints et plus homogènes via une stratégie de classification (Section 2). Les règles obtenues sont moins nombreuses et plus simples (Plasse (2006)).

5 Analyse multi-blocs

Dans certains contextes, on peut être amené à analyser des données dans lesquelles chaque individu est caractérisé par plusieurs groupes de variables, ces groupes étant définis a priori, typiquement, des données multivues. Par exemple, si on regarde pour différents patients, des résultats d'analyses de sang, des radiographies et des comptes rendus médicaux, alors, chaque individu est décrit par 3 groupes de variables, chaque groupe correspondant à un type d'examen particulier. On rencontre également fréquemment ce type de structure quand on effectue différentes mesures à des temps différents. Dans ce cas, à chaque temps considéré correspond un groupe de variables. L'analyse de ce type de données nécessite des méthodes adaptées à leur structure. A l'image de l'ACP normée qui équilibre le poids des variables, on cherchera à équilibrer le poids des groupes, ou ``blocs'' de variables en un certain sens. Nous présentons ici différentes méthodes d'analyse factorielle dédiées à l'étude de cette structure en blocs.

5.1 Objectifs de l'analyse multi-blocs

Lorsque que l'on dispose d'un tableau de données unique, l'analyse de celui-ci passe traditionnellement par des représentations graphiques mettant en avant la structure du tableau. Ces analyses sont effectuées à l'aide de méthodes factorielles bien connues (ACP, ACM, etc) en fonction de la nature des données (quantitatives, qualitatives ou mixtes). Si l'on dispose de plusieurs tableaux, ce type d'analyse est toujours envisageable, mais il devient rapidement fastidieux dès lors que le nombre de tableaux augmente, et il ne permet pas de résumer globalement l'information contenue dans l'ensemble des tableaux. En particulier, il ne renseigne pas sur les ressemblances et différences entre tableaux. Ainsi, l'analyse multi-blocs permet de répondre à de nouveaux objectifs tels que :

- comparer les groupes de variables (de groupes de variables sont proches si deux individus proches l'un de l'autre pour un groupe le sont également pour le second groupe)
- comparer de façon simultanée les typologies des individus vus par chaque groupe de variables pris un par un. En particulier, quand les différents tableaux correspondent à une dimension temporelle, l'analyse multi-blocs permettra de voir l'évolution des profils des individus au cours du temps.

Une analyse multi-blocs peut se résumer en 4 phases :

1. Etude de l'interstructure. Dans cette analyse, chacun des tableaux est associé à un objet dont le choix dépend de la méthode d'analyse employée. Ces objets caractérisent les tableaux et sont de mêmes nature (contrairement aux tableaux qui peuvent potentiellement être de dimensions différentes). L'analyse globale consiste à comparer ses objets entre eux, d'y repérer des groupes homogènes. Dans cette phase, la description précise des différences et ressemblances entre objets n'est pas effectuée, il s'agit juste d'une analyse globale des proximités ou des différences entre tableaux.
2. Recherche d'un compromis. Cette étape consiste à déterminer un espace commun de représentation résument les données globalement et non résument chaque tableau pris isolément.
3. Etude de l'intrastructure. L'analyse de l'intrastructure consiste en l'analyse de ce compromis et permet d'étudier plus finement les ressemblances ou différences entre tableaux.
4. Analyse des trajectoires. Cette dernière phase consiste à comparer les profils des individus ou les variables selon les différents groupes.

5.2 Méthodes

Il existe de nombreuses méthodes permettant d'analyser les données multi-blocs, nous présentons ici 3 d'entre elles : la double analyse en composantes principales (DACP), la structuration de tableaux à trois indices de la statistique (STATIS) et l'analyse factorielle multiple (AFM). D'autres méthodes basées sur l'existence d'un modèle sous-jacent aux différents tableaux de données ont également été développées par les anglo-saxons. Citons parmi ces modèles INDSCAL, IDIOSCAL et PARAFAC (cf. Kroonenberg (1983)).

5.2.1 Notations

On considère N tableaux X_t représentant p_t variables quantitatives (X^1, \dots, X^{p_t}) portant sur n_t individus (cf Figures 5.1 et 5.2). Sauf mention du contraire, on supposera que le nombre d'individus est identique d'un tableau à l'autre ($n_t = n$ pour tout t), permettant une concaténation des tableaux les uns à côté des autres et faisant ainsi apparaître N ``blocs'' de variables.

$$\mathbf{X}_t = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{p_{t1}} \\ X_{12} & X_{22} & \cdots & X_{p_{t2}} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{p_{tn}} \end{bmatrix}$$

Figure 5.1: Schéma d'un tableau (i.e. un bloc).

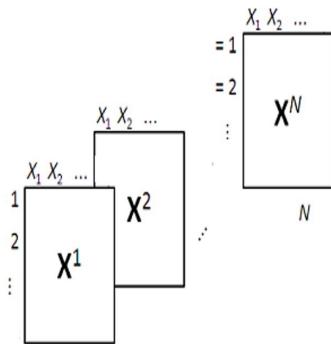


Figure 5.2: Schéma d'un ensemble de tableaux (i.e. données multi-blocs).

Par ailleurs, à chaque tableau on associe une métrique M_t définissant les proximités entre individus et une matrice de pondération de ces individus, notée D et supposée identique d'un tableau à l'autre. On note par ailleurs \mathbf{V}_t la matrice de variance-covariance associée au tableau X_t .

5.2.2 Double ACP

5.2.2.1 Méthode

La double ACP (Bourotche (1975)) s'applique généralement sur des données où les mêmes variables ont été mesurées sur des individus identiques ($p_t = p$ pour tout t) mais à des instants différents, chaque tableau correspondant ici à un temps donné. Il est possible d'étendre cette méthode à une 3ème dimension qui ne correspondrait pas au temps, mais l'interprétation devient alors beaucoup plus difficile (Dazy et al. (1996)).

L'étude de l'interstructure consiste en une (première) analyse en composantes principales effectuée sur les centres de gravité des nuages de points correspondants à chacun des tableaux. Chaque tableau est en quelque sorte représenté par le vecteur centre de gravité. Cette phase ne présente donc aucune difficulté théorique particulière. Souvent, on peut vérifier que le premier axe s'explique en termes d'évolution globale dans le temps : les centres de gravité sont alors positionnés selon un ordre chronologique.

Une fois cette structure analysée, on recherche un espace compromis résumant au mieux les différents

nuages (chaque nuage correspondant à un tableau). Pour cela, on choisit de chercher les axes qui maximisent la somme des inerties de projection des différents nuages (d'autres critères sont possibles, cf Dazy et al. (1996)). On remarque que dans le cas où il n'y aurait qu'un seul nuage, ce critère revient à celui utilisé en ACP. L'inertie de projection de chaque nuage vaut donc $\sum_{k=1}^q v'_k V_t v_k$ et on cherchera les facteurs principaux (v_1, \dots, v_q) maximisant l'inertie de projection de l'ensemble des nuages $\sum_{t=1}^N \sum_{k=1}^q v'_k V_t v_k$. Pour cela, on peut remarquer que le critère se réécrit $\text{Max}_{(v_1, \dots, v_q)} \sum_{k=1}^q v'_k V v_k$ avec $V = \sum_{t=1}^N V_t$. De cette façon, on voit apparaître une expression de l'inertie identique à celle habituellement rencontrée avec un seul tableau, la matrice V pouvant être considérée comme une matrice de variance-covariance `` compromis'' des variables sur les différents temps. Ainsi, l'espace compromis s'obtient assez naturellement en effectuant une (deuxième) ACP sur l'ensemble des nuages, préalablement centrés relativement à leur centre de gravité propre.

5.2.2.2 Exemple

Pour illustrer la méthode DACP, nous reprenons l'exemple traité dans Dazy et al. (1996) portant sur des données de criminalité entre 1974 et 1993 dans les départements de France métropolitaine. Ces données présentent la criminalité et la délinquance constatées par procès-verbal. Les crimes et délits sont répartis en 9 catégories pour l'ensemble des départements métropolitains, par année, de 1974 à 1993 :

- VO : vols et recels
- FX : faux et escroqueries
- DF : délits financiers
- CH : chèques sans provisions
- CR : coups, règlements de comptes, traumatismes
- ST : stupéfiants
- DD : destructions et dégradations
- ET : délits à la police des étrangers
- DV : divers

Ces données peuvent être représentées sous la forme d'un cube à trois dimensions : département, catégorie, année.

5.2.2.2.1 Interstructure

Le tableau suivant regroupe les centres de gravité des tableaux analysés.

| Années | VO | FX | DF | CH | CR | ST | DD | ET |
|--------|-------|------|-------|-------|------|-------|------|-------|
| 1974 | 15,46 | 1,31 | 0,27 | 3,31 | 1,41 | 0,040 | 1,06 | 0,088 |
| 1975 | 16,84 | 1,51 | 0,24 | 2,64 | 1,45 | 0,056 | 1,27 | 0,092 |
| 1976 | 16,44 | 1,63 | 0,27 | 2,11 | 1,44 | 0,068 | 1,25 | 0,076 |
| 1977 | 18,14 | 1,98 | 0,31 | 3,39 | 1,56 | 0,090 | 1,55 | 0,093 |
| 1978 | 18,26 | 2,13 | 0,37 | 3,88 | 1,54 | 0,123 | 1,70 | 0,113 |
| 1979 | 19,90 | 2,35 | 0,40 | 4,34 | 1,64 | 0,185 | 1,98 | 0,134 |
| 1980 | 21,89 | 2,67 | 0,44 | 5,90 | 1,74 | 0,201 | 2,41 | 0,179 |
| 1981 | 23,56 | 3,08 | 0,45 | 7,00 | 1,81 | 0,275 | 2,68 | 0,203 |
| 1982 | 27,20 | 3,88 | 0,73 | 8,33 | 1,86 | 0,388 | 3,18 | 0,158 |
| 1983 | 28,41 | 4,32 | 0,70 | 8,41 | 1,93 | 0,371 | 3,24 | 0,346 |
| 1984 | 30,18 | 4,53 | 0,95 | 7,64 | 1,95 | 0,401 | 3,39 | 0,341 |
| 1985 | 30,65 | 4,68 | 0,84 | 6,46 | 1,97 | 0,499 | 3,26 | 0,369 |
| 1986 | 28,72 | 4,74 | 0,66 | 4,89 | 1,78 | 0,727 | 2,98 | 0,375 |
| 1987 | 27,64 | 4,80 | 0,45 | 4,42 | 1,72 | 0,699 | 3,02 | 0,364 |
| 1988 | 27,18 | 4,54 | 0,50 | 3,96 | 1,88 | 0,666 | 2,98 | 0,365 |
| 1989 | 28,34 | 4,84 | 0,35 | 3,90 | 1,99 | 0,682 | 3,26 | 0,414 |
| 1990 | 30,99 | 5,17 | 0,34 | 3,58 | 2,03 | 0,741 | 3,77 | 0,366 |
| 1991 | 32,98 | 5,64 | 0,43 | 3,01 | 2,14 | 0,828 | 4,29 | 0,449 |
| 1992 | 34,76 | 5,14 | 0,38 | 0,35 | 2,21 | 0,955 | 4,99 | 0,501 |
| 1993 | 35,13 | 5,29 | 0,500 | 0,056 | 2,32 | 0,976 | 5,39 | 0,511 |

Figure 5.3: Centres de gravité des tableaux analysés

L'analyse en composantes principales de ce tableau centré fournit les valeurs propres suivantes :

| Axes | Valeurs propres | % d'inertie | % cumulé |
|------|-----------------|-------------|----------|
| 1 | 5,87 | 73,42 | 73,42 |
| 2 | 1,70 | 21,27 | 94,70 |
| 3 | 0,19 | 2,31 | 97,01 |
| 4 | 0,17 | 2,10 | 99,11 |

Figure 5.4: Valeurs propres de l'interstructure déterminés par la DACP

Le premier plan cumulant 94.70% d'inertie, on se limitera à son étude seule.

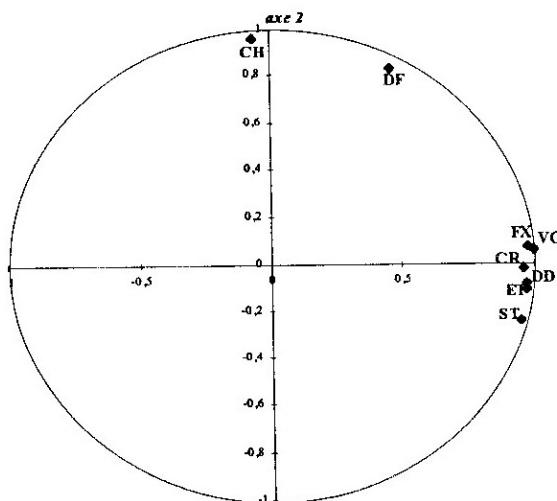


Figure 5.5: Cercle des corrélations pour le plan 1-2 (interstructure de la DACP)

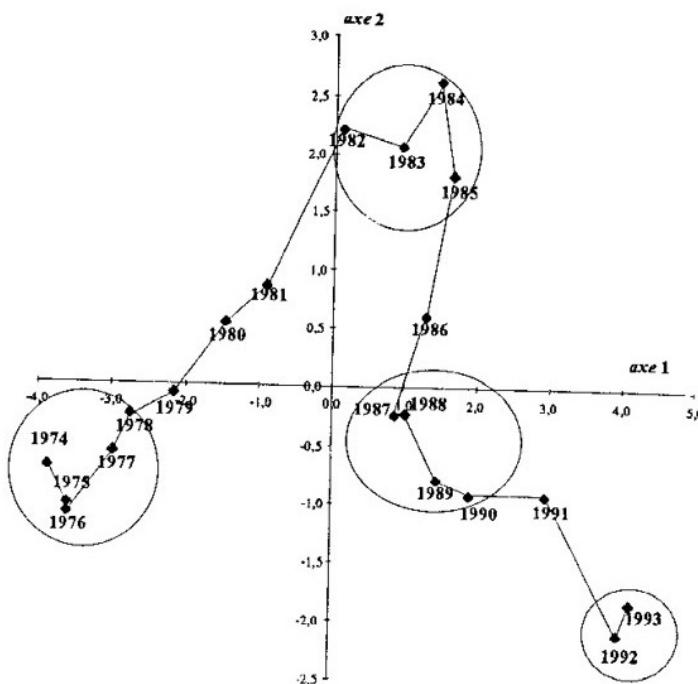


Figure 5.6: Graphe des individus plan 1-2 (interstructure de la DACP)

On observe sur le cercle des corrélations Figure 5.5) un effet taille pour l'axe 1 pour les variables VO, DD, ET, FX, CR et ST. Le second axe quant à lui est essentiellement lié aux variables CH et DF.

Le plan des individus (Figure 5.6) fait apparaître une évolution temporelle quasi linéaire le long de l'axe 1. Cela signifie que les variables qui sont fortement corrélées avec cet axe (cf cercle des corrélations Figure 5.5) varient de façon linéaire par rapport au temps. Ceci est confirmé par les Figures 5.7 et 5.8 représentant l'évolution de ces variables au cours du temps. On en déduit que la criminalité associée aux variables VO, ST, ET, FX, CR et DD a progressé.

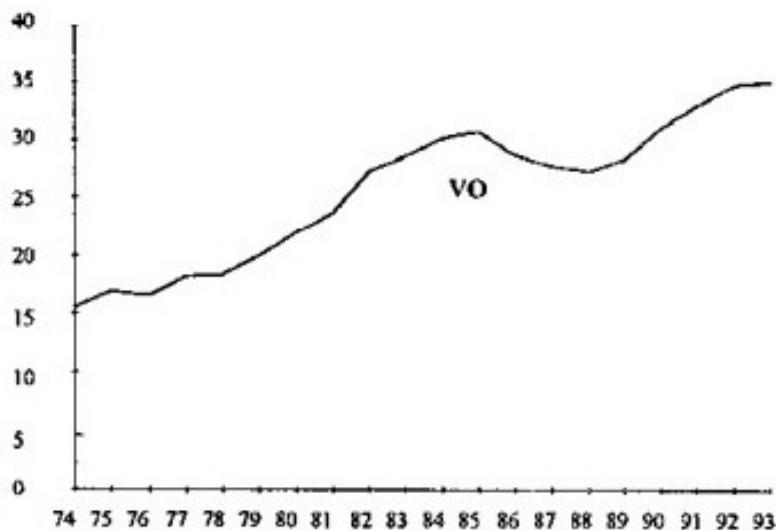


Figure 5.7: Evolution de la variable VO (vols et recels) en France métropolitaine

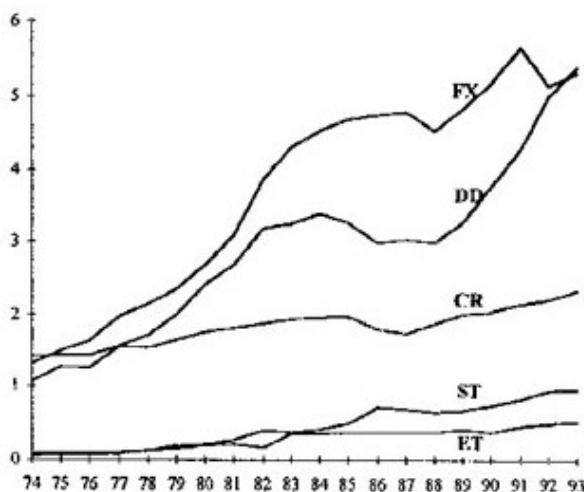


Figure 5.8: Evolution des variables ST, ET, FX, CR et DD en France métropolitaine

De la même façon, on peut dire que la criminalité associée aux variables DF et CH a eu tendance à augmenter jusqu'au milieu des années 1980 puis a diminué au-delà, ce qui est confirmé par la Figure 5.9.

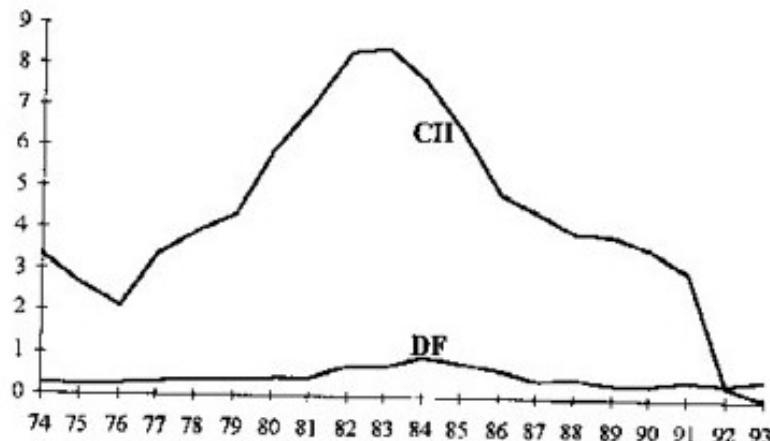


Figure 5.9: Evolution des variables CH et DF en France métropolitaine

La représentation de l'interstructure sur la Figure 5.6 fait également apparaître des groupes d'années qui se ressemblent, ces groupes sont matérialisés par des cercles. Les tableaux de données correspondant sont donc assez proches les uns des autres. Entre 1974 à 1978, on observe une lente augmentation de l'ensemble des variables corrélées à l'axe 1. Puis, entre les années 1978 et 1982, on observe une augmentation plus marquée de l'ensemble des variables. La structure des tableaux analysés se modifie donc fortement de 1978 à 1982. Les années comprises entre 1982 et 1985 sont caractérisées par des taux de chèques sans provisions et de délits financiers très supérieurs aux valeurs moyennes de ces mêmes taux sur la période. Ces derniers diminuent sur les années 1986-1987, tandis que les délits correspondants aux variables corrélées à l'axe 1 restent constants, voire diminuent pour cette même période. Enfin, entre 1989 et 1993, les délits correspondants aux variables corrélées à l'axe 1 augmentent de nouveau tandis que les taux de chèques sans provisions et de délits financiers tendent à diminuer.

5.2.2.2 Inrastructure

L'inrastructure s'analyse en effectuant une ACP sur les différents tableaux centrés et concaténés les uns en dessous des autres. La décroissance des valeurs propres suggère d'étudier les 3 premiers axes cumulant 65.3% de l'inertie (Figure 5.10). Nous nous limiterons néanmoins au premier plan, dans la mesure où le but ici est simplement d'illustrer la méthode DACP.

| Axes | Valeurs propres | % d'inertie | % cumulé |
|------|-----------------|-------------|----------|
| 1 | 3,13 | 39,18 | 39,18 |
| 2 | 1,11 | 13,89 | 53,07 |
| 3 | 0,97 | 12,19 | 65,26 |
| 4 | 0,83 | 10,34 | 75,59 |
| 5 | 0,73 | 9,12 | 84,71 |
| 6 | 0,58 | 7,31 | 92,02 |
| 7 | 0,39 | 4,90 | 96,92 |
| 8 | 0,25 | 3,08 | 100 |

Figure 5.10: Valeurs propres de l'inrastructure (DACP)

L'interprétation des axes se fait à l'aide de l'examen des corrélations avec les variables ``compromis'' (Figure 5.11).

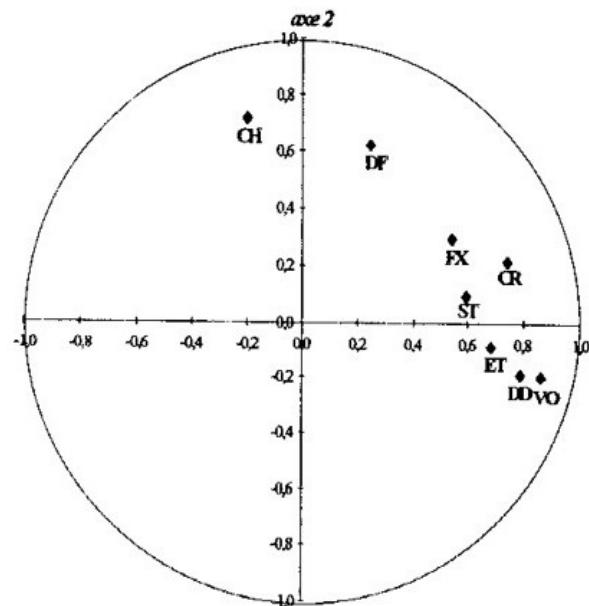


Figure 5.11: Cercle des corrélations pour le plan 1-2 (infrastructure de la DACP)

Comme précédemment, on observe un effet taille pour l'axe 1 pour les variables VO, DD, ET, FX, CR et ST. Le second axe quant à lui oppose les variables CH et DF d'une part, aux variables DD et VO d'autre part. Ainsi, l'axe deux met en opposition deux catégories de criminalité : la criminalité financière et la criminalité avec violence.

Remarque : la DACP ne prévoit pas la représentation de positions-compromis des individus (ici, les départements de France métropolitaine). On pourra néanmoins étudier la représentation des trajectoires.

5.2.2.2.3 Trajectoires

Les trajectoires s'obtiennent en projetant simultanément les individus de chaque tableau sur le plan principal. La Figure 5.12 représente 3 de ces trajectoires.

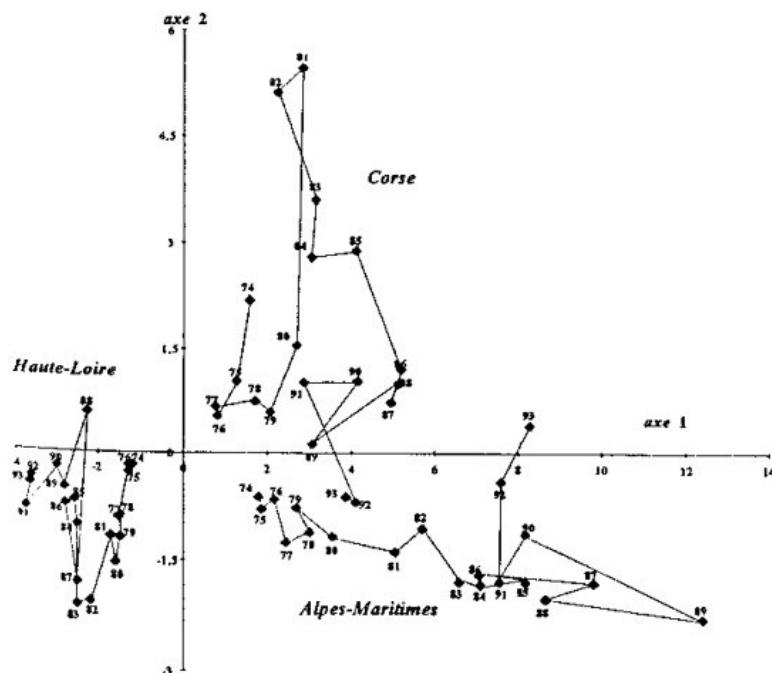


Figure 5.12: Trajectoires des départements Haute-Loire, Alpes-Maritimes et Corse.

A titre d'exemple, la trajectoire des Alpes-Maritimes se situe dans le coin inférieur droit du plan principal. Les

coordonnées selon l'axe 1 sont toujours positives, ce qui signifie que dans ce département la criminalité relative aux variables corrélées à l'axe 1 est toujours supérieure à la moyenne de France métropolitaine. Par ailleurs, on constate une évolution de cette trajectoire vers la droite ce qui signifie que chaque année cette criminalité à tendance à augmenter relativement à la moyenne de France métropolitaine. Les coordonnées selon l'axe 2 sont quant à elles toujours négatives, excepté en 1993. En effet, les Alpes-Maritimes présentent globalement un taux de chèque sans provision nettement inférieur au taux observé en France métropolitaine durant toute la période, mais ce taux s'est rapproché du taux national entre 1991 et 1993.

5.2.3 STATIS/STATIS duale

La méthode de structuration de tableaux à trois indices de la statistique (L'Hermier des Plantes (1976)) permet, tout comme la double ACP, d'analyser simultanément plusieurs tableaux de données quantitatives. A la différence de la double ACP, les tableaux peuvent avoir des dimensions différentes. On distingue alors deux versions de la méthode. La première (dite méthode STATIS) s'applique au cas où les individus sont les mêmes d'un tableau à l'autre (les variables pouvant être différentes). La seconde, dite STATIS duale, s'applique dans les cas où les variables sont les mêmes d'un tableau à l'autre (les individus pouvant être différents). La méthode STATIS duale étant très proche dans son fonctionnement de la méthode STATIS, nous ne l'évoquerons que brièvement.

5.2.3.1 STATIS

5.2.3.1.1 Méthode

Les tableaux étant potentiellement de dimensions différentes, l'étude de l'interstructure telle qu'effectuée en DACP n'est pas possible. Pour obtenir une visualisation de l'interstructure, il va donc falloir définir un autre objet propre à chaque tableau que le centre de gravité. Ainsi, la méthode STATIS consiste à utiliser les matrices de produits scalaires entre individus définies par $W_t = X_t M_t X_t'$, de dimensions $n \times n$ pour tous les tableaux. Par la suite, il s'agit de définir une distance entre ces objets et rechercher une représentation plane respectant ``au mieux'' ces distances. La distance entre deux objets est ici définie à partir du produit scalaire de Hilbert-Schmidt

$$\langle W_t, W_{t'} \rangle = \text{Trace}(DW_t D W_{t'})$$

selon

$$d^2(W_t, W_{t'}) = \|W_t - W_{t'}\|^2 = \langle W_t - W_{t'}, W_t - W_{t'} \rangle$$

En pratique, il arrive que les objets W_t aient des normes très différentes. Ceci peut grandement influencer l'analyse. Pour cette raison, il pourra être utile de normaliser les objets, i.e. de considérer les objets $\frac{W_t}{\|W_t\|}$. Dès lors, on peut remarquer que le produit scalaire entre deux objets normalisés correspond au coefficient RV entre les deux objets. Ce coefficient est très classique pour la mesure de liaison entre deux groupes de variables multidimensionnels, ce qui justifie le choix du produit scalaire de Hilbert-Schmidt. Il est compris entre 0 et 1, un RV égal à 1 correspondant à deux nuages de points identiques.

Pour obtenir une représentation plane, on procède alors comme en ACP : on définit une pondération des objets et on construit la matrice des produits scalaires entre objets, qui par sa diagonalisation permettra d'obtenir les composantes principales. Plus précisément, on note S la matrice de dimensions $N \times N$ des produits scalaires entre objets où l'élément $s_{tt'} = \langle W_t, W_{t'} \rangle$ ou $s_{tt'} = \langle \frac{W_t}{\|W_t\|}, \frac{W_{t'}}{\|W_{t'}\|} \rangle$ si on a normalisé. On note $\Delta = \text{diag}(\pi_1, \dots, \pi_N)$ la matrice des poids. Soit λ_k , la k -ième valeur propre de la matrice $S\Delta$ associée à u_k , le k -ième vecteur propre. Les coordonnées des points associés aux objets W_1, \dots, W_N sur

l'axe k sont contenues dans $c_k = \sqrt{\lambda_k} u_k$, k -ième composante principale.

A partir des deux premières composantes, il est alors possible d'obtenir une représentation synthétique en deux dimensions résumant aux mieux les distances (de Hilbert-Schmidt) entre objets. Cette représentation nous permet de visualiser l'interstructure, c'est-à-dire que la proximité entre deux points convenablement représentés dans ce plan traduit l'existence d'une structure des individus commune aux tableaux correspondants.

La procédure pour étudier l'interstructure est résumée dans le schéma de la Figure 5.13.

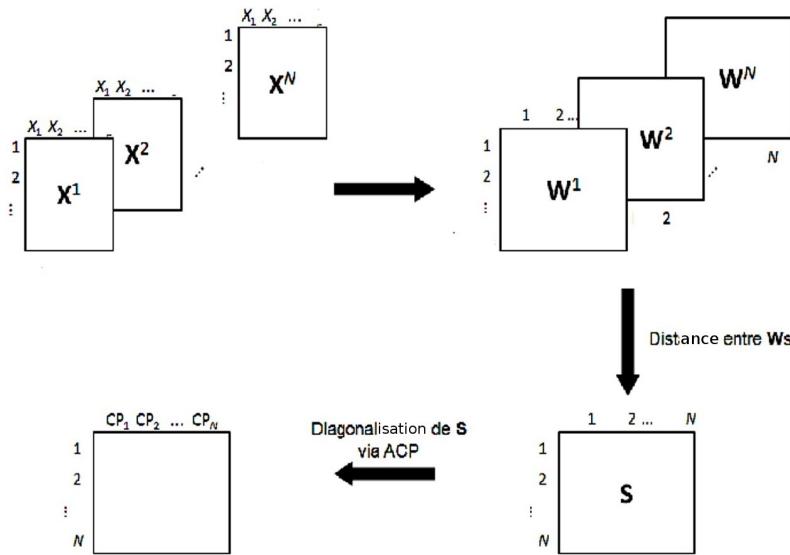


Figure 5.13: Résumé de l'étude interstructure.

L'étude de l'interstructure consiste d'abord à rechercher une matrice compromis notée W^{CO} qui résume au mieux l'ensemble de tableaux. Elle est définie comme une moyenne pondérée des objets W^t maximisant la corrélation au sens du produit scalaire de Hilbert Schmidt, en particulier, du coefficient RV dans le cas normé.

On montre (Lavit et al. (1994)) que la solution est donnée par :

$$W^{CO} = \sum_{t=1}^N \alpha_t W^t$$

avec $\alpha_t = \frac{u^1}{\sqrt{\lambda^1}} \pi_t$ la t -ième coordonnée du premier vecteur propre de $S\Delta$ normalisé.

Ces poids représentent le degré d'accord ou de ressemblance entre les tableaux et le compromis. Cette définition du compromis confère à STATIS une robustesse aux valeurs aberrantes : plus un tableau est différent des autres, moins il a d'influence sur le compromis.

La diagonalisation de la matrice compromis $W^{CO} D$ fournit les vecteurs propres permettant d'obtenir l'espace de représentation commun à l'ensemble des tableaux. Il est alors possible de visualiser sur le premier plan principal des points artificiels B_i ($i = 1, \dots, n$), correspondant aux individus, et appelés points compromis. Les coordonnées sur le k -ième axe factoriel sont les éléments du vecteur suivant :

$$z^k = \sqrt{\delta^k} v^k = \frac{1}{\sqrt{\delta^k}} W^{CO} D v^k$$

où δ_k est la valeur propre associée au k -ième vecteur propre v^k .

La procédure pour construire le compromis menant à l'analyse de l'intrastructure est résumée dans le schéma de la Figure 5.14.

Afin d'interpréter ce compromis, et donc visualiser l'intrastructure, on peut remarquer que chaque composante principale du compromis est un vecteur à n dimensions, tout comme les variables initiales. Ainsi, il va être possible de calculer la corrélation entre ces composantes et les variables des tableaux X_t et ainsi de construire un cercle des corrélations qui permettra d'expliquer les positions compromis des individus.

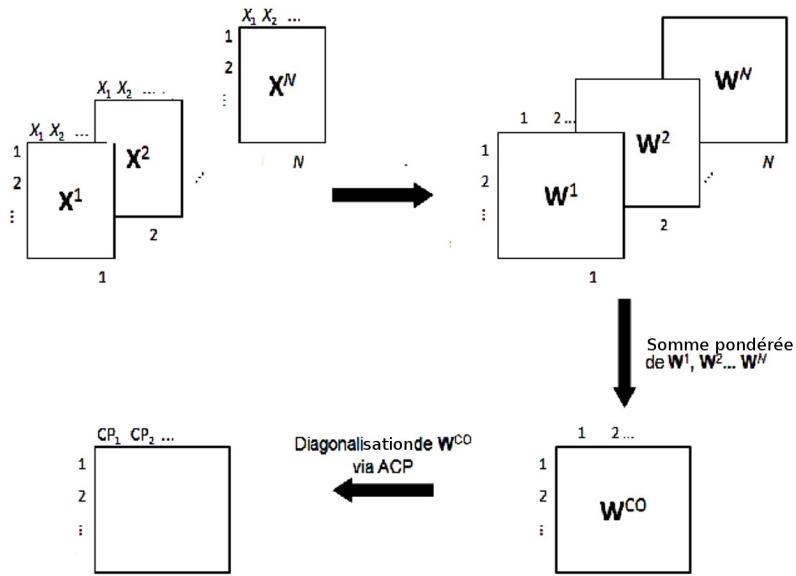


Figure 5.14: Résumé de la construction du compromis.

De plus, il est possible de représenter les individus de tous les tableaux W^t en les projetant sur le plan compromis comme des points supplémentaires. La coordonnée $z_{t,i}^k$ de l'individu i du tableau t sur le k -ième axe factoriel est :

$$z_{t,i}^k = \frac{1}{\sqrt{\delta^k}} w_i^t v^k$$

où w_i^t est la i -ième ligne de W_t et δ^k et v^k sont définies comme précédemment. Les différentes positions d'un individu selon les tableaux définissent sa trajectoire qui permet de mettre en évidence des écarts entre les N tableaux au niveau individuel.

5.2.3.1.2 Exemple

Pour illustrer la méthode STATIS, nous reprenons l'exemple traité dans Dazy et al. (1996) portant sur les résultats des élections présidentielles françaises de 1968 à 1988. Les données sont structurées en 4 tableaux (un par année électorale). Dans chacun de ces tableaux on a disposé en ligne les 95 départements et en colonne les taux de votes pour les différents candidats au premier et au second tour. Les variables du second tour seront utilisées en tant que variables illustratives. Une description des différentes variables est fournie dans les Figures 5.15 et 5.16. On notera que ces variables ne sont pas les mêmes au travers des différentes années.

En 1969 :

- GI1¹ : Vote Defferre
- DI1 : Vote Poher + Vote Pompidou
- GP1 : Vote Rocard + Vote Duclos + Vote Krivine
- AT1 : Vote Ducatel
- NP1 : Votes blancs et nuls + Abstention
- CN2² : Vote Poher

- DI2 : Vote Pompidou
- NP2 : Votes blancs et nuls + Abstention

En 1974 :

- GI1 : Vote Mitterrand
- DI1 : Vote Giscard + Vote Muller + Vote Royer + Vote Chaban
- GP1 : Vote Laguiller + Vote Krivine
- DP1 : Vote Le Pen + Vote Renouvin *(droite protestataire)*
- AT1 : Vote Hérault + Vote Sebag + Vote Dumont
- NP1 : Votes blancs et nuls + Abstention
- GI2 : Vote Mitterrand
- DI2 : Vote Giscard d'Estaing
- NP2 : Votes blancs et nuls + Abstention

En 1981 :

- GI1 : Vote Mitterrand + Vote Crépeau
- DI1 : Vote Giscard + Vote Chirac + Vote Debré + Vote Garaud
- GP1 : Vote Laguiller + Vote Marchais + Vote Bouchardieu
- AT1 : Vote Lalonde
- NP1 : Votes blancs et nuls + Abstention
- GI2 : Vote Mitterrand
- DI2 : Vote Giscard d'Estaing
- NP2 : Votes blancs et nuls + Abstention

En 1988 :

- GI1 : Vote Mitterrand
- DI1 : Vote Barre + Vote Chirac
- GP1 : Vote Lajoinie + Vote Juquin + Vote Laguiller + Vote Boussel
- DP1 : Vote Le Pen
- AT1 : Vote Waechter
- NP1 : Votes blancs et nuls + Abstention
- GI2 : Vote Mitterrand
- DI2 : Vote Chirac
- NP2 : Votes blancs et nuls + Abstention

Figure 5.15: Présentation des données.

Tableau 3.1 Statistiques élémentaires sur les tableaux de données

| <i>Scrutin 1969</i> | | | | | | | | |
|---------------------|------|-------|-------|------|-------|-------|-------|-------|
| (en %) | GI1 | DI1 | GP1 | AT1 | NP1 | DI2 | CN2 | NP2 |
| Moyenne | 3,74 | 52,77 | 19,48 | 0,98 | 23,04 | 37,79 | 28,33 | 33,88 |
| Écart-type | 1,07 | 5,83 | 5,18 | 0,23 | 1,68 | 5,35 | 4,37 | 5,36 |

| <i>Scrutin 1974</i> | | | | | | | | | |
|---------------------|-------|-------|------|------|------|-------|-------|-------|-------|
| (en %) | GI1 | DI1 | GP1 | DP1 | AT1 | NP1 | DI2 | GI2 | NP2 |
| Moyenne | 36,33 | 43,41 | 2,46 | 0,78 | 0,98 | 15,81 | 44,19 | 42,86 | 12,95 |
| Écart-type | 5,55 | 5,58 | 0,49 | 0,18 | 0,36 | 1,68 | 5,83 | 6,13 | 1,42 |

| <i>Scrutin 1981</i> | | | | | | | | |
|---------------------|-------|-------|-------|------|-------|-------|-------|-------|
| (en %) | GI1 | DI1 | GP1 | AT1 | NP1 | DI2 | GI2 | NP2 |
| Moyenne | 23,12 | 39,61 | 14,84 | 1,21 | 19,36 | 40,35 | 44,10 | 15,54 |
| Écart-type | 2,92 | 4,75 | 3,94 | 0,55 | 2,01 | 4,84 | 4,94 | 1,61 |

| <i>Scrutin 1988</i> | | | | | | | | | |
|---------------------|-------|-------|------|-------|------|-------|-------|-------|-------|
| (en %) | GI1 | DI1 | GP1 | DP1 | AT1 | NP1 | DI2 | GI2 | NP2 |
| Moyenne | 27,70 | 29,66 | 9,16 | 11,09 | 3,12 | 19,26 | 37,78 | 44,51 | 17,71 |
| Écart-type | 3,42 | 4,05 | 2,61 | 3,30 | 0,83 | 1,65 | 3,94 | 4,19 | 1,96 |

Figure 5.16: Statistiques élémentaires sur les tableaux de données.

5.2.3.1.2.1 Interstructure

Afin d'équilibrer l'influence des différents tableaux dans l'analyse, les objets normés sont utilisés ($W_t / \| W_t \|$). La matrice à diagonaliser pour obtenir la visualisation de l'interstructure sera la matrice des coefficients RV (Figure 5.17).

| Années | 1969 | 1974 | 1981 | 1988 |
|--------|-------|-------|-------|-------|
| 1969 | 1,000 | 0,571 | 0,576 | 0,490 |
| 1974 | 0,571 | 1,000 | 0,710 | 0,622 |
| 1981 | 0,576 | 0,710 | 1,000 | 0,746 |
| 1988 | 0,490 | 0,622 | 0,746 | 1,000 |

Figure 5.17: Matrice des coefficients RV.

Les coefficients les plus élevés sont ceux des trois dernières années, alors que ce coefficient est plus faible entre 1969 et les autres années. La structure des départements en 1969 apparaît donc différente de celle des trois autres scrutins. Par ailleurs, la structure globale des départements a peu évolué entre 1974 et 1988.

L'analyse des valeurs propres suite à la diagonalisation de la matrice des RV (Figure 5.18) nous amène à analyser les deux premiers axes, cumulant 85,19% de l'inertie.

| Axes | Valeurs propres | % d'inertie | % cumulé |
|------|-----------------|-------------|----------|
| 1 | 2,87 | 71,66 | 71,66 |
| 2 | 0,54 | 13,53 | 85,19 |
| 3 | 0,36 | 9,04 | 94,24 |
| 4 | 0,23 | 5,76 | 100 |

Figure 5.18: Valeurs propres de l'interstructure.

Le premier plan factoriel (Figure 5.19) nous permet d'identifier une proximité entre les objets caractérisant

les années de 1974 à 1988, tandis que l'objet représentant l'année 1969 est à l'écart, témoignant ainsi d'une différence de structure des individus entre les 3 derniers scrutins et celui de 1969.

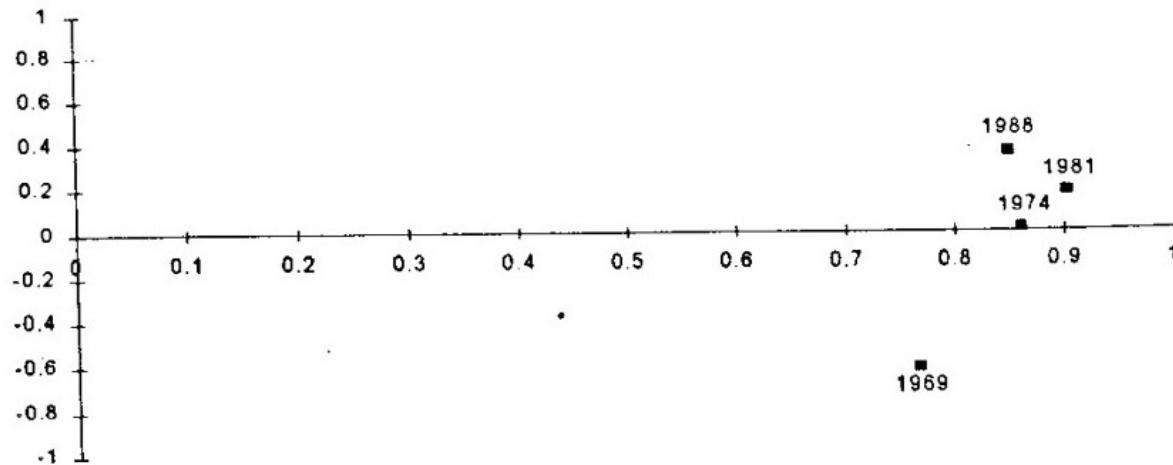


Figure 5.19: Représentation de l'interstructure.

5.2.3.1.2.2 Inrastructure

Les ressemblances et différences entre les structures des départements ayant été mises en évidence, on construit le tableau ``moyen'', ou compromis, \bar{W} , et on représente les positions-compromis des départements que l'on interprète à l'aide des corrélations entre les variables et les axes.

Les 10 premières valeurs propres de WD sont reportées en Figure 5.20 ($D = \frac{1}{95} I_{95}$ ici car tous les départements ont le même poids)

| Axes | Valeurs propres | % d'inertie | % cumulé |
|------|-----------------|--------------|--------------|
| 1 | 0,753 | 31,68 | 31,68 |
| 2 | 0,480 | 20,19 | 51,88 |
| 3 | 0,257 | 10,81 | 62,68 |
| 4 | 0,208 | 8,74 | 71,42 |
| 5 | 0,166 | 6,97 | 78,39 |
| 6 | 0,106 | 4,44 | 82,83 |
| 7 | 0,097 | 4,07 | 86,90 |
| 8 | 0,074 | 3,12 | 90,02 |

Figure 5.20: Valeurs propres de l'inrastructure.

Leur examen suggère d'analyser les 3 premiers axes cumulant 62,68% d'inertie. Nous nous limiterons toutefois aux deux premiers car l'objet ici est juste d'illustrer la méthode STATIS sur des données réelles.

Les corrélations entre les axes et les variables initiales sont présentées en Figure 5.21.

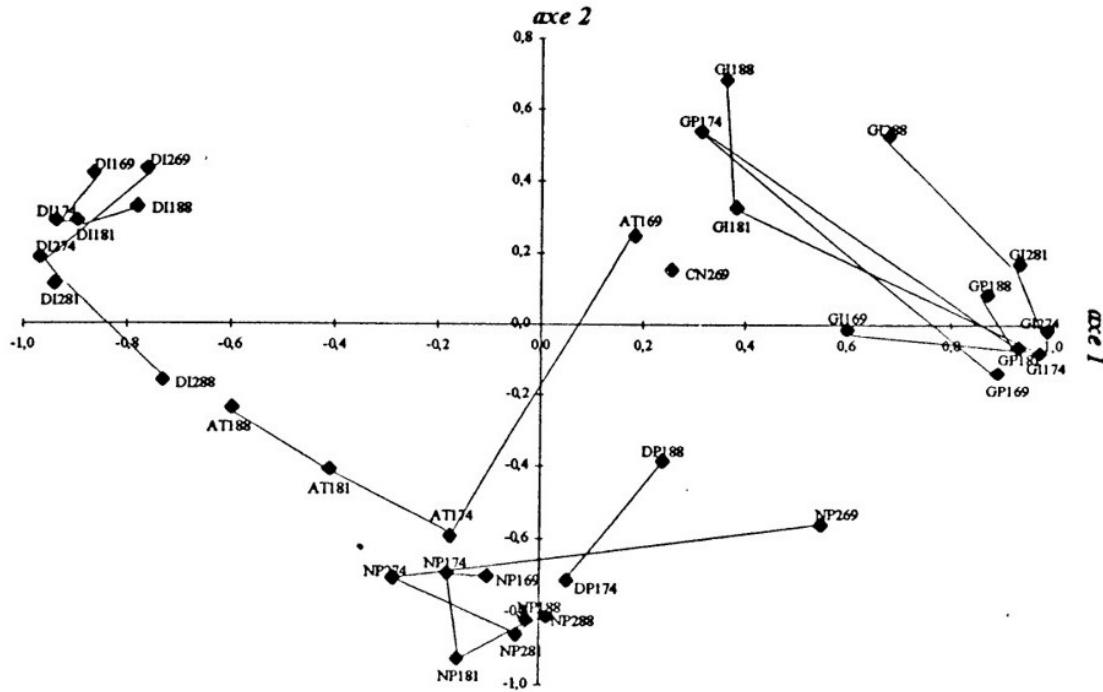


Figure 5.21: Corrélation des variables dans le plan principal.

L'axe 1 oppose la droite institutionnelle (variables DI1, DI2) à gauche (corrélations négatives) et les gauches institutionnelle et protestataire (variables GI1, GI2, GP1) à droite (corrélations positives) et ceci sur toute la période 1969-1888.

Les variables les plus corrélées à l'axe 2 sont la non-participation (NP1) et la droite protestataire (DP1). Dans une moindre mesure, la gauche institutionnelle a une corrélation positive avec le deuxième axe à partir de 1981. On peut ainsi remarquer une participation plus forte des électeurs dans les départements où le vote est à dominante gauche institutionnelle que dans les départements où le vote est à dominante gauche protestataire. Ce second axe s'interprète donc comme un axe protestataire où la non-participation et la droite protestataire sont plus présents.

A la lecture de cette interprétation, on est en mesure de décrire les profils des départements en étudiant les positions compromis des individus dans le plan principal (cf Figure 5.22). Notons que les points compromis correspondent à des positions moyennes des départements sur la période étudiée.

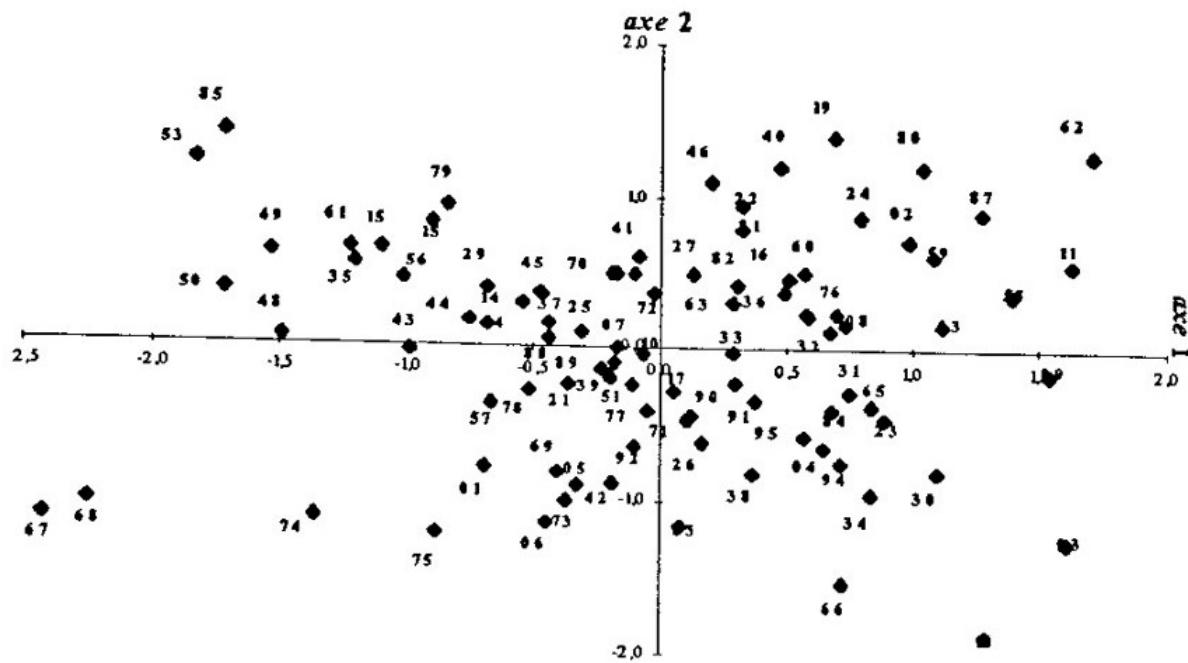


Figure 5.22: Positions compromis des individus (départements) dans le plan principal. Chaque département est identifié par son numéro.

Par exemple, les départements Bas-Rhin et Haut-Rhin (67 et 68) ont une coordonnée négative sur l'axe 1. Ces départements ont donc voté majoritairement pour la droite institutionnelle en moyenne sur la période étudiée. Au contraire, les départements de l'Ariège (09), de l'Aude (11), des Bouches-du-Rhône (13), de la Nièvre (58), du Pas-de-Calais (62), de la Haute-Vienne (87) et de la Seine-Saint-Denis (93) présentent des coordonnées positives, caractéristiques d'un vote majoritairement à gauche.

Sur l'axe 2, les coordonnées fortement négatives des départements des Alpes-Maritimes (06), des Bouches-du-Rhône (13), des Pyrénées-Orientales (66), de Paris (75), du Var (83) et de la Seine-Saint-Denis (93) indiquent que ces départements sont caractérisés par une non-participation et/ou un vote pour la droite protestataire important. Il serait utile de retourner aux données brutes pour être plus précis.

5.2.3.1.2.3 Trajectoires

Rappelons qu'une trajectoire de grande amplitude caractérise une modification importante de la structure des variables au cours du temps dans un département (relativement à l'évolution d'ensemble de la France), tandis qu'une trajectoire de petite amplitude est la marque d'une évolution stable par rapport à la moyenne de la France.

La Figure 5.23 représente les trajectoires de trois départements. Par exemple, le département du Haut-Rhin a une trajectoire située à gauche du plan. Cette position s'explique par la forte tradition du vote en faveur de la droite institutionnelle dans ce département. La trajectoire a une grande amplitude, ce qui caractérise une modification importante de la structure des variables au cours du temps pour ce département. La trajectoire descendante s'explique par l'émergence très forte du vote en faveur de la droite protestataire en 1988 au détriment de la droite institutionnelle.

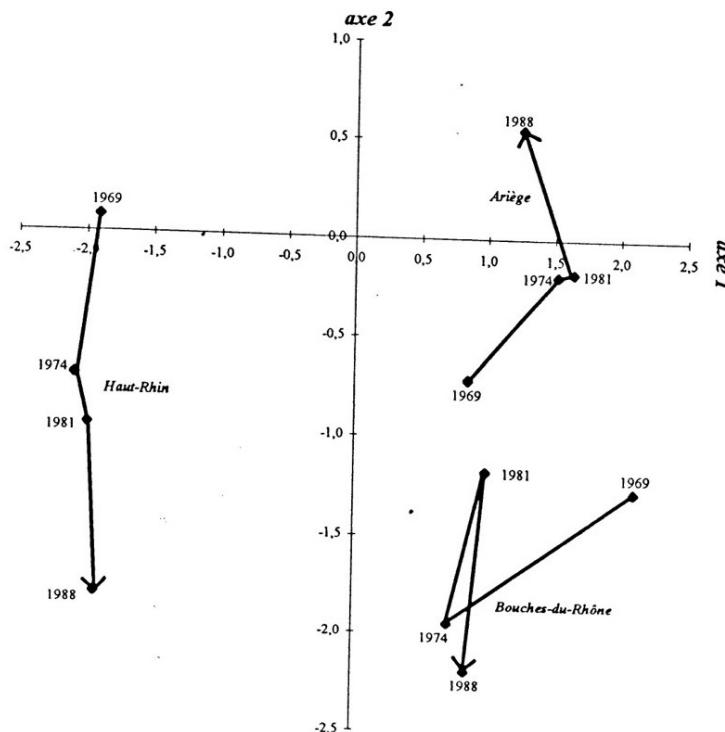


Figure 5.23: Trajectoires de trois départements (Haut-Rhin, Ariège et Bouches-du-Rhône) dans le plan principal. Chaque département est identifié par son numéro.

5.2.3.2 STATIS duale

La méthode STATIS duale est complètement semblable à la méthode STATIS techniquement. En pratique on applique cette méthode quand les individus diffèrent d'un tableau à l'autre tandis que les variables, elles, restent les mêmes, mais simplement mesurées à des temps différents. L'étude de l'interstructure s'appuie ici sur les matrices de variance-covariance, et non sur les matrices de produit-scalaire entre individus. Le compromis sera calculé sous forme d'une combinaison linéaire de ces matrices. La diagonalisation de celle-ci fournira une image approchée des variables (intrastructure). On pourra ensuite projeter les variables en tant qu'éléments supplémentaires afin d'obtenir les trajectoires des variables.

La méthode STATIS duale privilégie l'étude des variables, là où STATIS privilégie celle des individus. Dans le cas où les données forment un cube (comme pour la double ACP), on pourra utiliser les deux approches conjointement.

5.2.4 AFM

Dans le cas de variables quantitatives (l'AFM s'applique aussi sur des variables qualitatives), l'AFM consiste en une ACP particulière sur l'ensemble des tableaux disposés les uns à côté des autres. L'AFM s'applique donc uniquement pour des individus identiques d'un tableau. Cette pondération consiste diviser chaque variable la première valeur propre de l'ACP appliquée au bloc de variables auquel cette variable appartient. Cette méthode est à rapprocher de STATIS dans la mesure où les positions compromis des individus correspondent à celles obtenues si on avait appliqué une ACP sur l'ensemble des tableaux mais en pondérant cette fois par les coefficients $\sqrt{\alpha_t}$, où α_t correspond au coefficient de la combinaison linéaire associé à l'objet W_t . Dans STATIS, on accorde donc une influence plus forte aux groupes proches du compromis, tandis qu'en AFM on accorde une importance plus forte aux groupes dont les variables sont faiblement liées entre elles (i.e. des groupes multidimensionnels). Bien que ces pondérations soient assez différentes dans leur principe, en pratique, les configurations d'individus qui résultent sont très voisines, dès lors que les données possèdent une structure un tant soit peu forte (Pagès (2013)). En revanche, les représentations des groupes sont

généralement assez différentes. On pourra approfondir cette présentation de la méthode en consultant les supports de cours de F. Husson (http://math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/100459_AnDo_AFM_cours_Slides.pdf) et les vidéos 1 (<https://www.youtube.com/watch?v=1U-s8u1rcpo>)-2 (<https://www.youtube.com/watch?v=wCTaFaVKGAM>)-3 (<https://www.youtube.com/watch?v=abQllkEfleM>) associées.

Nous illustrons la méthode sur les données des élections. L'AFM s'effectue en commençant par les analyses séparées afin d'apprécier les pondérations de chaque groupe de variables. La Figure 5.24 rassemble les valeurs propres des différentes analyses. On voit que les premières valeurs propres sont assez proches pour les différents tableaux. Cette pondération n'aura donc pas une grande influence sur les résultats.

| Axes | Inertie | Pourcentage | Cumul |
|------|-------------|-------------|-------|
| 1 | 2,25 | 45,03 | 45,03 |
| 2 | 1,32 | 26,37 | 71,41 |
| 3 | 0,77 | 15,36 | 86,77 |
| 4 | 0,66 | 13,23 | 99,99 |

Tableau 1974

| Axes | Inertie | Pourcentage | Cumul |
|------|-------------|-------------|-------|
| 1 | 2,16 | 35,96 | 35,96 |
| 2 | 1,63 | 27,24 | 63,20 |
| 3 | 1,04 | 17,40 | 80,64 |
| 4 | 0,67 | 11,16 | 91,80 |

Tableau 1981

| Axes | Inertie | Pourcentage | Cumul |
|------|-------------|-------------|-------|
| 1 | 2,07 | 41,44 | 41,44 |
| 2 | 1,31 | 26,15 | 67,59 |
| 3 | 0,98 | 19,52 | 87,11 |
| 4 | 0,64 | 12,89 | 99,99 |

Tableau 1988

| Axes | Inertie | Pourcentage | Cumul |
|------|-------------|-------------|-------|
| 1 | 2,16 | 35,96 | 35,96 |
| 2 | 1,79 | 19,79 | 65,76 |
| 3 | 0,95 | 15,85 | 81,61 |
| 4 | 0,71 | 11,82 | 93,42 |

Figure 5.24: Valeurs propres des analyses séparées

Le compromis s'obtient en effectuant une ACP sur l'ensemble des tableaux en prenant en compte les pondérations précédentes. La Figure 5.25 rassemble les valeurs propres de cette ACP pondérée, définissant l'AFM.

| Axes | Inertie | Pourcentage | Cumul |
|------|-------------|-------------|-------|
| 1 | 3,54 | 34,77 | 34,77 |
| 2 | 2,30 | 22,60 | 57,37 |
| 3 | 1,16 | 11,36 | 68,73 |
| 4 | 0,84 | 8,20 | 76,92 |
| 5 | 0,50 | 4,95 | 81,87 |
| 6 | 0,45 | 4,39 | 86,26 |
| 7 | 0,34 | 3,36 | 89,62 |

Figure 5.25: Valeurs propres de l'AFM

A nouveau, l'examen de ces valeurs propres suggère d'analyser les 3 premiers axes. Nous nous limiterons aux

deux premiers.

L'analyse de ce compromis est effectuée en analysant les corrélations entre les variables initiales et les axes (Figure 5.26). Sans surprise, on arrive aux mêmes conclusions que pour la méthode STATIS (Section 5.2.3.1.2.2) : l'axe 1 oppose la droite institutionnelle à gauche et les gauches institutionnelle et protestataire à droite ; le second axe s'interprète comme un axe protestataire où la non-participation et la droite protestataire sont importantes. Les positions des individus sont également très proches de celles observées précédemment (graphique non présenté).

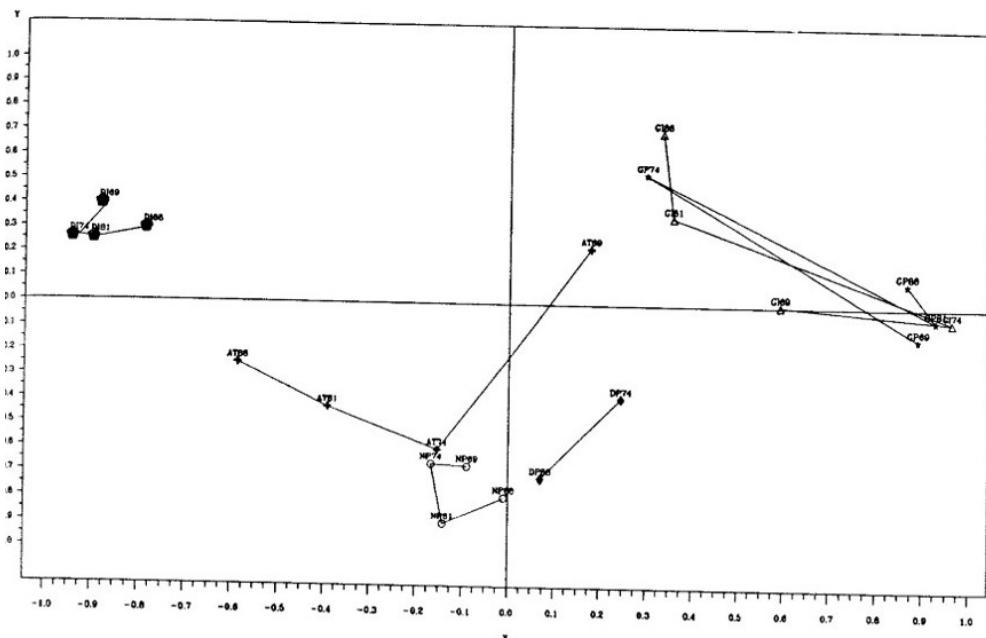


Figure 5.26: Corrélations des variables dans le plan principal

Une différence fondamentale entre STATIS et l'AFM, est due au fait que les axes de l'interstructure sont interprétables en AFM, alors qu'ils ne le sont pas dans STATIS. Dans la représentation fournie en AFM, la représentation d'un tableau sur un axe représente une liaison entre le tableau et la composante associée à l'axe. Ainsi, plus la coordonnée d'un tableau est forte, plus la composante est caractéristique de ce tableau. En revanche, la qualité de représentation des tableaux est moins bonne que pour STATIS (Dazy et al. (1996)). La représentation de l'interstructure est présentée en Figure 5.27. De par l'interprétation préalable des axes et la position des tableaux sur cette figure, on peut dire que les années 1981 et 1984 ont été marquées par un vote dominant en faveur des formations institutionnelles (droite ou gauche). A l'inverse, les élections de 1960 et 1988 sont caractérisées par une position moins dominante de ces formations. Par ailleurs, au cours du temps, les tableaux sont de plus en plus situés en haut du plan. Ceci signifie que globalement, la droite protestataire et la non-participation ont davantage marqué les scrutins.

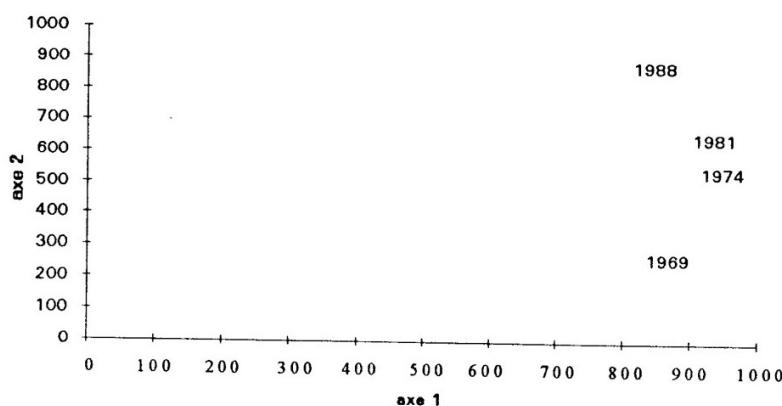


Figure 5.27: Représentation de l'interstructure dans le plan principal

Enfin, la Figure 5.28 représente les trajectoires de trois départements. Celles-ci ressemblent beaucoup à celles de la Figure 5.23 et, du fait de l'interprétation des axes également similaire, amène aux mêmes commentaires.

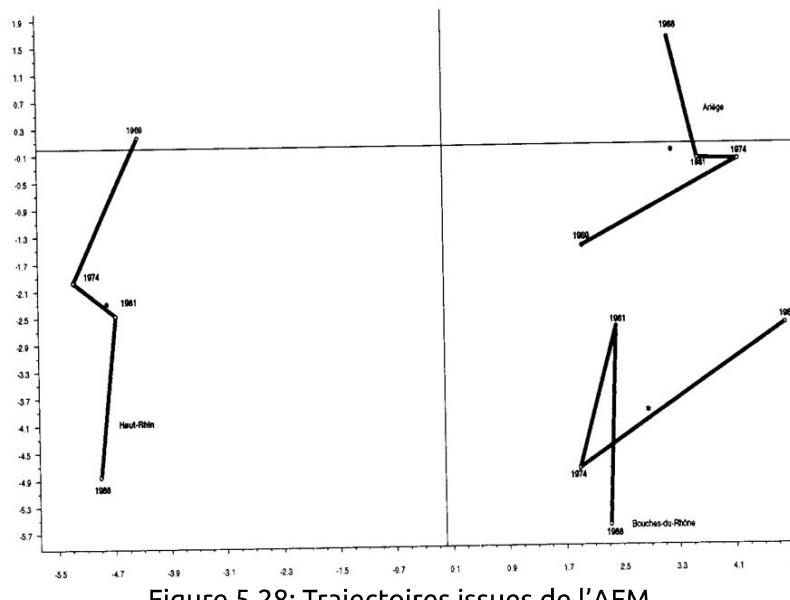


Figure 5.28: Trajectoires issues de l'AFM

5.3 Conclusion

Nous avons présenté 3 méthodes non supervisées dédiées à l'analyse des données multi-blocs. En dehors des considérations techniques, les différences les plus importantes entre ces méthodes sont liées aux données auxquelles elles s'appliquent. Il faut prendre en considération la nature des variables (quantitative ou qualitative) et les dimensions de chacun des tableaux (nombre d'individus ou de variables en fonction des tableaux). La double ACP s'applique dans un cadre bien particulier où les données peuvent être représentées sous la forme d'un cube. STATIS et AFM permettent d'avoir des tableaux avec des variables différentes en nombre. La méthode STATIS duale permettra de gérer des tableaux avec des individus différents. Nous avons essentiellement évoqué des tableaux différant par une dimension temporelle, mais ceci n'est pas une nécessité.

La prise en compte de cet aspect multi-blocs est aussi important dans un contexte supervisé. Dans ce cas, on pourra avoir recours à la régression multi-blocs dans un cadre de régression, ou à l'analyse discriminante multi-tableaux dans un cadre de classification.

Références

- Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. 1993. "Mining Association Rules Between Sets of Items in Large Databases." In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207–16. SIGMOD '93. New York, NY, USA: ACM. <http://doi.acm.org/10.1145/170035.170072> (<http://doi.acm.org/10.1145/170035.170072>).
- Agrawal, Rakesh, Ramakrishnan Srikant, and others. 1994. "Fast Algorithms for Mining Association Rules." In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, 1215:487–99.
- Bouroche, Jean-Marie. 1975. "Analyse Des Données Ternaires: La Double Analyse En Composantes Principales." PhD thesis.

- Brin, Sergey, Rajeev Motwani, Jeffrey D Ullman, and Shalom Tsur. 1997. "Dynamic Itemset Counting and Implication Rules for Market Basket Data." *Acm Sigmod Record* 26 (2). ACM: 255–64.
- Charrad, Malika, and Mohamed Ben Ahmed. 2011. "Simultaneous Clustering : a survey." In *4th International Conference on Pattern Recognition and Machine Intelligence. PReMI 2011*, LNCS 6744:370–75. Springer. Moscow, Russia. <https://hal.archives-ouvertes.fr/hal-01125890> (<https://hal.archives-ouvertes.fr/hal-01125890>).
- Chavent, Marie, Vanessa Kuentz-Simonet, Benoît Liquet, and Jérôme Saracco. 2012. "ClustOfVar: An R Package for the Clustering of Variables." *Journal of Statistical Software, Articles* 50 (13): 1–16. doi:10.18637/jss.v050.i13 (<https://doi.org/10.18637/jss.v050.i13>).
- Chen, Ning, and Nuno C. Marques. 2005. "An Extension of Self-Organizing Maps to Categorical Data." In *Progress in Artificial Intelligence*, edited by Carlos Bento, Amílcar Cardoso, and Gaël Dias, 304–13. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Cottrell, Marie, Smail Ibbou, and Patrick Letrémy. 2004. "SOM-Based Algorithms for Qualitative Variables." *Neural Networks* 17 (8-9). Elsevier: 1149–67.
- Dazy, Frédéric, Jean-François Le Barzic, Gilbert Saporta, and Françoise Lavallard. 1996. *L'analyse Des Données évolutives: Méthodes et Applications*. Editions technip.
- Fichet, B., and G. le Calve. 1984. "Structure géométrique des principaux indices de dissimilarité sur signes de présence-absence." *Stat. Anal. Données* 9 (3). Association des Statisticiens Universitaires, A.S.U., Paris: 11–44.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc. <https://web.stanford.edu/~hastie/Papers/ESLII.pdf> (<https://web.stanford.edu/~hastie/Papers/ESLII.pdf>).
- Kroonenberg, Pieter M. 1983. *Three-Mode Principal Component Analysis: Theory and Applications*. Vol. 2. DSWO press.
- Lavit, Christine, Yves Escoufier, Robert Sabatier, Pierre Traissac, and others. 1994. "The Act (Statis Method)." *Computational Statistics and Data Analysis* 18 (1). Amsterdam: North-Holland Pub. Co.: 97–120.
- Lebbah, Mustapha, Fouad Badran, and Sylvie Thiria. 2000. "Topological Map for Binary Data." In *ESANN*, 267–72.
- Lebbah, Mustapha, Aymeric Chazottes, Fouad Badran, and Sylvie Thiria. 2005. "Mixed Topological Map." In *ESANN*, 17:47.
- L'Hermier des Plantes, Henri. 1976. "Structuration Des Tableaux à Trois Indices de La Statistique: Théorie et Application d'une Méthode d'analyse Conjointe." PhD thesis, Université des sciences et techniques du Languedoc.
- Pagès, Jérôme. 2013. *Analyse Factorielle Multiple Avec R*. EDP sciences.
- Plasse, Marie. 2006. "Utilisation Conjointe Des Méthodes de Recherche de Règles d'association et de Classification: Contribution à L'amélioration de La Qualité Des Véhicules En Production Grâce à L'exploitation Des Systèmes d'information." PhD thesis, Paris, CNAM.
- Saporta, G. 2006. *Probabilités, Analyse Des Données et Statistique*. Editions Technip.
- Sarle, WS. 1990. "The Varclus Procedure. Sas/Stat User's Guide. Sas Institute." Inc., Cary, NC, USA.
- Savasere, Ashok, Edward Robert Omiecinski, and Shamkant B Navathe. 1995. "An Efficient Algorithm for

Mining Association Rules in Large Databases." Georgia Institute of Technology.

Toivonen, Hannu, and others. 1996. "Sampling Large Databases for Association Rules." In *VLDB*, 96:134–45.

Vigneau, Evelyne, and EM Qannari. 2003. "Clustering of Variables Around Latent Components."

Communications in Statistics-Simulation and Computation 32 (4). Taylor & Francis: 1131–50.

Zaki, Mohammed Javeed. 2000. "Scalable Algorithms for Association Mining." *IEEE Transactions on Knowledge and Data Engineering* 12 (3). IEEE: 372–90.