

STA211 : Analyse multi-blocs

Vincent Audigier, Ndeye Niang

20 mars, 2025

- 1 Introduction
- 2 Objectifs de l'analyse multi-blocs
- 3 Méthodes
 - 3.1 Notations
 - 3.2 Double ACP
 - 3.2.1 Méthode
 - 3.2.2 Exemple
 - 3.3 STATIS/STATIS duale
 - 3.3.1 STATIS
 - 3.3.2 STATIS duale
 - 3.4 AFM
- 4 Analyse multi-blocs sous R
 - 4.1 Données
 - 4.2 AFM
 - 4.3 STATIS
- 5 Conclusion
- Références

1 Introduction

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. 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.

2 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 (deux 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 résumant les données globalement et non résumant 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.

3 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)).

3.1 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 3.1 et 3.2). 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.

$$X_t = \begin{matrix} & \begin{matrix} 1 & 2 & \vdots & n \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ \vdots \\ n \end{matrix} & \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{p_t 1} \\ X_{12} & X_{22} & \cdots & X_{p_t 2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{p_t n} \end{bmatrix} \end{matrix}$$

Figure 3.1: Schéma d'un tableau (i.e. un bloc).

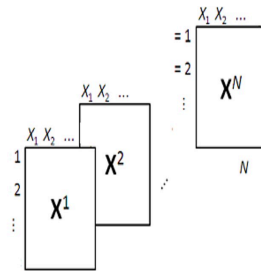


Figure 3.2: Schéma d'un ensemble de tableaux (i.e. données multi-blocs).

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 V_t la matrice de variance-covariance associée au tableau X_t .

3.2 Double ACP

3.2.1 Méthode

La double ACP (Bouroche (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 $Max_{(v_1, \dots, v_q)} \sum_{k=1}^q v'_k V v_k$ 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.

3.2.2 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.

3.2.2.1 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 3.3: 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 3.4: Valeurs propres de l'interstructure déterminés par la DACP

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

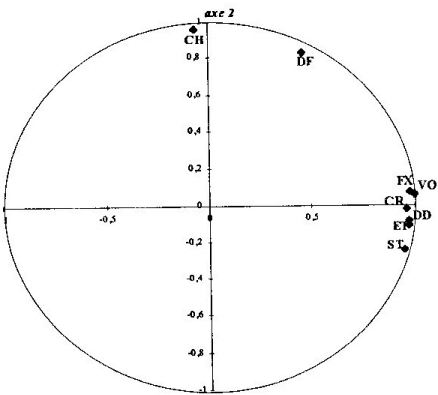


Figure 3.5: Cercle des corrélations pour le plan 1-2 (interstructure de la DACP)

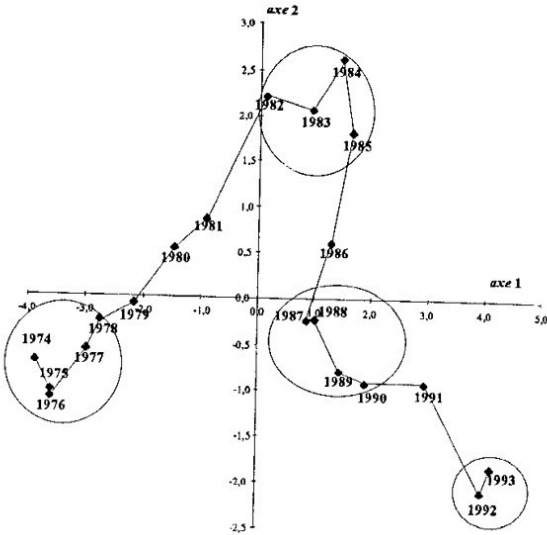


Figure 3.6: Graphe des individus plan 1-2 (interstructure de la DACP)

On observe sur le cercle des corrélations Figure 3.5) 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 3.6) 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 3.5) varient de façon linéaire par rapport au temps. Ceci est confirmé par les Figures 3.7 et 3.8 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.

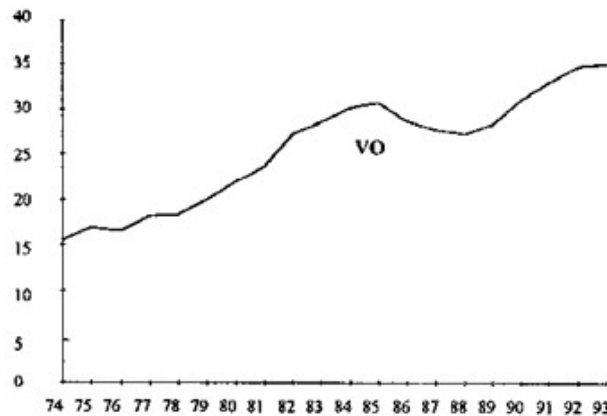


Figure 3.7: Evolution de la variable VO (vols et recels) en France métropolitaine

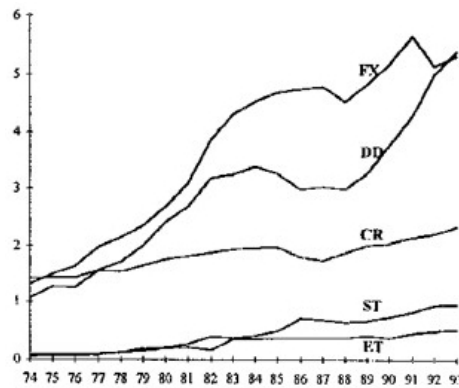


Figure 3.8: Evolution des variables ST, ET, FX, CR et DD en France métropolitaine

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 3.9.

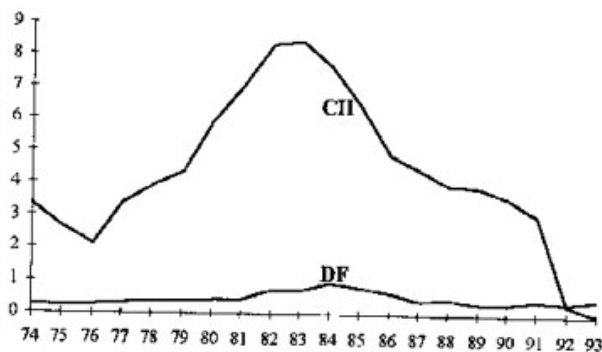


Figure 3.9: Evolution des variables CH et DF en France métropolitaine

La représentation de l'interstructure sur la Figure 3.6 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.

3.2.2.2 Intrastructure

L'intrastructure 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 3.10). 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 3.10: Valeurs propres de l'intrastructure (DACP)

L'interprétation des axes se fait à l'aide de l'examen des corrélations avec les variables ``compromis'' (Figure 3.11).

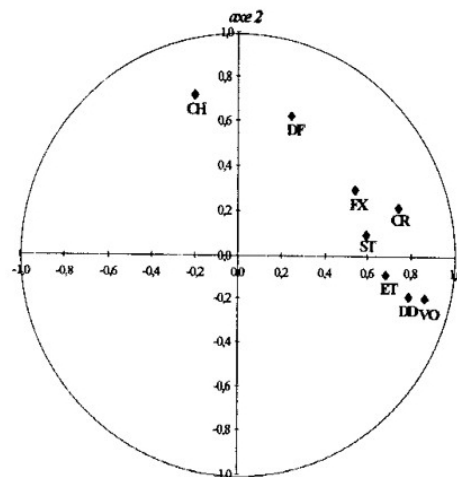


Figure 3.11: Cercle des corrélations pour le plan 1-2 (intrastructure de la DACP)

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.

3.2.2.3 Trajectoires

Les trajectoires s'obtiennent en projetant simultanément les individus de chaque tableau sur le plan principal. La Figure 3.12 représente 3 de ces trajectoires.

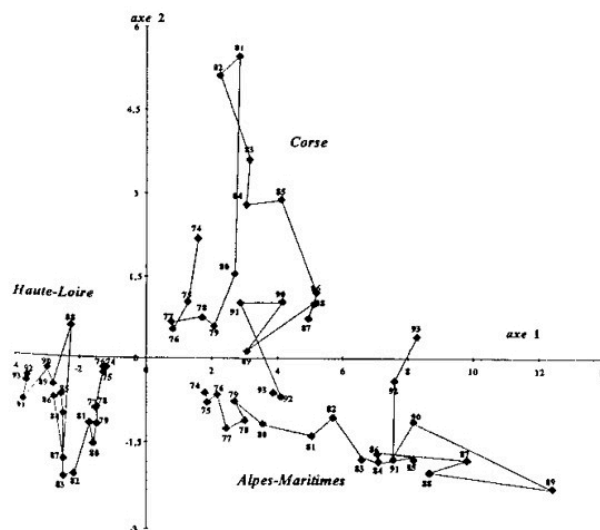


Figure 3.12: Trajectoires des départements Haute-Loire, Alpes-Maritimes et Corse.

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 a 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.

3.3 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'évoquons que brièvement.

3.3.1 STATIS

3.3.1.1 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 $W_t = X_t M_t X_t^T$, 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

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

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 pour lesquels les distances entre individus sont identiques à un facteur près.

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 des objets. Soit λ_k , la k -ième valeur propre de la matrice produit $S\Delta$ 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 3.13.

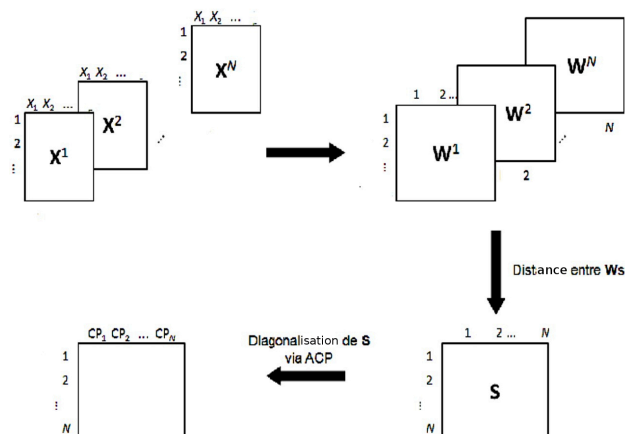


Figure 3.13: Résumé de l'étude interstructure.

L'étude de l'intrastructure 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 3.14.

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 compromises des individus.

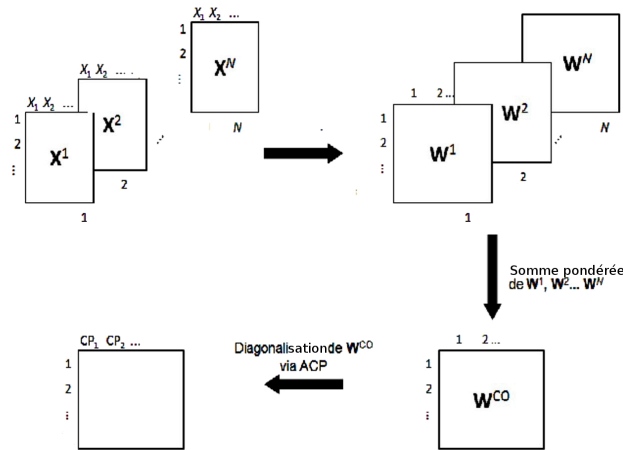


Figure 3.14: Résumé de la construction du compromis.

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.

3.3.1.2 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 3.15 et 3.16. 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 Héraut + 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 Bouchardeau
 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 3.15: Présentation des données.

Tableau 3.1 Statistiques élémentaires sur les tableaux de données

| Scrutin 1969 | | | | | | | | |
|--------------|------|-------|-------|------|-------|-------|-------|-------|
| (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 |

| Scrutin 1974 | | | | | | | | | |
|--------------|-------|-------|------|------|------|-------|-------|-------|-------|
| (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 |

| Scrutin 1981 | | | | | | | | |
|--------------|-------|-------|-------|------|-------|-------|-------|-------|
| (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 |

| Scrutin 1988 | | | | | | | | | |
|--------------|-------|-------|------|-------|------|-------|-------|-------|-------|
| (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 3.16: Statistiques élémentaires sur les tableaux de données.

3.3.1.2.1 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 3.17).

| 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 3.17: 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 évoluée entre 1974 et 1988.

L'analyse des valeurs propres suite à la diagonalisation de la matrice des RV (Figure 3.18) 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 3.18: Valeurs propres de l'interstructure.

Le premier plan factoriel (Figure 3.19) 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.

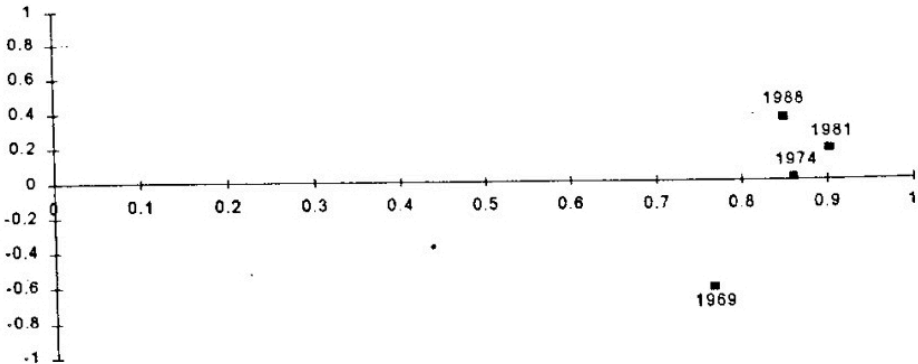


Figure 3.19: Représentation de l'interstructure.

3.3.1.2.2 Intrastructure

Les ressemblances et différences entre les structures des départements ayant été mises en évidence, on construit le tableau "moyen", ou compromis, 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 3.20 ($D = \frac{1}{95} I_{95}$ 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 3.20: Valeurs propres de l'intrastructure.

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 3.21.

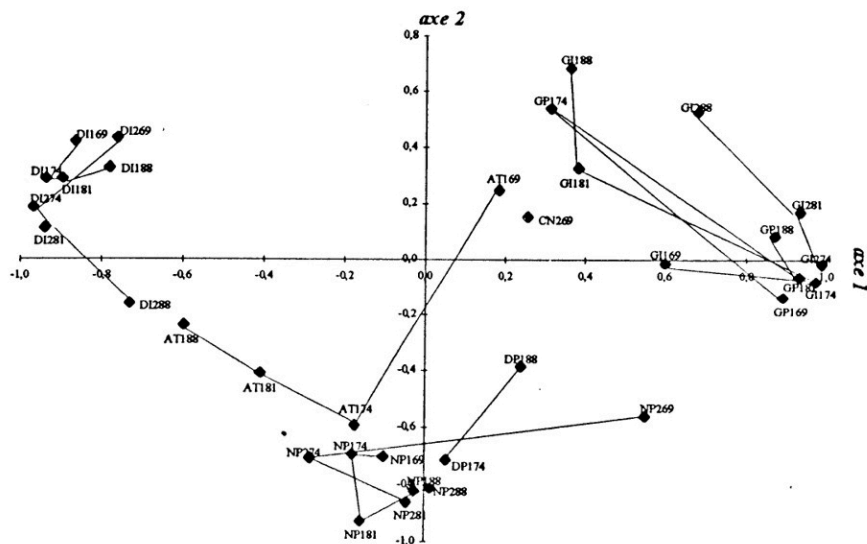


Figure 3.21: Corrélations des variables dans le plan principal.

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 comprises des individus dans le plan principal (cf Figure 3.22). Notons que les points compris correspondent à des positions moyennes des départements sur la période étudiée.

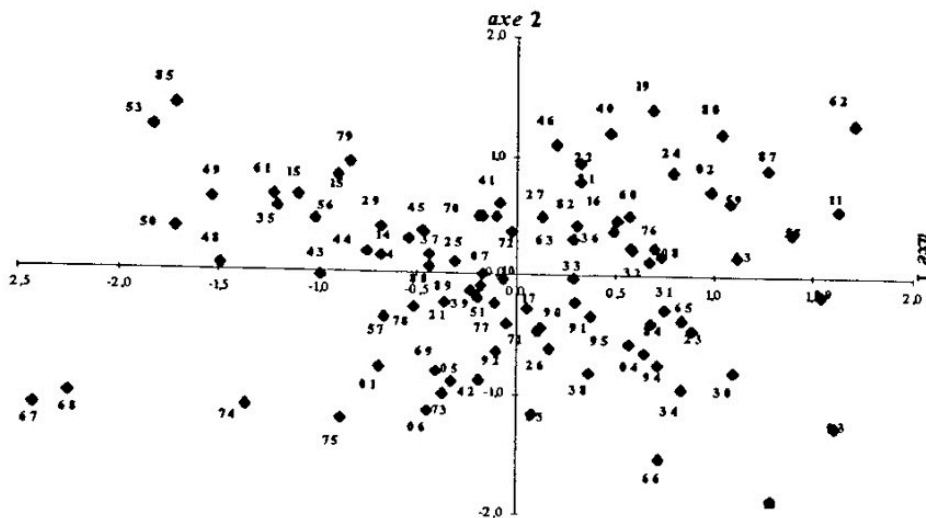


Figure 3.22: Positions comprises des individus (départements) dans le plan principal. Chaque département est identifié par son numéro.

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 le préciser.

3.3.1.2.3 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 3.23 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.

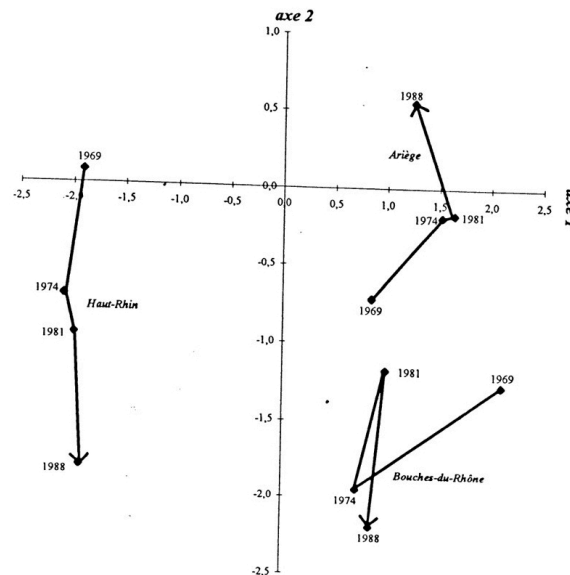


Figure 3.23: Trajectoires de trois départements (Haut-Rhin, Ariège et Bouches-du-Rhône) dans le plan principal. Chaque département est identifié par son numéro.

3.3.2 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.

3.4 AFM

Dans le cas de variables quantitatives (l'AFM s'applique aussi sur des variables qualitatives ou mixtes), 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 à 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 compromises 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 (https://husson.github.io/MOOC_AnaDo/AFM.html) 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 3.24 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 3.24: 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 3.25 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 3.25: 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 3.26). Sans surprise, on arrive aux mêmes conclusions que pour la méthode STATIS (Section 3.3.1.2.2) : 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).

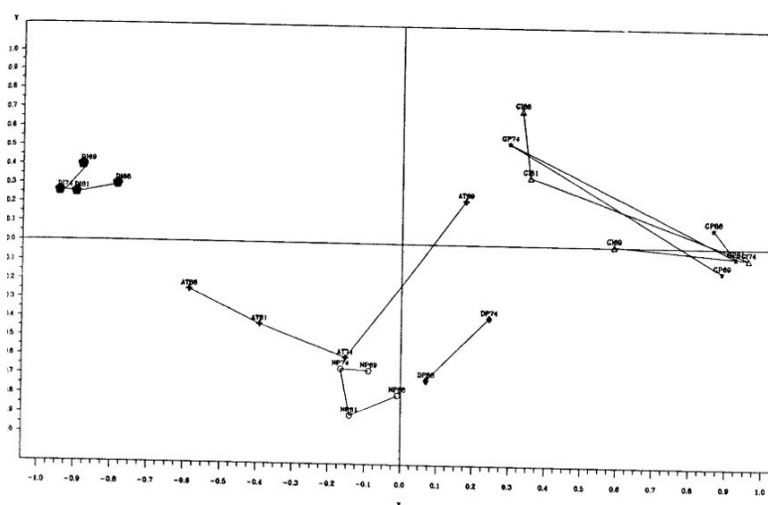


Figure 3.26: Corrélation des variables dans le plan principal

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 3.27. 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.

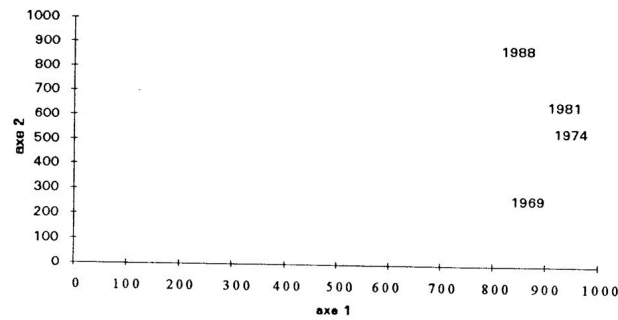


Figure 3.27: Représentation de l'interstructure dans le plan principal

Enfin, la Figure 3.28 représente les trajectoires de trois départements. Celles-ci ressemblent beaucoup à celles de la Figure 3.23 et, du fait de l'interprétation des axes également similaire, amène aux mêmes commentaires.

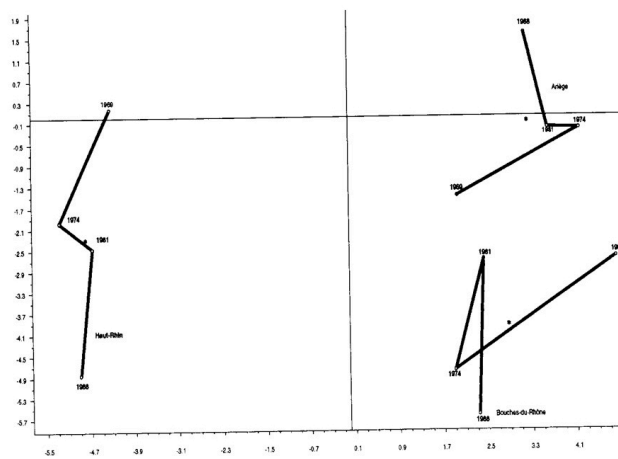


Figure 3.28: Trajectoires issues de l'AFM

4 Analyse multi-blocs sous R

4.1 Données

Nous reprenons ici un exemple issu de Pagès (2013).

```
#importation des données "à la main"
```

```
orange <- structure(
  list(
    Glucose = c(25.32, 17.33, 23.65, 32.42, 22.7, 27.16),
    Fructose = c(27.36, 20, 25.65, 34.54, 25.32, 29.48),
    Saccharose = c(36.45, 44.15, 52.12, 22.92, 45.8, 38.94),
    pH_brut = c(3.59, 3.89, 3.85, 3.6, 3.82, 3.68),
    pH_apres_centrifuga. = c(3.55, 3.84, 3.81, 3.58, 3.78, 3.66),
    Titre = c(13.98, 11.14, 11.51, 15.75, 11.8, 12.21),
    Acide_citrique = c(0.84, 0.67, 0.69, 0.95, 0.71, 0.74),
    Vitamine_C = c(43.44, 32.7, 37, 36.6, 39.5, 27),
    Intensite_odeur = c(2.82, 2.76, 2.83, 2.76, 3.2, 3.07),
    Typicite_odeur = c(2.53, 2.82, 2.88, 2.59, 3.02, 2.73),
    Pulpeux = c(1.66, 1.91, 4, 1.66, 3.69, 3.34),
    Intensite_gout = c(3.46, 3.23, 3.45, 3.37, 3.12, 3.54),
    Acide = c(3.15, 2.55, 2.42, 3.05, 2.33, 3.31),
    Amer = c(2.97, 2.08, 1.76, 2.56, 1.97, 2.63),
    Sucre = c(2.6, 3.32, 3.38, 2.8, 3.34, 2.9)
  ),
  class = "data.frame",
  row.names = c(
    "P1 Pampryl amb.",
    "P2 Tropicana amb.",
    "P3 Fruivita fr.",
    "P4 Joker amb.",
    "P5 Tropicana fr.",
    "P6 Pampryl fr."
  )
)
```

```
## visualisation des données
View(orange)
```

| | Glucose | Fructose | Saccharose | pH brut | pH apres centrifuga. | Titre | Acide_citrique | Vitamine_C | Intensite_odeur | Typicite_odeur | Pulpeux | Intensite_gout | Acide | Amer | Sucre |
|-------------------------|---------|----------|------------|---------|----------------------|-------|----------------|------------|-----------------|----------------|---------|----------------|-------|------|-------|
| P1 Pampryl amb. | 25.32 | 27.36 | 36.45 | 3.59 | 3.55 | 13.98 | 0.84 | 43.44 | 2.82 | 2.53 | 1.66 | 3.46 | 3.15 | 2.97 | 2.60 |
| P2 Tropicana amb. | 17.33 | 20.00 | 44.15 | 3.89 | 3.84 | 11.14 | 0.67 | 32.70 | 2.76 | 2.82 | 1.91 | 3.23 | 2.55 | 2.08 | 3.32 |
| P3 Fruivita fr. | 23.65 | 25.65 | 52.12 | 3.85 | 3.81 | 11.51 | 0.69 | 37.00 | 2.83 | 2.88 | 4.00 | 3.45 | 2.42 | 1.76 | 3.38 |
| P4 Joker amb. | 32.42 | 34.54 | 22.92 | 3.60 | 3.58 | 15.75 | 0.95 | 36.60 | 2.76 | 2.59 | 1.66 | 3.37 | 3.05 | 2.56 | 2.80 |
| P5 Tropicana fr. | 22.70 | 25.32 | 45.80 | 3.82 | 3.78 | 11.80 | 0.71 | 39.50 | 3.20 | 3.02 | 3.69 | 3.12 | 2.33 | 1.97 | 3.34 |
| P6 Pampryl fr. | 27.16 | 29.48 | 38.94 | 3.68 | 3.66 | 12.21 | 0.74 | 27.00 | 3.07 | 2.73 | 3.34 | 3.54 | 3.31 | 2.63 | 2.90 |

Le jeu de données porte sur 6 jus d'orange Pampryl amb., Tropicana amb., Fruivita fr., Joker amb., Tropicana fr., Pampryl fr.. Le suffixe 'fr' indique un jus d'orange servi frais, tandis que le suffixe 'amb' désigne un jus servi à température ambiante. Sur ces différents jus, on s'intéresse à deux types de descripteurs : des descripteurs chimiques (Glucose, Fructose, Saccharose, pH_brut, pH_apres_centrifuga., Titre, Acide_citrique, Vitamine_C) et des descripteurs sensoriels correspondant à une note moyenne suite à l'évaluation par un panel de 96 juges (Intensite_odeur, Typicite_odeur, Pulpeux, Intensite_gout, Acide, Amer, Sucre). Nous avons donc ici deux blocs de variables dont les effectifs diffèrent mais portant sur des mêmes individus. Ainsi, nous illustrons comment mettre en oeuvre une AFM ou une analyse STATIS sur ces données à l'aide des packages FactoMineR ou ade4 respectivement. L'interprétation d'une analyse multi-blocs ayant déjà fait l'objet des sections précédentes, nous reléguons le lecteur à Pagès (2013) pour l'interprétation des sorties fournies et nous focalisons sur leur obtention.

4.2 AFM

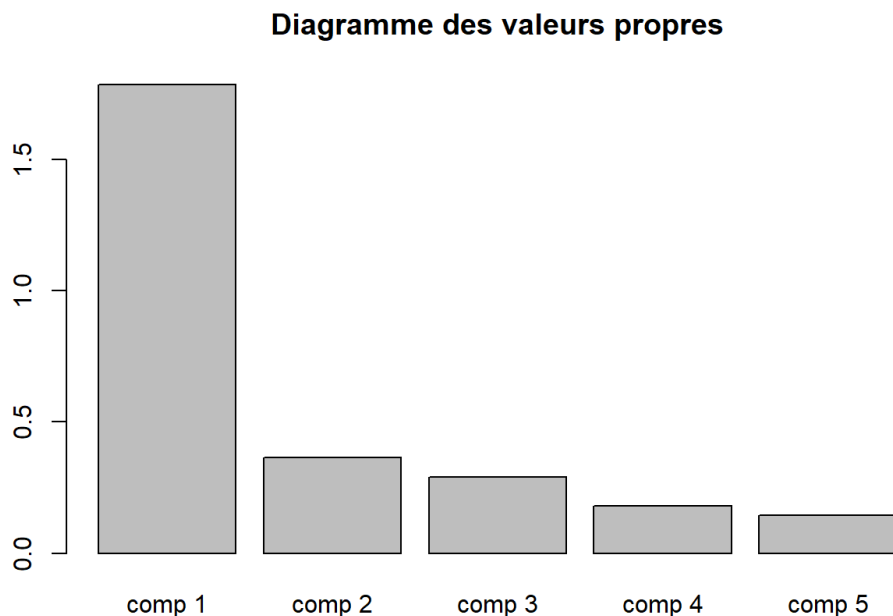
```
# AFM
library(FactoMineR)
res.afm <- MFA(
  orange,
  group = c(8, 7),
  type = c("s", "s"),# "s" pour standardiser
  name.group = c("Chimie", "Sensoriel"),
  graph = FALSE
)

# Graphiques

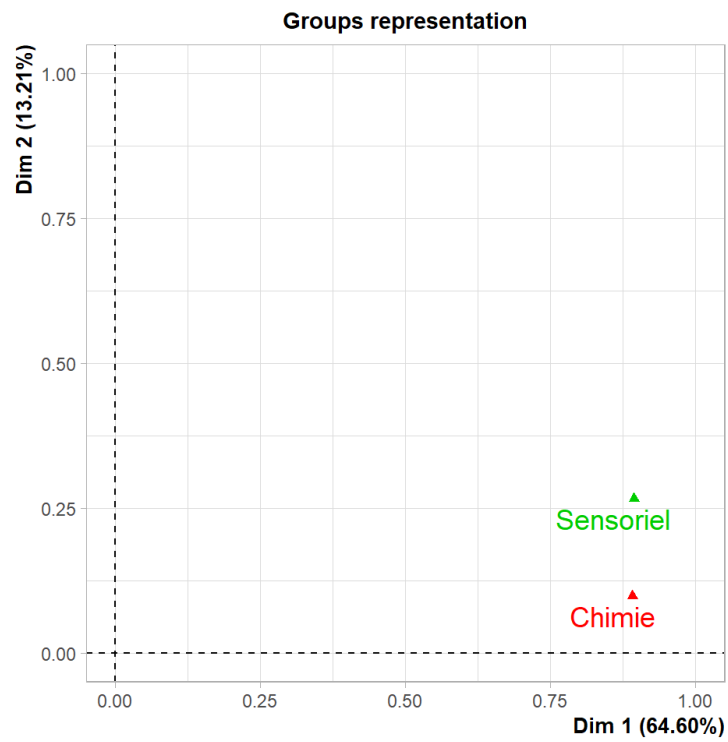
## Analyses séparées
lapply(res.afm$separate.analyses, "[", "eig")
```

```
## $Chimie
##      eigenvalue percentage of variance cumulative percentage of variance
## comp 1 6.21248019          77.6560024          77.65600
## comp 2 1.09951112          13.7438890          91.39989
## comp 3 0.36641085           4.5801356          95.98003
## comp 4 0.30941980           3.8677475          99.84777
## comp 5 0.01217804           0.1522255         100.00000
##
## $Sensoriel
##      eigenvalue percentage of variance cumulative percentage of variance
## comp 1 4.74369269          67.7670384          67.76704
## comp 2 1.33328986          19.0469979          86.81404
## comp 3 0.81984115          11.7120164          98.52605
## comp 4 0.08402330           1.2003328          99.72639
## comp 5 0.01915301           0.2736144         100.00000
```

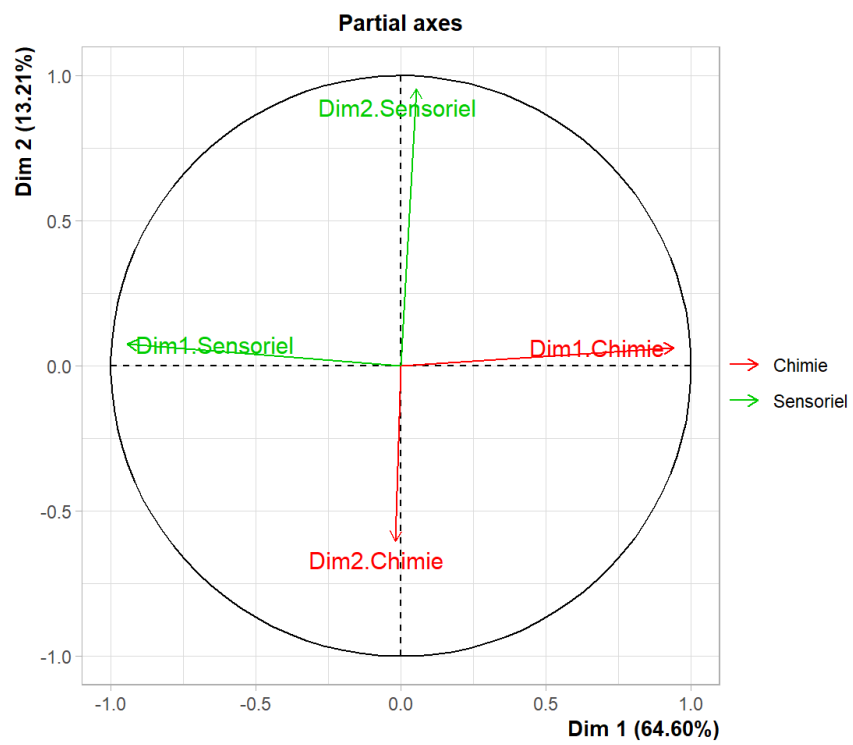
```
## Diagramme des valeurs propres de l'AFM
barplot(res.afm$eig[, 1], main = "Diagramme des valeurs propres")
```



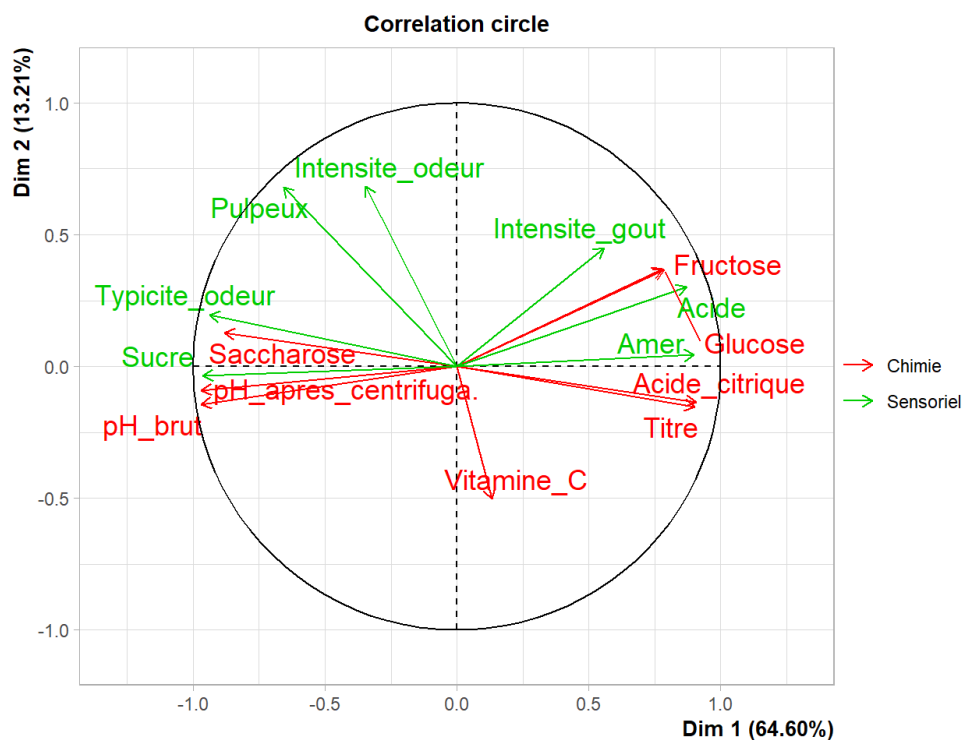
```
## Interstructure
plot(res.afm, choix = "group", cex = 1.2)
```



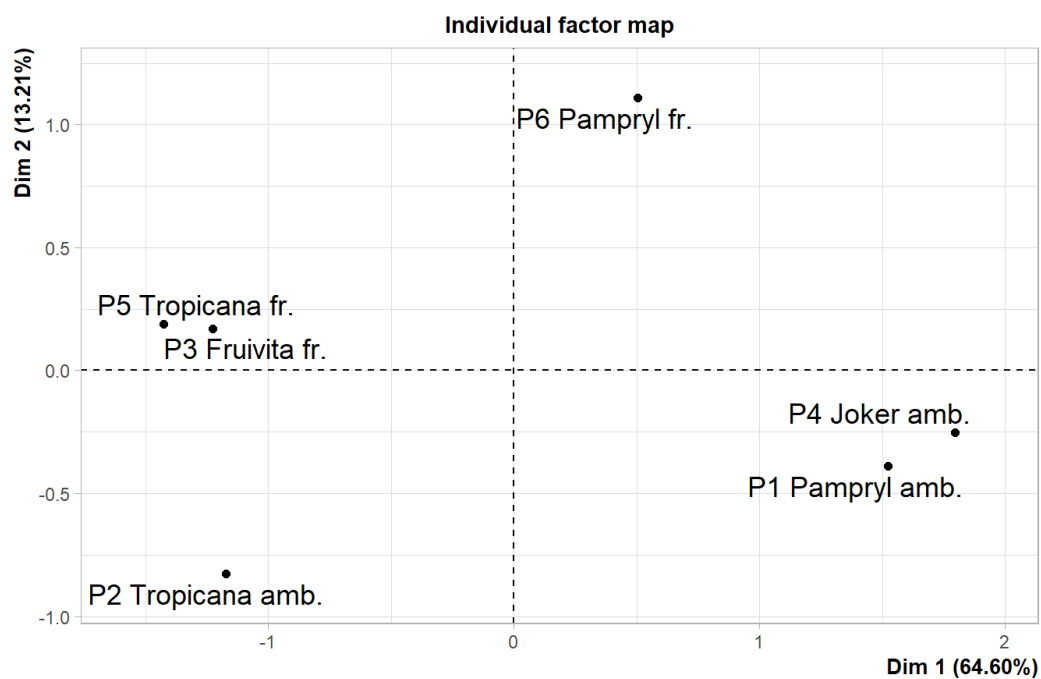
```
plot(res.afm, choix = "axes")# corrélations entre les composantes des analyses séparées et celles de L'AFM
```



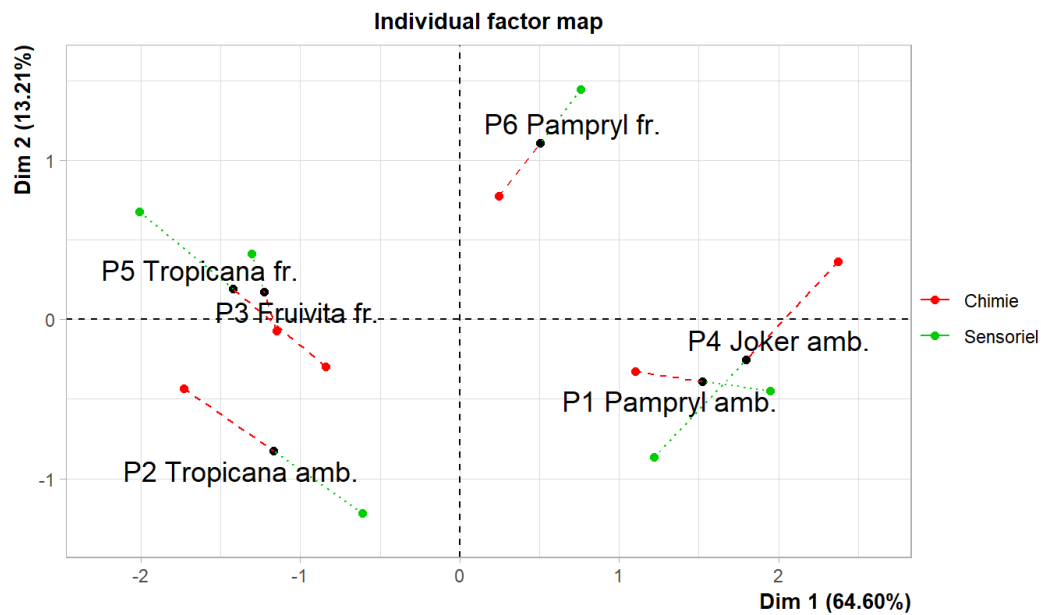
```
## Analyse du compromis
plot(res.afm, choix = "var", cex = 1.2)
```

```
plot(res.afm, choix = "ind", cex = 1.2)
```



```
## Trajectoires
plot(res.afm, choix = "ind", partial = "all", cex = 1.2)
```



4.3 STATIS

```
# STATIS
library(ade4)
library(adegraphics)

#####
## Mise en forme des données
#####

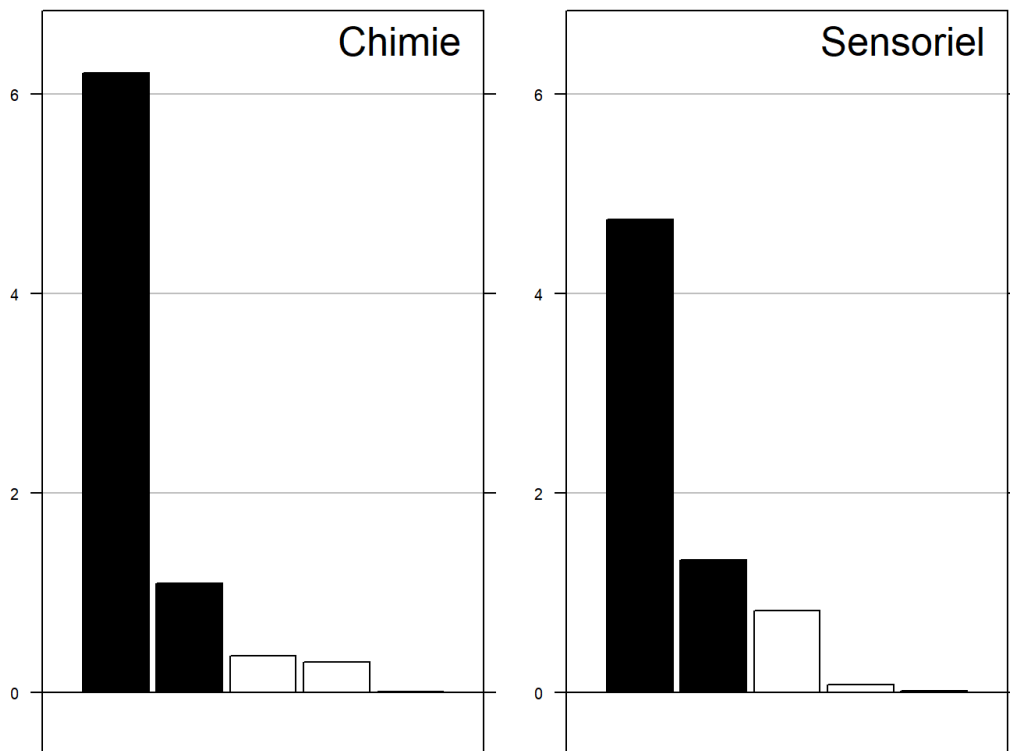
# Centrage et réduction des données
df <- as.data.frame(scalewt(orange))
blocks <- c(8, 7) #taille des groupes
# On crée un objet ktab avec la fonction ktab.data.frame
tab <- ktab.data.frame(
  df,
  blocks = blocks,
  tabnames = c("Chimie", "Sensoriel"),
  rownames = rownames(df)
)

#####
## Mise en oeuvre de l'analyse
#####

#Par défaut, Le paramètre scannf vaut TRUE et permet un choix interactif du nombre d'axes analyser. Nous préférons fixer c
e nombre a priori à l'aide de l'argument nf.
res.statis <- statis(tab, scannf = FALSE, nf = 3)

#####
## Graphiques et sorties
#####

# Analyses séparées
sep <- sepan(tab)
plot(sep, psub.cex = 2)
```



```
# Interstructure
```

```
## Tableau des RV
```

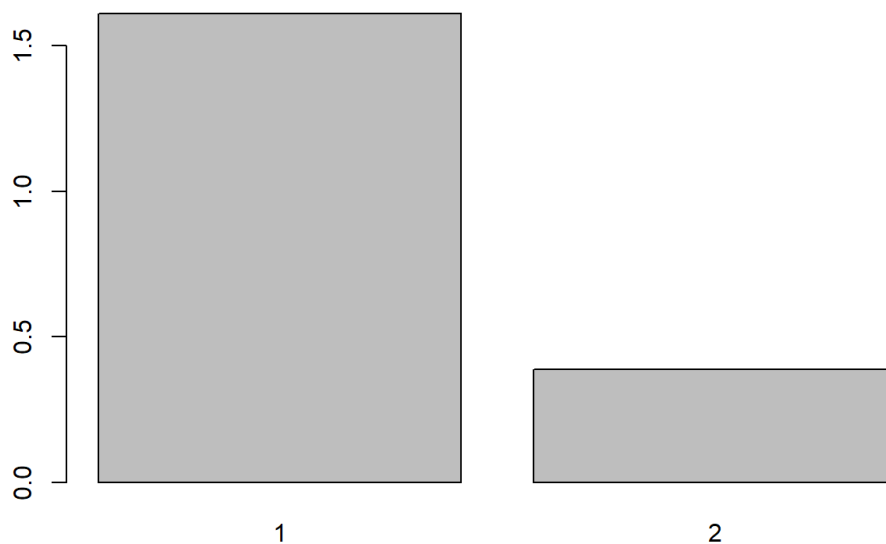
```
tabrv <- rbind(res.statis$RV, t(res.statis$RV.coo[, 1]))
rownames(tabrv)[3] <- "Compromis"
round(tabrv, 3)
```

```
##          Chimie Sensoriel
## Chimie   1.000   0.611
## Sensoriel 0.611   1.000
## Compromis 0.897   0.897
```

```
## Diagrammes des valeurs propres
```

```
barplot(res.statis$RV.eig,
        names.arg = seq.int(length(res.statis$RV.eig)),
        main = "Diagramme des valeurs propres de l'interstructure")
```

Diagramme des valeurs propres de l'interstructure

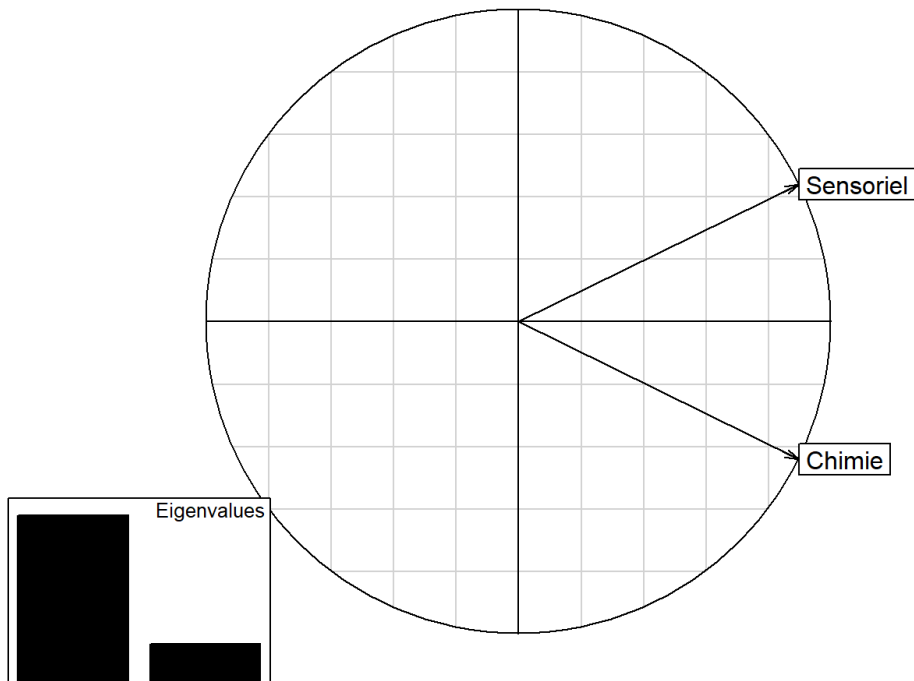


```
## Poids des tableaux  
res.statis$RV.tabw
```

```
## [1] 0.7071068 0.7071068
```

```
## Représentation de l'interstructure  
ade4::plot.statis(res.statis,  
  xax = 1,  
  yax = 2,  
  option = 1)
```

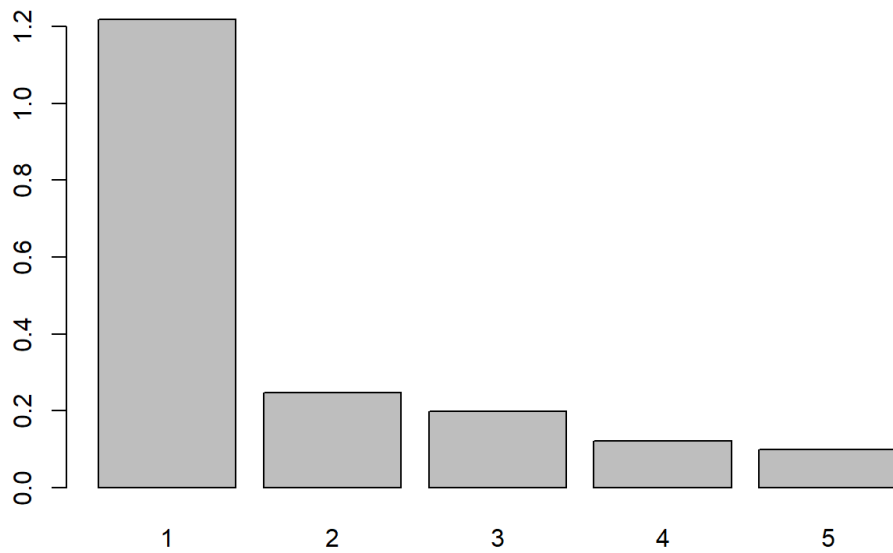
Interstructure



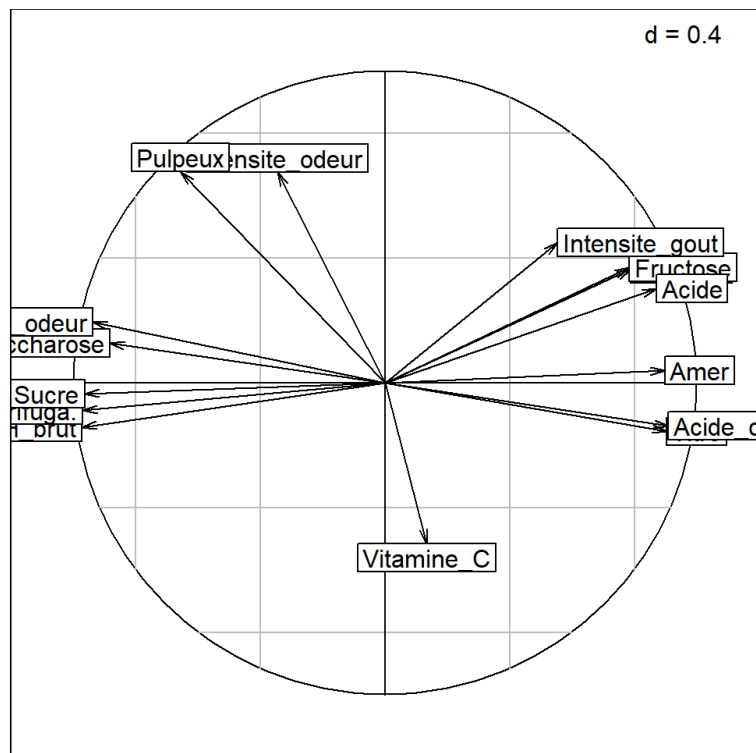
```
# Intrastructure
```

```
## Diagrammes des valeurs propres  
barplot(res.statis$C.eig,  
  names.arg = seq.int(length(res.statis$C.eig)),  
  main = "Diagramme des valeurs propres de l'intrastructure")
```

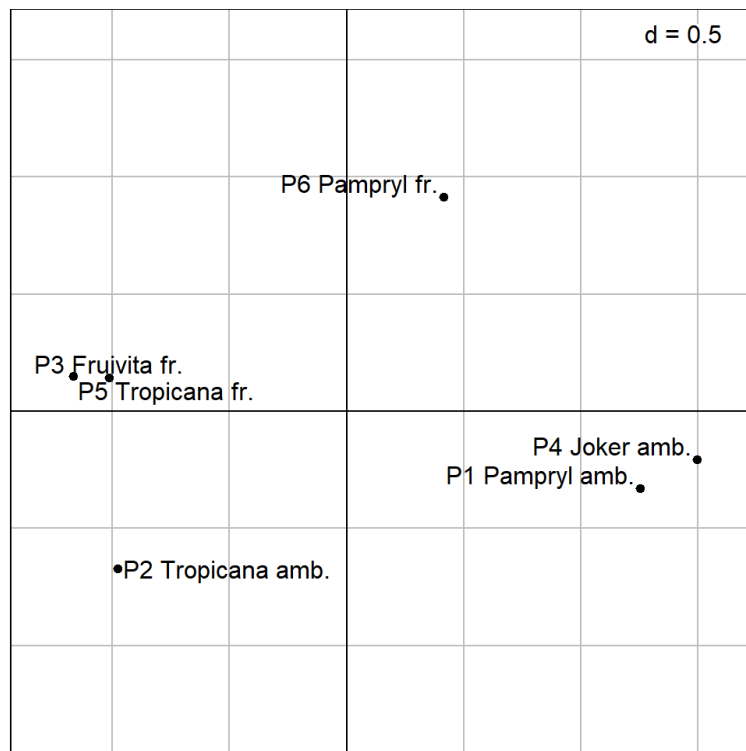
Diagramme des valeurs propres de l'intrastructure



```
## Correlations des variables
s.corcircle(res.statistC.Co)
```



```
## Coordonnées des individus compromis
s.label(res.statistC.li,
        plabels.optim = TRUE)
```



```
# Les trajectoires des observations ne semblent pas fournies par le package
```

On trouvera ici (<https://adeverse.github.io/ade4/articles/ChapKtables.html#statis-on-operators>) des éléments complémentaires quant à la mise en application de STATIS via le package `ade4`. En particulier, dans le cas d'un nombre d'individus différents entre les tableaux (STATIS duale).

Du point de vue implémentation, on notera que le package FactoMineR permet la gestion des tableaux supplémentaires (via la spécification de l'argument `num.group.sup`) et qu'il offre aussi une visualisation des trajectoires.

5 Conclusion

Nous avons présenté trois 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.

Reférences

- Bourroche, Jean-Marie. 1975. "Analyse Des Données Ternaires: La Double Analyse En Composantes Principales." PhD thesis.
- Dazy, Frédéric, Jean-François Le Barzic, Gilbert Saporta, and Françoise Lavallard. 1996. *L'analyse Des Données évolutives: Méthodes Et Applications*. Editions technip.
- Kroonenberg, Pieter M. 1983. *Three-Mode Principal Component Analysis: Theory and Applications*. Vol. 2. DSWO press.
- 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.
- Lavit, Christine, Yves Escoufier, Robert Sabatier, Pierre Traissac, et al. 1994. "The Act (Statis Method)." *Computational Statistics and Data Analysis* 18 (1): 97–120.
- Pagès, J. 2013. *Analyse Factorielle Multiple Avec r*. Pratique r. EDP Sciences.