ANALYSE FACTORIELLE MULTIPLE PROCUSTÉENNE

E. MORAND et J. PAGÈS [1]

RÉSUMÉ

Pour comparer deux nuages de points homologues, la méthode de référence est l'analyse procustéenne et, dans le cas de plus de deux nuages, l'analyse procustéenne généralisée (APG). L'analyse factorielle multiple (AFM) fournit aussi une représentation superposée permettant de comparer des nuages de points homologues. Dans cette représentation superposée les nuages à comparer subissent des déformations autres que les seules rotations.

Il est possible de compléter l'AFM par un ajustement procustéen de chacun des nuages initiaux sur le nuage moyen de l'AFM. On obtient ainsi une représentation de ces nuages qui à la fois respecte le modèle procustéen et s'inscrit dans le cadre de l'AFM, d'où le nom d'analyse factorielle multiple procustéenne (AFMP). Cette nouvelle représentation est précieuse lorsque les nuages initiaux sont bidimensionnels. Les propriétés de cette nouvelle représentation sont décrites.

Il arrive que l'on ne dispose pas d'une partie des données. Cette situation est fréquente dans le cas de l'application présentée. On décrit ici un algorithme d'imputation qui permet d'utiliser l'AFMP même dans ce cas. Cette situation est illustrée par une application en analyse sensorielle.

Mots-clés : Analyse Factorielle Multiple, Analyse Procrustéenne Généralisée, données manquantes, imputation.

ABSTRACT

To compare two homologous clouds of points, the classical method is the procrustes analysis and, in the case of more than two clouds of points, the Generalized Procrustes Analysis (GPA). The multiple factor analysis (MFA) also provides a superimposed representation in order to compare several homologous clouds of points. In this latter superimposed representation, the clouds of points to compare experience other distorsions different from only rotations.

MFA can be complemented by a procrustes adjustment of each initial cloud of points on the average cloud of the MFA. This superimposed representation of these initial clouds of points respects, at the same time, the procrustes pattern and is included in MFA framework, hence the name of Procrustes Multiple Fator Analysis (PMFA). This new representation is especially very

1. Laboratoire de mathématiques appliquées. Agrocampus Rennes , 65 rue de Saint Brieuc CS 84215 35042 Rennes cedex France. E-mail : morand@agrocampus-rennes.fr

useful when initial clouds are two-dimensional ones. The properties of this new representation are here described.

It happens that we only have a part of the data set. This situation is frequent in the case of the presented application. An algorithm of imputation is here described allowing the use, even in that case, of the PMFA. This situation is illustrated by an application in sensory analysis.

Keywords : Multiple Factor Analysis, Generalized Procrustes Analysis, missing data, imputation.

1. Données-Problème

1.1. Problématique

On étudie un tableau constitué d'individus décrits par plusieurs groupes de variables. Chaque groupe de variables constitue un sous-tableau du tableau complet. L'étude simultanée de plusieurs tableaux peut être abordée suivant, principalement, deux points de vue.

Le premier est celui de l'Analyse Canonique. On considère alors chaque sous-tableau comme un groupe de variables et on étudie les relations entre ces groupes de variables.

Le second consiste à considérer chaque ensemble de variables au travers d'un nuage d'individus appelé configuration partielle. C'est ce point de vue qui est employé dans l'Analyse Procustéenne Généralisée (APG) (Gower, 1975, Ten Berge, 1977). Dans la suite, on utilisera ce point de vue.

Dans l'APG, on compare l'ensemble des nuages de points entre eux. Pour ce faire, on construit à la fois une représentation superposée des configurations partielles et une représentation de référence qui est, en pratique, le barycentre des configurations partielles. La représentation superposée met en évidence les traits communs aux différentes configurations et ce sans déformer les configurations partielles initiales.

L'Analyse Factorielle Multiple (AFM) (Escofier et Pagès, 1998) permet, de la même manière, d'étudier un tableau multiple en considérant les sous-tableaux comme des nuages de points homologues. Elle fournit en premier lieu une représentation de référence à partir de l'ensemble des données qui s'inscrit dans le cadre d'une analyse factorielle classique. On compare ensuite les configurations partielles entre elles et par rapport à la représentation de référence qui est, là encore, la représentation moyenne des représentations partielles superposées. La représentation superposée des nuages est obtenue par projection des configurations partielles dans l'espace de la représentation moyenne des configurations partielles ; elle présente par construction un certain nombre d'avantageuses propriétés (par exemple la dualité). Cependant, les configurations partielles initiales sont déformées, ce qui, dans quelques cas, peut s'avérer être un inconvénient.

On présente ici une méthodologie qui allie des propriétés des deux méthodes : une représentation de référence avec les propriétés intéressantes de l'AFM et une représentation superposée des configurations partielles non déformées.

Cette méthodologie présente un intérêt particulier en analyse sensorielle lors d'un recueil de données par la méthode du napping (Pagès, 2003). Les configurations partielles sont alors des nuages de points en dimension 2. Par la suite, on s'appuiera fréquemment sur cette application.

Cette méthododologie sera d'abord définie dans le cas où le tableau de données est complet. Or, dans le cas de l'application précitée, il est difficile, en pratique, de positionner simultanément plus de 10 ou 12 produits par nappe. Toutefois, le nombre de produits étudiés est souvent plus important. Dans ce cas, chaque juge n'évalue qu'une partie des produits : on dispose alors d'un tableau de données incomplet dont les données manquantes présentent une structure particulière mais classique en APG (Commandeur, 1991).

Dans ce cas particulier où lorsqu'une donnée est manquante dans une ligne d'une configuration partielle alors toute la ligne est manquante pour cette configuration partielle, on propose ici un algorithme (dit d'imputation par rotations multiples) permettant de compléter les données.

L'algorithme proposé est testé dans différents cas de figure. Une fois le tableau completé par cet algorithme on est ramené à l'étude d'un tableau de données complet et la méthodologie présentée ici peut s'appliquer.

1.2. Notation

Un ensemble d'individus, $\{i;\ i = 1, \ldots, I\}$, est décrit par plusieurs groupes de variables. Ces données peuvent être regroupées sous forme d'un tableau unique structuré en sous-tableaux. On note :

- $\mathbf{X}$ le tableau complet ;
- J le nombre de sous-tableaux ;
- K_j le nombre de variables du groupe j ;
- $\mathbf{X}_j$ le tableau associé au groupe j ;
- i^j individu partiel, correspondant à la i^e ligne de $\mathbf{X}_j$.

L'Analyse Procrustéenne Généralisée nécessite des sous-tableaux de même dimension. On supposera dans la suite que cette hypothèse est vérifiée soit :

$$K_j = K \quad \text{pour tout } j = 1, \ldots, J$$

Il est toujours possible de se ramener à ce cas en prenant $K = max\{K_j;\ j = 1, \ldots, J\}$ et en complétant les configurations de dimension inférieure par des colonnes de 0. Les variables sont supposées centrées. On note le produit scalaire usuel entre deux matrices A et B de mêmes dimensions : $\langle \mathbf{A}, \mathbf{B} \rangle = Tr(\mathbf{AB'})$ et $\|\mathbf{A}\| = \sqrt{Tr(\mathbf{AA'})}$ la norme associée.

Pour le cas particulier du napping, une configuration partielle est une nappe. Chaque sous-tableau, $\mathbf{X}_j$, contient alors les $K_j = 2$ coordonnées des produits sur la nappe du dégustateur j.

2. Analyse Factorielle Multiple Procustéenne (AFMP)

Nous proposons, ici, une méthodologie dans laquelle on complète la représentation moyenne de l'AFM par une représentation superposée des configurations partielles obtenue par rotations procustéennes, d'où la dénomination d'Analyse Factorielle Multiple Procustéenne (AFMP).

2.1. Construction de l'AFMP

2.1.1. Représentation de référence

Dans un premier temps, on cherche une représentation de référence à l'aide de l'Analyse Factorielle Multiple du tableau $\mathbf{X}$. Cette représentation de référence (parfois appelée représentation compromis, consensus ou moyenne) est utilisée pour ajuster les configurations partielles.

La représentation de référence est constituée par les S premières composantes de la représentation moyenne de l'AFM. En pratique, on prend généralement $S = K$. Le tableau obtenu des K premières composantes principales, de dimension (I, K), est noté $\mathbf{F}_K$.

La configuration moyenne de l'AFM est choisie comme représentation de référence car elle présente des propriétés intéressantes, en particulier, une interprétation directe à partir des variables initiales. A contrario, en APG la représentation moyenne n'est pas primordiale (elle n'intervient pas, par exemple, lorsqu'on traite seulement deux configurations).

Dans le cas de l'APG, dès lors que $J > 2$, on a recours à un algorithme pour trouver la représentation de référence. L'algorithme converge, mais pas nécessairement, vers un optimum global. Ce problème ne se pose pas dans le cas de l'AFMP dont la configuration de référence est obtenue analytiquement.

2.1.2. Représentation superposée

On considère pour chaque groupe j, les tableaux $\mathbf{X}_j$ pondérés comme en AFM, c'est-à-dire les tableaux $\dfrac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j$ (où λ_1^j est la première valeur propre de l'ACP du tableau $\mathbf{X}_j$). On souhaite les faire coincider le mieux possible avec la représentation de référence $\mathbf{F}_K$ définie précédemment.

On effectue, pour cela, la rotation procustéenne de $\dfrac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j$ sur $\mathbf{F}_K$. On cherche la transformation orthogonale $\mathbf{H}_j$ qui minimise le critère (Saporta, 1990, p. 195) :

$$Tr\Big(\frac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j\mathbf{H}_j - \mathbf{F}_K\Big)\Big(\frac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j\mathbf{H}_j - \mathbf{F}_K\Big)'$$

On obtient ainsi les tableaux, dits « procustéanisés », par les relations :

$$\frac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j\mathbf{H}_j$$

avec $\mathbf{H}_j = \mathbf{V}'_j\mathbf{U}_j$ dans laquelle les matrice orthogonales $\mathbf{V}_j$ et $\mathbf{U}_j$ sont obtenues par la décomposition en valeurs singulières de $\frac{1}{\sqrt{\lambda_1^j}}\mathbf{X}'_j\mathbf{F}_K = \mathbf{V}_j\mathbf{S}_j\mathbf{U}'j$, $\mathbf{S}_j$ étant la matrice diagonale des valeurs singulières.

Remarque. — En se plaçant dans le cadre de l'algorithme d'APG proposé par Gower en 1975, ce calcul peut s'interpréter comme la dernière boucle de cet algorithme en prenant comme configuration consensus la configuration moyenne de l'AFM.

Homothétie (« scaling »)

On peut choisir, lors de l'ajustement de chaque sous-configuration sur la configuration de référence, d'associer à chaque transformation orthogonale $(\mathbf{H}_j)$ une homothétie.

Un fois obtenue la transformation orthogonale $(\mathbf{H}_j)$, le rapport d'inertie (ou « scaling factor ») (ρ_j) est calculé pour minimiser l'écart :

$$Tr\big(\rho_j\frac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j\mathbf{H}_j - \mathbf{F}_K\big)\big(\rho_j\frac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j\mathbf{H}_j - \mathbf{F}_K\big)'$$

la solution est (Gower, 1975) :

$$\rho_j = \frac{Tr(\frac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j\mathbf{H}_j\mathbf{F}'_K)}{Tr(\frac{1}{\lambda_1^j}\mathbf{X}_j\mathbf{X}'_j)}$$

Comme $\mathbf{H}_j$ est une matrice orthogonale, la matrice $\mathbf{H}_j\mathbf{H}'_j$ est la matrice identité donc l'expression précédente peut se réécrire sous la forme suivante :

$$\rho_j = \frac{\|\mathbf{F}_K\|}{\|\frac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j\mathbf{H}_j\|}\frac{Tr\frac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j\mathbf{H}_j\mathbf{F}'_K}{\sqrt{Tr\frac{1}{\lambda_1^j}\mathbf{X}_j\mathbf{X}'_j}\sqrt{Tr(\mathbf{F}_K\mathbf{F}'_K)}} = \frac{\|\mathbf{F}_K\|}{\|\frac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j\mathbf{H}_j\|}\beta_j$$

Ce rapport d'homothétie contient deux termes. Le premier $(\frac{\|\mathbf{F}_K\|}{\|\frac{1}{\sqrt{\lambda_1^j}}\mathbf{X}_j\|})$ correspond à une standardisation tenant compte d'une possible homothétie

entre les deux configurations. Le second, β_j, est un coefficient de similarité, dit de Procuste (Sibson, 1978, Qannari *et al.*, 1999), entre les deux configurations $\mathbf{X}_j$ et $\mathbf{F}_K$. Ce coefficient indique le niveau de conformisme de la configuration partielle à la représentation de référence.

Dans le cas du « napping », la standardisation permet de tenir compte de l'utilisation plus ou moins importante de la surface de la nappe par les différents juges.

Dimension des sous-tableaux

Dans ce qui précède, tous les sous-tableaux sont de même dimension (K_j constant) et la configuration de référence est de dimension $S = K$. C'est ce cas particulier, avec $K_j = S = 2$, qui a suscité l'AFMP et dont on a évalué l'interêt pratique (Morand & Pagès, 2006). Toutefois d'autres cas peuvent être envisageables :

- K_j n'est pas constant. On prend $K = \max\{\mathbf{K}_j, j = 1 \dots J\}$, on complète les tableaux, de dimension inférieure à K, par $K - K_j$ colonnes de 0. La dimension de la configuration de référence S est prise égale à K. Une autre solution consiste à se limiter à un nombre fixe de composantes principales par sous-tableaux ;
- $S > K$. Le choix d'un nombre d'axes supérieur à la dimension des configurations partielles peut être suggéré par les résultats de l'AFM. On complète alors $\mathbf{X}_j$ par $S - K$ colonnes de 0. On se ramène alors au cas classique décrit précédemment.

Propriétés

Par rapport à l'AFM, la représentation superposée obtenue en AFMP est telle que les nuages partiels n'ont subi aucune autre déformation que les transformations orthogonales et les homothéties. Cette représentation est particulièrement intéressante dans le cas bidimensionnel puisque la représentation plane des nuages partiels n'est absolument pas déformée. Cette propriété est précieuse dans le cadre du napping pour présenter les résultats aux dégustateurs. En effet, chaque dégustateur peut voir sa propre configuration telle qu'il l'avait fournie.

En contrepartie, il n'y a plus de relations de transition partielle. En effet, pour la représentation superposée obtenue en analyse factorielle multiple, la coordonnée sur l'axe principal de rang s de l'individu i vu par le groupe j s'exprime comme combinaison linéaire des coordonnées des seules variables du groupe j sur ce même axe. Cette propriété permettant de relier directement la représentation de l'individu vu par le groupe j aux variables du groupe j n'est pas, par construction, conservée en AFMP.

En APG la représentation superposée est optimale au sens d'un critère global de ressemblance entre toutes les configurations transformées : la statistique procustéenne généralisée.

Ce critère, avec nos notations, s'écrit (Gower, 1975, p. 36) :

$$Sr = Tr \sum_{j=1}^{J} (s_j \frac{1}{\sqrt{\lambda_1^j}} \mathbf{X}_j \mathbf{R}_j - \frac{1}{J} \sum_{i=1}^{J} s_i \frac{1}{\sqrt{\lambda_1^i}} \mathbf{X}_i \mathbf{R}_i)' \\ (s_j \frac{1}{\sqrt{\lambda_1^j}} \mathbf{X}_j \mathbf{R}_j - \frac{1}{J} \sum_{i=1}^{J} s_i \frac{1}{\sqrt{\lambda_1^i}} \mathbf{X}_i \mathbf{R}_i) \qquad (2.1)$$

Expression dans laquelle, dans le cadre de l'APG, les $\mathbf{R}_j$ et les s_j sont les transformations orthogonales et les homothéties qui minimisent ce critère .

Soit le modèle d'APG (Commandeur, 1991) suivant :

$$s_j \mathbf{X}_j \mathbf{R}_j = \mathbf{A} + \mathbf{E}_j$$

Où :

- $\mathbf{A}$: configuration compromis ;
- s_j : coefficient d'homothétie ;
- $\mathbf{R}_j$: matrice $(K \times K)$ orthogonale ;
- $\mathbf{E}_j$: matrice $(I \times K)$ des résidus.

En prenant $\mathbf{A} = \frac{1}{J} \sum_{i=1}^{J} s_i \frac{1}{\sqrt{\lambda_1^i}} \mathbf{X}_i \mathbf{R}_i$, Sr peut s'interprèter comme la somme des carrés des résidus du modèle d'APG.

Dans le cas de l'AFMP, on peut calculer le critère Sr en prenant $\mathbf{R}_j = \mathbf{H}_j$ (où $\mathbf{H}_j$ est la matrice de la transformation orthogonale qui permet d'ajuster au mieux $\mathbf{X}_j$ sur $\mathbf{F}_s$) et $s_j = \rho_j$ (où ρ_j est le rapport d'homothétie qui permet d'ajuster au mieux, après transformation orthogonale, $\mathbf{X}_j$ sur $\mathbf{F}_s$). Dans ce cas, la représentation de référence n'est pas la moyenne des représentations partielles superposées. Dans le cas de l'AFMP, par construction, on minimise la somme des carrés des distances entre chaque représentation partielle et la représentation de référence. Lorsque la représentation de référence est égale à la moyenne des représentations partielles transformées, on minimise le même critère que dans le cas de l'APG.

La représentation superposée obtenue en AFMP n'est pas optimale au sens du critère utilisé dans l'APG, mais des traitements systématiques ont montré que le critère n'était pas très différent, en pratique, de celui obtenu en APG. (Morand et Pagès, 2003, p. 107).

2.1.3. Dilatation

Dans le cas de l'AFMP, les rapports d'homothéties ρ_j peuvent se rapprocher de 0 si les configurations partielles sont trop différentes de la configuration de référence. Ceci peut entraîner un problème de lisibilité des graphiques. Dans le cas du napping, cette propriété peut être intéressante pour l'analyste car elle permet de visualiser rapidement les nappes différentes de la configuration obtenue à partir du jury complet. Cependant, cette propriété peut s'avérer

génante au moment de la présentation des résultats aux juges, ceux-ci pouvant avoir des difficultés à reconnaître leur nappe et à voir les différences entre leur représentation des produits et la représentation de référence. On utilise donc dans la pratique un ajustement procustéen sans homothétie.

Toutefois, pour obtenir une meilleur lisibilité des graphiques, on applique une dilatation identique pour toutes les configurations partielles. Le rapport de cette dilatation est calculé pour tenir compte du décalage entre les configurations partielles superposées et la configuration de référence. Dans l'esprit de l'AFM, on souhaite alors que l'inertie de la configuration partielle, $\mathbf{X}_j$, suivant son premier axe, λ_1^j soit égale à l'inertie de la configuration de référence λ_1. On dilate donc les configurations partielles d'un coefficient égal à la racine carré de la première valeur propre de l'AFM ($\sqrt{\lambda_1}$).

Cette dilatation :

- rapproche la configuration partielle de la configuration de référence pour le cas où celle-ci ne diffère que d'une transformation orthogonale.
- montre bien les différences lorsque la configuration partielle est très différente de la configuration de référence.

Finalement, on représente :

$$\widehat{\mathbf{X}_j} = \frac{\sqrt{\lambda_1}}{\sqrt{\lambda_1^j}} \mathbf{X}_j \mathbf{H}_j$$

En particulier, si tous les $\mathbf{X}_j$ sont identiques entre eux à une dilatation et une transformation orthogonale près, alors cette dilatation permet, au même titre que celle permettant d'ajuster au mieux $\mathbf{X}_j$ sur $\mathbf{F}_K$ (ρ_j décrit au paragraphe 2.1.2), de superposer exactement les $\mathbf{X}_j$ à $\mathbf{F}_K$.

3. Aides à l'interprétation

3.1. Notations

On note :

- $\widehat{\mathbf{X}_j}$ les configurations partielles transformées ;

$$\widehat{\mathbf{X}_j} = p_j \frac{1}{\sqrt{\lambda_1^j}} \mathbf{X}_j \mathbf{T}_j$$

  avec p_j le coefficient d'homothétie (s_j dans le cas de l'APG ou $\sqrt{\lambda_1}$ dans le cas de l'AFMP) ;
- $\mathbf{T}_j$ les transformations orthogonales ($\mathbf{R}_j$ dans le cas de l'APG ou $\mathbf{H}_j$ dans le cas de l'AFMP) ;
- $\mathbf{Y} = \frac{1}{J} \sum_j \widehat{\mathbf{X}_j}$ la moyenne des configurations partielles.

3.2. Représentation de référence et moyenne des configurations partielles

Dans le cas de l'AFMP, la configuration de référence, $\mathbf{F}_K$, n'est pas la moyenne, $\mathbf{Y}$, des configurations. Dans le cas de l'APG, on minimise la somme des carrés des écarts à la moyenne $\mathbf{Y}$. Ce critère peut aussi se calculer dans le cas de l'AFMP. Toutefois, dans ce cas, on minimise l'écart des représentations superposées à la représentation de référence. On note alors dans le cas de l'AFMP :

– l'inertie inter individus :

$$Sr(\mathbf{Y}) = \sum_j \text{Tr } (\widehat{\mathbf{X}_j} - \mathbf{Y})(\widehat{\mathbf{X}_j} - \mathbf{Y})'$$

– l'inertie des individus par rapport à la représentation de référence :

$$Sr(\mathbf{F}_K) = \sum_j \text{Tr } (\widehat{\mathbf{X}_j} - \mathbf{F}_K)(\widehat{\mathbf{X}_j} - \mathbf{F}_K)'$$

– Le coefficient dit « de distorsion » : $Dist(\mathbf{F}_K, \mathbf{Y}) = \text{Tr } (\mathbf{Y} - \mathbf{F}_K)(\mathbf{Y} - \mathbf{F}_K)'$

On a alors l'égalité suivante :

$$Sr(\mathbf{F}_K) = S_r(\mathbf{Y}) + J * Dist(\mathbf{F}_K, \mathbf{Y})$$

Plus $\mathbf{Y}$ est proche de $\mathbf{F}_K$ plus les critères minimisés en AFMP ($Sr(\mathbf{F}_K)$) et en APG ($Sr(Y)$) sont proches. Il est donc souhaitable d'évaluer, par un indice, la ressemblance entre $\mathbf{Y}$ et $\mathbf{F}_K$. L'indice employé en pratique est un coefficient de distortion « normalisé » rapporté à la somme des inerties des deux configurations :

$$\frac{Tr(\mathbf{Y} - \mathbf{F}_K)(\mathbf{Y} - \mathbf{F}_K)'}{Tr(\mathbf{Y}\mathbf{Y}') + Tr(\mathbf{F}_K\mathbf{F}'_K)}$$

Si cet indicateur est faible, alors le coefficient de distorsion est faible. Dans la pratique, un coefficient de distorsion faible indique que, dans le cas de l'AFMP, on minimise un critère proche de celui minimisé par l'APG. Ce dernier s'interprète comme un indicateur d'homogénéité globale entre les configurations partielles.

3.3. Comparaison de deux configurations

Lors de l'interprétation, il est nécessaire de pouvoir comparer, entre elles, deux configurations, soit deux configurations partielles soit une configuration partielle et la configuration de référence.

Coefficient RV

Un indicateur classique pour comparer deux configurations entre elles est le coefficient RV (Escoufier, 1973).

$$RV(\mathbf{X}_j;\mathbf{F}_K) = \frac{Tr(\mathbf{X}_j\mathbf{X}_j'\mathbf{F}_K\mathbf{F}_K')}{\sqrt{Tr(\mathbf{X}_j\mathbf{X}_j'\mathbf{X}_j\mathbf{X}_j')Tr(\mathbf{F}_K\mathbf{F}_K'\mathbf{F}_K\mathbf{F}_K')}}$$

Ce critère vaut 1 si $\mathbf{X}_j$ peut se déduire de $\mathbf{F}_K$ par une homothétie et/ou une transformation orthogonale. Ce critère vaut 0 si toutes les colonnes de $\mathbf{X}_j$ sont orthogonales à toutes les colonnes de $\mathbf{F}_K$.

Coefficient RV standardisé

Le coefficient RV standardisé (Kazi-Aoual *et al.*, 1995) s'écrit avec nos notations :

$$RV_{std} = \frac{RV(\mathbf{X}_j;\mathbf{F}_K) - E(RV(\mathbf{X}_j;\mathbf{F}_K))}{\sigma(RV(\mathbf{X}_j;\mathbf{F}_K))}$$

Expression dans laquelle $E(RV(\mathbf{X}_j;\mathbf{F}_K))$ et $\sigma(RV(\mathbf{X}_j;\mathbf{F}_K))$ sont respectivement l'espérance et l'écart-type du coefficient, calculés en considérant sa distribution obtenue lorsqu'on effectue **toutes** les permutations (considérées comme équiprobables) des lignes de l'une des deux matrices (cela correspond à une distribution sous l'hypothèse nulle d'indépendance des deux matrices, distribution que l'on peut approcher avec une bonne approximation par une loi gaussienne). On se place comme Schlich (Schlich, 1996, p. 266) sous l'hypothèse d'une distribution gaussienne, lorsque le RV_{std} est supérieur à 1.65 cela signifie que la valeur observée du RV est significativement différente (au seuil 5%) de celle que l'on obtiendrait en moyenne en permutant les lignes des configurations. Ceci permet en particulier de s'affranchir de la variation structurelle du RV due à la taille des matrices comparées.

Indice de similarité de Procruste

L'indice de similarité de procruste entre deux configurations $\mathbf{A}$ et $\mathbf{B}$, $S(\mathbf{A},\mathbf{B}) = Tr(\mathbf{A}'\mathbf{B}\mathbf{H})/\sqrt{Tr(\mathbf{A}\mathbf{A}')}\sqrt{Tr(\mathbf{B}\mathbf{B}')}$ dans laquelle $\mathbf{H}$ est la matrice de transformation orthogonale qui ajuste au mieux $\mathbf{B}$ à $\mathbf{A}$, est classiquement employé en APG pour comparer les sous-configurations deux à deux. Dans notre cas, on peut, pour chaque sous-configuration, calculer un indice de similarité entre $\mathbf{X}_j$ et $\mathbf{F}_K$:

$$S(\mathbf{X}_j,\mathbf{F}_K) = \frac{\text{Tr } \mathbf{X}_j\mathbf{H}_j\mathbf{F}_K'}{\sqrt{\text{Tr } (\mathbf{X}_j\mathbf{X}_j')}\sqrt{\text{Tr } (\mathbf{F}_K\mathbf{F}_K')}} = \frac{\text{Tr}\widehat{\mathbf{X}_j}\mathbf{F}_K'}{\sqrt{\text{Tr}(\widehat{\mathbf{X}_j}\widehat{\mathbf{X}_j}')}\sqrt{\text{Tr } (\mathbf{F}_K\mathbf{F}_K')}}$$

Cet indicateur vaut 1 si $\mathbf{X}_j$ est identique à $\mathbf{F}_K$ et 0 si les $\mathbf{X}_j$ et $\mathbf{F}_K$ sont dans deux sous-espaces orthogonaux. C'est à partir de la matrice de ces indices de similarités que les praticiens de l'APG obtiennent à l'aide d'une méthode MDS, une représentation des groupes (Arnold & Williams, 1986, Krzanowski, 1990, pp. 163-164).

3.4. Décomposition de l'inertie

On souhaite connaître les individus et les sous-configurations les plus contributifs à la somme des carrés des écarts à la représentation de référence. Dans le cas de l'APG, on décompose alors Sr suivant les individus et suivant les sous-configurations (Dijksterhuis & Punter, 1990). Ces résultats sont en général regroupés au sein de tableaux d'analyse de variance (Dijksterhuis & Gower, 1991).

Le résidu Sr s'écrit dans le cas de l'APG :

$$Sr = \sum_{j=1}^{J} Tr(\widehat{\mathbf{X}_j} - \mathbf{Y})'(\widehat{\mathbf{X}_j} - \mathbf{Y})$$

On sait que :

$$J * \sum_{j=1}^{J} \text{Tr } (\mathbf{X}_j - \mathbf{Y})(\mathbf{X}_j - \mathbf{Y})') = \sum_{u<v}^{J} \text{Tr } (\mathbf{X}_u - \mathbf{X}_v)(\mathbf{X}_u - \mathbf{X}_v)'$$

De par cette égalité, on interprète Sr comme un indicateur d'homogénéité globale entre les configurations partielles. Cet indicateur est valable aussi bien pour la représentation superposée de l'APG que pour celle de l'AFMP.

De plus, chaque terme de la somme $Tr(\widehat{\mathbf{X}_j} - \mathbf{Y})'(\widehat{\mathbf{X}_j} - \mathbf{Y})$ peut s'interpréter comme une proximité de la jème configuration partielle à la réprésentation moyenne. Plus une sous-configuration contribue à la somme des carrés des résidus moins elle est en accord avec la représentation moyenne. On considère, en général, qu'une configuration partielle dont la contribution au résidu Sr est très supérieure à la contribution moyenne Sr/J est « très en désaccord » avec le reste des configurations partielles. Il y a alors un intérêt à étudier cette configuration partielle « aberrante » voire à la supprimer du modèle pour améliorer l'ajustement.

Dans notre cas, on s'intéresse plutôt à l'écart entre la configuration partielle et la configuration de référence. On peut alors décomposer l'inertie des individus par rapport à la représentation de référence (en remplaçant $\mathbf{Y}$ par $\mathbf{F}_K$). Si la configuration de référence est proche de la moyenne des configurations, on peut empiriquement interpréter cet écart comme Sr. On procède de même que dans le cas de l'APG. Toutefois, si l'on souhaite étudier la proximité entre une configuration partielle et la configuration de référence, on utilise préférentiellement le coefficient RV.

On note $\widehat{\mathbf{X}_j}^{(i)}$ la i^e ligne de la j^e configuration transformée qui a pour dimension $(1 \times K)$ et $\mathbf{Y}^{(i)}$ la i^e ligne de la moyenne des configurations partielles. On peut décomposer chaque terme $Tr(\widehat{\mathbf{X}_j} - \mathbf{Y})'(\widehat{\mathbf{X}_j} - \mathbf{Y})$ comme une somme sur toutes les lignes (tableau 1) soit : $Tr(\widehat{\mathbf{X}_j} - \mathbf{Y})'(\widehat{\mathbf{X}_j} - \mathbf{Y}) = \sum_{i=1}^{I}(Tr(\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})(\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})')$

TABLEAU 1. — Décomposition de l'inertie par configuration partielle et par ligne.

	1 ...	j	... J	Total
1	⋮ ...	...	... ⋮	
⋮	⋮ ...	⋮	... ⋮	⋮
i	⋮ ... Tr $(\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})(\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})'$		... ⋮	$\sum_{j=1}^{J}$ Tr $(\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})(\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})'$
⋮	⋮ ...	⋮	... ⋮	⋮
I	⋮ ...	...	... ⋮	
Total	... ...	Tr $(\widehat{\mathbf{X}_j} - \mathbf{Y})(\widehat{\mathbf{X}_j} - \mathbf{Y})'$	...	Sr

Alors Sr s'écrit :

$$Sr = \sum_{i=1}^{I}(\sum_{j=1}^{J} \mathrm{Tr}(\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})(\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})')$$

Dans l'interprétation, on utilise, pour chaque individu, la somme des carrés des distances entre les points partiels $(\mathbf{X}_j^{(i)})$ et le point moyen :

$$\sum_{j=1}^{J} \mathrm{Tr}\ (\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})(\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})'.$$

Chacun des termes de la décomposition peut être réécrit comme suit :

$$J * \sum_{j=1}^{J} \mathrm{Tr}\ (\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})(\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})') = \sum_{u<v}^{J} \mathrm{Tr}(\widehat{\mathbf{X}_u}^{(i)} - \widehat{\mathbf{X}_v}^{(i)})(\widehat{\mathbf{X}_u}^{(i)} - \widehat{\mathbf{X}_v}^{(i)})'$$

Ce terme représente, à une constante multiplicative près, la somme des carrés des distances entre l'ensemble des J points $\{i^j, j = 1 \dots J\}$ pour un même individu i. Dans le cas de l'exemple des nappes, ce terme s'interprète comme un indicateur de l'homogénéité des avis des dégustateurs sur un même produit i. Cet indicateur est interprétable pour la représentation superposée de l'APG et de l'AFMP.

Dans le cas de l'APG, on peut écrire le théorème de Huygens pour chaque ligne, i :

$$\sum_{j=1}^{J} \mathrm{Tr}\ ((\widehat{\mathbf{X}_j}^{(i)})(\widehat{\mathbf{X}_j}^{(i)})') = JTr(\mathbf{Y}^{(i)}(\mathbf{Y}^{(i)})') + \sum_{j} \mathrm{Tr}\ (\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})(\widehat{\mathbf{X}_j}^{(i)} - \mathbf{Y}^{(i)})'$$

Dans le cas de l'APG, cela s'interprète comme la décomposition de l'inertie totale pour un individu i en une part expliquée par la configuration moyenne et une part résiduelle. Cette décomposition présente peu d'intérêt dans le cas de l'AFMP pour laquelle la configuration de référence n'est pas $\mathbf{Y}$.

On peut aussi décomposer chacun des termes $Tr(\widehat{\mathbf{X}_j} - \mathbf{Y})'(\widehat{\mathbf{X}_j} - \mathbf{Y})$ suivant les K axes ayant servi à réaliser les rotations procrustéennes. Dans le cas de l'exemple des nappes (K étant alors égale à 2), on peut ainsi savoir, pour un produit sur lequel il n'y a pas homogénéité des avis, si ce manque de cohésion est particulièrement visible sur certains axes.

4. Cas de données incomplètes

Dans tout ce qui précède, on a considéré que le tableau de données était complet. Or, en pratique, on ne dispose pas toujours de l'intégralité du tableau. On envisage ici une structure particulière pour les données manquantes : lorsqu'une donnée est manquante sur une ligne i du sous-tableau $\mathbf{X}_j$, toute la ligne est manquante pour ce sous-tableau.

Dans le cas de l'exemple de l'analyse sensorielle, en pratique, il est difficile de positionner plus de 12 produits sur une nappe. Or, il arrive fréquemment que l'on ait plus de 12 produits à comparer. On se retrouve donc dans un cas où tous les juges ne peuvent déguster tous les produits. La configuration du juge j, $\mathbf{X}_j$, contient alors des lignes non renseignées correspondant aux produits que le juge n'a pas dégustés.

4.1. Données, Notations

On définit ici quelques notations utilisées par Commandeur (Commandeur, 1991, p. 22).

- $\mathbf{X}_j$ la j^e configuration partielle incomplète ;
- $\mathbf{X}_j^c$ la j^e configuration partielle individuelle complétée (voir paragraphe suivant) ;
- $\lambda_1^{j,c}$ la première valeur propre de l'ACP de $\mathbf{X}_j^c$;
- $\mathbb{1}$ le vecteur de dimension $(I \times 1)$ ne contenant que des 1 ;
- $\mathbf{M}_j$ la matrice diagonale de dimension $(I \times I)$ dont les termes valent 1 si la ligne de $\mathbf{X}_j$ correspondante est présente et 0 si la ligne de $\mathbf{X}_j$ correspondante est manquante ;
- $\mathbf{C}_{\mathbf{M}_j} = \left(I - \frac{\mathbb{1}\ \mathbb{1}'\mathbf{M}_j}{\mathbb{1}'\mathbf{M}_j\ \mathbb{1}}\right)$

Le produit $\mathbf{M}_j\mathbf{C}_{\mathbf{M}_j}\mathbf{X}_j$ permet d'obtenir une configuration partielle centrée par rapport aux valeurs présentes et dont les lignes manquantes sont mises à 0 (Commandeur, 1991, p. 27).

4.2. AFMP en présence de valeurs manquantes

On cherche une représentation de référence des I points qui ressemble à toutes les configurations partielles incomplètes. On opte ici pour une méthode par imputation pour traiter le tableau de données incomplet. Cette méthode de traitement des valeurs manquantes présente l'avantage de permettre l'application de l'AFMP aussi bien que d'autres méthodes telles que l'APG et l'AFM sur le tableau de données complété. Une méthode d'imputation facile d'utilisation consiste à remplacer une valeur manquante par la moyenne de la variable (méthode dite de la moyenne inconditionnelle : MI). On essaie ici d'améliorer cette méthode par une imputation de type « plus proche voisin ».

4.2.1. Imputation par une méthode de type « plus proche voisin » et représentation de référence

Dans un tableau de données complet classique (tableau *individus* × *variables* sans structure de groupe sur les variables), lorsqu'une partie des variables n'est pas renseignée pour un individu, on a la possibilité de compléter ses informations par celles obtenues sur un individu proche. Dans notre cas, l'individu partiellement renseigné est une configuration partielle dans laquelle certains points sont absents. Pour compléter cette configuration, on dispose d'autres configurations lui ressemblant plus ou moins.

On se propose de compléter une configuration partielle en utilisant les informations contenues dans les autres configurations. On effectue pour cela des ajustements procustéens entre les configurations deux à deux. On entend par ajustement procustéen, une translation et une transformation orthogonale d'une configuration afin de la rapprocher d'une autre de façon optimale.

Compléter une configuration partielle par une autre

On souhaite compléter $\mathbf{X}_{j_1}$ par une configuration $\mathbf{X}_{j_2}$ contenant les points manquants dans $\mathbf{X}_{j_1}$. Pour cela, on effectue un ajustement procustéen avec homothétie de $\mathbf{X}_{j_2}$ sur $\mathbf{X}_{j_1}$. On complète alors les lignes manquantes de $\mathbf{X}_{j_1}$ par les lignes correspondantes présentes dans la configuration $\mathbf{X}_{j_2}$ « ajustée ».

L'algorithme est le suivant :

1. On calcule la matrice diagonale, $\mathbf{M}_{j_1}\mathbf{M}_{j_2}$, dont les termes valent :
   - 1 si la ligne correspondante est présente dans les deux configurations $\mathbf{X}_{j_1}$ et $\mathbf{X}_{j_2}$;
   - 0 sinon.

   On centre $\mathbf{X}_{j_1}$ et $\mathbf{X}_{j_2}$ par rapport à leur points communs. Soit la matrice $\widetilde{\mathbf{X}}_{j_1} = \mathbf{M}_{j_1}\mathbf{C}_{\mathbf{M}_{j_1}\mathbf{M}_{j_2}}\mathbf{X}_{j_1}$ constituée par les points présents dans $\mathbf{X}_{j_1}$ translatée au barycentre des points communs aux deux configurations et les lignes manquantes de $\mathbf{X}_{j_1}$ mises à 0. On a de même la matrice $\widetilde{\mathbf{X}}_{j_2} = \mathbf{M}_{j_2}\mathbf{C}_{\mathbf{M}_{j_1}\mathbf{M}_{j_2}}\mathbf{X}_{j_2}$;
2. On ajuste les deux configurations sur leurs seuls points communs. On rapproche donc $\mathbf{M}_{j_1}\widetilde{\mathbf{X}}_{j_2}$ de $\mathbf{M}_{j_2}\widetilde{\mathbf{X}}_{j_1}$ à l'aide d'une transformation orthogonale

$\widetilde{\mathbf{H}}_{j_2}$ et d'une homothétie $\widetilde{\rho}_{j_2}$. Ces transformations sont optimales au sens de l'ajustement procustéen avec homothétie (voir paragraphe 2.1.2) ;

3. On utilise les lignes de la configuration transformée, $\widetilde{\rho}_{j_2}\widetilde{\mathbf{X}}_{j_2}\widetilde{\mathbf{H}}_{j_2}$, absentes dans $\mathbf{X}_{j_1}$, pour compléter $\widetilde{\mathbf{X}}_{j_1}$. On obtient ainsi une configuration complétée $\widetilde{\mathbf{X}}_{j_1}^{c,j_2}$;
4. On translate les points de $\widetilde{\mathbf{X}}_{j_1}^{c,j_2}$ pour obtenir $\mathbf{X}_{j_1}^{c,j_2}$ matrice contenant les coordonnées des points de $\mathbf{X}_{j_1}$, les points complémentaires obtenus à partir de $\mathbf{X}_{j_2}$ et des lignes de 0 pour les points absents dans les deux configurations. Cette matrice s'obtient par la relation :

$$\mathbf{X}_{j_1}^{c,j_2} = (\mathbf{M}_{j_1} + \mathbf{M}_{j_2} - \mathbf{M}_{j_1}\mathbf{M}_{j_2})\big(\widetilde{\mathbf{X}}_{j_1}^{c,j_2} + \frac{\mathbb{1}\mathbb{1}'\mathbf{M}_{j_1}\mathbf{M}_{j_2}\mathbf{X}_{j_1}}{\mathbb{1}'\mathbf{M}_{j_1}\mathbf{M}_{j_2}\mathbb{1}}\big)$$

$\mathbf{M}_{j_1} + \mathbf{M}_{j_2} - \mathbf{M}_{j_1}\mathbf{M}_{j_2}$ est une matrice diagonale dont les termes valent :
- 0 si la ligne correspondante est manquante dans les deux configurations ;
- 1 si la ligne est présente dans au moins une des deux configurations.

Imputation par « rotations multiples » (IRM)

Pour chaque configuration $\mathbf{X}_{j_1}$ il est possible d'obtenir $J-1$ configurations complétées $\mathbf{X}_{j_1}^{c,j_2}(\{j_2 \neq j_1, j_2 = 1 \dots J\})$ en complétant $\mathbf{X}_{j_1}$ par chacune des configurations $\mathbf{X}_{j_2}$ restantes.

Ainsi pour chaque point, i^{j_1}, manquant de la configuration $\mathbf{X}_{j_1}$, on obtient un ensemble de points i^{j_2}, contenus dans les configurations $\mathbf{X}_{j_1}^{c,j_2}$, pouvant prétendre au remplacement du point manquant. On note cet ensemble qui peut contenir au maximum $J-1$ points :

$$E(i^{j_1}) = \{i^{j_2}, j_2 \neq j_1, j_2 = 1 \dots J, \quad i^{j_2} \text{ présent dans la configuration } \mathbf{X}_{j_2}\}$$

On remplace chaque point i^{j_1} manquant de la configuration $\mathbf{X}_{j_1}$ par une fonction des points contenus dans $E(i^{j_1})$. Cette fonction peut être :
- **le barycentre**. On remplace i^{j_1} par le barycentre des points de $E(i^{j_1})$.
- **la médiane**. On entend par médiane des points de $E(i^{j_1})$, le point dont la k^e coordonnée est la médiane des k^e coordonnées de l'ensemble de points de $E(i^{j_1})$. Cette imputation est robuste vis à vis de points « aberrants ».
- **une sélection**. On suppose ici qu'un point provenant d'une configuration très éloignée de celle que l'on souhaite compléter présente peu d'intérêt. On complète donc la configuration $\mathbf{X}_{j_1}$ à l'aide des configurations les plus proches. La proximité entre deux configurations partielles, $\mathbf{X}_{j_2}$ et $\mathbf{X}_{j_1}$, est mesurée à l'aide du coefficient RV standardisé calculé sur leurs points communs. On prend le barycentre des points de $E(i^{j_1})$ pour lesquels le coefficient RV standardisé entre $\mathbf{X}_{j_2}$ et $\mathbf{X}_{j_1}$ est supérieur strictement à un seuil. Empiriquement on a choisi de prendre un seuil égal à 0. Si aucune configuration $\mathbf{X}_{j_2}$ ne remplit cette condition, on remplace les points manquants à l'aide de la moyenne inconditionnelle.

Représentation de référence

Une fois les configurations individuelles reconstruites, on se retrouve dans le cas classique d'un tableau multiple complet pour lequel on peut trouver une configuration de référence par AFM ou APG.

Dans le cadre de l'AFMP, la représentation de référence, $\mathbf{F}_K^c$, est obtenue à l'aide d'une AFM sur le tableau complété. En pratique, une fois les données complétées, on effectue les analyses comme sur un tableau de données complet ; c'est la raison pour laquelle les configurations partielles sont pondérées à l'aide des poids de l'AFM des données complétées, $\lambda_1^{j,c}$.

4.2.2. Représentation superposée et dilatation

La représentation superposée de l'AFMP s'obtient en ajustant les configurations individuelles incomplètes sur la représentation moyenne $\mathbf{F}_K^c$ obtenue précédemment.

On ajuste $\frac{1}{\sqrt{\lambda_1^{j,c}}}\mathbf{M}_j\mathbf{X}_j^c$, les seuls points présents de la configuation $\mathbf{X}_j$, sur les points correspondants de la configuration $\mathbf{F}_K^c$.

L'ajustement procustéen peut être complété par une homothétie (voir paragraphe 2.1.2). On calcule pour cela l'homothétie de l'ajustement procustéen des données présentes dans la configuration individuelle sur les mêmes points de la configuration de référence.

Pour la représentation graphique, on peut comme dans le cas complet utiliser une dilatation égale à la première valeur propre de l'AFM du tableau de données complété. Dans la pratique, c'est cette dilatation qui a été utilisée.

4.3. Évaluation de la méthode d'imputation

4.3.1. Principe

Pour évaluer la méthode d'imputation dans ses différentes variantes décrite au paragraphe précédent, on utilise plusieurs ensembles de données réelles ou simulées auxquelles on enlève aléatoirement, à chaque configuration partielle, le même nombre de lignes (n). Pour chaque ensemble de données, on simule différents scénarios d'emplacement des valeurs manquantes.

On note $\mathbf{Z}$ la configuration de référence obtenue sur l'ensemble des données complètes et $\mathbf{Z}_c$ celle obtenue à partir des données complétées. L'objectif de l'imputation est d'obtenir un $\mathbf{Z}_c$ qui ressemble le plus à $\mathbf{Z}$.

On étudie cette ressemblance entre $\mathbf{Z}$ et $\mathbf{Z}_c$ en fonction du nombre de valeurs manquantes n. Dans le cas de notre exemple, l'étude de cette ressemblance en fonction de n permet de déterminer un nombre de produits manquants par juge à partir duquel $\mathbf{Z}$ et $\mathbf{Z}_c$ ne sont plus suffisamment ressemblantes. Concrètement, la détermination de ce nombre permet de fixer des recommandations pour les recueils de données par la méthode des nappes.

On peut aussi avoir l'intuition que la reconstitution de la configuration de référence par les méthodes d'imputation est meilleure si la structure commune

aux configurations partielles est forte. D'où l'idée de mesurer la « force » de la structure commune et d'en tenir compte dans l'évaluation de la qualité de la méthode d'imputation.

Pour évaluer la « force » de la structure commune des sous-tableaux de $\mathbf{X}$, on utilise la statistique procustéenne généralisée (voir paragraphe 2.1.2) ramenée à un pourcentage de l'inertie totale soit :

$$Sr' = \frac{Sr}{\sum_{j=1}^{J} \text{Trace } \mathbf{X}_j \mathbf{X}_j'}$$

Sr' vaut 0 si le modèle d'APG est parfaitement vérifié, c'est-à-dire si toutes les configurations $\mathbf{X}_j$ diffèrent les unes des autres d'une rotation et/ou d'une homothétie. Plus on s'éloigne de ce modèle plus ce critère augmente (sans jamais atteindre la valeur 1).

La capacité de la méthode d'imputation à reconstruire $\mathbf{Z}$ est donc une fonction du nombre de données manquantes, n, et de la force de la structure sous-jacente aux données, Sr'.

4.3.2. Construction des tableaux complets

Pour déterminer les propriétés d'une méthode d'imputation, on la teste dans un premier temps sur des tableaux complets simulés. On conçoit ces tableaux complets à l'aide d'un modèle connu que l'on perturbe. On a utilisé ici deux modèles.

Modèle APG

On se donne une configuration $\mathbf{A}$, un ensemble de J homothéties (ρ_j) et de J transformations orthogonales $(\mathbf{H}_j)$. A partir de $(\mathbf{A}, \rho_j, \mathbf{H}_j)$, on constitue les configurations partielles par la relation $\mathbf{X}_j = \rho_j \mathbf{A}\mathbf{H}_j$: c'est ce que nous appellons un modèle d'APG. On perturbe ce modèle initial en ajoutant un bruit. La procédure étant indépendante du choix des transformations orthogonales, on choisit ici $\mathbf{H}_j = I$.

Les différentes configurations sont obtenues en prenant comme modèle :

$$\mathbf{X}_j = \rho_j \mathbf{A} + \mathbf{E}_j(\sigma)$$

Dans lequel $\mathbf{E}_j(\sigma)$ est une matrice de même dimension que $\mathbf{X}_j$ contenant les réalisations indépendantes d'une loi normale centrée et de variance σ^2. Plus σ augmente moins la structure commune aux groupes est « forte ».

Modèle INDSCAL

On se donne une configuration $\mathbf{A}$ ainsi qu'un ensemble de J matrices diagonales, Δ_j, d'éléments non nuls. Chaque configuration partielle est obtenue par $\mathbf{X}_j = \mathbf{A}\Delta_j$: c'est ce que nous appelons le modèle INDSCAL. A ce modèle on ajoute, comme dans le cas du modèle APG, un bruit :

$$\mathbf{X}_j = \mathbf{A}\Delta_j + \mathbf{E}_j(\sigma)$$

4.3.3. Critères d'évaluation

D'une part, on étudie la capacité d'une méthode d'imputation à reconstituer la configuration de référence : pour cela on calcule le coefficient RV entre $\mathbf{Z}$ et $\mathbf{Z}_c$. D'autre part, on cherche à connaître la capacité de la méthode d'imputation à reconstituer les données : pour cela on calcule, pour chaque groupe j, le coefficient RV entre $\mathbf{X}_j$ et $\mathbf{X}_j^c$; on conserve comme indicateur synthétique la moyenne de ces RV.

Pour obtenir un tableau complet, on se donne un modèle et un bruit (σ). On prend un ensemble de Ω valeurs de σ comprises dans l'intervalle de 0 à $max(\sigma)$. La valeur maximum de σ est choisie de façon à obtenir des ensembles de données sans structure commune. Le couple, (modèle, bruit), détermine la force de structure commune, Sr', des données. Pour un tableau complet, et donc une valeur de Sr', on fait varier le nombre de valeurs manquantes n ($\{n = 1 \dots max(n)\}$). Pour chaque tableau complet et pour un nombre de valeurs manquantes fixé, n, on fait varier l'emplacement des valeurs manquantes. On a donc pour un couple de valeurs (Sr', n), N tableaux incomplets correspondant à différents emplacements de valeurs manquantes. Pour un couple (Sr', n) et une méthode d'imputation, il existe donc N valeurs de RV($\mathbf{Z}$, $\mathbf{Z}_c$). Pour une appréciation globale de la méthode d'imputation, on se restreint à conserver, pour un couple (Sr', n) et une méthode d'imputation, un seul RV($\mathbf{Z}$, $\mathbf{Z}_c$) égal à la moyenne des N RV($\mathbf{Z}$, $\mathbf{Z}_c$) obtenus. On fait de même pour le RV entre configurations partielles.

Dans le cas des tableaux complets simulés, on dispose d'un nombre considérable de résultats. Pour étudier l'ensemble des valeurs de RV obtenues, on les représente graphiquement dans le repère (RV($\mathbf{Z}$, $\mathbf{Z}_c$), $\frac{1}{J}\sum_{j=1}^{J} RV(\mathbf{X}_j, \mathbf{X}_j^c)$). Le nuage de point obtenu a pour cardinal le nombre de combinaisons (méthode d'imputation, modèle, n, Sr') soit $4 * 2 * max(n) * \Omega$. On résume l'information en représentant le barycentre des points issus de la même méthode d'imputation, celui des points issus du même modèle de simulation celui des points ayant le même nombre de valeurs manquantes et celui des points ayant la même valeur de Sr'.

Dans le cas d'un jeu réel de données, la force de la structure sous-jacente est fixe. On fait alors varier uniquement le nombre et l'emplacement des valeurs manquantes. Pour les données réelles, pour chaque méthode d'imputation et chaque valeur de n, on fait la moyenne des critères obtenus en faisant varier l'emplacement des valeurs manquantes dans les configurations partielles. Les résultats sont présentés sous forme d'un graphique par critère présentant l'évolution de la moyenne des coefficients RV en fonction du nombre de valeurs manquantes et de la méthode d'imputation.

4.4. Résultats

4.4.1. Simulations à partir des modèles

Cadre de la simulation

Pour les simulations, les deux modèles (APG, INDSCAL) sont employés. On fixe les paramètres, I, J, K des tableaux multiples simulés identiquement à ceux d'un cas réel de recueil par la méthode du napping : $J = 10$ le nombre de juges, $I = 16$ le nombre de produits et $K = 2$ les coordonnées sur la nappe. La configuration $\mathbf{A}$ initiale est une configuration plane de forme hexagonale centrée (les deux variables ont pour écart-type respectif 1.32 et 2.08). La configuration de référence est constituée des 2 premières composantes de l'AFM du tableau de données complet ($\mathbf{Z} = \mathbf{F}_2$). Lorsque le modèle d'APG est parfaitement vérifié, la configuration $\mathbf{Z}$ est identique à $\mathbf{A}$. Plus le bruit augmente moins la configuration de référence $\mathbf{Z}$ ressemble à la configuration sous-jacente $\mathbf{A}$. Pour les deux modèles APG et INDSCAL, un des $\mathbf{X}_j$ est pris égal à $\mathbf{A}$. Le modèle APG initial (voir paragraphe 4.3.2) est constitué à partir de $\{\rho_j\,;\, j = 1 \ldots 10\}$ variant de 0.75 à 3 avec un pas de 0.25. Dans le cas du modèle INDSCAL, les rapports entre dilatation sur l'axe 1 et dilatation sur l'axe 2 varient de 0.852 à 6.

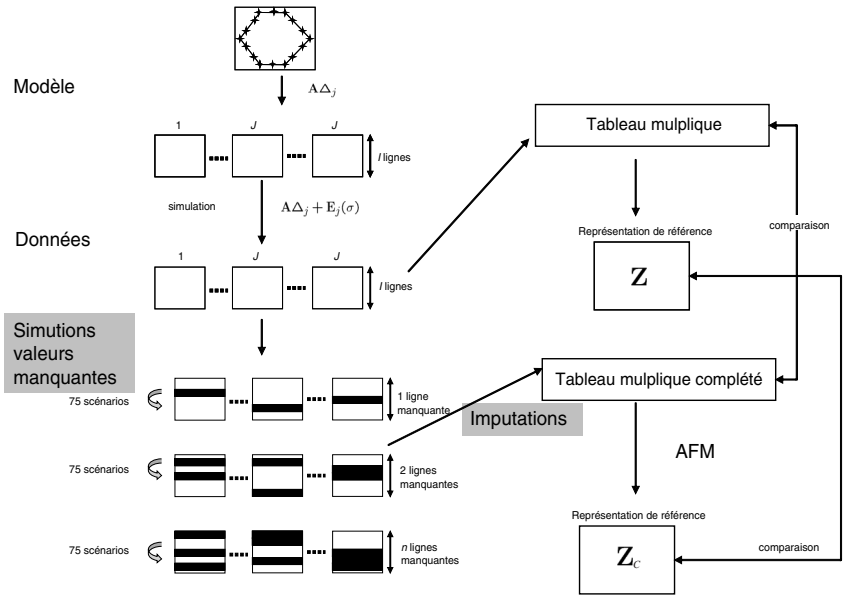


FIG 1. — Schéma d'une simulation d'un tableau complet pour un modèle particulier et d'un ensemble de scénarios d'emplacement de valeurs manquantes.

Les valeurs de Sr' varient en fonction du modèle de simulation des données et du bruit (σ). On fait varier σ de 0 à 10.2 avec un pas de 0.05 ($\Omega = 205$). Pour chaque σ, on crée un tableau complet à partir du modèle APG et un

tableau complet à partir d'un modèle INDSCAL. En procédant de la sorte, on obtient un éventail de valeurs de Sr' comprises entre 0 à 0.75 (voir figure 2). L'histogramme présente de nombreuses valeurs de Sr' entre 0,6 et 0,75 du fait du choix de σ qui engendre de nombreux jeux de données sans structure sous-jacente. De plus, en travaillant sur des données aléatoires de ce format, l'analyse procrustéenne généralisée explique au moins 25% de l'inertie totale.

Pour les simulations de valeurs manquantes effectuée ici, n varie de 1 à 9 et $N = 75$ (voir figure 1). On utilise 4 méthodes d'imputation : la méthode de référence (i.e. la moyenne inconditionnelle, MI) et les 3 variantes de l'algorithme IRM (barycentre, médiane, sélection).

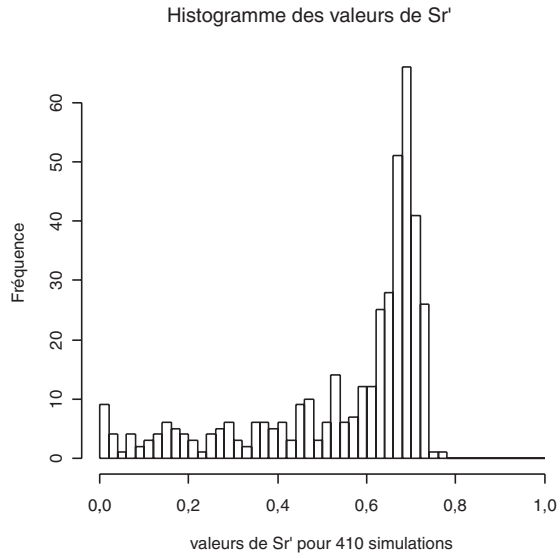


FIG 2. — Histogramme des valeurs prises par Sr' pour l'ensemble des tableaux simulés.

Pour simplifier l'étude des simulations, on a regroupé les Sr' en 3 classes correspondant à différentes intensités de structure commune. Ces 3 classes ont été déterminées de façon empirique d'une part à partir des nombreuses simulations effectuées avec différentes hypothèses (changement de la configuration de départ, variation de σ) et d'autre part à partir de l'expérience acquise sur de nombreux jeux de données. Les Sr' sont regroupés de façon à représenter 3 cas :

- une structure commune très forte : $0 \leqslant Sr' \leqslant 0,43$;
- une structure commune d'une force comparable, à celle que l'on peut obtenir dans la réalité : $0,43 < Sr' \leqslant 0,59$;
- aucune structure commune : $Sr' > 0,59$.

Évolution des critères en fonction des paramètres de simulation (cf. figure 3)

On étudie la capacité de reconstitution des données d'une part et de la configuration de référence d'autre part, en fonction des méthodes d'imputation et des différents paramètres des simulations.

En fonction de la méthode d'imputation (figure 3a). Les différentes variantes de l'algorithme IRM sont meilleures que la méthode MI du point de vue des deux critères. Les variantes « barycentre » et « médiane » de l'algorithme IRM sont équivalente en terme de performance.

En fonction du nombre de valeurs manquantes. Comme attendu, les critères se dégradent en fonction du nombre de valeurs manquantes (figure 3b). Toutefois on s'attend, a priori, à mieux reconstruire la configuration de référence que l'on ne reconstitue les données. En effet, on sait que le premier plan d'une analyse factorielle est très robuste à des changements au sein des données. A moins de 4 valeurs manquantes, en moyenne, on reconstitue presque aussi bien les données que l'on reconstruit la configuration de référence. Entre 5 et 7 valeurs manquantes, on reconstruit sensiblement moins bien les données, mais le critère RV($\mathbf{Z}$,$\mathbf{Z}_c$) diminue peu.

Pour 9 valeurs manquantes, le coefficient RV entre la configuration de référence réelle et celle obtenue après imputation est en moyenne de 0,6. On obtient une représentation de référence après imputation qui contient les grands traits de l'analyse du tableau complet. Par contre, les données reconstituées peuvent être très différentes des données réelles.

En fonction de Sr'. Lorsqu'il y a une structure sous-jacente forte (figure 3b), c'est-à-dire lorsque $Sr' \leqslant 0.43$, en moyenne on reconstruit bien les données et la configuration de référence. On observe une dégradation sensible des deux critères lorsque les données présentent une structure sous-jacente moins importante ($0,43 < Sr' \leqslant 0,59$). Pour un $Sr' > 0.59$, on observe une dégradation de la reconstitution de la configuration de référence pour aboutir in fine à une meilleure reconstitution des données que de la configuration de référence. La raison en est que, pour ces données s'apparentant à du bruit pur, le premier plan de l'AFM est instable.

En fonction du modèle choisi pour les simulations. Pour des données issues des modèles APG (A) et INDSCAL (I), les valeurs des critères sont très proches (figure 3b). Le modèle APG est un cas particulier du modèle INDSCAL, donc du point de vue de l'AFM, qui peut être considérée comme une méthode d'estimation des paramètres du modèle INDSCAL, il n'y a pas de distinction dans les résultats entre les deux modèles. Ceci peut expliquer la proximité des résultats en terme de reconstitution de la configuration de référence.

En fonction du couple (Sr',n) (figure 3c). Pour une classe de Sr' fixé on observe la dégradation des deux critères en fonction du nombre de valeurs manquantes. Cette dégradation est peu marquée pour des Sr' faibles. Lorsque les données présentent une structure sous-jacente forte, jusqu'à 8 valeurs manquantes, on reconstruit correctement la configuration de référence alors que l'on reconstitue de moins en moins bien les données. Lorsque l'on a

une structure sous-jacente moins importante, jusqu'à 4 valeurs manquantes on reconstruit bien la configuration de référence alors que les données sont de moins en moins bien reconstituées. Au-delà de 5 valeurs manquantes la reconstruction des données est de moins en moins bonne et on a dans la même proportion une dégradation de la reconstruction de la configuration de référence. Lorsque l'on n'a pas de structure sous-jacente, $Sr' > 0,59$, les deux critères décroissent régulièrement et dans les mêmes proportions en fonction du nombre de valeurs manquantes.

Dans tous les cas on observe une très forte dégradation des deux critères pour 9 valeurs manquantes. Toutefois si la structure sous-jacente est très forte ($Sr' < 0.43$), on reconstitue tout de même correctement la configuration de référence.

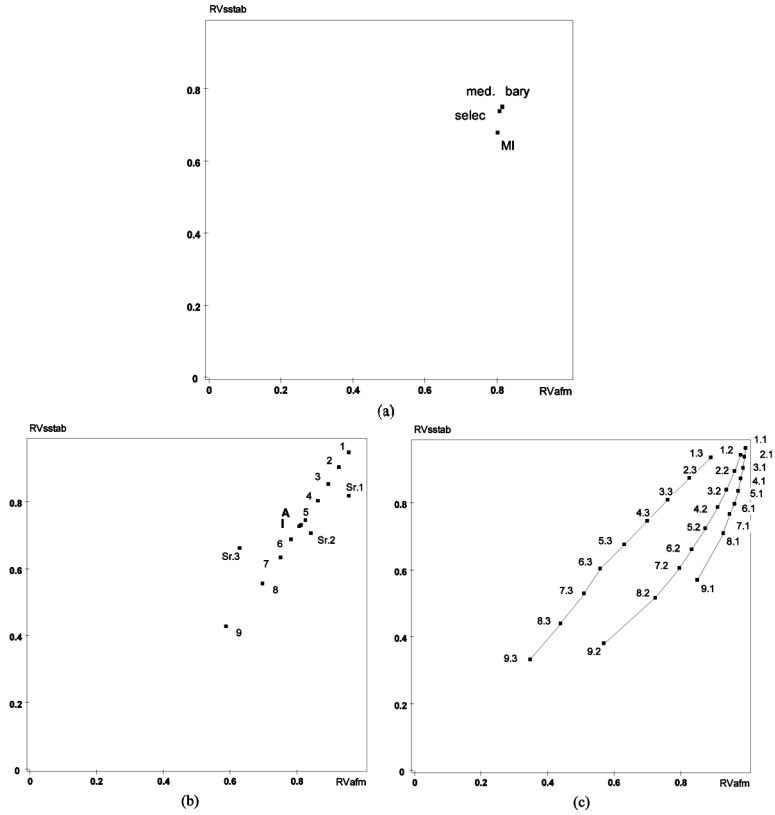


FIG 3. — Étude de l'évolution des RV (en abcisse reconstitution de la configuration de référence et en ordonnée reconstitution des données)

(a) par méthode ;

(b) par modèle (APG (A) ou INDSCAL (I)), nombre de valeurs manquantes, classe de Sr'. (la classe $Sr.3$ correspond à $Sr' > 0,59$ et $Sr.1$ à $Sr' < 0,43$) ;

(c) par couple (Sr', n) : 8.3 signifie $n = 8$ valeurs manquantes et $Sr' = Sr.3$.

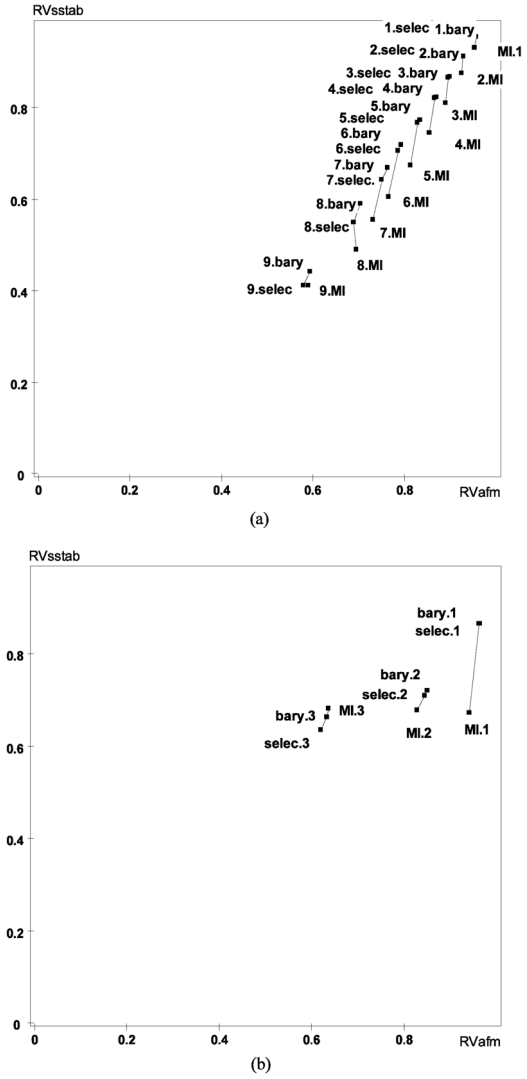


FIG 4. — Évolution des RV en fonction de la méthode d'imputation (reconstitution de la configuration de référence en abcisse et des données en ordonnée).

Les variantes barycentre et médiane de la méthode d'imputation par rotations multiples donnent des résultats identiques ; on ne représente donc que la variante barycentre (notée bary.).

(a) en fonction de n. Ex : 3.selec est le barycentre des points pour un remplacement par la variante sélection et 3 valeurs manquantes.

(b) en fonction de Sr. selec.1 est le barycentre des points pour un remplacement par la variante sélection et un $Sr' < 0,43$.

Comparaison des méthodes d'imputation (cf. figure 4)

En fonction du nombre de valeurs manquantes. Les deux critères se dégradent en fonction du nombre de valeurs manquantes et ce quelle que soit la méthode d'imputation employée. Dans l'ensemble les différentes variantes de l'algorithme IRM se comportent mieux que la méthode MI.

A 9 valeurs manquantes, il y a peu de différence entre la méthode MI et les variantes de l'algorithme IRM : toutes les méthodes d'imputations se comportent mal. Rappelons que lorsqu'on est en présence de 8 (50%) ou 9 valeurs manquantes (56%), les ajustements entre 2 configurations se font sur 4 voire 3 points communs ce qui est très peu. On atteint là une limite de la méthode d'imputation par rotations multiples. En effet, on considère empiriquement qu' il faut que les ajustements 2 à 2 se fassent au moins sur 5 ou 6 points.

En fonction de Sr' (figure 4b). Quelle que soit la valeur de Sr', la méthode MI reconstitue les données de la même manière. En effet, par construction, la méthode MI ne tient pas compte des autres configurations partielles et donc en particulier la capacité à reconstituer des données ne dépend pas de la valeur de Sr'. Pour cette méthode, seule la reconstruction de la configuration de référence varie.

Pour une structure sous-jacente forte ($Sr' < 0,43$), l'algorithme IRM reconstitue mieux les données que la méthode MI. Cette dernière méthode reconstitue un peu moins bien la configuration de référence que les différentes variantes de l'algorithme IRM. Lorsque Sr' est moyen l'algorithme IRM est meilleur que la méthode MI. Lorsque la structure sous-jacente est inexistante, les différentes méthodes se comportent sensiblement de la même façon. Il semble même que la variante sélection de l'algorithme IRM se comporte moins bien. La dégradation de performance de l'algorithme IRM en terme de reconstitution de données s'explique par le fait que l'algorithme IRM s'appuie sur l'hypothèse qu'il existe une structure sous-jacente aux données pour reconstituer les valeurs manquantes, or, lorsque $Sr' > 0,59$, cette hyptohèse n'est plus vérifiée.

D'après ces simulations, dans le cas d'une structure sous jacente forte, ($Sr' < 0.43$) quelle que soit la méthode d'imputation on va pouvoir reconstituer le premier plan de l'AFM. Si les données sont essentiellement constituées de bruit ($Sr' > 0,59$), il n'y a pas de structure sous-jacente et les méthodes d'imputation ne peuvent fonctionner. Dans les cas proches de la réalité, $0,43 < Sr' \leqslant 0,59$, l'algorithme IRM dans ces différentes variantes reconstitue mieux les données et la configuration de référence que la méthode MI.

D'après ces simulations, dans le cas d'une dégustation de 16 produits par 10 juges, la pratique consistant à ne pas faire déguster plus de 10 produits permettrait de bien reconstituer la configuration de référence.

4.4.2. Simulations à partir de données réelles

On étudie ici un cas réel de données recueillies par la méthode du napping : dix juges ont évalué 16 cocktails non alcoolisés à base de jus de fruits. Ces données sont regroupées au sein d'un tableau multiple de dimension $I = 16$,

$J = 10$ et $K = 2$. La statistique procustéenne généralisée, Sr', vaut 44%. Ces nappes ne sont pas identiques entre elles mais présentent une structure commune.

On enlève de $n = 1$ à $n = 9$ valeurs. Pour chaque valeur de n on simule 150 emplacements de valeurs manquantes. On dispose, pour chaque méthode d'imputation et chaque nombre n de valeurs manquantes, de 150 valeurs du $\mathrm{RV}(\mathbf{Z}, \mathbf{Z}_c)$ et du $\frac{1}{J}\sum_{j=1^J} RV(\mathbf{X}_j, \mathbf{X}_j^c)$. On représente la moyenne de ces 150 valeurs pour un critère par valeurs de n et type d'imputation. Les résultats sont regroupés figure 5.

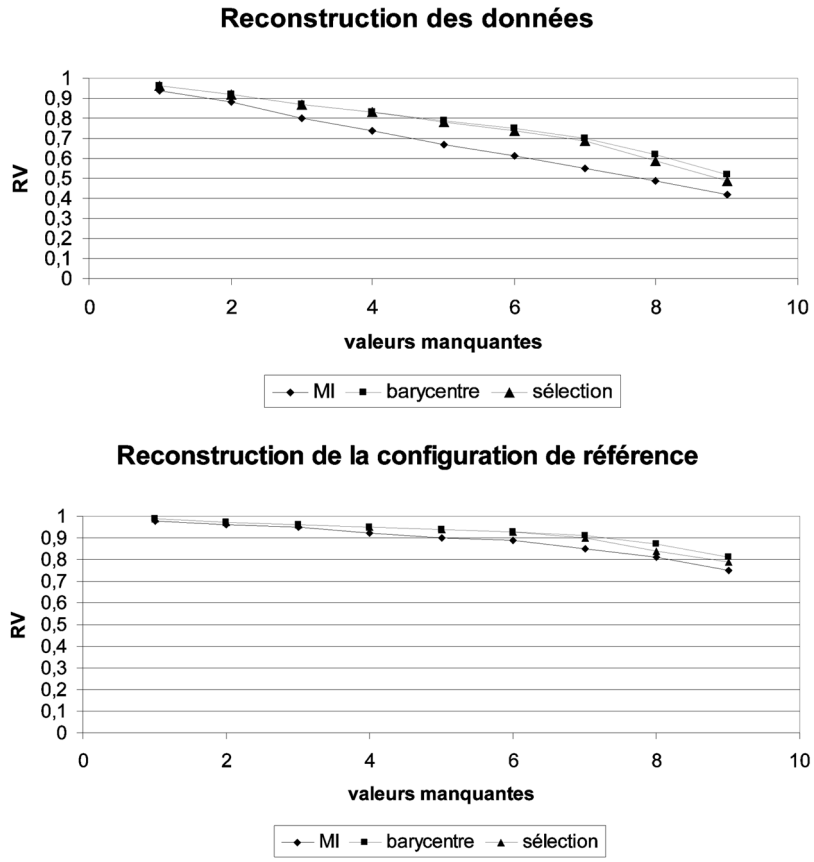


FIG 5. — Moyenne des RV en fonction des valeurs manquantes.

Les variantes barycentre et médiane de la méthode d'imputation par rotations multiples donnent des résultats identiques, on ne représente donc que la variante barycentre.

La méthode MI est moins performante en terme de reconstitution des données que l'IRM (figure 5 haut). La différence de performance est particulièrement marquée à partir de 4 cocktails manquants.

En terme de reconstruction de la configuration de référence (figure 5 bas), la méthode MI est aussi performante que la méthode IRM jusqu'à 3 valeurs manquantes. La méthode MI devient sensiblement moins performante à partir de 4 cocktails manquants.

Dans le cas de la reconstruction de la configuration de référence, le coefficient RV pour l'IRM décroit faiblement jusqu'à 6 valeurs manquantes (soit environ 40% de données manquantes).

4.5. Application de l'AFMP à un jeu de données réel

On s'intéresse au jeu de données précédent où 10 juges ont évalué 16 cocktails par la méthode du napping. Ces cocktails sont à base de jus d'orange, de citron, de nectar de mangue et de banane ainsi que d'un colorant alimentaire. Chacun des 16 cocktails correspond à une recette différente. Les différentes recettes ont été réalisées à partir d'un plan de mélange.

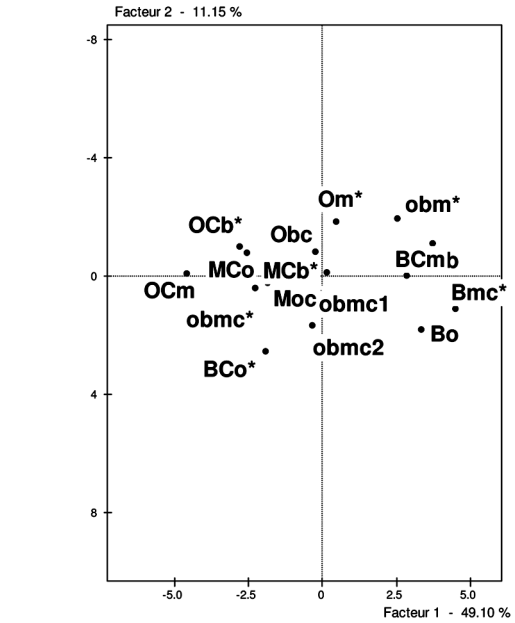
Ces données sont étudiées, dans un premier temps, par l'AFMP. Dans un deuxième temps, on simule, à partir de ces données, un jeu de données avec 4 cocktails manquants par nappe. Les données sont complétées par imputation par rotations multiples en utilisant comme variante de l'algorithme le barycentre. On compare les résultats des analyses exécutées sur le jeu de données réel et le jeu de donnée simulé. On prend comme configuration de référence pour l'AFMP les 2 premières composantes de l'AFM du tableau de données complet ($\mathbf{Z}$) ou complété ($\mathbf{Z}_c$).

4.5.1. Données complètes

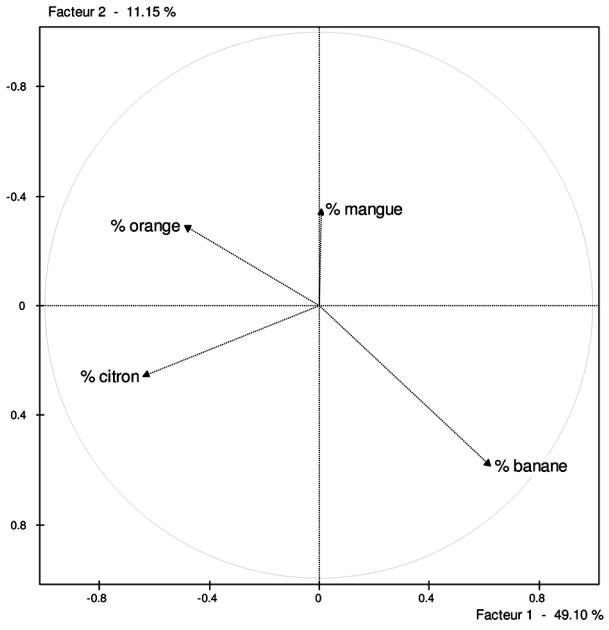
Les données sont regroupées au sein d'un tableau multiple de 10 (J) sous-tableaux de dimensions 16×2. Dans chaque sous-tableau, les variables sont centrées et non réduites. On ajoute à cette analyse, en groupe illustratif, la composition des cocktails.

Les résultats sur la représentation de référence sont regroupés au sein de la figure 6. Le premier axe de l'analyse est prépondérant (49.10% de l'inertie). Il oppose les produits à base de citron et d'orange aux autres. La seconde bissectrice sépare les produits à dominante banane des autres.

On souhaite ensuite connaître les juges qui ont fourni des nappes proches de cette disposition de référence des produits. On calcule le coefficient RV pour chaque juge entre sa configuration partielle et la configuration de référence. (voir tableau 2). Le juge 7 a fourni la représentation plane la plus proche de la configuration de référence. Le juge 3 a proposé la représentation la plus différente. À partir de la représentation superposée du juge 3 (figure 7), on note que celui-ci a regroupé les produits à base de citron et d'orange (OCm, OCb*, MCO, BCo*) à l'exception des produits Moc* et obmc* qui sont regroupés avec un produit, obm*, sans dominante.



(a)



(b)

FIG 6. — Configuration de référence dans le cas des données complètes.
Les différents cocktails sont désignés à l'aide d'un codage alphabétique rappelant la recette. Le plan de mélange est fourni en annexe.

TABLEAU 2. — Coefficients RV entre les configurations partielles et la configuration de référence. Les juges sont ordonnés de la nappe la plus proche à la plus éloignée de la configuration de référence.

juge	RV
7	0,834
2	0,805
4	0,724
8	0,696
9	0,676
1	0,632
6	0,576
10	0,547
5	0,526
3	0,476

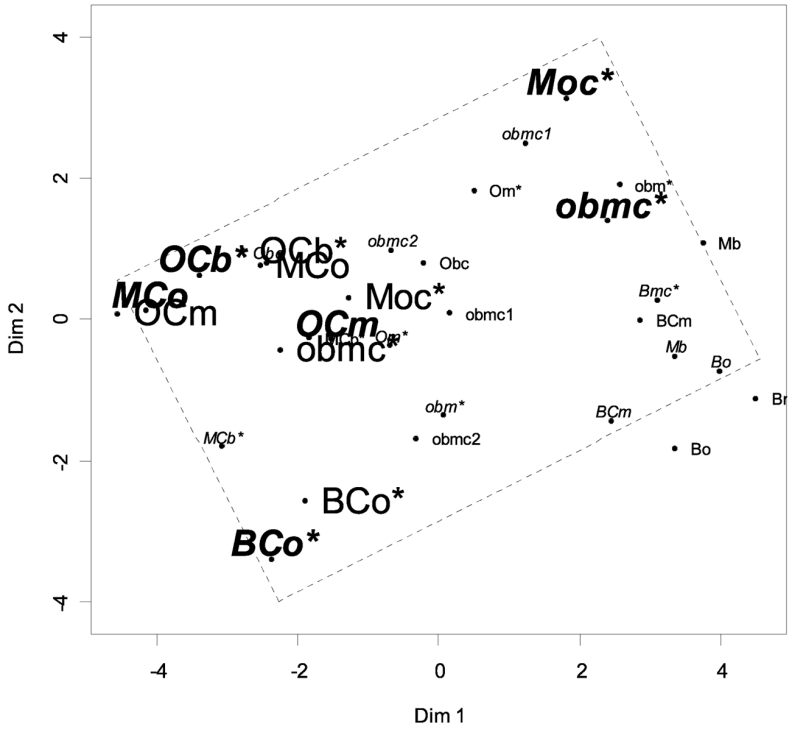
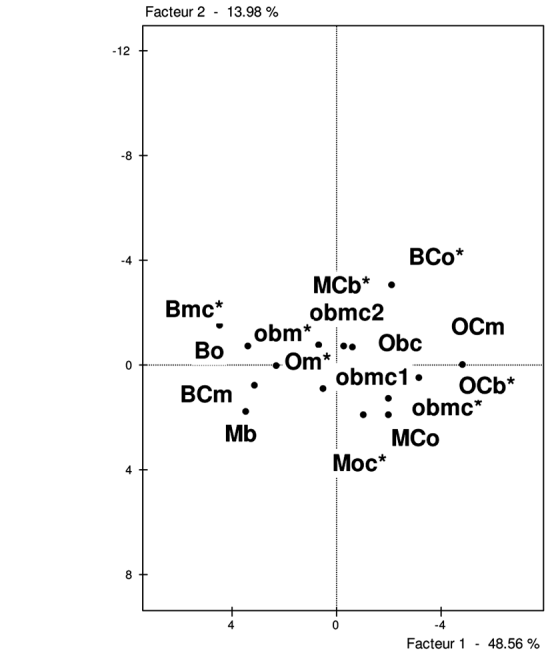
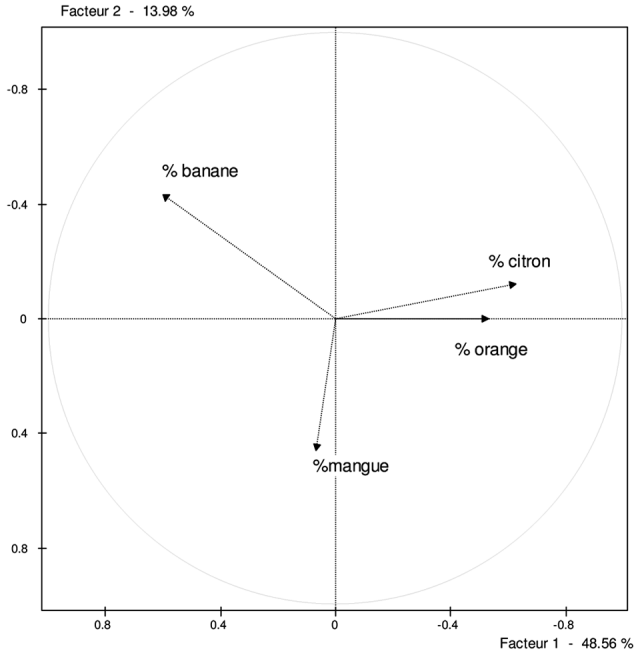


FIG 7. — Configuration du juge 3 par rapport à la configuration de référence. Les données du juge sont en italique (ex : *BCo**). Les produits repérés par une police plus importante sont commentés dans le texte. Le trait en pointillé rappelle le contour exact de la nappe du juge 3.



(a)



(b)

FIG 8. — Configuration de référence dans le cas des données incomplètes.

4.5.2. Données incomplètes

À partir du tableau de données complet précédent, on génère un tableau incomplet dans lequel chaque juge n'a dégusté que 12 produits. On compare l'analyse de ce tableau à celle du tableau complet décrit au paragraphe précédent.

Reconstitution de la configuration de référence. Le coefficient RV entre les configurations de référence vaut : $RV(\mathbf{Z}, \mathbf{Z}_c) = 0,94$. La configuration de référence après imputation est représentée figure 8. Le premier axe oppose, comme dans le cas complet, les produits à base de citron et d'orange aux autres. Toutefois, dans le cas incomplet (figure 8), l'axe 2 n'oppose pas l'orange au citron comme cela était le cas dans l'analyse du tableau complet.

Représentation superposée. À titre d'exemple, la configuration du juge 6 est étudiée (figure 9 en haut). Le juge 6 a positionné ses 12 produits en utilisant seulement une partie de la nappe (la moitié droite). Cette configuration est différente de la configuration de référence ($RV(\mathbf{X}_j^c, \mathbf{Z}_c) = 0,69$). En particulier, les produits à base de banane n'ont pas été séparés des autres. Par rapport au cas complet (voir tableau 2), le coefficient RV entre la configuration de référence et la configuration du juge 6 a sensiblement augmenté. Toutefois les RV standardisés sont du même ordre de grandeur (le RV standardisé vaut 5.62 dans le cas complet et 5.069 dans le cas incomplet) : cette augmentation du coefficient RV semble pouvoir être attribuée à la diminution du nombre de points de comparaison (12 points par configuration dans le cas incomplet contre 16 points dans le cas complet)

Reconstitution des configurations. En moyenne, $\frac{1}{10}\sum_{j=1}^{10} RV(\mathbf{X}_j, \mathbf{X}_j^c) = 0,827$. Dans un but méthodologique, on a représenté les seul points « estimés » par la méthode d'imputation (IRM variante barycentre) ainsi que les points réels correspondant pour le juge 6 (figure 9 en bas). Les produits Bmc* et BCo* ont été replacés très près de leur emplacement d'origine. En particulier, le produit Bmc* était placé sur le bord de la nappe, son estimation par l'IRM a elle aussi été positionnée loin du centre. La méthode IRM arrive pratiquement à reconstituer ce point « extrême ». A contrario, les deux autres points, eux aussi en bord de nappe, n'ont pu être reconstitués par l'IRM qui les a repositionnés au centre de la nappe. L'imputation par IRM permet pour les points extrêmes soit de les reconstituer comme l'imputation classique (produits OCb* et Moc*) soit de les repositionner à leur emplacement d'origine (produits BCo* et Bmc*).

Conclusion

La représentation superposée des nuages homologues obtenue par AFMP est particulièrement précieuse dans le cas bidimensionnel. Dans ce cas très particulier, les configurations partielles ne sont absolument pas déformées. Cette méthode permet un enrichissement de la représentation moyenne de l'AFM par une représentation superposée des configurations partielles contenant toute la spécificité de l'information initiale.

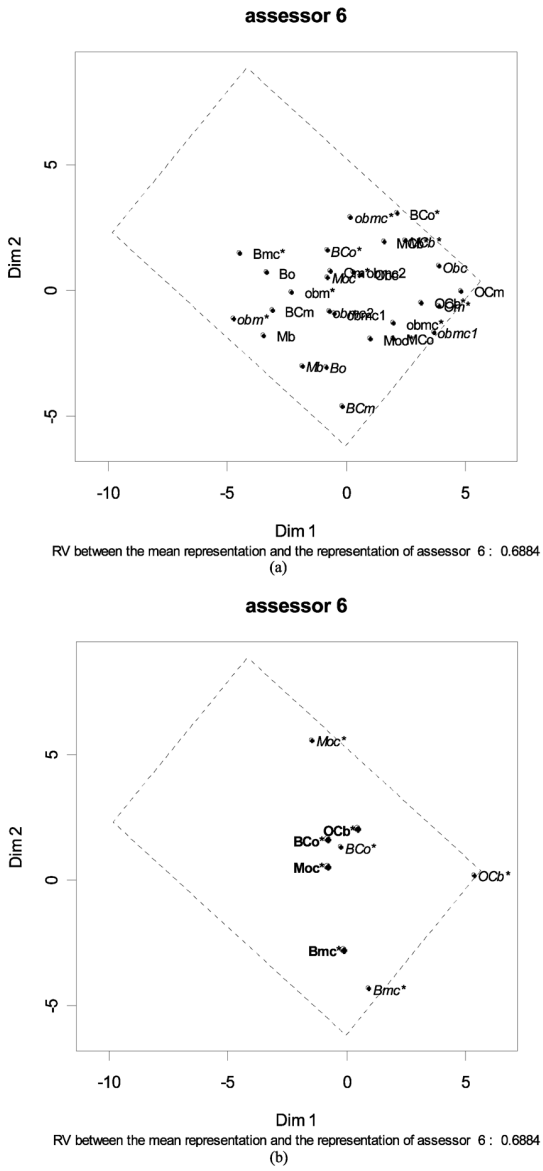


FIG 9. — Représentation superposée de la configuration du juge 6 et de la configuration de référence. En haut, configuration du juge 6 avec 4 données manquantes. En bas, représentation des seuls produits supprimés de la nappe du juge 6 (en italique) et de leurs estimations par la méthode d'imputation IRM (variante barycentre).

Souvent, on ne dispose pas de tous les points dans toutes les configurations partielles. Si ces données manquantes sont peu nombreuses, une imputation par la moyenne inconditionnelle donne de bons résultats. Sinon, et si en plus

il existe une structure sous-jacente aux configurations partielles suffisamment forte, l'algorithme proposé ici permet de reconstituer la représentation de référence avec une qualité suffisante en pratique.

Dans l'exemple du recueil par la méthode des nappes, le nombre de produits par nappe est généralement limité à 10. Avec la méthode proposée ici, il est possible à la fois de respecter cette contrainte et d'étudier des ensembles de produits plus vastes.

Logiciels

Le cas réel a été traité avec le package SensoMineR de R (la procédure pmfa disponible dans ce package contient l'implémentation de l'AFMP pour la méthode du napping®). Les différents modèles de simulation et les différentes méthodes d'imputation ont donné lieu à une programmation spécifique sous R.

Remerciements : Il est agréable de remercier ici C. Bosquillon De Jenlis, E. Le Coz et G. Sehan étudiantes, de la spécialité statistique appliquée d'Agrocampus Rennes, qui ont réalisé les cocktails et recueilli les données de napping®.

Annexes

Tableau 4. — Plan de mélange des recettes des cocktails.

Recette	%Orange	%Banane	%Mangue	%Citron	Colorant
MCb*	20	20	50	10	Oui
Mb	20	30	50	0	Non
BCm	20	50	20	10	Non
BmC*	20	50	25	5	Oui
MCo	40	0	50	10	Oui
BCo*	40	50	0	10	Oui
Moc*	45	0	50	5	Oui
obmc*	45	25	25	5	Oui
obmc1	45	25	25	5	Non
obmc2	45	25	25	5	Non
obm*	46,7	26,7	26,7	0	Oui
Bo	50	50	0	0	Non
OCm	70	0	20	10	Non
Om*	70	0	30	0	Oui
OCb*	70	20	0	10	Oui
Obc	70	25	0	5	Non

Références

Arnold G.M. et Williams A. A. (1986) *The use of generalised procrustes analysis in sensory analysis*, Statistical procedure in food research, Elsevier, p. 233-253.

Commandeur J.J.F. (1991) *Matching configurations*. DSWO press, Leiden.

Dijksterhuis G. B. et Punter P. (1990) Interpreting generalized procrustes analysis « Analysis of Variance » tables. *Food Quality and Preference*, 2, p. 255-265.

Dijksterhuis G. B. et Gower J.C. (1991) The interpretation of generalized procrustes analysis and allied methods. *Food Quality and Preference*, 3, p. 67-87.

Escofier B. et Pagès J. (1998) *Analyses factorielles simples et multiples*, Dunod.

Escoufier Y. (1973) Le traitement des variables vectorielles. *Biometrics*, 29, p. 751-760.

Gower J.-C. (1975) Generalized Procrustes Analysis. *Psychometrika*, 40, p. 33-51.

Kazi-Aoual F., Hitier S., Sabatier R. et Lebreton J.-D. (1995) Refined Approximation to Permutation Tests for Multivariate Inference. *Computational Statistics and Data Analysis*, 20, p. 643-656.

Krzanowski W. J. (1990) *Principles of Multivariate Analysis. A User's Perspective*, Oxford Statistical Science Series, Oxford.

Morand E. et Pagès J. (2003) Incorporation de rotations procrustéennes dans une analyse factorielle multiple procrustéenne. *Revue de Nouvelles Technologies de l'Information*, C1(1), p. 101-111.

Morand E. et Pagès J. (2006) Procrustes multiple factor analysis to analyse the overall perception of food products. *Food Quality and Preference*, 17, p. 36-42.

Pagès J. (2003) Recueil direct de distances sensorielles : application à l'évaluation de dix vins blancs de Val-de-Loire. *Science des Aliments*, 23(5/6), p. 679-888.

Qannari E.M, Macfie H.J.H et Courcoux P.(1999) Performance indices and isotropic scaling factors in sensory profiling. *Food Quality and Preference*, 10, p. 17-21.

R development core team (2004). R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http ://www.R-project.org.

Saporta G. (1990) *Probabilités analyse des données et statistiques*, Technip.

Schlich P. (1996) *Defining and validating assessor compromises about products distances and attribute correlations*, Multivariate Analysis of Data in Sensory Science, Elsevier, p259-306.

Sibson R. (1978) Studies in the robustness of multidimensional scaling : Procrustes Analysis. *J. Royal Statist. Soc. B*, 40, p. 234-238.

Ten Berge J.M.F. (1977) Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, 42(2), p. 267-276.