Incorporation de rotations procrustéennes dans une analyse factorielle multiple

Elisabeth Morand*, Jérôme Pagès*

*Laboratoire de mathématiques appliquées
Agocampus Rennes
65, rue de Saint-Brieuc
CS 84215
35042 Rennes cedex
morand@agrocampus-rennes.fr

Résumé. Pour comparer deux nuages de points homologues, la méthode de référence est l'analyse procrustéenne et, dans le cas de plus de deux nuages, l'Analyse Procrustéenne Généralisée (APG). L'Analyse Factorielle Multiple (AFM) fournit aussi une représentation superposée de nuages de points homologues. Cette dernière représentation bénéficie, par rapport à celle issue de l'APG, d'avantages (elle s'inscrit dans le cadre d'une analyse factorielle riche en aides à l'interprétation) et d'inconvénients (les nuages à comparer subissent des déformations autres que les seules projections et rotations).
Il est possible de compléter l'AFM par un ajustement procrustéen 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 procrustéen et s'inscrit dans le cadre de l'AFM.
Cette nouvelle représentation est précieuse lorsque les nuages initiaux sont bidimensionnels. Une application dans ce cas particulier est présentée.

1. Introduction

Un objectif classique de l'étude de tableaux multiples est la comparaison de plusieurs configurations d'un même ensemble de points. Pour réaliser cette comparaison, une méthode de référence est l'analyse procrustéenne généralisée (APG) [Gower, 1975]. Elle permet d'obtenir une représentation superposée mettant en évidence les traits communs aux différentes configurations et ce sans déformer les configurations initiales. L'analyse factorielle multiple [Escofier et Pagès, 1998] permet aussi de mettre en évidence des traits communs aux différentes configurations mais au prix d'une déformation des configurations initiales.

Après un rappel des notations utilisées et des caractéristiques de l'APG et de l'AFM, nous montrerons l'intérêt d'inclure des rotations procrustéennes dans l'AFM. On obtient alors une nouvelle méthodologie que nous nommons Analyse Factorielle Multiple Procrustéenne (AFMP). Une application comparée de l'AFM et de l'AFMP sur des analyses sensorielles de vins est décrite afin d'illustrer l'intérêt de cette méthodologie.

2. Données

Les données sont constituées d'un ensemble d'individus, $\{i\ ;\ i=1, I\}$, décrits par plusieurs groupes de variables. Ces données peuvent être regroupées sous forme d'un tableau unique structuré en sous-tableaux. On note (figure 1) :

- X le tableau complet ;
- K l'ensemble des variables ;
- J l'ensemble des sous-tableaux ;
- K_j l'ensemble des variables du groupe j ;
- X_j le tableau associé au groupe j.

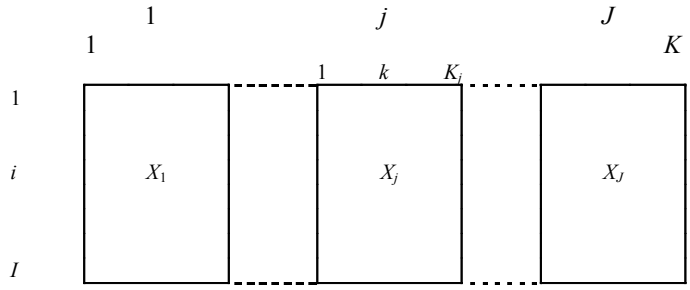

FIG. 1 - *Structure des données*

Au tableau X correspond le nuage des individus, noté N_I, situé dans l'espace R^K. A chaque groupe de variables, correspond un nuage d'individus, dit partiel et noté N_I^j, situé dans un espace de dimension K_j . Si l'on plonge le nuage N_I^j dans l'espace R^K, les coordonnées de chacun des individus de ce nuage se trouvent au sein du tableau, noté $\widetilde{X}_j$, de dimensions (I, K), dans lequel X_j est complété par des 0.

3. Analyse procustéenne généralisée

On se place, dans un premier temps, dans le cas de deux tableaux X_1 et X_2 . On suppose de plus que $K_1=K_2$ (cas auquel il est toujours possible de se ramener en complétant le plus petit des deux tableaux par des colonnes de 0). On cherche à superposer « au mieux » les deux nuages à l'aide de rotations orthogonales. Pour ce faire, on fixe X_1 et on cherche H rotation orthogonale ($HH'=I$) qui minimise la distance entre X_1 et X_2H :

$$Tr(X_1 - X_2H)(X_1 - X_2H)'$$

Dans ce cadre, il existe une solution analytique qui s 'appuie sur la décomposition en valeurs singulières de $X_2'X_1=VSU'$ [Saporta, 1990]. La solution est alors $H=VU'$.

Dans le cas général de J nuages (J>2) de même dimension, on peut écrire un modèle, dit procrustéen, selon lequel toutes les configurations se déduisent les unes des autres par des

rotations orthogonales ou, ce qui revient au même, d'une même configuration (*Z)* à une erreur près (E'_j). Soit :

$$X_j = ZH'_j + E'_j \quad \text{pour } j=1...J$$

ou encore :

$$X_j H_j = Z + E_j \quad \text{pour } j=1...J$$

Il n'existe pas de solution analytique à ce problème. On utilise donc une solution algorithmique. Dans cet algorithme, on pose Z moyenne des configurations à chaque pas.

Soit :

$$Z = \frac{1}{J} \sum_{k=1}^{J} X_j$$

A chaque pas, on effectue une procrustéenne de chaque configuration individuelle sur le consensus *Z*. On met alors à jour les configurations individuelles en les remplaçant par les configurations obtenues après rotation

et on minimise:

$$Tr \sum_{j=1}^{J} (X_j H_j - Z)(X_j H_j - Z)' \qquad (3.1)$$

Si $K_j > 2$, on effectue une ACP du consensus final *Z* et on projette les représentations individuelles sur les axes principaux du consensus.

Cet algorithme a été proposé par Gower en 1975. Une amélioration de cet algorithme a été proposée par Ten Berge en 1977.

4. Analyse factorielle multiple

Le cœur de cette analyse est constitué par une ACP effectuée sur le tableau complet *X*, dont les variables sont pondérées. La pondération utilisée consiste à diviser chaque variable du groupe j par la racine carrée de λ_1^j (en notant λ_1^j l'inertie projetée sur le premier axe de l'analyse séparée du groupe *j*). On obtient ainsi une représentation du nuage N_I , comme dans toute ACP, mais dans laquelle le rôle des groupes a été équilibré. A cette représentation, on superpose les nuages N_I^j en introduisant les tableaux $\tilde{X}_j$ en supplémentaires dans l'ACP du tableau complet *X*. Cette représentation présente quelques propriétés intéressantes, en particulier :

- elle s'inscrit dans une méthode générale qui fournit de nombreux points de vue sur l'analyse simultanée de plusieurs tableaux en particulier de nombreuses aides à l'interprétation ;
- il existe pour cette représentation des relations de transition dites partielles (détaillées ci-après).

La coordonnée, sur l'axe principal de rang *s*, de l'individu *i* vu par le groupe *j*, notée $F_s(i^j)$, s'exprime comme combinaison linéaire des coordonnées des seules variables du groupe *j* sur ce même axe. Ceci se traduit par la formule de transition suivante, dans laquelle on reconnaît la restriction au groupe *j* de la relation de transition classique :

$$F_s(i^j) = F_s^j(i) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\sqrt{\lambda_1^j}} \sum_{k \in K_j} x_{ik} G_s(k)$$

en notant :

- $F_s(i^j)$ projection de l'individu i^j sur l'axe principal de rang s du nuage des individus N_I ;
- $G_s(k)$ projection sur l'axe de rang s de la variable k ;
- λ_s est l'inertie projetée du nuage N_I .

En contrepartie de cette précieuse propriété, cette représentation présente des déformations autres que celles induites par les projections. En effet, chaque nuage partiel est projeté sur deux axes n'appartenant pas à son sous-espace initial (figure 2).

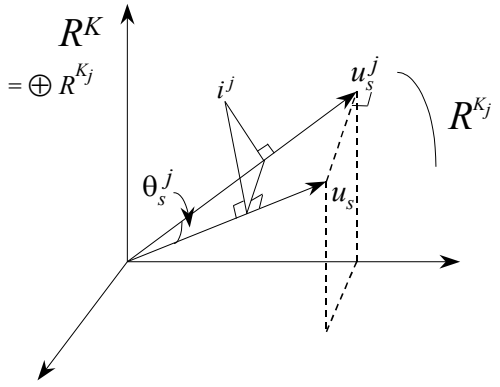

FIG. 2 - *Projection de l'individu i du nuage j dans N_I*

u_s : axe principal de rang s du nuage des individus
u_s^j : composante de u_s dans l'espace du nuage partiel j.
i^j, $i^{ème}$ ligne de $\widetilde{X}_j$ et appartenant à R^{K_j}, est dans un premier temps projeté sur u_s^j puis, en multipliant les coordonnées par cos(θ_s^i), sur u_s

Nous proposons ci-après une représentation procrustéenne à la fois intégrée à l'AFM, d'où la dénomination d'analyse factorielle multiple procrustéenne (AFMP), et telle que les nuages initiaux ne subissent aucune autre déformation que celles résultant des projections.

Toutefois, il est à noter que dans le cas où les nuages partiels sont dans un espace à plus de deux dimensions, la représentation procrustéenne, bien que non déformée dans l'espace total, le sera lors de sa représentation finale en deux dimensions, par projection sur un espace de dimension inférieure.

5. Analyse factorielle multiple procrustéenne

5.1 Principe

On calcul N_I nuage moyen de l'AFM. Chaque nuage partiel N_I^j est ensuite ajusté, à l'aide d'une rotation procrustéenne, sur le nuage moyen N_I.

On peut utiliser soit l'intégralité de la représentation moyenne de l'AFM soit ses S premières composantes. Le choix de la dimension, S, est à discuter suivant les cas. Le tableau ainsi obtenu, de dimension (I, S), est noté $\tilde{F}$. On considère ensuite, pour chaque groupe j, le sous-tableau X_j de X (figure 1.), pondéré comme en AFM, l'expression des sous tableaux est donc $\frac{1}{\sqrt{\lambda_1^j}} X_j$. Nous cherchons à les rapprocher de $\tilde{F}$ par une rotation procrustéenne de $\frac{1}{\sqrt{\lambda_1^j}} X_j$ sur $\tilde{F}$. Cela revient donc (voir paragraphe 3.1) à chercher une rotation orthogonale H_j qui minimise $Trace[(\tilde{F} - \frac{1}{\sqrt{\lambda_1^j}} X_j H_j)(\tilde{F} - \frac{1}{\sqrt{\lambda_1^j}} X_j H_j)']$.

On obtient alors les tableaux «procrustéanisés», $\hat{X}_j$, par les relations suivantes :

$$\hat{X}_j = \frac{1}{\sqrt{\lambda_1^j}} X_j H_j$$

où $H_j = V_j U_j'$ avec :

V_j la matrice orthogonale des vecteurs propres normés de la matrice : $\frac{1}{\lambda_1^j} X_j' \tilde{F} \tilde{F}' X_j$

U_j la matrice orthogonale des vecteurs propres normés de la matrice : $\frac{1}{\lambda_1^j} \tilde{F}' X_j X_j' \tilde{F}$

Ce calcul revient à effectuer la dernière boucle de l'algorithme de Gower (1975) en prenant comme consensus le nuage moyen issu de l'AFM.

Remarque : Nous avons travaillé ici sous l'hypothèse que tous les sous-tableaux étaient de même dimension (K_j constant). C'est ce cas particulier, avec $K_j = S = 2$, qui a suscité l'AFMP et dont on a évalué l'impact pratique.

Toutefois, pour les cas où K_j ne serait pas constant d'autres stratégies sont envisageables et méritent d'être examinées comme :

- prendre $K' = \max K_j$ et compléter les tableaux, de dimension inférieure à K', par $K' - K_j$ colonnes de 0 ;
- prendre K_j composantes principales du consensus ; on ajuste alors chaque configuration individuelle sur les $S = K_j$ premières composantes de la représentation

moyenne de l'AFM, la dimension *S* étant différente d'une configuration individuelle à l'autre ;
- se limiter à un nombre fixe de composantes principales par tableaux.

5.2 Propriétés

5.2.1 Représentation superposée

Pour la représentation des nuages partiels obtenue en AFMP, il n'y a plus de relations de transition partielles. En contrepartie, les nuages partiels n'ont subi aucune déformation autre que les rotations orthogonales.

Remarque. Cette représentation est particulièrement intéressante dans le cas bidimensionnel, puisque la représentation des nuages partiels n'est absolument pas déformée.

5.2.2 Critère

Dans le cadre de l'analyse procrustéenne généralisée, le critère que l'on minimise est l'inertie intra configuration (voir (3.1)). En revanche, dans le cadre de l'AFMP, cet aspect n'est pas pris en compte dans la construction. On souhaite donc savoir comment la construction de l'AFMP se comporte vis-à-vis de ce critère. Une approche empirique a été utilisée. Ainsi, les deux méthodes ont été appliquées sur plusieurs jeux de données et à chaque fois ce critère a été calculé afin de comparer la valeur du critère pour une AFMP à celle obtenue pour une APG réalisée sur les mêmes données.

Les analyses procrustéennes généralisées ont été réalisées avec la fonction « procGPA » contenue dans le package shapes du logiciel R. De plus, chaque analyse procrustéenne généralisée a été effectuée sur des données pondérées comme en AFM, et ce afin de comparer les aspects algorithmiques et non les aspects de pondération.

Les résultats sont regroupés au sein du tableau 1. On peut remarquer que ce critère est du même ordre de grandeur pour les deux analyses, les deux méthodes sont comparables du point de vue de ce critère. De plus, on peut penser que les représentations obtenues avec les deux méthodes soient ressemblantes.

Jeu de données	I	K_j	J	APG	AFMP
1	10	2	3	25.26078	25.28557
2	10	2	4	29.57061	29.86092
3	10	2	4	40.22659	43.89492
4	10	2	5	50.15532	51.61395
5	10	2	6	42.8738	44.24953
6	10	2	4	50.58664	51.14354
7	10	2	5	45.53425	46.84337

TAB 1 – *Comparaison de critère pour l'APG et l'AFMP*

6. Exemple

On présente ici un exemple dans le cadre où tous les sous tableaux sont de dimension 2 (K_j=K'=2) pour illustrer l'intérêt de la nouvelle méthode par rapport à la représentation superposée usuelle de l'AFM.

6.1 Données

On a demandé à 11 dégustateurs de fournir chacun une représentation euclidienne de 10 vins blancs de Val de Loire (5 Chenins, numérotés de 1 à 5, et 5 Sauvignons, numérotés de 6 à 10) en les positionnant sur une nappe. On dispose alors de J=11 représentations euclidiennes. Ainsi, des vins proches sur une nappe sont des vins qui paraissent similaires au juge. Les données analysées sont les coordonnées $X_j(i)$ et $Y_j(i)$ de chaque vin i, mesurées sur la nappe du juge j.

		X_1		.....	X_j		.....	X_{11}	
		x_1	y_1		x_j	y_j		x_{11}	y_{11}
	1								
vins	i	$x_1(i)$	$y_1(i)$		$x_j(i)$	$y_j(i)$		$x_{11}(i)$	$y_{11}(i)$
I=10									

FIG. 3 - *Structure des données de l'exemple*

A titre d'exemple, nous représentons ici la nappe fournie par le juge 9. Remarquons en particulier, sur cette nappe, les vins 7 et 10 relativement excentrés.

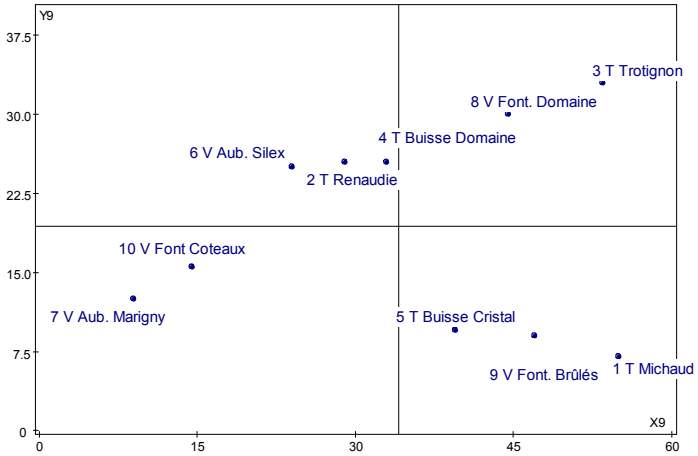

FIG. 4 - Représentation des 10 vins par le juge 9

6.2 Résultats de l'AFM

Ces données sont traitées par une AFM dans laquelle chaque nappe constitue un groupe de deux variables. De plus, pour conserver l'importance relative de l'abscisse et de l'ordonnée, les données ne sont pas réduites.

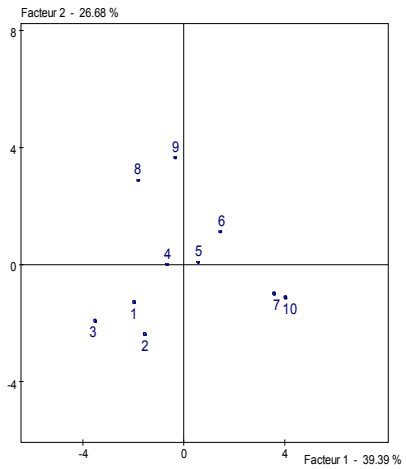

FIG. 5A - Représentation du nuage moyen de l'AFM

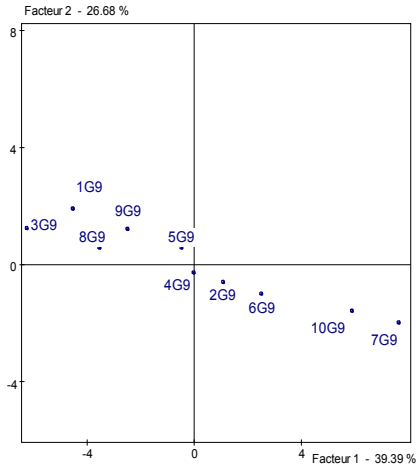

FIG. 5B - Représentation du nuage partiel du juge 9 pour l'AFM

Sur la représentation moyenne des individus ainsi obtenue (figure 5a) il est commode d'orienter les commentaires selon les deux bissectrices. La première bissectrice sépare les Sauvignons (vins 1 à 5) des Chenins (vins 6 à 10). La seconde bissectrice sépare les 5 Chenins entre eux en mettant en évidence la particularité des vins 7 et 10.

Sur la représentation partielle du juge 9 fournie par l'AFM (figure 5b) les vins 8 et 9 sont opposés aux vins 7 et 10 ce qui est en accord avec la représentation initiale. La dispersion quasi nulle le long de la première bissectrice renvoie à la non distinction des Chenins et des Sauvignons. Mais on peut remarquer que les vins 8 et 9 sont très proches sur cette représentation alors que dans les données initiales ils étaient séparés. Cette représentation a déformé le nuage initial. Elle est étirée le long de la seconde bissectrice essentiellement pour faire apparaître ses points communs avec la représentation globale.

6.4 Résultats de l'AFMP

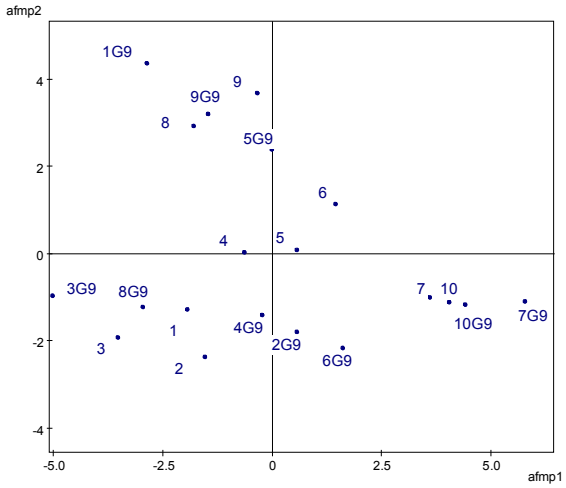

FIG. 6 - *Représentation superposée du nuage moyen et du nuage partiel obtenu par rotation procrustéenne du nuage partiel du juge 9 sur le nuage moyen*

Sur ces mêmes données, on effectue des rotations procrustéennes des nuages partiels sur les deux premières dimensions du nuage moyen. Par rapport à la représentation précédente, la représentation ainsi obtenue est, à une rotation près, la configuration fournie par le juge 9 (figure 4). On retrouve dans cette représentation du nuage partiel du juge 9 (figure 6) le long de la seconde bissectrice, une configuration des chenins proche de celle de la représentation globale, en particulier on a bien les vins 7 et 10 opposés aux vins 8 et 9. En revanche, on peut observer la séparation des vins 8 et 9 comme sur les données initiales alors que dans la configuration moyenne ces vins sont regroupés. Cette séparation n'était pas visible dans la représentation partielle fournie par l'AFM car cette séparation est propre au juge 9 et n'existe pas sur la représentation globale.

Les distinctions faites par le juge 9, par exemple l'opposition $\{1,9,5\} \leftrightarrow \{3,8,4\}$, apparaissent par construction dans l'AFMP mais n'apparaissent pas dans l'AFM.

7. Conclusion

La représentation superposée de nuages de points homologues décrite ici est fondée sur l'utilisation d'une méthode usuelle l'AFM complétée par des rotations procrustéeennes. Elle est particulièrement précieuse dans le cas bidimensionnel. En effet dans ce cas précis les nuages partiels ne sont absolument pas déformés. Ceci permet donc un enrichissement de la représentation moyenne de l'AFM par une représentation des nuages partiels contenant toute la spécificité de la configuration partielle initiale.

8. Bibliographie

[Escofier et Pagès, 1998] B. Escofier , J.Pagès, *Analyses factorielles simples et multiples ; objectifs méthodes et interprétation*, 284p, Dunod, Paris, 1998.

[Gower, 1975] J.C Gower, Generalized Procrustes Analysis*, Psychometrika*, vol.40, n°1 , p. 33-51,1975.

[Pagès, 2003] J. Pagès, Recueil direct de distances sensorielles : application à l'évaluation de dix vins blancs du Val-de-Loire*, Sciences des aliments*, vol.23, n°5/6 , p. 679-688,1975.

[Saporta, 1990] G.Saporta, *Probabilités, analyse des données et statistique*, p 195, Editions Technip, Paris, 1990.

[Ten Berge, 1977] Jos M. F. Ten Berge, Orthogonal procrustes rotations for two or more matrices, *Psychometrika*, vol.42, n°2 , p. 267-276,1977.

R Development Core Team (2003). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.

Summary

To compare two homologous clouds of points, the method mainly used 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 superposed representation of several homologous clouds of points. This latter representation gets, against the one stemmed from, advantages (it is included in the framework of a factor analysis rich in viewpoints in analysis of several clouds of points) and drawbacks (the clouds to compare experience different distortions than the only projections and rotations)

We can complement the MFA by a procrustes adjustment of each initial cloud on the average cloud of the MFA. We thus get representation of those clouds which both respect the procrustes pattern and is included into the framework of the MFA.

This new representation is especially very useful when the initial clouds are two-dimensional.

An application of this particular case is presented.