Une solution analytique pour la rotation planaire en Analyse Factorielle des Correspondances Multiples

Marie Chavent, Vanessa Kuentz, Jérôme Saracco

Une solution analytique pour la rotation planaire en Analyse Factorielle des Correspondances Multiples

Marie Chavent[1,2] & Vanessa Kuentz[1,2] & Jérôme Saracco[1,2,3]

[1] *Université de Bordeaux, IMB, CNRS, UMR 5251,*
351 Cours de la libération
33405 Talence
(e-mail : Marie.Chavent,Vanessa.Kuentz@math.u-bordeaux1.fr*)*

[2] *INRIA Bordeaux Sud-Ouest, CQFD team*

[3] *Université Montesquieu - Bordeaux IV, GREThA, CNRS, UMR 5113,*
Avenue Léon Duguit
33608 Pessac Cedex
(e-mail : Jerome.Saracco@u-bordeaux4.fr*)*

Résumé

L'Analyse en Composantes Principales (ACP) et l'Analyse Factorielle des Correspondances Multiples (AFCM) sont respectivement deux méthodes de description statistique multidimensionnelle de données quantitatives et qualitatives. Une rotation peut ensuite être appliquée à la matrice des scores des composantes principales. La définition d'un critère de rotation permet alors d'obtenir une structure simple, facilitant ainsi l'interprétation des résultats. Une solution analytique en deux dimensions a été proposée pour le critère varimax en ACP. Nous proposons ici une solution analytique en deux dimensions pour la rotation en AFCM utilisant un critère inspiré de varimax et basé sur la notion de rapport de corrélation.

Mots-clés : Analyse Factorielle des Correspondances Multiples, rotation, rapport de corrélation.

Abstract

Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) are well-known multivariate methods for statistical description of respectively quantitative and qualitative data. A rotation may then be applied to the principal component scores matrix. The definition of a criterion enables to achieve simple structure, thereby simplifying the interpretation of the results. A two-dimensional analytic solution has been proposed for varimax rotation in PCA. We propose here a two-dimensional analytic solution for rotation in MCA using a varimax-based criterion relying on correlation ratio.

Keywords: Multiple Correspondence Analysis, rotation, correlation ratio.

1 Introduction

L'Analyse en Composantes Principales (ACP) (voir par exemple Jolliffe, 2002) pour des données quantitatives et l'Analyse Factorielle des Correspondances Multiples pour des données qualitatives (AFCM) (voir par exemple Greenacre et Blasius, 2006) sont deux méthodes usuelles de réduction de dimension. Elles utilisent une décomposition en valeurs singulières pour écrire la matrice des données comme le produit de deux matrices. Cependant cette approximation n'est pas unique, ainsi les composantes principales sont déterminées à une rotation près. Nous traiterons ici le cas de rotations orthogonales.

Concernant les variables quantitatives, l'ACP approxime la matrice standardisée des données par le produit de la matrice des composantes principales et la matrice des saturations ("loadings"). Cette dernière joue un rôle fondamental dans l'interprétation des résultats car elle contient les corrélations entre les variables et les composantes. Appliquer la matrice de rotation $\mathbf{T}$ à la matrice des saturations $\mathbf{A}$ et à la matrice des composantes principales garantit que la matrice des saturations après rotation $\mathbf{B} = \mathbf{AT}$ contient toujours les corrélations entre les variables et les composantes après rotation. L'idée est d'appliquer une rotation à la matrice des saturations et à celle des composantes afin que les corrélations contenues dans la matrice $\mathbf{B}$ soient fortes ou faibles. Différents critères ont été proposés dont le plus connu est varimax introduit par Kaiser (1958). Définir la matrice de rotation optimale se résume donc à un problème d'optimisation sous contraintes du critère varimax $f(\mathbf{AT})$.

Pour les variables qualitatives, l'AFCM est présentée ici comme une ACP appliquée à la matrice des profils lignes. La matrice des fréquences (divisée par les poids des lignes et des colonnes) est approximée par le produit de deux matrices : la matrice des composantes principales des lignes (objets) et celle des composantes principales des colonnes (modalités). Notons $\mathbf{A}$ cette dernière matrice et $\mathbf{B} = \mathbf{AT}$ sa version après rotation. Cependant $\mathbf{B}$ ne relie pas directement les variables aux composantes principales après rotation et par conséquent ne joue pas le rôle précédent de la matrice des saturations. Pour cette raison, le critère varimax f n'est pas appliqué directement à la matrice $\mathbf{B}$ (comme en ACP) mais à la matrice $\mathbf{C} = g(\mathbf{B})$ dont les valeurs sont les rapports de corrélation entre les variables et les composantes principales des objets après rotation. Notons que ce critère est celui utilisé par Kiers (1991) dans sa méthode PCAMIX pour la recherche d'une structure simple pour des données quantitatives et/ou qualitatives. L'objectif est d'obtenir dans chaque colonne de $\mathbf{C}$ des valeurs élevées ou faibles. Définir la meilleure matrice de rotation se résume donc à un problème d'optimisation du critère $f \circ g(\mathbf{AT})$.

Dans le cas quantitatif et qualitatif, lorsqu'on retient deux composantes principales, la rotation a lieu dans un plan. La matrice de rotation s'écrit en fonction d'un angle de rotation et on obtient un problème d'optimisation réel non contraint. En ACP, l'expression analytique de l'angle θ optimisant le critère varimax f est donnée par Kaiser (1958). Nous proposons ici une solution analytique en deux dimensions pour le critère $h = f \circ g$ pour la rotation en AFCM.

Dans la Section 2, nous présentons la rotation en AFCM puis nous donnons dans la Section 3 l'écriture de la solution analytique en dimension deux.

2 Rotation en AFCM

Notations. Soit $\mathbf{X}$ une matrice de données qualitatives, où $x_{ij} \in \mathcal{M}_j$ avec $\mathcal{M}_j$ l'ensemble des modalités de x_j. Soit $\mathbf{O} = (o_{is})_{n\times q}$ la matrice $\mathbf{X} = (x_{ij})_{n\times p}$ convertie en une matrice indicatrice à n lignes et q colonnes, où $q = \sum_{j=1}^{p} q_j$ avec q_j le nombre de modalités de x_j. A chaque ligne i de $\mathbf{O}$, un élément vaut 1 si l'objet possède la modalité s de la variable qualitative correspondante ; sinon l'élément vaut 0. Ainsi la somme des éléments d'une ligne vaut p. Comme la matrice $\mathbf{O}$ peut être considérée comme une sorte de table de contingence, la matrice des fréquences $\mathbf{F} = (f_{is})_{n\times q}$ peut être construite où $f_{is} = \frac{o_{is}}{np}$ car $\sum_{i,s} o_{is} = np$. Soit $\mathcal{G}_s = \{e_i \in \mathcal{E} | o_{is} = 1\}$ et $n_s = \text{card}(\mathcal{G}_s)$. Les sommes marginales de la matrice des correspondances $\mathbf{F}$ sont utilisées pour définir les poids des lignes et des colonnes : $\mathbf{D}_n = \text{diag}\{f_{i.}, \text{pour } i = 1, ..., n\}$ avec $f_{i.} = \frac{1}{n}$, et $\mathbf{D}_q = \text{diag}\{f_{.s}, \text{pour } s = 1, ..., q\}$ avec $f_{.s} = \frac{n_s}{np}$. Notons m le rang de $\mathbf{F}$ et $\widetilde{\mathbf{F}} = \mathbf{D}_n^{1/2}\mathbf{F}\mathbf{D}_q^{-1/2}$ la matrice $\mathbf{F}$ dont les éléments ont été divisés par $\sqrt{f_{i.}f_{.s}}$.

AFCM. L'AFCM est présentée ici comme une ACP sur la matrice des profils lignes $\mathbf{R} = \mathbf{D}_n^{-1}\mathbf{F}$ obtenue en divisant chaque ligne i de la matrice des fréquences $\mathbf{F}$ par la marge $f_{i.}$. D'un point de vue géométrique, l'AFCM de la matrice non centrée des profils lignes étudie les n lignes de $\mathbf{R}$ avec les métriques $\mathbf{D}_n$ sur $\mathbb{R}^n$ et $\mathbf{D}_q^{-1}$ sur $\mathbb{R}^q$. Dans une première étape, l'AFCM cherche un axe de vecteur directeur $\mathbf{w}$ (de $\mathbf{D}_q^{-1}$-norme égale à 1) tel que le vecteur $\mathbf{y} \in \mathbb{R}^n$ des $\mathbf{D}_q^{-1}$-projections des n lignes de $\mathbf{R}$ sur cet axe, soit de $\mathbf{D}_n$-norme maximale. La première composante principale est $\mathbf{y}_2 = \mathbf{R}\mathbf{D}_q^{-1/2}\mathbf{v}_2$ où $\mathbf{v}_2$ est le vecteur propre associé à la première valeur propre non triviale λ_2 de $\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}$. Les autres composantes principales sont définies de façon similaire par $\mathbf{y}_\alpha = \mathbf{R}\mathbf{D}_q^{-1/2}\mathbf{v}_\alpha$, $\alpha = 3 \dots m$, où $\mathbf{v}_\alpha$ est le vecteur propre associé à la αème plus grande valeur propre de $\widetilde{\mathbf{F}}^t\widetilde{\mathbf{F}}$.

Formule de reconstruction en AFCM. Notons $\mathbf{Y}$ la matrice de dimension (n, m) dont les colonnes sont les m composantes principales (incluant celle associée à la première valeur propre triviale) et $\Lambda = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_m})$ la matrice diagonale de leurs écarts-types. Cette matrice de correspondance divisée par les poids des lignes et des colonnes $\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_q^{-1}$ peut s'écrire comme le produit de deux matrices :

$$\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_q^{-1} = \mathbf{Y}^*\mathbf{A}^t\,, \tag{1}$$

où $\mathbf{Y}^* = \mathbf{Y}\Lambda^{-1}$ est la matrice dont les colonnes sont les composantes principales standardisées et $\mathbf{A}$ est la matrice dont le terme général $a_{s\alpha}$ vaut $\bar{\mathbf{y}}^*_{\alpha,s}$ la moyenne de la composante

principale standardisée $\mathbf{y}_\alpha^*$ calculée sur les objets possédant la modalité s. Des détails concernant cette formule sont disponibles dans Chavent, Kuentz, Saracco (2009).

Structure simple. Soit $\mathbf{T}$ une matrice de rotation orthonormale de dimension (r, r) où $r \leq m$ est le nombre de composantes principales retenues. Les matrices $\mathbf{Y}^*$ et $\mathbf{A}$ sont désormais de dimension respective (n, r) et (q, r). Comme $\mathbf{Y}^*\mathbf{A}^t = \mathbf{Y}^*\mathbf{T}\mathbf{T}^t\mathbf{A}^t$, l'approximation de la matrice $\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_q^{-1}$ par le produit de deux matrices n'est pas unique. Notons $\mathbf{Z} = \mathbf{Y}^*\mathbf{T}$ la matrice des composantes principales standardisées après rotation et $\mathbf{B} = \mathbf{A}\mathbf{T}$ la matrice $\mathbf{A}$ après rotation dont le terme général $b_{s\alpha}$ vaut maintenant $\bar{\mathbf{z}}_{\alpha,s}$ la moyenne de la composante principale standardisée après rotation $\mathbf{z}_\alpha$ calculée sur les objets appartenant à la modalité s. La matrice $\mathbf{B}$ ne donne pas directement une idée de la liaison entre les variables et les composantes principales après rotation. On transforme donc la matrice $\mathbf{B} = \mathbf{A}\mathbf{T}$ en $\mathbf{C} = g(\mathbf{A}\mathbf{T})$ de dimension (p, r) dont le terme général :

$$c_{j\alpha} = p||b_\alpha^{(j)}||^2_{\mathbf{D}_q^{(j)}}, \tag{2}$$

où l'exposant (j) signifie qu'on considère seulement les éléments de $\mathbf{D}_q$ et b_α correspondant aux modalités de la variable x_j. Ainsi $\mathbf{D}_q^{(j)}$ (resp. $b_\alpha^{(j)}$) est une sous-matrice (resp. sous-vecteur) de $\mathbf{D}_q$ (resp. b_α) de dimension (q_j, q_j) (resp. longueur q_j). On montre alors que $c_{j\alpha} = \eta^2(\mathbf{x}_j, \mathbf{z}_\alpha) = \frac{s_j^2}{s^2(\mathbf{z}_\alpha)}$, où $s_j^2 = \frac{1}{n}\sum_{s\in\mathcal{M}_j} n_s(\bar{\mathbf{z}}_{\alpha,s} - \bar{\mathbf{z}}_\alpha)^2$ est la variance inter-classe empirique de $\mathbf{z}_\alpha$ dans la partition de $\mathcal{E}$ définie par les modalités de x_j. Ainsi $\mathbf{C}$ est la matrice des rapports de corrélation empiriques entre les variables x_j et les composantes principales standardisées après rotation z_α. Elle va jouer le rôle de la matrice des saturations dans le cadre de la rotation varimax en ACP. L'idée est donc de choisir la matrice $\mathbf{T}$ pour que ces rapports de corrélation soient les plus proches possibles de 0 ou de 1.

Le critère de rotation. Le critère varimax f n'est pas appliqué directement à $\mathbf{B} = \mathbf{A}\mathbf{T}$, comme en ACP, mais à la matrice des rapports de corrélation $\mathbf{C} = g(\mathbf{A}\mathbf{T})$. Le critère h ainsi obtenu est défini par :

$$\begin{aligned}
h(\mathbf{A}\mathbf{T}) = f \circ g(\mathbf{A}\mathbf{T}) = f(\mathbf{C}) &= \sum_{\alpha=1}^{q}\left\{\frac{\sum_{j=1}^p c_{j\alpha}^2}{p} - \left(\frac{\sum_{j=1}^p c_{j\alpha}^2}{p}\right)^2\right\} \\
&= \sum_{\alpha=1}^{q}\left\{\frac{\sum_{j=1}^p (p\sum_{s\in\mathcal{M}_j} f_{.s} b_{s\alpha}^2)^2}{p} - \left(\frac{\sum_{j=1}^p (p\sum_{s\in\mathcal{M}_j} f_{.s} b_{s\alpha}^2)}{p}\right)^2\right\} \\
&= \sum_{\alpha=1}^{q}\left\{p\sum_{j=1}^{p}\left(\sum_{s\in\mathcal{M}_j} f_{.s} b_{s\alpha}^2\right)^2 - \left(\sum_{j=1}^{p}\sum_{s\in\mathcal{M}_j} f_{.s} b_{s\alpha}^2\right)^2\right\} \qquad (3)
\end{aligned}$$

Pour une matrice $\mathbf{A}$ donnée, le problème d'optimisation s'écrit donc :

$$\left\{ \begin{array}{l} \max\limits_{\mathbf{T}} h(\mathbf{AT}), \\ \text{s.c. } \mathbf{TT}^t = \mathbb{I}_r. \end{array} \right. \tag{4}$$

3 Une solution analytique pour la rotation en dimension deux

En dimension $r = 2$, la matrice de rotation orthogonale $\mathbf{T}$ s'écrit en fonction d'un angle de rotation θ :

$$\mathbf{T} = \left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array} \right]. \tag{5}$$

Le problème d'optimisation (4) s'écrit alors comme un problème non contraint :

$$\left\{ \max_{\theta \in \mathbb{R}} h(\theta). \right. \tag{6}$$

L'expression de $h(\theta)$ et les détails du calcul dre la solution analytique sont disponibles dans Chavent, Kuentz, Saracco (2009). On écrit la dérivée de h par rapport à θ :

$$\frac{\partial h}{\partial \theta} = 2(a + b\cos(4\theta) + c\sin(4\theta)), \tag{7}$$

où :

$$\begin{aligned} a &= (p-1)\sum_{j=1}^{p} \sum_{s \in \mathcal{M}_j} \sum_{t \in \mathcal{M}_j} f_{.s} f_{.t} \alpha_{st} \beta_{st} - \sum_{j=1}^{p} \sum_{l \neq j}^{p} \sum_{s \in \mathcal{M}_j} \sum_{t \in \mathcal{M}_l} f_{.s} f_{.t} \alpha_{st} \beta_{st}, \\ b &= (p-1)\sum_{j=1}^{p} \sum_{s \in \mathcal{M}_j} \sum_{t \in \mathcal{M}_j} f_{.s} f_{.t} \delta_{st} \gamma_{st} - \sum_{j=1}^{p} \sum_{l \neq j}^{p} \sum_{s \in \mathcal{M}_j} \sum_{t \in \mathcal{M}_l} f_{.s} f_{.t} \delta_{st} \gamma_{st}, \\ c &= (p-1)\sum_{j=1}^{p} \sum_{s \in \mathcal{M}_j} \sum_{t \in \mathcal{M}_j} f_{.s} f_{.t} \frac{1}{2}(\gamma_{st}^2 - \delta_{st}^2) - \sum_{j=1}^{p} \sum_{l \neq j}^{p} \sum_{s \in \mathcal{M}_j} \sum_{t \in \mathcal{M}_l} f_{.s} f_{.t} \frac{1}{2}(\gamma_{st}^2 - \delta_{st}^2), \end{aligned} \tag{8}$$

et $\alpha_{st} = a_{s1}a_{t1} + a_{s2}a_{t2}, \beta_{st} = a_{s2}a_{t1} - a_{s1}a_{t2}, \gamma_{st} = a_{s2}a_{t1} + a_{s1}a_{t2}$ et $\delta_{st} = a_{s1}a_{t1} - a_{s2}a_{t2}$.

Pour résoudre $a + b\cos(4\theta) + c\sin(4\theta) = 0$, on divise chaque terme par $(b^2 + c^2)^{1/2}$ et on introduit l'angle $\varphi \in]-\pi, +\pi]$tel que $\cos(\varphi) = \frac{b}{(b^2+c^2)^{1/2}}$ et $\sin(\varphi) = \frac{c}{(b^2+c^2)^{1/2}}$. On obtient :

$$\frac{a}{(b^2 + c^2)^{1/2}} + \cos(\varphi)\cos(4\theta) + \sin(\varphi)\sin(4\theta) = \frac{a}{(b^2 + c^2)^{1/2}} + \cos(4\theta - \varphi) = 0$$

Cette équation possède alors deux solutions :

$$\hat{\theta} = \frac{1}{4}(\pm \arccos(-\frac{a}{\sqrt{b^2+c^2}}) + \varphi), \tag{9}$$

correspondant respectivement au minimum et maximum de h (sous la condition que $|a| \leq (b^2+c^2)^{1/2}$, ce qui est nécessairement vérifiée car h est périodique et dérivable).

4 Conclusion

Nous avons proposé une solution analytique en deux dimensions pour la rotation en AFCM. Nous illustrerons sur des données simulées la validité de l'approche. Nous montrerons l'intérêt de la rotation pour l'interprétation des résultats d'une AFCM sur des données réelles issues d'une enquête de satisfaction des plaisanciers sur le Canal des Deux Mers (commanditée par Voies Navigables de France en 2008). Des détails sur ces applications sont disponibles dans Chavent, Kuentz, Saracco (2009). En outre ce résultat pourra être utilisé pour étendre une approche de type VARCLUS pour la classification de variables qualitatives.

Bibliographie

[1] Chavent, M., Kuentz, V., Saracco, J. (2009), A two-dimensional analytic solution for rotation in Multiple Correspondence Analysis, *Submitted paper*.
[2] Greenacre, M.J. et Blasius, J. (2006), *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC Press, London.
[3] Jolliffe, I.T. (2002), *Principal Component Analysis (Second Edition)*, Springer-Verlag, New York.
[4] Kaiser, H.F. (1958), The varimax criterion for analytic rotation in factor analysis, *Psychometrika*, 23(3), 187–200.
[5] Kiers, H.A.L. (1991), Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables, *Psychometrika*, 56, 197–212.