

REVUE DE STATISTIQUE APPLIQUÉE

R. LAFOSSE

M. HANAFI

Concordance d'un tableau avec K tableaux : définition de $K + 1$ uples synthétiques

Revue de statistique appliquée, tome 45, n° 4 (1997), p. 111-126

http://www.numdam.org/item?id=RSA_1997__45_4_111_0

© Société française de statistique, 1997, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

CONCORDANCE D'UN TABLEAU AVEC K TABLEAUX : DÉFINITION DE $K + 1$ UPLES SYNTHÉTIQUES

R. Lafosse (1), M. Hanafi (2)

(1) *Lab. de Statistique et Probabilités, UMR C55830 Université Paul Sabatier,
118 rte de Narbonne, 31062 Toulouse Cedex 4 - France lafosse@cict.fr*

(2) *Écologie des Eaux Douces et des Grands Fleuves, ESA-CNRS 5023,
Université Lyon 1 69622 Villeurbanne Cedex - France. hanafi@biomserv.univ-lyon1.fr*

RÉSUMÉ

Une généralisation à plus de deux tableaux de l'analyse inter batterie de Tucker (1958) est proposée en adoptant un point de vue non symétrique, l'un des tableaux étant pris pour référence. Un choix particulier du tableau référence permet de définir une analyse symétrique se rapprochant d'analyses connues.

Mots-clés : *Corrélation linéaire, Analyses factorielles, Concordance entre matrices.*

ABSTRACT

The interbattery Tucker's analysis (1958) is extended over two matrices with an asymmetric viewpoint, one of the matrices being a reference matrix. A particular choice of the reference matrix leads to a symmetric analysis close to known analyses.

Keywords : *Linear correlation. Factor analyses. Agreements between matrices.*

1. Introduction

Les termes de l'analyse de Tucker (1958) sont utilisés pour définir de nouvelles analyses relatives à deux tableaux de variables : régressions PLS initialisées par Wold (1975), analyse de co-inertie de Chessel et Mercier (1993) définissant des cartes factorielles simultanées des individus et des variables, ou analyse de la concordance d'un tableau avec un autre (Lafosse, 1997).

Cette dernière analyse étudie le rôle joué par l'un ou l'autre des deux tableaux dans la concordance fondée sur les couples de Tucker et l'objet de cet article est de pouvoir prolonger la notion de concordance d'un tableau avec un autre en une concordance d'un tableau avec plusieurs autres tableaux, tous les tableaux ayant le même nombre de lignes.

Dans un premier temps est proposée une définition non symétrique des couples de Tucker. On s'appuie sur une notion de composante partenaire liée au critère de covariance.

Cette définition prolongée à $K + 1$ tableaux permet d'associer par des corrélations privilégiées K composantes relatives à K tableaux à une composante définie au préalable sur le $K + 1^{\text{ème}}$. Ainsi une notion de monogamie proposée dans l'analyse de la concordance d'un tableau avec un autre tableau se prolonge-t-elle en une notion de polygamie. Une analyse de concordance sera donc fondée sur les propriétés des $K + 1$ uples successifs de composantes polygames.

On vérifie que ces propriétés concourent à rendre assez isolables les uns des autres ces $K + 1$ uples. Une problématique multivariée relative aux données de départ pourrait ainsi être décomposée en études séparées plus simples, chacune à réaliser entre une composante et ses K partenaires.

Un choix particulier du $K + 1^{\text{ème}}$ tableau, permet de transformer l'analyse en une analyse symétrique de K tableaux. Le premier $K + 1$ uple alors défini dans cette analyse est celui de l'analyse en composantes principales généralisée de Casin (1996) ou encore celui de l'analyse de co-inertie multiple (ACOM) de Chessel et Hanafi (1996). Une tentative de comparaison nous conduit à considérer qu'il s'agit de trois analyses bien différentes, malgré les nombreuses analogies relevées entre l'ACOM et cette analyse de concordance symétrique.

Pour simplifier l'exposé, les métriques implicitement considérées dans les espaces d'individus sont les métriques identité. D'autres métriques pourraient cependant être considérées, tout comme cela est fait en analyse de co-inertie généralisée. Cela conduirait à définir des analyses différentes de celle définie dans ce papier, de type analyse explicative ou analyse canonique, tout comme cela est fait en concordance dans le cas de deux tableaux (Lafosse, 1997). D'autres structures que celles ici considérées sont alors mises en concordance.

2. Analyse de Tucker

2.1. Introduction

On rappelle ici comment Tucker (1958) introduit son analyse inter batterie. Soient $\mathbf{X}$ $n \times p_1$ et $\mathbf{Y}$ $n \times p_2$ deux matrices. Les ensembles respectifs de p_1 et p_2 colonnes définissent deux ensembles de variables centrées mesurées sur un même ensemble de n individus.

La matrice diagonale des poids des individus est notée $\mathbf{D}$.

On note $\mathbf{u}_i$ des vecteurs normés de $\mathbb{R}^{p_1}$ de coefficients de combinaisons linéaires des variables de $\mathbf{X}$ et $\mathbf{v}_i$ des vecteurs normés de $\mathbb{R}^{p_2}$ de coefficients de combinaisons linéaires des variables de $\mathbf{Y}$.

L'analyse de Tucker est fondée sur la définition de couples successifs de facteurs communs $(\mathbf{X}\mathbf{u}_i, \mathbf{Y}\mathbf{v}_i)$ caractérisant une invariance des scores des individus lorsqu'on passe de $\mathbf{X}$ à $\mathbf{Y}$.

L'ensemble de tous les scores est contenu dans la matrice $n \times (p_1 + p_2)$ concaténée des deux tableaux. L'analyse des scores des individus d'un tableau de variables centrées se fait habituellement par l'analyse en composantes principales (ACP) de cette matrice. La matrice de variances-covariances de la matrice concaténée contient celle de $\mathbf{X}$, celle de $\mathbf{Y}$, et aussi la matrice des inter covariances $\mathbf{X}'\mathbf{D}\mathbf{Y}$ qui possède l'information sur l'analogie de scores entre $\mathbf{X}$ et $\mathbf{Y}$. Par suite, Tucker définit les couples synthétiques de l'information contenue dans $\mathbf{X}'\mathbf{D}\mathbf{Y}$, optimaux au sens d'Eckart et Young (1936), à partir de la décomposition en valeurs singulières de $\mathbf{X}'\mathbf{D}\mathbf{Y}$.

On considérera cette décomposition sous la forme :

$$\mathbf{X}'\mathbf{D}\mathbf{Y} = \mathbf{U}\mathbf{\Delta}\mathbf{V}',$$

avec $\mathbf{U}$ qui est $p_1 \times r$, $\mathbf{V}$ qui est $p_2 \times r$ et $\mathbf{\Delta}$ qui est $r \times r$, r étant le rang de $\mathbf{X}'\mathbf{D}\mathbf{Y}$.

Les systèmes orthonormés $\{\mathbf{u}_i\}$ et $\{\mathbf{v}_i\}$ étant ceux formés respectivement des colonnes de $\mathbf{U}$ et $\mathbf{V}$, les couples de composantes obtenus par Tucker sont ainsi les couples $(\mathbf{X}\mathbf{u}_i, \mathbf{Y}\mathbf{v}_i)$. Les vecteurs $\mathbf{u}_i$ et $\mathbf{v}_i$ vérifient donc les équations :

$$\mathbf{X}'\mathbf{D}\mathbf{Y}\mathbf{v}_i = \lambda_i \mathbf{u}_i \quad \text{et} \quad \mathbf{Y}'\mathbf{D}\mathbf{X}\mathbf{u}_i = \lambda_i \mathbf{v}_i, \quad (1)$$

où la valeur positive $\lambda_i = \text{cov}(\mathbf{X}\mathbf{u}_i, \mathbf{Y}\mathbf{v}_i)$ constitue le $i^{\text{ème}}$ terme diagonal de $\mathbf{\Delta}$.

Des équations (1) se déduisent les équations aux valeurs propres :

$$\mathbf{X}'\mathbf{D}\mathbf{Y}\mathbf{Y}'\mathbf{D}\mathbf{X}\mathbf{u}_i = \lambda_i^2 \mathbf{u}_i \quad \text{et} \quad \mathbf{Y}'\mathbf{D}\mathbf{X}\mathbf{X}'\mathbf{D}\mathbf{Y}\mathbf{v}_i = \lambda_i^2 \mathbf{v}_i. \quad (1 \text{ bis})$$

On dira que les vecteurs de $\{\mathbf{u}_i\}$ sont vecteurs singuliers à gauche de $\mathbf{X}'\mathbf{D}\mathbf{Y}$ et que ceux de $\{\mathbf{v}_i\}$ sont vecteurs singuliers à droite. Par la suite on considère les r couples $(\mathbf{u}_i, \mathbf{v}_i)$ comme définissant l'ensemble unique de toutes les solutions, sachant que les couples $(-\mathbf{u}_i, -\mathbf{v}_i)$ constituent aussi une solution (un ordre de multiplicité supérieur à 1 d'une valeur singulière conduisant à une infinité de solutions).

L'ACP est donc le point de départ pris par Tucker pour définir ses composantes. On propose ci-après une autre introduction à l'analyse de Tucker partant d'une définition des composantes de l'ACP.

2.2. Composantes principales d'un tableau et composantes de Tucker

On rappelle ici une définition de composantes de l'ACP introduite à partir d'un critère de covariance. On se sert ensuite de cette approche pour proposer une introduction à l'analyse de Tucker fondée également sur le même type de critère de covariations.

Considérons le problème de la recherche d'une composante $\mathbf{X}\mathbf{a}$, avec $\mathbf{a}$ normé dans $\mathbb{R}^{p_1}$, synthétisant le système de covariations des variables centrées $\mathbf{x}_j$ d'un tableau centré $\mathbf{X}$ $n \times p_1$. Pour obtenir une composante traduisant le plus les covariations internes de $\mathbf{X}$, on cherche à définir la composante ayant les covariations les plus fortes avec les variables de $\mathbf{X}$. On considère ainsi le problème de l'optimisation

du critère :

$$f_1(\mathbf{a}) = \sum_{j=1}^{p_1} \text{cov}^2(\mathbf{X}\mathbf{a}, \mathbf{x}_j).$$

Comme $f_1(\mathbf{a}) = \mathbf{a}'\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{a}$, la solution est bien donnée par le premier axe principal $\mathbf{a}_1$ de l'ACP du triplet statistique $(\mathbf{X}, \mathbf{I}_{p_1}, \mathbf{D})$.

Comme pour les critères de définition plus usuels de l'ACP, l'optimum du critère f_1 est invariant par isométrie dans $\mathbb{R}^{p_1}$. Ainsi par exemple il ne change pas quand $\mathbf{X}$ est remplacé par $-\mathbf{X}$.

Maintenant on se pose le problème de définir un premier couple $(\mathbf{X}\mathbf{a}, \mathbf{Y}\mathbf{b})$, $\mathbf{a}$ et $\mathbf{b}$ étant respectivement normés dans $\mathbb{R}^{p_1}$ et $\mathbb{R}^{p_2}$, synthétisant une communauté de structures des tableaux $\mathbf{X}$ et $\mathbf{Y}$ relative à chacun des deux systèmes de covariations internes. Une façon simple de procéder d'après la définition précédente consiste à vouloir simultanément que la composante $\mathbf{X}\mathbf{a}$ caractérise le système de covariations des variables $\mathbf{y}_k$ colonnes de $\mathbf{Y}$, et que $\mathbf{Y}\mathbf{b}$ caractérise le système de covariations de $\mathbf{X}$. C'est ainsi que l'on considère le critère à optimiser :

$$f_2(\mathbf{a}, \mathbf{b}) = \sum_{j=1}^{p_1} \text{cov}^2(\mathbf{Y}\mathbf{b}, \mathbf{x}_j) + \sum_{k=1}^{p_2} \text{cov}^2(\mathbf{X}\mathbf{a}, \mathbf{y}_k). \quad (2)$$

Comme

$$f_2(\mathbf{a}, \mathbf{b}) = \mathbf{a}'\mathbf{X}'\mathbf{D}\mathbf{Y}\mathbf{Y}'\mathbf{D}\mathbf{X}\mathbf{a} + \mathbf{b}'\mathbf{Y}'\mathbf{D}\mathbf{X}\mathbf{X}'\mathbf{D}\mathbf{Y}\mathbf{b}, \quad (2 \text{ bis})$$

d'après (1 bis) la solution est bien donnée par le premier couple de Tucker $(\mathbf{X}\mathbf{u}_1, \mathbf{Y}\mathbf{v}_1)$, celui associé à la plus grande valeur singulière de $\mathbf{X}'\mathbf{D}\mathbf{Y}$.

Avec l'écriture 2 bis du critère, il apparaît immédiatement que les couples de Tucker, et l'optimum, sont indépendants de toute rotation préalablement effectuée sur $\mathbf{X}$ dans $\mathbb{R}^{p_1}$ ou sur $\mathbf{Y}$ dans $\mathbb{R}^{p_2}$.

On remarque que le critère :

$$f_3(\mathbf{a}, \mathbf{b}) = \left[\sum_{j=1}^{p_1} \text{cov}^2(\mathbf{Y}\mathbf{b}, \mathbf{x}_j) \right] \left[\sum_{k=1}^{p_2} \text{cov}^2(\mathbf{X}\mathbf{a}, \mathbf{y}_k) \right]$$

aurait pu tout aussi bien être posé, puisqu'il mène aussi à la même solution.

L'approche non symétrique du problème proposée ci-après est une façon de lever cette ambiguïté.

2.3. Définition non symétrique des couples de Tucker

On s'inspire du critère (2) pour proposer une façon non symétrique d'introduire les couples de Tucker.

Dans un premier temps, il s'agit de rechercher une composante $\mathbf{Y}\mathbf{b}$ pouvant caractériser au mieux le système de covariations généré par les variables $\mathbf{x}_j$ du

tableau $\mathbf{X}$, $j = 1, \dots, p_1$. C'est ainsi que le critère (2) est maintenant remplacé par le critère suivant proposé par Tenehaus (1995) dans une introduction non symétrique à l'analyse de Tucker :

$$g(\mathbf{b}) = \sum_{j=1}^{p_1} \text{cov}^2(\mathbf{Yb}, \mathbf{x}_j) = \mathbf{b}'\mathbf{Y}'\mathbf{D}\mathbf{X}\mathbf{X}'\mathbf{D}\mathbf{Y}\mathbf{b}. \quad (3)$$

Sous contrainte que $\mathbf{b}$ soit de norme 1, la solution est évidemment la composante $\mathbf{Yv}_1$, c'est-à-dire la composante du premier couple de Tucker associée au tableau $\mathbf{Y}$.

La composante $\mathbf{Yv}_1$ ayant été jugée comme le meilleur indicateur que l'on puisse trouver pour résumer à partir de $\mathbf{Y}$ l'ensemble des covariations de $\mathbf{X}$, on se demande maintenant s'il existe une composante $\mathbf{Xa}$ qui jouerait pour $\mathbf{X}$ le rôle de correspondant naturel de $\mathbf{Yv}_1$.

Une composante $\mathbf{Yb}$ constitue un ensemble de coordonnées de points formant un nuage d'inertie $\mathbf{b}'\mathbf{Y}'\mathbf{D}\mathbf{Y}\mathbf{b}$. De la même façon une composante $\mathbf{Xa}$ est associable à un nuage projeté et à la part d'inertie de $\mathbf{X}$ égale à $\mathbf{a}'\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{a}$. Toute base orthonormée de $\mathcal{R}^{p_1}$ découpe l'inertie de $\mathbf{X}$ en parts distinctes. Deux composantes $\mathbf{Xa}_1$ et $\mathbf{Xa}_2$ seront dites relatives à deux parts d'inertie distinctes de $\mathbf{X}$ quand $\mathbf{a}_1$ et $\mathbf{a}_2$ sont orthonormés.

Par suite on considère qu'une composante $\mathbf{Xa}$ s'associe tout particulièrement à $\mathbf{Yb}$ et caractérise dans $\mathbf{X}$ ce que représente la variable $\mathbf{Yb}$, quand elle vérifie les deux propriétés suivantes :

- $\mathbf{Xa}$ est corrélée à $\mathbf{Yb}$.
- Toute autre composante du tableau $\mathbf{X}$ associée à une part d'inertie distincte de $\mathbf{a}'\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{a}$ est de corrélation nulle avec $\mathbf{Yb}$.

De la sorte cette part d'inertie $\mathbf{a}'\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{a}$ s'associe de façon privilégiée à la part $\mathbf{b}'\mathbf{Y}'\mathbf{D}\mathbf{Y}\mathbf{b}$.

La deuxième propriété revient à dire que l'on veut :

$$\mathbf{b}'\mathbf{Y}'\mathbf{D}\mathbf{X}\mathbf{a}_i = 0, \quad \forall \mathbf{a}_i \perp \mathbf{a}$$

ce qui entraîne

$$\mathbf{b}'\mathbf{Y}'\mathbf{D}\mathbf{X}\mathbf{a}_i\mathbf{a}_i' = 0, \quad \forall \mathbf{a}_i \perp \mathbf{a}$$

soit encore :

$$\mathbf{b}'\mathbf{Y}'\mathbf{D}(\mathbf{X} - \mathbf{X}\mathbf{a}\mathbf{a}') = 0.$$

C'est donc définir le vecteur normé $\mathbf{a}$ à partir du vecteur normé $\mathbf{b}$ par la relation :

$$\mathbf{X}'\mathbf{D}\mathbf{Y}\mathbf{b} = \mathbf{a}\mathbf{a}'\mathbf{X}'\mathbf{D}\mathbf{Y}\mathbf{b} = \mathbf{a} \text{ cov}(\mathbf{Xa}, \mathbf{Yb}). \quad (4)$$

Le vecteur $\mathbf{a}$ est donc tel que la composante $\mathbf{Xa}$ soit celle la plus covariée possible avec $\mathbf{Yb}$. De là provient l'intérêt porté dans cet article aux critères basés sur la covariance.

La composante $\mathbf{Xa}$ existe, sauf quand aucune variable de $\mathbf{X}$ n'est corrélée à $\mathbf{Yb}$. Nous dirons que $\mathbf{Xa}$ ainsi définie à partir de $\mathbf{Yb}$ est la *composante partenaire* de $\mathbf{Yb}$ dans $\mathbf{X}$.

D'après (1), $\mathbf{Xu}_1$ est la composante partenaire de $\mathbf{Yv}_1$ dans $\mathbf{X}$.

Le premier couple $(\mathbf{Xu}_1, \mathbf{Yv}_1)$ est donc défini à partir d'une composante de $\mathbf{Y}$ synthétisant les covariations de $\mathbf{X}$ et de la composante partenaire. Cette définition pourrait convenir à l'introduction de la notion de monogamie considérée en analyse de la concordance d'un tableau avec un autre (Lafosse, 1997).

D'après (4), on remarque que le lien entre la composante $\mathbf{Yb}$ et sa composante partenaire $\mathbf{Xa}$ s'effectue à partir de l'opérateur $\mathbf{W_XD} = \mathbf{XX'D}$:

$$\mathbf{W_XDYb} = \text{cov}(\mathbf{Xa}, \mathbf{Yb})\mathbf{Xa}. \quad (5)$$

Cette remarque prend un sens particulier quand on travaille avec la métrique de Mahalanobis $(\mathbf{X'X})^{-1}$ au lieu de la métrique implicite identité, la matrice $\mathbf{W_XD}$ devenant alors celle d'un projecteur sur le sous-espace engendré par les variables. Par la suite, les relations de type (5) sont fournies chaque fois que cela s'avère intéressant.

3. Une généralisation de l'analyse de Tucker

La généralisation à plus de deux tableaux ci-après proposée est dissymétrique, l'analyse de $K + 1$ tableaux faisant jouer au $K + 1^{\text{ème}}$ tableau le rôle d'un tableau de référence, les autres K tableaux jouant un rôle symétrique entre eux. Elle vient en prolongement de l'approche non symétrique proposée en section 2.3. Dans un premier temps on définit un premier $K + 1$ -uple. Dans un deuxième temps on définit les $K + 1$ -uples successifs. On vérifie autant que possible si critère et propriétés relatifs aux couples monogames de l'analyse de la concordance d'un tableau avec un autre s'étendent aux $K + 1$ -uples polygames qui sont maintenant définis.

3.1. Définition de la première solution

Soient $\mathbf{Y}$ un tableau centré $n \times q$ et K tableaux centrés $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_i, \dots, \mathbf{X}_K$ de dimensions respectives $n \times m_i$. On pose $m = \sum m_i$.

Le tableau $\mathbf{X}$ dénote désormais le tableau $n \times m$ formé des tableaux $\mathbf{X}_i$ concaténés :

$$\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \dots \ \mathbf{X}_K].$$

On veut définir un $K + 1$ uple $(\mathbf{Yb}, \mathbf{X}_1\mathbf{a}_1, \dots, \mathbf{X}_i\mathbf{a}_i, \dots, \mathbf{X}_K\mathbf{a}_K)$ synthétisant une communauté de structures des tableaux $\mathbf{X}_i$ avec la structure du tableau $\mathbf{Y}$.

Pour cela on définit d'abord $\mathbf{Yb}$ comme une composante synthétisant à partir de $\mathbf{Y}$ l'ensemble des covariations des variables des tableaux $\mathbf{X}_i$. Ensuite on définit les composantes $\mathbf{X}_i\mathbf{a}_i$ en termes de composantes partenaires de $\mathbf{Yb}$.

Pour repérer la communauté de structure que le tableau $\mathbf{Y}$ peut avoir avec l'ensemble des tableaux $\mathbf{X}_i, i = 1, \dots, K$, on recherche une composante $\mathbf{Yb}$ pouvant

caractériser tous les systèmes de covariations internes générés par les variables $\mathbf{x}_{ij}$ de chaque tableau $\mathbf{X}_i$, $j = 1, \dots, m_i$. C'est ainsi que l'on considère le critère à optimiser avec la contrainte de norme sur $\mathbf{b}$:

$$g(\mathbf{b}) = \sum_{i=1}^K \sum_{j=1}^{m_i} \text{cov}^2(\mathbf{Y}\mathbf{b}, \mathbf{x}_{ij}) \quad (6)$$

Soit encore :

$$g(\mathbf{b}) = \sum_{i=1}^K (\mathbf{b}' \mathbf{Y}' \mathbf{D} \mathbf{X}_i \mathbf{X}_i' \mathbf{D} \mathbf{Y} \mathbf{b}).$$

En écrivant :

$$g(\mathbf{b}) = \mathbf{b}' \mathbf{Y}' \mathbf{D} \left(\sum_{i=1}^K \mathbf{X}_i \mathbf{X}_i' \right) \mathbf{D} \mathbf{Y} \mathbf{b} = \mathbf{b}' \mathbf{Y}' \mathbf{D} \mathbf{X} \mathbf{X}' \mathbf{D} \mathbf{Y} \mathbf{b}, \quad (6 \text{ bis})$$

on voit que la solution est connue, consistant à faire la décomposition en valeurs singulières de $\mathbf{Y}' \mathbf{D} \mathbf{X}$. Le vecteur $\mathbf{b}_1$ solution est donc le vecteur singulier à gauche de $\mathbf{Y}' \mathbf{D} \mathbf{X}$ correspondant à la plus grande valeur singulière λ_1 , et $g(\mathbf{b}_1)$ est égal à λ_1^2 .

La notion de monogamie introduite précédemment pour deux tableaux est maintenant prolongée en une notion de polygamie. Il s'agit de définir dans chacun des tableaux $\mathbf{X}_i$ les composantes partenaires $\mathbf{X}_i \mathbf{a}_{i1}$ constituant les correspondants de la composante $\mathbf{Y} \mathbf{b}_1$ de $\mathbf{Y}$. En se référant à (4), on définit donc les vecteurs normés $\mathbf{a}_{i1}$ par les relations :

$$\mathbf{a}_{i1} \text{ cov}(\mathbf{X}_i \mathbf{a}_{i1}, \mathbf{Y} \mathbf{b}_1) = \mathbf{X}_i' \mathbf{D} \mathbf{Y} \mathbf{b}_1, \quad i = 1, \dots, K. \quad (7)$$

Propriété 3.1

Le critère de définition de la composante $\mathbf{Y} \mathbf{b}_1$ et de ses composantes partenaires $\mathbf{X}_i \mathbf{a}_{i1}$ répond à un double problème d'optimisation de covariation, $g(\mathbf{b}_1)$ étant l'optimum de chacun des deux critères :

$$\begin{aligned} h(\mathbf{a}, \mathbf{b}) &= \text{cov}^2(\mathbf{X} \mathbf{a}, \mathbf{Y} \mathbf{b}), \\ k(\mathbf{b}, \mathbf{e}_1, \dots, \mathbf{e}_i, \dots, \mathbf{e}_K) &= \sum_{i=1}^K \text{cov}^2(\mathbf{X}_i \mathbf{e}_i, \mathbf{Y} \mathbf{b}), \end{aligned}$$

sous les contraintes de norme sur $\mathbf{a}$, $\mathbf{b}$, et sur les vecteurs $\mathbf{e}_i$.

On détermine d'abord les deux écritures de l'optimum qui correspondent à ces deux problèmes d'optimisation.

On note $\mathbf{a}_1$ le vecteur normé de $\mathfrak{R}^m$, singulier à droite de $\mathbf{Y}' \mathbf{D} \mathbf{X}$, associé à la plus grande valeur singulière λ_1 :

$$\lambda_1 \mathbf{a}_1 = \mathbf{X}' \mathbf{D} \mathbf{Y} \mathbf{b}_1. \quad (8)$$

D'après (7) et (8) le sous vecteur de $\mathbf{a}_1$ associé à $\mathbf{X}_i$ est proportionnel à $\mathbf{a}_{i1}$; si μ_{i1} est la norme de ce sous vecteur, il s'écrit $\mu_{i1} \mathbf{a}_{i1}$ puisque $\mathbf{a}_{i1}$ est normé. Ainsi on a :

$$\lambda_1 \mu_{i1} \mathbf{a}_{i1} = \mathbf{X}_i' \mathbf{D} \mathbf{Y} \mathbf{b}_1, \quad i = 1, \dots, K. \quad (9)$$

Donc :

$$\lambda_1 \mu_{i1} = \text{cov}(\mathbf{X}_i \mathbf{a}_{i1}, \mathbf{Y} \mathbf{b}_1). \quad (10)$$

Le vecteur $\mathbf{a}_1$ est normé et donc $\sum_{i=1}^K \mu_{i1}^2 = 1$. Ainsi on a :

$$g(\mathbf{b}_1) = \text{cov}^2(\mathbf{X} \mathbf{a}_1, \mathbf{Y} \mathbf{b}_1) = \lambda_1^2 = \lambda_1^2 \sum_{i=1}^K \mu_{i1}^2.$$

On déduit alors de (10) que :

$$g(\mathbf{b}_1) = \text{cov}^2(\mathbf{X} \mathbf{a}_1, \mathbf{Y} \mathbf{b}_1) = \sum_{i=1}^K \text{cov}^2(\mathbf{X}_i \mathbf{a}_{i1}, \mathbf{Y} \mathbf{b}_1). \quad (11)$$

On sait que $g(\mathbf{b}_1) = \text{cov}^2(\mathbf{X} \mathbf{a}_1, \mathbf{Y} \mathbf{b}_1)$ est l'optimum, sous contraintes de norme sur $\mathbf{a}$ et $\mathbf{b}$, du critère :

$$h(\mathbf{a}, \mathbf{b}) = \text{cov}^2(\mathbf{X} \mathbf{a}, \mathbf{Y} \mathbf{b}).$$

Montrons que sous contraintes de norme sur $\mathbf{b}$ et sur les $\mathbf{e}_i$, $g(\mathbf{b}_1)$ est aussi le maximum du critère :

$$k(\mathbf{b}, \mathbf{e}_1, \dots, \mathbf{e}_i, \dots, \mathbf{e}_K) = \sum_{i=1}^K \text{cov}^2(\mathbf{X}_i \mathbf{e}_i, \mathbf{Y} \mathbf{b}).$$

On note $\text{diag}(\mathbf{e} \mathbf{e}')$ la matrice des blocs diagonaux $\mathbf{e}_i \mathbf{e}_i'$, les autres blocs étant nuls. On note $\mathbf{c}_i$ les vecteurs blocs dans $\mathfrak{R}^{m_i}$ du vecteur $\mathbf{c}$ normé dans $\mathfrak{R}^m$, $\lambda \mathbf{c}$ étant l'image de $\mathbf{b}$ par $\mathbf{X}' \mathbf{D} \mathbf{Y}$.

La valeur λ est inférieure ou égale à la plus grande valeur singulière λ_1 de la matrice $\mathbf{X}' \mathbf{D} \mathbf{Y}$ et ainsi toute solution vérifie :

$$\begin{aligned} \sum_{i=1}^K \text{cov}^2(\mathbf{X}_i \mathbf{e}_i, \mathbf{Y} \mathbf{b}) &= \mathbf{b}' \mathbf{Y}' \mathbf{D} \mathbf{X} \text{diag}(\mathbf{e} \mathbf{e}') \mathbf{X}' \mathbf{D} \mathbf{Y} \mathbf{b} \\ &= \lambda^2 \mathbf{c}' \text{diag}(\mathbf{e} \mathbf{e}') \mathbf{c} \leq \lambda_1^2 \mathbf{c}' \text{diag}(\mathbf{e} \mathbf{e}') \mathbf{c} = \lambda_1^2 \sum_{i=1}^K \langle \mathbf{e}_i, \mathbf{c}_i \rangle^2. \\ \lambda_1^2 \sum_{i=1}^K \langle \mathbf{e}_i, \mathbf{c}_i \rangle^2 &= \lambda_1^2 \sum_{i=1}^K \|\mathbf{c}_i\|^2 \cos^2(\mathbf{e}_i, \mathbf{c}_i) \leq \lambda_1^2 \sum_{i=1}^K \|\mathbf{c}_i\|^2 = \lambda_1^2. \end{aligned}$$

Comme la valeur $g(\mathbf{b}_1)$ atteint la valeur λ_1^2 pour $(\mathbf{b}, \mathbf{e}_1, \dots, \mathbf{e}_i, \dots, \mathbf{e}_K) = (\mathbf{b}_1, \mathbf{a}_{11}, \dots, \mathbf{a}_{i1}, \dots, \mathbf{a}_{K1})$, ce $K+1$ -uple ne peut être qu'une solution au problème d'optimisation posé.

Remarques

La composante $\mathbf{Xa}_1$ apparaît comme moyenne pondérée des composantes partenaïres, constituant un représentant compromis pour ces partenaïres. Les poids, proportionnels aux valeurs $\text{cov}(\mathbf{X}_i \mathbf{a}_{i1}, \mathbf{Yb}_1)$, indiquent les importances respectives prises par les tableaux $\mathbf{X}_i$ lors du calcul d'un $K + 1$ uple. D'après les relations (10), on a en effet :

$$\sum_{i=1}^K \frac{\text{cov}(\mathbf{X}_i \mathbf{a}_{i1}, \mathbf{Yb}_1)}{\lambda_1} \mathbf{X}_i \mathbf{a}_{i1} = \sum_{i=1}^K \mathbf{X}_i (\mu_{i1} \mathbf{a}_{i1}) = \mathbf{Xa}_1. \quad (12)$$

En posant $\mathbf{W}_Y = \mathbf{Y}\mathbf{Y}'$, on a la relation de type (5) :

$$\mathbf{W}_Y \mathbf{DXa}_1 = \lambda_1 \mathbf{Yb}_1. \quad (13)$$

3.2. Solutions successives

On note $(\mathbf{b}_j, \mathbf{a}_{1j}, \dots, \mathbf{a}_{ij}, \dots, \mathbf{a}_{Kj})$ le $K + 1$ uple constituant la $j^{\text{ème}}$ solution, formée de $k + 1$ vecteurs normés comme la première. On note $\mathbf{s}_j$ le vecteur de $\mathbb{R}^m$ formé par concaténation des $\mathbf{a}_{ij}$ et défini par :

$$\mathbf{s}'_j = [\mathbf{a}'_{1j} \dots \mathbf{a}'_{ij} \dots \mathbf{a}'_{Kj}],$$

On note $\text{diag}(\mathbf{s}_j \mathbf{s}'_j)$ la matrice $m \times m$ formée des K blocs diagonaux $\sum_{h=1}^j \mathbf{a}_{ih} \mathbf{a}'_{ih}$,

les autres blocs étant nuls.

La construction du premier $k + 1$ uple $(\mathbf{b}_1, \mathbf{a}_{11}, \dots, \mathbf{a}_{i1}, \dots, \mathbf{a}_{K1})$ implique, vue la façon de définir les composantes partenaïres, de rechercher la deuxième solution de sorte que les nouveaux vecteurs $\mathbf{a}_{i2}$ se trouvent respectivement orthogonaux à $\mathbf{a}_{i1}$ dans $\mathbb{R}^{m_i}$, pour chaque i , $i = 1, \dots, K$.

La deuxième solution $(\mathbf{b}_2, \mathbf{a}_{12}, \dots, \mathbf{a}_{i2}, \dots, \mathbf{a}_{K2})$ est donc calculée après avoir remplacé chaque tableau $\mathbf{X}_i$ par le tableau déflationné $(\mathbf{X}_i - \mathbf{X}_i \mathbf{a}_{i1} \mathbf{a}'_{i1})$, $i = 1, \dots, K$. Le vecteur $\mathbf{b}_2$ est ainsi le vecteur singulier à droite associé à la plus grande valeur singulière λ_2 de la matrice :

$$[\mathbf{X}'\mathbf{DY} - \text{diag}(\mathbf{s}_1 \mathbf{s}'_1) \mathbf{X}'\mathbf{DY}].$$

On note $\mathbf{a}_2$ le vecteur singulier normé dans $\mathbb{R}^m$ associé à $\mathbf{b}_2$, c'est-à-dire défini par :

$$[\mathbf{X}'\mathbf{DY} - \text{diag}(\mathbf{s}_1 \mathbf{s}'_1) \mathbf{X}'\mathbf{DY}] \mathbf{b}_2 = \lambda_2 \mathbf{a}_2. \quad (14)$$

On se place donc dans la situation où une deuxième solution existe, ce qui se produit quand le rang de la matrice à décomposer en valeurs singulières n'est pas nul.

Relativement à chaque tableau $\mathbf{X}_i$, $\mathbf{a}_{i2}$ et $\mathbf{a}_{i1}$ sont vecteurs normés des sous vecteurs respectifs associés de $\mathbf{a}_2$ et $\mathbf{a}_1$. Comme chaque sous vecteur $\mathbf{a}_{i2}$ est respectivement orthogonal à chaque sous vecteur $\mathbf{a}_{i1}$, $\mathbf{a}_2$ est orthogonal à $\mathbf{a}_1$. A part une différence de norme des sous vecteurs, la situation est la même entre $\mathbf{a}_2$ et $\mathbf{s}_1$ de sorte que $\mathbf{a}_2$ est aussi orthogonal à $\mathbf{s}_1$. Pour les mêmes raisons $\mathbf{a}'_2 \text{diag}(\mathbf{s}_1 \mathbf{s}'_1)$ est une matrice nulle. D'après (14), l'existence d'une deuxième solution revient donc à supposer que :

$$\lambda_2 = \text{cov}(\mathbf{X}\mathbf{a}_2, \mathbf{Y}\mathbf{b}_2) \neq 0. \quad (15)$$

Les vecteurs $\mathbf{a}_{i2}$ sont alors définis par les égalités :

$$\mathbf{a}_{i2} \text{cov}(\mathbf{X}_i \mathbf{a}_{i2}, \mathbf{Y} \mathbf{b}_2) = (\mathbf{X}_i - \mathbf{X}_i \mathbf{a}_{i1} \mathbf{a}'_{i1})' \mathbf{D} \mathbf{Y} \mathbf{b}_2, \quad i = 1, \dots, K. \quad (16)$$

Propriété 3.2.1

La deuxième solution est associable à l'optimisation des deux mêmes critères que la première, les contraintes d'orthogonalité des vecteurs $\mathbf{a}_{i2}$ avec les vecteurs respectifs $\mathbf{a}_{i1}$, $i = 1, \dots, K$, s'ajoutant aux contraintes de norme.

En effet l'égalité (11) qui s'écrit pour la deuxième solution avec les tableaux déflationnés, peut en fait aussi s'écrire avec les tableaux initiaux. A cause des contraintes d'orthogonalité, on a bien :

$$\sum_{i=1}^K \text{cov}^2([\mathbf{X}_i - \mathbf{X}_i \mathbf{a}_{i1} \mathbf{a}'_{i1}] \mathbf{a}_{i2}, \mathbf{Y} \mathbf{b}_2) = \sum_{i=1}^K \text{cov}^2(\mathbf{X}_i \mathbf{a}_{i2}, \mathbf{Y} \mathbf{b}_2),$$

de sorte qu'effectivement :

$$\lambda_2^2 = \text{cov}^2(\mathbf{X}\mathbf{a}_2, \mathbf{Y}\mathbf{b}_2) = \sum_{i=1}^K \text{cov}^2(\mathbf{X}_i \mathbf{a}_{i2}, \mathbf{Y} \mathbf{b}_2). \quad (17)$$

Propriété 3.2.2

La deuxième composante $\mathbf{Y}\mathbf{b}_2$ est définie à partir d'un vecteur $\mathbf{b}_2$ orthogonal à $\mathbf{b}_1$ et détermine ainsi une part d'inertie de $\mathbf{Y}$ distincte de la première part associée à $\mathbf{b}_1$.

On a en effet :

$$\begin{aligned} \mathbf{b}'_1 (\lambda_2 \mathbf{b}_2) &= \mathbf{b}'_1 [\mathbf{Y}' \mathbf{D} \mathbf{X} - \mathbf{Y}' \mathbf{D} \mathbf{X} \text{diag}(\mathbf{s}_1 \mathbf{s}'_1)] \mathbf{a}_2 \\ &= \mathbf{b}'_1 \mathbf{Y}' \mathbf{D} \mathbf{X} \mathbf{a}_2 = \text{cov}(\mathbf{X} \mathbf{a}_1, \mathbf{Y} \mathbf{b}_1) \mathbf{a}'_1 \mathbf{a}_2 = 0. \end{aligned}$$

Et ainsi $\mathbf{b}_2$ est orthogonal à $\mathbf{b}_1$.

Cette propriété est à la base du découpage de l'information par les $K + 1$ -uples successifs, puisque les mesures de covariances introduites dans l'analyse sont alors

constituées à partir de parts de variance distinctes quand on passe d'un $K + 1$ -uple à l'autre.

Dans Lafosse (1997) la définition de couples concordants en analyse de deux tableaux est fondée sur une notion d'indépendance entre couples monogames. La notion d'indépendance entre les deux premières solutions définissant les deux premiers $K + 1$ -uples polygames provient de l'ensemble des corrélations nulles ou faibles considérées dans la propriété suivante.

Propriété 3.2.3

Les corrélations entre $\mathbf{Xa}_2$ et $\mathbf{Yb}_1$, et entre $\mathbf{Xa}_1$ et $\mathbf{Yb}_2$ sont nulles.

La composante $\mathbf{Yb}_1$ est de corrélation nulle avec les partenaires $\mathbf{X}_i\mathbf{a}_{i2}$ de la deuxième composante, $i = 1, \dots, K$.

De plus en un certain sens, les covariations entre la deuxième composante $\mathbf{Yb}_2$ et les partenaires $\mathbf{X}_i\mathbf{a}_{i1}$ de la première sont minimisées.

L'ensemble de ces propriétés concourent ainsi à une deuxième solution peu corrélée à la première.

On a déjà remarqué que $\mathbf{a}'_2 \text{diag}(\mathbf{s}_1\mathbf{s}'_1)$ est nulle et que $\mathbf{b}_2$ était orthogonal à $\mathbf{b}_1$. On a donc les égalités suivantes traduisant la corrélation nulle de $\mathbf{Yb}_1$ et de $\mathbf{Xa}_2$, puis de $\mathbf{Yb}_2$ et de $\mathbf{Xa}_1$:

$$\mathbf{Y}'\mathbf{D}\mathbf{Xa}_j = \lambda_j \mathbf{b}_j, \quad j = 1, 2. \quad (18)$$

$$\mathbf{a}'_2[\mathbf{X}'\mathbf{D}\mathbf{Y} - \text{diag}(\mathbf{s}_1\mathbf{s}'_1)\mathbf{X}'\mathbf{D}\mathbf{Y}]\mathbf{b}_1 = \mathbf{a}'_2\mathbf{X}'\mathbf{D}\mathbf{Yb}_1 = \lambda_2\mathbf{b}'_2\mathbf{b}_1 = 0.$$

On a également :

$$\mathbf{a}'_1\mathbf{X}'\mathbf{D}\mathbf{Yb}_2 = \lambda_1\mathbf{b}'_1\mathbf{b}_2 = 0.$$

Et d'après (7), comme $\mathbf{a}_{i2}$ est orthogonal à $\mathbf{a}_{i1}$, on a aussi :

$$0 = \mathbf{a}'_{i2}\mathbf{X}'_i\mathbf{D}\mathbf{Yb}_1, \quad i = 1, \dots, K.$$

Par ailleurs, d'après le calcul de $\mathbf{b}_2$ et sachant l'idempotence des projecteurs $(\mathbf{I}_{mi} - \mathbf{a}_{i1}\mathbf{a}'_{i1})$, $\mathbf{b}_2$ est solution du problème d'optimisation du critère :

$$\begin{aligned} g(\mathbf{b}) &= \mathbf{b}'\mathbf{Y}'\mathbf{D}\mathbf{X}(\mathbf{I} - \text{diag}(\mathbf{s}_1\mathbf{s}'_1))\mathbf{X}'\mathbf{D}\mathbf{Yb} \\ &= \mathbf{b}'\mathbf{Y}'\mathbf{D}\left(\sum_{i=1}^K \mathbf{X}_i\mathbf{X}'_i\right)\mathbf{D}\mathbf{Yb} - \mathbf{b}'\mathbf{Y}'\mathbf{D}\left(\sum_{i=1}^K \mathbf{X}_i\mathbf{a}_{i1}\mathbf{a}'_{i1}\mathbf{X}'_i\right)\mathbf{D}\mathbf{Yb} \end{aligned}$$

soit encore :

$$g(\mathbf{b}) = \mathbf{b}'\mathbf{Y}'\mathbf{D}\left(\sum_{i=1}^K \mathbf{X}_i\mathbf{X}'_i\right)\mathbf{D}\mathbf{Yb} - \sum_{i=1}^K \text{cov}^2(\mathbf{Yb}, \mathbf{X}_i\mathbf{a}_{i1}).$$

Ainsi donc optimiser $g(\mathbf{b})$ lors du calcul du deuxième $K + 1$ uple, c'est maximiser le critère considéré au moment de définir la première solution $\mathbf{b}_1$ (et donc maximiser dans leur ensemble les covariances de $\mathbf{Yb}_2$ avec les composantes partenaires $\mathbf{X}_i\mathbf{a}_{i2}$) et simultanément minimiser dans leur ensemble les covariances de cette nouvelle solution $\mathbf{Yb}_2$ avec les composantes partenaires de $\mathbf{Yb}_1$. C'est donc veiller en particulier à ce que les corrélations de $\mathbf{Yb}_2$ avec les composantes partenaires de $\mathbf{Yb}_1$ ne soient pas trop élevées, alors qu'il est impossible que toutes ces corrélations soient nulles.

En effet, si les covariances $\text{cov}(\mathbf{Yb}_2, \mathbf{X}_i\mathbf{a}_{i1})$ étaient nulles pour tout i , on aurait d'après (16) :

$$\mathbf{a}_{i2}\text{cov}(\mathbf{X}_i\mathbf{a}_{i2}, \mathbf{Yb}_2) = \mathbf{X}_i'\mathbf{D}\mathbf{Yb}_2, \quad i = 1, \dots, K.$$

Alors, le couple $(\mathbf{a}_2, \mathbf{b}_2)$ vérifiant (18) devrait aussi vérifier la relation $\lambda_2\mathbf{a}_2 = \mathbf{X}'\mathbf{D}\mathbf{Yb}_2$. Le couple $(\mathbf{a}_2, \mathbf{b}_2)$ serait alors un couple de Tucker pour les matrices $\mathbf{X}$ et $\mathbf{Y}$, tout comme $(\mathbf{a}_1, \mathbf{b}_1)$. La façon de réaliser les déflations des tableaux $\mathbf{X}_i$ devrait donc correspondre à la déflation de $\mathbf{X}$ par le vecteur $\mathbf{a}_1$, et cela n'est en général possible que lorsque K est égal à 1.

D'après (18), on encore la relation de type (5) suivante :

$$\mathbf{W}_\mathbf{Y}\mathbf{D}\mathbf{Xa}_2 = \lambda_2\mathbf{Yb}_2.$$

De proche en proche sont ainsi définies S solutions successives, la $S + 1^{\text{ème}}$ n'existant pas lorsque les S déflations successives des tableaux $\mathbf{X}_i$ mènent finalement à une matrice de rang nul au moment du calcul de $\mathbf{b}_{S+1}$.

Pour la dernière solution, il existe au moins un i pour lequel un vecteur égal à $\text{cov}(\mathbf{X}_i\mathbf{a}_{iS}, \mathbf{Yb}_S)\mathbf{a}_{iS}$ n'est pas nul.

Quand une image de $\mathbf{b}_{j+1}$, $(\mathbf{X}_i - \mathbf{X}_i \sum_{h=1}^j \mathbf{a}_{ih}\mathbf{a}_{ih}')'\mathbf{D}\mathbf{Yb}_{j+1}$, est nulle, il n'y a pas

de nouvelle déflation de la matrice $(\mathbf{X}_i - \mathbf{X}_i \sum_{h=1}^j \mathbf{a}_{ih}\mathbf{a}_{ih}')$ avant de passer au calcul de la solution $j + 2$.

Finalement, S parts d'inertie distinctes de $\mathbf{Y}$ sont ainsi constituées.

Exceptés les cas précédents d'images nulles pouvant concerner certaines valeurs de l'indice i , la $j^{\text{ème}}$ part constituée est en relation avec une part d'inertie de chaque tableau $\mathbf{X}_i$ de façon privilégiée dans la mesure où une corrélation non nulle existe entre $\mathbf{Yb}_j$ et $\mathbf{X}_i\mathbf{a}_{ij}$, alors que les corrélations entre $\mathbf{Yb}_j$ et $\mathbf{X}_i\mathbf{a}_{i,j+1}$, entre $\mathbf{Yb}_j$ et $\mathbf{X}_i\mathbf{a}_{i,j+2}, \dots$, et entre $\mathbf{Yb}_j$ et $\mathbf{X}_i\mathbf{a}_{iS}$ sont nulles pour tout i et alors que les corrélations entre $\mathbf{Yb}_j$ et $\mathbf{X}_i\mathbf{a}_{i,j-1}$, entre $\mathbf{Yb}_j$ et $\mathbf{X}_i\mathbf{a}_{i,j-2}, \dots$, et entre $\mathbf{Yb}_j$ et $\mathbf{X}_i\mathbf{a}_{i1}$ sont faibles.

Le premier $K + 1$ uple défini correspond à un concept de polygamie restreinte, $\mathbf{Yb}_1$ ayant un partenaire dans chaque tableau $\mathbf{X}_i$ et n'ayant aucune liaison avec les partenaires des autres composantes $\mathbf{Yb}_j$, pour $j \geq 2$.

Les $K + 1$ uples suivant correspondent à des composantes $\mathbf{Yb}_j$ qui n'ont pas de liens de corrélations avec les partenaires des composantes $\mathbf{Yb}_{j+h}$ définies ensuite ($h > 0$), mais qui ont des liens de corrélations non nuls (mais minimisés) avec les partenaires des composantes $\mathbf{Yb}_{j-h}$ précédemment définies. Cette polygamie prend ainsi un sens plus large.

On pourrait dire que premier marin a des relations intenses mais avec les femmes les moins fidèles, le plus souvent à raison d'une femme dans chaque port.

Ceux qui suivent son exemple ont, de proche en proche, des relations moins intenses, avec des femmes plus fidèles.

Le dernier marin n'a que des femmes fidèles, avec des relations de faible intensité et sans toujours réussir à en trouver une dans chaque port.

Ce dernier marin traduit un autre type de polygamie restreinte, puisque tous les partenaires femmes considérées n'ont de relation qu'avec ce seul marin.

4. Lien avec l'analyse de co-inertie généralisée

4.1. Comparaison conceptuelle

L'analyse de co-inertie multiple (ACOM) de Chessel et Hanafi (1996) pose pour K tableaux $\mathbf{X}_i$ de variables centrées le problème de la recherche d'une variable $\mathbf{V}$, meilleur représentant compromis de composantes $\mathbf{X}_i \mathbf{e}_i$ associées à chacun des tableaux $\mathbf{X}_i$ soumis à l'analyse, en ce sens qu'elle maximise dans $\mathcal{R}^n$ le critère (ici donné dans son expression la plus simple) :

$$g(\mathbf{V}, \mathbf{X}_1 \mathbf{e}_1, \dots, \mathbf{X}_i \mathbf{e}_i, \dots, \mathbf{X}_K \mathbf{e}_K) = \sum_{i=1}^K \text{cov}^2(\mathbf{V}, \mathbf{X}_i \mathbf{e}_i), \quad (19)$$

sous les contraintes de norme pour $\mathbf{V}$ dans $\mathcal{R}^n$ et pour $\mathbf{e}_i$ dans $\mathcal{R}^{m_i}$, $i = 1, \dots, K$.

Comme le remarque les auteurs, en remplaçant les covariances par les corrélations on obtient le critère de Carroll (1968). Chez Carroll l'ensemble des solutions successives « $\mathbf{V}_j$ » forme un système de variables canoniques compromis non corrélées entre elles. En ACOM, en plus des corrélations nulles obtenues entre les variables solutions $\mathbf{V}_j$, on a aussi l'orthogonalité dans $\mathcal{R}^{m_i}$ de chacun des K systèmes $\{\mathbf{e}_{ij}\}$, $i = 1, \dots, K$. En effet, cette analyse associe aux solutions un découpage de l'inertie de chaque tableau $\mathbf{X}_i$, ce qui n'est pas une préoccupation en analyse canonique d'espaces vectoriels engendrés par les variables.

Posons maintenant

$$\mathbf{Y} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \dots \ \mathbf{X}_K].$$

Ce choix particulier de $\mathbf{Y}$ fait de l'analyse de la concordance de $\mathbf{Y}$ avec les tableaux $\mathbf{X}_i$ une analyse symétrique de K tableaux. Nous nommons cette analyse *analyse de concordance symétrique*.

Le critère (3) revient à vouloir définir une combinaison linéaire de l'ensemble des m variables synthétisant l'ensemble de toutes les covariations de ces variables.

Comparant le calcul de la première solution de l'ACOM à celui de la première solution en concordance symétrique, on constate que la composante V_1 est la composante Yb_1 dont la variance aurait été ramenée à 1. Cette composante Yb_1 est aussi la première de l'ACP généralisée de Casin (1996).

Cette première solution Yb_1 est la première composante principale de l'ACP de Y . De ce fait, sa corrélation est nulle avec toute combinaison linéaire b_j des variables de Y associée à une part d'inertie $b_j'Y'DYb_j$ distincte de $b_1'Y'DYb_1$ (propriété de l'ACP). Par suite, les composantes successives Yb_j de l'analyse de concordance symétrique, $j > 1$, sont de corrélation nulle avec Yb_1 .

La différence entre l'ACOM, l'ACPG de Casin, et l'analyse de la concordance symétrique de $Y = [X_1 \ X_2 \ \dots \ X_K]$ avec les tableaux $X_1, X_2, \dots$ et X_K , se produit à partir des deuxièmes solutions, à cause de contraintes différentes.

L'ACPG est conceptuellement très éloignée des deux autres, sachant que les contraintes introduites pour définir les solutions successives induisent la nullité des corrélations entre composantes relatives à chaque tableau X_i , tout comme le sont les composantes de l'ACP d'un tableau.

La comparaison avec l'ACOM est plus intéressante car les contraintes sont les mêmes en grande partie, mais la différence introduite correspond à des objectifs recherchés bien différents :

Les orthogonalités introduites en ACOM permettent des représentations simultanées par projection. Celles introduites en concordance symétrique visent l'indépendance des $K + 1$ uples successifs par une recherche de corrélations croisées nulles, et surtout un découpage de l'inertie de chacun des $K + 1$ tableaux.

Dans l'analyse ACOM la déflation est réalisée sur chaque tableau X_i .

Dans l'analyse de concordance symétrique, ces déflations ne sont pas réalisées quant au tableau concaténé $[X_1 \ X_2 \ \dots \ X_K]$ représentant Y , ce dernier tableau restant inchangé. Lors du calcul du $j^{\text{ème}}$ K -uple, sous contrainte de normes et d'orthogonalités, il s'agit de trouver la meilleure combinaison linéaire Yb_j de l'ensemble des variables initiales. En ACOM, tout se passe comme s'il s'agissait d'une analyse de concordance symétrique sous les mêmes contraintes, le problème étant de trouver la meilleure combinaison linéaire de l'ensemble des colonnes de tous les tableaux déflationnés.

De ce fait on n'a pas pour les variables Yb_j la propriété qu'ont les variables V_j d'être de corrélation nulle deux à deux. Par contre l'orthogonalité du système b_j permet un découpage de la variabilité totale de Y .

A part ces différences, les propriétés des termes respectifs des deux analyses sont analogues.

4.2. Comparaison numérique

Reprenant les données traitées par Chessel et Hanafi et publiées par Friday (1987), quelques mesures associées à la deuxième solution de l'ACOM et de l'analyse de concordance symétrique sont ici mises côte à côte. Nombres de ces mesures n'ont cependant d'intérêt que pour l'une des deux analyses. Malgré les analogies relevées précédemment entre les deux analyses, la difficulté à proposer des mesures indiquant

TABLE 4.2

	1	2	3	4	5	6
	Variance de Yb	Total des covariances au carré	Total des variances expliquées	Total des corrélations au carré avec les partenaires	Total des covariances croisés au carré	Total des corrélations croisées au carré
Concordance symétrie	1.8711	3.0913	1.6521	6.6736	.4124	.6336
ACOM	1.6580	2.7489	1.6580	6.6609	.3382	.5868

une supériorité de l'une par rapport à l'autre prouve bien qu'il s'agit de deux analyses différentes sur le fond.

Les 10 tableaux de 16 lignes analysés ont des nombres de colonnes qui diffèrent, mais ont chacun leur variance totale égale à 1.

On a programmé l'ACOM comme une analyse de concordance modifiée où les tableaux X_1 de Y sont aussi déflationnés. C'est pourquoi on obtient une composante notée Yb qui nous permet de donner un sens aux colonnes 2 et 5 de la table 4.2. C'est dire que la composante V_2 de l'ACOM est obtenue en divisant la composante Yb par la racine carrée de 1,658.

La composante Yb en concordance symétrique désigne Yb_2 .

La colonne 1 dit que la part de variabilité dégagée par la deuxième composante sur les variables initiales est plus importante que celle sur les colonnes des tableaux déflationnés.

La colonne 2 contient la somme des covariances au carré entre Yb et ses partenaires respectifs dans X_1 . C'est donc une mesure intéressante en concordance.

La colonne 3 contient les covariances précédentes après avoir ramené à un les variances de Yb : c'est donc pour l'ACOM la somme des covariances au carré de V_2 avec ses partenaires. C'est une mesure intéressante en ACOM.

La colonne 4 contient la somme des corrélations au carré des composantes Yb avec leurs partenaires. C'est une mesure intéressante, mais qui ne répond à aucun problème d'optimisation posé dans l'une ou l'autre analyse.

La colonne 5 correspond aux sommes des covariations au carré de Yb avec les partenaires trouvées pour la première solution (qui sont les mêmes pour les deux analyses). On a en concordance une valeur plus élevée, due en partie à la variance plus élevée de Yb .

La colonne 6 n'est pas sensible comme la colonne 5 à la variance de Yb , mais la somme des corrélations au carré de Yb avec les partenaires de la première solution est cependant plus élevée en concordance. C'est que la recherche d'une composante Yb ayant une variance élevée en concordance n'a pas permis de se rapprocher des corrélations nulles autant que dans l'ACOM. On peut donc se demander si en ACOM les covariations croisées ne sont pas minimisées d'une certaine manière, comme en concordance.

Les covariations croisées de Yb1 avec les partenaires de ces deuxièmes solutions n'ont pas été fournies, étant toutes nulles.

Remarque

L'introduction des métriques de Mahalanobis respectives dans les espaces $\mathbb{R}^{m_i}$ permet avec l'analyse ACOM d'obtenir alors pour chaque i , $i = 1, \dots, K$, des systèmes de composantes de corrélation nulle deux à deux relativement à chaque tableau.

Dans la concordance d'un tableau avec un autre ces changements de métriques donnent lieu aussi à des analyses très différentes. Dans la concordance d'un tableau avec plusieurs, cette introduction de métriques de Mahalanobis peut être considérée, soit relativement à $\mathbf{Y}$, soit relativement à chacun des tableaux $\mathbf{X}_i$, soit encore relativement à $\mathbf{Y}$ et à chacun des tableaux $\mathbf{X}_i$. Mais ces possibilités correspondent à des analyses très différentes et le vocabulaire utilisé dans cet article devrait en être partiellement revu.

Références

- CARROLL, J.D. (1968) *A generalization of canonical analysis to three or more sets of variables*. Pcr. 76 th. Convention, American Psychology Association, 227-228.
- CASIN Ph. (1996) *L'analyse en composantes principales généralisée*. Rev. Stat. Appliquée 44, 3, 63-81.
- CHESEL D. et HANAFI M. (1996) *Analyses de la co-inertie de K nuages de points*. Rev. Stat. Appliquée 44, 2, 35-60.
- CHESEL D. et MERCIER P. (1993) *Couplage de triplets statistiques et liaisons espèces-environnement*. In Biométrie et Environnement, Lebreton J.D. et Asselain B. (Eds). Paris, Masson, 15-44.
- ECKART C. et YOUNG G. (1936) *The approximation of one matrix by another of lower rank*. Psychometrika, 1, 211-218.
- FRIDAY L.E. (1987) *The diversity of macro invertebrate and macrophyte communities in ponds*. Freshwater biology, 18, 87-104.
- LAFOSSE R. (1997) *Analyse de concordance de deux tableaux : monogamies, simultanités et découpages*. Rev. Stat. Appliquée, 45,3, 45-72.
- MAHALANOBIS P.C. (1936) *On the generalized distance in statistics*. Proc. Nat. Inst. Sci. India, 12.
- TENENHAUS M. (1995) *A partial least squares approach to multiple regression, redundancy analysis and canonical analysis*. Cahier de recherche HEC, CR 550/1995.
- TUCKER L.R. (1958) *An interbattery method of factor analysis*. Psychometrika, 23, 111-136.
- WOLD H. (1975) *Soft modeling by latent variables : the non-linear iterative partial least squares approach*. In Perspectives in probability and statistics. Papers in honour of M.S. Bartlett, Gani J. (Ed), London : Academic Press.