Statistique et analyse des données

Patrice De La Salle

Analyse conjointe de plusieurs tableaux

Statistique et analyse des données, tome 4, n° 1 (1979), p. 13-30

<http://www.numdam.org/item?id=SAD_1979__4_1_13_0>

Statistiques et Analyse des Données
1 - 1979 pp. 13, 30.

ANALYSE CONJOINTE DE PLUSIEURS TABLEAUX

par

Patrice DE LA SALLE

Centre Océanologique de Bretagne - B.P. 337 - 29273 Brest Cedex

INTRODUCTION

Le but de cette étude est de proposer une méthode d'analyse conjointe de plusieurs tableaux en un sens que nous allons définir. Lorsque l'on considère un tableau de données, de la forme n observations par p variables, il est bien des cas où celles-ci peuvent se regrouper en plusieurs paquets, par association suivant le sujet auquel elles se rapportent ; par exemple, dans l'étude d'un biotope marin, on est conduit à mesurer des paramètres de nature hydrobiologique (température, salinité, oxygène dissous...), chimique (nitrates, phosphates, silicates...), biologique (caractéristiques de la faune, de la flore dans le milieu...), etc... ; autre exemple, lors d'un bilan scolaire de fin d'année, les variables étudiées, qui sont les notes des élèves obtenues au cours de toutes les interrogations, peuvent être regroupées par matière : français, math, anglais, physique.... Si le nombre de tableaux considérés est important, il n'est pas toujours facile d'obtenir une configuration des observations qui tienne compte de toutes les variables mesurées (des problèmes d'homogénéité, de dimension apparaissent, les représentations planes n'ont pas toujours une interprétation claire...). Par ailleurs, il est des cas où c'est moins l'analyse de toutes les variables de tous les tableaux qu'il nous importe de faire que l'étude globale des positions respectives de ces tableaux, les uns par rapport aux autres (par exemple, dans le cas où chaque tableau représente une variable qualitative et où les colonnes de la matrice correspondante sont les modalités de cette variable qualitative). Si donc, on veut diminuer le nombre de variables de façon à ne représenter un tableau que par un seul paramètre, il conviendra de déterminer ce paramètre de telle sorte que l'analyse de la matrice construite à partir des nouvelles variables obtenues, redonne une configuration des observations aussi proche que possible de celle que l'on aurait obtenue si l'on avait pris en compte toutes les variables de tous les tableaux. Définissant chaque paramètre comme une combinaison linéaire des variables du tableau qu'il représente, le principe de la méthode est donc de déterminer les coefficients appropriés de chacune d'elles de façon à ce que l'objectif que l'on s'est fixé soit satisfait.

Dans une première partie, nous voyons comment considérer la proximité entre deux ensembles d'observations placées dans un espace à p dimensions. La deuxième partie est consacrée à l'application de la méthode dans le cas d'un seul tableau ; nous montrons comment on retrouve ainsi les résultats de l'analyse en composantes principales (ACP) usuelle. Enfin, dans la troisième partie, nous considérons un ensemble de plusieurs tableaux et développons la méthode de résolution générale de notre problème.

1 - COMPARAISON DE DEUX TABLEAUX

1.1 - Equivalences et opérateurs

Considérons un tableau de données X, à p lignes et n colonnes, représentant les mesures de p variables sur n individus ; les variables sont supposées centrées. Soit M la métrique définie sur l'espace $\mathbb{R}^p$ des individus, et D_p la métrique définie sur l'espace $\mathbb{R}^n$ des variables. Appelons X' la matrice transposée de X (dimensions n x p). Les composantes principales associées au nuage des individus sont les vecteurs propres de l'opérateur X' M X D_p. Celui-ci ne dépend que des poids attribués aux individus (définissant la métrique D_p) et des distances entre les individus (comptées selon la métrique M) ; il est donc parfaitement représentatif du triplet (X, M, D_p).

Si nous considérons deux triplets (X_1, M_1, D_p) et (X_2, M_2, D_p), nous dirons que ces triplets sont équivalents si les opérateurs $X'_1 M_1 X_1 D_p$ et $X'_2 M_2 X_2 D_p$ sont proportionnels ; les nuages de points correspondants se déduisent alors l'un de l'autre par une similitude.

Les opérateurs considérés sont D_p-symétriques. Ils engendrent un sous-espace vectoriel de $L(\mathbb{R}^n, \mathbb{R}^n)$, ensemble des applications linéaires de $\mathbb{R}^n$ dans $\mathbb{R}^n$. Pour que l'équivalence précédente ait une valeur opérationnelle, il faut munir ce sous-espace d'une métrique euclidienne.

1.2 - Distance entre opérateurs D_p-symétriques

Soit un opérateur D_p-symétrique A ; il ne peut être considéré comme associé à un triplet (X, M, D_p) (i.e. de la forme X' M X D_p) que si la forme quadratique X' M X est semi-définie positive. Sur le sous-espace H des opérateurs D_p-symétriques de $L(\mathbb{R}^n, \mathbb{R}^n)$, considérons le produit scalaire défini à partir de la trace, proposé par Y. ESCOUFIER (cf. (5), (6)) :

$$\forall A, B \in H, \quad < A, B > = tr\ (AB)$$

Sur H, on obtient ainsi une forme bilinéaire symétrique, définie positive ; en effet, si $\{\lambda_i, i = 1...p\}$ est l'ensemble des valeurs propres de l'opérateur A, on a :

$$|| A ||^2 = tr\ (A^2) = \sum_{i=1}^{p} \lambda_i^2$$

Dans le cas de deux triplets (X, M, D_p) et (Y, N, D_p), le produit scalaire entre les opérateurs $X' M X D_p$ et $Y' N Y D_p$ s'écrit :

$$< X' M X D_p, Y' N Y D_p > = tr (X' M X D_p Y' N Y D_p)$$
$$= tr (M X D_p Y' N Y D_p X')$$
$$= tr (M V_{XY} N V_{YX})$$

avec $V_{XY} = X D_p Y' = V'_{YX}$

Ce produit scalaire est toujours positif. En effet :

soient x^j $(j = 1...p)$ les variables de X (dimensions p x n)
y^k $(k = 1...q)$ les variables de Y (dimensions q x n)
M (dimensions p x p) la métrique diagonale de terme général m_j $(j = 1...p)$
N (dimensions q x q) la métrique diagonale de terme général n_k $(k = 1...q)$

on a :

$$X' M X D_p = \sum_{j=1}^{p} m_j x^j x^{j'} D_p$$

$$Y' N Y D_p = \sum_{k=1}^{q} n_k y^k y^{k'} D_p$$

d'où le produit scalaire :

$$< X' M X D_p, Y' N Y D_p > = tr (X' M X D_p, Y' N Y D_p)$$
$$= tr \left(\sum_{j=1}^{p} m_j x^j x^{j'} D_p \sum_{k=1}^{q} n_k y^k y^{k'} D_p \right)$$
$$= \sum_{j=1}^{p} \sum_{k=1}^{q} m_j n_k (x^{j'} D_p y^k)^2 > 0 \qquad \text{CQFD}$$

La distance entre les deux opérateurs a pour expression :

$$||X' M X D_p - Y' N Y D_p||^2 = ||X' M X D_p||^2 + ||Y' N Y D_p||^2 - 2 <X' M X D_p, Y' N Y D_p>$$
$$= ||M V_X||^2 + ||N V_Y||^2 - 2 tr (M V_{XY} N V_{YX})$$
$$= tr (M V_X)^2 + tr (N V_Y)^2 - 2 tr (M V_{XY} N V_{YX})$$

avec $V_X = X D_p X'$, $V_Y = Y D_p Y'$

1.3 - Ajustement à un opérateur

Soit X un ensemble de p variables, mesurées sur n observations. Appelons P_X le projecteur orthogonal sur le sous-espace engendré par les p variables de X :

$$P_X = X' (X D_p X')^{-1} X D_p$$

Soit un opérateur $W D_p$ (dimensions n x n), définissant un système de distances entre les n observations.

Théorème 1 :

La variable y, combinaison linéaire des caractères de X, qui permet de reconstruire une configuration des n observations aussi proche que possible de celle obtenue à partir de l'opérateur $W D_p$, est le vecteur propre de l'application $P_X W D_p$, associé à sa plus grande valeur propre ; de plus, la norme de y est égale à la racine carrée de cette valeur propre.

Ce théorème a été énoncé par RAO (15) sous le titre "Principal components of instrumental variables" et fut repris par P. ROBERT et Y. ESCOUFIER (13) sous l'appellation "Principal components of Y with respect to X". Nous en donnons ici une démonstration à partir des distances entre opérateurs.

Démonstration :

La distance entre les deux configurations des n observations est donnée par la distance entre les deux opérateurs associés : $W D_p$ et $y y' D_p$. On a :

$$||W D_p - y y' D_p||^2 = ||W D_p||^2 + ||y y' D_p||^2 - 2 < W D_p, y y' D_p >$$

Il s'agit donc de trouver la variable y qui minimise :

$$||y y' D_p||^2 - 2 < W D_p , y y' D_p >$$

y est une combinaison linéaire des variables de X. Posons :

$$y = X' b \quad \text{avec } b \in \mathbb{R}^p$$

Commençons par maximiser le produit scalaire $< W D_p , y y' D_p >$ sous la contrainte :

$$||y||^2 = a \qquad \text{avec } a \in \mathbb{R}^+ \qquad (*)$$

Soit : $Q = < W D_p , y y' D_p > - \lambda (||y||^2 - a)$

Dérivant Q par rapport à b, il vient :

$$\frac{\partial Q}{\partial b} = 2 X D_p W D_p X' b - 2 \lambda X D_p X' b$$

d'où :

$$\frac{\partial Q}{\partial b} = 0 \Rightarrow X D_p W D_p X' b = \lambda X D_p X' b \qquad (1)$$

(*) Dans toute la suite, le symbole $||-||^2$ représentera toujours le carré de la norme calculée par rapport à la métrique D_p.

Supposons que $(X\ D_p\ X')^{-1}$ existe :

$$\Rightarrow (X\ D_p\ X')^{-1}\ X\ D_p\ W\ D_p\ X'\ b = \lambda\ b$$

$$\Rightarrow X'\ (X\ D_p\ X')^{-1}\ X\ D_p\ W\ D_p\ y = \lambda\ y$$

$$\Leftrightarrow P_X\ W\ D_p\ y = \lambda\ y \qquad (2)$$

Pour déterminer la valeur propre λ, il suffit de voir que :

$$\begin{aligned} < W\ D_p\ ,\ y\ y'\ D_p > &= tr\ (W\ D_p\ y\ y'\ D_p) \\ &= b'\ X\ D_p\ W\ D_p\ X'\ b \\ &= \lambda\ b'\ X\ D_p\ X'\ b \qquad \text{d'après (1)} \\ &= \lambda\ a \end{aligned}$$

Donc pour maximiser le produit scalaire $< W\ D_p\ ,\ y\ y'\ D_p >$, il faut prendre pour λ la plus grande valeur propre de $P_X\ W\ D_p$.

Déterminons la norme de y. On a :

$$\begin{aligned} ||y\ y'\ D_p||^2 - 2 < W\ D_p\ ,\ y\ y'\ D_p > &= ||y'\ D_p\ y||^2 - 2\ \lambda\ a \\ &= a^2 - 2\ \lambda\ a \\ &= a\ (a - 2\ \lambda) \end{aligned}$$

Cette expression est donc minimum pour :

$$a = \lambda \qquad \text{CQFD}$$

Remarques

_ Si l'on cherche une deuxième combinaison linéaire des variables de X, orthogonale à la première, telle que l'opérateur associé $\sum_{k=1}^{2} y^k\ y^{k'}\ D_p$ soit le plus proche possible de l'opérateur $W\ D_p$, on trouve aisément que cette combinaison linéaire est le vecteur propre de l'application $P_X\ W\ D_p$, associé à la deuxième valeur propre ; et sa norme au carré est égale à cette valeur propre. On peut itérer ainsi le processus jusqu'à extraire tous les vecteurs propres de $P_X\ W\ D_p$.

_ L'application $P_X\ W\ D_p$, restreinte au sous-espace engendré par les variables de X, est D_p-symétrique. En effet, soit :

$$y^1 = X'\ b_1 \quad \text{et} \quad y^2 = X'\ b_2$$

On a :

$$\begin{aligned} y^{1\prime}\ D_p\ P_X\ W\ D_p\ y^2 &= y^{1\prime}\ P_X'\ D_p\ W\ D_p\ y^2 \qquad \text{car } P_X \text{ est } D_p \text{ symétrique} \\ &= y^{1\prime}\ D_p\ W\ D_p\ P_X\ y^2 \qquad \text{car } P_X\ y^1 = y^1\ ,\ P_X\ y^2 = y^2 \end{aligned}$$

$$= y^{1\prime}\ D_p\ W\ P_X'\ D_p\ y^2 \qquad \text{car } P_X \text{ est } D_p \text{ symétrique}$$

$$= y^{1\prime}\ D_p\ W'\ P_X'\ D_p\ y^2 \qquad \text{car W est symétrique}$$

d'où :

$$D_p\ P_X\ W = W'\ P_X'\ D_p = (P_X\ W)'\ D_p$$

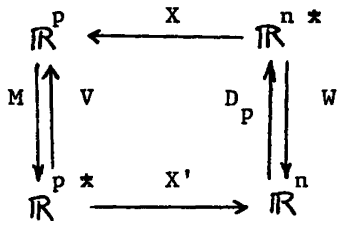
On en déduit que les valeurs propres de $P_X\ W\ D_p$ sont toutes réelles et que deux vecteurs propres associés à deux valeurs propres distinctes sont D_p-orthogonaux.

2 - ANALYSE D'UN SEUL TABLEAU

2.1 - Définitions et notations

Soit X le tableau de données à analyser, de dimensions p x n (nous considérons, par exemple, qu'il s'agit des mesures de p caractères centrés sur n individus). Soient X' la matrice transposée de X, M et D_p les métriques définies respectivement sur $\mathbb{R}^p$ et $\mathbb{R}^n$, V la forme quadratique d'inertie $X\ D_p\ X'$ et W la forme bilinéaire $X'\ M\ X$.

On a, classiquement, le schéma de dualité suivant :



2.2 - Résolution

La diagonalisation de la matrice $W\ D_p$ fournit les composantes principales qui définissent les positions des n individus sur les différents axes factoriels. Considérons une combinaison linéaire y des caractères de X. D'après le théorème 1, la variable y dont on peut déduire des représentations des individus aussi voisines que possible de celles obtenues en diagonalisant $W\ D_p$, est le vecteur propre de $P_X\ W\ D_p$ associée à sa plus grande valeur propre, λ_1.

On a :

$$P_X = X'\ (X\ D_p\ X')^{-1}\ X\ D_p$$

$$W\ D_p = X'\ M\ X\ D_p$$

$$\Rightarrow\ P_X\ W\ D_p = X'\ (X\ D_p\ X')^{-1}\ X\ D_p\ X'\ M\ X\ D_p$$

$$= X'\ M\ X\ D_p$$

$$= W\ D_p$$

Donc y, premier vecteur propre de $W D_p$ et de norme égale à λ_1, n'est autre que la première composante principale de l'ACP du triplet (X, M, D_p). En poursuivant ainsi la recherche des différentes combinaisons linéaires, orthogonales entre elles et recréant au mieux le nuage des n individus, on retrouve tous les résultats de l'ACP de (X, M, D_p).

3 - ANALYSE DE PLUSIEURS TABLEAUX

3.1 - Définitions et notations

Considérons une suite de T tableaux X_k $(k = 1.....T)$, de dimensions respectives $p_k \times n$ (chacun représentant, par exemple, les mesures de p_k caractères centrés sur les mêmes n individus). Soit M_k la métrique définie sur $\mathbb{R}^{p_k}$. Appelons X le tableau obtenu en superposant les T tableaux X_k, et M la métrique, diagonale par blocs, formée à partir des matrices M_k. Le choix de cette métrique M se justifie par le fait que l'on veut conserver les caractéristiques propres à chaque tableau, et en particulier la forme du nuage des n individus (définie par la métrique M_k) dans chaque sous-espace $\mathbb{R}^{p_k}$.

$$X = \begin{array}{c} \\ 1 \\ \vdots \\ p_1 \\ 1 \\ \vdots \\ p_2 \\ \vdots \\ 1 \\ \vdots \\ p_T \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & . & . & . & n \end{array} \\ \left[\begin{array}{c} X_1 \\ \hline X_2 \\ \hline \vdots \\ \hline X_T \end{array}\right] \end{array} \qquad M = \begin{array}{c} \\ 1 \\ \vdots \\ p_1 \\ 1 \\ \vdots \\ p_2 \\ \vdots \\ 1 \\ \vdots \\ p_T \end{array} \left[\begin{array}{ccc} M_1 & & \\ & M_2 & 0 \\ & & \\ 0 & & M_T \end{array}\right]$$

On a le schéma de dualité suivant :

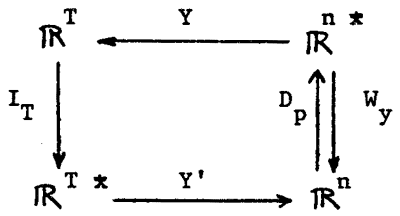
$$\begin{array}{ccc} \mathbb{R}^{\Sigma p_k} & \xleftarrow{\quad X \quad} & \mathbb{R}^{n\,*} \\ V \uparrow \downarrow M & & D_p \uparrow \downarrow W \\ \mathbb{R}^{\Sigma p_k\,*} & \xrightarrow{\quad X' \quad} & \mathbb{R}^{n} \end{array}$$

avec $V = X\ D_p\ X'$

$$W = X'\ M\ X = \sum_{k=1}^{T} X'_k\ M_k\ X_k = \sum_{k=1}^{T} W_k$$

3.2 - Résolution théorique

Soit $y^k = X'_k\ b^k$ une combinaison linéaire des variables du tableau X_k ($b^k \in \mathbb{R}^{p_k}$), et Y', la matrice de dimensions n x T, dont les colonnes sont les variables y^k ($\in \mathbb{R}^n$). Soit I_T la matrice identité d'ordre T sur $\mathbb{R}^T$. On a le schéma de dualité :



avec $W_Y = Y'\ I_T\ Y = \sum_{k=1}^{T} y^k\ y^k,$

L'analyse du tableau complet X conduit à diagonaliser la matrice $X'\ M\ X\ D_p$, dont les vecteurs propres sont les composantes principales de l'ACP de (X, M, D_p). Celles-ci fournissent les meilleures représentations des proximités entre les n individus.

Le principe de la méthode est de chercher simultanément T combinaisons linéaires y^1, $y^2 \ldots . y^T$ (chacune appartient à un sous-espace de $\mathbb{R}^n$ différent) de telle sorte que l'analyse du tableau Y donne une représentation des n individus aussi voisine que possible de celle obtenue à partir de l'ACP de (X, M, D_p). En d'autres termes, nous cherchons T variables telles que l'opérateur associé $Y'\ Y\ D_p$ soit le plus proche possible de l'opérateur $X'\ M\ X\ D_p$, au sens de la métrique définie en 1.2.

La distance entre ces deux opérateurs s'écrit :

$$||W\ D_p - Y'\ Y\ D_p||^2 = ||W\ D_p||^2 + ||Y'\ Y\ D_p||^2 - 2 < W\ D_p\ ,\ Y'\ Y\ D_p >$$

Nous procédons en deux étapes. Tout d'abord, nous cherchons les T variables y^k (k = 1 .... T) qui rendent maximum le produit scalaire $< W\ D_p\ ,\ Y'\ Y\ D_p >$. On a :

$$< W\ D_p\ ,\ Y'\ Y\ D_p > = \sum_{k=1}^{T} < W\ D_p\ ,\ y^k\ y^{k\prime}\ D_p >$$

Pour maximiser chacun des produits scalaires $< W\ D_p\ ,\ y^k\ y^{k\prime}\ D_p >$, il suffit de se reporter au théorème 1 ; la variable y^k cherchée est alors vecteur propre de l'application $P_k\ W\ D_p$ par rapport à sa plus grande valeur propre (avec P_k projecteur orthogonal sur le sous-espace engendré par les variables du tableau X_k , i.e. : $P_k = X'_k\ (X_k\ D_p\ X'_k)^{-1}\ X_k\ D_p$).

La deuxième étape du calcul consiste à déterminer les normes des différents vecteurs propres y^k, précédemment obtenus. Pour celà, considérons la distance entre les opérateurs $W D_p$ et $Y' Y D_p$, $||W D_p - Y' Y D_p||^2$. Les normes des y^k seront calculées de telle façon que cette distance soit la plus faible possible. Soit donc, à minimiser la quantité :

$$D = ||Y' Y D_p||^2 - 2 < W D_p , Y' Y D_p >$$

On a :

$$||Y' Y D_p||^2 = tr\ (Y' Y D_p\ Y' Y D_p)$$

$$= tr\ (\sum_{k=1}^{T} y^k\ y^{k'} D_p \sum_{k_1=1}^{T} y^{k_1}\ y^{k_1'} D_p)$$

$$= \sum_{k=1}^{T} \sum_{k_1=1}^{T} (y^{k'} D_p\ y^{k_1})^2$$

$$= \sum_{k=1}^{T} \sum_{k_1=1}^{T} a_k\ a_{k_1}\ \rho^2_{kk_1}$$

avec ρ_{kk_1} : coefficient de corrélation entre les variables y^k et y^{k_1}.

D'autre part :

$$< W D_p , Y' Y D_p > = \sum_k tr\ (W D_p\ y^k\ y^{k'} D_p)$$

$$= \sum_k y^{k'} D_p\ W D_p\ y^k$$

$$= \sum_k y^{k'} D_p\ P_k\ W D_p\ y^k \qquad \text{car } P_k\ y^k = y^k$$

$$= \sum_k \lambda_k\ y^{k'} D_p\ y^k \qquad \text{car } P_k\ W D_p\ y^k = \lambda_k\ y^k$$

$$= \sum_k \lambda_k\ a_k$$

avec λ_k : plus grande valeur propre de $P_k\ W D_p$.

D'où :

$$D = \sum_{k=1}^{T} \sum_{k_1=1}^{T} a_k\ a_{k_1}\ \rho^2_{kk_1} - 2 \sum_k \lambda_k\ a_k$$

Appelons a le vecteur des a_k ($k = 1 \ldots T$), λ le vecteur des λ_k ($k = 1 \ldots T$) et R la matrice carrée des $\rho^2_{kk_1}$ (dimensions T x T). Il vient :

$$D = a' R\ a - 2\ a' \lambda$$

Dérivons D par rapport à a :

$$\frac{\partial D}{\partial a} = 2 R\ a - 2 \lambda$$

d'où :

$$\frac{\partial D}{\partial a} = 0 \iff R\,a = \lambda \qquad (3)$$

On obtient bien ainsi un minimum pour D. En effet, quel que soit ε :

$$c = a + \varepsilon \Rightarrow c'\,R\,c - 2\,c'\,\lambda = a'\,R\,a - 2\,a'\,\lambda + \varepsilon'\,R\,a + a'\,R\,\varepsilon + \varepsilon'\,R\,\varepsilon - 2\,\varepsilon'\,\lambda$$

Comme R est symétrique, on a, d'après (3) :

$$\varepsilon'\,R\,a = a'\,R\,\varepsilon = \varepsilon'\,\lambda$$

d'où : $c'\,R\,c - 2\,c'\,\lambda = a'\,R\,a - 2\,a'\,\lambda + \varepsilon'\,R\,\varepsilon$

Comme R est définie positive :

$$\varepsilon'\,R\,\varepsilon > 0 \Rightarrow c'\,R\,c - 2\,c'\,\lambda > a'\,R\,a - 2\,a'\,\lambda$$

L'égalité (3) définit un système de T équations à T inconnues, qui s'écrit sous forme matricielle :

$$\begin{bmatrix} 1 & \rho^2_{12} & \rho^2_{13} & \cdots\cdots\cdots & \rho^2_{1T} \\ \rho^2_{21} & 1 & \rho^2_{23} & \cdots\cdots\cdots & \rho^2_{2T} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \rho^2_{T1} & \rho^2_{T2} & \cdots\cdots\cdots\cdots & & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_T \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdot \\ \cdot \\ \cdot \\ \lambda_T \end{bmatrix}$$

Appelons R la matrice des coefficients, a le vecteur des a_k et λ le vecteur des λ_k ; on obtient, si la matrice R est régulière, le résultat suivant :

$$a = R^{-1}\,\lambda \qquad (4)$$

On détermine ainsi de façon unique les combinaisons linéaires y^1, y^2 ..... y^T cherchées : elles sont, respectivement, vecteurs propres des applications $P_1\,W\,D_p$, $P_2\,W\,D_p$, ... , $P_T\,W\,D_p$ par rapport à leur plus grande valeur propre, et leurs normes sont définies par l'équation (4).

Remarques

— Les nombres a_k sont des normes, donc doivent être positifs strictement. Or, la solution du système (4) ne fournit pas nécessairement des a_k positifs. On peut néanmoins obtenir une solution approchée du système (4) ne comportant que des a_k positifs, en utilisant la méthode itérative de résolution d'un système linéaire de GAUSS-SEIDEL (cf. (12)). Ce procédé itératif converge toujours - dans ce cas - vers la solution exacte du système, car la matrice R est symétrique définie positive. Pour obtenir une solution ne comportant que des a_k positifs ou nuls,

il suffit de fixer les a_k qui seraient négatifs au cours du processus itératif à une valeur positive (donnée) ou nulle ; le système linéaire devient alors d'ordre inférieur à T, et sa solution ne comporte par construction que des éléments positifs ou nuls. La solution obtenue est unique, car celà revient à minimiser la fonction convexe D sur le domaine convexe $\{a_k \geq 0 , k = 1....T\}$.

— Nous n'avons extrait que le premier vecteur propre de chaque matrice $P_k W D_p$ pour définir la matrice Y. Il est toujours possible de trouver un autre système de variables $y^1, y^2 .... y^T$ en prenant le deuxième vecteur propre de chaque application $P_k W D_p$; chacun d'eux est orthogonal au premier (car $P_k W D_p$ est D_p- symétrique dans le sous-espace engendré par les variables du tableau X_k) et leurs normes sont définies par le système (4), où l'on remplace les ρ_{kk_1} par les coefficients de corrélation des seconds vecteurs propres entre eux, et les λ_k par les secondes plus grandes valeurs propres des matrices $P_k W D_p$. On peut ainsi trouver p systèmes de vecteurs $y^1, y^2....y^T$ (où $p = \min. (p_1, p_2....p_T)$), mais qui répondront de moins en moins bien au critère de distance entre les opérateurs que l'on a posé au départ.

3.3 - Résolution pratique

La recherche des variables $y^1, y^2....y^T$ conduit à diagonaliser les matrices $P_k W D_p$, qui sont de dimensions n x n. Lorsque le nombre d'individus est grand (par exemple, de l'ordre de quelques centaines), ce calcul peut devenir très vite couteux en temps et en mémoire machine. Aussi d'un point de vue pratique, utiliserons-nous les équations, fonctions des b^k et non des y^k. Par analogie avec l'équation (1) trouvée dans la démonstration du théorème 1, nous pouvons écrire :

$$X_k D_p W D_p X'_k b^k = \lambda_k X_k D_p X'_k b^k \qquad \forall k = 1....T$$

$$\Leftrightarrow (X_k D_p X'_k)^{-1} X_k D_p W D_p X'_k b^k = \lambda_k b^k$$

Soit V_{kk} la forme quadratique d'inertie associée au tableau X_k :

$$V_{kk} = X_k D_p X'_k$$

alors la matrice $(X_k D_p X'_k)^{-1} X_k D_p W D_p X'_k$ est V_{kk}- symétrique.

En effet :

$$V_{kk} (X_k D_p X'_k)^{-1} X_k D_p W D_p X'_k = X_k D_p W D_p X'_k$$

$$\text{et } X_k D_p W' D_p X'_k (X_k D_p X'_k)^{-1} V_{kk} = X_k D_p W' D_p X'_k$$

$$= X_k D_p W D_p X'_k \quad \text{car W est symétrique.}$$

Donc, $(X_k D_p X'_k)^{-1} X_k D_p W D_p X'_k$ a des valeurs propres réelles (qui sont d'ailleurs identiques à celles de $P_k W D_p$), et deux vecteurs propres correspondant à deux valeurs propres distinctes sont V_{kk}-orthogonaux. Le vecteur b^k cherché ($\in \mathbb{R}^{p_k}$) est alors le premier vecteur propre de la matrice $(X_k D_p X'_k)^{-1} X_k D_p W D_p X'_k$; on en déduit immédiatement la variable y^k ($= X'_k b^k$)

pour tout k. L'avantage de cette procédure est de diagonaliser des matrices de dimensions $p_k \times p_k$, ce qui représente une moindre dépense en temps et en mémoire machine.

Une fois obtenues les y^k (k=1....T), on construit la matrice Y, de dimensions T x n, dont les lignes sont les variables y^k. Une analyse en composantes principales du triplet (Y, I_T, D_p) permet, d'une part, de visualiser au mieux les proximités entre les n individus, d'autre part de donner une représentation des T variables, chacune correspondant à un tableau.

On peut mesurer la qualité de reconstruction du nuage des n individus ainsi constitué, par rapport à celui que l'on aurait obtenu si l'on avait étudié le tableau X tout entier, à l'aide du coefficient de ressemblance entre les opérateurs $W D_p$ et $Y' Y D_p$, défini par Y. ESCOUFIER (6), (13) :

$$RV = \frac{< W D_p , Y' Y D_p >}{\sqrt{||W D_p||^2 \; ||Y' Y D_p||^2}}$$

On a :

$$< W D_p , Y' Y D_p > = \sum_{k=1}^{T} \lambda_k a_k$$

et

$$||Y' Y D_p||^2 = \sum_{k=1}^{T} \sum_{k_1=1}^{T} a_k \; a_{k_1} \; \rho^2_{kk_1} = a' R a$$

Dans le cas où les a_k (k = 1....T) réalisent effectivement un minimum pour la fonction D, on a, d'après (3) :

$$||Y' Y D_p||^2 = a' R a = a' \lambda = \sum_{k=1}^{T} \lambda_k a_k$$

Alors, le coefficient RV s'écrit sous la forme simple suivante :

$$RV = \frac{\sqrt{\sum_{k=1}^{T} \lambda_k a_k}}{||W D_p||}$$

Le coefficient RV ne s'exprime donc qu'en fonction des a_k et des λ_k ; il varie entre 0 et 1, et selon qu'il est proche de l'une ou l'autre valeur, il permet d'apprécier la reconstitution du nuage des n individus à partir des T variables y^k.

3.4 - Cas particuliers

Cas d'un seul tableau

Si nous ne considérons qu'un seul tableau, nous retrouvons, bien sûr les résultats du paragraphe II. En effet, y^1 est vecteur propre de

$$P_1 W D_p = X'_1 (X_1 D_p X'_1)^{-1} X_1 D_p X'_1 M_1 X_1 D_p$$

$$= X'_1 M_1 X_1 D_p$$

par rapport à sa plus grande valeur propre.

D'autre part, sa norme est déterminée par l'équation (3)

$$\sum_{k_1 \neq k} a_{k_1} \rho^2_{kk_1} + a_k = \lambda_k$$

qui s'écrit ici :

$$a_1 = \lambda_1$$

y^1 est donc bien la première composante principale de l'ACP de (X_1, M_1, D_p).

Cas de deux tableaux

Soient X_1 et X_2 deux tableaux de données de dimensions respectives p_1 x n et p_2 x n. Soient

$$V_{11} = X_1 D_p X'_1 \quad \text{et} \quad V_{22} = X_2 D_p X'_2$$

les formes quadratiques d'inertie correspondantes (les caractères sont supposés centrés dans les deux tableaux). Plaçons-nous dans le cas où la métrique définie sur l'espace des individus est la métrique de MAHANALOBIS, soit :

$$M_1 = V_{11}^{-1} \quad , \quad M_2 = V_{22}^{-1}$$

Alors :

$$W D_p = \sum_{k=1}^{2} X'_k M_k X_k D_p$$

$$= X'_1 V_{11}^{-1} X_1 D_p + X'_2 V_{22}^{-1} X_2 D_p$$

$$= P_1 + P_2$$

P_1 et P_2 sont les projecteurs sur les sous-espaces engendrés respectivement par les variables de X_1 et de X_2.

Les combinaisons linéaires cherchées, y^1 et y^2, satisfont au système d'équations :

$$\begin{cases} P_1 (P_1 + P_2) y^1 = \lambda_1 y^1 \\ P_2 (P_1 + P_2) y^2 = \lambda_2 y^2 \end{cases}$$

$$\Leftrightarrow \begin{cases} P_1 P_2 y^1 = (\lambda_1 - 1) y^1 \\ P_2 P_1 y^2 = \lambda_2 - 1) y^2 \end{cases}$$

Or, $P_1 P_2$ et $P_2 P_1$ ont mêmes valeurs propres. En effet :
si u est vecteur propre de $P_1 P_2$ par rapport à la valeur propre μ, on a :

$$P_1 P_2 u = \mu u \Rightarrow P_2 P_1 (P_2 u) = \mu P_2 u$$

donc $P_2 u$ est vecteur propre de $P_2 P_1$ par rapport à μ ; et réciproquement.

Posons :

$$\mu = \lambda_1 - 1 = \lambda_2 - 1$$

on obtient alors le système d'équations :

$$\begin{cases} P_1 P_2 y^1 = \mu y^1 \\ P_2 P_1 y^2 = \mu y^2 \end{cases}$$

avec μ, plus grande valeur propre de $P_1 P_2$ (ou $P_2 P_1$). Les solutions y^1 et y^2 du système ne sont autres que les homothétiques des deux premiers caractères de l'analyse canonique de X_1 et X_2.

Leurs normes sont données par le système des équations (3) :

$$\begin{cases} a_1 + \rho^2_{12} a_2 = \lambda \\ \rho^2_{12} a_1 + a_2 = \lambda \end{cases} \qquad \text{avec } \lambda = \lambda_1 = \lambda_2$$

Résolvant ce système, il vient :

$$a_1 = a_2 = \frac{\lambda}{\rho^2_{12} + 1}$$

D'autre part, on sait que :

$$\begin{cases} P_1 P_2 y^1 = \mu y^1 \\ P_2 P_1 y^2 = \mu y^2 \end{cases} \Rightarrow \rho_{12} = \sqrt{\mu} = \sqrt{\lambda - 1}$$

On en déduit :

$$||y^1||^2 = ||y^2||^2 = 1$$

Donc, les combinaisons linéaires y^1 et y^2 cherchées ne sont autres que les deux premiers caractères de l'analyse canonique de X_1 et X_2.

Il apparait ainsi que les deux premiers caractères canoniques sont les combinaisons linéaires (de chaque paquet de variables) qui permettent de reconstruire au mieux le nuage des n observations, lorsque celles-ci sont placées dans un espace $\mathbb{R}^{p1}$ (resp. $\mathbb{R}^{p2}$) muni de la métrique V_{11}^{-1} (resp. V_{22}^{-1}).

3.5 - Cas des variables qualitatives

Considérons T variables qualitatives x^1, x^2..... x^T, et associons à chacune d'elles un tableau X_k (k = 1.....T), constitué par ses p_k modalités (dimensions p_k x n).

Codage et ACP de variables qualitatives

La démarche utilisée dans la méthode que nous venons de développer nous permet tout naturellement de réaliser une analyse en composantes principales (ACP) sur variables qualitatives. Les variables y^k (k = 1.....T), combinaisons linéaires des P_k modalités de x^k (k = 1.....T) que nous trouvons, constituent des codages de chaque variable qualitative. Ces codages ne sont pas obtenus indépendamment les uns des autres, mais sont calculés à partir de la matrice globale réunissant tous les tableaux disjonctifs X_k. Par conséquent, ils tiennent compte des interactions existant entre toutes les variables qualitatives, deux à deux. De plus, ces codages sont calculés de telle façon que l'ACP de la matrice Y qui les regroupe, reconstitue le plus fidèlement possible le nuage des n observations, placées au départ dans un espace à $\sum_{k=1}^{T} p_k$ dimensions. L'existence de ces deux propriétés contribue fortement à l'intérêt de ces codages de variables qualitatives ; par ailleurs, ils permettent, en plus, d'avoir, par l'ACP de Y, une représentation optimale de ces variables qualitatives, dans laquelle celles-ci ne sont plus exprimées que par un seul point.

Analyse canonique

Nous envisageons ici l'étude de la liaison entre une variable qualitative x^o (à p_o modalités) et un paquet de variables qualitatives x^1, x^2..... x^T (respectivement à p_1 , p_2.... p_T modalités). Soit X_o le tableau des modalités de x^o et X celui regroupant toutes les modalités des variables x^1, x^2..... x^T.

Deux méthodes sont classiquement utilisées pour résoudre cette question (cf. (4)) :

- l'analyse canonique usuelle qui décrit les positions relatives des sous-espaces E_o et E, engendrés respectivement par x^o et $\{x^1, x^2.... x^T\}$. Si P_o et P sont les projecteurs associés à E_o et E, les caractères canoniques

$$y^o = X'_o\, b^o \in E_o$$

$$\text{et } y = X'\, b = \sum_{k=1}^{T} X'_k\, b^k \in E$$

sont respectivement vecteurs propres de $P_o\, P$ et $P\, P_o$.

- l'analyse des correspondances du tableau $X_o\, D_p\, X'$. Cette matrice regroupe les T tableaux de probabilités, associés aux T couples de variables (x^o, x^k). C'est par analogie avec le cas de deux variables * que cette méthode a été mise en oeuvre, bien qu'il ne soit alors plus très facile de donner une signification concrète aux caractères trouvés. Notre démarche analytique va permettre de redonner quelque intérêt à cette méthode, en explicitant plus clairement les résultats obtenus.

Posons :

$$V_{11} = X\, D_p\, X' \quad \left(\text{dimensions } \sum_{k=1}^{T} p_k \times \sum_{k=1}^{T} p_k\right)$$

$$V_{21} = X_o\, D_p\, X' = V'_{12} \quad \left(\text{dimensions } p_o \times \sum_{k=1}^{T} p_k\right)$$

$$V_{22} = X_o\, D_p\, X'_o \quad (\text{dimensions } p_o \times p_o)$$

Soit V_1 la matrice diagonale ayant mêmes éléments diagonaux que V_{11} et

$$V_2 = T\, V_{22}$$

Tout couple de facteurs (b, b^o), où $b \in \mathbb{R}^{\Sigma p_k}$ et $b^o \in \mathbb{R}^{p_o}$, issu de l'analyse des correspondances de V_{21} vérifie les équations :

$$\begin{cases} V_{21}\, b = \sqrt{\mu}\, V_2\, b^o \\ V_{12}\, b^o = \sqrt{\mu}\, V_1\, b \end{cases} \qquad (5)$$

Posant :

$$y = X'\, b = \sum_{k=1}^{T} X'_k\, b^k = \sum_{k=1}^{T} y^k$$

$$y^o = X'_o\, b^o$$

les équations (5) s'écrivent :

$$\begin{cases} P_o\, y = T \sqrt{\mu}\, y^o \\ P_k\, y^o = \sqrt{\mu}\, y^k \end{cases}$$

---

* L'analyse des correspondances du tableau de contingence associé est théoriquement la meilleure analyse pour mettre en évidence les liaisons entre deux variables qualitatives.

d'où :

$$\begin{cases} \left(\sum_{k=1}^{T} P_k\right) P_o\, y = \lambda\, y \\ P_o \left(\sum_{k=1}^{T} P_k\right) y^o = \lambda\, y^o \end{cases} \qquad \text{avec } \lambda = T\,\mu$$

Le couple (y, y^o) maximise le produit scalaire $y' D_p y^o$ (ou plus exactement $\frac{1}{T} y' D_p y^o$ cf. (3)) sous les conditions de normalisation :

$$\frac{1}{T} \sum_{k=1}^{T} ||y^k||^2 = 1 \quad \text{et} \quad ||y^o||^2 = 1$$

La dernière équation nous permet de donner une interprétation nouvelle de y^o : sous les conditions de normes précédentes, y^o est la combinaison linéaire des modalités de la variable x^o qui permet de reconstruire au mieux le nuage des n observations, définies par les $\sum_{k=1}^{T} p_k$ modalités des variables $x^1, x^2 \ldots x^T$. En termes d'explication d'une variable qualitative x^o par un paquet de variables qualitatives $x^1, x^2 \ldots x^T$, le pouvoir d'explication de l'une par les autres est donc directement issu du critère de proximité des deux ensembles d'observations, déterminés, l'un à partir de la variable à expliquer, l'autre à partir des variables explicatives.

BIBLIOGRAPHIE

(1) BENZECRI J.P. L'analyse des données. Tome II. Dunod. 1973.

(2) CAILLEZ F. et J.P. PAGES. Introduction à l'analyse des données. SMASH. 1976.

(3) CAZES P. Etude de quelques propriétés extrêmales des facteurs issus d'un sous-tableau d'un tableau de Burt. Publication du Laboratoire de Statistique. 1975.

(4) CAZES P., A. BAUMERDER, S. BONNEFOUS, J.P. PAGES. Codage et analyse des tableaux logiques. Introduction à la pratique des variables qualitatives. Cahier du BURO n° 27. Paris VI. 1977.

(5) ESCOUFIER Y. Echantillonnage dans une population de variables aléatoires réelles. Publ. ISUP 19, fasc. 4. 1970.

(6) ESCOUFIER Y. Le traitement des variables vectorielles. Biometrics 29. 1973.

(7) HORST P. Relations among m sets of variables. Psychometrika 26. 1961.

(8) HOTELLING H. Relations between two sets of variates. Biometrika 28. 1935.

(9) KETTENRING R.J. Canonical analysis of several sets of variables. Biometrika 58. 1971.

(10) LEBART L., A. MORINEAU, N. TABARD. Techniques de la description statistique. Dunod. 1977.

(11) PAGES J.P., Y. ESCOUFIER, P. CAZES. Opérateurs et analyses des tableaux à plus de deux dimensions. C.R. séminaire BURO. 1974.

(12) PHAM D. Techniques du calcul matriciel. Dunod. 1962.

(13) ROBERT P. et Y. ESCOUFIER. A unifying tool for linear multivariate statistical methods : the RV-coefficient. Applied Statistics 25, n° 3. 1976.

(14) SAPORTA G. Liaisons entre plusieurs ensembles de variables et codages de données qualitatives. Thèse 3ème cycle. Paris VI. 1975.

(15) RAO C.R. The use and interpretation of principal component analysis in applied research. Sankhya. A 26, 4. 1964.