CAHIERS DU BURO

J. P. PAGES
Y. ESCOUFIER
P. CAZES

Opérateurs et analyse des tableaux à plus de deux dimensions

Cahiers du Bureau universitaire de recherche opérationnelle. Série Recherche, tome 25 (1976), p. 61-89

<http://www.numdam.org/item?id=BURO_1976__25__61_0>

OPÉRATEURS ET ANALYSE DES TABLEAUX A PLUS DE DEUX DIMENSIONS

J. P. PAGES , Y. ESCOUFIER , P. CAZES

Note des auteurs

Cet article peut être considéré en partie comme le compte-rendu d'un séminaire sur les opérateurs organisé par le B.U.R.O. au cours de l'année 1974 ; ce séminaire avait été animé non seulement par les auteurs de l'article mais aussi par J.M. BRAUN, J. DAUDIN et B. LETOURNEAU à qui l'on doit certaines des idées qui sont exposées ici et qui ont été reprises en collaboration avec F. TESTU [16].

SOMMAIRE

(*) Ingénieur au C.E.A. et professeur à l'I.S.U.P.

(**) Maître de conférences à l'université des sciences et techniques du Languedoc.

(***) Maître-assistant à l'université Pierre et Marie Curie (Paris VI) et professeur à l'I.S.U.P.

INTRODUCTION

On peut classer l'ensemble des techniques d'analyse des données qui relèvent de l'algèbre linéaire en deux groupes suivant que leur premier objectif est de décrire les proximités entre "individus" ou de décrire les liaisons entre "caractères" ; la nature (qualitatif, quantitatif) des caractères sur lesquels opèrent ces techniques, cette première classification étant admise, permet alors de les différencier.

Les notions d'individu et de caractère diffèrent de par le type des équivalences qui sont retenues pour les décrire ; si dans le cas des individus ce sont les positions relatives des uns par rapport aux autres qui nous intéressent, dans le cas des caractères les équivalences considérées reflètent le souci de reconstruire les uns à partir des autres.

Il est évident que la tactique à utiliser pour décrire les liaisons entre caractères ne résulte pas uniquement de la nature a priori des caractères considérés ; le choix d'une technique est avant tout fonction des objectifs assignés à l'étude et du nombre d'observations que l'on possède. Si on est amené à s'interroger sur la nature de la liaison entre deux paquets de caractères quantitatifs, par exemple, la stratégie à emprunter pour décrire cette liaison ne sera pas la même suivant que l'on dispose d'un petit nombre ou d'un grand nombre de données :

– si le nombre d'individus considérés est petit, il n'est pas possible d'explorer d'autres liaisons que les liaisons linéaires, aussi est-ce la régression ou l'analyse canonique que l'on utilisera ;

– si le nombre d'individus est suffisamment grand, l'analyse factorielle discriminante permettra de mettre en évidence des liaisons fonctionnelles ; pour la pratiquer, l'un des paquets de variables quantitatives sera rendu qualitatif, à l'aide d'une analyse en composantes principales par exemple, en regroupant les individus en classes homogènes relativement aux variables du paquet ;

– si on dispose d'un grand nombre d'individus, en utilisant l'analyse factorielle des correspondances, ayant rendu qualitatifs les deux paquets de caractères quantitatifs, on pourra mettre en évidence des liaisons (relations) d'un type encore plus général que le type des liaisons fonctionnelles.

Le choix d'une stratégie qui semble raisonnable pour décrire des proximités entre individus ou des liaisons entre caractères ne constitue en aucune façon une garantie quant à l'intérêt des résultats qui seront obtenus avec la technique sélectionnée pour analyser ces données ; la qualité d'une analyse dépend très largement de la capacité que l'on a d'interpréter les résultats.

On sait fort bien par exemple, quand on effectue une analyse en composantes principales, que la qualité des résultats n'est qu'en partie mesurée par la "part d'inertie expliquée" par le plan principal ou par les trois premiers axes principaux ; l'analyse sera décrétée bonne ou intéressante uniquement si ce que l'on voit sur le plan principal par exemple permet de faire des remarques pertinentes, la significativité des faits mis en évidence ne pouvant pas être mise en doute (stabilité). Comme cela est fort bien dit dans un article récent de R.N. SHEPARD [15], pour obtenir les résultats les plus satisfaisants possibles il faut effectuer son analyse de façon à réaliser un harmonieux

compromis entre "qualité de la représentation" d'une part et "facilité d'interprétation" et "stabilité" d'autre part.

Il est à noter que les documents statistiques obtenus à l'aide des techniques d'analyse des données n'ont pas pour unique rôle d'aider à la découverte ou à la compréhension des phénomènes, ils constituent aussi des éléments de discours, et des articles récents dans le Nouvel Observateur illustrent bien ceci [1], qui permettent avec peu de mots d'exprimer des idées ou de rendre plus flagrants certains faits (les inégalités sociales dans le cas du Nouvel Observateur).

Les objectifs que l'on poursuit et les données que l'on a à sa disposition conduisent parfois à s'interroger sur les ressemblances entre tableaux de données ; dans les paragraphes qui suivent nous proposons, en guise d'introduction à l'analyse des tableaux à plus de deux dimensions, une manière d'aborder la comparaison des tableaux de données qui s'inspire très largement des travaux effectués ces dernières années par Y. ESCOUFIER [6] [7] [8], puis par J.M. BRAUN [3].

Les propositions qui sont faites dans cet article, et dont on retrouvera certaines dans un travail récent de G. SAPORTA [13], ne peuvent être considérées que comme des préliminaires à une méthodologie de l'analyse des tableaux à plusieurs dimensions à l'aide de techniques qui relèvent de l'algèbre linéaire.

Il est évident que, comme précédemment, la stratégie à utiliser pour décrire les tableaux à plus de deux dimensions doit tenir compte des objectifs de l'étude, de la nature et du nombre d'observations que l'on possède ; les résultats obtenus n'auront d'intêret que dans la mesure où ils sont interprétables et stables.

Dans le paragraphe 1, on introduit des équivalences entre tableaux et les opérateurs qui permettent de rendre opérationnelles ces équivalences ; le lecteur aura l'occasion dans ce paragraphe de revoir rapidement l'ensemble des techniques d'analyse linéaire présentées comme en [5]. Dans le paragraphe 2, le sous-espace des opérateurs sera muni du produit scalaire proposé par Y. ESCOUFIER [6] ; on retrouvera alors quelques indices bien connus des statisticiens "classiques". La pratique des opérateurs à l'aide de l'analyse en composantes principales est décrite au paragraphe 3 où des remarques sont faites à propos des méthodes d'analyse (INDSCAL et IDIOSCAL) proposées par J.D. CARROLL [4].

On utilise les notations de [5] :

. I désigne un ensemble dont les n éléments sont munis des poids p_i : $p_i > 0$; $\sum_{i \in I} p_i = 1$;

. $P(I)$ est l'ensemble des parties de I ;

. X désigne un tableau "individus x caractères" centré à p lignes (caractères) et n colonnes (individus) :

. $\underline{x}_i$ désigne la i^e colonne de X ; $\underline{x}_i \in E = R^p$

. $\underline{x}^j$ désigne la j^e colonne de X' ; $\underline{x}^j \in F = R^n$;

. M désigne la métrique choisie dans E pour mesurer les proximités entre individus, souvent cette métrique n'est autre que "la métrique diagonale des inverses des carrés des écarts-types" notée D_{1/σ^2} ;

. D_p désigne la "métrique des poids" : $D_p(\underline{x}^j, \underline{x}^{j'})$ n'est autre que la covariance (notée cov) entre les caractères x^j et $x^{j'}$;

. D désigne un "tableau de distances" entre éléments de I ; les éléments de D peuvent n'obéir qu'aux axiomes des indices de dissimilarité.

1 – EQUIVALENCES ENTRE TABLEAUX DE DONNEES ET OPERATEURS EN ANALYSE LINEAIRE

11 – Equivalence entre tableaux de distances

Si D est un "tableau de distances" où sont décrites les proximités entre les n éléments, supposés munis des masses p_i d'un ensemble I, on sait qu'au couple (D, D_p) (cf. [5] chapitre 12 § 12.2) on peut associer une matrice W symétrique dont les éléments $w_{ii'}$ sont donnés par les formules :

$$w_{ii'} = \frac{1}{2}\,(d^2_{i.} + d^2_{i'.} - d^2_{..} - d^2_{ii'}) \ ,$$

avec :

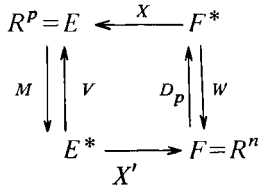
$$d^2_{i.} = \Sigma\{p_{i'}\, d^2_{ii'} / i' \in I\}$$

$$d^2_{..} = \Sigma\{p_i\, p_{i'}\, d^2_{ii'} / (i, i') \in I \times I\} \ .$$

Si W est semi-définie positive, il existe une infinité de triplets (X, M, D_p) qui fournissent de I une image euclidienne, c'est-à-dire tels que :

$$\| \underline{x}_i - \underline{x}_{i'} \|_M = d_{ii'} \quad \text{pour tout } (i, i') \in I \times I \ .$$

Au tableau centré X et aux métriques euclidiennes M et D_p correspond alors le schéma de dualité :



Si E et F sont munis respectivement des bases $\{\underline{e}_j / j = 1, \ldots, p\}$ et $\{\underline{f}_i / i = 1, \ldots, n\}$ et si E^* et F^* sont munis respectivement des bases duales $\{\underline{e}_j^* / j = 1, \ldots, p\}$ et $\{\underline{f}_i^* / i = 1, \ldots, n\}$:

$$X(\underline{f}_i^*) = \underline{x}_i \; ; \qquad X'(\underline{e}_j^*) = \underline{x}^j \; ;$$

$$W = X' \circ M \circ X \; ; \qquad V = X \circ D_p \circ X' \; .$$

La matrice associée à l'application V, qui a même rang que X' et que W, est la matrice de covariance entre les caractères $\underline{x}^j$.

Les composantes principales associées au nuage des individus $\underline{x}_i$ ne sont autres que les vecteurs propres de l'"opérateur"

$$U = W \circ D_p$$

correspondant aux valeurs propres non nulles :

$$W \circ D_p \, \underline{c}^i = \lambda_i \, \underline{c}^i \qquad i = 1, \ldots, p \; .$$
$$\| \underline{c}^i \| = \sqrt{\lambda_i}$$

Si D_1 et D_2 sont deux tableaux de distances entre éléments de I, en analyse linéaire il est naturel de considérer que ces tableaux sont équivalents soit s'ils sont identiques, soit, de façon plus générale, s'ils sont proportionnels.

Aux tableaux de distances D_1 et D_2 sont associés respectivement les opérateurs sur F :

$$U_1 = W_1 \circ D_p \quad ; \quad U_2 = W_2 \circ D_p \; .$$

Les deux équivalences peuvent s'exprimer ainsi :

Equivalence (a)

les couples (D_1, D_p) et (D_2, D_p) sont équivalents si les opérateurs associés U_1 et U_2 sont identiques.

Equivalence (b)

les couples (D_1, D_p) et (D_2, D_p) sont équivalents si les opérateurs associés sont homothétiques.

Rappelons qu'en classification automatique, suivant en cela SHEPARD [14] et KRUSKAL [11], on considère souvent que deux tableaux de distances sont équivalents s'ils induisent une même préordonnance sur I.

12 – Equivalence entre tableaux "individus x caractères" dans une optique de description

Quand on désire préciser en analyse linéaire les proximités entre les éléments de l'ensemble I symbolisés par les n colonnes d'un tableau de données numériques X à p lignes, on est amené à faire le choix d'une métrique M dans l'espace des individus E. Le choix de la métrique M est fonction des objectifs de l'analyse, de la connaissance que l'on a du phénomène étudié, de la confiance accordée aux différents caractères sélectionnés, de principes sages ou de traditions, etc.

Dans cette optique de description, la *donnée du tableau X est donc inséparable de la donnée de la métrique euclidienne M.*

Il y a évidemment en général une infinité de façons de sélectionner des caractères pour repérer les éléments de l'ensemble I ; à ces multiples façons d'envisager X correspondent les multiples façons de choisir la métrique M.

121 -- Equivalences et opérateurs

L'opérateur : $U = W \circ D_p$

qui est uniquement fonction :

– des distances $d_{ii'} = \| \underline{x}_i - \underline{x}_{i'} \|_M$

– et des masses p_i

et dont les vecteurs propres permettent de simplifier l'image euclidienne de I fournie par X et M (analyse en composantes principales) est représentatif du triplet (X, M, D_p).

Aux équivalences précédemment introduites entre couples (D, D_p) correspondent naturellement des équivalences entre triplets (X, M, D_p) :

Equivalence (a')

les triplets (X_1, M_1, D_p) et (X_2, M_2, D_p) sont équivalents si les opérateurs associés U_1 et U_2 sont identiques.

Equivalence (b')

les triplets (X_1, M_1, D_p) et (X_2, M_2, D_p) sont équivalents si les opérateurs associés U_1 et U_2 sont homothétiques.

Tous les triplets équivalents au sens de (a') admettent le même système $\{\underline{c}^i / i = 1, \ldots, p\}$ de composantes principales ; si C désigne le tableau dont le

transposé est $C' = (\underline{c}^1, \underline{c}^2, \ldots, \underline{c}^p)$ et si $\mathcal{J}$ désigne ici la métrique identité, tous ces triplets sont équivalents au triplet $(C, \mathcal{J}, D_p)$ (cf. [5] chapitre 9 § 12) et s'en déduisent par des transformations linéaires injectives T. En effet, la matrice de covariance entre les composantes principales $\underline{c}^i$ n'étant autre que la matrice diagonale D_λ dont les éléments diagonaux sont les valeurs propres non nulles λ_i de l'opérateur U :

$$\begin{array}{ccccc} E_1 & \xleftarrow{T} & E & \xleftarrow{C} & F^* \\ {\scriptstyle M}\downarrow\uparrow{\scriptstyle V} & & {\scriptstyle \mathcal{J}}\downarrow\uparrow{\scriptstyle D_\lambda} & & {\scriptstyle W}\downarrow\uparrow{\scriptstyle D_p} \\ E_1^* & \xrightarrow{T'} & E^* & \xrightarrow{C'} & F \end{array}$$

Si $T' \circ M \circ T = \mathcal{J}$, les deux triplets (TC, M, D_p) et $(C, \mathcal{J}, D_p)$ sont équivalents.

Si on se restreint à la métrique euclidienne $\mathcal{J}$, l'équivalence (a') conduit à considérer comme équivalents deux nuages de points se déduisant l'un de l'autre par des symétries et des rotations et l'équivalence (b') conduit à considérer comme équivalents deux nuages de points identiques à une similitude près.

122 – Images obtenues par les opérateurs

Si la métrique euclidienne choisie sur E est la métrique euclidienne classique $\mathcal{J}$, on montre immédiatement, tous les caractères étant centrés (ils sont considérés comme points de H, hyperplan de F D_p-orthogonal à la droite des constantes), que :

$$\underline{z} \in H \Rightarrow U(\underline{z}) = \sum_{j=1}^{p} D_p(\underline{x}^j, \underline{z})\, \underline{x}^j = \sum_{j=1}^{p} \operatorname{cov}(\underline{x}^j, \underline{z})\, \underline{x}^j \ .$$

Dans le cas d'une métrique quelconque M, si T est un isomorphisme vérifiant :

$$M = T' \circ \mathcal{J} \circ T$$

$$\begin{array}{ccccc} E & \xleftarrow{T} & E & \xleftarrow{X} & F^* \\ {\scriptstyle \mathcal{J}}\downarrow & & {\scriptstyle M}\downarrow\uparrow{\scriptstyle V} & & {\scriptstyle W}\downarrow\uparrow{\scriptstyle D_p} \\ E^* & \xrightarrow{T'} & E^* & \xrightarrow{X'} & F \end{array}$$

Le changement de variable

$$\underline{y}^j = X' \circ T'(\underline{e}_j^*) = \sum_{k=1}^{p} t_{jk}\, \underline{x}^k \quad ; \quad j = 1, \ldots, p$$

où t_{jk} désigne l'élément (j, k) de la matrice associée à T, permet d'écrire :

$$\underline{z} \in H \Rightarrow U(\underline{z}) = \sum_{j=1}^{p} D_p(\underline{y}^j, \underline{z})\, \underline{y}^j = \sum_{j=1}^{p} \operatorname{cov}(\underline{y}^j, \underline{z})\, \underline{y}^j .$$

13 – Equivalence entre tableaux "individus x caractères" dans une optique de prévision

Si on dépasse le cadre de la description pour s'intéresser plus particulièrement à l'information apportée par les caractères considérés (les lignes du tableau X) – il s'agit par exemple de déterminer dans quelle mesure ces caractères permettent de prévoir le futur (optique de prévision) – il est bien évident que l'on n'est pas enclin a priori à utiliser les équivalences (a') et (b').

Tous les caractères centrés qui sont combinaisons linéaires des caractères $\underline{x}^j$ sont dans le sous-espace vectoriel $X'(E^*)$ de F qui symbolise en analyse linéaire "le potentiel de prévision des caractères $\underline{x}^j$" (cf. [5] chap. 11 § 11.1).

Aussi est-il logique ici de considérer que la donnée du tableau X est équivalente à la donnée de ce sous-espace, donc de l'opérateur A de D_p-projection sur $X'(E^*)$ et d'introduire comme équivalence entre tableaux, l'équivalence :

Equivalence (c)

les couples (X_1, D_p) et (X_2, D_p) sont équivalents si les projecteurs associés A_1 et A_2 coïncident.

Si X' est injective, on sait que le projecteur D_p-symétrique A a pour expression :

$$A = X' \circ (X \circ D_p \circ X')^{-1} \circ X \circ D_p ,$$

soit :

$$A = X' \circ V^{-1} \circ X \circ D_p = W \circ D_p$$

$$\begin{array}{ccc} E & \xleftarrow{\;X\;} & F^* \\ {\scriptstyle V}\uparrow\downarrow{\scriptstyle V^{-1}} & & {\scriptstyle D_p}\uparrow\downarrow{\scriptstyle W} \\ E^* & \xrightarrow[X']{} & F \end{array}$$

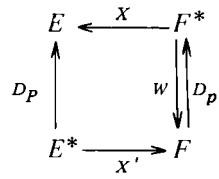
Cette dernière expression montre que dans une optique de prévision on est amené en analyse linéaire à faire correspondre au couple (X , D_p) l'opérateur qui est associé, dans une optique de description, au triplet (X , M , D_p), la métrique M choisie dans E étant la "métrique de Mahalanobis" V^{-1} :

les couples (X_1 , D_p) et (X_2 , D_p) sont équivalents au sens de (c) si les triplets (X_1 , V_1^{-1} , D_p) et (X_2 , V_2^{-1} , D_p) le sont au sens de (a').

14 – Equivalence entre variables qualitatives

Toute variable quantitative ξ induisant sur l'ensemble I une partition identique ou plus grossière que la partition induite par la variable qualitative x, dont J est l'ensemble des modalités, est représentée dans le vectoriel F par un point du sous-espace engendré par les variables indicatrices des p modalités de la variable x (cf. [5] chapitre 7).

Si les variables indicatrices constituent les colonnes du tableau X', on est amené à considérer le schéma de dualité :



où : $D_P = X \circ D_p \circ X'$ est l'application dont la matrice associée est la matrice diagonale des poids P_j avec :

$$P_j = \Sigma\{ p_i / i \in I \ ; \ x_i^j = 1 \} .$$

Au caractère quantitatif :

$$\underline{\xi} = X'(\underline{a}) \qquad \underline{a} \in E^*$$

correspond la forme linéaire $\underline{a}$ dont les coordonnées dans la base $\{\underline{e}_j^* / j = 1, \ldots , p\}$ sont les *codages* affectés aux modalités de la variable x.

Rappelons que (cf. [5] chapitre 13 § 13.4), $(J , P(J))$ étant muni de la loi de probabilité P définie par les probabilités élémentaires P_j :

. E^* est assimilé à l'espace des variables aléatoires réelles définies sur $(J , P(J))$;

. E est assimilé à l'ensemble des mesures sur $(J , P(J))$.

La restriction de la métrique $D_{1/P}$ à l'hyperplan affine des mesures de masse totale égale à 1, permet de munir le simplexe S des lois de probabilité sur $(J , P(J))$ d'une distance appelée distance du chi-deux de centre P :

$$\underline{q} \in S \quad \underline{r} \in S \Rightarrow d^2_{\chi^2_P}(\underline{q}, \underline{r}) = D_{1/P}(\underline{q} - \underline{r}) .$$

Le tableau X considéré ici n'est pas centré ; la droite des constantes Δ_j, engendrée par le vecteur $\underline{j}$ de F dont toutes les coordonnées sont égales à 1, est dans $X'(E^*)$; le sous-espace H supplémentaire D_p-orthogonal à Δ_j dans $X'(E^*)$ est "le sous-espace des caractères quantitatifs centrés que l'on sait reconstruire à partir de la variable qualitative x" :

$$X'(E^*) = \Delta_j \oplus H \quad ; \quad \Delta_j \underset{D_p}{\perp} H .$$

Assimilant la variable qualitative x au sous-espace H des caractères quantitatifs centrés qu'elle permet de reconstruire (potentiel de prévision), on identifiera x à l'opérateur B de D_p-projection sur H ; d'où l'équivalence entre caractères qualitatifs :

Equivalence (c')

les couples (x_1, D_p) et (x_2, D_p) sont équivalents si les D_p-projecteurs associés B_1 et B_2 coïncident.

Remarques :

. si A et A_j désignent respectivement les D_p-projecteurs sur $X'(E^*)$ et Δ_j :

$$B = A - A_j ;$$

matriciellement :

$$B = [X'(XD_pX')^{-1}X - \underline{j}\,\underline{j}']D_p = WD_p .$$

. $B_1 = B_2 \Leftrightarrow A_1 = A_2$.

15 – Equivalence entre une variable qualitative et un paquet de variables quantitatives

Les équivalences (c) et (c') permettent de définir immédiatement une équivalence entre une variable qualitative y et un paquet de variables quantitatives (tableau centré X) : c'est cette équivalence que l'on utilise en analyse factorielle discriminante (cf. [5] chapitre 12).

Equivalence (c'')

le couple (y, D_p), auquel est associé le D_p-projecteur B, est équivalent au couple (X, D_p), auquel est associé le D_p-projecteur A, si ces deux projecteurs coïncident.

16 – Bilan

Toutes les équivalences précédemment introduites font intervenir des opérateurs D_p-symétriques définis sur l'espace des caractères F ; ces opérateurs engendrent un sous-espace vectoriel G du vectoriel $L(F, F)$ des applications linéaires de F dans F.

Pour rendre opérationnelles ces équivalences il faut se donner un moyen pour mesurer les proximités entre ces opérateurs ; on obtiendra ce moyen en munissant le sous-espace G d'une métrique euclidienne.

2 – PRODUIT SCALAIRE ET DISTANCE ENTRE OPERATEURS D_p-SYMETRIQUES

Le sous-espace G de $L(F, F)$ des opérateurs D_p-symétriques contient tous les opérateurs précédemment rencontrés ; un élément de G ne peut être considéré comme un opérateur du type $W \circ D_p$ associé à un triplet (X, M, D_p) que si la forme quadratique W est semi-définie positive.

La forme bilinéaire symétrique P définie sur $L(F, F)$ à partir de la trace (tr) :

$$\left.\begin{array}{l} A \in L(F, F) \\ B \in L(F, F) \end{array}\right\} \Rightarrow \underline{P(A, B) = tr(AB)}$$

permet de définir une distance euclidienne sur G ; en effet, restreinte à G, la forme bilinéaire P est définie positive, car si les p valeurs propres de U sont notées λ_i (elles sont rangées par valeurs décroissantes) :

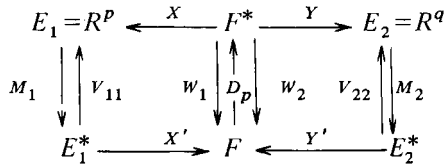
$$\underline{U \in G \Rightarrow P(U, U) = tr(U^2) = \sum_{i=1}^{p} \lambda_i^2 \ .}$$

Considérons :

– les p caractères $\underline{x}^i$ (tableau X) et la métrique M_1 ;

– les q caractères $\underline{y}^j$ (tableau Y) et la métrique M_2 .

Aux deux triplets (X, M_1, D_p) et (Y, M_2, D_p) correspond le double schéma de dualité :



On adopte les notations :

$$V_{11} = X \circ D_p \circ X' \; ; \quad V_{22} = Y \circ D_p \circ Y' \; ; \quad V_{12} = X \circ D_p \circ Y' = V'_{21} \; ;$$

$$W_1 = X' \circ M_1 \circ X \; ; \qquad W_2 = Y' \circ M_2 \circ Y \; ;$$

$$U_1 = W_1 \circ D_p \; ; \qquad U_2 = W_2 \circ D_p \; .$$

Le produit scalaire entre les opérateurs U_1 et U_2 a pour expression :

$$P(U_1 , U_2) = tr(U_1 U_2) = tr(V_{21} M_1 V_{12} M_2) \; ;$$

en particulier :

$$P(U_1 , U_1) = \| U_1 \|^2 = tr(U_1^2) = tr((V_{11} M_1)^2) \; ,$$

$$P(U_2 , U_2) = \| U_2 \|^2 = tr(U_2^2) = tr((V_{22} M_2)^2) \; .$$

Le cosinus de l'angle entre les opérateurs U_1 et U_2 sera noté $R_{U_1 U_2}$:

$$R_{U_1 U_2} = \frac{P(U_1,U_2)}{\| U_1 \| \, \| U_2 \|} = \frac{tr(V_{21} M_1 V_{12} M_2)}{\sqrt{tr((V_{11} M_1)^2) \, tr((V_{22} M_2)^2)}} .$$

21 – Optique de description : proximité entre triplets

Rappelons qu'ici, l'objectif étant de décrire l'ensemble I, la donnée d'un ensemble de caractères (tableau X) va de pair avec la donnée d'une métrique euclidienne M définie sur le vectoriel des individus.

Dans la pratique les deux métriques euclidiennes les plus utilisées sont :

– la métrique euclidienne classique $M = \mathcal{I}$;

– la "métrique diagonale des inverses des carrés des écarts-types" $M = D_{1/\sigma^2}$.

La deuxième a en particulier l'avantage de conduire en analyse en composantes principales à des résultats ne dépendant pas des échelles de mesure.

Une métrique intéressante a été introduite par Joreskog [9] qui cherchait à résoudre sans itérations les équations rencontrées en "analyse factorielle en

facteurs communs et spécifiques". V désignant la matrice de covariance, supposée régulière, entre les caractères sélectionnés, la "métrique de Joreskog" admet pour matrice la matrice diagonale $M = \text{diag}(V^{-1})$ dont les éléments diagonaux sont les éléments diagonaux de V^{-1}.

La métrique de Joreskog permet aussi d'obtenir en analyse en composantes principales des résultats indépendants du choix des unités de mesure.

Remarque :

Le i^e terme diagonal de la matrice $M = \text{diag}(V^{-1})$ a pour expression :

$$M_{ii} = \frac{1}{\sigma_i^2(1 - r_i^2)},$$

où : σ_i est l'écart-type de la variable x^i,

r_i est le coefficient de corrélation multiple entre la variable x^i et l'ensemble des variables autres que x^i.

Si les caractères considérés sont regroupés en p paquets, le tableau des données X étant alors la superposition de p tableaux $X_1, X_2, \ldots, X_p$, il est possible de choisir pour mesurer la proximité entre individus, si on accorde à chacun des paquets une importance proportionnelle au nombre de variables du paquet, la métrique M de matrice :

$$M = \begin{bmatrix} V_{11}^{-1} & 0 & \ldots & 0 \\ 0 & V_{22}^{-1} & & \\ \vdots & & & \vdots \\ 0 & \ldots 0 & \ldots & V_{pp}^{-1} \end{bmatrix}$$

où V_{jj} désigne la matrice de variance associée au j^e paquet.

Au triplet (X, M, D_p) correspond alors l'opérateur U qui s'écrit : $U = \sum_{j=1}^{p} A_j$,

où A_j est l'opérateur de D_p-projection sur le sous-espace engendré par les variables du j^e paquet.

Dans le cas particulier où chacun des paquets ne comprend qu'une seule variable on retrouve pour métrique M la métrique D_{1/σ^2}.

S'il n'existe que deux paquets de variables, on obtient en projetant les composantes principales associées au triplet (X, M, D_p) sur les plans engendrés respectivement par les variables du premier et du deuxième paquet, des vecteurs qui sont homothétiques aux caractères canoniques que l'on obtiendrait en effectuant une analyse canonique sur ces deux paquets [12] : choisir la métrique M revient donc en quelque sorte à décider d'effectuer une analyse "canonique généralisée" des p paquets de variables considérées.

Suivant que l'on choisit pour équivalence entre triplets l'équivalence (a') ou l'équivalence (b') on ne normera pas ou on normera les opérateurs :

– *choix de l'équivalence (a')* : ici la distance $d_{U_1 U_2}$ entre les opérateurs U_1 et U_2 est donnée par :

$$d^2_{U_1 U_2} = ||U_1||^2 + ||U_2||^2 - 2P(U_1, U_2),$$

soit :

$$d^2_{U_1 U_2} = tr((V_{11}M_1)^2 + tr((V_{22}M_2)^2) - 2tr(V_{21}M_1 V_{12}M_2) ;$$

– *choix de l'équivalence (b')* : ici la distance entre les opérateurs sera uniquement fonction du cosinus de l'angle qu'ils forment :

$$d^2_{U_1 U_2} = 2(1 - R_{U_1 U_2}).$$

211 – Choix de la métrique euclidienne classique

Si on munit, pour mesurer les proximités entre individus, les vectoriels E_1 et E_2 de la métrique euclidienne classique :

$$P(U_1, U_2) = tr(U_1 U_2) = tr(V_{21} V_{12}),$$

d'où :

$$P(U_1, U_2) = \Sigma\{D_p^2(x^i, y^j)/i = 1, \ldots, p \quad j = 1, \ldots, q\}$$
$$= \Sigma\{\text{cov}^2(x^i, y^j)/i = 1, \ldots, p ; j = 1, \ldots, q\}.$$

On en déduit :

$$||U_1||^2 = \Sigma\{\text{cov}^2(x^i, x^{i'})/i = 1, \ldots, p \ ; \ i' = 1, \ldots, p\}$$
$$||U_2||^2 = \Sigma\{\text{cov}^2(y^j, y^{j'})/j = 1, \ldots, q \ ; \ j' = 1, \ldots, q\}$$

Le cosinus $R_{U_1 U_2}$ de l'angle entre les deux opérateurs a pour expression :

$$R_{U_1U_2} = \frac{tr(V_{21}V_{12})}{\sqrt{tr(V_{11}^2)\ tr(V_{22}^2)}} .$$

$R_{U_1U_2}$ n'est autre que le carré du coefficient de corrélation entre x et y quand on se trouve dans le cas particulier où les deux tableaux X et Y ne comportent qu'une seule ligne (caractère x pour X et caractère y pour Y).

212 – Choix de la métrique D_{1/σ^2}

Si la métrique diagonale des inverses des carrés des écarts-types est choisie aussi bien dans E_1 que dans E_2 :

$$P(U_1 , U_2) = \Sigma\{\text{cor}^2 (x^i , y^j)/i = 1 , \ldots , p \ ; \ j = 1 , \ldots , q\} ;$$

d'où :

$$||U_1||^2 = \Sigma\{\text{cor}^2 (x^i , x^{i'})/i = 1 , \ldots , p \ ; \ i' = 1 , \ldots , p\},$$

$$||U_2||^2 = \Sigma\{\text{cor}^2 (y^j , y^{i'})/j = 1 , \ldots , q \ ; \ j' = 1 , \ldots , q\} ;$$

ici :

$$R_{U_1U_2} = \frac{tr(R_{21}R_{12})}{\sqrt{tr(R_{11}^2)\ tr(R_{22}^2)}},$$

où R_{11}, R_{22}, R_{12} et R_{21} sont les matrices de corrélation correspondant respectivement aux matrices de variance V_{11}, V_{22}, V_{12} et V_{21} .

22 – Optique de prévision

221 – Proximités entre paquets de caractères quantitatifs

On a vu au paragraphe 13 que l'opérateur de D_p-projection A associé au couple (X , D_p) n'est autre que l'opérateur U associé au triplet (X , V^{-1} , D_p).

Le produit scalaire entre les opérateurs U_1 et U_2 associés respectivement aux couples (X , D_p) et (Y , D_p) a donc pour expression :

$$P(U_1 U_2) = tr(V_{21} V_{11}^{-1} V_{12} V_{22}^{-1}) .$$

Si r_i désigne le $i^{\text{ème}}$ coefficient de corrélation canonique entre les deux groupes de caractères considérés et si $k = \inf\{p , q\}$:

$$P(U_1 U_2) = \sum_{i=1}^{k} r_i^2 \ ;$$

de plus : $\|U_1\|^2 = p$; $\|U_2\|^2 = q$.

Le cosinus de l'angle entre U_1 et U_2 : $R_{U_1U_2} = \dfrac{\sum_{i=1}^{k} r_i^2}{\sqrt{pq}}$ ne vaut 1 que si les sous-espaces $X'(E_1^*)$ et $Y'(E_2^*)$ coïncident ; les opérateurs U_1 et U_2 sont alors à distance nulle.

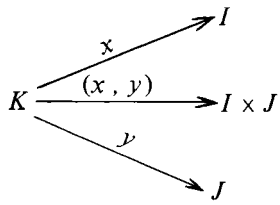
Si les sous-espaces $X'(E_1^*)$ et $Y'(E_2^*)$ sont D_p-orthogonaux les opérateurs U_1 et U_2 sont orthogonaux.

222 – Proximités entre caractères qualitatifs

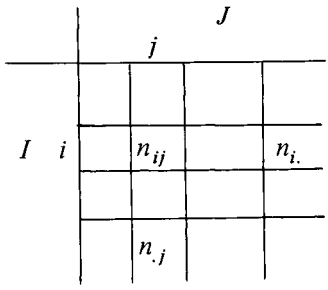
Notons ici comme au chapitre 13 de [5] :

. $K = \{k/1 = 1, \ldots, n\}$ l'ensemble des "individus" supposés tous munis de la masse $\dfrac{1}{n}$;

. $I = \{i/i = 1, \ldots, p\}$ et $J = \{j/j = 1, \ldots, q\}$ les ensembles de modalités des variables qualitatives x et y.



Au couple (x, y) est associé un tableau de contingence à p lignes et n colonnes :



$n_{ij} \doteq \text{card}\,[(x, y)^{-1}((i, j))]$;

$n_{i.} = \sum_{j=1}^{q} n_{ij}$;

$n_{.j} = \sum_{i=1}^{p} n_{ij}$.

On note P_{IJ} la loi de probabilité sur $(I \times J, P(I \times J))$ définie par les probabilités élémentaires :

$$p_{ij} = \frac{n_{ij}}{n} \quad ; \quad i = 1, \ldots, p \quad ; \quad j = 1, \ldots, q .$$

Les lois marginales sur $(I, P(I))$ et $(J, P(J))$ sont notées P_I et P_J ; elles sont définies respectivement par les probabilités élémentaires :

$$p_{i.} = \frac{n_{i.}}{n} \quad ; \quad i = 1, \ldots, p$$

$$p_{.j} = \frac{n_{.j}}{n} \quad ; \quad j = 1, \ldots, q .$$

Aux deux variables qualitatives x et y correspond le double schéma de dualité :

$$\begin{array}{ccccc} R^p = E_1 & \xleftarrow{X} & F & \xrightarrow{Y} & E_2 = R^q \\ D_{1/p_I} \downarrow \uparrow D_{p_I} & & \uparrow D_p = \frac{1}{n} I & & D_{p_J} \uparrow \downarrow D_{1/p_J} \\ E_1^* & \xrightarrow{X'} & F & \xleftarrow{Y'} & E_2^* \end{array}$$

E_1 et E_2 sont considérés comme les vectoriels des mesures sur $(I, P(I))$ et $(J, P(J))$ respectivement ; les simplexes S_1 et S_2 des lois de probabilité sur $(I, P(I))$ et $(J, P(J))$ sont munis respectivement de la distance du chi-deux de centre p_I (métrique D_{1/p_I}) et de la distance du chi-deux de centre p_J (métrique D_{1/p_J}).

Les projecteurs A_1 et A_2 sur $X'(E_1^*)$ et $Y'(E_2^*)$ ne sont autres que les opérateurs associés respectivement aux triplets $(X, D_{1/p_I}, D_p)$ et $(Y, D_{1/p_J}, D_p)$:

. $P(A_1, A_2) = tr\,(D_{1/p_I}\, P\, D_{1/p_J}\, P') = \Phi^2 + 1$

où Φ^2 est le phi-deux associé au tableau P, à p lignes et n colonnes, des probabilités p_{ij} ;

. $\Phi^2 = \frac{\chi^2}{n}$,

où χ^2 est le chi-deux associé au tableau de contingence des n_{ij}.

Si B_1 et B_2 désignent les opérateurs de D_p-projection sur les sous-espaces H_1 et H_2 des caractères centrés (cf. § 14) :

$$P(B_1 , B_2) = P(A_1 - A_j , A_2 - A_j) = P(A_1 , A_2) - P(A_j , A_2) .$$

La droite des constantes Δ_j étant dans le sous-espace $Y'(E_2^*)$:

$$P(A_j , A_2) = 1 = P(A_j , A_1) .$$

Le produit scalaire entre les opérateurs $U_1 = B_1$ et $U_2 = B_2$ associés aux variables qualitatives x et y a donc pour expression :

$$P(U_1 , U_2) = \Phi^2 = \frac{\chi^2}{n} .$$

Les sous-espaces H_1 et H_2 étant de dimensions respectives $(p - 1)$ et $(q - 1)$:

$$||U_1||^2 = p - 1 \quad ; \quad ||U_2||^2 = q - 1 .$$

D'où : $$R_{U_1 U_2} = \frac{\Phi^2}{\sqrt{(p-1)(q-1)}} = \frac{\chi^2}{n\sqrt{(p-1)(q-1)}} .$$

Le cosinus de l'angle entre les opérateurs U_1 et U_2 n'est autre que le coefficient T^2 introduit par Tschuprow pour mesurer le degré d'association entre deux variables qualitatives (réf. [10]).

223 – Proximité entre un caractère qualitatif et un paquet de caractères quantitatifs

Aux p caractères quantitatifs x^j est associé le tableau centré X ; au caractère qualitatif y est associé le tableau Y dont les lignes correspondent aux variables indicatrices des modalités de y (cf. [5] chapitre 12).

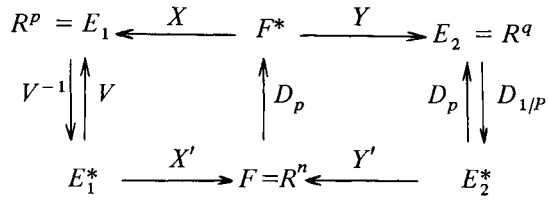
Le caractère qualitatif y induit sur l'ensemble des individus I, dont les n éléments sont munis des poids p_i, une partition $\{I_j / j = 1 , \ldots , q\}$.

A la classe I_j correspond :

– une masse : $P_j = \Sigma \{p_i / i \in I_j\}$;

– un centre de gravité : $g_j = \frac{1}{P_j} \Sigma \{p_i \, \underline{x}_i / i \in I_j\}$.

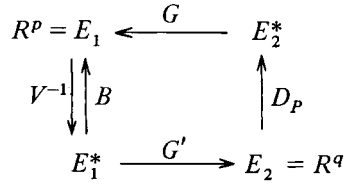
On a le schéma de dualité :



Les projecteurs A_1 et A_2 sur $X'(E_1^*)$ et $Y'(E_2^*)$ ne sont autres que les opérateurs associés aux triplets (X, V^{-1}, D_p) et $(Y, D_{1/P}, D_p)$; d'où :

$$P(A_1, A_2) = tr(V^{-1} B)$$

où $B = G D_p G'$ est la matrice d'inertie "inter-classe", G désignant le tableau à p lignes dont les q colonnes sont les vecteurs $\underline{g}_j$.



$tr(V^{-1} B)$ n'est autre que l'inertie à l'origine du nuage des centres de gravité $\underline{g}_j$, E_1 étant muni de la métrique de Mahalanobis V^{-1}.

La droite des constantes Δ_j étant D_p-orthogonale à $X'(E_1^*)$:

$$P(A_1, A_2 - A_j) = P(A_1, A_2).$$

Le produit scalaire entre l'opérateur $U_1 = A_1$ associé aux variables quantitatives x^j et l'opérateur $U_2 = A_2 - A_j$ associé à la variable qualitative y a donc pour expression :

$$\underline{P(U_1, U_2) = tr(V^{-1} B).}$$

On a de plus :

$$\|U_1\|^2 = p \; ; \qquad \|U_2\|^2 = q - 1.$$

Le cosinus de l'angle entre les deux opérateurs a donc pour expression :

$$R_{U_1 U_2} = \frac{tr(V^{-1} B)}{\sqrt{p(q-1)}}.$$

Remarques :

Si un seul caractère quantitatif x est considéré ($p = 1$), $tr(V^{-1}B)$, qui est le rapport de la variance inter-classe de x à sa variance totale, n'est autre que le carré du rapport de corrélation $\eta_{x/y}$ de x connaissant y ; en effet, la meilleure explication de x sachant que y a pris la modalité 1 n'est autre que la moyenne de x sur la classe 1 ; rappelons que le rapport précédent n'est pas symétrique :

$$\eta_{y/x} \neq \eta_{x/y} .$$

. dans le cas général ($p \neq 1$), si V est diagonale, $tr(V^{-1}B)$ n'est autre que la somme des carrés des rapports de corrélation $\eta_{xj/y}$.

3 – PRATIQUE DES OPERATEURS

Le fait d'avoir introduit sur le sous-espace des applications D_p-symétriques, le produit scalaire P permet de rendre opérationnelles les équivalences (a) et (b), (a') et (b'), (c), (c') et (c'') qui s'expriment en terme de proximité entre opérateurs.

31 – Description des opérateurs à l'aide de l'analyse en composantes principales

Deux façons de faire peuvent être retenues pour décrire un nuage de k opérateurs par l'analyse en composantes principales, suivant que l'on considère ces opérateurs comme des "individus" ou comme des "caractères".

311 – Opérateurs "individus"

On note :

– Δ le tableau des distances au sens de P entre les opérateurs ;

– π_j le poids associé au $j^{\text{ième}}$ opérateur $\left(\pi_j > 0 , \sum_{j=1}^{k} \pi_j = 1\right)$ et D_π la matrice diagonale des poids associée à l'ensemble des opérateurs.

Si ce sont les distances entre opérateurs qui nous intéressent –il s'agit d'en dresser un bilan– on représentera ces opérateurs dans un espace de petite dimension à l'aide de l'analyse en composantes principales en tirant les vecteurs propres (les composantes principales) associés aux plus grandes valeurs propres de l'opérateur $U = W \circ D_\pi$, la matrice associée à l'application W étant la matrice semi-définie positive des produits scalaires entre opérateurs centrés :

$$w_{ii'} = P(U_i - U\ ,\ U_{i'} - U) \qquad \text{avec} \qquad U = \sum_{i=1}^{k} \pi_i U_i\ .$$

Si :
$$P_{ii'} = P(U_i\ ,\ U_{i'})\ ;$$

$$P_{i.} = \sum_{i'=1}^{k} \pi_{i'} P(U_i\ ,\ U_{i'})\ ;\ P_{..} = \sum_{i=1}^{k} \sum_{i'=1}^{k} \pi_i\ \pi_{i'} P(U_i\ ,\ U_{i'})\ :$$

$$\underline{w_{ii'} = P_{ii'} + P_{..} - P_{i.} - P_{i'.}}\ .$$

Voici le schéma de dualité considéré ici au niveau de la description des opérateurs :

$$\begin{array}{ccc} R^{n^2} = E & \xleftarrow{\ Z\ } & F^* \\ P \big\downarrow\big\uparrow V & & W \big\downarrow\big\uparrow D_\pi \\ E^* & \xrightarrow{\ Z'\ } & F = R^k \end{array}$$

. $E = L(R^n\ ,\ R^n)$;

. $Z(F^*) \subset G \subset E$,

où G est le sous-espace des opérateurs D_p-symétriques sur F ;

. la matrice associée à l'application Z est le tableau des données obtenu en rangeant les opérateurs-individus "centrés" en colonnes.

312 – Opérateurs-caractères

Rappelons que l'analyse en composantes principales, effectuée sur un tableau centré "individus x caractères" X, fournit une double description des n individus-colonnes et des p caractères-lignes.

$$\begin{array}{ccc} E & \xleftarrow{\ X\ } & F^* \\ M \big\downarrow\big\uparrow V & & W \big\downarrow\big\uparrow D_p \\ E^* & \xrightarrow{\ X'\ } & F \end{array}$$

Le premier axe Δ_{u_1} est engendré par le vecteur M-normé $\underline{u}_1$ rendant maximum, sous la contrainte :

$$||\underline{u}|| = \sqrt{M(\underline{u})} = 1 ,$$

la quantité :

$$I_{\Delta_u^\perp} = \sum_{i=1}^{n} p_i M^2 (\underline{u} , \underline{x}_i)$$

qui est le moment d'inertie par rapport à l'hyperplan $\Delta_u^\perp$, M-orthogonal à la droite Δ_u engendrée par le vecteur M-normé $\underline{u}$ (cf. [5] chapitre 8).

Si $\underline{v} = M(\underline{u})$ et $\underline{c} = X'(\underline{v}) = X' \circ M(\underline{u})$:

$$I_{\Delta_u^\perp} = M(\underline{u} , V \circ M\underline{u}) = V(\underline{v}) = D_p(\underline{c}) .$$

Si λ_1 désigne la plus grande valeur propre de $V \circ M$:

$$V \circ M\underline{u}_1 = \lambda_1 \underline{u}_1 \quad ; \quad I_{\Delta_{u_1}^\perp} = \lambda_1 .$$

De façon duale :

$$\underline{v}_1 = M(\underline{u}_1) \quad ; \quad M \circ V\underline{v}_1 = \lambda_1 \underline{v}_1 ;$$

$$\underline{c}^1 = X' \circ M(\underline{u}_1) \quad ; \quad W \circ D_p \underline{c}^1 = \lambda_1 \underline{c}^1 .$$

On démontre sans difficulté le théorème suivant :

Théorème : *le caractère $\underline{c}$ de $X'(E^*)$, D_p-normé, rendant maximum l'indice :*

$$J_c = \sum_{j=1}^{p} \sum_{j'=1}^{p} m_{jj'} D_p(\underline{x}^j , \underline{c}) \, D_p(\underline{x}^{j'} , \underline{c})$$

est homothétique à la première composante principale $\underline{c}^1$.

Compte tenu de la remarque :

$$J_c = M(\underline{u}) \quad \text{si} \quad \underline{u} = X \circ D_p(\underline{c}) ,$$

rendre maximum $M(\underline{u})$, sous la contrainte "$D_p(\underline{c}) = 1$", revient à rendre maximum $D_p(\underline{c})$, sous la contrainte "$M(\underline{u}) = 1$".

Dans le cas particulier où la métrique M est diagonale :

$$J_c = \sum_{j=1}^{p} m_{jj} D_p^2 (\underline{x}^j , \underline{c}) = I_{\Delta_{\underline{c}}^\perp} ,$$

où $I_{\Delta_c^{\perp}}$ est le moment d'inertie du nuage des caractères $\underline{x}^j$ affectés des masses m_{jj} par rapport à l'hyperplan $\Delta_c^{\perp}$, D_p -orthogonal à la droite Δ_c engendrée par le vecteur $\underline{c}$.

En général :

. soit $M = \mathcal{I}$ et $J_c = \sum_{j=1}^{p} \text{cov}^2(x^j, c)$;

. soit $M = D_{1/\sigma^2}$ et $J_c = \sum_{j=1}^{p} \text{cor}^2(x^j, c)$.

Remarque :

Le sous-espace engendré par les l premières composantes principales n'est pas le sous-espace affine de dimension l le plus proche du nuage des caractères (ce sous-espace passe par le "centre de gravité" du nuage des caractères).

Par analogie avec ce que l'on fait en analyse en composantes principales, si ce sont les produits scalaires entre opérateurs qui intéressent au premier chef, car ils sont interprétés aussi bien lorsqu'ils sont nuls (orthogonalité) que lorsqu'ils sont maximum (colinéarité), c'est la matrice des produits scalaires entre opérateurs que l'on diagonalisera ; cette matrice sera assimilée à une matrice de variance ($M = \mathcal{I}$) si les opérateurs considérés ne sont pas normés, et à une matrice de corrélation ($M = D_{1/\sigma^2}$) si les opérateurs considérés sont normés.

Voici le schéma de dualité qui est ici considéré :

$$\begin{array}{ccc} R^k = E & \xleftarrow{\quad Z \quad} & F^* \\ M \downarrow \uparrow V & & W \downarrow \uparrow D_p \\ E^* & \xrightarrow{\quad Z' \quad} & F = R^{n^2} \end{array}$$

. $F = L(R^n, R^n)$;

. $Z'(E^*) \subset G \subset F$;

. la matrice associée à l'application Z est obtenue en rangeant les opérateurs-caractères U_j en lignes ;

. les éléments de la matrice V sont les produits scalaires $P(U_j, U_{j'})$;

. $M = \mathcal{I}$ ou $M = (\text{diag}(V))^{-1}$;

. D_p est ici la métrique de matrice :

$$D_p = \begin{bmatrix} p_1 & & & & & & & & 0 \\ & \ddots & & & & & & & \\ & & p_n & & & & & & \\ & & & p_1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & p_n & & & \\ & & & & & & \ddots & & \\ 0 & & & & & & & p_1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & p_n \end{bmatrix}$$

où les p_i sont les poids affectés aux éléments de l'ensemble I sur lequel ont été définis les opérateurs U_j.

Les opérateurs U_j, admettant pour vecteur propre de valeur propre nulle le vecteur $\underline{j}$ dont toutes les coordonnées sont égales à 1, sont bien "centrés".

On obtiendra ici une description des opérateurs en les représentant par leurs projections dans le sous-espace engendré dans F par les deux ou trois premières composantes principales ; si $\underline{u}_i$ désigne le vecteur propre, dont les coordonnées sont notées u_i^j, associé à la i$^{\text{ème}}$ valeur propre λ_i de la matrice des produits scalaires entre opérateurs (métrique I) ou entre opérateurs normés (métrique $D_{1/\sigma 2}$), la coordonnée de l'opérateur U_j par rapport à l'axe $\Delta_c i$ engendré par la i$^{\text{ème}}$ composante principale n'est autre que $\sqrt{\lambda_i}\, u_i^j$ (cf. [5] chapitre 8 § 82).

32 – Pratique des opérateurs dans une optique de description

Il est possible de comparer avec les opérateurs des descriptions multiples de l'ensemble I fournies indifféremment par des couples (D_j, D_p) ou des triplets (X_j, M_j, D_p) ; ici on utilisera de préférence pour décrire les opérateurs la méthode proposée en 311, les distances entre opérateurs-individus étant mieux mises en évidence.

Avec ses modèles "INDSCAL" et "IDIOSCAL" [2] et [4], J.D. CARROLL, analysant la subjectivité des avis d'un ensemble d'experts, cherche à retrouver les critères, et les métriques utilisées sur ces critères, retenus par ces experts pour placer des objets les uns par rapport aux autres.

Il fait donc l'hypothèse qu'il existe un tableau de données "objets x critères" X, dont les n colonnes symbolisent les objets (supposés munis des masses p_i) et les p lignes les critères, tel qu'à chaque expert "j", dont l'avis est traduit sous forme du tableau de distances entre objets D_j, corresponde une métrique euclidienne M_j qui serait la métrique utilisée par l'expert "j" sur les colonnes de X pour exprimer son avis "D_j" sur les proximités entre objets.

A l'expert "j" est alors associé le schéma de dualité :

$$\begin{array}{ccc} R^p = E & \xleftarrow{\quad X \quad} & F^* \\ M_j \downarrow & & W_j \downarrow\uparrow D_p \\ E^* & \xrightarrow{\quad X' \quad} & F = R^n \end{array} \qquad W_j = X' \circ M_j \circ X.$$

Les opérateurs correspondants respectivement au couple (D_j, D_p) et au triplet (X, M_j, D_p) étant supposés coïncider quel que soit "j", il s'agit de retrouver X et les M_j.

Si on affecte chacun des experts d'un poids π_j, on peut aisément avoir une idée de la diversité des avis des experts en décrivant, comme il est indiqué précédemment, par l'analyse en composantes principales, l'ensemble des opérateurs $U_j = W_j \circ D_p$.

L'opérateur "moyen" s'écrit :

$$U = \sum_{j=1}^{k} \pi_j U_j = W \circ D_p \qquad \text{avec :} \quad W = \sum_{j=1}^{k} \pi_j W_j = X' \circ M \circ X \ ;$$

$$M = \sum_{j=1}^{k} \pi_j M_j \ .$$

$$\begin{array}{ccccc} E & \xleftarrow{\quad T \quad} & E & \xleftarrow{\quad X \quad} & F^* \\ I \downarrow & & M \downarrow & & W \downarrow\uparrow D_p \\ E^* & \xrightarrow{\quad T' \quad} & E^* & \xrightarrow{\quad X' \quad} & F \end{array}$$

Les composantes principales correspondant au triplet (X, M, D_p), vecteurs propres de l'opérateur moyen U, permettent d'obtenir une description moyenne simplifiée des objets et de considérer le triplet équivalent (C, I, D_p) (cf. 121).

Si la matrice de changement de base associée au système des axes principaux est notée T :

$$C = T \circ X .$$

On a pour tout "j" :

$$W_j = X' \circ M_j \circ X = C' \circ N_j \circ C \qquad \text{avec : } N_j = T^{-1\prime} \circ M_j \circ T^{-1} .$$

Les triplets (X, M_j, D_p) et (C, N_j, D_p) étant équivalents, pour résoudre le problème de Carroll il suffit de rechercher les métriques N_j, ayant obtenu C en retenant les vecteurs propres associés au deux ou trois plus grandes valeurs propres positives de l'opérateur moyen U ; les éléments des matrices W_j étant obtenus à partir des formules rappelées au paragraphe 11, exploitant les équations matricielles :

$$W_j = C'N_jC \quad ; \quad j = 1, \ldots, k ,$$

on "estimera" les éléments des matrices N_j, qui doivent être définies positives, en effectuant k régressions linéaires.

Si le modèle est bon, la classification des experts suivant les métriques M_j doit être voisine de celle suivant les opérateurs W_j.

33 – Pratique des opérateurs dans une optique de prévision

En associant des opérateurs, qui sont ici des projecteurs D_p-symétriques, aux paquets de variables quantitatives et aux variables qualitatives, on se donne un moyen pour comparer globalement les "potentiels de prévision" que représentent ces variables.

Ici, les produits scalaires entre opérateurs intéressant l'analyste aussi bien quand ils sont maximum (dépendance maximum) que lorsqu'ils sont nuls (orthogonalité), c'est la méthode précisée en 312 qui sera utilisée pour décrire l'ensemble des opérateurs.

Dans le cas particulier où les opérateurs sont associés à des variables qualitatives, ces opérateurs étant normés, on obtiendra en diagonalisant la matrice des T^2 de Tchuprov (cf. 222) une description globale des liaisons entre les variables qualitatives ; deux opérateurs orthogonaux correspondront à deux variables qualitatives indépendantes.

Conclusion

A tous les tableaux de données envisagés dans la première partie, un opérateur peut être associé ce qui permet d'étudier rigoureusement l'équivalence entre deux tableaux de données dans les termes d'une égalité entre opérateurs. L'approche apparaît intéressante d'autant plus qu'elle permet de retrouver dans des situations particulières des coefficients déjà proposés leur fournissant ainsi un cadre général. Ces résultats sont une invitation à poursuivre l'explication d'autres types de tableaux de données par cette technique.

BIBLIOGRAPHIE

[1] J. ALIA (1973) – Le prix d'un Français – *Le Nouvel Observateur* n° 463.

[2] J.M. BOUROCHE et A.M. DUSSAIX (1973) – Le modèle INDSCAL et IDIOSCAL. Méthodes et utilisation METRA – Direction scientifique – Note de travail n° 152.

[3] J.M. BRAUN (1974) – Etude des séries chronologiques multiples par l'analyse des données. Rapport C.E.A. – R – 4561.

[4] J.D. CARROLL et J.J. CHANG (1970) – Analysis of individual differences in multidimensional scaling via an n-way generalization of "ECKART-YOUNG" decomposition. *Psychometrika,* vol. 35 n° 3.

[5] F. CAILLIEZ, J.P. MAILLES, J.P. NAKACHE, J.P. PAGES (1972) – Analyse des données multidimensionnelles. Centre d'Etudes Economiques d'Entreprises (C.3E), Tomes 2 et 3.

[6] Y. ESCOUFIER (1970) – Echantillonnage dans une population de variables aléatoires réelles. Thèse de Doctorat d'Etat. Université de Montpellier.

[7] Y. ESCOUFIER (1971) – Liaison entre groupes d'aléas. *Revue de Statistique Appliquée* n° 19 (2).

[8] Y. ESCOUFIER (1973) – Vecteurs aléatoires équivalents du point de vue de l'analyse en composantes principales. Université de Montpellier – Rapport technique n° 7301.

[9] K.G. JORESKOG (1963) – Statistical estimation in factor analysis. Almqvist. Wiksell (Stockholm).

[10] M.G. KENDALL et A. STUART (1953) – The advanced theory of statistics. c. GRIFFIN, London.

[11] J.B. KRUSKAL (1964) – Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika,* vol. 29.

[12] D.N. LAWLEY et A.E. MAXWELL (1963) – Factor analysis as a statistical method. Butterworths.

[13] G. SAPORTA (1975) – Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Thèse de 3ème cycle Université de Paris VI.

[14] R.N. SHEPARD (1962) – The analysis of proximities : multidimensional scaling with an unknown distance function. *Psychometrika*, vol. 27.

[15] R.N. SHEPARD (1974) – Representation of structure in similarity data : problems and prospects. *Psychometrika*, vol. 39 n° 4.

[16] F. TESTU (1975) – Application de l'analyse de données à l'étude des statistiques de causes de décés. Thèse de 3e cycle. Université Pierre et Marie Curie.