MATHÉMATIQUES ET SCIENCES HUMAINES

A. LECLERC

Quelques propriétés optimales en analyse des données, en termes de corrélation entre variables

Mathématiques et sciences humaines, tome 70 (1980), p. 51-67

<http://www.numdam.org/item?id=MSH_1980__70__51_0>

Math. Sci. hum. (18e année, n°70, *1980*, p.51-67)

QUELQUES PROPRIETES OPTIMALES EN ANALYSE DES DONNEES, EN TERMES DE CORRELATION ENTRE VARIABLES

A. LECLERC *

I. INTRODUCTION

Les propriétés optimales des méthodes d'analyse des données sont nombreuses, et peuvent se ranger sous plusieurs chapitres ; on pourrait distinguer, entre autres : des propriétés en terme de distance (à minimiser), de variance ou d'inertie (à maximiser) ; enfin on peut présenter certaines méthodes comme recherche de variables sur les observations, maximisant une fonction des coefficients de corrélation entre ces variables. Des articles plus ou moins récents présentent des propriétés optimales sous un aspect plus général, ce qui a un intérêt évident (6, 13, 25). Notre objectif est plus limité ; nous ne cherchons pas à être exhaustif, mais à rapprocher quelques propriétés simples de différentes méthodes, dont les présentations les plus courantes accentuent souvent les différences.

On trouvera donc ici une liste commentée de quelques propriétés optimales, en termes de corrélation entre variables, principalement pour l'analyse en composantes principales (A. C. P.) et diverses variantes de l'analyse des correspondances (A. F. C.) et de l'analyse canonique. La plupart des résultats présentés ne sont pas originaux, mais ils restent mal connus et dispersés dans la littérature, chaque article ou document ne traitant en général que d'une méthode.

II. GENERALITES, NOTATIONS

On suppose que l'on connait, pour un échantillon de taille n, les réalisations d'un certain nombre de variables aléatoires quantitatives ou qualitatives.

On distinguera le cas où l'on a un ensemble de p variables aléatoires :

$X_1 \ldots\ldots X_p$

Et le cas où l'on a deux ensembles de variables :

*Groupe de Recherche INSERM, U 88, 91, bd de l'Hôpital, F-75634 PARIS Cedex 13

$X_1 \ldots\ldots X_p$

$Y_1 \ldots\ldots Y_r$

Ce dernier cas est celui qui se présente, par exemple, dans une enquête où l'on a d'une part des caractéristiques socio-démographiques, d'autre part des réponses à des questions propres au thème de l'enquête ; on peut avoir p = 1 et (ou) r = 1.

Une variable quantitative peut être considérée comme une application de l'ensemble des observations sur l'ensemble des réels.

Une variable qualitative X_i prend ses valeurs dans l'ensemble des modalités ; on affectera d'une étoile X_i^+, une variable quantitative construite à partir de la variable qualitative X_i. X_i^+ est obtenue en appliquant à X_i une fonction-codage, qui à chaque modalité de X_i associe un réel. On parlera, pour simplifier, du codage X_i^+ ; à une variable qualitative X_i peuvent être associées plusieurs variables quantitatives qui seront alors notées $X_i^{+1} \ldots\ldots X_i^{+k}$

Pour des variables X_i quantitatives, ou pour des variables codées X_i^+ ou Y_j^+, on peut alors parler de moyenne, de variance (VAR), de coefficient de corrélation (CORR) et de covariance (COV). Dans la suite il s'agira toujours de valeurs empiriques, et les variables codées seront, dans la plupart des cas, définies à un changement de moyenne et d'écart - type près, c'est-à-dire que l'on pourra les prendre centrées réduites.

On appellera $a_1 \ldots a_p$ des coefficients numériques associés aux X_i, $b_1 \ldots b_r$ des coefficients associés aux Y_j.

On va voir qu'un certain nombre de problèmes qui se formulent comme la recherche de coefficient et (ou) de codages, à associer aux variables de départ, de façon à maximiser une fonction linéaire ou quadratique des coefficients de corrélations, sous certaines contraintes, ont comme solution une méthode d'analyse des données.

La suite de l'exposé renvoie à des méthodes ou à des "sous-méthodes", principalement différentes variantes de l'analyse des correspondances (sur tableau de Burt, tableau disjonctif complet, juxtaposition de tableaux de contingence) dont on pourra trouver une présentation rapide dans (20). Pour des résultats de base, on pourra se référer à (2, 4, 14, 26) et à (7, 15) pour quelques résultats ou démonstrations, particulièrement sur les tableaux de Burt et l'analyse des correspondances sur tableau disjonctif complet

III. UN SEUL GROUPE DE VARIABLES QUANTITATIVES

III.1. Soient $X_1 \ldots X_p$ des variables quantitatives. On suppose que l'on cherche des coefficients a_{i1} maximisant :

$$\sum_{i=1}^{p} \sum_{j=1}^{p} a_{i1}\ a_{j1}\ \mathrm{CORR}\,(X_i, X_j) \qquad [1]$$

sous la contrainte $\Sigma\, a_{i1}^2 = 1$

La solution est l'analyse en composantes principales, normée, des variables X_i. Le vecteur des a_{i1} est le vecteur propre, normé, de la matrice de corrélation, associé à la plus grande valeur propre. Celle-ci est égale à la quantité à maximiser.

Aux valeurs propres suivantes sont associés des vecteurs propres maximisant une expression analogue à 1, fonction de coefficients a_{iq} vérifiant des contraintes supplémentaires d'orthogonalité (voir tableau 1).

III.2. Si l'on cherche une fonction Z_1, combinaison linéaire des variables X_i, telle que soit maximum :

$$\sum_{i=1}^{p} \mathrm{CORR}^2\,(Z_1, X_i) \qquad [2]$$

la solution est aussi donnée par le premier facteur de l'analyse en composantes principales normée . Aux différents facteurs sont associés des variables Z_k, non corrélées entre elles. La démonstration est évidente. Un résultat plus général est donné par Okamoto (25).

IV. UN SEUL GROUPE DE VARIABLES QUALITATIVES

IV.1. Soient $X_1 \ldots X_p$ des variables qualitatives.
On cherche une fonction Z_1 sur les observations, somme de p termes dépendant chacun de la valeur prise par une variable pour l'observation, et des codages associant à chaque variable de départ une variable quantitative X_i^{+1} , tels que soit maximum :

$$\sum_{i=1}^{p} \mathrm{CORR}^2\,(Z_1, X_i^{+1}) \qquad [3]$$

La solution est donnée par le premier facteur de l'analyse des correspondances du tableau disjonctif complet, croisant l'ensemble des observations avec l'ensemble des modalités de réponse, en un tableau de 0 et de 1 se présentant ainsi (3) :

	X_1		X_2		.....	X_P	
1	1	0	0	0	1	1	0
2	0	1	0	1	0	1	0
•							

(ici, la variable X_1 serait à 2 modalités de réponses, la variable X_2 à 3 modalités ; à l'observation 1 serait associée la première modalité pour X_1, la troisième pour X_2 .....)

On peut prendre comme variable Z_1 la coordonnée d'une observation sur le premier axe factoriel, et le codage X_i^{+1} associe à chaque modalité de X_i sa coordonnée sur le premier axe factoriel.

Chaque facteur de l'analyse fournit une fonction Z_q et un ensemble de p codages, maximisant sous contraintes une expression semblable à [3]. On trouvera dans le tableau 1 l'expression de la quantité maximisée, et des contraintes imposées aux variables.

Une démonstration de ces résultats se trouve dans (22), avec l'équivalence entre l'analyse des correspondances sur tableau disjonctif complet, et la méthode d'analyse canonique généralisée proposée par J.D. Carroll (5) appliquée au cas particulier de tableaux logiques. Ces résultats sont aussi présents dans (23) et (27) mais pas directement sous cette forme. On peut aussi les considérer comme un cas particulier de ce que l'on a dans le cas d'une juxtaposition de tableaux de contingence (18) ce qui donne plus facilement l'expression des contraintes imposées aux variables.

IV.2. On cherche maintenant à associer à chaque variable un codage X_i^{+1} et un coefficient a_{i1} positif de façon à maximiser :

$$\sum_{i=1}^{p} \sum_{j=1}^{p} a_{i1}\, a_{j1}\, \mathrm{CORR}\,(X_i^{+1}, X_j^{+1}) \qquad [4]$$

sous la contrainte $\Sigma\, a_{i1}^2 = 1$

La solution est encore l'analyse des correspondances du tableau disjonctif complet ou, ce qui donne des facteurs équivalents, l'analyse des correspondances du tableau de Burt, croisant l'ensemble des modalités des variables X_i avec elles-mêmes.

Chaque facteur fournit un ensemble de coefficients, et de codages, vérifiant certaines contraintes (voir le tableau 1) et maximisant une expression de même forme que [4] .

Les variables codées, X_i^{+q} , sont obtenues à partir des coordonnées des différentes modalités de X_i sur l'axe q. Un coefficient a_{iq} est la racine carré de la somme des contributions absolues des modalités de X_i au facteur q.

Ces résultats peuvent se trouver dans (23) ou (27), plutôt comme cas particulier de résultats plus généraux, et dans (3) sous une forme un peu différente. On en trouvera une démonstration en annexe.

IV.3. Ce qui précède peut s'écrire de façon légèrement différente :

- trouver des codages et des coefficients qui maximisent :

$$\sum_{i=1}^{p} \sum_{j \neq i} a_{iq}\, a_{jq}\, \mathrm{CORR}(X_i^{+q}, X_j^{+q}) \qquad [5]$$

sous la contrainte $\Sigma\, a_{i1}^2 = 1$, et des contraintes d'orthogonalité.

- et : trouver des codages optimaux (que l'on appelle toujours X_i^{+q} bien qu'il ne s'agisse pas des mêmes variables) maximisant :

$$\sum_{i=1}^{p} \sum_{j=1}^{p} \mathrm{COV}\,(X_i^{+q}, X_j^{+q}) \qquad [6]$$

sous la contrainte : $\sum_{i=1}^{p} \mathrm{VAR}\,(X_i^{+q}) = 1$

à laquelle s'ajoutent des contraintes d'orthogonalité.

V. UN SEUL GROUPE DE VARIABLES QUALITATIVES OU QUANTITATIVES

Il existe d'autres solutions au problème de la description des relations entre variables qualitatives ; Masson (23) propose plusieurs alternatives à l'analyse des correspondances sur tableau disjonctif complet, qui ont l'inconvénient de n'avoir pas de solution mathématique simple. Saporta (28) propose d'appliquer l'analyse des correspondances à un tableau déduit du tableau disjonctif complet par pondération des différents sous-tableaux, ce qui maximise des expressions de la forme :

$$\sum_{i=1}^{p} \mathrm{CORR}^4\,(Z_1, X_i^{+1})$$

Les différentes méthodes, qui possèdent des propriétés optimales en termes de corrélation, sont intéressantes d'un point de vue théorique ; leur usage ne présente pas d'avantage, pour le praticien, sur ce qui est d'un usage plus courant , c'est-à-dire l'analyse des correspondances sur tableau disjonctif complet.

Le rapprochement des expressions [1]et [4] , [2] et[3] , montre l'analogie très étroite de cette dernière méthode avec l'analyse en composantes principales. (Si les contraintes sur les coefficients a_{iq} ne sont pas les mêmes, cela tient à ce qu'un changement de signe éventuel d'une variable interviendra en ACP par l'intermédiaire d'un coefficient négatif, et en AFC au niveau du codage). On peut considérer l'AFC sur tableau disjonctif complet comme une extension de l'ACP à des variables qualitatives, celles-ci étant transformées en variables quantitatives,

de façon optimale, puis soumises à une analyse en composantes principales. Ceci a été mis en évidence par Masson en 1974 (23), puis repris et développé par Saporta (27). Hill (9) le rappelle aussi, en se basant sur une propriété de maximisation de variance. La présentation la plus accessible de cette équivalence est cependant celle de l'article de B. Escofier (8), qui propose une technique permettant d'analyser simultanément variables qualitatives et quantitatives, en utilisant les programmes classiques d'AFC. Ceci maximise (pour le premier facteur) :

$$\sum_{i=1}^{P} \text{CORR}^2 \ (Z_1, X_i^{+1}) \qquad [7]$$

où X_i^{+1} est, soit la variable de départ (si elle est quantitative), soit un codage propre au facteur (pour les variables qualitatives).

Hill propose la même méthode (10), avec une solution mathématique par itération, beaucoup plus lourde. Une alternative au traitement simultané de variables qualitatives et quantitatives est proposée par Tenenhaus (29) avec un codage unique des variables qualitatives, mieux adapté au cas où les modalités sont ordonnées. La méthode cependant est longue et les résultats proches de ce que l'on trouve par d'autres méthodes (19).

Dans le cas de variables dichotomiques, à chaque variable ne peut être associé qu'un codage (à des changements de moyenne et de variance près). Il y a équivalence entre l'AFC du tableau dédoublé (tableau à 2p colonnes) et l'ACP du tableau non dédoublé, à p colonnes. Ce résultat est démontré par Nakhlé, sous une forme plus générale (24).

VI. DEUX GROUPES DE VARIABLES QUANTITATIVES

L'on suppose que l'on cherche à décrire les relations entre deux groupes de variables quantitatives :

$$X_1 \ldots X_p \text{ et } Y_1 \ldots Y_r$$

on sait que l'analyse canonique est la recherche de coefficients $a_{11} \ldots a_{p1}$ et $b_{11} \ldots b_{r1}$ maximisant

$$\text{COV}\left[\sum_{i=1}^{p} a_{i1} X_i, \sum_{j=1}^{r} b_{j1} y_j\right] = \sum_{i=1}^{p}\sum_{j=1}^{r} a_{i1} b_{j1} \text{COV}(X_i, Y_j) \qquad [8]$$

sous les contraintes :

$$\text{VAR}\left[\sum_{i=1}^{p} a_{i1} X_i\right] = 1 \text{ et VAR}\left[\sum_{j=1}^{r} b_{j1} Y_j\right] = 1$$

$$\underline{\text{ou}} \ \text{VAR}\left[\sum_{i=1}^{p} a_{i1} X_i\right] + \text{VAR}\left[\sum_{j=1}^{r} b_{j1} Y_j\right] = 2$$

(pour l'équivalence entre les deux ensembles de contraintes, on pourra se référer par exemple au chapitre III.1. de 15).

Le processus se poursuit par la recherche de couples de fonctions

$$\sum_{i=1}^{p} a_{iq} X_i \quad \text{et} \quad \sum_{j=1}^{r} b_{jq} Y_j$$

de covariance maximale et orthogonales aux précédentes (11).

VII. DEUX GROUPES DE VARIABLES QUALITATIVES

On suppose maintenant que les variables sont qualitatives. On cherche à associer à chacune une variable quantitative X_i^{+1} ou Y_j^{+1} et un coefficient a_{i1} ou b_{j1} positif, tels que soit maximum :

$$\text{COV}[\sum_{i=1}^{p} a_{i1} X_i^{+1}, \sum_{j=1}^{r} b_{j1} Y_j^{+1}] \qquad [9]$$

Deux solutions au moins existent à ce problème, selon les contraintes que l'on impose :

VII.1. Reprenant ce que l'on a pour l'analyse canonique "classique" on peut introduire les contraintes :

$$\text{VAR}[\sum_{i=1}^{p} a_{i1} X_i^{+1}] = 1$$

$$\text{VAR}[\sum_{j=1}^{r} b_{j1} Y_j^{+1}] = 1$$

Il s'agit très exactement d'une analyse canonique classique sur les variables indicatrices des modalités des variables des deux ensembles, à ceci près que les contraintes imposent de rejeter des solutions triviales, qui seraient la somme des variables indicatrices des modalités d'une même variables. On obtient une suite de couples de combinaisons linéaires des variables indicatrices, avec des conditions d'orthogonalité qui se déduisent de l'analyse canonique "classique" (voir tableau 1).

VII.2. Avec les contraintes :

$$\sum_{i=1}^{p} a_{i1}^2 = 1 \quad \text{et} \quad \sum_{j=1}^{r} b_{j1}^2 = 1 \qquad \text{(et les codages réduits)}$$

on a une autre solution : l'analyse des correspondances du tableau "juxtaposition de tableaux de contingence" croisant les modalités des variables X_i avec les modalités des variables Y_j (17).

A chaque facteur de l'analyse correspondent des pondérations a_{iq} et b_{jq}, et des codages X_i^{+q} et Y_j^{+q}, maximisant une expression semblable à [9], sous des contraintes précisées dans le tableau 1. Les variables quantitatives X_i^{+q} ou Y_j^{+q} sont données par les coordonnées des différentes modalités le long de l'axe q. Les coefficients a_{iq} ou b_{jq} sont les racines carrés des sommes des contributions absolues des modalités associées à une variable.

Une démonstration de cette propriété optimale de l'analyse des correspondances se trouve dans (16).

On peut remarquer que l'on introduit ici une contrainte sur la somme des variances des variables $a_{i1} X_i^{+1}$ (car on peut toujours choisir des codages de variance 1), alors qu'en VII.1. la contrainte est sur la variance de la somme. Des contraintes sur la somme des variances existent aussi dans l'AFC d'un tableau disjonctif complet et, on l'a vu, en analyse canonique.

Pour revenir à la comparaison des contraintes VII.1. et VII.2., du point de vue de la pratique, on peut penser que la première méthode (analyse canonique) est plus puissante, car elle tient compte des associations entre variables d'un même groupe. La seconde méthode (analyse des correspondances sur juxtaposition de tableaux de contingence) est très proche de la démarche simple de l'examen des tableaux croisés, avec calculs de Khi-2, et on dispose d'aides à l'interprétation assez variées (17).

VIII. CAS PARTICULIERS

Les cas particuliers où l'un ou l'autre groupe de variables qualitatives - ou les deux - ne comportent en fait qu'une variable, méritent d'être examinés.

VIII.1. On suppose que l'on cherche à décrire les relations entre p variables qualitatives $X_1 \ldots X_p$ et une variable qualitative Y.

On peut formuler ainsi un problème de maximisation :
trouver des codages centrés réduits $X_i^{+1} \ldots X_p^{+1}$, et Y^{+1}, et des coefficients $a_{i1} \ldots a_{p1}$ tels que soit maximum :

$$\sum_{i=1}^{p} a_{i1} \, \mathrm{CORR}\,(X_i^{+1}, Y^{+1}) \qquad [10]$$

sous une contrainte de type "analyse canonique" :

$$\mathrm{VAR}\,[\sum_{i=1}^{p} a_{i1} X_i^{+1}] = 1$$

ou de type "analyse des correspondances" :

$$\Sigma a_{il}^2 = 1$$

Si l'on retient la première contrainte, on est dans un cas particulier de l'analyse canonique.

En adoptant la seconde contrainte, on obtient une formulation différente pour l'expression à maximiser. En effet, quel que soit le codage des variables, les coefficients a_{il} qui maximisent [10]sont de la forme : $a_{il} = c\ CORR\ (X_i^{+l}, Y^{+l})$, le coefficient étant choisi de façon à respecter la norme de a_l.

Le problème de la maximisation s'écrit donc plus simplement : trouver des codages des variables maximisant :

$$\sum_{i=1}^{p} CORR^2 (X_i^{+l}, Y^{+l}) \qquad [11]$$

On est dans un cas particulier de l'analyse des correspondances sur juxtaposition de tableaux de contingence, où la variable qualitative Y est croisée successivement avec les p variables qualitatives X_i. Chaque facteur q fournit un ensemble de (p+1) codages des variables : $X_1^{+q} \ldots X_p^{+q}, Y^{+q}$. Ces codages sont donnés par les coordonnées des modalités des variables sur l'axe factoriel q.

Les propriétés de cette application particulière de l'analyse des correspondances, y compris la propriété optimale [11] sont développées dans (18). Ces résultats ont été donnés aussi par Cazes (6) sous une forme un peu différente .

VIII.2. Dans le cas particulier où l'on a aussi p = 1, c'est-à-dire dans le cas où l'on cherche à décrire les relations entre deux variables qualitatives X et Y, les deux types de contraintes "analyse canonique" ou "analyse des correspondances" sont équivalentes. On retrouve les résultats connus de l'analyse des correspondances appliquée à un tableau de contingence dont on peut trouver une démonstration par exemple dans (7);les facteurs q donnent des codages X^{+q} et Y^{+q} tels que soient maximum les quantités :

$$CORR^2 (X^{+q}, Y^{+q})$$

sous des contraintes d'orthogonalité :

$$CORR\ (X^{+q}, X^{+q'}) = CORR\ (Y^{+q}, Y^{+q'}) = 0$$

Il est équivalent dans ce cas de parler d'analyse des correspondances ou d'analyse canonique (12, page 568 ; 2, page 182).

IX. CONSEQUENCES ET CONCLUSIONS

On a rappelé rapidement les analogies existant entre des méthodes complémentaires. La présentation en termes de maximisation d'une fonction des coefficients de corrélation peut avoir l'intérêt de souligner certains points

importants dans l'application des méthodes d'analyse des données :

- Certaines propriétés des méthodes sont basées sur les coefficients de corrélation entre des variables codées , à valeurs discrètes. Ces coefficients ne peuvent être considérés que comme des mesures, parmi d'autres, d'association entre variables, la non-normalité des variables restreignant largement les possibilités d'interprétation. On pourrait imaginer des méthodes utilisant des mesures d'association entre variables mieux adaptées à la nature de celles-ci. Pour les méthodes proposées, les résultats sont sensibles à ce qui peut entraîner des perturbations (ou un aléa trop important) des coefficients de corrélation entre variables discrètes, en particulier les modalités à effectif trop faible (1).
- On voit que toutes les méthodes donnent des variables quantitatives sur les observations , variables dont certaines peuvent être considérées comme normales. Ceci est souvent utile dans la pratique, où l'on peut chercher à disposer de variables quantitatives se prêtant à des opérations ultérieures, telles que la mise en évidence de différences entre groupes ou un codage quantitatif de modalités ordonnées.
- Dans la recherche des codages optimaux, de variables qualitatives, en vue de la maximisation d'un critère, toutes les variables ne sont pas à égalité. Pour une variable à nombreuses modalités, l'éventail des codages possibles est plus large, il y a plus de chance a priori pour qu'il s'en trouve un qui soit fortement corrélé aux autres variables. Dans toute analyse, une variable à beaucoup de modalités pèse plus lourd, même si cela ne peut pas être exactement quantifié (21). Dans la pratique il est préférable, dans la mesure du possible, de ne pas analyser conjointement des variables trop hétérogènes vis-à-vis du nombre de modalités, et de réserver des méthodes traitant conjointement des variables qualitatives et quantitatives (8, 29) aux cas où un codage plus homogène est impossible.

L'examen des méthodes relativement nombreuses proposées pour décrire des relations entre variables ou les ressemblances entre individus, incite à penser que les méthodes les plus classiques, Analyse en Composantes Principales et Analyse des Correspondances sous différentes formes, complétées par des méthodes de classification, restent les outils les plus utiles. Ces méthodes, outre les propriétés mathématiques qu'elles possèdent, ont l'avantage d'avoir été éprouvées par la pratique, ce qui a conduit à un certain nombre de règles d'utilisation (codage et choix des variables, interprétation et validation des résultats) qui permettent actuellement une utilisation dans de bonnes conditions.

REFERENCES

1 AGRESTI A., "The effect of category choice on some ordinal measures of association" J. Amer. Statis. Assoc., 71, 353 (1976), 49-55

2 BENZECRI, J.P. : L'analyse des données. II. L'analyse des correspondances, Dunod, 1973.

3 BENZECRI, J.P. "Sur l'analyse des tableaux binaires associés à une correspondance multiple". Les Cahiers de l'Analyse des Données, II, 1 (1977), 55-71

4 CAILLIEZ F., PAGES J.P. Introduction à l'analyse des données. Paris, SMASH, 1976.

5 CARROLL J.D. "A generalization of canonical correlation analysis to three or more sets of variables". In Proc. 76th Conv. Amer., Psych. Assoc., 227-228

6 CAZES P. "Etude de quelques propriétés extrêmales des facteurs issus d'un sous tableau d'un tableau de Burt". Les Cahiers de l'Analyse des Données, II, 2 (1977), 143-160.

7 CEHESSAT R. Exercices commentés de statistique et informatique appliquées. Paris, Dunod, 1976.

8 ESCOFIER B. "Traitement simultané de variables qualitatives et quantitatives en Analyse Factorielle". Les Cahiers de l'Analyse des Données, IV, 2 (1979), 137-146.

9 HILL M.O. "Correspondence analysis : a neglected multivariate method". Appl. Statist. 23, 3 (1974) 340-354.

10 HILL M.O., SMITH A.J.E. "Principal component analysis of taxonomic data with multi-state discrete characters". Taxon 25 (2/3) (1976) 249-255.

11 HOTELLING H. "relations between two sets of variables". Biometrika, 28, (1936), 321-377.

12 KENDALL M.G. The advanced theory of statistics, vol II, Londres, Griffin, 1961.

13 KOBILINSKI A. "Ordre entre formes quadratiques Application à l'optimalité de sous espace en analyse des données". Rev. Statist. Appl., 27, 1 (1979) 45-54.

14 LEBART L., MORINEAU A., FENELON J.P. Traitement des données statistiques. Paris, Dunod, 1979.

15 LEBART L., MORINEAU A., TABARD N. Techniques de la description statistique. Paris, Dunod, 1977.

16 LECLERC A. Etude de certains types de tableaux par l'analyse des correspondances. Thèse de 3ème cycle, Université Paris VI, 1973.

17 LECLERC A. "L'analyse des correspondances sur juxtaposition de tableaux de contingence". Rev. Stat. Appl., 23, 3 (1975), 5-16.

18 LECLERC A. "Une étude de la relation entre une variable qualitative et un groupe de variables qualitatives". Internat. Stat. Rev., 44, 2 (1976), 241-248.

19 LECLERC A., MACQUIN A. "La description des relations entre variables qualitatives à modalités ordonnées : comparaison de trois méthodes". Séminaire IRIA. Classification automatique et perception par ordinateur, 1978, 31-44.

20 LECLERC A., AIACH P. "Mesures de l'importance des valeurs propres en analyse des données. Application à l'analyse en composantes principales de variables catégorisées". Rev. Stat. Appl., 26 , 1 (1978) 5-21.

21 LECLERC A. Pratique des analyses de données. Cours d'option du CSA, Paris, Université Paris VI, document multigraphié, 1977.

22 MACQUIN A. La méthode PRINQUAL. Présentation et comparaison avec l'analyse des correspondances Rapport d'application, Cycle de Stat. Appliquée, Paris, Université Paris VI, 1977.

23 MASSON M. Processus linéaires. Analyse non linéaire des données. Thèse d'Etat, Université Paris VI, 1974.

24 NAKHLE F. "Sur l'analyse d'un tableau de notes dédoublées" Les Cahiers de l'Analyse des Données, I, 3 (1976) 243-257.

25 OKAMOTO M. Optimality of principal components, in Multivariate analysis II, Pr. Krishnaiah, ed. New York Academic Press, 1969, 673-685.

26 RAO C.R. "The use and interpretation of principal component analysis in applied research". SANKHYA, 26, A (1964) 329-358.

27 SAPORTA G. Liaison entre plusieurs ensembles de variables et codage de données qualitatives. Thèse de 3ème cycle, Paris, Université Paris VI.

28 SAPORTA G. "Pondération optimale de variables qualitatives en analyse des données". Statistique et Analyse des Données, 1979, 3, 19-32.

29 TENENHAUS M., VACHETTE J.L. PRINQUAL : un programme d'analyse en composantes principales d'un ensemble de variables nominales ou numériques. Document multigraphié, Centre d'Enseignement Supérieur des Affaires, Jouy en Josas, 1977.

ANNEXE

Proposition 1

On peut formuler de deux façons analogues le même problème de maximisation :

A. Trouver des coefficients $a_{i1} \geq 0$ et des codages X_i^{+1}, maximisant :

$$\sum_{i=1}^{p} \sum_{j=1}^{p} a_{i1}\, a_{j1}\, \mathrm{CORR}\,(X_i^{+1}, X_j^{+1})$$

sous la contrainte $\sum_{i=1}^{p} a_{i1}^2 = 1$

B. Trouver des codages W_i^{+1}, maximisant :

$$\sum_{i=1}^{p} \sum_{j=1}^{p} \mathrm{COV}\,(W_i^{+1}, W_j^{+1})$$

sous la contrainte $\sum_{i=1}^{p} \mathrm{VAR}\,(W_i^{+1}) = 1$

Démonstration

On suppose que l'on ait la solution de A. On peut prendre les codages X_i^{+1} centrés réduits. On définit $W_i^{+1} = a_{i1}\, X_i^{+1}$.

La solution trouvée maximise :

$$\sum_{i=1}^{p} \sum_{j=1}^{p} \mathrm{COV}\,(W_i^{+1}, W_j^{+1})$$

avec une contrainte qui peut s'écrire :

$$\sum_{i=1}^{p} \mathrm{VAR}\,(W_i^{+1}) = 1$$

pour des codages W_i^{+1} de la forme $a_{i1} X_i^{+1}$, où a_{i1} est positif, mais où X_i^{+1} est quelconque (centré réduit) c'est-à-dire pour l'ensemble des codages centrés possibles. Or dans B les W_i^{+} ne sont définis qu'à un centrage près.

Inversement, supposons que l'on ait la solution de B , on peut prendre les W_i^{+1} centrés.

Soit $a_{i1} = \sqrt{\mathrm{VAR}\,(W_i^{+1})}$

$$W_i^{+1} = a_{i1}\, X_i^{+1}$$

Le problème B s'écrit alors dans les termes du problème A.

Pour la suite, on adoptera la formulation B, en parlant de codages X_i^{+1} au lieu de W_i^{+1}.

Proposition 2

On cherche des codages X_i^{+1} (fonctions de moyenne nulle sur les observations) maximisant :

$$\sum_{i=1}^{p} \sum_{j=1}^{p} \text{COV}\,(X_i^{+1},\ X_j^{+1})$$

sous la contrainte :

$$\sum_{i=1}^{p} \text{VAR}\,(X_i^{+1}) = 1$$

La solution est donnée par le premier facteur de l'analyse des correspondances du tableau disjonctif complet.

Démonstration

Soit n le nombre d'observation,

m une colonne du tableau disjonctif complet T.

m représente une modalité d'une variable X_i, on note T_i le sous tableau de T, restreint aux modalités de X_i.

On note X_i^{+1} le vecteur colonne des codages associés aux modalités de X_i par le premier facteur.

U_i est un vecteur colonne de 1 dont le nombre de lignes est égal au nombre de modalités de X_i.

U est un vecteur colonne de 1 à n lignes.

L'expression matricielle du problème de maximisation est :

$$\text{Max} : \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{p} (X_i^{+1})'\, T_i'\, T_j\, X_j^{+1} \qquad (1)$$

sous les contraintes :

$$\frac{1}{n} \sum_{i=1}^{p} (X_i^{+1})'\, T_i'\, T_i\, X_i^{+1} = 1 \qquad (2)$$

et : $$\frac{1}{n}\, U_i'\, T_i'\, T_i\, X_i^{+1} = 0 \qquad (3)$$

qui s'écrit aussi $\frac{1}{n}\, U'\, T_i\, X_i^{+1} = 0$

$T_i'T_j$ est un tableau de contingence croisant les modalités de X_i et celles de X_j.

$T_i'T_i$ est une matrice diagonale dont les éléments sont les effectifs des modalités de X_i.

Négligeant les coefficients 1/n, le Lagrangien s'écrit :

$$\mathcal{L} = \sum_{i=1}^{p}\sum_{j=1}^{p} (X_i^{+1})' T_i'T_j X_j^{+1} - a\left[\sum_{i=1}^{p} (X_i^{+1})' T_i' T_i X_i^{+1}\right] - 1$$

$$- \sum_{i=1}^{p} b_i U_1' T_1' T_i X_I^{+1}$$

En dérivant successivement par rapport aux vecteurs X_i^{+1} on obtient p équations :

$$\sum_{j=1}^{p} T_i' T_j X_j^{+1} + (1 - 2a) T_i' T_i X_i^{+1} - b_i T_1' T_i U_i = O$$

Les coefficients b sont nuls ; on le voit en multipliant à gauche par U_1', et en utilisant la relation (3) (noter que $U_1' T_1' = U'$)
Reste donc :

$$\sum_{j=1}^{p} T_1'T_j X_j^{+1} = (2a - 1) T_1'T_i X_i^{+1}$$

Si l'on multiplie à gauche par (X_i^{+1}) et que l'on somme sur i, on voit que la quantité à maximiser est (2a - 1).

L'ensemble des p équations :

$$\sum_{j=1}^{p} (T_i' T_i)^{-1} T_1'T_j X_j^{+1} = (2a - 1) X_i^{+1}$$

signifie que le vecteur colonne X^{+1} obtenu en mettant à la suite tous les vecteurs X_i^{+1} est vecteur propre, par rapport à la plus grande valeur propre, du tableau obtenu en multipliant chaque ligne du tableau de Burt (tableau des $T_i'T_j$) par :

1/ (effectif de la modalité associé à la ligne)

On identifie ce vecteur X^{+1} au premier facteur (appelé classiquement φ^1) de l'analyse des correspondances du tableau T.

Les facteurs suivants donnent les fonctions-codages suivantes. Les propriétés d'orthogonalité des facteurs peuvent s'écrire en termes de relation entre fonctions-codages associées à des facteurs différents.

Tableau 1

Type de variable et Méthode	un facteur q	
	fournit	maximisant
p variables quantitatives ACP normée	3.1. p coefficients a_{iq}	$\sum_{i=1}^{p}\sum_{j=1}^{p} a_{iq}a_{jq}\mathrm{CORR}(X_i, X_j)$
	3.2. Une fonction Z_q	$\sum_{i=1}^{p}\mathrm{CORR}^2(Z_q, X_i)$
p variables qualitatives AFC sur tableau disjonctif complet	4.1. une fonction Z_q p codages X_i^{+q}	$\sum_{i=1}^{p}\mathrm{CORR}^2(Z_q, X_i^{+q})$
	4.2. p coefficients $a_{iq} \geqslant 0$ p codages X_i^{+}	$\sum_{i=1}^{p}\sum_{j=1}^{p} a_{iq}a_{jq}\mathrm{CORR}(X_i^{+q}, X_j^{+q})$ + 2 variantes : cf 4.3.
p x r variables quantitatives analyse canonique	6. p coefficients a_{iq} r coefficients b_{jq}	$\mathrm{COV}\left[\sum_{i=1}^{p} a_{iq}X_i, \sum_{j=1}^{r} b_{jq}Y_j\right]$
p x r variables qualitatives Analyse canonique sur fonctions indicatrices	7.1. p coefficients $a_{iq} \geqslant 0$ r coefficients $b_{jq} \geqslant 0$ p codages X_i^{+q} r codages Y_j^{+q}	$\mathrm{COV}\left[\sum_{i=1}^{p} a_{iq}X_i, \sum_{j=1}^{r} b_{jq}Y_j\right]$
p x r variables qualitatives AFC sur juxtaposition de tableaux de contingence	7.2. idem codages centrés-réduits	idem
p variables qualitatives x 1 variable qualitative AFC sur juxtaposition de tableaux de contingence	8.1. p codages X_i^{+q} 1 codage Y^{+q}	$\sum_{i=1}^{p}\mathrm{CORR}^2(X_i^{+q}, Y^{+q})$
1 variable qualitative x 1 variable qualitative AFC sur tableau de contingence	8.2. 1 codage X^{+q} 1 codage Y^{+q}	$\mathrm{CORR}^2(X^{+q}, Y^{+q})$

Tableau 1

sous des contraintes		références démonstration
de normalisation	d'orthogonalité	
$\sum_{i=1}^{p} a_{iq}^2 = 1$	$\sum_{i=1}^{p} a_{iq}a_{iq'} = 0$	triviale
	$CORR(Z_q, Z_{q'}) = 0$	triviale
	$CORR(Z_q, Z_{q'}) = 0$ $\sum_{i=1}^{p} CORR(X_i^{+q}, Z_q)CORR(X_i^{+q'}, Z_{q'}) CORR(X_i^{+q}, X_i^{+q'}) = 0$	(20) et (16) (22, 18)
$\sum a_{iq}^2 = 1$	$\sum_{i=1}^{p} a_{iq}a_{iq'}CORR(X_i^{+q}, X_i^{+q'}) = 0$	annexe
$VAR\left[\sum_{i=1}^{p} a_{iq}X_i\right] = 1$ $VAR\left[\sum_{j=1}^{r} b_{jq}y_j\right] = 1$ ou : somme des VAR = 2	$CORR\left[\sum_{i=1}^{p} a_{iq}X_i, \sum_{i=1}^{p} a_{iq'}X_i\right] = 0$ idem pour les (b_{jq}, y_j)	analyse canonique (11, 15)
$VAR \sum_{i=1}^{p} a_{iq}X_i^{+q} = .1$ $VAR \sum_{j=1}^{r} b_{jq}Y_j^{+q} = 1$	$CORR\left[\sum_{i=1}^{p} a_{iq}X_i^{+q}, \sum_{i=1}^{p} a_{iq}, X_i^{+q'}\right] = 0$ idem pour les (b_{jq}, Y_i^{+q})	analyse canonique
$\sum_{i=1}^{p} a_{iq}^2 = 1$ $\sum_{i=1}^{r} b_{jq}^2 = 1$	$\sum_{i=1}^{p} a_{iq}a_{iq'}$, $CORR(X_i^{+q}, X_i^{+q'}) = 0$ idem pour les (b_{jq}, Y_j^{+q})	(16)
	$CORR(Y^{+q}, Y^{+q'}) = 0$ $\sum_{i=1}^{p} CORR(X_i^{+q}, Y^{+q}) CORR(X_i^{+q}, Y^{+q'}) CORR(X_i^{+q}, X_i^{+q'}) = 0$	(18)
	$CORR(Y^{+q}, Y^{+q'}) = 0$ $CORR(X^{+q}, X^{+q'}) = 0$	(2, 7)