

A. LECLERC

P. AIACH

Mesure de l'importance des valeurs propres en analyse des données. Application à l'analyse en composantes principales de variables catégorisées

Revue de statistique appliquée, tome 26, n° 1 (1978), p. 5-21

http://www.numdam.org/item?id=RSA_1978__26_1_5_0

© Société française de statistique, 1978, tous droits réservés.

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

NUMDAM

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

MESURE DE L'IMPORTANCE DES VALEURS PROPRES EN ANALYSE DES DONNÉES. APPLICATION A L'ANALYSE EN COMPOSANTES PRINCIPALES DE VARIABLES CATÉGORISÉES

A. LECLERC,* P. AIACH**

1. INTRODUCTION

On propose ici des aides à l'interprétation des résultats d'une analyse des correspondances ou d'une analyse en composantes principales, basées sur l'appréciation de l'importance des valeurs propres. Il est habituel d'apprécier globalement les résultats d'une analyse à partir des valeurs propres associées à chaque facteur. On parle en général de pourcentage d'inertie extrait, ou de pourcentage de variance expliquée par un facteur. Ce faisant, on compare une valeur propre à la somme de celles-ci. Or il existe au moins une autre mesure globale de l'importance d'un facteur, basée sur le carré de la valeur propre, et une autre comparaison portant sur l'écart entre une valeur propre et une borne supérieure connue.

Selon la nature du tableau analysé, et le traitement auquel il est soumis, on explicite ici la signification de certaines relations entre les valeurs propres et l'on en déduit la meilleure mesure de l'importance d'un facteur. Le fait que l'on constate systématiquement, pour certains types d'analyse, des valeurs propres faibles, trouve ici une explication.

Le cas de l'analyse en composantes principales de variables catégorisées, c'est-à-dire non continues mais à modalités ordonnées, est étudié de façon particulière. Un exemple est donné, où un premier facteur peut être jugé significatif alors qu'il n'explique que 26 % de la variance totale, avec 28 variables.

Les limites des propositions faites ici sont de deux types :

- aucun test statistique n'est donné. On peut rappeler qu'en analyse des données, des tests existent pour l'analyse des correspondances d'un tableau de contingence (Lebart, 1975) et l'analyse en composantes principales de variables multinormales. L'article de Krishnaiah (1976) fait le bilan des tests basés sur l'hypothèse de normalité, qui sont assez lourds à utiliser ;
- l'objet de cette étude est limité aux valeurs propres ; dans beaucoup de cas, particulièrement en analyse des correspondances, d'autres éléments d'interprétation, tels que les contributions absolues et relatives, sont indispensables et souvent plus utiles.

(*) INSERM U 88, 91, bd de l'hôpital 75 634 Paris cedex 13

(**) INSERM Centre de Recherches du Vésinet, 44, Chemin de Ronde 78110 Le Vésinet.

2. RELATIONS ENTRE LES VALEURS PROPRES. NOTION DE QUANTITE CRITERE

Soit une matrice symétrique (p, p) définie positive de terme général $s_{jj'}$; soient $\lambda_1 \dots \lambda_p$ ses valeurs propres, que l'on suppose toutes différentes, rangées par ordre décroissant. Les λ_q sont positifs.

On appelle C_1 la trace de la matrice :

$$C_1 = \sum_{j=1}^p s_{jj}$$

et C_2 , la somme des carrés des éléments de la matrice :

$$C_2 = \sum_{j=1}^p \sum_{j'=1}^p s_{jj'}^2 = \|S\|^2$$

alors les deux relations suivantes sont vérifiées :

$$\sum_{q=1}^p \lambda_q = C_1$$

$$\sum_{q=1}^p \lambda_q^2 = C_2$$

En effet, la première relation exprime l'égalité, pour la matrice S , de la trace et de la somme des valeurs propres. La seconde exprime cette égalité pour $S^2 = S^t S$ dont la trace est C_2 et les valeurs propres λ_q^2 .

On voit que C_1 et C_2 peuvent s'écrire comme somme de termes positifs dépendant chacun d'une valeur propre λ_q . Pour toute combinaison linéaire de C_1 et C_2 il en est de même, à cette différence près que les termes de la somme peuvent être négatifs.

On appellera *quantité critère* C_1 ou C_2 ou une combinaison linéaire de C_1 et C_2 . On cherchera la quantité critère dont l'interprétation est la plus simple d'un point de vue géométrique (inertie ...) ou statistique (fonction de coefficients de corrélation entre variables ...) et l'on mesurera l'importance d'une valeur propre par la façon dont la quantité critère est expliquée : part de critère expliquée ou différence entre le critère et son approximation. Le procédé généralise les calculs habituels de pourcentages d'inertie, ou de variance, associés à une valeur propre, où la quantité critère utilisée est toujours C_1 .

3. APPLICATION A PLUSIEURS TYPES D'ANALYSE

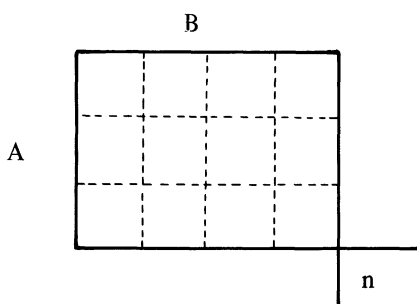
La signification des relations précédentes et le choix de la meilleure quantité critère dépend du contenu de la matrice à diagonaliser, donc du type d'analyse réalisé. On envisagera plusieurs cas, en mettant en évidence à chaque fois la relation la plus pertinente pour l'interprétation des valeurs propres.

Pour simplifier les notations, on note toujours S une matrice de terme général $s_{jj'}$, dont les valeurs propres sont $\lambda_1 \dots \lambda_p$. C'est à chaque fois la matrice diagonalisée dans le cas particulier envisagé.

ACP et AFC désignent respectivement l'Analyse en Composantes Principales et l'Analyse Factorielle des Correspondances. L'exposé des méthodes pourra se trouver dans le livre de Lebart et Fénélon (1975), et certaines démonstrations dans le R. Cehessat (1976) sous forme d'exercices.

3.1. AFC d'un tableau de contingence

Le tableau analysé est un tableau de contingence croisant les modalités d'une variable qualitative A avec les modalités d'une variable qualitative B , pour un échantillon de taille n . Chaque observation est comptée une fois et une seule dans le tableau :



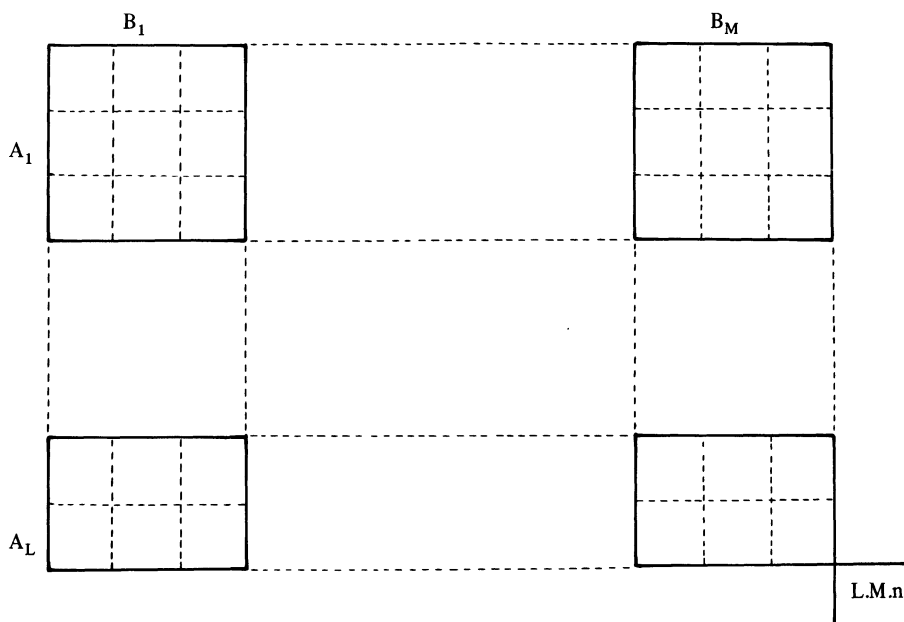
La quantité C_1 , somme des valeurs propres non triviales, est, à un facteur près, le Khi-2 calculé sur le tableau de contingence :

$$\lambda_1 + \dots + \lambda_p = \chi^2/n = \Phi^2$$

Ce Khi-2 peut être considéré comme une mesure de l'écart entre le tableau observé et le tableau au hasard ; le rapport d'une valeur propre à la somme de celles-ci note la part apportée par un facteur dans l'explication de cet écart. *Le critère C_1 et les mesures qui lui sont liées ont dans ce cas une bonne interprétation.*

3.2. AFC d'une juxtaposition de tableaux de contingence

Le tableau analysé croise les modalités de plusieurs variables qualitatives $A_1 \dots A_L$ avec les modalités de plusieurs autres variables qualitatives $B_1 \dots B_M$, pour un échantillon de taille n . Chaque observation est comptée une fois dans chaque tableau de contingence $A_1 \times B_m$, et au total $L.M.$ fois.



Cette analyse répond au problème suivant : décrire les relations entre deux groupes de variables qualitatives, constituées de variables de nature différente (caractéristiques socio-économiques de ménages, et réponses données à un questionnaire, par exemple).

Dans ce cas (Leclerc, 1975, 1976) la quantité C_1 est, à un facteur près, la somme des Khi-2 calculés sur les tableaux de contingence.

$$\lambda_1 + \dots + \lambda_p = \frac{1}{L.M.n} \sum_{l=1}^L \sum_{m=1}^M \chi_{l,m}^2 = \frac{1}{L.M} \sum \sum \Phi_{1,m}^2$$

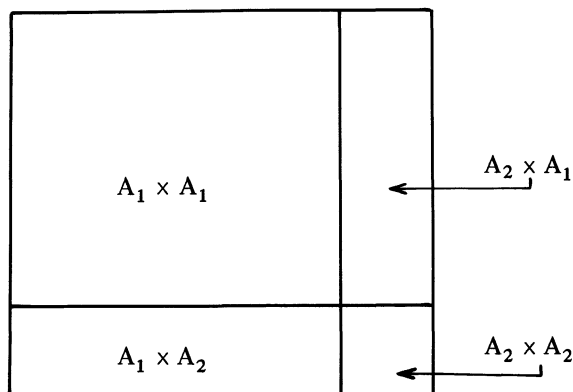
où $\chi_{l,m}^2$ est le Khi-2 calculé sur le tableau de contingence $A_l \times B_m$.

Le rapport d'une valeur propre à la quantité C_1 note la part apportée par le facteur dans l'explication de la somme des Khi-2, mesure globale de l'écart entre le tableau observé et le tableau tel qu'il serait si les variables A_l et B_m étaient indépendantes. Ici encore, *les mesures associées à C_1 ont une bonne signification.*

3.3. Cas de deux variables

Dans le cas de deux variables A_1 et A_2 , l'analyse des relations entre A_1 et A_2 peut se faire par l'AFC du tableau de contingence $A_1 \times A_2$, mais beaucoup d'autres approches sont possibles et fournissent des valeurs propres différentes (Cazes, 1977). L'AFC d'un tableau sous forme disjonctive complète (cf. 3.5.) est une de ces approches, ainsi que l'AFC d'un tableau de Burt, qui est détaillée ici :

On considère le tableau obtenu en croisant les variables qualitatives A_1 et A_2 avec elles-mêmes. Ce tableau est symétrique, et comporte quatre sous-tableaux : $A_1 \times A_1$ et $A_2 \times A_2$ contiennent les effectifs des modalités sur la diagonale, et des 0 ailleurs, $A_1 \times A_2$ est un tableau de contingence.



Si les plus grandes valeurs propres dans l'analyse du tableau de contingence $A_1 \times A_2$ sont $\lambda_1 \dots \lambda_q$, les plus grandes valeurs propres dans l'analyse du tableau de Burt sont (Lebart, 1977)

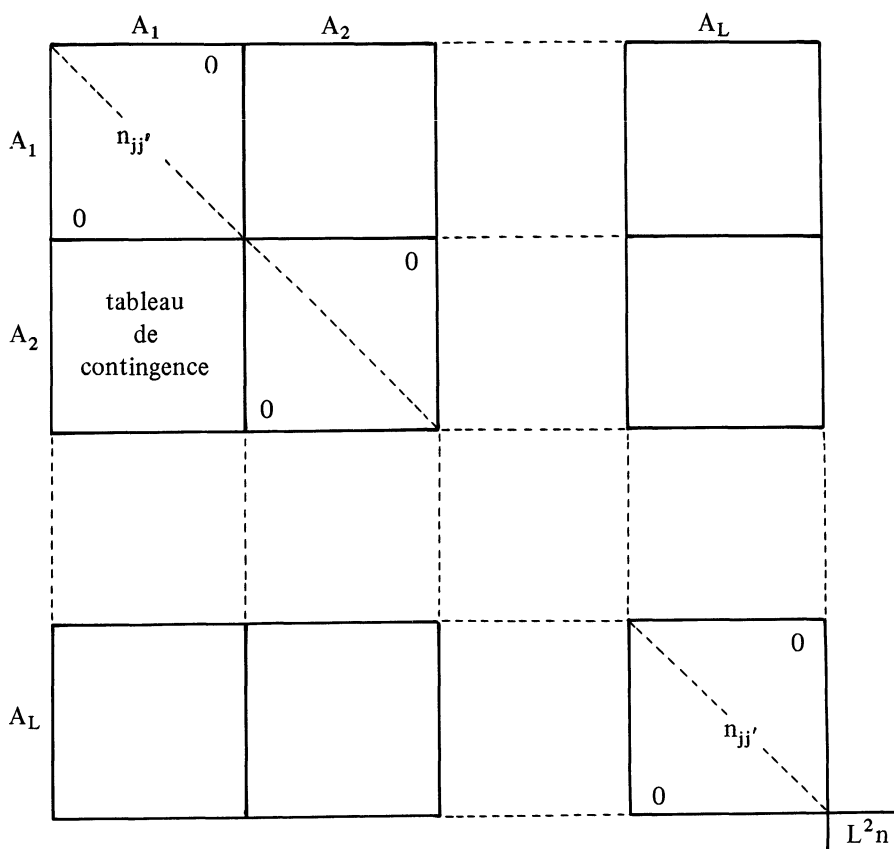
$$\mu_q = \frac{(1 + \sqrt{\lambda_q})^2}{4}$$

Si l'on mesure l'importance des valeurs propres par le critère C_1 , l'analyse du tableau de Burt donnera une estimation pessimiste par rapport à l'analyse du tableau de contingence.

3.4. AFC d'un tableau de Burt : cas général

Si l'on cherche à décrire toutes les relations entre L variables qualitatives $A_1 \dots A_L$, la méthode précédente (tableau de Burt) peut être utilisée, sans comparaison possible dans ce cas à un tableau de contingence.

Le tableau de Burt aura la forme suivante :



L'inertie totale, pour l'analyse d'un tel tableau, se calcule comme dans le cas d'une juxtaposition de tableaux de contingence :

$$C_1 = \sum_{l=1}^L \sum_{l'=1}^L \chi_{ll'}^2 / L^2 n$$

mais les quantités $\chi_{ll'}^2$ ont une forme particulière, d'où l'expression de C_1 :

$$C_1 = (nbm - L) / L^2 + \sum_{l=1}^L \sum_{l' \neq l} \chi_{ll'}^2 / L^2 n$$

où les $\chi_{ll'}^2$ sont des Khi-2 calculés sur les tableaux de contingence $A_l \times A_{l'}$, et où nbm est le nombre total de colonnes du tableau.

A défaut de disposer d'une mesure de l'importance des facteurs basée sur la part expliquée d'une somme de Khi-2, comme pour les tableaux de contingence, on peut admettre comme mesure la comparaison des valeurs propres à C_1 .

On voit de plus que dans l'expression de C_1 le second terme, somme de Khi-2 qui mesure globalement l'association entre les variables, sera prépondérant si le nombre de variables L est élevé.

3.5. AFC d'un tableau sous forme disjonctive complète

Soit un tableau croisant l'ensemble des observations, en ligne, avec les modalités des variables qualitatives $A_1 \dots A_L$ en colonne. A l'intersection d'une ligne et d'une colonne on trouve 1 si l'observation possède la modalité, 0 sinon.

	A ₁			A ₂			A _L			
Observations	0	1	0	0	0	0	1	1	0	0

L'analyse de ce tableau répond au même problème que l'analyse du tableau de Burt, la seule différence entre les résultats des analyses des deux tableaux, à partir des mêmes données, est que les valeurs propres sont ici $\sqrt{\lambda_1} \dots \sqrt{\lambda_p}$, alors qu'elles sont $\lambda_1 \dots \lambda_p$ si l'on analyse le tableau de Burt.

La relation entre valeurs propres dans l'analyse du tableau de Burt devient une relation entre carrés de valeurs propres pour un tableau sous forme disjonctive complète. Dans ce dernier cas, *une mesure de l'importance des facteurs à partir du critère C₂ peut donc être retenue.*

Elle donne des facteurs une vue plus optimiste qu'une mesure basée sur C₁.

Un examen de la quantité-critère C₁ montre d'autre part que son intérêt est limité, du moins pour ce qui concerne les variables :

$$C_1 = (nbm - L)/L$$

C₁ n'a aucune signification particulière comme mesure de l'association entre les variables. Son interprétation géométrique est l'inertie totale du nuage des points-observations. Or ces points, à quelques transformations près, ont des coordonnées 0 ou 1 ; ils sont placés aux sommets d'un hypercube formant un nuage auquel la notion d'inertie, de direction d'allongement maximum... s'applique mal. Un nuage ainsi construit ne peut pas être résumé par un sous espace de petite dimension, d'où des pourcentages d'inertie faibles.

Si l'on cherche néanmoins à mesurer la qualité de la représentation du nuage des points-observations, alors C₁ peut être utilisé.

On peut penser ici à une quantité-critère autre que C₁ ou C₂. En effet :

$$C_2 - C_1/L = \sum_1 \sum_{l \neq 1} \chi_{ll}^2$$

Cette somme de Khi-2 ne peut pas se mettre sous la forme d'une somme de termes positifs associés aux facteurs, ce qui limite son intérêt.

3.6. Analyse en composantes principales

La quantité C_1 est la somme des variances, ou le nombre de variables si l'analyse est effectuée sur des variables centrées et réduites.

La quantité C_2 est la somme des carrés des covariances, ou la somme des carrés des coefficients de corrélation empirique $r_{jj'}$.

Dans le cas d'une analyse normée, on a :

$$C_2 - C_1 = \sum_{j=1}^p \sum_{j' \neq j} r_{jj'}^2$$

C_1 est traditionnellement la quantité critère la plus utilisée.

Ceci est justifié dans le cas où le nuage des points-observations est approximativement un ellipsoïde, pour lequel on peut parler de directions d'allongement, prenant en compte un certain pourcentage de la variance totale (ou de l'inertie). Il faut pour cela que les valeurs possibles pour les coordonnées appartiennent à un éventail assez large de valeurs réelles, donc que les variables soient continues.

Si les variables ne prennent qu'un nombre limité de valeurs, on va voir qu'alors le critère C_2 peut être préféré.

3.7. ACP normée de variables catégorisées

On peut envisager d'appliquer l'analyse en composantes principales normée à des variables non continues, mais dont les modalités sont ordonnées, telles que des notes entières, par exemple : accord ou désaccord avec une affirmation ; ou note 0 la réponse "pas d'accord du tout" . . . 5 la réponse "tout-à-fait d'accord".

Cette pratique a plusieurs justifications :

- la distance entre deux observations est satisfaisante, et on en obtiendra la meilleure approximation (Pearson, 1901) au sens suivant : la somme étendue à tous les couples de points, des carrés des différences entre distances "réelles" et distances approchées par l'analyse, est minimum ;
- Beaucoup de propriétés statistiques des résultats d'ACP ne supposent pas la normalité des variables (Gill, 1976 ; Pousse, 1976).

Dans la pratique, la méthode donne de bons résultats (Aïach, 1975 ; Davidson, 1973) ; elle est peu coûteuse car la matrice à diagonaliser n'est pas de trop grande dimension. Peu d'autres méthodes sont utilisables ; parmi celles-ci il faut noter la méthode "Prinqual" (Tenenhaus, 1977) qui procède à un recodage optimum des modalités. Sur l'exemple exposé dans le paragraphe 4. Prinqual a été testé, et n'a pas donné de résultats fondamentalement différents de ceux de l'ACP (Macquin, 1977).

On examine donc ici, dans le cas de l'ACP, la signification des quantités critères C_1 et C_2 , ce qui apporte certaines explications sur les résultats constatés et permet de conclure à l'intérêt de comparaisons à la quantité-critère C_2 .

a. *Comparaison des valeurs propres à C_1*

On constate dans la pratique de faibles pourcentages de variance expliqués. On a déjà dit que ces pourcentages, dans le cas de variables à peu de modalités, avaient peu de sens du point de vue géométrique.

D'un point de vue statistique, le fait que les variables sont catégorisées apporte une explication ; en effet les valeurs propres sont astreintes à satisfaire à la fois :

$$\begin{aligned}\sum \lambda_q &= C_1 = p \\ \sum \lambda_q^2 &= C_2 = \sum_{j=1}^p \sum_{j'=1}^p r_{jj'}^2, \\ \lambda_q &\geq 0\end{aligned}$$

Le lieu des solutions possibles $(\lambda_1 \dots \lambda_p)$ est l'intersection d'un hyperplan, et d'une hypersphère de rayon $\sqrt{C_2}$. C'est une hypersphère centrée sur la première bissectrice, et dont le rayon diminue avec C_2 . Si C_2 est petit, les solutions sont proches de : $\lambda_q = 1, \forall q$, et les parts d'inertie associées aux premières valeurs propres sont faibles.

C'est ce qui se passe si les variables sont catégorisées ; les $r_{jj'}$, à degré d'association égal entre les variables j et j' , sont d'autant plus petits que les variables ont peu de catégories. Ceci a été étudié par Agresti (1976) et par Doreian (1972) qui ont décrit la façon dont varie un coefficient de corrélation entre deux variables, selon que celles-ci sont continues, de distribution numultinormale, ou transformées par "découpage" en variables catégorisées, avec un nombre plus ou moins grand de modalités.

Les parts d'inertie associées aux premières valeurs propres seront d'autant plus faibles que les variables ont peu de modalités, à nombre de variables identique, et un jugement basé sur C_1 reflète en partie le nombre de modalités des variables.

b. *Signification de C_2*

$$C_2 \text{ s'écrit : } \quad C_2 = P + \sum_{j \neq j'} r_{jj'}^2$$

D'après les études citées sur les coefficients de corrélation entre variables catégorisées, ces coefficients sont des mesures satisfaisantes d'association si la répartition des observations selon les modalités ne diffère pas trop d'une variable à une autre. Si toutes les variables n'ont pas un même nombre de modalités, celles qui en ont le plus auront une contribution plus importante.

c. *Utilisation d'autres critères*

Le critère $C_2 - C_1$ a une forme simple :

$$\sum_{j \neq j'} r_{jj'}^2$$

Il a l'inconvénient de s'écrire comme une somme de contributions $\lambda_q^2 - \lambda_q$ positives pour les premiers facteurs, puis négatives. Il paraît donc peu utilisable.

Un autre calcul peut être effectué ; c'est celui de la comparaison de λ_1 à la borne supérieure B :

$$B = 1 + \left(\frac{p-1}{p} \sum_j \sum_{j' \neq j} r_{jj'}^2 \right)^{1/2}$$

Cette borne est atteinte si et seulement si tous les $r_{jj'}$ sont égaux et approchée si les $r_{jj'}$ sont presque égaux (Leclerc, 1977)

d. Pratique de l'ACP sur variables catégorisées

Il semble préférable d'avoir le même nombre de modalités pour toutes les variables, et d'éviter des répartitions trop déséquilibrées entre modalités ; moyennant quoi, *l'analyse en composantes principales peut être appliquée, et ses résultats jugés selon le critère C_2 . On pourra aussi comparer λ_1 à la borne B.* L'exemple donné en 4 reprend ces calculs.

3.8. Cas de variables dichotomiques

Pour étudier les relations entre plusieurs variables dichotomiques, deux procédés parmi d'autres sont possibles :

- appliquer l'AFC au tableau observation x variables dédoublées ;
- appliquer l'ACP normée au tableau non dédoublé.

Les deux méthodes sont équivalentes, les valeurs propres sont les mêmes, à un facteur multiplicatif près. Si l'on procède à une ACP la quantité C_2 a la valeur :

$$C_2 = P + \sum_j \sum_{j' \neq j} \chi_{jj'}^2 / n$$

où n est le nombre d'observations et $\chi_{jj'}^2$ le Khi-2 du tableau de contingence à 4 cases croisant les modalités de j et j'.

On peut comparer les λ_q^2 à C_2 , et interpréter le résultat comme la part prise par un facteur dans l'explication d'un indice d'association basé sur des khi-2 ; on peut aussi comparer λ_1 à sa borne supérieure, qui s'écrit :

$$1 + \left[\frac{p-1}{p} \sum_j \sum_{j' \neq j} \chi_{jj'}^2 / n \right]^{1/2}$$

4. UN EXEMPLE

4.1. Les données et les résultats de l'ACP

Une étude sur les processus cumulatifs d'inégalités sociales a été réalisée à partir d'entretiens semi-directifs auprès de 134 ménages avec enfants dans un

quartier de Paris (Aiach, 1975). Une partie de l'étude a utilisé les résultats d'une ACP normée sur 28 variables, notes en 1, 2, 3 (1, 2 pour quelques-unes) données par le sociologue responsable du travail, au vu des réponses fournies par la famille. Ces notes concernent divers aspects des conditions de vie (revenu, travail, logement, relations, activités extérieures, vacances, . . .). La note 1 a été attribuée aux familles dont les conditions de vie sont jugées les plus défavorables pour le critère, la note 3 à celles qui sont dans les conditions les plus favorables, la note 2 étant une note moyenne. Quelques notes, comme le revenu, se basent sur une quantité. La plupart ordonnent les modalités d'une variable qualitative. Par exemple (variable 25) : 1 = pas de possibilité pour la famille de disposer d'une maison pour les vacances ; 2 = possibilité, chez des amis ou dans la famille ; 3 = le ménage possède une résidence secondaire.

D'autres caractéristiques des ménages (catégorie socio-professionnelle du chef de ménage, nombre d'enfants . . .) ne sont pas intervenues dans l'analyse mais ont servi comme identificateurs des ménages.

On trouvera en annexe la liste des variables, la matrice de corrélation, les corrélations variables-facteurs et le graphique des ménages sur le plan des deux premiers axes factoriels, chaque ménage ayant comme "identité" la CSP du chef de ménage.

Le but de l'analyse était de dégager les variables les plus caractéristiques d'une ou plusieurs dimensions d'inégalité sociale.

Le premier facteur qui n'explique que 26 % de la variance, a été interprété comme niveau socio-économique : si l'on place les ménages le long du premier axe factoriel, et que l'on regarde pour chacun d'eux quelle est la CSP du chef de ménage, on retrouve la typologie des CSP (variable qui n'était pas intervenue dans l'analyse) selon un ordre classique : cadres supérieurs, cadres moyens, . . . , manoeuvres .

Les variables les plus significatives (pour cette population) du niveau socio-économique sont le revenu (V_1 , V_{27}), les loisirs extérieurs (V_{14} , V_{22}) et la voiture (V_{26}). Toutes les variables sont corrélées positivement au premier facteur.

Les facteurs suivants n'ont trouvé aucune interprétation certaine, bien que le second facteur soit un peu lié à la nationalité du chef de ménage (Macquin, 1977).

Des modifications de codage minimales (regroupement des notes 2 avec les notes 1, puis avec les 3) ont été essayés, et ont très peu modifié les résultats.

Les inégalités sociales, dans cette population, seraient donc sous-tendues par une seule variable résumante, indice global de niveau socio-économique. peut-on l'affirmer en s'appuyant sur des mesures plus objectives de l'importance du premier facteur ? On va examiner successivement les mesures basées sur C_1 , C_2 , $C_2 - C_1$ et B.

4.2. Mesures de l'importance du premier facteur

a. Mesures basées sur C_1

Les pourcentages de variance expliquée sont données par le tableau suivant :

	valeurs propres	% variance	% cumulé
1	7.32	26.1	26.1
2	1.82	6.5	32.6
3	1.55	5.5	38.2
4	1.48	5.3	43.5
5	1.41	5.0	48.5
6	1.30	4.6	53.2
7	1.20	4.2	57.5
8	1.12	4.0	61.5
9	1.02	3.6	65.2
10	0.92	3.2	68.5
11	0.90	3.2	71.7
12	0.77	2.7	74.5
13	0.74	2.6	77.1
14	0.72	2.5	79.7
15	0.67	2.3	82.1
16	0.59	2.1	84.2
17	0.57	2.0	86.3
18	0.55	1.9	88.2
19	0.44	1.6	89.8
20	0.43	1.5	91.4
21	0.42	1.5	92.9
22	0.38	1.3	94.3
23	0.37	1.3	95.6
24	0.32	1.1	96.8
25	0.30	1.0	97.9
26	0.23	0.8	98.7
27	0.19	0.7	99.5
28	0.14	0.5	100.0

Le pourcentage associé au premier facteur n'est pas élevé. Il y a cependant décroissance nette après celui-ci.

b. Mesures basées sur C_2

C_2 , somme des carrés des valeurs propres, vaut 74.56. On peut comparer à C_2 les carrés des premières valeurs propres :

Facteur	λ_q^2	Pourcentage de C_2
1	53.58	72
2	3.33	4
3	2.40	3
3	2.22	3

Selon ce critère, le premier facteur apparaît comme un bon résumé des relation entre les variables. L'écart entre le premier facteur et les autres est encore plus manifeste.

c. Utilisation de $C_2 - C_1$

$C_2 - C_1$, somme des carrés des coefficients de corrélations empiriques, vaut 46.56. C'est aussi la somme des $\lambda_q^2 - \lambda_q$. Ces quantités sont positives si $\lambda_q \geq 1$, négatives ensuite, atteignant leur minimum pour la valeur propre la plus proche de 0.5 :

Facteur	$\lambda_q^2 - \lambda_q$
1	46.26
2	1.51
.	.
.	.
9	0.03
10	- 0.07
.	.
.	.
18	- 0.24
.	.
.	.
28	- 0.12

La quantité $C_2 - C_1$ est bien approchée par le facteur 1 seul. Pour en tirer des conclusions sur le rôle particulier du premier facteur, il faudrait en savoir plus sur la répartition de $C_2 - C_1$ entre les facteurs sous certaines hypothèses.

d. Comparaison de λ_1 à B

La borne B vaut ici 8.63 ; elle est approchée de près par λ_1 . On sait que B est atteinte si et seulement si tous les coefficients de corrélation sont égaux, donc si toutes les variables sont équivalentes dans la recherche d'un facteur résumant qui est alors une somme pondérée des variables. Dans ce cas, les valeurs propres autres que la première sont égales, et d'autant plus grandes que le coefficient de corrélation commun est petit. On peut penser qu'ici on approche cette situation, bien que tous les coefficients de corrélation empiriques sont loins d'être égaux.

6. CONCLUSION ET DISCUSSION

Les résultats rassemblés ici peuvent se résumer ainsi : la trace de la matrice à diagonaliser est une bonne mesure de l'importance d'un facteur dans les cas où l'on applique l'AFC à un tableau de contingence, à une juxtaposition de tableaux de contingence, à un tableau de Burt, ou l'ACP à des variables continues.

Par contre, la comparaison des valeurs propres à la somme des carrés des éléments de la matrice à diagonaliser a une meilleure signification dans l'analyse d'un tableau sous forme disjonctive complète, et dans l'ACP de variables catégorisées, du moins si l'on s'intéresse plus aux variables (et à obtenir de bons résumés de celles-ci) qu'à des observations à représenter au mieux. Dans le cas de l'ACP sur variables catégorisées, la comparaison de la plus grande valeur propre à une borne supérieure atteinte dans le cas d'égalité de coefficients de corrélation peut aussi être effectuée.

Ces résultats veulent être une incitation à ne pas utiliser des critères de qualités des facteurs sans tenir compte de la nature du tableau analysé, et à différencier les critères selon ce qu'ils permettent de mesurer. Ils sont encore très limités et devraient être améliorés par la recherche de tests statistiques, même approchés, sur la signification des valeurs propres.

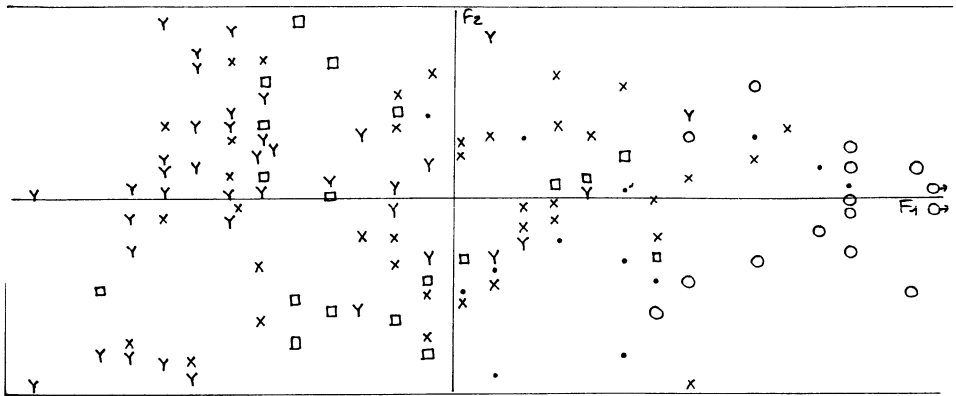
LISTE DES VARIABLES

- 1 Revenu par Unité de Consommation
- 2 Statut d'occupation du logement
- 3 Confort du logement
- 4 Surface par habitant dans le logement
- 5 Temps de transport
- 6 Possibilité de promotion dans le travail
- 7 Conditions de travail
- 8 Age de fin d'études du chef de ménage
- 9 Diplôme du chef de ménage
- 10 Fréquence de rencontre des parents
- 11 Aide familiale
- 12 Bien reçus au moment du mariage
- 13 Retards scolaires des enfants
- 14 Activités extérieures des enfants
- 15 Maladies des enfants
- 16 Vacances des enfants
- 17 Santé des parents
- 18 Rencontre d'amis ou de voisins
- 19 Capital relations
- 20 Loisirs à la maison
- 21 Lecture
- 22 Loisirs extérieurs
- 23 Sorties en week end
- 24 Vacances des parents
- 25 Résidence de vacances
- 26 Voiture
- 27 Salaire du chef de ménage
- 28 Dépenses remises

MATRICE DE CORRELATION

	..1	..2	..3	..4	..5	..6	..7	..8	..9	..10	..11	..12	..13	..14	..15	..16	..17	..18	..19	..20	..21	..22	..23	..24	..25	..26	..27	..28	
1	100																												
2	11	100																											
3	42	21	100																										
4	39	22	53	100																									
5	-4	-6	00	11	100																								
6	17	-3	9	-6	-3	100																							
7	26	8	16	21	13	28	100																						
8	35	-1	31	35	14	28	27	100																					
9	28	12	37	42	11	15	6	61	100																				
10	32	-5	31	31	22	5	22	30	11	100																			
11	22	1	18	10	-4	22	3	25	21	15	100																		
12	13	7	29	15	-4	20	13	30	27	27	26	100																	
13	30	11	30	27	-1	22	33	34	31	6	15	22	100																
14	48	8	44	41	9	16	25	31	27	34	19	22	16	100															
15	22	-3	00	-8	-17	19	4	3	-8	13	5	21	12	18	100														
16	39	1	35	22	-5	8	22	25	21	9	23	17	21	41	18	100													
17	33	7	21	13	18	13	31	21	16	12	6	25	24	27	19	100													
18	00	-7	5	-8	-4	6	12	1	2	-8	6	-8	-14	3	-3	-6	100												
19	38	7	35	32	5	7	17	47	43	27	15	21	25	34	4	21	13	25	100										
20	27	3	13	15	-5	22	9	21	19	13	34	14	13	22	3	32	25	12	24	100									
21	30	1	27	31	12	14	6	33	28	24	14	14	15	28	5	15	16	4	31	13	100								
22	50	1	38	39	8	17	19	32	36	44	25	26	19	50	10	33	25	11	47	31	27	100							
23	21	4	19	17	-2	5	15	16	19	17	7	15	12	21	9	19	7	1	18	20	5	16	100						
24	33	-3	40	27	-2	13	17	26	30	7	23	20	20	40	9	69	23	8	22	31	11	26	17	100					
25	26	12	26	27	-2	14	21	24	33	3	16	5	24	17	1	32	7	14	24	24	3	23	26	35	100				
26	49	7	37	34	8	17	33	30	37	38	10	19	17	48	9	35	27	-5	39	30	18	43	48	32	22	100			
27	78	11	49	37	-4	25	34	34	23	34	22	9	35	51	14	44	28	6	35	30	25	47	25	41	33	55	100		
28	30	20	27	16	21	18	35	22	25	25	5	18	27	36	7	31	12	3	33	25	18	29	31	31	26	41	45	100	

REPRESENTATION DES MENAGES DANS LE PLAN
DES DEUX PREMIERS AXES FACTORIELS



○ cadres supérieurs
• cadres moyens
□ employés
x ouvrier qualifié ou contremaître
∇ ouvrier spécialisé ou manoeuvre

CORRELATIONS VARIABLES-FACTEURS
(en 0/00)

	1 # F	2 # F
1	723	80
2	153	- 117
3	650	- 161
4	582	- 406
5	81	- 522
6	317	344
7	429	- 40
8	620	- 144
9	580	- 159
10	469	- 455
11	364	282
12	385	- 51
13	473	140
14	677	- 79
15	175	539
16	584	360
17	416	217
18	43	130
19	599	- 178
20	457	312
21	421	- 249
22	674	- 119
23	394	39
24	578	325
25	456	188
26	688	- 82
27	762	119
28	560	- 43

REFERENCES

- [1] AGRESTI A. (1976). – The effect of category choice on some ordinal measures of association. *JASA* 71, n° 353.
- [2] AIACH P. (1975). – Vivre à Folie-méricourt ; étude des processus cumulatifs d'inégalités. Rapport CORDES.
- [3] CAZES P., BAUMERDER A., BONNEFOUS S., PAGES J.P. (1977). – Codage et analyse des tableaux logiques. Introduction à la pratique des variables qualitatives. *Cahiers du BURRO*, n° 27, 3-47.
- [4] CEHESAT R., (1976). – Exercices commentés de statistique et informatique appliquées. Dunod, 418 p.
- [5] DAVIDSON F., CHOQUET M., DEPAGNE M., (1973). – Les lycéens devant la drogue et les autres produits psychotropes. Monographie INSERM, 207 p.
- [6] DOREIAN P. (1972). Multivariate analysis and categorized data. *Quality and quantity* VI, n° 2.
- [7] GILL R.D. (1976). – Consistency of maximum likelihood estimators of the factor analysis model, when the observations are not multivariate normally distributed. Communication au Congrès Européen des Statisticiens, Grenoble, septembre 1976.
- [8] KRISHNAIAH P.R. (1976). – Somme recent developments on complex multivariate distribution. *J. of Mult. Analysis*, 6. 1-30.
- [9] LEBART L., FENELON J.P. (1975). – Statistique et informatique appliquées. Dunod, 457 p.
- [10] LEBART L. (1975). – Validité des résultats en analyse des données. Rapport DGRST, 158 p.
- [11] LEBART L. MORINEAU A., TABARD N. (1977). – Techniques de la description statistique. Dunod, 352 p.
- [12] LECLERC A. (1975). – L'analyse des correspondances sur juxtaposition de tableaux de contingence. *Revue de Statistique Appliquée*, 23, n° 3, 5-16.
- [13] LECLERC A. (1976). – Une étude de la relation entre une variable qualitative et un groupe de variables qualitatives. *Int. Stat. Rev.*, 44 n° 2, 241-248.
- [14] LECLERC A. – Une borne supérieure pour les valeurs propres d'une matrice symétrique et son application à l'analyse en Composante Principales. (à paraître).
- [15] MACQUIN A. (1977). – La méthode Prinqual. Présentation et comparaison avec l'Analyse des Correspondances, et avec l'Analyse en Composantes Principales. Rapport d'Application, Cycle de Statistique Appliquée, 50 p.
- [16] PEARSON (1901). – On lines and plans of points in space. *Phil. Mag.* 2.
- [17] POUSSE A. (1976). – Analyse en composantes principales de probabilités et échantillonnage. Communication au Congrès Européen des Statisticiens. Grenoble, Septembre 1976.
- [18] TENENHAUS M. (1977). – Analyse en composantes principales d'un ensemble de variables nominales ou numérique, note n° D 345 n 77 C.E.S.A.