Statistique et analyse des données

P. Cazes
S. Bonnefous
A. Baumerder
J. P. Pages

Description cohérente des variables qualitatives prises globalement et de leurs modalités

Statistique et analyse des données, tome 1, n° 2 (1976), p. 48-62
<http://www.numdam.org/item?id=SAD_1976__1_2_48_0>

DESCRIPTION COHERENTE des VARIABLES QUALITATIVES PRISES GLOBALEMENT et de LEURS MODALITES

par
P. CAZES[1], S. BONNEFOUS, A. BAUMERDER et J.P. PAGES[2]

Une analyse en composantes principales effectuée sur l'ensemble des opérateurs associés à k variables qualitatives fournit une description claire des liaisons (au sens du chi-deux) entre ces variables.

Exploitant une idée d' Yves ESCOUFIER, on montre comment on peut utiliser les résultats obtenus dans cette première analyse pour décrire au mieux simultanément les modalités des différentes variables et les individus considérés.

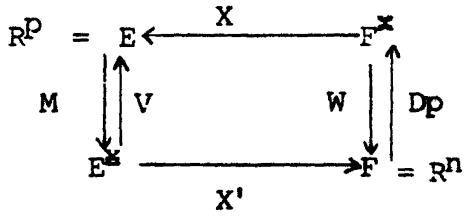
(1) Laboratoire de Statistique (PARIS VI)

(2) Laboratoire de Statistiques et d'Etudes Economiques et Sociales-
Departement de protection - (CEA)

1. NOTATIONS .

Au tableau "individus x caractères" X sont associées les métriques M et N qui permettent respectivement de mesurer les proximités entre les n colonnes (individus) et entre les p lignes (caractères) du tableau; la métrique N n'est autre en général que la métrique des poids Dp (distance en moyenne quadratique).

Au triplet (X, M, Dp) correspond alors le schéma de dualité (1) :



avec $V = X \circ Dp \circ X'$ $W = X' \circ M \circ X$

Si X est centré la forme quadratique associée V n'est autre que la forme quadratique d'inertie qui traduit la dispersion du nuage des individus dans E autour de son centre de gravité.

L'opérateur $W \circ Dp$ est l'opérateur qu' Y. ESCOUFIER (2) (3) (4) a associé au triplet (X, M, Dp); cet opérateur peut être considéré comme un vecteur du sous espace G des endomorphismes Dp-symétriques définis sur F.

L'espace G a été muni par Y. ESCOUFIER du produit scalaire P défini matriciellement à partir de la trace (tr) de la façon suivante :

$$U_1 \in G \qquad U_2 \in G \qquad P(U_1, U_2) = tr\ (U_1\ U_2)$$

Si les valeurs propres de l'opérateur U sont notées λ_i on a alors :

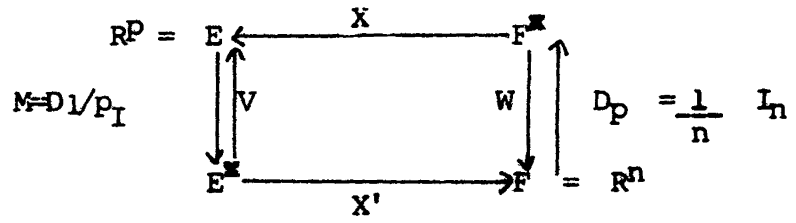
$$\| U \|^2 = P(U,U) = \sum_{i=1}^{m} \lambda_i^2$$

2. OPERATEURS ASSOCIES à des VARIABLES QUALITATIVES .

La donnée d'une variable qualitative x, dont I désigne l'ensemble des p modalités, est équivalente à la donnée du sous espace de F engendré par les variables indicatrices associées à ces modalités (1). Ce sous - espace

contient la droite des constantes Δ_j engendrée par le vecteur $\underline{j}$ dont toutes les coordonnées sont égales à 1 .

A x correspond alors le schéma de dualité :



X est ici le tableau logique (p x n) des indicatrices (rangées en lignes) des modalités de x ; les poids associés aux n individus ont tous été pris égaux à $\frac{1}{n}$; la forme quadratique V induite par D_p admet pour matrice la matrice diagonale D_{p_I} dont les éléments diagonaux ne sont autres que les probabilités p_i de prendre les différentes modalités de x (ces probabilités définissent sur (I, $\mathcal{P}(I)$) la loi de probabilité notée p_I) ; restreinte au simplexe des lois de probabilité sur (I, $\mathcal{P}(I)$), la métrique $M = D_{1/p_I}$ (équivalente à celle de Mahalanobis dans le cas du quantitatif), induite par D_{p_I} sur E, n'est autre que la métrique du chi-deux de centre p_I .

La donnée de la variable qualitative x est équivalente à la donnée de l'opérateur $A = W \circ D_p$ associé au triplet (X, D_{1/p_I}, D_p);
A n'est autre que le D_p-projecteur correspondant au sous espace engendré par les variables indicatrices. Se plaçant orthogonalement à la droite des constantes Δ_j , on préfèrera associer à x, plutôt que A, le D_p-projecteur $B = A - A_{\Delta_j}$ où A_{Δ_j} désigne le D_p-projecteur sur la droite Δ_j (5).

Remarque : B est comme A de la forme " $W \circ D_p$"; matriciellement :

$$B = (X' MX - \underline{j}\,\underline{j}') D_p$$

Le produit scalaire au sens de P entre les opérateurs $B_1 = A_1 - A_{\Delta_j}$ et $B_2 = A_2 - A_{\Delta_j}$, associés respectivement aux variables qualitatives x (à p modalités) et y (à q modalités) n'est autre alors que le "phi-deux" de Yule :

$$P(B_1, B_2) = \frac{\lambda^2}{n} = \phi^2$$

. /..

On a de plus :

$$\|B_1\|^2 = P(B_1, B_1) = p - 1 ; \|B_2\|^2 = P(B_2, B_2) = q - 1$$

Pour s'affranchir en partie des degrés de liberté on travaillera sur les opérateurs normés ; le produit scalaire entre ces opérateurs normés n'est autre alors que le cosinus de l'angle entre les opérateurs :

$$\text{Cos}(B_1, B_2) = \frac{\phi^2}{\sqrt{(p-1)(q-1)}} = T^2$$

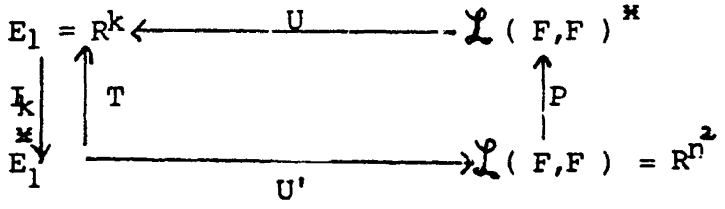
où T est l'indice qu'avait introduit Tschuprow (5) (6) .

3. DESCRIPTION GLOBALE d'un ENSEMBLE de k VARIABLES QUALITATIVES .

Aux k variables qualitatives x_i à p_i modalités sont associés les opérateurs normés $B_i = \frac{A_i - A_{\Delta_i}}{\sqrt{p_i - 1}}$

Assimilant ces opérateurs à des "caractères" on décrit au mieux les angles entre ces opérateurs en effectuant une analyse en composantes principales.

Le schéma de dualité considéré est alors le suivant :



. U de dimension (k x n^2) est le tableau des opérateurs normés rangés en lignes

. l'application P de $\mathcal{L}(F,F)$ dans son dual symbolise ici le produit scalaire défini précédemment sur G ; en toute rigueur ce produit scalaire aurait dû être représenté par un isomorphisme du sous espace des opérateurs D_p-symétriques G (dans ce sous espace les coordonnées de l'opérateur sont les $\frac{n(n+1)}{2}$ éléments sub-diagonaux de la matrice associée) dans sons dual G^*, isomorphisme qui n'est autre que la restriction de P à G.

. on a noté T la matrice des produits scalaires entre les opérateurs normés (matrice des "T^2" de Tschuprow) ; cette matrice, qui est associée à une application de E_1^* dans son dual, est analogue à une matrice de corrélation .

. /..

Les composantes principales forment un système P orthogonal dans G ; on représentera, comme on le fait pour les caractères en analyse en composantes principales, les opérateurs initiaux dans ce système .
La matrice des produits scalaires entre les opérateurs "caractères" (en lignes) et les composantes principales normées (en colonnes) est la matrice $R = VD_{\sqrt{\lambda}}$ où V est la matrice des vecteurs propres normés de T rangés en colonnes et $D_{\sqrt{\lambda}}$ la matrice diagonale dont les éléments diagonaux sont les racines carrées des valeurs propres de T (elles sont rangées par valeurs décroissantes). (1)

Dans le plan des deux premières composantes principales, par exemple, on jugera des colinéarités et des orthogonalités entre les k opérateurs.

4. VERS une DESCRIPTION COHERENTE des VARIABLES QUALITATIVES et de LEURS MODALITES .

4.1 Interprétation de la première composante principale obtenue dans l'analyse globale de k variables qualitatives .

A la première composante principale, vecteur de $\mathcal{L}(F,F)$, correspond l'opérateur :

$$C = \sum_{i=1}^{k} \alpha_i B_i$$

ou α_i est la ième coordonnée du premier facteur principal et :

$$B_i = \frac{A_i - A_{\Delta j}}{\sqrt{p_i - 1}}$$

La première composante principale, renormée à 1, n'est autre que l'opérateur Dp-symétrique rendant maximum la quantité (5) :

$$J_C = \sum_{i=1}^{k} \cos^2 (C, B_i)$$

On peut dire qu'au sens de l'indice J_C l'opérateur C est le plus représentatif des k opérateurs normés B_i ; c'est donc l'opérateur Dp-symétrique qui représente au mieux l'ensemble des variables qualitatives considérées au sens du critère J_C .

D'après le théorème de Frobénius (7) le premier facteur principal, qui est vecteur propre de la matrice T dont tous les coefficients sont positifs ou nuls, peut être choisi de coordonnées toutes positives.

. /..

Les projecteurs B_i s'écrivant (cf.paragraphe 2)

$$B_i = W_i \circ D_p$$

l'opérateur C vérifie : $C = W \circ D_p$

avec : $$W = \sum_{i=1}^{k} \alpha_i \; W_i$$

La forme quadratique W , combinaison linéaire à coefficients positifs des formes quadratiques semi-définies positives W_i, est semi définie positive; les valeurs propres de l'opérateur C sont donc toutes positives ou nulles.

4.2 Pratique usuelle dans la description des modalités de k variables qualitatives .

On sait que lorsque l'on veut décrire la liaison entre deux variables qualitatives, effectuer l'analyse factorielle des correspondances du tableau de contingence revient à effectuer l'analyse factorielle des correspondances du tableau des indicatrices associées aux deux variables considérées, ce dernier tableau étant considéré comme s'il était un tableau de contingence.

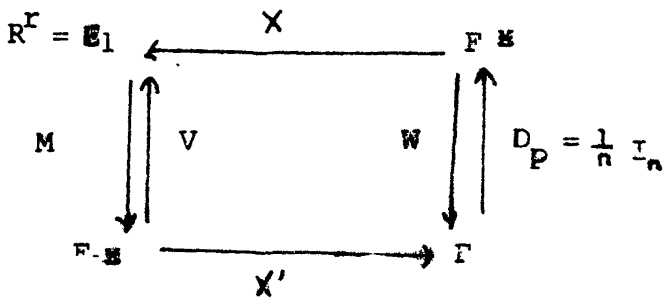
Cette remarque est en partie à l'origine de la pratique actuelle consistant à décrire les modalités de k variables qualitatives en effectuant une analyse factorielle des correspondances sur le tableau des $r=\sum_{i=1}^{k} p_i$ indicatrices associées aux k variables (tableau disjonctif complet) dont les ensembles de modalités sont notés I_1 , I_2 ... I_k, considéré comme s'il était un tableau de contingence.

Les composantes principales obtenues dans cette analyse, fournissant les coordonnées des individus dans le système des axes principaux, ne sont autres que les vecteurs propres, normés à la racine carrée des valeurs propres correspondantes, de l'opérateur :

$$U = \sum_{i=1}^{k} \sqrt{p_i - 1} \; B_i$$

Ces vecteurs et valeurs propres coïncident avec les vecteurs et valeurs propres de l'opérateur $\sum_{i=1}^{k} A_i$ après avoir éliminé le vecteur propre trivial $\underline{j}$ de valeur propre k .

Effectuer l'analyse factorielle des correspondances du tableau des indicatrices X revient à effectuer l'analyse en composantes principales du triplet (X, M, D_p), le schéma de dualité suivant étant considéré :



. la matrice associée à l'application $V = X \circ D_p \circ X'$ est " le tableau de BURT " :

$$V = \begin{pmatrix} D_{p_1} & P_{12} & P_{13} & \cdots & P_{1k} \\ P_{21} & \ddots & & & \\ \vdots & & \ddots & & \\ P_{k1} & & & & D_{p_k} \end{pmatrix}$$

où $P_{jj'}$ est le tableau des probabilités définies sur $I_j \times I_{j'}$,

. la métrique M admet pour matrice :

$$M = \begin{pmatrix} D_{1/p_1} & & 0 \\ & \ddots & \\ 0 & & D_{1/p_k} \end{pmatrix}$$

remarque: dans l'analyse en composantes principales précédente le tableau des indicatrices X n'est pas centré.

4.3 Pratique nouvelle optimale

Au sens du critère J_C l'opérateur $C = W \circ D_p = \sum_{i=1}^{k} \alpha_i B_i$.

est l'opérateur D_p-symétrique, n'admettant que des valeurs propres positives ou nulles, le plus représentatif des k variables considérées . Aussi est-il cohérent, ayant décrit dans un premier temps les variables qualitatives prises globalement en diagonalisant la matrice des carrés des coefficients de Tschuprow, de décrire l'ensemble des modalités des variables qualitatives considérées en diagonalisant

l'opérateur : $\sum_{i=1}^{k} \frac{\alpha_i}{\sqrt{p_i - 1}} A_i$ plutôt que l'opérateur $\sum_{i=1}^{k} A_i$

Procéder ainsi revient en particulier à effectuer l'analyse en composantes principales du triplet (X, M, D_p) , la matrice associée à la métrique M choisie dans l'espace des individus E_1 à $\sum_{i=1}^{k} p_i$ dimensions ayant ici pour expression :

$$M = \begin{pmatrix} \frac{\alpha_1}{\sqrt{p_1 - 1}} D_{1/p_1} & & 0 \\ & \ddots & \\ 0 & & \frac{\alpha_k}{\sqrt{p_k - 1}} D_{1/p_k} \end{pmatrix}$$

Cette analyse est équivalente à l'analyse des correspondances effectuée

$$Z = \begin{pmatrix} a_1 X_1 \\ a_2 X_2 \\ \vdots \\ a_k X_k \end{pmatrix}$$

$$\text{avec} \quad a_i = \frac{\alpha_i}{\sqrt{p_i - 1}}$$

remarque : Si $k = 2$ et $p_1 = p_2$ on trouve $\alpha_1 = \alpha_2$; l'analyse précédente revient alors à l'analyse factorielle des correspondances classique .

5. EXEMPLE .

Ces données proviennent d'une enquête au cours de laquelle 44 personnes ont été interrogées sur des questions d'actualité. Treize variables dichotomiqu (codées 0 et 1) ont été retenues :

Code question	n°	Libellé
Avortement	1	Etes-vous pour la libéralisation de l'avortement ?
Nucléaire	2	Etes-vous favorable aux centrales électriques nucléaires?
Médecins	3	Faites-vous confiance aux médecins ?
Grandes Ecoles	4	Etes-vous pour la suppression des grandes écoles ?
Dieu	5	Croyez-vous en Dieu ?
Milices	6	Doit-on autoriser les municipalités à constituer des milices ?
Larzac	7	Faut-il soutenir le mouvement pour le Larzac ?
Force de frappe	8	Etes-vous pour la force de frappe ?
Justice	9	Faites-vous confiance à la justice ?
Pop-Music	10	Aimez-vous la pop-music ?
Peine de mort	11	Etes-vous pour la peine de mort ?
Mitterand	12	Pensez-vous que François Mitterand ferait un meilleur Président que Valéry Giscard d'Estaing ?
Héritages	13	Faut-il limiter le patrimoine obtenu par héritage ?

. /..

5.1 Méthode usuelle .

Les données etant dichotomiques, effectuer une analyse en composantes principales sur le tableau des réponses affirmatives avec la métrique D $1/\sigma^2$ est équivalent à effectuer l'analyse factorielle des correspondances sur ce dernier tableau dédoublé . L'analyse en composantes principales a l'avantage de fournir avec le "cercle des corrélations" une description des chi-deux (figure 1).

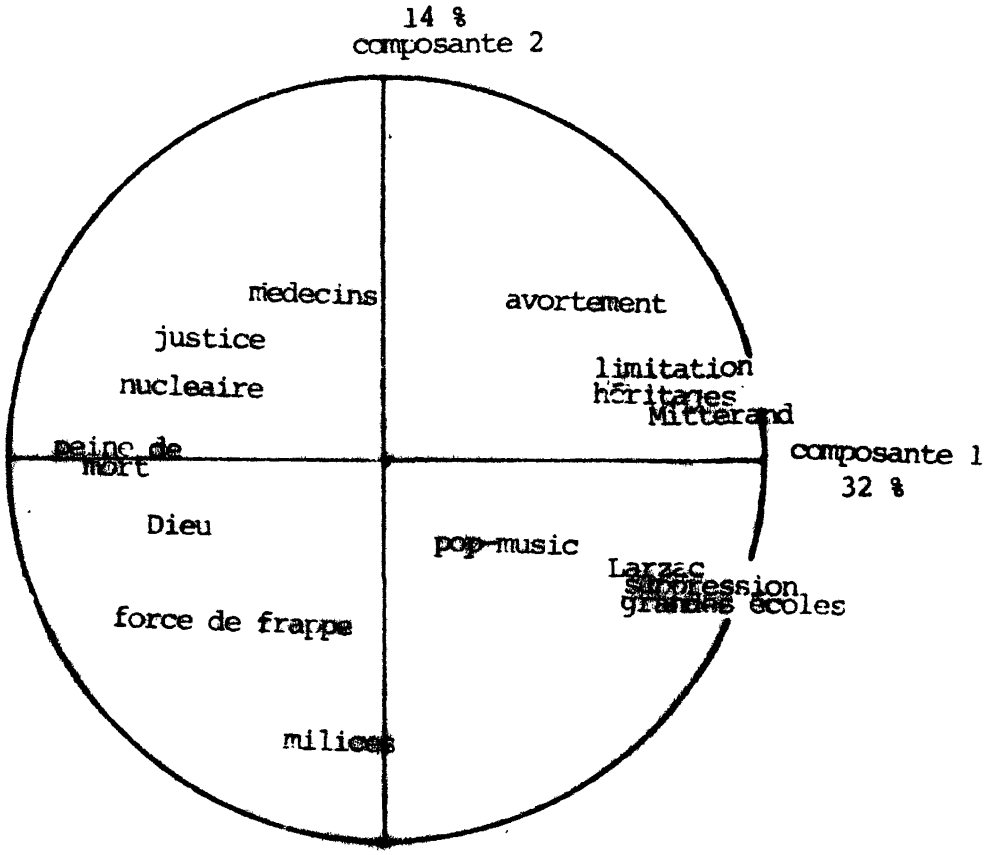


Figure 1

La figure 2 fournit le plan principal obtenu par l'analyse factorielle des correspondances du tableau dédoublé . La question dans la forme ou elle est posée au niveau du questionnaire apparait en écriture droite et sa négation logique en écriture italique .

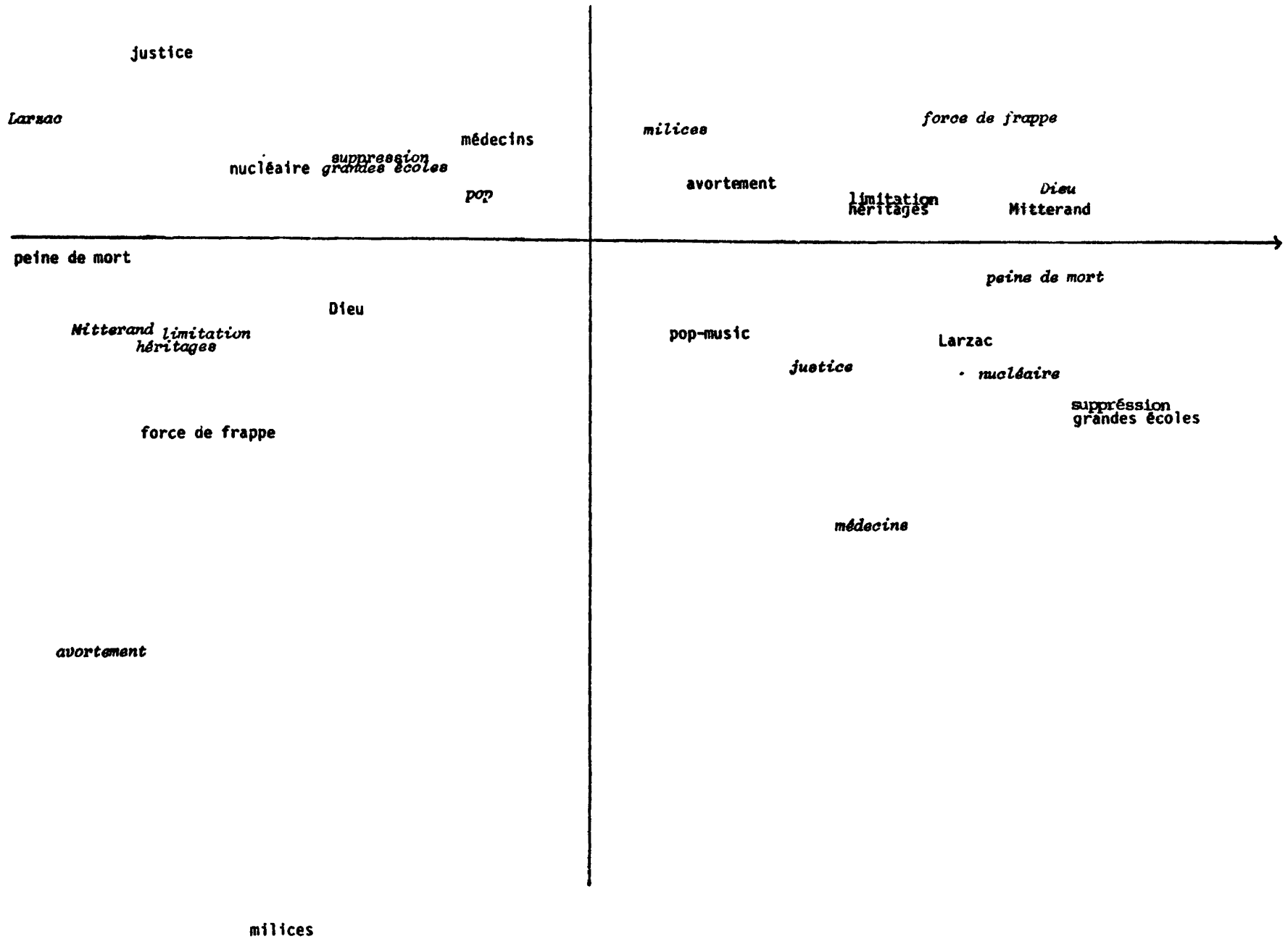

Plan principal 1-2
justice
Larzac
médecins
suppression
nucléaire grandes écoles
pop
milices
force de frappe
avortement
limitation
héritages
Dieu
Mitterand
peine de mort
peine de mort
Dieu
Mitterand limitation
héritages
pop-music
Larzac
justice
nucléaire
suppression
grandes écoles
force de frappe
médecins
avortement
milices

Figure 2

5.2 Pratique nouvelle.

Les variables étant toutes dichotomiques les T^2 de Tchuprow qui ne sont autres que les phi-deux de Yule coïncident avec les carrés des coefficients de corrélation :

1	1.000											
2	.002	1.000										
3	.011	.105	1.000									
4	.033	.136	.110	1.000								
5	.040	.050	.012	.024	1.000							
6	.103	.002	.011	.006	.005	1.000						
7	.027	.148	.058	.223	.095	.008	1.000					
8	.139	.140	.000	.058	.077	.249	.119	1.000				
9	.011	.129	.023	.081	.098	.027	.212	.085	1.000			
10	.015	.000	.011	.003	.008	.000	.024	.005	.152	1.000		
11	.073	.268	.005	.163	.191	.021	.583	.156	.051	.013	1.000	
12	.051	.105	.054	.135	.250	.124	.244	.244	.067	.001	.292	1.000
13	.035	.062	.045	.119	.141	.035	.075	.152	.009	.008	.110	.201	1.000

Matrice des T^2 de Tschuprow

Voici les plus grandes valeurs propres obtenues par la diagonalisation de la matrice T (elles sont rangées par valeurs décroissantes).

valeur propre	pourcentage d'inertie	pourcentage d'inertie cumulée
$\lambda_1 = 2,34$	18%	18%
$\lambda_2 = 1,31$	10%	28%
$\lambda_3 = 1,14$	9%	37%

Toutes les coordonnées du premier facteur principal sont positives, ce sont : 0.112, 0.278, 0.096, 0.260, 0.248, 0.123, 0.446, 0.294, 0.198, 0.042, 0.471, 0.395, 0.229 .

. /..

Le cercle des corrélations (figure 3) obtenu dans l'analyse en composantes principales des opérateurs donne une description des liaisons au sens du chi-deux entre les différentes variables considérées qui doit être comparée à celle fournie par la figure 1.

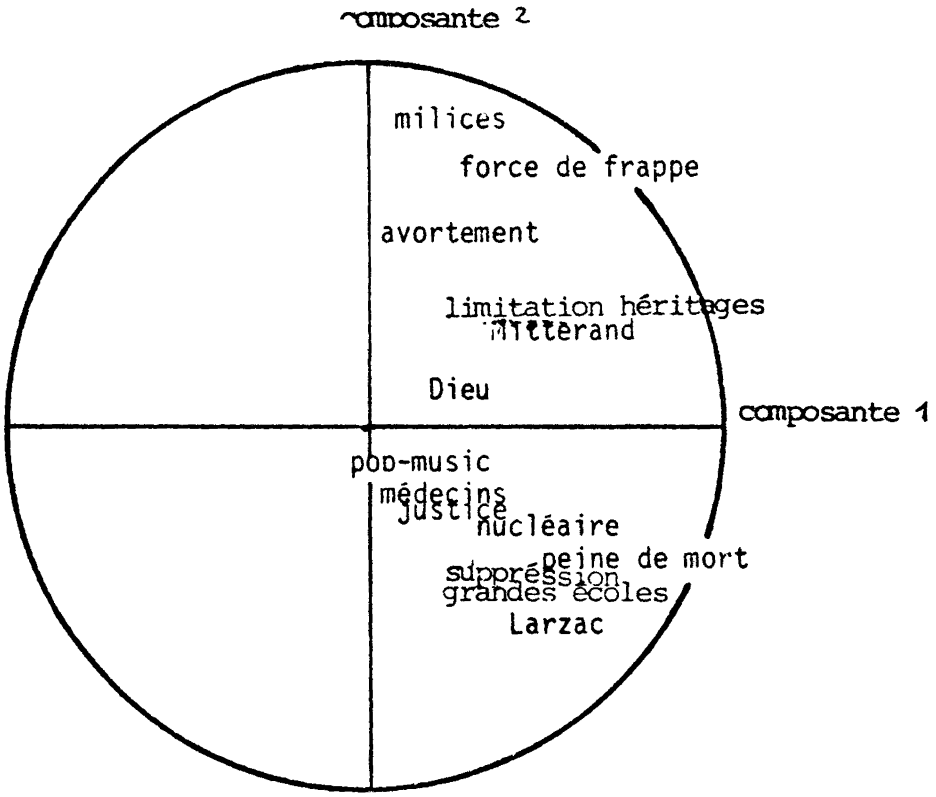


Figure 3

Pour décrire au mieux les modalités des différentes variables on effectue une analyse factorielle des correspondances sur le tableau :

$$Z = \begin{pmatrix} \frac{\alpha_1}{\sqrt{p_1 - 1}} X_1 \\ \frac{\alpha_2}{\sqrt{p_2 - 1}} X_2 \\ \vdots \\ \frac{\alpha_k}{\sqrt{p_k - 1}} X_k \end{pmatrix}$$

Ici:

$p_i = 2$ pour tout $i = 1, 2, \ldots\ldots k$

X_i est le tableau des indicatrices, à 2 lignes et 44 colonnes

ble i.

Les résultats obtenus sont les suivants :

. valeurs propres

VALEUR PROPRE	POURCENTAGE d'INERTIE	POURCENTAGE d'INERTIE CUMULEE
$\lambda_1 = 0,437$	44%	44%
$\lambda_2 = 0,117$	12%	56%
$\lambda_3 = 0,079$	8%	64%

On consultera à la figure 4 le plan principal ; la description des modalités qu'il fournit doit être comparée à celle donnée par la figure 2 .

5.3 Conclusion .

La part d'inertie expliquée par le plan principal est de 55% quand on décrit de facon optimale les modalités des différentes variables ; elle n'est que de 46% quand on effectue l'analyse factorielle des correspondances sur le tableau des indicatrices .

On pouvait s'y attendre; en effet la pratique proposée conduit à favoriser très largement les variables bien liées ; si une variable possède un chi-deux nul avec l'ensemble des autres variables,elle n'est pas prise en compte dans l'analyse optimale ce qui parait logique ; rapprocher les modalités de cette variable des modalités des autres variables ne présenterait aucun intérêt .

Pour l'exemple qui a été donné, l'analyse factorielle des correspondances a fourni des résultats au niveau de la description des modalités des différentes variables, qui sont trés similaires à ceux obtenus dans l'analyse optimale .

Il serait interessant de comparer la méthode proposée aux analyses optimales développées par YOUNG,DE LEEUW,BOUROCHE et SAPORTA (8,9,10) .

Plan principal 1-2

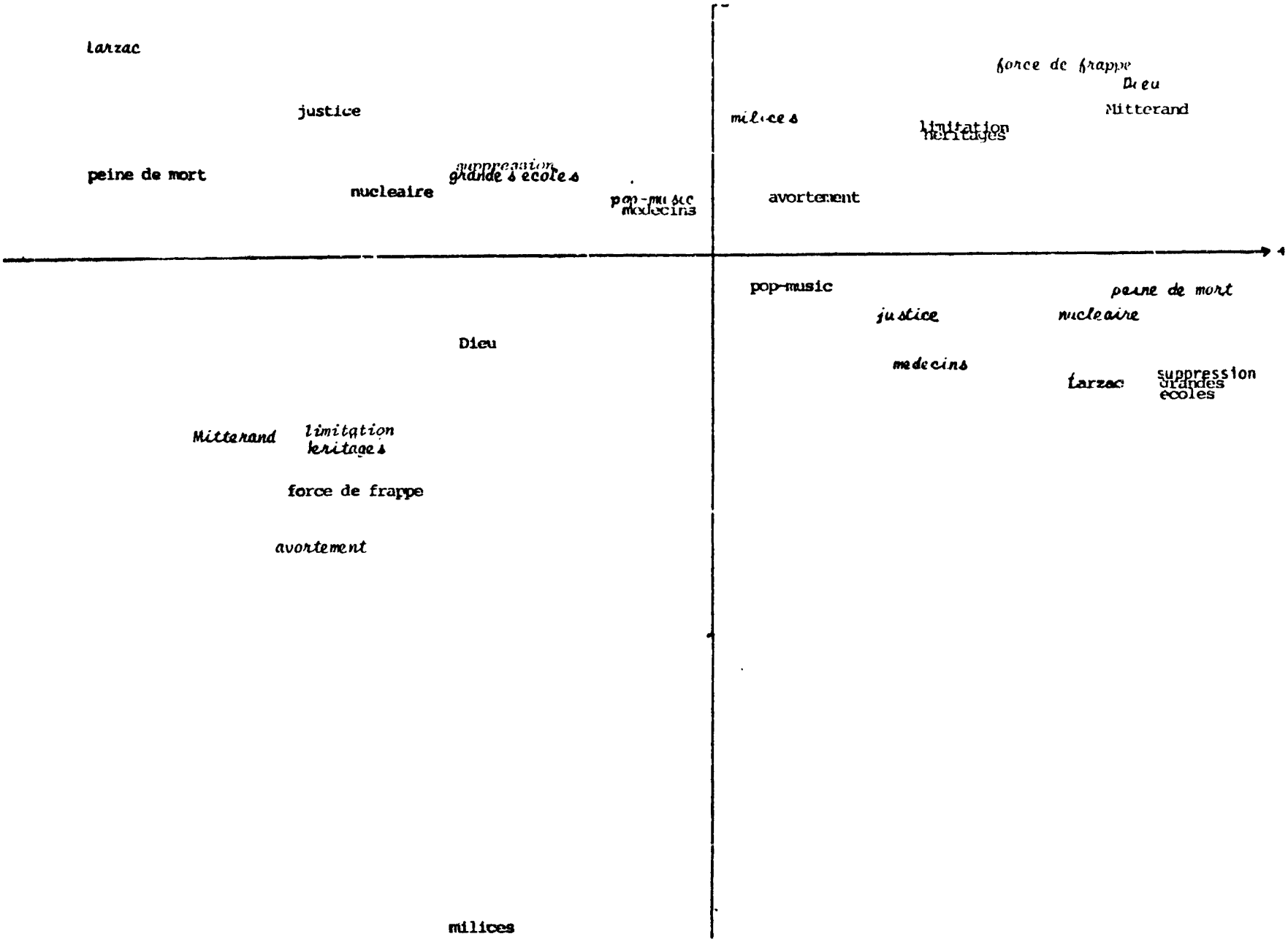


Figure 4

BIBLIOGRAPHIE

(1) F. CAILLIEZ et J.P. PAGES
Introduction à l'analyse de données.
S.M.A.S.H.- 1976

(2) Y. ESCOUFIER
Echantillonnage dans une population de variables aléatoires réelles .
Thèse de Doctorat d'Etat - Université de Montpellier- 1970

(3) J.M. BRAUN
Etude des séries chronologiques multiples par l'analyse des données
Rapport CEA R 4561 - 1974

(4) J.P. PAGES
A propos des opérateurs d'Escoufier .
Séminaires IRIA sur la classification automatique et la perception sur ordinateur - 1974

(5) J.P. PAGES, Y. ESCOUFIER, P. CAZES
Opérateurs et analyse des tableaux à plus de deux dimensions .
Cahiers du Bureau Universitaire de Recherche Opérationnelle -1976

(6) G. SAPORTA
Liaisons entre plusieurs ensembles de variables et codage de données qualitatives .
Thèse de Doctorat de 3ème cycle - Université de Paris VI - 1975

(7) F.R. GANTMACHER
Théorie des matrices .
Tome 2 - Dunod- 1966

(8) F.W. YOUNG,J. DE LEEUW ,Y. TAKANE
Additive structure in qualitative data .
Psychométrika Laboratory - University of North Carolina - 1975

(9) J.M. BOUROCHE,G.SAPORTA,M.TENENHAUS
Generalized canonical analysis of qualitative data .
U.S. Japan,Seminar on theory ,methods and applications of multidimensional scaling - 1975

(10) G. SAPORTA
Le traitement de variables qualitatives par codages .
n°11 ,COREF - 1976