REVUE DE STATISTIQUE APPLIQUÉE

PH. CASIN
J. C. TURLOT

Une présentation de l'analyse canonique généralisée dans l'espace des individus

Revue de statistique appliquée, tome 34, n° 3 (1986), p. 65-75
<http://www.numdam.org/item?id=RSA_1986__34_3_65_0>

© Société française de statistique, 1986, 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/

UNE PRESENTATION DE L'ANALYSE CANONIQUE GENERALISEE DANS L'ESPACE DES INDIVIDUS

Ph. CASIN ([1]) et J.C. TURLOT ([2])

(1) Docteur 3e cycle
Université de Strasbourg I
(2) Ingénieur C.N.R.S.
Centre d'études statistiques
I.R.M.A., Strasbourg I

Nous tenons à remercier tout particulièrement Monsieur Michel TENENHAUS, Professeur associé au C.E.S.A., pour les conseils, critiques et suggestions qu'il nous a adressés lors de l'élaboration de cet article.

RÉSUMÉ

Nous montrons que l'analyse canonique généralisée de CARROLL est une analyse discriminante et nous discutons de certaines conséquences de ce résultat.

SUMMARY

We prove that CARROLL'S generalized canonical analysis is a discriminant analysis and we examine some consequences of this result.

Mots-clés : **Analyse canonique généralisée, Analyse canonique, Analyse discriminante.**

1. INTRODUCTION

L'analyse canonique généralisée (A.C.G.) — au sens de J.D. CARROLL [1] — admet pour cas particuliers toutes les méthodes d'analyse des données usuelles, qu'elles soient « explicatives » comme la régression simple ou multiple, l'analyse de variance ou de covariance, l'analyse discriminante, ou « descriptives » comme l'analyse canonique simple, l'analyse des correspondances simples ou multiples ou l'analyse en composantes principales (A.C.P.) normée.

Après avoir rappelé les présentations traditionnelles de l'A.C.G. dans l'espace des variables et dans l'espace des facteurs (§ 2), nous montrons que l'A.C.G. est une analyse discriminante particulière : les représentations graphiques que l'on obtient alors permettent une analyse plus fine des résultats (§ 3) et coïncident, dans les cas particuliers de l'A.C.G. les plus courants, avec les représentations utilisées habituellement. Cette présentation de l'A.C.G. permet, d'autre part, de reconsidérer les relations existant entre les différentes

techniques d'analyse linéaire, et apporte un éclairage nouveau de certaines de leurs propriétés (§ 4).

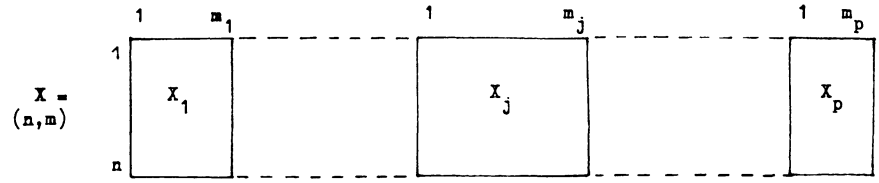
Enfin, on décrit un algorithme de calcul et les principaux résultats nécessaires à l'interprétation d'une A.C.G. (§ 5).

2. RAPPELS SUR L'ANALYSE CANONIQUE GENERALISEE

a. Position du problème, notations

On dispose de p groupes de $q_j (1 \leqslant j \leqslant p)$ variables numériques ou qualitatives; soit X_j le tableau dont les lignes sont les observations de n individus sur ces q_j variables. X_j est un tableau de dimensions $n \times m_j$ (m_j peut être différent de q_j, chaque variable qualitative étant remplacée dans X_j par ses variables indicatrices).

On note $X = [X_1, ..., X_p]$ la juxtaposition des p tableaux. X est de dimensions $n \times m$, où $m = \sum_{j=1}^{p} m_j$; enfin, on suppose — sans perdre en généralité — que chaque tableau X_j est de rang m_j et que les variables numériques sont centrées. Dans R^n, muni de la métrique diagonale des poids des individus D_p, les m_j colonnes du tableau X_j engendrent un espace de dimension m_j, noté W_j.



L'A.C.G., au sens de CARROLL [1], se propose de « résumer » chaque groupe de variables par une variable unique $Z_j^1 \in W_j$ $(1 \leqslant j \leqslant p)$ et de chercher une variable auxiliaire Z^1 de manière à rendre maximale la quantité $\sum_{j=1}^{p} R^2(Z^1, Z_j^1)$; ainsi Z^1 est la variable la plus liée en moyenne à chacun des groupes de variables, l'intensité de la liaison étant mesurée par le carré du coefficient de corrélation linéaire entre Z^1 et Z_j^1 $(1 \leqslant j \leqslant p)$.

b. La solution dans l'espace des variables

Pour $Z \in R^n$ donné, le maximum de $R^2(Z, Z^j)$ est atteint lorsque Z^j coïncide avec la projection D_p-orthogonale de Z sur W_j (CARROLL [1]); il s'ensuit que si on note P_j l'opérateur de D_p-projection sur W_j, on a :

$$Z_j^1 = P_j Z^1 = X_j({}^tX_j D_p X_j)^{-1} {}^tX_j D_p Z^1$$

Z^1 et Z_j^1 sont définis à une constante multiplicative près; Z^1 est choisie D_p-normée à 1 et on retiendra pour Z_j^1 la variable de W_j obtenue par projection orthogonale de Z^1 sur W_j.

De la relation :

$$R^2(Z^1, Z_j^1) = \frac{{}^tZ^1 D_p P_j Z^1}{{}^tZ^1 D_p Z^1} \quad 1 \leqslant j \leqslant p$$

on déduit que Z^1 est le vecteur propre de $\sum_{j=1}^{p} P_j$ associé à sa plus grande valeur propre.

Plus généralement, Z^i est solution de :

$$\max_{Z^i} \sum_{j=1}^{p} R^2(Z^i, Z_j^i)$$

sous les contraintes :

$${}^tZ^i D_p Z^i = 1 \qquad 1 \leqslant i \leqslant m$$

$${}^tZ^i D_p Z^k = 0 \qquad 1 \leqslant i, k \leqslant m \qquad i \neq h$$

Z^i est alors le vecteur propre de $\sum_{j=1}^{p} P_j$ associé à sa $i^{-\text{ème}}$ valeur propre :

$$\left(\sum_{j=1}^{p} P_j \right) Z^i = \lambda_i Z^i$$

$\left(\sum_{j=1}^{p} P_j \right)$, somme d'opérateurs symétriques semi-définis positifs est symétrique semi-défini positif et possède n valeurs propres réelles positives ou nulles.

c. La solution dans l'espace des facteurs

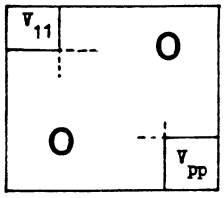
Les composantes canoniques Z^i du tableau X peuvent être déduites de la diagonalisation d'un opérateur de dimension m (SAPORTA [13]).

En effet,

$$\sum_{j=1}^{p} P_j = \sum_{j=1}^{p} X_j ({}^tX_j D_p X_j)^{-1} {}^tX_j D_p = X \, (\text{diag } V_{XX})^{-1} {}^tX D_p$$

si on note diag V_{XX} la matrice m x m suivante :

diag V_{XX} =



où : $V_{ji} = {}^tX_j D_p X_i$

Posons

$$V_{XX} = {}^tX D_p X$$

V_{XX} est, par conséquent, la matrice m x m comportant p^2 blocs, et dont le bloc (j, i) ($1 \leqslant i \leqslant p$, $1 \leqslant j \leqslant p$) est égal à V_{ji}.

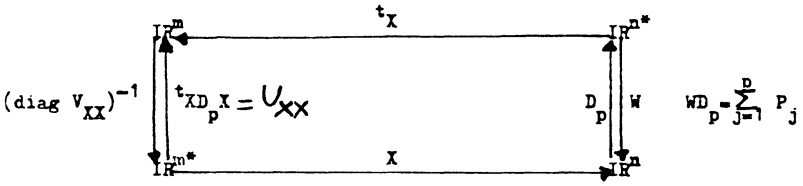
Si $b^i \in R^m$ est vecteur propre de $(\text{diag } V_{XX})^{-1} V_{XX}$:

$$(\text{diag } V_{XX})^{-1} {}^tX D_p X b^i = \lambda_i b^i$$

on obtient, en multipliant à gauche les deux membres de l'équation par X :

$$\left(\sum_{j=1}^{p} P_j\right) Xb^i = \lambda_i Xb^i$$

Xb^i s'identifie par conséquent avec la composante canonique d'ordre i, Z^i, et b^i constitue le $i^{-ème}$ facteur de l'A.C.P. de X, R^m étant muni de la métrique $(\text{diag } V_{XX})^{-1}$. On peut alors décrire l'A.C.G. par le schéma de dualité suivant :



d. La nécessité d'une approche différente

Considérer l'A.C.G. comme une A.C.P. dans l'espace des individus muni de la métrique de MAHALANOBIS par blocs ne fournit pas de règles d'interprétation de la configuration des individus sur les axes principaux.

Nous allons montrer dans le paragraphe suivant que l'on peut présenter l'A.C.G. d'une manière moins abstraite dans R^m, comme le résultat de la maximisation d'un critère simple, ce qui permettra de dégager quelques principes d'interprétation des résultats.

3. L'ANALYSE CANONIQUE GÉNÉRALISÉE EST UNE ANALYSE DISCRIMINANTE

a. Position du problème, notations

Soit G le tableau suivant :

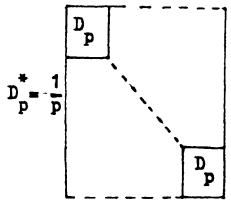
$G_{(np,m)} =$

X_1		0
0 0		0 0
0		X_p

Le tableau X (n x m) décrit dans R^m un nuage de points N; si on suppose que R^m est muni de la métrique euclidienne usuelle, l'écriture de G revient à considérer les projections orthogonales de N sur chacun des sous-espaces

engendrés par les variables des différents tableaux, c'est-à-dire les individus tels qu'ils sont décrits par chaque groupe de variables. On note N_j le nuage des individus décrits par le tableau X_j. (Il s'agit d'une présentation classique que l'on trouve notamment dans ESCOFIER et PAGES [5], GLACON [6]).

La matrice des poids des np individus considérés est la matrice D_p^* suivante :



Si on cherche à déterminer les directions selon lesquelles les nuages N_j sont les plus ressemblants, il est naturel de désirer que les p projections sur un axe Δ de R^m du même individu décrit selon les p tableaux soient proches les unes des autres et que les n groupes ainsi constitués soient le plus distincts possible. Ce problème est résolu de façon classique par l'analyse factorielle discriminante.

b. Equivalence entre analyse canonique généralisée et analyse discriminante

L'équation de décomposition de la variance :

$$V_T = V_E + V_I$$

V_T, V_E et V_I désignant respectivement les matrices de variance totale, entre les classes et à l'intérieur des classes, s'écrit ici, puisque $1/p\ X$ est le tableau des coordonnées des centres de gravité :

$$\frac{1}{p}\begin{vmatrix} V_{11} & & \\ & \ddots & \\ & & V_{pp} \end{vmatrix} = \frac{1}{p^2}\begin{vmatrix} V_{11} & \cdots & V_{1p} \\ \vdots & & \vdots \\ V_{p1} & \cdots & V_{pp} \end{vmatrix} + \frac{1}{p^2}\begin{vmatrix} (p-1)\,V_{11} & \cdots & -V_{1p} \\ \vdots & & \vdots \\ -V_{p1} & \cdots & (p-1)\,V_{pp} \end{vmatrix}$$

On a, par conséquent :

$$V_T = \frac{1}{p}\,\text{diag}\,V_{XX} \quad , \quad V_E = \frac{1}{p^2}\,V_{XX}$$

L'axe Δ, de vecteur directeur u^1, satisfait au critère énoncé ci-dessus si

$$\frac{{}^t u^1\,V_E\,u^1}{{}^t u^1\,V_T\,u^1} = \frac{{}^t u^1\,V_{XX}\,u^1}{p\,{}^t u^1\,(\text{diag}\,V_{XX})\,u^1}$$

a une valeur maximale.

u^l, premier vecteur propre de $V_T^{-1} V_E$ s'identifie au premier facteur canonique (à une normalisation près) et les contraintes d'orthogonalité sur les facteurs discriminants successifs,

$$ {}^t u^j V_T^{-1} u^{j'} = \delta_j^{j'} \quad 1 \leqslant j \leqslant m \quad 1 \leqslant j' \leqslant m $$

$\delta_j^{j'}$ désignant le symbole de KRONECKER, sont équivalentes aux contraintes sur les facteurs canoniques :

$$ {}^t b^j (\text{diag } V_{XX})^{-1} b^{j'} = \delta_j^{j'} \quad 1 \leqslant j \leqslant m \quad 1 \leqslant j' \leqslant m $$

Ainsi, l'A.C.G. est une analyse discriminante, où chaque individu constitue une classe dont les éléments sont les descriptions de l'individu par les p tableaux; ce résultat est le prolongement de l'approche de l'A.C.G. en « termes d'analyse de la variance » proposée par SAPORTA [13].

Remarques

1) Les principaux calculs nécessaires à l'interprétation d'une A.C.G. peuvent ainsi être effectués au moyen d'un programme d'analyse discriminante (ROMEDER [12], par exemple).

2) L'A.C.G. de X peut aussi être présentée comme l'analyse canonique de G et A dans R^{np}

$$ \underset{(np,n)}{A} = \begin{vmatrix} I_n \\ \vdots \\ I_n \end{vmatrix} $$ où I_n désigne la matrice identité d'ordre n.

3) Les valeurs propres de l'A.C.G. sont des pouvoirs discriminants.

c. Les représentations associées à l'analyse de G

En analyse discriminante, on est amené à représenter simultanément les centres de gravité des classes (en projection sur les axes), c'est-à-dire $1/p\ Xb^i$, et les np points par leurs coordonnées sur les axes factoriels, c'est-à-dire Gb^i.

Or :

$$ Gb^i = \begin{vmatrix} X_1 b_1^i \\ X_j b_j^i \\ X_p b_p^i \end{vmatrix} = \frac{1}{\lambda_i} \begin{vmatrix} Z_1^i \\ Z_j^i \\ Z_p^i \end{vmatrix} $$

car : $Z_j^i = \lambda_i X_j b_j^i$ (MASSON [9], SAPORTA [13])

b_j^i désignant le bloc de b^i relatif au j-ème paquet de variables.

Cette représentation revient, en termes d'A.C.G., à superposer les $1/\lambda_i$ Z_j^i, $j = 1, ..., p$ et $1/p\ Z^i$ (qui est égal à $1/p\ Xb^i$) et permet de mettre en évidence les proximités entre les différentes descriptions des individus, selon les p tableaux; par rapport aux graphiques habituellement proposés en A.C.G., on dispose de deux types d'informations utiles :

i) Si on ne tient compte que des centres de gravité, on est réduit à ne donner qu'une interprétation unique des axes factoriels; or le point de vue de l'A.C.G. est très différent : il s'agit d'interpréter chacune des variables Z_j^i, $j = 1, ..., p$, l'interprétation de la variable auxiliaire Z^i qui n'appartient généralement à aucun groupe de variables et qui n'est pas forcément la première composante principale des Z_j^i, $1 \leqslant j \leqslant p$, étant délicate et de peu d'importance.

ii) La dispersion des projections des p observations d'un individu autour de leur centre de gravité permet de mettre en évidence les points « aberrants »; si un groupe a en projection une variance très supérieure à celle des autres groupes, l'individu correspondant a un comportement particulier vis-à-vis des liaisons étudiées.

Remarque

Les variables Gb^i, $i = 1, ..., m$, étant issues de l'analyse canonique de G et A dans R^{np} sont deux à deux D_p^*-orthogonales, ce qui permet de retrouver la propriété d'« orthogonalité faible » des Z_j^i (DEVILLE et SAPORTA [4]) :

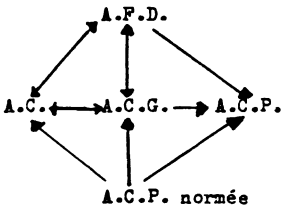
$$\sum_{j=1}^{p} {}^tZ_j^i D_p Z_j^{i'} = 0 \quad \text{si} \quad i \neq i'$$

4. UNE PRÉSENTATION DES CAS PARTICULIERS DE L'A.C.G.

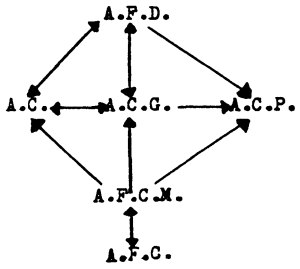
a. Les cas particuliers de l'A.C.G.

Dans ce quatrième paragraphe, nous examinerons successivement les représentations obtenues pour les différentes méthodes classiques d'analyse des données, quand on les considère comme des cas particuliers de l'A.C.G. (ou, ce qui est équivalent, de l'analyse discriminante), selon les schémas suivants :

i) Variables quantitatives



ii) Variables qualitatives



Les résultats de l'analyse d'un tableau X par une méthode située à l'origine d'une flèche peuvent aussi être obtenus par la méthode située à l'extrémité de la même flèche, les différents sigles utilisés désignant respectivement :

A.F.C.M. : analyse factorielle des correspondances multiples
A.F.C. : analyse factorielle des correspondances simples
A.F.D. : analyse factorielle discriminante
A.C. : analyse canonique simple
A.C.P. : analyse en composantes principales
A.C.G. : analyse canonique généralisée

On notera, en particulier, les résultats suivants peu connus :

i) L'analyse canonique est une analyse discriminante.
ii) L'A.C.P. normée est un cas particulier de l'analyse canonique simple et de l'analyse discriminante.

b. L'analyse canonique et l'analyse des correspondances simples

Dans le cas de l'analyse canonique, on dispose de deux ensembles de variables X_1 et X_2 engendrant dans R^n deux sous-espaces W_1 et W_2 de dimensions respectives m_1 et m_2. Si (S_1^i, S_2^i) représente le $i^{-\text{ième}}$ couple canonique, associé à la valeur propre r_i, on montre (KETTENRING [8], SAPORTA [13]) que la valeur propre λ_i et la variable canonique Z^i, au sens de CARROLL, sont obtenues ainsi :

$i = 1, \inf(m_1, m_2)$	$\lambda_i = 1 + r_i$	$Z^i = \frac{1}{2}(S_1^i + S_2^i)$
$i = 1 + \inf(m_1, m_2), \sup(m_1, m_2)$	$\lambda_i = 1$	$Z^i = S_1^i$ si $m_1 > m_2$ $= S_2^i$ sinon
$i = \sup(m_1, m_2) + 1,$ $m_1 + m_2$	$\lambda_i = 1 - r_{i'},$ $i' = \inf(m_1, m_2) - i$	$Z^i = \frac{1}{2}(S_1^{i'} - S_2^{i'})$

En pratique, on ne s'intéresse au plus qu'à l'interprétation des $\inf(m_1, m_2)$ premiers facteurs.

Selon la présentation donnée de l'A.C.G. au paragraphe précédent, on représente simultanément Z^i, $P_1 Z^i$ et $P_2 Z^i$ (à une homothétie près), soit : Z^i, S_1^i et S_2^i. Dans le cas de l'A.F.C., les deux variables considérées sont qualitatives et sont identifiées à l'ensemble de leur modalités. Les individus sont alors caractérisés par les modalités qu'ils prennent sur X_1 et X_2 : un individu, pour un tableau donné, a alors sa représentation confondue avec celle de la modalité qu'il choisit. Superposant S_1^i et S_2^i, on obtient les graphiques habituels de l'A.F.C.

c. L'analyse des correspondances multiples et l'A.C.P. normée

L'analyse des correspondances multiples est le cas particulier de l'A.C.G. pour lequel chaque groupe de variables X_j décrit les m_j indicatrices d'une variable qualitative; la présentation donnée de l'A.C.G. permet de retrouver les principales propriétés de l'A.F.C.M.

En effet, les tableaux G et A décrivant ici les modalités de deux variables qualitatives, l'analyse canonique de G et A est une analyse des correspondances simples et en possède les propriétés, le tableau de contingence associé étant ${}^tAG = X$.

Il est plus intéressant de remarquer que puisque G et A sont deux variables qualitatives, l'A.F.C.M. est une double analyse discriminante, ce qui prouve l'équivalence des deux problèmes suivants :

i) rechercher un codage des modalités discriminant au mieux les individus,

ii) rechercher un codage des individus discriminant au mieux les modalités qui conduisent tous deux à effectuer une analyse canonique de G et A. Cette équivalence montrée par le calcul par GUTTMAN [7] trouve ici une interprétation géométrique.

Remarques

1) L'équivalence entre l'A.F.C.M. de X et l'A.F.C. du tableau (G, A) a été montrée dans CAZES, BAUMERDER, BONNEFOUS et PAGES [3], et se déduit ainsi du schéma donné plus haut (4) a) schéma ii) :

A.F.C.M. (X) $\leftrightarrow$ A.F.D. de G et A $\leftrightarrow$ A.C.G. (G, A) $\leftrightarrow$ A.F.C. (G, A)

2) L'A.F.C.M. et l'A.C.P. normée étant deux analyses canoniques particulières dans R^{np}, toute méthode généralisant l'analyse canonique les admet comme cas particuliers.

Dans le cas de l'A.C.P. normée, on est amené à représenter les np données contenues dans le tableau X et les n lignes de ce tableau : on obtient ainsi, en plus de la représentation usuelle des n individus la représentation des p axes de variables.

5. LE PROGRAMME S.A.T.I.R.E.

Dans ce paragraphe, nous allons décrire brièvement l'algorithme de calcul et les principales sorties du programme S.A.T.I.R.E. (Support à l'Analyse des Tableaux d'Individus Répétés ou Evoluant). On trouvera la liste des instructions de ce programme dans CASIN [2].

a. L'algorithme de calcul

Pour obtenir les résultats de l'A.C.G. en effectuant un minimum de calculs, on recherche par le procédé d'orthogonalisation de SCHMIDT une base D_p-orthogonale de chaque espace W_j, Y_j égale à $X_j A_j$, la détermination des composantes canoniques étant indépendante des bases retenues sur les W_j. La diagonalisation de $YD_p {}^tY$ ou de ${}^tYD_p Y$, Y étant la matrice $[Y_1, ..., Y_p]$ permet alors d'obtenir rapidement les composantes ou les facteurs canoniques.

b. Les sorties du programme

- Les pouvoirs discriminants;
- Les axes canoniques Z^i;
- Les coordonnées des descriptions des individus selon les p tableaux : Z_j^i, $j = 1,..., p$;
- Les corrélations entre les variables canoniques Z^i et les variables de départ;
- Les corrélations entre les variables canoniques Z^i et les variables Z_j^i, $i = 1,..., p$;
- Les facteurs b^i, calculés à partir des variables de départ normées à 1;
- Les variances résiduelles absolues (ie pour chacun des n individus l'inertie de sa classe pour l'axe considéré);
- Les variances résiduelles relatives (ie pour chacun des n individus le rapport de la variance résiduelle absolue à l'inertie du centre de gravité).

6. CONCLUSION

La présentation que nous donnons ici de l'analyse canonique généralisée devrait renforcer la cohérence des techniques d'analyse de données; la simplicité du critère utilisé — le critère de discrimination usuel — permet une meilleure compréhension des techniques d'analyse linéaire, et facilite l'interprétation de leurs résultats, en particulier dans le cas de l'analyse canonique.

BIBLIOGRAPHIE

[1] J.D. CARROLL. — "A generalization of canonical correlation analysis to three or more sets of variables", Proc. 76th Amer. Psych. Assoc., 1968.

[2] Ph. CASIN. — "L'analyse canonique généralisée et l'analyse des tableaux évolutifs...", Thèse de 3e cycle, Faculté des sciences économiques et de gestion, Strasbourg I, 1985.

[3] P. CAZES, A. BAUMERDER, S. BONNEFOUS, J.P. PAGES. — "Codage et analyse des tableaux logiques. Introduction à la pratique des variables qualitatives", Cahiers du B.U.R.O. n° 27, 1977.

[4] J.C. DEVILLE, G. SAPORTA. — "Analyse harmonique qualitative", Data Analysis and Informatics, North-Holland Publishing Company, 1980.

[5] B. ESCOFIER, J. PAGES. — "L'analyse factorielle multiple. Une méthode de comparaison de groupes de variables", Third International Symposium Data Analysis and Informatics, 1983.

[6] F. GLACON. — "Analyse conjointe de plusieurs matrices de données", Thèse de 3ecycle, Université de Grenoble, 1981.

[7] L. GUTTMAN. — "The quantification of a class of attributes; a theory and method of scale construction", Social Science Research Council, 1941.

[8] R.J. KETTENRING. — "Canonical Analysis of several sets of variables", Biometrika n° 58, 1971.

[9] M. MASSON. — "Méthodologies générales du traitement statistique de l'information de masse", Nathan 1980.

[10] L. LEBART, A. MORINEAU, J.P. FENELON. — "Traitement des données statistiques", Dunod 1979.

[11] J.J. Mc KEON. — "Canonical analysis : Some relations between canonical correlation, factor analysis, discriminant function analysis, and scaling theory", Psychometric Monograph n° 13, 1966.

[12] J.M. ROMEDER. — "Méthodes et programmes d'analyse discriminante", Dunod, 1973.

[13] G. SAPORTA. — "Liaisons entre plusieurs ensembles de variables et codage des données qualitatives", Thèse de 3e cycle, Université de Paris VI, 1975.

[14] M. TENENHAUS, F.W. YOUNG. — "An analysis and synthesis of multiple correspondance analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data", Psychometrika vol. 50, n° 1, 1985.