

REVUE DE STATISTIQUE APPLIQUÉE

MICHEL TENENHAUS

Analyse en composantes principales d'un ensemble de variables nominales ou numériques

Revue de statistique appliquée, tome 25, n° 2 (1977), p. 39-56

http://www.numdam.org/item?id=RSA_1977__25_2_39_0

© Société française de statistique, 1977, 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/>

ANALYSE EN COMPOSANTES PRINCIPALES D'UN ENSEMBLE DE VARIABLES NOMINALES OU NUMÉRIQUES

Michel TENENHAUS

C.E.S.A.

1, rue de la Libération, 78350 Jouy en Josas

RESUME

Nous disposons d'un ensemble de données décrites par des variables nominales ou numériques. Nous recherchons un codage des variables nominales (à chaque modalité de chaque variable nominale sera associé un nombre réel) tel que l'analyse en composantes principales à m facteurs effectuée sur les variables nominales codées et les variables numériques soit optimale en terme de variance expliquée.

Nous proposons une procédure itérative utilisant alternativement l'analyse en composantes principales et l'analyse canonique pour déterminer de tels codages. Il est cependant possible que la solution obtenue soit un optimum local.

L'analyse en composantes principales des variables nominales codées et des variables numériques permet alors de structurer le phénomène étudié et de représenter dans des espaces de faible dimension variables, modalités et observations.

INTRODUCTION

Les données d'une enquête sociologique sont en général formées d'un grand nombre d'observations décrites à l'aide de caractéristiques nominales, ordinales ou numériques.

Si nous supposons que toutes ces variables contribuent de manière à priori équivalente à la description du phénomène étudié, la structuration de ces données se ramène fréquemment à la représentation dans des espaces de faible dimension des variables, des modalités et des observations.

Les méthodes statistiques utilisées pour résoudre ce problème dépendent de la nature des variables. Si toutes les variables sont numériques on utilise l'analyse en composantes principales. Si toutes les variables sont nominales l'analyse des correspondances du tableau disjonctif complet (tableau formé des valeurs des variables indicatrices associées aux modalités des variables nominales) est une méthode largement utilisée. Cette méthode ne permet cependant pas la représentation des variables. C'est le besoin de les reprendre en compte pour saisir leurs interrelations globales qui a incité G. Saporta [12] à développer l'analyse en composantes principales sur les coefficients de Tschuprow et P. Cazes, S. Bonnefous, A. Baumerder et J.P. Pagès [6] à mettre au point une nouvelle méthode : "L'analyse des correspondances optimale".

Lorsque toutes les variables sont ordinales, ou bien de nature quelconque, c'est une méthode de codage optimal qui nous semble appropriée. L'idée est simple : les variables qualitatives sont codées, en respectant leurs structures, de manière à optimiser l'analyse en composantes principales effectuée sur les variables qualitatives codées et les variables numériques.

Y. Takane, F.W. Young et J. de Leeuw [14] ont mis au point une méthode de recherche de ces codages optimaux en utilisant une procédure de moindres carrés alternés. Ils ont également écrit un programme : "PRINCIPALS, a program for principal component analysis of qualitative or quantitative data which uses the alternating least square method". Ce programme est actuellement limité à 20 variables et 200 observations.

Nous avons utilisé cette méthode en préalable à une typologie de données décrites à l'aide de variables hétérogènes (J.M. Bourouche et M. Tenenhaus [2]).

De son côté J. De Leeuw développe une bibliothèque de programmes d'analyse de données qualitatives : le projet HOMALS [9]. Il utilise principalement les idées contenues dans sa thèse [8] et cherche à construire des programmes pouvant traiter un grand nombre d'observations et de variables.

Une méthode fait le lien entre l'analyse des correspondances et les méthodes de codage optimal : l'analyse canonique généralisée de J.D. Carroll [5]. M. Masson [10] et G. Saporta [11] ont montré indépendamment que l'analyse canonique généralisée des ensembles formés des variables indicatrices associées aux variables nominales redonne les équations de l'analyse des correspondances du tableau disjonctif complet.

Par ailleurs M. Masson [10] a utilisé l'analyse canonique généralisée pour étudier les liaisons non linéaires entre variables numériques, ce qui revient à les découper en classes et, les considérant comme des variables nominales, à rechercher des codages rendant les liaisons maximales.

Pour notre part c'est en étudiant les méthodes de "Conjoint Measurement" (P. Bertier et M. Tenenhaus [1]) que nous avons constaté que l'analyse canonique généralisée des variables nominales transformées en variables indicatrices était basée sur le même critère que l'analyse en composantes principales des variables nominales codées, à ceci près que les codages obtenus étaient multi-dimensionnels. C'est le désir d'obtenir un seul codage par variable nominale qui nous a incité à introduire, à la suite de l'analyse canonique généralisée, la procédure itérative décrite dans cet article.

Si notre objectif est exactement le même que celui de Y. Takane, F.W. Young et J. de Leeuw [14], (mais nous nous limitons à des variables nominales ou numériques), notre méthode a l'avantage de ne pas partir d'une solution arbitraire et donc, pratiquement, de converger plus rapidement. De plus dans le cadre des projets de recherche COREF-DGRST n° 757-0230 et C.E.S.A. n° 3017, J.L. Vachette [15] a mis au point un programme, PRINQUAL, correspondant à la méthode décrite dans cet article et pouvant traiter un grand nombre de variables et d'observations.

Nous illustrons cette méthode par son utilisation sur des données qui nous ont été communiquées par J.P. Pagès.

Enfin signalons les notes de G. Saporta [13] sur les problèmes de codage : “le codage : astuce ou méthode fondamentale ?” de J.M. Bouroche, G. Saporta et M. Tenenhaus [3] sur des méthodes de traitement de variables qualitatives, l'article de P. Cazes, A. Baumerder, S. Bonnefous et J.P. Pagès [7] “Codage et analyse des tableaux logiques : Introduction à la pratique des variables qualitatives” et la synthèse générale de toutes ces techniques effectuée dans le rapport de recherche COREF-DGRST [10].

1 – ANALYSE EN COMPOSANTES PRINCIPALES D'UN ENSEMBLE DE VARIABLES NOMINALES

1.1. Position du problème

Nous disposons de k variables nominales D_i mesurées sur un ensemble E de n observations.

Nous désirons coder ces variables de manière à obtenir la meilleure analyse en composante principales à m facteurs possible de ces variables codées.

Nous notons δ_i un codage des modalités de D_i . C'est une application réelle définie sur l'ensemble $\mathcal{O}_i$ des modalités de D_i :

$$E \xrightarrow{D_i} \mathcal{O}_i \xrightarrow{\delta_i} R$$

$\delta_i \circ D_i$

La variable D_i codée est donc représentée par la variable réelle $\delta_i \circ D_i$.

Pour éviter des problèmes de dégénérescence nous imposons aux variables $\delta_i \circ D_i$ d'être centrées-réduites :

$$\sum_{e \in E} \delta_i \circ D_i(e) = 0 \quad \text{et} \quad \sum_{e \in E} (\delta_i \circ D_i(e))^2 = n$$

L'analyse en composantes principales à m facteurs des variables centrées réduites $\delta_i \circ D_i$ consiste à rechercher m variables réelles $Z_1, \dots, Z_m$ définies sur E , centrées réduites et non corrélées maximisant

$$\sum_{i=1}^k \sum_{j=1}^m \text{cor}^2(\delta_i \circ D_i, Z_j).$$

La valeur de cet indice au maximum représente la variance expliquée par les m premières composantes principales.

D'où l'idée de rechercher les codages optimaux δ_i en maximisant cet indice à la fois sur les δ_i et les Z_j .

Le problème à résoudre s'écrit donc :

$$\begin{array}{l}
\text{Maximiser} \\
\delta_1, \dots, \delta_k \\
\delta_i \circ D_i \text{ centrée réduite} \\
Z_1, \dots, Z_m \\
\text{centrées réduites} \\
\text{non corrélées}
\end{array}
\sum_{i=1}^k \sum_{j=1}^m \text{cor}^2(\delta_i \circ D_i, Z_j)$$

Chaque variable $\delta_i \circ D_i$ est combinaison linéaire des p_i variables indicatrices $D_{i1}, \dots, D_{ip_i}$ associées à D_i :

$$\delta_i \circ D_i = \sum_{l=1}^{p_i} \delta_i(l) D_{il}.$$

Examinons à quelles conditions sur les codages $\delta_i(l)$ une telle variable est centrée réduite :

La contrainte " $\delta_i \circ D_i$ centrée" est équivalente à la condition

$$\sum_{l=1}^{p_i} n_{il} \delta_i(l) = 0$$

où n_{il} représente la fréquence sur la population E de la modalité l de la variable D_i . Par conséquent, en tenant compte de cette condition, une variable $\delta_i \circ D_i$ centrée peut s'écrire :

$$\delta_i \circ D_i = \sum_{l=1}^{p_i-1} \delta_i(l) \left(D_{il} - \frac{n_{il}}{n_{ip_i}} D_{ip_i} \right) = \Delta_i d_i$$

$$\text{où} \quad \Delta_i = \left[D_{i1} - \frac{n_{i1}}{n_{ip_i}} D_{ip_i}, \dots, D_{i(p_i-1)} - \frac{n_{i(p_i-1)}}{n_{ip_i}} D_{ip_i} \right]$$

$$\text{et} \quad d_i = \begin{bmatrix} \delta_i(1) \\ \vdots \\ \delta_i(p_i - 1) \end{bmatrix}$$

Nous considérons maintenant les variables observées sur E comme des vecteurs de $\mathbf{R}^n$.

La contrainte " $\delta_i \circ D_i$ centrée-réduite" est alors équivalente aux conditions

$$\delta_i \circ D_i = \Delta_i d_i \text{ et } d_i' \left(\frac{1}{n} \Delta_i' \Delta_i \right) d_i = 1.$$

Si nous posons :

$$F(d_1, \dots, d_k, Z_1, \dots, Z_m) = \sum_{i=1}^k \sum_{j=1}^m \text{cor}^2(\Delta_i d_i, Z_j)$$

le problème à résoudre peut s'écrire :

$$\begin{array}{ll} \text{Maximiser} & F(d_1, \dots, d_k, Z_1, \dots, Z_m) \\ & d_1, \dots, d_k \\ \text{avec} & d_i' \left(\frac{1}{n} \Delta_i' \Delta_i \right) d_i = 1 \\ & Z_1, \dots, Z_m \\ & \text{centrées réduites} \\ & \text{non corrélées} \end{array}$$

1.2. L'algorithme de résolution

La méthode est itérative et consiste à effectuer alternativement une analyse en composantes principales et un ensemble d'analyses canoniques.

Nous rappelons tout d'abord ces méthodes sous la forme particulière qui nous est utile et avec tous les résultats nécessaires à l'algorithme.

1.2.1. Rappels

Proposition 1()*

Pour des d_i donnés, vérifiant les contraintes $d_i' \left(\frac{1}{n} \Delta_i' \Delta_i \right) d_i = 1$, le maximum de $F(d_1, \dots, d_k, Z_1, \dots, Z_m)$ sur des $Z_1, \dots, Z_m$ centrées réduites et non corrélées est atteint pour les m premières composantes principales réduites, des k variables centrées réduites $\delta_i \circ D_i = \Delta_i d_i$ et la valeur du maximum est la variance expliquée par les m premières composantes principales.

Démonstration

Nous avons :

$$\begin{aligned} F(d_1, \dots, d_k, Z_1, \dots, Z_m) &= \sum_{i=1}^k \sum_{j=1}^m \text{cor}^2(\Delta_i d_i, Z_j) = \\ &= \sum_{i=1}^k \sum_{j=1}^m \left(\frac{1}{n} d_i' \Delta_i' Z_j \right)^2 = \frac{1}{n^2} \sum_{j=1}^m Z_j' D D' Z_j \end{aligned}$$

où D est la matrice $[\Delta_1 d_1, \dots, \Delta_k d_k]$ des variables codées observées sur E .

Le Maximum $F(d_1, \dots, d_k, Z_1, \dots, Z_m)$ est
 $Z_1, \dots, Z_m$
 centrées réduites
 non corrélées

(*) On trouvera une proposition plus générale que celle-ci dans F. Caillez et J.P. Pagés [4], p. 490.

clairement atteint pour les m premiers vecteurs propres $Z_1^*, \dots, Z_m^*$, normés à $\sqrt{n}$, de $\frac{1}{n} DD'$. Si nous notons μ_j la j -ième valeur propre de $\frac{1}{n} DD'$ nous avons $F(d_1, \dots, d_k, Z_1^*, \dots, Z_m^*) = \sum_{j=1}^m \mu_j$. Mais μ_j est également la j -ième valeur propre de la matrice des corrélations entre les variables $\Delta_i d_i$, $R = \frac{1}{n} D'D$, et $\sum_{j=1}^m \mu_j$ représente bien la variance expliquée par les m premières composantes principales des k variables centrées réduites $\Delta_i d_i$.

Proposition 2

Pour des Z_j données, centrées-réduites et non corrélées, le maximum de $F(d_1, \dots, d_k, Z_1, \dots, Z_m)$ sur des $d_1, \dots, d_k$ vérifiant les contraintes $d_i' \left(\frac{1}{n} \Delta_i \Delta_i \right) d_i = 1$, est atteint pour des d_i^* obtenus par analyse canonique entre les variables formées par les colonnes de Δ_i et les variables $Z_1, \dots, Z_m$: d_i^* est le vecteur canonique associé à Δ_i correspondant à la première corrélation canonique $\sqrt{\lambda_i}$.

La valeur du maximum $F(d_1^*, \dots, d_k^*, Z_1, \dots, Z_m)$ est égale à $\sum_{i=1}^m \lambda_i$.

Démonstration

Les variables Z_j étant non corrélées nous pouvons écrire :

$$\begin{aligned} F(d_1, \dots, d_k, Z_1, \dots, Z_m) &= \sum_{i=1}^k \sum_{j=1}^m \text{cor}^2(\Delta_i d_i, Z_j) \\ &= \sum_{i=1}^k R^2(\Delta_i d_i ; Z_1, \dots, Z_m) \\ &= \sum_{i=1}^k \max_{e_i} \text{cor}^2(\Delta_i d_i, Z e_i) \end{aligned}$$

où R^2 est le carré du coefficient de corrélation multiple, e_i un vecteur de $\mathbf{R}^m$ et $Z = [Z_1, \dots, Z_m]$.

Puis nous avons :

$$\text{Maximum}_{d_1, \dots, d_k} F(d_1, \dots, d_k, Z_1, \dots, Z_m) = \sum_{i=1}^k \max_{e_i, d_i} \text{cor}^2(\Delta_i d_i, Z e_i).$$

Il est donc clair que chaque vecteur d_i^* s'obtient par analyse canonique entre les colonnes de Δ_i et celles de Z . Et si λ_i est le carré de la première corrélation canonique nous avons bien :

$$\text{Maximum}_{d_1, \dots, d_k} F(d_1, \dots, d_k, Z_1, \dots, Z_n) = \sum_{i=1}^k \lambda_i.$$

Les conditions $d_i' \left(\frac{1}{n} \Delta_i' \Delta_i \right) d_i = 1$ sont habituelles en analyse canonique.

Chaque analyse canonique est particulièrement simple à réaliser sur le plan numérique car elle correspond en réalité à une analyse discriminante où les groupes sont définis par la variable nominale D_i et les variables discriminantes les Z_j .

Notons $\bar{Z}(i, l) = [\bar{Z}_1(i, l), \dots, \bar{Z}_m(i, l)]$ le vecteur moyenne des Z_j sur les observations e appartenant à $D_i^{-1}(l)$.

Les vecteurs d_i^* s'obtiennent alors à l'aide de la proposition suivante.

Proposition 3()*

Le maximum de $\text{cor}^2(\Delta_i d_i, Ze_i)$ pour des d_i et e_i vérifiant

$$d_i' \left(\frac{1}{n} \Delta_i' \Delta_i \right) d_i = e_i' e_i = 1$$

est atteint pour e_i^* vecteur propre de

$$\begin{aligned} B_i &= \frac{1}{n} \sum_{l=1}^{p_i} n_{il} \bar{Z}(i, l)' \bar{Z}(i, l) \\ &= \frac{1}{n} Z' \Delta_i (\Delta_i' \Delta_i)^{-1} \Delta_i' Z \end{aligned}$$

correspondant à la plus grande valeur propre λ_i et pour :

$$d_i^* = \frac{1}{\sqrt{\lambda_i}} \begin{bmatrix} \bar{Z}(i, 1) \\ \vdots \\ \bar{Z}(i, p_i - 1) \end{bmatrix} e_i^* = \frac{1}{\sqrt{\lambda_i}} (\Delta_i' \Delta_i)^{-1} \Delta_i' Z e_i^*.$$

Démonstration

Le vecteur e_i^* , premier facteur discriminant, est le vecteur propre correspondant à la plus grande valeur propre λ_i de $T^{-1} B_i$ où T est la matrice variance – covariance totale des Z_j et B_i la matrice variance – covariance inter-classes.

Ici $T = I$ puisque les Z_j sont centrées, réduites et non corrélées et B_i est bien la matrice définie dans la proposition. Le vecteur d_i^* s'obtient en utilisant les résultats classiques de l'analyse canonique.

(*) Cette proposition rappelle, dans un cas particulier, des résultats exposés par F. Cailliez et J.P. Pagès [4], chapitre 12.

1.2.2. L'algorithme

1) Etape d'initialisation

Nous pouvons choisir des $d_i^{(0)}$ vérifiant les contraintes de manière arbitraire ou bien utiliser les résultats d'une analyse canonique généralisée (voir la 2^{ème} partie de cet article).

2) Etape t

A l'étape précédente, nous avons obtenu des $d_i^{(t-1)}$ vérifiant les contraintes.

A l'étape t nous recherchons des $d_i^{(t)}$ et des $Z_j^{(t)}$.

a) Recherche des $Z_j^{(t)}$

Nous recherchons des $Z_j^{(t)}$ centrées réduites et non corrélées rendant maximum $F(d_1^{(t-1)}, \dots, d_k^{(t-1)}, Z_1, \dots, Z_m)$.

La solution de ce problème est donnée par la proposition 1 et nous notons $\mu^{(t)}$ la valeur du maximum.

b) Recherche des $d_i^{(t)}$

Nous recherchons des $d_i^{(t)}$ vérifiant les contraintes $d_i^{(t)} \left(\frac{1}{n} \Delta_i' \Delta_i \right) d_i^{(t)} = 1$ rendant maximum $F(d_1, \dots, d_k, Z_1^{(t)}, \dots, Z_m^{(t)})$.

La solution de ce problème est donnée par la proposition 2 et nous notons $\lambda^{(t)}$ la valeur du maximum.

3) Convergence de l'algorithme

Deux indices caractérisent l'étape t : $\mu^{(t)}$ et $\lambda^{(t)}$. La convergence de ces deux suites vers une même limite assure la convergence de l'algorithme.

Proposition 4

Les suites $\mu^{(t)}$ et $\lambda^{(t)}$ sont croissantes et convergent vers une même limite L .

Démonstration

Les deux suites $\mu^{(t)}$ et $\lambda^{(t)}$ sont croissantes, bornées supérieurement et imbriquées : nous avons en effet : $\mu^{(t)} \leq \lambda^{(t)} \leq \mu^{(t+1)} \leq k$ puisque

$$\begin{aligned} F(d_1^{(t-1)}, \dots, d_k^{(t-1)}, Z_1^{(t)}, \dots, Z_m^{(t)}) &\leq F(d_1^{(t)}, \dots, d_k^{(t)}, Z_1^{(t)}, \dots, Z_m^{(t)}) \leq \\ &\leq F(d_1^{(t)}, \dots, d_k^{(t)}, Z_1^{(t+1)}, \dots, Z_m^{(t+1)}) \end{aligned}$$

(nous maximisons à chaque fois le même indice) et, puisque, par ailleurs :

$$F(d_1, \dots, d_k, Z_1 \dots Z_m) = \sum_{i=1}^k R^2(\Delta_i d_i ; Z_1 \dots Z_m) \leq k$$

Ces deux suites sont donc convergentes et convergent vers une même limite L .

Cette limite L dépend-elle de la solution initiale choisie ? Donne-t-elle bien le maximum cherché du critère ?

Nous ne disposons pas actuellement de réponses analytiques et la convergence a peut-être lieu vers un optimum local. Ce point qui paraît fondamental constitue souvent la faiblesse des méthodes itératives.

4) Analyse de la solution "optimale"

Notons $\Delta_i d_i^*$ et Z_j^* les variables codées et les facteurs finalement retenus après un nombre suffisant d'itérations.

$$\begin{aligned} \text{a) Maximum } F(d_1^*, \dots, d_k^*, Z_1 \dots Z_m) = \\ Z_1, \dots, Z_m \\ \text{centrées réduites} \\ \text{non corrélées} \\ = F(d_1^*, \dots, d_k^*, Z_1^*, \dots, Z_m^*) = \mu^* = L \end{aligned}$$

Les Z_j^* sont les composantes principales, réduites, des variables codées $\Delta_i d_i^*$.

$\frac{\mu_j^*}{k} = \frac{1}{k} \sum_{i=1}^k \text{cor}^2(\Delta_i d_i^*, Z_j^*)$ représente le pourcentage de variance expliquée par Z_j^* .

$\frac{\mu^*}{k} = \sum_{j=1}^m \frac{\mu_j^*}{k}$ représente le pourcentage de variance expliquée par les m composantes principales $Z_1^*, \dots, Z_m^*$.

Z_j^* est interprétée à l'aide des corrélations entre les $\Delta_i d_i^*$ et les Z_j^* .

Les observations e peuvent être représentées dans R^m à l'aide des vecteurs $Z^*(e) = (Z_1^*(e), \dots, Z_m^*(e))$.

Les modalités (i, l) (modalité l de la variable D_i) peuvent être représentées dans R^m par $\bar{Z}^*(i, l)$ centre de gravité des $Z^*(e)$, e appartenant à $D_i^{-1}(l)$:

$$\text{b) Maximum } \sum_{j=1}^m \text{cor}^2(\Delta_i d_i, Z_j^*) = \sum_{j=1}^m \text{cor}^2(\Delta_i d_i^*, Z_j^*) = \lambda_i^*$$

et :

$$\sum_{i=1}^k \lambda_i^* = L.$$

λ_i^* représente la qualité de représentation de la variable $\Delta_i d_i^*$ à l'aide des composantes principales $Z_1^*, \dots, Z_m^*$ (dans le langage de l'analyse factorielle en facteur commun et spécifique λ_i^* est la communauté de la variable $\Delta_i d_i^*$).

c) Comme en analyse en composantes principales il est possible de représenter dans un plan (cercle des corrélations) les variables D_i à l'aide des couples

$$(\text{cor}(\Delta_i d_i^*, Z_1^*), \text{cor}(\Delta_i d_i^*, Z_2^*)).$$

2 – CHOIX DE LA SOLUTION INITIALE

Il est possible de choisir arbitrairement des $d_i^{(0)}$ vérifiant les contraintes ou bien de rechercher des $Z_j^{(0)}$ à partir de l'analyse canonique généralisée, puis des $d_i^{(0)}$ à l'aide de la proposition 2.

En effet, nous pouvons considérer le problème :

$$\begin{aligned} & \text{Maximiser} \quad \sum_{i=1}^k \sum_{j=1}^m \text{cor}^2 (\Delta_i d_{ij}, Z_j) \\ & d_{ij} \\ & i = 1, \dots, k \\ & j = 1, \dots, m \\ & Z_1 \dots Z_m \\ & \text{centrées réduites} \\ & \text{non corrélées} \end{aligned}$$

La proposition suivante, due à J.D. Carroll (5) donne la solution de ce problème.

Proposition 5

Le maximum, sur des $d_{ij} \in \mathbb{R}^{p_i-1}$ et des $Z_1, \dots, Z_m$ centrées réduite et non corrélées de $\sum_{i=1}^k \sum_{j=1}^m \text{cor}^2 (\Delta_i d_{ij}, Z_j)$ est atteint pour les Z_j vecteurs propres normés à $\sqrt{n}$ de $Q = \sum_{i=1}^k \Delta_i (\Delta_i' \Delta_i)^{-1} \Delta_i'$ correspondant aux m plus grandes valeurs propres $\theta_1, \dots, \theta_m$ de Q et pour $d_{ij} = (\Delta_i' \Delta_i)^{-1} \Delta_i' Z_j$.

Démonstration

1) Les Z_j étant centrées-réduites par hypothèse, nous avons (résultat classique de régression multiple) :

le Maximum $\text{cor}^2 (\Delta_i d_{ij}, Z_j)$ est atteint pour $d_{ij} = (\Delta_i' \Delta_i)^{-1} \Delta_i' Z_j$ et est égal à $\frac{1}{n} Z_j' \Delta_i (\Delta_i' \Delta_i)^{-1} \Delta_i' Z_j$.

2) Nous pouvons donc écrire :

$$\begin{aligned} & \text{maximum} \quad \sum_{i=1}^k \sum_{j=1}^m \text{cor}^2 (\Delta_i d_{ij}, Z_j) = \\ & d_{ij} \\ & Z_1 \dots Z_m \\ & \text{centrées réduites} \\ & \text{non corrélées} \end{aligned}$$

$$\begin{aligned}
&= \underset{\substack{Z_1 \dots Z_m \\ \text{centrées réduites} \\ \text{non corrélées}}}{\text{maximum}} \quad \frac{1}{n} \sum_{j=1}^m Z_j' \sum_{i=1}^k \Delta_i (\Delta_i' \Delta_i)^{-1} \Delta_i' Z_j \\
&= \underset{\substack{Z_1 \dots Z_m \\ \text{centrées réduites} \\ \text{non corrélées}}}{\text{maximum}} \quad \frac{1}{n} \sum_{j=1}^m Z_j' Q Z_j \quad \text{où} \quad Q = \sum_{i=1}^k \Delta_i (\Delta_i' \Delta_i)^{-1} \Delta_i'
\end{aligned}$$

Ce problème est classique et le maximum est obtenu pour les Z_j vecteurs propres de Q correspondant aux m plus grandes valeurs propres $\theta_1, \dots, \theta_m$ de Q et de normes $\sqrt{n}$.

Le 1) nous donne les $d_{ij} : d_{ij} = (\Delta_i' \Delta_i)^{-1} \Delta_i' Z_j$

$$\begin{aligned}
\text{et} \quad & \underset{\substack{Z_1 \dots Z_m \\ \text{centrées réduites} \\ \text{non corrélées}}}{\text{maximum}} \quad \frac{1}{n} \sum_{j=1}^m Z_j' Q Z_j = \sum_{j=1}^m \theta_j .
\end{aligned}$$

Il existe cependant un inconvénient pratique à la méthode de Carroll : les dimensions de la matrice à diagonaliser sont en effet $n \times n$ où n est le nombre d'observations. Or les Z_j peuvent être obtenus en diagonalisant une matrice de taille $\sum_{i=1}^k (p_i - 1) \times \sum_{i=1}^k (p_i - 1)$.

Posons :

$$\begin{aligned}
\Delta &= [\Delta_1, \dots, \Delta_k] \\
Y &= \begin{bmatrix} \Delta_1' \Delta_1 & 0 & \\ 0 & \ddots & \\ & & \Delta_k' \Delta_k \end{bmatrix}
\end{aligned}$$

Nous pouvons remarquer que :

$$Q = \sum_{i=1}^k \Delta_i (\Delta_i' \Delta_i)^{-1} \Delta_i' = \Delta Y^{-1} \Delta'$$

La proposition suivante montre que les Z_j peuvent être obtenues en diagonalisant la matrice $Y^{-1} \Delta' \Delta$.

Proposition 6 ()*

Si b_j est le vecteur propre de $Y^{-1} \Delta' \Delta$ associé à la valeur propre θ_j et tel que $\frac{\theta_j}{n} b_j' Y b_j = 1$, $Z_j = \Delta b_j$ est le vecteur propre de Q associé à la valeur propre θ_j et de norme $\sqrt{n}$.

Démonstration

Si $Y^{-1} \Delta' \Delta b_j = \theta_j b_j$ nous avons :

$$\Delta Y^{-1} \Delta' \Delta b_j = \theta_j \Delta b_j, \quad \text{d'où} \quad Q Z_j = \theta_j Z_j$$

et

$$\begin{aligned} Z_j' Z_j &= b_j' \Delta' \Delta b_j = b_j' Y Y^{-1} \Delta' \Delta b_j \\ &= \theta_j b_j' Y b_j = n. \end{aligned}$$

Le vecteur $d_{.j} = \begin{bmatrix} d_{1j} \\ \vdots \\ d_{kj} \end{bmatrix}$ est ensuite particulièrement simple à calculer.

Proposition 7

Le vecteur $d_{.j}$ est égal à $\theta_j b_j$

Démonstration

Nous avons $d_{ij} = (\Delta_i' \Delta_i)^{-1} \Delta_i' Z_j$ d'où :

$$\begin{aligned} d_{.j} &= \begin{bmatrix} \Delta_1' \Delta_1 & & 0 \\ & \ddots & \\ 0 & & \Delta_k' \Delta_k \end{bmatrix}^{-1} \begin{bmatrix} \Delta_1' \\ \vdots \\ \Delta_k' \end{bmatrix} Z_j \\ &= Y^{-1} \Delta' Z_j \\ &= Y^{-1} \Delta' \Delta b_j \\ &= \theta_j b_j \end{aligned}$$

d'après la définition de b_j .

Recherche de la solution initiale

A partir des Z_j obtenues à l'aide des propositions précédentes et en utilisant la proposition 2, nous obtenons les vecteurs $d_i^{(0)}$.

(*) Nous retrouvons dans cette proposition les résultats de J. de Leeuw [8] et de G. Saporta [11].

3 – EXTENSION DE LA METHODE A L'ETUDE D'UN ENSEMBLE DE VARIABLES NOMINALES OU NUMERIQUES

Nous disposons maintenant de variables $D_1, \dots, D_k$ nominales et de variables $X_1, \dots, X_l$ numériques.

Le critère s'écrit :

$$\begin{array}{l} \text{Maximiser} \quad \left(\sum_{i=1}^k \sum_{j=1}^m \text{cor}^2(\Delta_i d_i, Z_j) + \sum_{i=1}^l \sum_{j=1}^m \text{cor}^2(X_i, Z_j) \right) \\ d_1, \dots, d_k \\ \text{vérifiant les} \\ \text{contraintes} \\ Z_1 \dots Z_m \\ \text{centrées-réduites} \\ \text{non corrélées} \end{array}$$

A l'étape t du processus itératif les $Z_j^{(t)}$ sont les composantes principales des variables $\Delta_1 d_1^{(t-1)}, \dots, \Delta_k d_k^{(t-1)}, X_1^*, \dots, X_l^*$ (où X_i^* représente la variable X_i centrée réduite). Puis les $d_i^{(t)}$ sont obtenus comme précédemment.

Tous les autres résultats restent valables.

Pour la recherche d'une solution initiale il suffira de considérer à la place des matrices Δ et Y les matrices

$$\begin{aligned} \tilde{\Delta} &= [\Delta_1, \dots, \Delta_k, X_1^*, \dots, X_l^*] \\ \text{et} \quad \tilde{Y} &= \begin{bmatrix} \Delta_1' \Delta_1 & & & & 0 \\ & \ddots & & & \\ & & \Delta_k' \Delta_k & & \\ 0 & & & X_1^{*'} X_1^* & \\ & & & & \ddots \\ & & & & & X_l^{*'} X_l^* \end{bmatrix} \\ &= \begin{bmatrix} \Delta_1' \Delta_1 & & & & 0 \\ & \ddots & & & \\ & & \Delta_k' \Delta_k & & \\ 0 & & & n & \\ & & & & \ddots \\ & & & & & n \end{bmatrix} \end{aligned}$$

et de procéder comme précédemment.

4 – UN EXEMPLE

Jean-Pierre Pagès a mis à notre disposition une enquête d'opinion effectuée par le C.E.A. Vingt questions furent extraites du questionnaire auquel répondirent 90 petits commerçants de la région de Grenoble.

Voici les 20 questions.

- Q.1. La libération de l'avortement est une bonne chose (AVOR)
- Q.2. Il faut continuer à construire des centrales nucléaires (CENT)
- Q.3. Les médecins méritent notre confiance (MEDE)
- Q.4. Dieu existe (DIEU)
- Q.5. La force de frappe est indispensable à la France (FRAP)
- Q.6. Les centrales nucléaires sont dangereuses (DCEN)
- Q.7. Il faut supprimer la peine de mort (MORT)
- Q.8. La publicité rend service (PUBL)
- Q.9. Il y a trop de travailleurs immigrés en France (TRIM)
- Q.10. La France devrait supprimer son aide aux pays sous-développés (PASD)
- Q.11. Avec le temps le progrès technique résout tous les problèmes (PROT)
- Q.12. Les étudiants vivent en parasite de la Société (ETUD)
- Q.13. Il faut encourager la natalité (NATA)
- Q.14. La famille doit rester la cellule de base de la Société (FAM)
- Q.15. La prostitution doit être réprimée sévèrement (PROS)
- Q.16. Si les conditions de sécurité sont insuffisantes dans une entreprise, il faut inculper le chef d'entreprise (SECU)
- Q.17. Les très grosses entreprises doivent être nationalisées (NATI)
- Q.18. On devrait arrêter les importations de pétrole en provenance des pays arabes (PETR)
- Q.19. Le M.L.F. mérite d'être soutenu (M.L.F.)
- Q.20. Il est nécessaire de censurer certains films ou certains écrits (CENS).

Chaque personne interrogée devait exprimer son opinion sur ces propositions en donnant une note sur une échelle à 5 points (1 = tout à fait d'accord, 2 = plutôt d'accord, 3 = indifférent, 4 = plutôt pas d'accord, 5 = pas du tout d'accord).

Nous avons considéré ces variables comme nominales et utilisé le programme PRINQUAL cité précédemment pour analyser ces données.

Nous avons choisi l'option "analyse canonique généralisée" pour définir les codages initiaux. L'algorithme a ensuite convergé en 4 itérations pour une analyse à 3 facteurs. L'analyse à 2 facteurs a donné un pourcentage de variance expliquée de 33 %. Pour 3 facteurs nous avons obtenu 42 %. Les résultats étant cependant plus clairs pour l'analyse à 2 facteurs c'est celle-ci que nous allons décrire.

Nous pouvons représenter les variables à l'aide des corrélations variables-facteurs, (figure 1).

Ces deux dimensions s'interprètent clairement :

- le premier facteur est essentiellement corrélié à des opinions de nature sociologique : famille, avortement, dieu, publicité, MLF, censure, peine de mort, etc. ;
- le deuxième facteur est lui, plus particulièrement corrélié à des opinions de nature économique et technologique : natalité, progrès technique, médecine, pétrole, centrale nucléaire, force de frappe, aide aux pays en voie de développement.

Nous pouvons également représenter les différentes modalités des variables.

Ainsi, nous pouvons déterminer différentes configurations d'opinion.

Par exemple, en ce qui concerne l'opinion sur les centrales nucléaires, nous observons les schémas suivants (figure 2) :

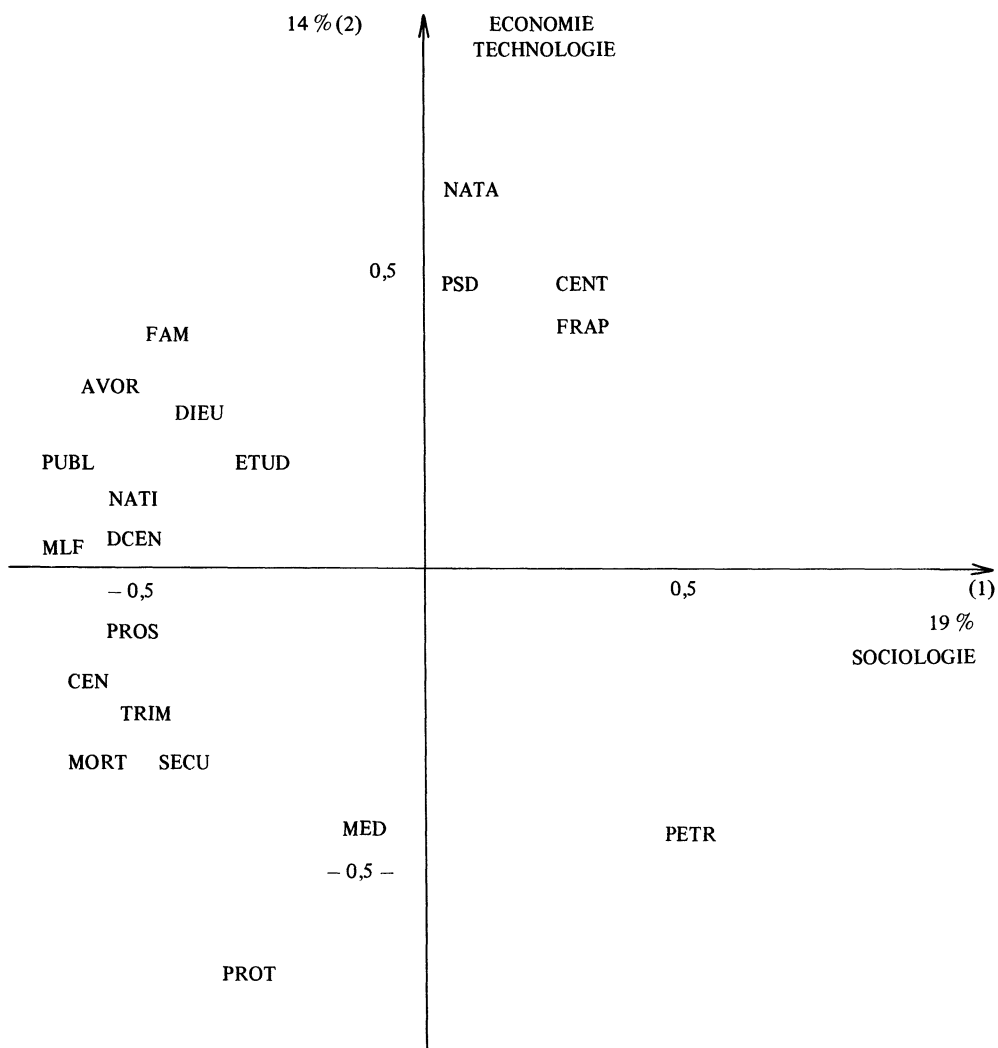


Figure 1

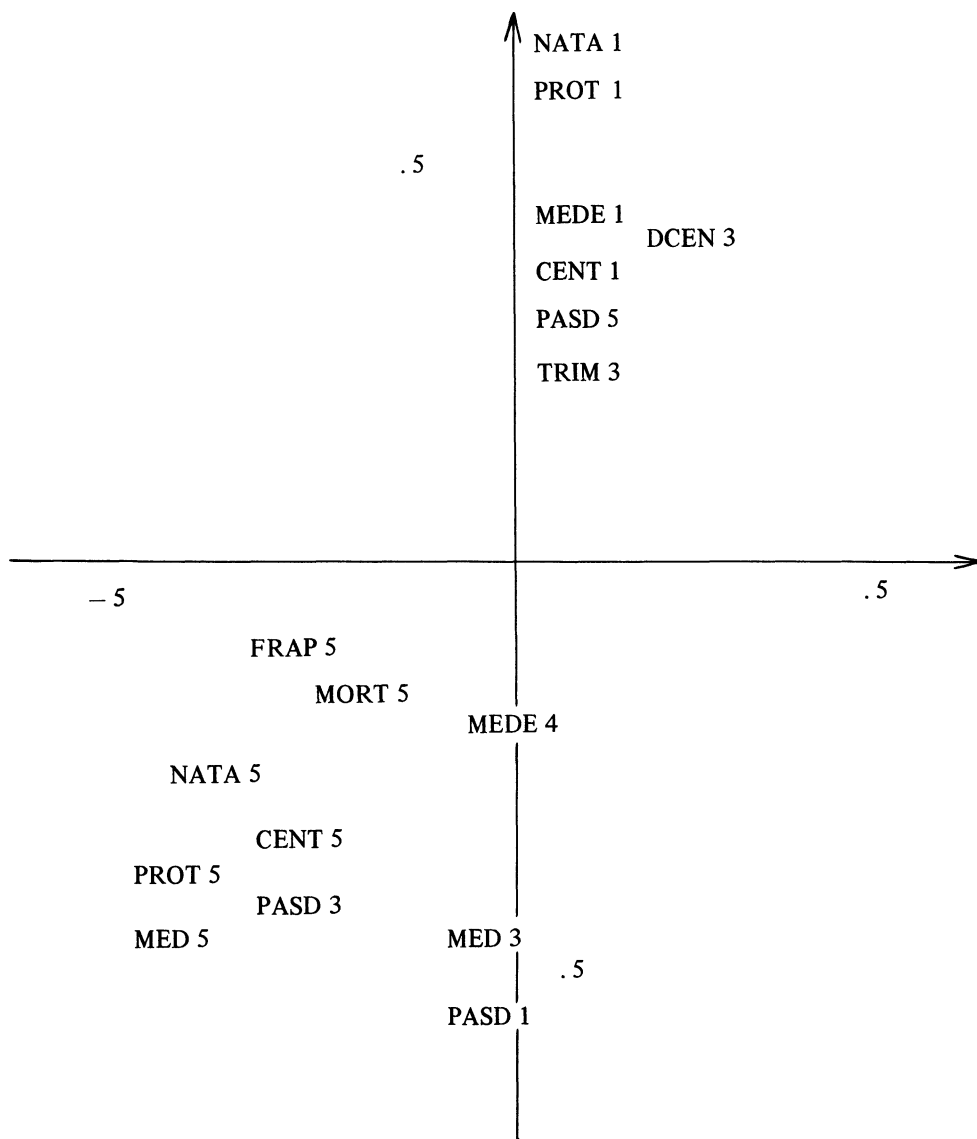


Figure 2

(NATA 1 = Modalité 1 de la variable NATA)

Le quadrant en haut à droite nous indique que les personnes tout à fait favorables aux centrales nucléaires sont également tout à fait favorables au développement de la natalité, font confiance au progrès technique, à la médecine, n'ont pas d'opinion sur le danger des centrales nucléaires et sur la présence des travailleurs immigrés en France et pensent que la France doit continuer à aider les pays sous-développés.

Par contre, d'après le quadrant en bas à gauche, les personnes tout à fait défavorables aux centrales nucléaires sont aussi défavorables au développement de

la natalité, ne font pas confiance au progrès technique, à la médecine, pensent que la France n'a pas besoin de force de frappe, qu'elle devrait plutôt supprimer son aide aux pays sous développés et, enfin, sont favorables à la peine de mort.

Ces interprétations sont basées sur des proximités entre centres de gravité et il faudrait pour les confirmer retourner à l'analyse des dossiers individuels.

Disons, cependant, pour conclure que cette méthode a permis de structurer les données tant au niveau des variables que des modalités.

REFERENCES

- [1] BERTIER P. et TENENHAUS M. — Les méthodes de "Conjoint Measurement" Séminaire AFCET (1974). Paru dans l'*Analyse des Données en Marketing*, J.M. Bouroche, Masson, 1977.
- [2] BOUROCHE J.M. et TENENHAUS M. — Méthode de typologie sur variables hétérogènes. *Note de Travail* n° 10, Coref 1976.
- [3] BOUROCHE J.M., SAPORTA G., TENENHAUS M. — Méthode et exemples d'analyse des Données qualitatives. *Note de Travail*, n° 13, Coref 1976.
- [4] CAILLIEZ F. et PAGES J.P. — Introduction à l'analyse des données. SMASH, 1976.
- [5] CARROLL J.D. — Generalisation of canonical correlation analysis to three or more sets of variables. Proceedings, 76 th Annual Convention, APA, 1968.
- [6] CAZES P., BONNEFOUS S., BAUMERDER A. et PAGES J.P. — Description cohérente des variables qualitatives prises globalement et de leurs modalités. *Statistique et Analyse des Données*, n° 2, 1976.
- [7] CAZES P., BAUMERDER A., BONNEFOUS S. et PAGES J.P. — Codage et analyse des tableaux logiques. Introduction à la pratique des variables qualitatives. *Cahiers du B.U.R.O. (à paraître)*.
- [8] DE LEEUW J. — Canonical analysis of categorical data. Université de Leyde, 1973.
- [9] DE LEEUW J. — Homals. Hand out "Optimal scaling symposium", Spring meeting, Psychometric Society, 1976.
- [10] MASSON M. — Processus linéaire et analyse de données non linéaire. Thèse d'état, Paris, 1974.
- [11] SAPORTA G. — Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Thèse de 3^e cycle, Paris, 1975.
- [12] SAPORTA G. — Quelques applications des opérateurs d'Escouffier au traitement des variables qualitatives. *Statistique et Analyse des Données*, n° 1, 1976.
- [13] SAPORTA G. — Le traitement des variables qualitatives par codage. *Note de Travail*, n° 11, Coref, 1976.

- [14] TAKANE Y., YOUNG F.W. et DE LEEUW J. — How to use Principals, A program for principal component analysis of qualitative data which uses the alternating least square method. University of North Carolina, 1975.
- [15] TENENHAUS M. et VACHETTE J.L. — Prinqual, un programme d'analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Cahier de recherche CESA*, 1977. Projet de recherche COREF/DGRST, 1976.
- [16] COREF-DGRST. — Rapport de recherche sur le traitement des variables qualitatives. Février 1977.