



# Problemes d'optimisation relatifs aux tableaux multiples .Point de vue de la classification automatique

Henri Ralambondrainy

## ► To cite this version:

Henri Ralambondrainy. Problemes d'optimisation relatifs aux tableaux multiples .Point de vue de la classification automatique. RR-0576, INRIA. 1986. inria-00075978

**HAL Id: inria-00075978**

**<https://inria.hal.science/inria-00075978>**

Submitted on 24 May 2006

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



CENTRE DE ROCQUENCOURT

Institut National  
de Recherche  
en Informatique  
et en Automatique

Domaine de Voluceau  
Rocquencourt  
BP 105  
78153 Le Chesnay Cedex  
France  
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 576

**PROBLÈMES D'OPTIMISATION  
RELATIFS  
AUX TABLEAUX MULTIPLES  
POINT DE VUE DE LA  
CLASSIFICATION AUTOMATIQUE**

**Henri RALAMBONDRAIN**

**Octobre 1986**

**PROBLEMES D'OPTIMISATION RELATIFS AUX TABLEAUX MULTIPLES  
POINT DE VUE DE LA CLASSIFICATION AUTOMATIQUE**

**THE STUDY OF HETEROGENEOUS DATA BY THE CLUSTERING METHODS**

**Henri RALAMBONDRAINY**

**I.N.R.I.A. Domaine de Voluceau Rocquencourt B.P.105 78380 LE CHESNAY**

**Résumé:**

*Un tableau multiple recense les valeurs prises par plusieurs groupes de variables relativement à un même ensemble d'individus. Les méthodes factorielles usuelles ne sont pas adaptées à l'analyse de tableaux multiples et des extensions ont été proposées par différents auteurs. Nous montrons la similitude des problèmes d'optimisation résolus par les méthodes factorielles et les méthodes de Classification Automatique maximisant l'inertie inter-classes dans l'étude de tels tableaux. Nous approfondissons ensuite l'étude des tableaux multiples par la méthode des Nuées Dynamiques (variante du centre de gravité).*

**Abstract:**

*The traditional multivariate analysis are not fitted to treat data which contain different kind of variables (heterogeneous data). So some authors proposed extension to factorial analysis methods. We study here the relationship between the factorial analysis and the clustering methods generalised to heterogeneous data. We propose new approach to treat heterogeneous data by clustering methods "Nuees Dynamiques".*



## I - INTRODUCTION

Un tableau multiple recense les valeurs prises par plusieurs groupes de variables relativement à un même ensemble d'individus. Le tableau est dit "mixte" s'il comporte des variables de type quantitatif et qualitatif. Les méthodes factorielles usuelles ne sont pas adaptées à l'analyse de tels tableaux et des extensions ont été proposées par différents auteurs. Citons, entre autres :

- L'Analyse des Correspondances Multiples qui généralise l'Analyse des Correspondances Simples aux tableaux de variables qualitatives, ces dernières pouvant être éventuellement pondérées pour l'analyse [Caz80].

- L'Analyse de la structure intra dans la méthode STATIS proposée par Escoufier [Esc80] permet l'étude de plusieurs groupe de variables de type quantitatif.

- L'Analyse Factorielle Multiple d'Escoufier-Pagès traite les tableaux mixtes [EsP84] en imposant que la métrique relative aux individus soit diagonale.

Ces différentes méthodes sont des Analyses en Composantes Principales où chaque groupe de variables est pondéré par un coefficient positif qui dépend de la méthode. Elles peuvent aussi être considérées comme des Analyses Canoniques Généralisées. Ainsi l'Analyse des Correspondances Multiples est une Analyse Canonique Généralisée, recherchant des variables bien liées au sens de la corrélation multiple (liaison  $R^2$ ) avec chaque groupe de variables. L'Analyse Factorielle Multiple peut être vue comme une Analyse Canonique Généralisée fondée sur une autre mesure de liaison, notée  $L^2$ , qui tient compte de l'inertie des différents groupes de variables.

Dans un premier temps, nous mettons en évidence les problèmes d'optimisation généraux sous-jacents à ces différentes méthodes en ne faisant aucune hypothèse sur la nature des variables, ni sur le type de métrique relatif aux individus. Benzécri présente les méthodes factorielles comme la recherche de la meilleure approximation de rang  $K$  fixé d'un tenseur [Ben73]. Notre démarche [Ral86] consiste à reprendre, dans le cadre des tableaux multiples, cette présentation. L'Analyse en Composantes Principales pondérée apparaîtra alors comme l'interprétation du problème précédent du point de vue des "individus" tandis que l'Analyse Canonique Généralisée, au sens d'une mesure de liaison  $L$  que nous proposons généralisant les liaisons  $R^2$  et  $L^2$ , sera le point de vue des "variables".

Dans un second temps, nous approfondirons l'étude des tableaux multiples par la méthode de Classification Automatique des Nuées Dynamiques (variante du centre de gravité) [Did79]. Nous faisons le lien avec les méthodes factorielles en remarquant qu'une partition est une variable qualitative et en montrant qu'une méthode de Classification Automatique maximisant l'inertie inter-classes est une Analyse Canonique Généralisée recherchant une variable qualitative bien liée, au sens de la mesure de liaison  $L$ , aux différents groupes de variables constituant le tableau multiple.

Le paragraphe II introduit les cadres de référence, définit la mesure d'information associée à un tableau simple et présente les méthodes factorielles comme des techniques de réduction de cette mesure d'information. Les résultats sont ensuite étendus à un tableau multiple et les paragraphes III et IV concernent l'étude d'un tableau multiple respectivement par l'Analyse en Composantes Principales et l'Analyse Canonique Généralisée. Le paragraphe V aborde le problème du point de vue de la Classification Automatique et présente un exemple d'application de la méthode proposée.

## II - LES CADRES DE REFERENCES POUR L'ETUDE D'UN TABLEAU SIMPLE

### II.1 - Les notations

On note les ensembles  $I = \{1, \dots, n\}$ ,  $J = \{1, \dots, p\}$  et le tableau des données  $X = \{x_{ij} \mid j \in J, i \in I\}$ . L'ensemble des individus  $\{x_i = (x_{ij} \mid j \in J), i \in I\}$  forme un nuage  $N_E^I$  dans l'espace  $E \approx \mathbb{R}^p$  et l'ensemble des variables  $\{x_j = (x_{ij} \mid i \in I), j \in J\}$ , un nuage  $N_F^J$  dans l'espace  $F \approx \mathbb{R}^n$ . Les espaces  $E$  et  $F$  sont respectivement munis d'une métrique quelconque  $M = \{M_{ij} \mid j, j' \in J\}$  et de la métrique diagonale des poids  $D_p = \{p_i \mid i \in I\}$ . Le tableau  $X$  est supposé centré. On adopte la convention d'écriture de représenter l'ensemble et son cardinal par la même lettre.

### II.2 - Opérateurs représentatifs d'un tableau

Il existe plusieurs possibilités de représentation d'un tableau ou d'une matrice rectangulaire par un vecteur d'un espace euclidien. On considérera les opérateurs suivants :

• L'application linéaire  $U = XM^tXD_p = WD_p$ , élément de  $L(F, F)$  isomorphe à l'espace produit tensoriel  $F^* \otimes F$ , appelée "opérateur variable" et l'application linéaire  $Z = {}^tXD_pXM = VM$ , élément de  $L(E, E)$  isomorphe à l'espace produit tensoriel  $E^* \otimes E$ , qui est appelée "opérateur individu" (cf figure 1). Les opérateurs  $U$  et  $Z$ , dits d'Escoufier, sont caractéristiques du triplet  $(X, M, D_p)$  et leurs éléments propres engendrent respectivement les composantes principales et les axes principaux d'inertie du nuage des individus [CaP76] [Esc80].

• Le tenseur  $X_{E \otimes F}$ , associé au tableau  $X$ , élément de l'espace produit tensoriel  $E \otimes F$  muni du produit scalaire  $M \otimes D_p$  induit par les produits scalaires  $M$  et  $D_p$ .

Dans le premier cas, deux tableaux différents peuvent être équivalents si les opérateurs associés sont égaux c.a.d. les nuages des individus ont mêmes axes d'inertie tandis que dans le deuxième cas, deux tableaux sont équivalents si les tenseurs associés sont égaux c.a.d. lorsque les tableaux sont identiques.

Nous définirons ensuite une mesure d'information  $I(X)$  relatif au triplet  $(X, M, D_p)$  et étudierons la réduction de  $I(X)$  dans les différents espaces de référence choisis.

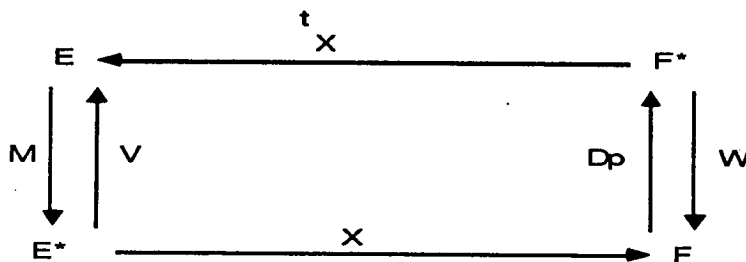


Figure 1 : Schéma de dualité

### II.3 - L'espace $F^* \otimes F$ et l'opérateur $U$

Compte tenu de l'isomorphisme  $L(F, F) \approx F^* \otimes F$ , nous utiliserons la représentation tensorielle des applications linéaires. Les aspects géométriques sont en effet bien mieux mis en évidence et les calculs facilités. Rappelons que le produit tensoriel de rang 1 entre une forme linéaire  $\alpha$  appartenant à  $F^*$  et un vecteur  $v$  de  $F$  est l'application linéaire  $\alpha \otimes v$  de  $L(F, F)$  telle que pour  $x \in F$ , on a :

$$\alpha \otimes v(x) = \alpha(x) v.$$

Si on note  $\{f_i \mid i \in I\}$  la base canonique de  $F$ ,  $\{f_i^* \mid i \in I\}$  la base duale de  $F^*$ , alors l'ensemble  $\{f_i^* \otimes f_{i'} \mid i, i' \in I\}$  est une base de  $F^* \otimes F$ . Soit un tenseur  $T \in F^* \otimes F$  qui s'exprime dans cette base de la manière suivante:

$$T = \sum_{i, i'} T_{ii'} f_i^* \otimes f_{i'}$$

Par définition la trace du tenseur  $T$  notée  $\text{Trace} T$  est le nombre unique, indépendant de la représentation tensorielle de  $T$  choisie, défini en remplaçant chaque produit tensoriel  $f_i^* \otimes f_{i'}$  par la contraction :

$$\langle f_i^*, f_{i'} \rangle = f_i^*(f_{i'}), \text{ ainsi : } \text{Trace } T = \sum_{i, i'} T_{ii'} f_i^*(f_{i'}) = \sum_i T_{ii}$$

On retrouve la définition de la trace relative à une matrice. On vérifie facilement que la Trace, que l'on notera désormais  $\langle, \rangle$ , est une forme linéaire sur  $F^* \otimes F$ . L'opérateur variable  $U$  est Dp-symétrique et semi-défini positif. On montre dans [Esc80] que la Trace du produit de composition est un produit scalaire sur le sous-espace  $U$  des opérateurs Dp-symétriques de  $F^* \otimes F$ . Soient  $U, U'$  deux éléments de  $U$  le produit scalaire entre ces opérateurs s'écrit:

$$\langle U, U' \rangle = \text{Trace}(U \circ U') = \text{Trace } U \cdot U'$$

en représentant l'application linéaire et la matrice associée par la même lettre. Soit  $v$  un vecteur de  $F$ , on note  $v^*$  la forme linéaire de  $F^*$  telle que :  $v^* = \langle \cdot, v \rangle_{Dp}$ . Si  $v, v', w, w'$  sont des vecteurs de  $F$  alors, par définition, on a :

$$\langle v^* \otimes v', w^* \otimes w' \rangle = \text{Trace}(v^* \otimes v' \circ w^* \otimes w') = \langle v, w' \rangle_{Dp} \langle v', w \rangle_{Dp}$$

On note  $v$  un vecteur normé; le Dp-projecteur  $A_v$  associé à  $v$  a alors pour expression :

$$A_v = v^* v_{Dp} = v^* \otimes v.$$

Si  $F_K$  est un espace vectoriel de dimension  $K$ ,  $\{v_k \mid k \in K\}$  une base Dp-orthonormée de  $F_K$ , alors notons  $A_F^K$  le Dp-projecteur associé à  $F_K$ . L'expression tensorielle de  $A_F^K$  est :

$$A_F^K = \sum_k v_k^* v_{kDp} = \sum_k v_k^* \otimes v_k.$$

Nous allons donner quelques expressions tensorielles de l'opérateur variable  $U$ . L'opérateur  $U$  s'écrit matriciellement:

$$U = XM^t X D_p = \sum \{ M_{ij}^t x_j^t x_j D_p \mid j, j' \in J \}$$

par suite pour  $x \in F$ , on a :

$$U(x) = \sum \{ M_{ij}^t x_j^t x_j D_p x \mid j, j' \in J \} = \sum \{ M_{ij}^t \langle x_j, x \rangle_{D_p} x_j^t \mid j, j' \in J \}$$

On a donc une représentation tensorielle de  $U$  :

$$U = \sum \{ M_{ij}^t x_j^t \otimes x_j^t \mid j, j' \in J \}$$

On note  $\{\phi_r \mid r \in R\}$  les vecteurs propres de  $U$  associés aux valeurs propres non nulles, on les complète par des vecteurs du noyau de  $U$  pour former une base  $D_p$ -orthonormée de  $F$  :  $\{\phi_i \mid i \in I\}$ . La famille de vecteurs  $\{\phi_i^* \otimes \phi_i \mid i, i' \in I\}$  forme une base de  $F^* \otimes F$  dans laquelle la matrice associée à  $U$  est diagonale. Les valeurs propres de  $U$ , supposées classées par ordre décroissant,  $\{\lambda_i \mid i \in I\}$  sont les éléments diagonaux de cette matrice. On a donc la représentation tensorielle de  $U$  suivante :

$$U = \sum_r \lambda_r \phi_r^* \otimes \phi_r$$

Calculons la norme de l'opérateur  $U$ , on a :

$$\|U\|^2 = \text{Trace } U^2 = \sum_r \lambda_r^2$$

#### II.4 - L'espace $E^* \otimes E$ et l'opérateur $Z$

L'opérateur individu  $Z$  est  $M$ -symétrique et le sous-espace  $\mathbf{Z}$  des applications  $M$ -symétriques de  $E^* \otimes E$  est muni du produit scalaire  $\text{Trace}$  du produit de composition. Un calcul analogue au paragraphe précédent permet d'obtenir les représentations tensorielles de  $Z$  suivantes :

$$Z = \sum_i p_i x_i^* \otimes x_i$$

$$Z = \sum_r \lambda_r \Psi_r^* \otimes \Psi_r$$

où les vecteurs  $M$ -orthonormés  $\{\Psi_r \mid r \in R\}$  sont les vecteurs propres de  $Z$  relatifs aux valeurs propres non nulles. Notons que les formes linéaires  $x_i^*$  et  $\Psi_r^*$  sont ici relatives à la métrique  $M$  :  $x_i^* = \langle \cdot, x_i \rangle_M$  et  $\Psi_r^* = \langle \cdot, \Psi_r \rangle_M$ .

La norme au carré de  $Z$  est égale à celle de  $U$ , en effet, on a :

$$\|Z\|^2 = \text{Trace } Z^2 = \sum_r \lambda_r^2 = \|U\|^2$$

Nous allons donner maintenant quelques expressions géométriques de l'inertie le long d'un espace vectoriel. Soit  $u$  un vecteur de norme unité de  $E$ ,  $\Delta u$  la droite engendré par  $u$ ; l'inertie par rapport à l'espace  $\Delta u^\perp$  orthogonal à  $\Delta u$  du nuage des individus  $N_E^I$  s'écrit :

$$I_{\Delta u}^\perp(N_E^I) = \sum_i p_i \langle x_i, u \rangle_M^2$$

Soit  $A_u = u^* \otimes u$  le  $M$ -projecteur associé à  $u$  calculons :

$$\langle Z, A_u \rangle = \langle Z, u^* \otimes u \rangle = \langle \sum_i p_i x_i^* \otimes x_i, u^* \otimes u \rangle$$

$$\langle Z, A_u \rangle = \sum_i p_i \langle x_i^* \otimes x_i, u^* \otimes u \rangle = \sum_i p_i \langle x_i, u \rangle_M^2 \quad \text{ainsi, on a :}$$

$$I_{\Delta u}^\perp(N_E^I) = \langle Z, A_u \rangle = \langle Z, u^* \otimes u \rangle$$

Soit  $E_K$  un e.v. de dimension  $K$ ,  $\{u_k | k \in K\}$  une base  $M$ -orthonormée de  $E_K$  et  $A_E^K$  le  $M$ -projecteur

associé à  $E_K$  qui a pour expression :  $A_E^K = \sum_k u_k^* \otimes u_k$  (cf §II.3).

En utilisant l'égalité suivante:  $I_E^\perp(N_E^I) = \sum_k I_{\Delta u_k}^\perp(N_E^I)$ , il vient :

$$I_E^\perp(N_E^I) = \sum_k \langle Z, u_k^* \otimes u_k \rangle = \langle Z, \sum_k u_k^* \otimes u_k \rangle, \text{ d'où :}$$

$$I_E^\perp(N_E^I) = \langle Z, A_E^K \rangle.$$

Si l'on choisit  $E_K = E$ , on a  $A_E = e$  le tenseur unité associé à l'application identité  $\text{Id}_E$  et alors :

$$I_G(N_E^I) = \langle Z, e \rangle = \text{Trace } Z$$

$I_G(N_E^I)$  désignant l'inertie totale du nuage  $N_E^I$ .

## II.5 - L'espace $E \otimes F$ et le tenseur $X_{E \otimes F}$

En Analyse des Données, on s'intéresse à l'application linéaire  $X \in L(E^*, F)$  ou  ${}^t X \in L(F^*, E)$  suivant que l'on étudie l'ensemble des variables ou celui des individus. Il est préférable de considérer le tenseur relatif au tableau  $X$  :

$$X_{E \otimes F} = \sum_i \sum_j x_{ij}^j e_i \otimes f_j$$

car  $X$  et  ${}^t X$  ne sont que des expressions différentes de ce tenseur; en effet on a  $L(E^*, F) \approx L(F^*, E) \approx E \otimes F$ . L'espace  $E$  étant muni de la métrique  $M$  et l'espace  $F$  de la métrique  $D_p$ , l'espace  $E \otimes F$  est muni de la métrique  $M \otimes D_p$  telle que si  $x \otimes y, x' \otimes y'$  sont deux éléments de  $E \otimes F$ , on a :

$$\langle x \otimes y, x' \otimes y' \rangle_{M \otimes D_p} = \langle x, x' \rangle_M \langle y, y' \rangle_{D_p}$$

Soient les tenseurs  $T_{E \otimes F} = \sum \{T_{ij}^i e_j \otimes f_i \mid j \in J, i \in I\}$  et  $S_{E \otimes F} = \sum \{S_{ij}^j e_j \otimes f_i \mid j \in J, i \in I\}$ .

On a :

$$\langle T_{E \otimes F}, S_{E \otimes F} \rangle_{M \otimes Dp} = \sum \{T_{ij}^i S_{ij}^j M_{ij}^j p_i \mid i \in I, j, j' \in J\} = \text{Trace}(TM^tSDp) \quad \text{en particulier :}$$

$$\|X_{E \otimes F}\|_{M \otimes Dp}^2 = \langle X_{E \otimes F}, X_{E \otimes F} \rangle_{M \otimes Dp} = \text{Trace}(XM^tXDp) \quad \text{et}$$

$$\|X_{E \otimes F}\|_{M \otimes Dp}^2 = \text{Trace } U = \text{Trace } Z = I_G(N_E^I)$$

## II.6 - Définition et réduction de la mesure d'information associée à un tableau

### Définition II.6.1

La mesure d'information associée à un triplet  $(X, M, Dp)$  est :

$$I(X) = \|X_{E \otimes F}\|_{M \otimes Dp}^2 = \langle Z, e \rangle = \langle U, f \rangle$$

où  $e$  et  $f$  sont les tenseurs unités associés aux applications identités  $\text{Id}_E$  et  $\text{Id}_F$ . La mesure d'information  $I(X)$  s'interprète comme l'inertie du nuage des individus  $I_G(N_E^I)$  dans l'espace  $E$ .

### II.6.1 - Réduction de $I(X)$ dans l'espace euclidien $(E \otimes F, M \otimes Dp)$

On se donne un entier  $K \leq \inf(p, n)$ ; la réduction de la mesure d'information  $I(X)$  est la recherche d'un tenseur  $T_{E \otimes F}$  de rang  $K$  "le plus proche" possible du tenseur  $X_{E \otimes F}$  constituant donc une bonne approximation de  $X_{E \otimes F}$ . Le tenseur  $T_{E \otimes F}$  est solution du problème suivant :

#### Problème II.6.0

$$\begin{aligned} \min & \|X_{E \otimes F} - T_{E \otimes F}\|_{M \otimes Dp}^2 \\ \text{rang } T_{E \otimes F} &= K \end{aligned}$$

C'est la formulation géométrique par Benzécri du problème d'Eckart et Young (la recherche d'un tableau de rang  $K$  le plus proche d'un tableau  $X$  donné). L'Analyse en Composantes Principales est présentée comme la recherche de la meilleure approximation de rang  $K$  du tenseur  $X_{E \otimes F}$ . La solution est fournie par le tenseur:

$$T_{E \otimes F} = \sum_k \sqrt{\lambda_k} \phi_k \otimes \psi_k$$

où  $(\phi_k, \psi_k)$  est le  $k$ -ième couple de facteurs de l'Analyse en Composantes Principales du tableau  $X$ . La formule de reconstitution des données en est la conséquence immédiate.

Nous allons substituer au problème II.6.0 un problème d'optimisation équivalent portant sur les espaces vectoriels associés à un tenseur.

Soient  $\text{TeL}(E^*, F)$  et  ${}^t\text{TeL}(F^*, E)$  les applications linéaires associées au tenseur  $T_{E \otimes F}$ . Comme le rang de  $T_{E \otimes F}$  est  $K$ , l'espace vectoriel  $E_K = {}^tT(F^*)$  est de dimension  $K$ . Le tenseur  $T_{E \otimes F}$  appartient à l'espace  $E_K \otimes F$ . Soit  $X_{E \otimes F}^K$  la projection  $M \otimes Dp$  orthogonale du tenseur  $X_{E \otimes F}$  sur l'e.v.  $E_K \otimes F$ .

On a nécessairement  $X_{E \otimes F} = T_{E \otimes F}$  sinon le tenseur  $X_{E \otimes F}$  de rang  $\leq K$  serait une meilleure approximation de rang  $\leq K$  de  $X_{E \otimes F}$  que  $T_{E \otimes F}$ . Le problème II.6.0 est donc équivalent à la recherche d'un e.v.  $E_K$  de dimension  $K$  tel que la projection de  $X_{E \otimes F}$  sur  $E_K \otimes F$  soit de norme maximale (cf figure 2). Le problème d'optimisation s'énonce comme suit :

**Problème II.6.1**

$$\max_K \|X_{E \otimes F}\|_{M \otimes Dp}^2$$

$$E_K \subset E \quad \dim E_K = K$$

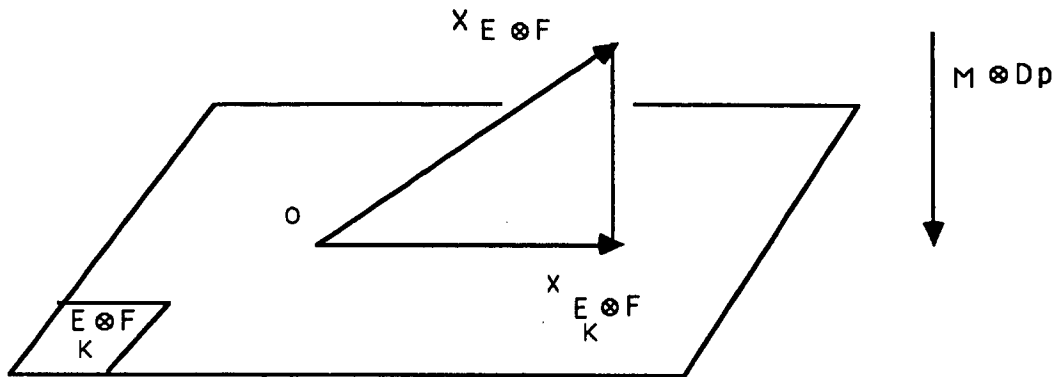


Figure 2: Réduction de  $I(X) = \|X_{E \otimes F}\|_{M \otimes Dp}^2$  dans  $E \otimes F$

**II.6.2- Réduction de  $I(X)$  dans  $E^* \otimes E$  et  $F^* \otimes F$**

Le tenseur  $e$  se décompose comme suit :  $e = A_{E_K} + (e - A_{E_K})$  et il est facile de vérifier que les projecteurs  $A_{E_K}$  et  $e - A_{E_K}$  sont orthogonaux dans  $E^* \otimes E$ . On a donc la décomposition suivante de la mesure d'information  $I(X)$  (cf figure 3) :

$$I(X) = \langle Z, e \rangle = \langle Z, A_{E_K} \rangle + \langle Z, e - A_{E_K} \rangle$$

Il est naturel de rechercher le projecteur  $A_E^K$  solution du problème :

**Problème II.6.2**

$$\max_K \langle Z, A_E^K \rangle$$

$A_E^K$  projecteur de rang K de  $E^* \otimes E$

Symétriquement, on recherchera dans  $F^* \otimes F$  le projecteur  $A_F^K$  solution du problème suivant.

**Problème II.6.3**

$$\max_K \langle U, A_F^K \rangle$$

$A_F^K$  projecteur de rang K de  $F^* \otimes F$

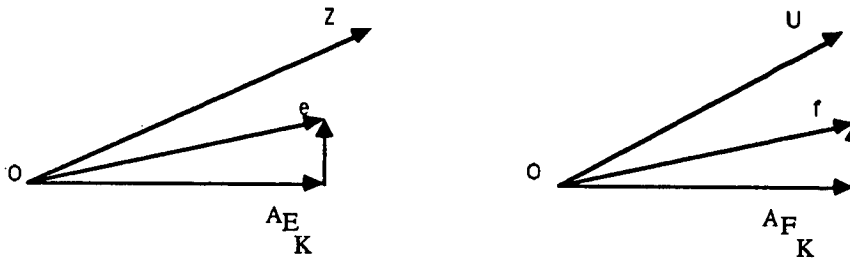


Figure 3 : Réduction de la mesure d'information  $I(X) = \langle U, f \rangle = \langle Z, e \rangle$  dans  $E^* \otimes E$  et  $F^* \otimes F$

**II.6.3.- Lien entre les problèmes de réduction de la mesure d'information  $I(X)$**

Il est donné par la proposition suivante :

**Proposition II.6.3**

Soient  $E_K$  un e.v. de dimension K de E et  $\{u_k, k \in K\}$  une base M-orthonormée de  $E_K$ . On note  $c^k$  le vecteur de F tel que  $c^k = (c_i^k = \langle x_i, u_k \rangle, i \in I)$  et ceci pour  $k \in K$ . La projection  $M \otimes D_p$  orthogonale de  $X_{E \otimes F}$  sur l'e.v.  $E_K \otimes F$  est notée  $X_{E \otimes F}^K$  les égalités suivantes sont vérifiées :

$$1) X_{E \otimes F}^K = \sum_k u_k \otimes c^k$$

$$2) \|X_{E \otimes F}^K\|_{M \otimes D_p}^2 = \sum_k \|c^k\|_{D_p}^2 = \langle Z, A_E^K \rangle$$

*Démonstration*

On sait que  $X_{E \otimes F} = \sum_i \sum_j x_i^j e_j \otimes f_i = \sum_i (\sum_j x_i^j e_j) \otimes f_i = \sum_i x_i \otimes f_i$  et que la

projection de  $X_{E \otimes F}$  sur  $E_K \otimes F$  s'écrit, en utilisant les propriétés du produit tensoriel :

$$X_{E \otimes F} = A_{E \otimes F} (X_{E \otimes F}) = A_{E \otimes F} \otimes \text{Id}_F (\sum_i x_i \otimes f_i) = \sum_i A_{E \otimes F} (x_i) \otimes \text{Id}_F (f_i)$$

Or  $A_{E \otimes F} (x_i) = \sum_k \langle x_i, u_k \rangle_M u_k = \sum_k c_i^k u_k$  et  $\text{Id}_F (f_i) = f_i$  par suite:

$$X_{E \otimes F} = \sum_K \sum_k c_i^k u_k \otimes f_i = \sum_k u_k \otimes (\sum_i c_i^k f_i) = \sum_k u_k \otimes c^k$$

Calculons la norme de ce vecteur :

$$\|X_{E \otimes F}\|_{M \otimes Dp}^2 = \langle \sum_k u_k \otimes c^k, \sum_{k'} u_{k'} \otimes c^{k'} \rangle_{M \otimes Dp} = \sum_k \sum_{k'} \langle u_k, u_{k'} \rangle_M \langle c^k, c^{k'} \rangle_{Dp}$$

Comme les vecteurs  $\{u_k, k \in K\}$  sont  $M$ -orthonormés, on a le résultat :

$$\|X_{E \otimes F}\|_{M \otimes Dp}^2 = \sum_k \|c^k\|_{Dp}^2$$

Calculons l'expression :

$$\langle Z, A_E \rangle = \langle \sum_K p_i x_i^* \otimes x_i, \sum_k u_k^* \otimes u_k \rangle = \sum_i \sum_k p_i \langle x_i^* \otimes x_i, u_k^* \otimes u_k \rangle$$

$$\langle Z, A_E \rangle = \sum_K \sum_i p_i \langle x_i, u_k \rangle_M^2$$

d'où  $\langle Z, A_E \rangle = \sum_K \sum_i p_i (c_i^k)^2 = \sum_k \|c^k\|_{Dp}^2$ . Ce qui achève la démonstration.

Soit  $T_{E \otimes F}$  un tenseur de rang  $K$  solution du problème II.6.0 et  $E_K = {}^t T(F^*)$  l'espace vectoriel de dimension  $K$  associé. On a  $T_{E \otimes F} = T_{E \otimes F}^K$  qui appartient à  $E_K \otimes F$ . Nécessairement  $T_{E \otimes F}$  est la projection  $M \otimes Dp$

orthogonale de  $X_{E \otimes F}$  sur  $E_K \otimes F$  sinon ce tenseur projection serait une meilleure approximation de rang  $\leq K$  que  $T_{E \otimes F}$ . En vertu de l'égalité  $\|T_{E \otimes F}^K\|_{M \otimes Dp}^2 = \langle Z, A_E \rangle$ , le projecteur  $A_E$  est solution du

problème II 6.2. Symétriquement, soit  $F_K = T(E^*)$ , on a  $T_{E \otimes F} = T_{E \otimes F}^K$  qui appartient à  $E \otimes F_K$  et

$T_{E \otimes F}^K$  est la projection  $M \otimes Dp$  orthogonale de  $X_{E \otimes F}$  sur  $E \otimes F_K$  et  $A_F$  est solution du problème II.6.3

pour les mêmes raisons.

L'étude de ces problèmes d'optimisation revient à étudier les invariants de deux formes quadratiques [Ben73] et l'e.v.  $E_K$  (resp.  $F_K$ ) est unique (à condition que les valeurs propres de  $Z$  ne soient pas multiples); il est engendré par les  $K$ -premiers vecteurs propres de  $Z$  (resp.  $U$ ). Si on désigne par  $\{u_k, k \in K\}$  et  $\{v_k, k \in K\}$  les  $K$  premiers vecteurs propres de  $Z$  et  $U$  associés aux valeurs propres non nulles  $\{\lambda_k, k \in K\}$ , le tableau suivant résume les résultats.

|                                 |                                                                                                                           |                                     |                                                                                                           |
|---------------------------------|---------------------------------------------------------------------------------------------------------------------------|-------------------------------------|-----------------------------------------------------------------------------------------------------------|
| <i>Espaces de référence</i>     | $(E \otimes F, M \otimes D_p)$                                                                                            | $(E^* \otimes E, \langle, \rangle)$ | $(F^* \otimes F, \langle, \rangle)$                                                                       |
| <i>Tenseurs représentatifs</i>  | $X_{E \otimes F}$                                                                                                         | $Z$                                 | $U$                                                                                                       |
| <i>Mesure d'information</i>     | $\ X_{E \otimes F}\ ^2_{M \otimes D_p}$                                                                                   | $\langle Z, e \rangle$              | $\langle U, f \rangle$                                                                                    |
| <i>Problèmes d'optimisation</i> | $\min \ X_{E \otimes F} - T_{E \otimes F}\ ^2_{M \otimes D_p}$<br>$\text{rang } T_{E \otimes F} = K$                      | $\max_K \langle Z, A_E \rangle$     | $\max_K \langle U, A_F \rangle$                                                                           |
| <i>Solutions</i>                | $T_{E \otimes F} = \sum_k \sqrt{\lambda_k} u_k \otimes v_k$<br>$\ T_{E \otimes F}\ ^2_{M \otimes D_p} = \sum_k \lambda_k$ | $E_K = \sum_k \Delta u_k$           | $F_K = \sum_k \Delta v_k$<br>$A_E = \sum_{K,k} u_k^* \otimes u_k$<br>$A_F = \sum_{K,k} v_k^* \otimes v_k$ |

Tableau II.6.3

*Remarque :*

Les propriétés consignées dans ce tableau sont vraies pour une métrique  $D_p$  quelconque et un tableau  $X$  non centré.

### III - ETUDE D'UN TABLEAU MULTIPLE PAR L'ANALYSE EN COMPOSANTES PRINCIPALES

#### III.1 - Notations

Un tableau multiple est un ensemble de groupes de variables  $\{X_q \mid q \in Q\}$  centrées. On note :

$J_q$  l'ensemble des indices relatifs aux variables du groupe  $q$   
 $E_q \approx \mathbb{R}^{J_q}$  l'espace partiel des individus mesurés sur les variables du groupe  $q$   
 $M_q$  une métrique définie sur  $E_q$   
 $J = \bigcup_q J_q$  l'ensemble des indices relatifs à toutes les variables.

On considère l'ensemble des triplets  $(X_q, M_q, D_p)$  pour  $q \in Q$  auxquels on associe des coefficients de pondération positifs  $\{c_q, q \in Q\}$ ; les cadres de référence pour un triplet  $(X_q, M_q, D_p)$  sont :

$$(E_q \otimes F, M_q \otimes D_p); \quad (E^* \otimes E_q, \langle, \rangle); \quad (F^* \otimes F, \langle, \rangle)$$

et les tenseurs représentatifs :  $(X_{E_q \otimes F}, Z_q, U_q)$ .

#### III.2 - L'espace des individus $E$ et la métrique pondérée $M$

Pour représenter globalement les individus relativement à l'ensemble des variables, on considère l'espace  $E = \bigoplus_q E_q$  somme directe des espaces  $E_q$ . Dans cet espace, un individu  $x_i$  a  $Q$  composantes  $x_{iq} \in E_q$ :

$x_i = \bigoplus_q x_{iq}$ . Au vecteur  $x_i$  correspond la ligne  $i$  du tableau  $X$  juxtaposition des tableaux  $X_q$  :

$$X = (X_1, \dots, X_Q)$$

On note  $\pi_q$  la projection canonique de  $E$  sur  $E_q$  :

$$\begin{aligned} \pi_q : E &\rightarrow E_q \\ x_i &\rightarrow \pi_q(x_i) = x_{iq} \end{aligned}$$

la projection du nuage  $N_E^I = \{x_i \mid i \in I\}$  dans  $E_q$  est  $N_{E_q}^I = \{x_{iq} \mid i \in I\}$

##### Définition III.2.1

Le produit scalaire  $M$  pondéré par les coefficients positifs  $\{c_q \mid q \in Q\}$  est défini sur  $E \times E$  de la manière suivante. Soient  $x = \bigoplus_q x_q \in E$ ,  $y = \bigoplus_q y_q \in E$ , on a :

$$\langle x, y \rangle_M = \sum_q c_q \langle x_q, y_q \rangle_{M_q}$$

Il est facile de vérifier que  $M$  est un produit scalaire car les coefficients  $c_q$  sont positifs et les  $M_q$  sont des produits scalaires. Matriciellement  $M$  est une matrice diagonale par blocs. Les blocs diagonaux étant

constitués par les matrices relatives aux métriques  $c_q M_q$ .

La forme bilinéaire  $M_q$  définie sur  $E_q \times E_q$  se prolonge aisément sur  $E \times E$  en posant pour  $x, y \in E$   $M_q(x, y) = M_q(\pi_q(x), \pi_q(y))$ , on peut donc écrire la forme bilinéaire  $M$  comme suit :

$$M = \sum_q c_q M_q$$

### Proposition III.2.1

Soit le triplet  $(X, M, D_p)$ , on a les propriétés suivantes relativement aux cadres de référence :

$$1) (E \otimes F = \oplus_q E_q \otimes F, M \otimes D_p = \sum_q c_q M_q \otimes D_p) ; \quad X_{E \otimes F} = \oplus_q X_{E_q \otimes F}$$

$$2) (E^* \otimes E = \oplus_{q, q'} E^* \otimes E_q, \langle, \rangle) ; \quad \pi_q^* \otimes \pi_{q'}(Z) = c_q Z_q$$

où  $\pi_q^* \otimes \pi_{q'}$  est la projection canonique de  $E^* \otimes E$  sur  $E^* \otimes E_q$

$$3) (F^* \otimes F, \langle, \rangle) ; \quad U = \sum_q c_q U_q$$

$$4) I(X) = \sum_q c_q I(X_q)$$

Ces résultats s'obtiennent en appliquant les propriétés classiques relatives aux produits tensoriels d'espaces vectoriels ou de métriques [Sch81]. Les égalités relatives aux tenseurs  $X_{E \otimes F}$ ,  $Z$ ,  $U$  se démontrent facilement en revenant aux définitions.

Par exemple, pour la propriété 3), comme l'espace des variables  $F$  est le même pour tous les groupes, on considère la représentation tensorielle de l'opérateur variable  $U = \sum \{ M_{jj'}^i x^j \otimes x^{j'} \mid j, j' \in J \}$ ;  $M$  est une matrice diagonale par blocs telle que  $M_{jj'}^i = c_q M_{jj'}^i$  pour  $j, j' \in J_q$ , et  $M_{jj'}^i = 0$  si  $j \in J_q$  et  $j' \in J_{q'}$  avec  $q \neq q'$ ; on en déduit que  $U = \sum \{ c_q M_{jj'}^i x^j \otimes x^{j'} \mid j, j' \in J_q, q \in Q \} = \sum \{ c_q U_q \mid q \in Q \}$ ; l'égalité 3) est donc démontrée. La linéarité de la trace permet d'avoir l'égalité 4).

Le problème d'optimisation qui nous servira de référence est celui relatif aux variables :

#### Problème II.6.3

$$\max_K \langle U, A_F \rangle = \max_q \sum_q c_q \langle U_q, A_F \rangle_K$$

$A_F$  projecteur de rang  $K$  de  $F^* \otimes F$

#### IV - ETUDE D'UN TABLEAU MULTIPLE PAR L'ANALYSE CANONIQUE GENERALISEE

##### IV.1 - Les liaisons $R^2$ , $L^2$ et $L$

Pour étudier un ensemble de groupes de variables  $X_q$ , Carroll propose la recherche d'un ensemble de vecteurs  $\{v_k \mid k \in K\}$  Dp-orthonormés liés à chaque groupe de variables au sens du carré de la corrélation multiple  $R^2$ . La famille de vecteurs  $\{v_k \mid k \in K\}$  est solution des  $K$  problèmes suivants :

###### Problème IV.1.0

$$\max_q \sum R^2(v_k, X_q)$$

$$\langle v_k, v_{k'} \rangle_{Dp} = 1 \text{ si } k=k', \quad 0 \text{ sinon}$$

$$v_k \in F, k \in K$$

Escoufier-Pagès font remarquer que l'Analyse Canonique Généralisée de Carroll pose des problèmes d'interprétation car les vecteurs canoniques peuvent exprimer une variance très faible des groupes de variables  $X_q$ . Ils proposent donc de rechercher des combinaisons linéaires de variables d'un groupe décrivant mieux ces groupes au sens de la variance expliquée.

Pour cela, on se restreint à des métriques diagonales  $M_q = \Delta_q$  et on définit la liaison entre une variable  $v_k$  et un groupe  $X_q$  comme suit :

$$L^2(v_k, X_q) = I^{\perp}_{\Delta_v} (N_{F,q}^{Jq}) : \text{l'inertie en projection du nuage } \{x^j \mid j \in J_q\} \text{ des variables du groupe } q \text{ sur } v_k.$$

On recherche donc une famille  $\{v_k \mid k \in K\}$  Dp-orthonormés de vecteurs solutions des  $K$  problèmes suivants :

###### Problème IV.1.1

$$\max_q \sum c_q L^2(v_k, X_q)$$

$$\langle v_k, v_{k'} \rangle_{Dp} = 1 \text{ si } k=k', \quad 0 \text{ sinon}$$

$$v_k \in F, k \in K$$

Nous allons proposer une liaison  $L$  généralisant les liaisons  $R^2$  et  $L^2$ . Définissons la liaison  $L$  entre une variable  $v$  normée et un triplet  $(X_q, M_q, D_p)$  de la manière suivante :

$$L(v, X_q, M_q) = \langle A_v, U_q \rangle$$

où  $A_v$  est le Dp-projecteur associé à la droite  $\Delta_v$ .

###### Proposition IV.1

Les égalités suivantes sont vraies :

$$L(v, X_q, V_{qq}^{-1}) = R^2(v, X_q)$$

$$L(v, X_q, \Delta_q) = L^2(v, X_q)$$

où  $V_{qq}^{-1}$  est la métrique de Mahalanobis associée au groupe  $X_q$  et  $\Delta_q$  une métrique diagonale sur  $E_q$ .

*Démonstration*

Au triplet  $(X_q, V_{qq}^{-1}, D_p)$  est associé l'opérateur variable  $U_q$  tel que  $U_q = X_q V_{qq}^{-1} {}^t X_q D_p = A_q$  projecteur associé à l'espace  $E_q$ . Par suite :

$$\begin{aligned} \langle A_v, A_q \rangle &= \text{Trace}(A_v \circ A_q) = \text{Trace}(v {}^t v D_p A_q) = \text{Trace}({}^t v D_p A_q v) = {}^t v D_p A_q v \\ \langle A_v, A_q \rangle &= \langle v, A_q v \rangle_{D_p} = R^2(v, X_q) \quad \text{car } \|v\|_{D_p}^2 = 1 \end{aligned}$$

Si la métrique  $M_q = \Delta_q$  est diagonale par un calcul similaire à celui effectué au paragraphe II.4, on a  $\langle A_v, U_q \rangle = I_{\Delta_v}^{\perp} (N_{F_q}^J) = L^2(v, X_q)$ . Ce qui achève la démonstration.

L'Analyse Canonique Généralisée au sens de  $L$  revient donc à rechercher une famille de vecteurs :  $\{v_k \mid k \in K\}$   $D_p$ -orthonormés solutions des  $K$  problèmes suivants :

Problème IV.1.2

$$\max_q \sum_q c_q L(v_k, X_q, M_q) = \sum_q c_q \langle A_v, U_q \rangle = \langle A_v, U \rangle_k$$

$$\langle v_k, v_{k'} \rangle_{D_p} = 1 \quad \text{si } k=k', \quad 0 \text{ sinon}$$

$$v_k \in F, \quad k \in K$$

Soient  $\{v_k \mid k \in K\}$  une famille de vecteurs solution de l'Analyse Canonique Généralisée au sens de  $L$ . L'espace engendré par la famille  $\{v_k \mid k \in K\}$  est noté  $F_K = \oplus_k v_k$ . Le projecteur associé à  $F_K$  a pour expression :

$$A_F = \sum_{k \in K} A_v \quad \text{car les vecteurs } \{v_k \mid k \in K\} \text{ sont } D_p\text{-orthonormés. Comme chaque vecteur } v_k \text{ est solution d'un problème du type IV.1.2, le projecteur } A_F \text{ maximise l'expression : } \sum_{k \in K} \langle A_v, U \rangle = \langle \sum_{k \in K} A_v, U \rangle = \langle A_F, U \rangle_K.$$

Le projecteur  $A_F$  est donc solution du problème II.6.3. On a donc démontré l'équivalence entre l'Analyse

en Composantes Principales et l'Analyse Canonique Généralisée aux tableaux multiples.

#### IV.2 - Liaison $L$ entre plusieurs groupes de variables

Soient les triplets  $(X_1, M_1, D_p)$  et  $(X_2, M_2, D_p)$  et  $U_1, U_2$  les opérateurs variables associés. On définit la liaison  $L$  entre ces groupes de variables comme suit :

$$L[(X_1, M_1), (X_2, M_2)] = \langle U_1, U_2 \rangle$$

On peut normaliser cette mesure de liaison par les normes  $\|U_1\|$  et  $\|U_2\|$ , on a alors l'équivalent d'un coefficient de corrélation  $R_v$  [Esc80] entre opérateurs.

$$L[(X_1, M_1), (X_2, M_2)] = \frac{\langle U_1, U_2 \rangle}{\|U_1\| \|U_2\|} = R_v(U_1, U_2)$$

Deux groupes de variables ayant une forte liaison  $R_v \approx 1$  signifie que les nuages des variables ont mêmes composantes principales [CaP76]. Nous utiliserons toutefois la liaison  $L$  sous la forme non normalisée.

Soit une variable  $v$  de norme unité, on lui associe le triplet  $(v, Id, D_p)$  l'opérateur variable de ce triplet est le projecteur  $A_v$  et l'on a bien :

$$L(v, X_q, M_q) = \langle A_v, U_q \rangle = L[(v, Id), (X_q, M_q)]$$

L'Analyse Canonique Généralisée au sens de  $L$  peut donc être considérée comme la recherche d'une famille de variables  $X_K = \{v_k \mid k \in K\}$   $D_p$ -orthonormées la plus liée au sens de  $L$  aux groupes  $X_q, q \in Q$ . Il suffit de considérer les triplets  $(X_K, Id, D_p)$  et  $(X, M, D_p)$  avec  $X = (X_1, \dots, X_Q)$  et  $M = \{c_q M_q \mid q \in Q\}$ . La liaison que l'on note  $B$  s'écrit alors :

$$B = L[(X_K, Id), (X, M)] = \langle A_F, U \rangle = \sum_K c_q \langle A_F, U_q \rangle = \sum_q c_q L[(X_K, Id), (X_q, M_q)]$$

où  $A_F$  est le projecteur associé à l'espace  $F_K = \bigoplus_k v_k$ .

#### IV.3 - Expressions de la liaison $L$

En utilisant les diverses expressions de l'opérateur  $U_q$  (cf. § II.3) :

$$U_q = \sum \{ M_{jj'}^q x^{*j} \otimes x^{j'} \mid jj' \in J_q \} = \sum \{ \lambda_r^q \phi_r^q \otimes \phi_r^q \mid r \in R_q \}$$

où  $R_q$  est l'ensemble des indices relatifs aux valeurs propres  $\lambda_r^q$  non nulles de  $U_q$ , et  $\phi_r^q$  le vecteur propre normé de  $U_q$  associé à  $\lambda_r^q$ . On a alors :

$$B = \sum_q c_q \langle A_F, U_q \rangle = \sum_q c_q \langle \sum_k v_k^* \otimes v_k, \sum \{ M_{jj'}^q x^{*j} \otimes x^{j'} \mid jj' \in J_q \} \rangle$$

$$B = \sum \{ c_q M_{jj'}^q \langle v_k, x^j \rangle_{D_p} \langle v_k, x^{j'} \rangle_{D_p} \mid jj' \in J_q, q \in Q, k \in K \}$$

ou encore :

$$B = \sum_q c_q \langle \sum_k v_k^* \otimes v_k, \sum_r \{\lambda_r^{q_r} \phi_r^{q_r} \otimes \phi_r^{q_r} \mid r \in R_q\} \rangle$$

$$B = \sum \{c_q \lambda_r^{q_r} \phi_r^{q_r} \otimes \phi_r^{q_r} \mid r \in R_q, q \in Q, k \in K\}$$

soit :

$$B = \sum \{c_q \lambda_r^{q_r} \text{corr}^2(v_k, \phi_r^{q_r}) \mid r \in R_q, q \in Q, k \in K\}$$

les facteurs  $\phi_r^{q_r}$  sont en effet normés et centrés. La contribution d'un triplet  $q$  au critère optimisé par l'Analyse Canonique Généralisée au sens de  $L$  dépendra donc de l'importance et du nombre des valeurs propres non nulles de l'opérateur variable  $U_q$ .

## V- ETUDE DES TABLEAUX MULTIPLES PAR LA CLASSIFICATION AUTOMATIQUE

### V.1 - Notations

Soit  $P_K$  l'ensemble des partitions à  $K$  classes,  $P \in P_K$  une partition à laquelle est associée la variable qualitative  $X_K = \{x^k \mid k \in K\}$  où  $x^k$  est la variable indicatrice relative à la classe  $k$ . On note  $I_k$  l'ensemble des indices des individus de la classe  $k$ ,  $D_P = \{p_k = \sum_i \{p_i \mid i \in I_k\}, k \in K\}$  l'ensemble des poids des classes et

$g_k = (g_j^k \mid j \in J)$  le centre de gravité de la classe  $k$ . On a la proposition :

#### Proposition V.1

*Les variables indicatrices sont  $D_P$ -orthogonales et la norme au carré de  $x_k$  est le poids  $p_k$  de la classe  $k$ . Ce qui peut se résumer par l'égalité :*

$$\langle x^k, x^{k'} \rangle_{D_P} = p_k \delta^{k,k'} \text{ pour } k, k' \in K \text{ où } \delta \text{ est le symbole de Kronecker.}$$

*Démonstration*

Calculons  $\langle x^k, x^{k'} \rangle_{D_P} = \sum_i p_i x_i^k x_i^{k'} = 0$  si  $k \neq k'$  car les classes  $k$  et  $k'$  sont disjointes,

sinon  $\sum_i p_i x_i^k x_i^k = p_k$ . La famille de vecteurs :  $\{x^k / \sqrt{p_k}, k \in K\}$  est donc  $D_P$ -orthonormée.

Nous allons interpréter une méthode de Classification Automatique maximisant l'inertie inter-classes sous les deux points de vue suivants .

## V.2 - Point de vue de l'Analyse Canonique Généralisée :

Une partition  $P$  est équivalente à la variable qualitative  $X_K$  de ses indicatrices à laquelle on associe la métrique du chi-deux. Le triplet considéré est  $(X_K, D_{1/p}, D_p)$  où  $D_{1/p}$  est l'inverse de  $D_p$ . L'opérateur

variable associé à ce triplet s'écrit :  $U_K = \sum_k x^k \otimes x^k / p_k = A_F$  le projecteur associé à l'espace  $F_K = \bigoplus_k x^k$ .

On peut naturellement se poser le problème de la recherche d'une variable qualitative partition la plus liée au sens de  $L$  aux triplets  $(X_q, M_q, D_p)$  pondérés par les coefficients  $c_q$  pour  $q \in Q$ . Le problème d'optimisation s'énonce comme suit :

### Problème V.2

$$\max_{X_K} L[(X_K, D_{1/p}), \{(X_q, M_q), q \in Q\}] = \sum_q c_q \langle A_F, U_q \rangle_K$$

$X_K$  variable qualitative à  $K$  modalités

Nous allons montrer que ce problème revient à rechercher une partition maximisant l'inertie inter-classes, mais énonçons d'abord le deuxième point de vue.

## V.3 - Point de vue : approximation d'un tenseur d'ordre $K$

Nous savons que  $B = \langle A_F, U \rangle = \sum_q c_q \langle A_F, U_q \rangle = \|X_{E \otimes F}\|_{M \otimes D_p}^2$  où  $X_{E \otimes F}$  est la projection

orthogonale de  $X_{E \otimes F}$  sur l'espace  $E \otimes F_K$  (cf § II.6.1). On peut donc aussi énoncer le problème précédent comme la recherche d'un e.v.  $F_K$  de dimension  $K$  engendré par les variables indicatrices d'une variable qualitative tel que la projection du tenseur  $X_{E \otimes F}$  sur l'espace  $E \otimes F_K$  soit de norme maximale (cf figure 2, § II.6.1). Un e.v.  $F_K$  de dimension  $K$  est engendré par une variable qualitative s'il existe une famille de vecteurs  $\{x^k, k \in K\}$  de  $F$  formant une base  $D_p$ -orthogonale de  $F_K$  telle que :  $\{x_i^k \in \{0,1\} \mid k \in K, i \in I\}$  et  $\sum_k \{x^k \mid k \in K\} = \mathbf{1}$ , où  $\mathbf{1}$  est le vecteur de  $F$  dont toutes les composantes valent 1.

Le problème d'optimisation précédent s'énonce comme suit :

### Problème V.3 :

$$\max \|X_{E \otimes F}\|_{M \otimes D_p}^2$$

$F_K \subset F$  avec  $\dim F_K = K$

$F_K$  engendré par une variable qualitative

La proposition suivante permet de faire le lien entre l'Analyse en Composantes Principales, l'Analyse Canonique Généralisée et les méthodes de Classification Automatique maximisant l'inertie inter-classes.

### Proposition V.3.1

Les égalités suivantes sont vraies :

$$1) X_{E \otimes F} = \sum_K g_k \otimes x^k$$

$$2) \|X_{E \otimes F}\|_{M \otimes Dp}^2 = \sum_K p_k \|g_k\|_M^2 = \text{Inertie inter-classes}$$

Démonstration :

Notons  $\{v^k = x^k / \sqrt{p_k} \mid k \in K\}$  la famille de vecteurs  $Dp$ -orthonormés de  $F$  qui engendrent  $F_K$  et  $d_k$  le vecteur de  $E$  dont la  $j$ -ème composante est :  $d_k^j = \langle x^j, v^k \rangle_{Dp}$  et ceci pour  $k \in K$ . Les espaces euclidiens  $(E, M)$  et  $(F, Dp)$ , les individus et les variables jouant un rôle symétrique, la proposition II.6.3 permet d'écrire :

$$X_{E \otimes F} = \sum_K d_k \otimes v^k = \sum_K d_k \otimes x^k / \sqrt{p_k}$$

Or  $\langle x^j, x^k \rangle_{Dp} = \sum_i p_i x_i^j x_i^k = p_k g_k^j$  par définition de la variable indicatrice  $x^k$ .

On a donc la  $j$ -ème composante de  $d_k$  :  $d_k^j = \langle x^j, x^k \rangle_{Dp} / \sqrt{p_k} = \sqrt{p_k} g_k^j$  et  $d_k / \sqrt{p_k} = g_k$  d'où le résultat :  $X_{E \otimes F} = \sum_K g_k \otimes x^k$ .

Calculons la norme de ce vecteur :

$$\langle X_{E \otimes F}, X_{E \otimes F} \rangle_{M \otimes Dp} = \langle \sum_K g_k \otimes x^k, \sum_{K'} g_{K'} \otimes x^{K'} \rangle_{M \otimes Dp} = \sum_K \sum_{K'} \langle g_k, g_{K'} \rangle_M \langle x^k, x^{K'} \rangle_{Dp}$$

Comme les vecteurs  $x^k$  sont  $Dp$ -orthogonaux et leurs normes au carré sont égales à  $p_k$ , on a l'égalité cherchée :

$$\|X_{E \otimes F}\|_{M \otimes Dp}^2 = \sum_K p_k \|g_k\|_M^2 = \text{Inertie inter-classes}$$

Si on désigne par  $E_K$  l'espace associé au tenseur  $X_{E \otimes F}_K$  dans  $E$ , le tableau ci-dessous résume les résultats en les mettant en parallèle à ceux obtenus par une A.C.P. classique sur le triplet  $(X, M, D_p)$ .

| Classification Automatique                                         | Analyse en Composantes Principales                           |
|--------------------------------------------------------------------|--------------------------------------------------------------|
| $F_K = \bigoplus_k \Delta x^k$                                     | $F_K = \bigoplus_k \Delta c^k$                               |
| $E_K = \sum_k \Delta g_k$                                          | $E_K = \bigoplus_k \Delta u^k$                               |
| $X_{E \otimes F}_K = \sum_k g_k \otimes x^k$                       | $X_{E \otimes F}_K = \sum_k u_k \otimes c^k$                 |
| $\ X_{E \otimes F}_K\ _{M \otimes D_p}^2 = \sum_k p_k \ g_k\ _M^2$ | $\ X_{E \otimes F}_K\ _{M \otimes D_p}^2 = \sum_k \lambda_k$ |

Tableau V.3

On voit donc que le  $k$ -ème couple  $(g_k, x^k)$  joue le même rôle que  $k$ -ème couple de facteurs  $(u_k, c^k)$  à la différence près que les centres de gravités  $g_k$  ne sont point  $M$ -orthogonaux et que les couples  $(u_k, c^k)$  sont ordonnés et uniques. Chacun d'eux est solution d'un problème d'optimisation, ce qui n'est pas le cas pour les couples  $(g_k, x^k)$ . On peut donc considérer une méthode de Classification Automatique, maximisant l'inertie inter-classes, comme une A.C.P. "sous contraintes" puisque, on impose à l'e.v.  $F_K$  d'être engendré par une variable qualitative. Ce résultat a été mis en évidence par Lerman [Ler79] et Govaert [Gov83] sur la base des travaux de Howard [How69] dans l'étude d'un tableau simple et pour une métrique  $M$  diagonale. Nous avons généralisé le résultat dans le cas d'un tableau simple pour une métrique  $M$  quelconque et pour un tableau multiple lorsque la métrique  $M$  est diagonale par blocs.

*Remarque :*

On ne sait rien sur l'unicité de l'espace vectoriel  $F_K$  solution du problème V.3 ni sur l'unicité d'une base de vecteurs  $D_p$ -orthogonales associée à une variable qualitative engendrant un espace vectoriel  $F_K$  donné. En d'autres termes, on ne sait pas si la partition optimale maximisant l'inertie inter-classes est unique ou non.

La proposition suivante résume les résultats précédents.

### Proposition V.3.2

Une méthode de Classification Automatique relative à un ensemble de triplets  $(X_q, M_q, D_p)$  pondérés par les coefficients  $\{c_q, q \in Q\}$  maximisant l'inertie inter-classes  $B$  est équivalent à la recherche d'une variable qualitative  $X_K$  maximisant la liaison  $L$  entre les triplets  $(X_K, D_{1/p}, D_p)$  et  $(X_q, M_q, D_p)$  pour

$q \in Q$ . L'inertie inter-classes  $B$  a pour expression :

$$B = L[(X_K, D_{1/p}), \{(X_q, M_q), q \in Q\}] = \sum_q c_q \langle A_F, U_q \rangle = \sum_q c_q B_q$$

où  $B_q$  désigne l'inertie inter-classes de la partition calculée sur le groupe de variables  $q$ .

#### V.4 - Expression de l'inertie inter-classes B

On a, dans le cadre de la classification automatique, les équivalents des expressions du critère maximisé par l'Analyse Canonique Généralisée donnés au paragraphe IV.3 selon les expressions de l'opérateur  $U_q$  :

$$B = \sum_q c_q \langle A_F, U_q \rangle = \sum_q c_q \langle \sum_k x^k \otimes x^k / p_k, \sum \{ M_{jj'}^q x^j \otimes x^{j'} \mid j, j' \in J_q \} \rangle$$

d'où

$$B = \sum \{ c_q M_{jj'}^q \langle x^k, x^j \rangle_{D_p} \langle x^k, x^{j'} \rangle_{D_p} / p_k \mid j, j' \in J_q, q \in Q, k \in K \}$$

Or on sait que  $\langle x^k, x^j \rangle_{D_p} = p_k g_k^j$  où  $g_k^j$  est la moyenne de la variable  $x^j$  dans la classe  $k$ , il vient :

$$B = \sum \{ c_q p_k M_{jj'}^q g_k^j g_k^{j'} \mid j, j' \in J_q, q \in Q, k \in K \}.$$

En utilisant l'expression de l'opérateur  $U_q$  en fonction de ses éléments propres, on a par ailleurs :

$$\begin{aligned} B &= \sum_q c_q \langle A_F, U_q \rangle = \sum_q c_q \langle \sum_k x^k \otimes x^k / p_k, \sum \{ \lambda_r^q \phi_r^q \otimes \phi_r^q \mid r \in R_q \} \rangle \\ B &= \sum \{ c_q \lambda_r^q \langle x^k, \phi_r^q \rangle_{D_p}^2 / p_k \mid r \in R_q, q \in Q, k \in K \}. \end{aligned}$$

Les vecteurs  $x^k$  ont pour normes au carré  $p_k$ , les facteurs  $\phi_r^q$  sont normés et de moyennes nulles et l'on a  $\langle x^k, \phi_r^q \rangle_{D_p} = p_k \phi_{rk}^q$  où  $\phi_{rk}^q$  est la moyenne du facteur  $\phi_r^q$  dans la classe  $k$ , on a alors :

$$B = \sum \{ c_q \lambda_r^q \cos^2(x^k, \phi_r^q) \mid r \in R_q, q \in Q, k \in K \} = \sum \{ c_q \lambda_r^q p_k \phi_{rk}^q{}^2 \mid r \in R_q, q \in Q, k \in K \}$$

• On considère un ensemble de variables quantitatives  $X$  muni de la métrique  $D_{1/\sigma^2}$  (métrique diagonale de l'inverse des variances); le triplet considéré est  $(X, D_{1/\sigma^2}, D_p)$ ; l'inertie inter-classes d'une partition  $P$  relative à  $X$  s'écrit :

$$B = \sum_j \sum_k \langle x^j, x^k \rangle_{D_p}^2 / \|x^k\|_{D_p}^2 \|x^j\|_{D_p}^2 = \sum_j \sum_k \cos^2(x^j, x^k)$$

Ce résultat est à mettre en parallèle à celui d'une A.C.P. normée du tableau  $X$  où le critère  $B$  optimisé par les  $K$  premières composantes principales  $c^k$  est  $B = \sum_k \sum_j \text{corr}^2(x^j, c^k)$ . La différence résulte du fait que les variables indicatrices  $x^k$  sont des variables binaires non centrées tandis que les composantes principales  $c^k$  sont des variables continues et centrées.

• Soit un ensemble de variables qualitatives  $\{X_q, q \in Q\}$ , on note  $X_q^*$  la variable  $X_q$  centrée. Les triplets considérés sont  $(X_q^*, D_{1/p}, D_p)$  et l'opérateur variable est  $A_q^*$  le projecteur associé à l'espace  $F_q^*$

orthogonal à la droite des constantes (cf. [CaP76]). L'inertie inter-classes  $B$  a pour expression dans ce cas :

$$B = \sum_q c_q \langle A_F, A_q^* \rangle = \sum_q c_q \Phi_{Kq}^2$$

où  $\Phi_{Kq}^2$  est le phi-deux entre la variable qualitative  $X_q$  et la variable qualitative partition  $X_K$ . Si l'on

choisit comme coefficient de pondération pour la variable  $X_q$ ,  $c_q = 1 / \sqrt{(K-1)(\text{card } J_q - 1)}$  alors l'inertie inter-classes s'écrit :

$$B = \sum_q \frac{\sum K_q^2}{\sqrt{(K-1)(\text{card } J_q - 1)}} = \sum_q T_{Kq}^2$$

où  $T_{Kq}^2$  est le coefficient de Tschuprow mesurant la liaison entre les variables qualitatives  $X_q$  et  $X_K$ .

### V.5 Choix des coefficients de pondération :

Pour équilibrer le rôle joué par les différents groupes de variables, on considérera l'expression de la mesure d'information  $I(X) = \sum_q c_q I(X_q)$  ou le critère optimisé par les méthodes factorielles ou de Classification

Automatique considérées :  $B = \sum_q c_q \langle A_F, U_q \rangle_K$

#### 1) Normalisation des inerties des groupes :

On choisit les coefficients  $c_q$  tels que :  $c_1 I(X_1) = c_2 I(X_2) = \dots = c_Q I(X_Q) = 1$ , on a :  $c_q = 1/I(X_q)$  pour  $q \in Q$ . Le critère  $B$  optimisé s'interprète géométriquement. Considérons uniquement le groupe de variables  $q$ , on a :

-L' inertie du nuage des individus :  $I(X_q) = \|X_{E_q \otimes F}\|_{M_q \otimes D_p}^2$ .

-L'inertie inter-classes :  $B_q = \|X_{E_q \otimes F}\|_{M_q \otimes D_p}^2 \langle U_q, A_F \rangle_K$  où  $X_{E_q \otimes F}$  est la projection  $M_q \otimes D_p$  orthogonale de  $X_{E_q \otimes F}$  sur  $E_q \otimes F$ .

-Le pourcentage d'inertie expliquée par la partition  $P$  pour la variable  $q$  s'écrit :

$$B_q / I(X_q) = \|X_{E_q \otimes F}\|_{M_q \otimes D_p}^2 / \|X_{E_q \otimes F}\|_{M_q \otimes D_p}^2 = \cos^2(X_{E_q \otimes F}, E_q \otimes F)_K$$

Le critère  $B$  s'exprime alors pour les coefficients  $c_q$  choisis, comme suit :

$$B = \sum_q B_q / I(X_q) = \sum_q \cos^2(X_{E_q \otimes F}, E_q \otimes F)_K$$

#### 2) Normalisation par la norme de l'opérateur variable :

L'expression de  $B$  en fonction de  $A_F$  et  $U_q$  suggère le choix suivant :

$$c_q = 1 / \|A_F\| \|U_q\|_K$$

le critère  $B$  s'écrit alors comme une somme de termes compris entre 0 et 1 :

$$B = \sum_q c_q \langle U_q, A_F \rangle_K = \sum_q \langle U_q, A_F \rangle_K / \|A_F\| \|U_q\|_K = \sum_q R_v(A_F, U_q)_K$$

on a donc

$$B = \sum_q \cos(A_F, U_q)_K$$

*Remarque :*

Nous avons associé à une variable qualitative la métrique du chi-deux. Il aurait été possible de considérer d'autres métriques, comme la métrique identité par exemple, et d'envisager à priori des méthodes de Classification Automatique maximisant le coefficient Rv. Cela a été fait par Nin [Nin81] dans le cas d'un tableau simple. Le critère optimisé n'est plus l'inertie inter-classes et les algorithmes proposés, basés sur la technique du gradient réduit, sont plus complexes que celui des centres mobiles.

D'autres choix sont possibles dans le cadre des méthodes factorielles. Celui d'Escoufier qui égalise le premier moment d'inertie de chaque groupe de variables  $X_q$ , on a donc :  $c_q = 1/\lambda_q^1$  ; on peut aussi prendre comme dans la méthode STATIS les coefficients de l'opérateur compromis. Ces différentes stratégies nécessitent d'effectuer des A.C.P. sur chaque groupe de variables ou sur le tableau des produits scalaires entre les opérateurs. Les choix que l'on propose sont plus simples à mettre en œuvre. En effet, pour un triplet  $(X_q, c_q M_q, D_p)$ , on considère le triplet  $(Y_q, Id_q, D_p)$  où  $Y_q = \{y^j, | j \in J_q\}$  est le tableau centré déduit de  $X_q$  pour se ramener à la métrique euclidienne usuelle  $Id_q$  sur  $Eq$ . On a alors les expressions suivantes :

$$I(X_q) = I(Y_q) = \sum \{ \text{var } y^j | j \in J_q \} \quad \text{et} \quad \|U_q\|^2 = \|Z_q\|^2 = \sum \{ \text{covar}^2(y^j, y^{j'}) | j, j' \in J_q \} \quad (\text{cf } \S II.4)$$

qui permettent le calcul des coefficients  $c_q$  à partir de la matrice de variances-covariances relative à  $Y_q$ .

*Remarque :*

Si l'on a une variable  $x^q$  par groupe,  $X = \{x^q | q \in Q\}$  alors le coefficient  $c_q$  s'écrit :

$c_q = 1/I(X_q) = 1/\|U_q\| = 1/\text{var } x^q$ . Les choix proposés correspondent à la réduction classique.

**V.6 - Aides à l'interprétation d'une partition**

Pour analyser un ensemble de triplets  $(X_q, M_q, D_p)$ ,  $q \in Q$ , nous étudions le triplet  $(X, M, D_p)$  où  $X = (X_1, \dots, X_Q)$  et  $M = \sum_q c_q M_q$ . Ce dernier triplet est lui-même équivalent au triplet  $(Y, Id, D_p)$  en

considérant la décomposition de Choleski de la métrique  $M = T^t T$  et en posant  $Y = XT$ .

Nous rappelons alors les aides à l'interprétation proposées par Modulad [Mod82] pour l'étude d'une partition en interprétant les divers critères dans l'espace des variables  $(F, D_p)$ . On note le tableau centré  $Y = \{y^j, | j \in J\}$  et  $U = \sum \{ y^j \otimes y^j | j \in J \}$  l'opérateur variable associé à  $(Y, Id, D_p)$ . L'inertie inter-classes  $B$  s'écrit alors :

$$B = \langle A_F, U \rangle = \langle \sum_K x^k \otimes x^k / \|x^k\|^2_{D_p}, \sum_j y^j \otimes y^j \rangle$$

$$B = \sum_k \sum_j \langle x^k, y^j \rangle_{D_p}^2 / \|x^k\|^2_{D_p}$$

On peut écrire  $B = \sum_k \sum_j B_{jk}^j$  avec  $B_{jk}^j = \langle x^k, y^j \rangle_{D_p}^2 / \|x^k\|^2_{D_p}$

On note de même  $T = \sum_j T^j = \sum_j \|y^j\|^2_{D_p}$  l'inertie totale,  $T^j$  étant la variance de la variable  $y^j$  :

$$T^j = \|y^j\|^2_{D_p} = (\sigma^j)^2$$

On définit alors :

• La contribution de la variable  $j$  et de la classe  $k$  à l'inertie inter-classes :

$$\text{CTR}(j,k) = B_k^j / B$$

• La contribution de la variable  $j$  à l'inertie inter-classes :

$$\text{CTR}(j) = \sum_k \text{CTR}(j,k)$$

Ces indices mesurent l'importance d'une variable dans la détermination d'une classe ou d'une partition.

On définit par ailleurs :

• La liaison entre une variable  $y^j$  et une classe  $x^k$  qui est mesurée par le critère suivant :

$$\text{COR}(j,k) = B_k^j / T^j = \langle x^k, y^j \rangle_{D_P}^2 / \|x^k\|_{D_P}^2 \|y^j\|_{D_P}^2 = \cos^2(y^j, x^k)$$

qui est le "pouvoir discriminant" de la variable  $y^j$  par rapport à la classe  $k$ .

• La corrélation de la variable  $j$  et la partition  $P$  qui est notée :

$$\text{COR}(j) = \sum_k \text{COR}(j,k)$$

Le programme INTERP de Modulad édite le tableau recensant les indices CTR et COR pour une partition donnée (cf figure 5).

En complément à ces critères, on considère le tableau répertoriant les indices suivants pour des variables de type *quantitatif* (cf figure 4) :

•  $\text{stud}(j,k) = (g_k^j - g^j) / \sigma^j$  l'écart normalisé entre la moyenne de la variable  $j$  dans la classe  $k$  et la moyenne calculée dans la population totale (avant centrage).

•  $\text{moy}_k(j)$ ,  $\text{sig}_k(j)$ ,  $\text{min}_k(j)$ ,  $\text{max}_k(j)$  : la moyenne, l'écart-type, le minimum et le maximum de la variable  $j$  dans la classe  $k$ .

•  $\text{moy}_g(j)$ ,  $\text{sig}_g(j)$ ,  $\text{min}_g(j)$ ,  $\text{max}_g(j)$  : la moyenne, l'écart-type le minimum et le maximum de la variable  $j$  dans la population totale.

Pour les variables de type *qualitatif* le programme INPAQL de Modulad édite les indices suivants pour interpréter une partition (cf figure 6):

•  $\chi^2_1(k,j)$  le chi-deux a un degré de liberté mesurant la liaison entre la classe  $k$  et la modalité  $j$ . Le tableau de contingence considéré étant, si l'on note:  $n_j$  l'effectif de la modalité  $j$ ,  $n_k$  l'effectif de la classe  $k$ ,  $n_{jk}$  l'effectif de la modalité  $j$  dans la classe  $k$ :

|           | $x^k$          | $1 - x^k$                |
|-----------|----------------|--------------------------|
| $y^j$     | $n_{jk}$       | $n_j - n_{jk}$           |
| $1 - y^j$ | $n_k - n_{jk}$ | $n - n_j - n_k + n_{jk}$ |

- l'effectif de la modalité  $j$  dans la classe  $k$  (effe)
- la fréquence de la modalité  $j$  dans la classe  $k$  (mc/cl)
- la fréquence de la modalité  $j$  dans la population (mt/n)
- la fréquence de la modalité  $j$  dans la classe  $k$  par rapport à son effectif total (mc/mt)

On n'édite que les modalités dont la liaison est significative au seuil de 5 % c.à.d. si  $\chi^2_1(k,j) \geq 3.85$ .

### V.7 - Exemple d'application

Les données sont relatives aux étudiants de première année de la MIAE de l'Université de Paris-Dauphine. La promotion comporte 111 étudiants dont on connaît le sexe, la formation d'origine (DUT, GEA, MASS, MD, SSM et autres) et les notes dans 11 matières (Cobol (Cobo), Fortran (Fort), Structures de données (Stru), Méthodologie (Métho), Fichiers (Fich), Algorithmique (Algo), Méthode Numérique (Mnum), Statistiques (Stat), Techniques d'expression (Expr), Gestion financière (Gest), Anglais (Angl)). On a donc deux groupes de variables un premier groupe  $X$  comportant 11 variables quantitatives notes et un second groupe  $Y$  composé des 2 variables qualitatives sexe et formation d'origine totalisant 8 modalités.

Classiquement pour étudier un tel tableau, on se ramène à un ensemble de variables de même type, quantitatif ou qualitatif. On peut coder les variables de type quantitatif en qualitatif en les découpant en classes puis on analyse l'ensemble des variables, rendues ainsi qualitatives, par une Analyse des Correspondances Multiples. L'autre possibilité consiste à effectuer une Analyse des Correspondances Multiples sur les variables qualitatives et de remplacer ces dernières par un nombre restreint de facteurs. Les méthodes traitant des tableaux quantitatifs peuvent alors s'appliquer. La première procédure, si elle présente l'avantage de pouvoir mettre en évidence d'éventuelles liaisons non linéaires, nécessite de passer par une étape de codage pas toujours évidente (choix du nombre des classes, des bornes) dont dépendront les résultats. D'autre part découper les variables quantitatives en classes augmente la dimension de l'espace des individus, ce qui n'est pas souhaitable dans notre exemple compte tenu du nombre restreint d'individus que l'on dispose. La seconde procédure présente l'inconvénient de perdre l'information contenue dans les facteurs non sélectionnés. Nous allons voir que la méthode proposée, qui consiste à traiter simultanément les deux groupes de variables en les pondérant de manière adéquate, permet de résoudre simplement le problème de l'hétérogénéité des variables:

Les triplets analysés sont :  $(X, D_{1/02}, D_p)$  celui relatif aux variables quantitatives et  $\{(Y^*, D_{1/p}, D_p) \mid q = 1, 2\}$  ceux relatifs aux deux variables qualitatives centrées. L'inertie inter-classes

B s'écrit, quand on recherche une partition P à K classes :

$$B = c_1 \sum \{\cos^2(x^k, x^j) \mid k \in K, j = 1, 11\} + c_2 (\Phi_{K1}^2 + \Phi_{K2}^2)$$

Le calcul de  $c_1$  et  $c_2$ , en égalisant l'inertie des deux groupes ou en normalisant la norme des opérateurs, conduit aux mêmes valeurs :  $c_1 = 0,214$  et  $c_2 = 0,786$ . Une A.C.P. a d'abord été effectuée sur le tableau juxtaposant les deux sous-tableaux relatifs aux variables quantitatives et aux variables indicatrices (11+8=19 variables), pondérés par les coefficients  $c_1$  et  $c_2$ . Cette analyse suggère le nombre de classes (3) et les individus nécessaires pour initialiser les Nuées Dynamiques. La partition optimale P obtenue est ensuite interprétée grâce aux indices recensés dans les figures 4,5,6 et conduit aux résultats suivants :

*Classe des faibles : 23 %*

Elle est constituée d'élèves dont les notes, dans les différentes matières, sont en général inférieures aux notes moyennes relatives à l'ensemble de la population (indice stud). Le pourcentage d'étudiants de formation "autres", dans cette classe, est deux fois plus important que le pourcentage moyen. Ils ne réussissent pas particulièrement dans les matières statistiques et méthodes numériques.

*Classe des moyens : 61 %*

Elle est constituée en grande partie d'étudiants de formation GEA, MD et MASS.

*Classe des forts : 17 %*

Cette classe est caractérisée par des étudiants dont les notes, dans les différentes matières, sont en général supérieures aux notes moyennes relatives à l'ensemble de la population. Les étudiants la composant sont uniquement de formation DUT et ils réussissent bien en Fortran, Cobol et Structures de Données.

Les variables les plus contributives à la détermination de la partition sont : Statistiques, méthodes numériques et DUT.

## VI - CONCLUSION

Nous avons donc formulé un ensemble de méthodes d'Analyse de Données comme différentes expressions d'un même problème qui est l'approximation d'un tenseur d'ordre K. Cette présentation débouche sur des aspects pratiques pour traiter des tableaux multiples par la Classification Automatique, en suggérant les coefficients de pondération appropriés et des aides à l'interprétation spécifiques. Les limites d'une telle approche résident dans le choix du modèle vectoriel et inertiel pour la représentation des phénomènes. Ainsi les méthodes proposées ne sont valides que si la représentation d'une variable ou groupe de variables par un opérateur, ou ce qui revient au même par un espace vectoriel est adéquat. Ce qui n'est pas toujours le cas (variables ordinales par exemple). Les axes de recherche restent nombreux dans l'étude des tableaux multiples. Dans le cadre des Nuées Dynamiques, on pourra considérer d'autres modes de représentation d'une classe.

### Remerciements :

*Je remercie Mme Blin directrice de la Miage de l'université Paris-Dauphine de m'avoir permis d'accéder à ces données.*

*Je remercie les professeurs Cazes et Diday de l'université Paris-Dauphine pour l'aide, et les conseils qu'ils m'ont apportés.*

classe : 1 effectif : 26, 23% variance : .82217

```
*****
* vari *.libelle.....*stud.* moyk * moyg * sigk * sigg * mink maxk*
*****
* stat *.statistiques.....*-1.20 * 7.3 * 13. * 2.9 * 4.2 * 3.0 14.*
* mnum *.méthodes numériques...*-1.10 * 9.5 * 13. * 3.3 * 3.3 * 2.0 16.*
* algo *.algorithmique.....*-0.79 * 9.2 * 12. * 3.0 * 3.1 * 4.0 15.*
* gest *.gestion financière...*-0.79 * 11. * 13. * 2.4 * 2.9 * 5.5 14.*
* meto *.méthodologie.....*-0.79 * 7.4 * 10. * 2.8 * 3.5 * 3.0 13.*
* stru *.structures.de.données.*-0.75 * 4.9 * 8.1 * 3.1 * 4.3 * 1.0 12.*
* fort *.fortran.....*-0.64 * 8.5 * 11. * 3.4 * 4.3 * 3.0 17.*
* fich *.fichiers.....*-0.63 * 7.6 * 9.7 * 3.1 * 3.4 * 2.0 13.*
* cobo *.cobol.....*-0.62 * 11. * 13. * 2.6 * 3.3 * 4.5 17.*
* angl *.anglais.....*-0.41 * 12. * 13. * 2.7 * 2.6 * 6.0 15.*
* expr *.techniquesd'expression*-0.40 * 12. * 12. * 2.1 * 2.0 * 6.5 16.*
*****
```

classe : 2 effectif : 68, 61% variance : .88016

```
*****
* vari *.libelle.....* stud * moyk * moyg * sigk * sigg * mink maxk*
*****
* cobo *.cobol.....* .01 * 13. * 13. * 3.0 * 3.3 * 3.0 20.*
* expr *.techniquesd'expression* .02 * 12. * 12. * 1.9 * 2.0 * 8.0 18.*
* fort *.fortran.....* .03 * 11. * 11. * 4.1 * 4.3 * 1.5 20.*
* stru *.structures.de.données.* .04 * 8.3 * 8.1 * 3.8 * 4.3 * 1.0 17.*
* angl *.anglais.....* .05 * 13. * 13. * 2.6 * 2.6 * 5.0 18.*
* fich *.fichiers.....* .08 * 10. * 9.7 * 3.0 * 3.4 * 1.0 18.*
* meto *.méthodologie.....* .13 * 11. * 10. * 3.0 * 3.5 * 5.0 18.*
* algo *.algorithmique.....* .21 * 12. * 12. * 2.7 * 3.1 * 4.0 18.*
* gest *.gestion financière...* .30 * 14. * 13. * 2.7 * 2.9 * 5.0 19.*
* mnum *.méthodes numériques...* .37 * 14. * 13. * 2.3 * 3.3 * 7.0 18.*
* stat *.statistiques.....* .44 * 14. * 13. * 3.0 * 4.2 * 7.5 19.*
*****
```

classe : 3 effectif : 17, 15% variance : .83311

```
*****
* vari *.libelle.....* stud * moyk * moyg * sigk * sigg * mink maxk*
*****
* gest *.gestion financière...* .01 * 13. * 13. * 2.5 * 2.9 * 6.5 18.*
* stat *.statistiques.....* .13 * 13. * 13. * 3.3 * 4.2 * 5.0 18.*
* mnum *.méthodes numériques...* .26 * 14. * 13. * 2.1 * 3.3 * 9.0 18.*
* algo *.algorithmique.....* .39 * 13. * 12. * 2.8 * 3.1 * 8.0 16.*
* angl *.anglais.....* .40 * 14. * 13. * 2.0 * 2.6 * 10. 18.*
* expr *.techniquesd'expression* .54 * 14. * 12. * 1.2 * 2.0 * 12. 16.*
* fich *.fichiers.....* .63 * 12. * 9.7 * 3.2 * 3.4 * 8.0 18.*
* meto *.méthodologie.....* .70 * 13. * 10. * 3.7 * 3.5 * 6.0 19.*
* fort *.fortran.....* .83 * 15. * 11. * 3.3 * 4.3 * 9.0 20.*
* cobo *.cobol.....* .90 * 16. * 13. * 2.8 * 3.3 * 11. 20.*
* stru *.structures de données.* .96 * 12. * 8.1 * 3.8 * 4.3 * 5.0 19.*
*****
```

Figure 4 : Indices complémentaires d'interprétation d'une partition relativement à des variables quantitatives.

classe : 1 effectif : 26, 23%

```
*****
*vari*..libelle.....*moda*.khi2.proba.effe.mc..mt...mc*
*able*.....*lite.*.....
*.....cl...n...mt*
*****
*.form.* formation prealable.....*autr.* 13 0.000 11 42 18 55*
*****
```

classe : 2 effectif : 68, 61%

```
*****
*vari*..libelle.....*moda*.khi2.proba.effe.mc..mt...mc*
*able*.....*lite.*.....
*.....cl...n...mt*
*****
*.form.* formation prealable.....*gea.* 6 0.009 16 24 16 89*
*.form.* formation prealable.....*md...* 4 0.030 7 10 6 100*
*.form.* formation prealable.....*mass.* 4 0.045 12 18 13 86*
*****
```

classe : 3 effectif : 17, 15%

```
*****
*vari*..libelle.....*moda*.khi2.proba.effe.mc..mt...mc*
*able*.....*lite.*.....
*.....cl...n...mt*
*****
*.form.* formation prealable.....* dut * 111 0.000 17 100 15 100*
*****
```

Figure 6 : Interprétation d'une partition relativement à des variables qualitatives

| I | ****I | **** I    | **** I | **** I | **** I | **** I | **** I | **** I | **** I |
|---|-------|-----------|--------|--------|--------|--------|--------|--------|--------|
| I | ****I | PARTITION | I      | CLASSE | -1- I  | CLASSE | -2- I  | CLASSE | -3- I  |
| I | ****I | **** I    | **** I | **** I | **** I | **** I | **** I | **** I | **** I |
| I | VAR I | COR I     | CTR I  | COR I  | CTR I  | COR I  | CTR I  | COR I  | CTR I  |
| I | ****I | **** I    | **** I | **** I | **** I | **** I | **** I | **** I | **** I |
| I | coboI | 21.2 I    | 4.5 I  | 8.9 I  | 4.7 I  | 0.0 I  | 0.0 I  | 12.3 I | 5.6 I  |
| I | fortI | 20.1 I    | 4.3 I  | 9.5 I  | 5.0 I  | 0.1 I  | 0.1 I  | 10.6 I | 4.8 I  |
| I | algoI | 19.7 I    | 4.2 I  | 14.8 I | 7.8 I  | 2.6 I  | 4.5 I  | 2.3 I  | 1.0 I  |
| I | mnumI | 39.9 I    | 8.6 I  | 30.5 I | 16.1 I | 8.4 I  | 14.4 I | 1.1 I  | 0.5 I  |
| I | statI | 48.1 I    | 10.3 I | 35.9 I | 19.0 I | 12.0 I | 20.5 I | 0.2 I  | 0.1 I  |
| I | metoI | 23.3 I    | 5.0 I  | 14.8 I | 7.8 I  | 1.0 I  | 1.7 I  | 7.5 I  | 3.4 I  |
| I | struI | 27.3 I    | 5.8 I  | 13.0 I | 6.9 I  | 0.1 I  | 0.2 I  | 14.2 I | 6.5 I  |
| I | fichI | 15.9 I    | 3.4 I  | 9.3 I  | 4.9 I  | 0.4 I  | 0.7 I  | 6.1 I  | 2.8 I  |
| I | anglI | 6.6 I     | 1.4 I  | 3.9 I  | 2.0 I  | 0.2 I  | 0.3 I  | 2.5 I  | 1.1 I  |
| I | gestI | 20.3 I    | 4.3 I  | 14.8 I | 7.8 I  | 5.5 I  | 9.5 I  | 0.0 I  | 0.0 I  |
| I | exprI | 8.2 I     | 1.8 I  | 3.8 I  | 2.0 I  | 0.0 I  | 0.0 I  | 4.4 I  | 2.0 I  |
| I | hommI | 3.0 I     | 0.5 I  | 1.3 I  | 0.5 I  | 0.0 I  | 0.0 I  | 1.7 I  | 0.6 I  |
| I | femmI | 3.0 I     | 0.7 I  | 1.3 I  | 0.7 I  | 0.0 I  | 0.0 I  | 1.7 I  | 0.8 I  |
| I | .dutI | 100.0 I   | 33.2 I | 4.2 I  | 3.5 I  | 11.1 I | 29.5 I | 84.7 I | 59.9 I |
| I | .geaI | 6.6 I     | 2.2 I  | 1.3 I  | 1.0 I  | 2.4 I  | 6.4 I  | 3.0 I  | 2.1 I  |
| I | massI | 4.1 I     | 1.4 I  | 0.5 I  | 0.4 I  | 1.4 I  | 3.9 I  | 2.2 I  | 1.6 I  |
| I | .md.I | 4.3 I     | 1.6 I  | 1.6 I  | 1.4 I  | 1.6 I  | 4.9 I  | 1.0 I  | 0.8 I  |
| I | ssm.I | 8.7 I     | 2.3 I  | 1.3 I  | 0.8 I  | 0.4 I  | 0.9 I  | 7.1 I  | 4.0 I  |
| I | autrI | 13.7 I    | 4.4 I  | 9.4 I  | 7.4 I  | 0.9 I  | 2.4 I  | 3.4 I  | 2.3 I  |
| I | ****I | **** I    | **** I | **** I | **** I | **** I | **** I | **** I | **** I |

Figure 5 : Corrélations et contributions des variables aux classes.

# BIBLIOGRAPHIE

- <Ben73> BENZECRI J.P.  
- "L'Analyse des Données" - Tome 2 "L'analyse des correspondances" - Dunod 1973
- <CaP76> CAILLIEZ F. et PAGES J.P.  
- "Introduction à l'Analyse des Données" - Smash 1976
- <Caz80> CAZES P.  
- "L'analyse de certains tableaux rectangulaires décomposés en blocs"-  
Les cahiers de l'Analyse des Données - vol 4, numéro 4, 1980, pp 387-406-
- <Did79> DIDAY E.  
- "Optimisation en Classification Automatique" - Inria 1979
- <EsP84> ESCOPIER B. et PAGES J.  
- "L'Analyse Factorielle Multiple" - Cahier du B.U.R.O. - numéro 42, 1984 -
- <Esc80> ESCOPIER Y.  
- "L'analyse conjointe de plusieurs matrices" - Société Française de Biométrie 1980
- <Gov83> GOVAERT G.  
- "Classification croisée" - Thèse de doctorat d'état - Université Paris VI 1983-
- <How69> HOWARD N.  
- "Least squares classification and principals components analysis: a comparaison in quantitative ecological analysis in the social science" - Dogan et Rokkan ed. Cambridge MIT Press 1969-
- <Ler79> LERMAN J.C.  
- "Les présentations factorielles de la classification "-  
Rairo, Recherche Opérationnelle - vol 13, numéro 2, 1979-
- <Mod82> Rapport MODULAD - INRIA 1982-
- <Nin81> NIN G.  
- "Maximisation de la trace et du Rv appliquée à la Classification Automatique"-  
Thèse de 3-ème cycle - Université de Provence 1981-
- <Ral86> RALAMBONDRAIN Y.  
- "Contribution à l'Analyse des Données : - Partie I : étude des tableaux n-aires,  
-Partie II : Le système SICLA"  
Doctorat d'état - Université Paris-Dauphine 1986-
- <Sch81> SCHWARTZ L.  
- "Les tenseurs" - Editions Hermann, 1981-

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique

