REVUE DE STATISTIQUE APPLIQUÉE

P. CAZES
A. CHOUAKRIA
E. DIDAY
Y. SCHEKTMAN

Extension de l'analyse en composantes principales à des données de type intervalle

Revue de statistique appliquée, tome 45, n° 3 (1997), p. 5-24
<http://www.numdam.org/item?id=RSA_1997__45_3_5_0>

Rev. Statistique Appliquée, 1997, **XLV** (3), 5-24

ENTENSION DE L'ANALYSE EN COMPOSANTES PRINCIPALES À DES DONNÉES DE TYPE INTERVALLE

P. Cazes (1), A. Chouakria (1), E. Diday (1), Y. Schektman (2)

(1) LISE-CEREMADE, Université Paris IX Dauphine,
Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16
(2) Maison de la Recherche, Université Toulouse le Mirail.

RÉSUMÉ

Les méthodes d'analyse factorielle sont largement utilisées; cependant, elles ne sont applicables qu'à des objets caractérisés par des variables monovaluées (la valeur prise par une variable pour un objet est une valeur unique). Dans cet article nous étendons l'analyse en composantes principales à des objets caractérisés par des variables multivaluées décrivant de la variation ou de l'imprécision (la valeur prise par une variable pour un objet est un intervalle de valeurs). Ce travail entre dans le cadre de l'analyse de données symboliques dont l'objectif est d'étendre les méthodes d'analyse de données classiques à des objets plus complexes, dits symboliques. Un exemple d'application sera présenté afin d'illustrer la fiabilité et l'efficacité de la nouvelle méthode.

Mots-clés : *Analyse en composantes principales, Analyse de données symboliques, Intervalle.*

ABSTRACT

Factorial analysis methods are extensively studied, however, these methods deal only with objects which are described by single-value features (the feature value, for each object, is single). The aim of the present paper is to extend the well known principal component analysis method to a particular kind of symbolic objects characterized by multi-values features of interval type data. An example is reported to corroborate the effectiveness of the method.

Keywords : *Symbolic Data Analysis, Principal Component Analysis, Interval.*

1. Introduction

Les bases de données statistiques manipulent des objets conventionnels décrits par des variables monovaluées (la valeur prise par une variable pour un objet est une valeur unique). Les évolutions récentes dans les systèmes de bases de données permettent de stocker de nouveaux types de données (intervalles, ensembles)

introduisant de l'imprécision ou de la variation. Des contraintes de domaines peuvent être exprimées et des liens de hiérarchie et de composition peuvent être stockés. Ces évolutions dans les systèmes de base de données ont donné lieu à de nombreuses applications manipulant des objets décrits de façon plus proches de la réalité et donc plus complexes que ceux habituellement traités. En gestion des stocks, par exemple, on décrit une situation de rupture de stock comme suit : «Niveau-de-stock = [100, 150], Quantité-en-cours-de-commande = [50, 100], Durée-de-livraison = [30, 45], État-fournisseur = {Critique, Mauvais}, État-écoulement-produit = {Moyen, Rapide} ». On peut décrire des contraintes entre des variables, par exemple, si «État-fournisseur = {Critique}» alors «Durée-de-livraison $\geqslant$ 40». On peut avoir des taxinomies, par exemple, dans la variable couleur les modalités Blanc et Jaune sont remplaçables par la modalité Claire. Des objets incluant dans leurs descriptions de telles informations sont dits **Symboliques** (Diday, 1987, 1995) (car dans chaque case du tableau de données peuvent apparaître des valeurs multiples, parfois pondérées et liées entre elles par des règles). L'extension des méthodes d'analyse des données à de tels objets est appelé «Analyse de Données Symboliques». Plusieurs auteurs se sont intéressés à l'extension des méthodes de réduction de dimension et de transformation de variables à des données complexes. Nagabushan (1988) a présenté une méthode de réduction à deux dimensions s'appliquant à des objets décrits par des variables à valeurs intervalles; cette méthode est basée sur les développements en séries de Taylor. Ichino (1994) s'est également intéressé aux problèmes de réduction de dimension; il propose une extension de la méthode d'Analyse en Composantes Principales (ACP) à des objets décrits par des variables de type intervalle, de type ensemble et même structurées. Ichino se base, pour étendre la méthode d'ACP classique à des données complexes, sur la généralisation de la distance de Minkowsky.

On présentera dans cet article, une nouvelle méthode dite **Méthode des Sommets** qui étend l'ACP à des données de type intervalle. Nous présenterons, par la suite, une seconde méthode, dite **Méthode des Centres**, plus adaptée, à première vue, aux données décrites par un nombre élevé de variables. Nous comparons ensuite ces deux méthodes, avant d'indiquer une généralisation probabiliste, puis de proposer une extension au cas de l'analyse des correspondances multiples. Un exemple d'application termine l'article.

2. Les données de type intervalle

Le résultat d'une observation ou d'une mesure peut être de deux types :

- Monovalué : le résultat de la mesure ou de l'observation de la variable est une valeur unique. Par exemple, les variables : âge d'une personne, niveau-de-gris d'un pixel.
- Multivalué : le résultat de la mesure ou de l'observation de la variable est un ensemble de valeurs ou un intervalle de valeurs selon que la variable est discrète ou continue.

  En botanique, par exemple, si les objets à étudier sont des plantes, la taille de la tige d'une plante est une valeur unique. Si, par contre, les objets auxquels on s'intéresse sont des espèces de plantes, la taille de la tige d'une espèce est

un intervalle de valeurs. Cet intervalle représente le domaine de variation de la taille de la tige sur tous les specimens appartenant à l'espèce en question.

L'ACP classique traite des tableaux de données de la forme I × J où I représente l'ensemble des objets et J celui des variables. La case du tableau, croisement de la i^{eme} ligne et de la j^{eme} colonne, contient la valeur observée x_{ij}, supposée unique, de la j^{eme} variable pour le i^{eme} objet. Nous proposons une nouvelle méthode qui permet de traiter des tableaux de données où x_{ij} est un intervalle de valeur, introduisant la variation ou l'imprécision : $x_{ij} = [\underline{x_{ij}}, \overline{x_{ij}}]$ où $\underline{x_{ij}}$, $\overline{x_{ij}}$ sont respectivement, la plus petite et la plus grande valeur observée, de la j^{eme} variable pour le i^{eme} objet.

2.1. Données du problème

Soient $S_1, \ldots, S_m$ m objets décrits par n variables $X_1, \ldots, X_n$ de type intervalle.

$$\begin{pmatrix} S_1 \\ \vdots \\ S_m \end{pmatrix} = \begin{pmatrix} x_{S_1 1} & \cdots & x_{S_1 n} \\ \vdots & \ddots & \vdots \\ x_{S_m 1} & \cdots & x_{S_m n} \end{pmatrix} \quad (1)$$

où $x_{S_i j} = [\underline{x_{ij}}, \overline{x_{ij}}]$ est la valeur de la variable X_j pour l'objet S_i.

2.2. Objectif

Classiquement, étant donné un ensemble d'objets décrits chacun par un vecteur $(x_{i1}, \cdots, x_{in})$, l'objectif de toute méthode de réduction de dimension, en particulier l'ACP, est de réduire le nombre de variables descriptives, tout en préservant la «structure de distribution» des objets. Soient $Y_1, \ldots, Y_p$ $(p < n)$ les nouvelles variables descriptives obtenues après réduction : chaque objet S_i sera décrit par un vecteur $(y_{i1}, \cdots, y_{ip})$ dans un espace de dimension plus faible.

De façon similaire, partant d'un ensemble d'objets S_i caractérisés chacun par un n-uple $\left([\underline{x_{i1}}, \overline{x_{i1}}], \cdots, [\underline{x_{in}}, \overline{x_{in}}]\right)$, l'objectif est de pouvoir décrire ces objets par un nombre restreint de variables nouvelles. Ces variables nouvelles devront non seulement préserver la structure de distribution des objets mais également conserver l'information de variation ou d'imprécision apportée par les variables de départ. Il s'agit en fait de décrire la structure de distribution des S_i dans un espace de dimension faible défini par des variables de type intervalle $Y_1, \ldots, Y_p$ $(p < n)$: chaque objet S_i sera alors décrit par un p-uple $\left([\underline{y_{i1}}, \overline{y_{i1}}], \cdots, [\underline{y_{ip}}, \overline{y_{ip}}]\right)$

3. Méthode des Sommets

Soit un objet S décrit par le n-uple $\left([\underline{x_1}, \overline{x_1}], \cdots, [\underline{x_n}, \overline{x_n}]\right)$. Cet objet peut être visualisé dans l'espace de description, par un hypercube à 2^n sommets. La longueur des côtés de l'hypercube est donnée par l'étendue des intervalles associés à chaque variable de description. Pour $n = 2$ l'objet S décrit par (2) est représenté par un

rectangle comme indiqué sur la figure 1.

$$S = \left([\underline{x_1}, \overline{x_1}], [\underline{x_2}, \overline{x_2}]\right) \tag{2}$$

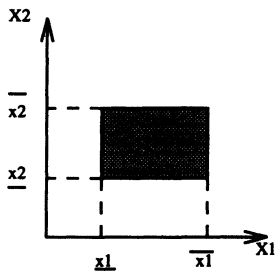

FIGURE 1
Représentation de l'objet S dans un espace à deux dimensions

Un hypercube, dans un espace de dimension n, peut être décrit par une matrice à 2^n lignes et n colonnes où la i^{eme} ligne correspond aux coordonnées du i^{eme} sommet. Dans ce cas-ci l'objet S défini en (2) sera décrit par la matrice suivante.

$$\begin{pmatrix} \underline{x_1} & \underline{x_2} \\ \underline{x_1} & \overline{x_2} \\ \overline{x_1} & \underline{x_2} \\ \overline{x_1} & \overline{x_2} \end{pmatrix} \tag{3}$$

Notons qu'un objet peut être caractérisé soit par un vecteur de composantes de type intervalles (2), soit par une matrice réelle (ou à éléments réels) (3).

3.1. Algorithme de la méthode des Sommets

1. Décrire chaque objet S_i par une matrice de données numériques M_i à 2^n lignes et n colonnes dont les éléments sont les coordonnées (n) des sommets (2^n) des hypercubes associés.
2. Construire une nouvelle matrice M à $2^n \times m$ lignes et n colonnes en concaténant les m matrices M_i précédentes. Ainsi au tableau (m x n) suivant,

$$S = \begin{pmatrix} S_1 \\ \vdots \\ S_m \end{pmatrix} = \begin{pmatrix} [\underline{x_{11}}, \overline{x_{11}}] & \cdots & [\underline{x_{1n}}, \overline{x_{1n}}] \\ \vdots & \ddots & \vdots \\ [\underline{x_{m1}}, \overline{x_{m1}}] & \cdots & [\underline{x_{mn}}, \overline{x_{mn}}] \end{pmatrix} \tag{4}$$

où chaque élément est un intervalle, on fait correspondre la matrice à $2^n \times m$ lignes et n colonnes suivante :

$$M = \begin{pmatrix} M_1 \\ \vdots \\ M_m \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \underline{x_{11}} & \cdots & \underline{x_{1n}} \\ \vdots & \ddots & \vdots \\ \overline{x_{11}} & \cdots & \overline{x_{1n}} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \underline{x_{m1}} & \cdots & \underline{x_{mn}} \\ \vdots & \ddots & \vdots \\ \overline{x_{m1}} & \cdots & \overline{x_{mn}} \end{bmatrix} \end{pmatrix} \tag{5}$$

De plus, à chacune des lignes de M (i.e à chaque sommet), on attribue un poids, à savoir $p_i/2^n$, s'il s'agit d'une ligne de la sous matrice M_i de M, p_i étant le poids de l'objet S_i ($\sum\{p_i |\ i = 1, m\} = 1$). On donne ainsi la même importance à chacun des 2^n sommets associés à S_i.

3. Appliquer l'ACP classique à la matrice M de données numériques définie en (5). Soient $Y_1, \ldots, Y_p$ $(p \leqslant n)$ les p premières composantes principales (à valeurs numériques) issues de cette ACP et $\lambda_1, \ldots, \lambda_p$ les valeurs propres associées.

4. Déterminer les composantes principales à valeurs intervalles $Y_1^I, \ldots, Y_p^I$ à partir des composantes numériques $Y_1, \ldots, Y_p$.

   Soit L_{S_i} l'ensemble des numéros de lignes dans la matrice M associés à l'objet S_i et y_{kj}, $k \in L_{S_i}$, la valeur de la j^{eme} composante principale numérique Y_j associée au sommet de l'objet S_i correspondant à la k^{eme} ligne de M. La valeur de la j^{eme} composante principale de type intervalle Y_j^I pour l'objet S_i est alors $y_{S_ij}^I = [\underline{y_{ij}}, \overline{y_{ij}}]$ avec;

$$\underline{y_{ij}} = \min_{k \in L_{S_i}} (y_{kj})$$
$$\overline{y_{ij}} = \max_{k \in L_{S_i}} (y_{kj})$$

3.2. Paramètres d'aide à l'interprétation

Les paramètres d'interprétation se généralisent très naturellement :

Pour mesurer la qualité de la représentation de l'objet S_i sur l'axe factoriel de direction u_j, on peut proposer l'une des deux formules suivantes :

$$COR_I^1(S_i, u_j) = \frac{\sum_{k \in L_{S_i}} p_k . y_{kj}^2}{\sum_{k \in L_{S_i}} p_k . d^2(k, G)} = \frac{\sum_{k \in L_{S_i}} y_{kj}^2}{\sum_{k \in L_{S_i}} d^2(k, G)} \tag{6}$$

$$COR_I^2(S_i, u_j) = \frac{1}{2^n} . \sum_{k \in L_{S_i}} \frac{y_{kj}^2}{d^2(k, G)} \tag{7}$$

L_{S_i} désignant toujours l'ensemble des 2^n sommets associés à l'objet S_i, p_k le poids du sommet k (égal à $p_i/2^n$), y_{kj} la coordonnée du sommet k sur l'axe factoriel de direction u_j, G le centre de gravité et $d(k,G)$ la distance entre k et G.

La première formule est le rapport entre la contribution de L_{S_i} à l'inertie λ_j de l'axe factoriel j et la contribution de L_{S_i} à l'inertie totale, tandis que la seconde correspond à la moyenne des cosinus carrés des angles entre chacun des 2^n sommets k de L_{S_i} et l'axe factoriel j.

On mesure de même la contribution de S_i

- à l'inertie λ_j du j^{eme} axe factoriel par :

$$CTR_I(S_i,u_j) = \frac{\sum\limits_{k\in L_{S_i}} p_k.y_{kj}^2}{\lambda_j} = \frac{p_i}{2^n.\lambda_j}.\sum_{k\in L_{S_i}} y_{kj}^2. \tag{8}$$

- à l'inertie totale du nuage des $m.2^n$ sommets associés aux m objets par :

$$INR_I(S_i) = \frac{\sum\limits_{k\in L_{S_i}} p_k.d^2(k,G)}{I_T} = \frac{p_i}{2^n}.\frac{\sum\limits_{k\in L_{S_i}} d^2(k,G)}{\sum\limits_{j=1}^{n}\lambda_j} \tag{9}$$

I_T désignant l'inertie totale.

Les deux contributions précédentes reviennent à sommer les contributions correspondantes des 2^n sommets associés à l'objet S_i.

4. Méthode des Centres

La méthode des Sommets risque de devenir coûteuse quand le nombre de variables descriptives est élevé. Nous proposons une nouvelle approche qui se base pour la détermination des axes factoriels sur l'information apportée par les centres d'hypercubes. Les intervalles de variation des composantes principales seront déterminés à partir des variations des variables de départ. On considère ici la matrice des centres d'hypercubes donnée en (10).

$$\begin{pmatrix} x_{11}^c & \cdots & x_{1n}^c \\ \vdots & \ddots & \vdots \\ x_{m1}^c & \cdots & x_{mn}^c \end{pmatrix} \tag{10}$$

avec

$$x_{ij}^c = \frac{\overline{x_{ij}} + \underline{x_{ij}}}{2} \tag{11}$$

Algorithme de la méthode des Centres

1. Transformer la matrice donnée en (4) en la matrice donnée en (10). Soient $X_1^c, \ldots, X_n^c$ les nouvelles variables numériques ainsi obtenues.
2. Appliquer l'ACP classique sur la matrice des centres obtenue à l'étape 1.
3. Déduire pour chaque objet les intervalles de variation sur les axes factoriels. Soit y_{ik}^c la coordonnée (numérique), sur le k^{eme} axe principal du point c_i (centre de l'hypercube associé à l'objet S_i) de coordonnées $(x_{i1}^c, \ldots, x_{in}^c)$. Cette valeur est obtenue à l'aide de la formule donnée en (12), où $\overline{X_j^c}$ est la moyenne de la variable X_j^c et u_{jk} la j^{eme} composante du k^{eme} vecteur axial factoriel :

$$y_{ik}^c = \sum_{j=1}^{n} (x_{ij}^c - \overline{X_j^c}).u_{jk} \tag{12}$$

Soit un point quelconque $x^r = (x_{i1}^r, \ldots, x_{in}^r)$ variant à l'intérieur de l'hypercube S_i ; l'objectif est de déterminer la plus petite valeur $\underline{y_{ik}}$ et la plus grande valeur $\overline{y_{ik}}$ prise par y_{ik}^r (coordonnée du point x^r sur l'axe factoriel k) quand les variables x_{ij}^r varient dans l'intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$ pour $j = 1, \ldots, n$.

Comme y_{ik}^r est une fonction linéaire des n variables $x_{i1}^r, \ldots, x_{in}^r$, variant indépendamment dans $[\underline{x_{ij}}, \overline{x_{ij}}]$ pour $j = 1, \ldots, n$, les valeurs extrêmes de y_{ik}^r sont données par les formules (13) et (14).

$$\underline{y_{ik}} = \sum_{j=1}^{n} \min_{\underline{x_{ij}} \leqslant x_{ij}^r \leqslant \overline{x_{ij}}} (x_{ij}^r - \overline{X_j^c}).u_{jk} \tag{13}$$

$$\overline{y_{ik}} = \sum_{j=1}^{n} \max_{\underline{x_{ij}} \leqslant x_{ij}^r \leqslant \overline{x_{ij}}} (x_{ij}^r - \overline{X_j^c}).u_{jk} \tag{14}$$

Soit encore,

$$\underline{y_{ik}} = \sum_{j, u_{jk}<0} (\overline{x_{ij}} - \overline{X_j^c})u_{jk} + \sum_{j, u_{jk}>0} (\underline{x_{ij}} - \overline{X_j^c})u_{jk} \tag{15}$$

$$\overline{y_{ik}} = \sum_{j, u_{jk}<0} (\underline{x_{ij}} - \overline{X_j^c})u_{jk} + \sum_{j, u_{jk}>0} (\overline{x_{ij}} - \overline{X_j^c})u_{jk} \tag{16}$$

5. Comparaison des deux méthodes

Nous allons voir que les matrices de variances V_s et V_c que l'on diagonalise dans la méthode des sommets et celle des centres respectivement ne diffèrent que par leurs termes diagonaux. Plaçons nous d'abord dans la méthode des centres, et

supposons, ce qui ne restreint pas la généralité que les variables sont centrées, soit :

$$\forall\ j = 1, n\ \ :\ \overline{X_j^c} = \sum_{i=1}^{m} p_i x_{ij}^c = 0 \tag{17}$$

p_i étant, rappelons-le, le poids de l'individu i.

Alors le terme général $(V_c)_{jj'}$ de la matrice variance dans la méthode des centres s'écrit :

$$(V_c)_{jj'} = \sum_{i=1}^{m} p_i x_{ij}^c x_{ij'}^c \tag{18}$$

Dans la méthode des sommets, chacun des 2^n sommets associés à l'individu i est affecté de la masse $p_i/2^n$. Si l'on considère la variable X_j, les valeurs $\underline{x_{ij}}$ et $\overline{x_{ij}}$ apparaîtront chacune 2^{n-1} fois. La moyenne $\overline{X_j}$ de X_j s'écrira donc :

$$\overline{X_j} = \sum_{i=1}^{m} \frac{p_i}{2^n}(2^{n-1}\underline{x_{ij}} + 2^{n-1}\overline{x_{ij}}) = \sum_{i=1}^{m} p_i x_{ij}^c = \overline{X_j^c} = 0 \tag{19}$$

On obtient donc la même moyenne que dans la méthode des centres, soit 0 puisqu'on a centré les variables.

De même la variance $(V_s)_{jj}$ de X_j s'écrit :

$$(V_s)_{jj} = \sum_{i=1}^{m} \frac{p_i}{2^n}[2^{n-1}(\underline{x_{ij}})^2 + 2^{n-1}(\overline{x_{ij}})^2] \tag{20}$$

$$= \sum_{i=1}^{m} \frac{p_i}{2}[(\underline{x_{ij}})^2 + (\overline{x_{ij}})^2] \tag{21}$$

soit encore (puisque $(a^2 + b^2)/2 = [(a+b)^2 + (a-b)^2]/4$) :

$$(V_s)_{jj} = \sum_{i=1}^{m} p_i[(x_{ij}^c)^2 + (\overline{x_{ij}} - \underline{x_{ij}})^2/4] \tag{22}$$

$$= (V_c)_{jj} + \sum_{i=1}^{m} p_i(\overline{x_{ij}} - \underline{x_{ij}})^2/4 \tag{23}$$

On obtient donc la variance calculée dans la méthode des centres (variance interclasses) augmentée d'un terme traduisant l'imprécision (variance intraclasses) puisque :

$$(\underline{x_{ij}} - \overline{x_{ij}})^2/4 = (1/2)(\underline{x_{ij}} - x_{ij}^c)^2 + (1/2)(\overline{x_{ij}} - x_{ij}^c)^2 \tag{24}$$

Dans le calcul de la covariance entre X_j et $X_{j'}$ dans la méthode des sommets, vu que le produit des coordonnées de chacun des quatre sommets du rectangle défini par $(\underline{x_{ij}}, \overline{x_{ij}})$ et $(\underline{x_{ij'}}, \overline{x_{ij'}})$ apparaît 2^{n-2} fois, on a, après mise en facteurs :

$$(V_s)_{jj} = \sum_{i=1}^{m} \frac{p_i}{2^n} 2^{n-2} (\underline{x_{ij}} + \overline{x_{ij}})(\underline{x_{ij'}} + \overline{x_{ij'}}) \tag{25}$$

$$= \sum_{i=1}^{m} p_i x_{ij}^c x_{ij'}^c = (V_c)_{jj'} \tag{26}$$

On obtient bien, comme annoncé, la même covariance que dans la méthode des centres. Il résulte des résultats précédents que si dans la méthode des Sommets, on n'effectue pas les calculs de contribution donnés au paragraphe 3.2, et si l'on se contente sur chaque axe factoriel de calculer pour un individu i la projection des sommets extrêmes (ce qui revient à déterminer les composantes principales à valeurs intervalles) la complexité dans la méthode des Sommets est en O(n) et est identique à celle de la méthode des Centres. En effet pour un individu i, sur l'axe factoriel k, les valeurs extrêmes $\underline{y_{ik}}$ et $\overline{y_{ik}}$ des projections des 2^n sommets associés à l'individu i sont données par les mêmes formules que dans la méthode des Centres (i.e. par les formules (15) et (16), où $\overline{X_j^c} = 0$ par hypothèse, et où u_{jk} est la j^{eme} composante du k^{eme} vecteur propre normé de la matrice V_s).

Remarque :

Si on a des données hétérogènes, on effectuera des ACP normées, et on diagonalisera donc les matrices de corrélation R_c et R_s respectivement associées à V_c et V_s.

C'est ce que l'on a fait dans l'application exposée à la fin de l'article. On peut noter que dans ce cas, on obtient bien sûr des corrélations différentes dans la méthode des centres et dans celle des sommets.

Une autre façon de procéder est de calculer les variances $(V_c)_{jj} = s_j^2$ à l'aide de la formule (18) (où $j = j'$) et de réduire les données en considérant les intervalles $(\frac{\underline{x_{ij}}}{s_j}, \frac{\overline{x_{ij}}}{s_j})$. Dans ce cas, la méthode des centres qui revient à diagonaliser la matrice R_c correspond à une ACP normée, alors que ce n'est pas le cas dans la méthode des sommets, les termes diagonaux de la matrice à diagonaliser étant dans ce dernier cas supérieurs à 1. Notons qu'on aurait pu de façon symétrique effectuer la réduction des données à partir des écarts-type $\sqrt{(V_s)_{jj}}$ calculés dans la méthode des sommets.

6. Généralisation à des modèles probabilistes

La méthodologie précédente, au niveau de la méthode des sommets, revient en fait à associer à chaque individu i et à chaque variable j la variable aléatoire $X_{ij} = x_{ij}^c + E_{ij}$, où les $(E_{ij},\ j = 1, n)$ pour i fixé sont des variables aléatoires de moyenne nulle, indépendantes (indépendance locale) discrètes prenant les valeurs $(\underline{x_{ij}} - x_{ij}^c)$ et $(\overline{x_{ij}} - x_{ij}^c)$ avec les probabilités 1/2.

On peut bien sûr faire d'autres hypothèses, comme par exemple une répartition uniforme, triangulaire ou normale tronquée de E_{ij} dans l'intervalle $[\underline{x_{ij}} - x_{ij}^c, \overline{x_{ij}} - x_{ij}^c]$. Avec ces hypothèses, si on considère que X_j est la variable aléatoire dont les lois de probabilité conditionnelles (relativement aux occurences des objets S_i) sont les lois de probabilité des X_{ij} $(i = 1, m)$ alors, on obtient comme moyenne de chaque variable X_j et comme covariance entre tout couple de variable X_j, $X_{j'}$ (j $\neq$j') les mêmes valeurs que dans la méthode des centres.

Par contre la variance de la variable X_j s'écrit :

$$Var(X_j) = (V_c)_{jj} + \sum_{i=1}^{m} p_i \alpha_{ij} (\overline{x_{ij}} - \underline{x_{ij}})^2 \qquad (27)$$

où le coefficient α_{ij} dépend de la loi de probabilité retenue pour E_{ij}. Par exemple, pour une loi uniforme, α_{ij} vaut 1/12. Notons que la méthode des centres rentre également dans le cadre précédent : il suffit, de prendre pour E_{ij} la répartition de Dirac de centre 0; le coefficient α_{ij} est alors nul.

Le tableau 1 donne la valeur de α_{ij} pour différentes lois. On voit que cette valeur est maximale dans le cas de la méthode des sommets et minimale dans le cas de la méthode des centres.

TABLEAU 1

Valeurs du coefficient α_{ij} pour quelques lois symétriques (f(-t)=f(t)). On s'est ramené à l'intervalle (-1/2,1/2).

Loi	Définition	α_{ij}	Méthode
Discrète	$f(-1/2) = f(1/2) = 1/2$	1/4	Sommets
Triangulaire Inverse	$f(t) = 4 \mid t \mid$	1/8	
Uniforme	$f(t) = 1$	1/12	
Triangulaire	$f(t) = 2 - 4 \mid t \mid$	1/24	
Normale Tronquée	Centrée Tronquée à 3 écarts-types	1/36	
Dirac	$f(0) = 1$	0	Centres

L'interêt de faire des hypothèses de lois continues dans les intervalles $[\underline{x_{ij}}, \overline{x_{ij}}]$ est d'une part d'avoir des méthodes intermédiaires entre la méthode des centres et celle des sommets, et d'autre part de pouvoir sur chaque axe factoriel donner des intervalles de confiance de niveau donné (95% ou 99% par exemple) pour chaque individu, intervalles plus petits que ceux obtenus à partir de la méthode des sommets ou celle des centres (formules (15)(16)).

Remarque : d'un point de vue théorique, on peut abandonner sans difficulté l'hypothèse d'indépendance locale : (27) reste vraie, mais on a Cov(X_j,$X_{j'}$) $\neq$

$(V_c)_{jj'}$. D'un point de vue pratique, ce choix peut s'avérer meilleur puisque dans le cas particulier où l'on a des variables identiques, on n'obtient qu'un seul axe factoriel (et non pas plusieurs). Les calculs détaillés correspondant aux différentes extensions précédentes quoique simples sortent du cadre de ce papier. Elles seront exposées dans un article ultérieur.

7. Extension au cas de l'analyse des correspondances multiples

Au lieu d'appliquer l'ACP, on peut effectuer une analyse des correspondances multiples avec un codage flou. Le principe de ce codage est défini ci dessous.

Soit V une variable prenant ses valeurs dans l'intervalle [a,b], intervalle découpé en t classes c_1 =[a_0,a_1], c_2 =[a_1,a_2],...,c_t=[a_{t-1},a_t]; avec a_0=a, a_t=b . Si un objet S_i est caractérisé par un intervalle $[\underline{vi}, \overline{vi}]$ de V de longueur non nulle, on adoptera le codage suivant :

$$\forall l = 1..t : \; k(i, c_l) = \frac{\text{Longueur}\,([a_{l-1}, a_l] \cap [\underline{vi}, \overline{vi}])}{(\overline{vi} - \underline{vi})} \tag{28}$$

soit,

1. Si $[a_{l-1}, a_l] \cap [\underline{vi}, \overline{vi}] = \phi$ alors $k(i, c_l) = 0$
2. Si $[a_{l-1}, a_l] \cap (\underline{vi}, \overline{vi}] \neq \phi$ alors

$$k(i, c_l) = \frac{\text{Min}(a_l, \overline{vi}) - \text{Max}(a_{l-1}, \underline{vi})}{(\overline{vi} - \underline{vi})} \tag{29}$$

3. Si $\underline{vi} = \overline{vi} = vi$, on adoptera le codage disjonctif usuel, à savoir :

$$k(i, c_l) \begin{cases} 1 & \text{si} \quad vi \in c_l \\ 0 & \text{sinon} \end{cases} \tag{30}$$

Si on a n variables $V1, V2, ...Vn$ et si on désigne par Jq l'ensemble des classes de la variable Vq et par J l'union disjointe des Jq, on fera l'analyse des correspondances du tableau k_{IJ} (I désignant l'ensemble des m objets) juxtaposition des blocs k_{IJ_q} obtenus en appliquant le codage précédent à chaque variable. Notons que le tableau k_{IJ} se réduit au tableau disjonctif complet usuel si on a des variables numériques au lieu de variables intervalles. Notons également que si l'on a des variables qualitatives définies de façon imprécise, ou un mélange de ces variables avec des variables intervalles, la méthodologie précédente reste valable en définissant pour chaque variable qualitative un codage flou tenant compte de l'imprécision.

Si l'on veut obtenir des composantes principales de type intervalle, il suffit d'associer à chaque objet Si un ensemble de sommets L_{Si} et de prendre sur un axe factoriel les coordonnées extrêmes des projections de cet ensemble de sommets. De façon précise, si l'objet Si possède un codage non nul pour u_1 sommets de J_1, u_2

sommets de J_2,..., u_n sommets pour J_n, il lui correspondra $u_1u_2...u_n$ sommets (en codage disjonctif complet).

Un exemple de sommets associé à un objet Si est donné sur le tableau 2 dans le cas où l'on a 2 variables la première à 3 modalités, la seconde à 4 modalités.

TABLEAU 2
Exemple de sommets associé au codage flou d'un objet Si

		V1			V2				Poids
		Modalités			Modalités				
		1	2	3	1	2	3	4	
Si		0.4	0.6	0	0	0.2	0.5	0.3	
L_{si}	(1,2)	1	0	0	0	1	0	0	0.08
	(1,3)	1	0	0	0	0	1	0	0.2
	(1,4)	1	0	0	0	0	0	1	0.12
	(2,2)	0	1	0	0	1	0	0	0.12
	(2,3)	0	1	0	0	0	1	0	0.3
	(2,4)	0	1	0	0	0	0	1	0.18

Cette méthodologie est analogue à la méthode des centres étudiée au paragraphe 4, mais où il faut remplacer l'ACP par l'Analyse Factorielle des Correspondances Multiples (AFCM). On peut également définir en AFCM l'analogue de la méthode des sommets développée au paragraphe 3; il suffit dans l'exemple précédent de remplacer la ligne associé à l'objet Si par les 6 lignes associées à chacun des 6 sommets correspondant à Si, chaque ligne étant pondérée(i.e multipliée) par un poids (donné dans la dernière colonne du tableau 2) égal au produit des codages de Si pour le sommet correspondant. On obtient ainsi un tableau $K_{L_{Si},J}$.

Opérant ainsi pour chaque individu, on obtient un tableau $K_{LJ}(L = \cup\{Lsi \mid i = 1, m\})$ concaténation des $K_{L_{Si}J}$, tableau dont on fera l'analyse des correspondances.

Compte tenu du principe d'équivalence distributionnelle, cette analyse revient à effectuer l'analyse des correspondances du tableau K_{TJ} suivant, où

$$T = \Pi\{J_q|\text{q= 1..n }\} \tag{31}$$

$$\forall t = (j_1, j_2, ..., j_n) \in T, \forall j'_q \in J_q \subset J \tag{32}$$

$$K(t, j'_q) = \begin{cases} 0 & \text{si } j'_q \neq j_q \\ \sum_{i=1}^{m} k_(i, j_1)k_(i, j_2)...k_(i, j_n) & \text{sinon} \end{cases} \tag{33}$$

On obtiendra comme précédemment des composantes principales de type intervalle en considérant l'ensemble des sommets L_{Si} associé à un objet Si. Cette méthodologie est en cours d'étude.

Remarques :

1. L'analyse interclasses du tableau K_{LJ}, L étant muni de la partition définie par les Si n'est rien d'autre que l'analyse des correspondances du tableau K_{IJ}. L'analyse de ce dernier tableau correspond donc bien à l'analyse des correspondances des centres des classes du tableau K_{LJ}.
2. Contrairement à la méthode des sommets en ACP, la taille du tableau analysé ne dépend pas de l'ensemble I des objets (soit $m2^n$ lignes) mais du produit $T = \Pi\{J_q \mid q = 1, n\}$ (soit $r_1.r_2....r_n$ lignes si l'on désigne par r_q le cardinal de J_q).

8. Exemple des Huiles

8.1. Description des données

Afin d'illustrer les méthodes proposées nous utilisons les données d'Ichino (1994) de la table 3. Chaque ligne du tableau représente une classe d'huile décrite par 4 variables quantitatives : «Specific gravity», «Freezing point», «Iodine value», «Saponification». L'intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$, croisement de la i^{eme} ligne et de la j^{eme} colonne signifie que la valeur de la j^{eme} variable pour toute huile appartenant à la i^{eme} classe d'huile, appartient à l'intervalle $[\underline{x_{ij}}, \overline{x_{ij}}]$.

TABLEAU 3
La description des 8 classes d'huile par 4 variables de type intervalle

Nom	Label	GRA	FRE	IOD	SAP
Linseed	L	[0.93,0.94]	[−27.00,−18.00]	[170.00,204.00]	[118.00,196.00]
Perilla	P	[0.93,0.94]	[−5.00,−4.00]	[192.00,208.00]	[188.00,197.00]
Cotton	Co	[0.92,0.92]	[−6.00,−1.00]	[99.00,113.00]	[189.00,198.00]
Sesame	S	[0.92,0.93]	[−6.00,−4.00]	[104.00,116.00]	[187.00,193.00]
Camellia	Ca	[0.92,0.92]	[−21.00,−15.00]	[80.00,82.00]	[189.00,193.00]
Olive	O	[0.91,0.92]	[0.00,6.00]	[79.00,90.00]	[187.00,196.00]
Beef	B	[0.86,0.87]	[30.00,38.00]	[40.00,48.00]	[190.00,199.00]
Hog	H	[0.86,0.86]	[22.00,32.00]	[53.00,77.00]	[190.00,202.00]

Afin de réduire l'espace de description des 8 classes d'huile on utilise la méthode des sommets puis celle des centres.

8.2. Résultats

Pour chacune des méthodes (Sommets, puis Centres) utilisées, on a effectué une ACP normée, puisque les variables sont hétérogènes.

1. Méthode des Sommets

Les valeurs propres et les pourcentages d'inerties figurent dans la table 4, tandis que les deux premières composantes principales de type intervalle sont données dans la table 5. Chaque classe d'huile caractérisée par les deux composantes principales de type intervalle est visualisée dans le plan factoriel (1,2) par un rectangle (figure 2). Les corrélations entre les variables initiales et les composantes principales sont données dans la table 6, et visualisées sur la figure 3.

TABLEAU 4
Valeurs propres et % d'inertie

	Méthode des Sommets		
Numéro	Valeur-propres	% d'inertie	Cumul
1	2.7316	68.29	68.29
2	0.8093	20.23	88.52
3	0.3801	9.50	98.02
4	0.0790	1.98	100

TABLEAU 5
Les deux premières composantes principales de type intervalle des 8 classes d'huile

	Méthode des Sommets	
Label	CP1	CP2
L	[−3.58,−1.43]	[−3.04,1.10]
P	[−1.76,−1.22]	[0.36,0.95]
Co	[−0.45,−0.01]	[0.16,0.67]
S	[−0.71,−0.23]	[0.09,0.53]
Ca	[−0.58,−0.32]	[0.27,0.53]
O	[−0.09,0.56]	[−0.14,0.49]
B	[2.26,2.93]	[−0.87,−0.23]
H	[1.95,2.68]	[−0.80,−0.07]

TABLEAU 6
Corrélations entre variables descriptives et composantes principales

	Méthode des Sommets			
	CP1	CP2	CP3	CP4
GRA	−0.93	0.27	−0.11	0.21
FRE	0.92	−0.16	0.33	0.17
IOD	−0.86	0.06	0.51	−0.06
SAP	0.54	0.84	0.06	−0.03

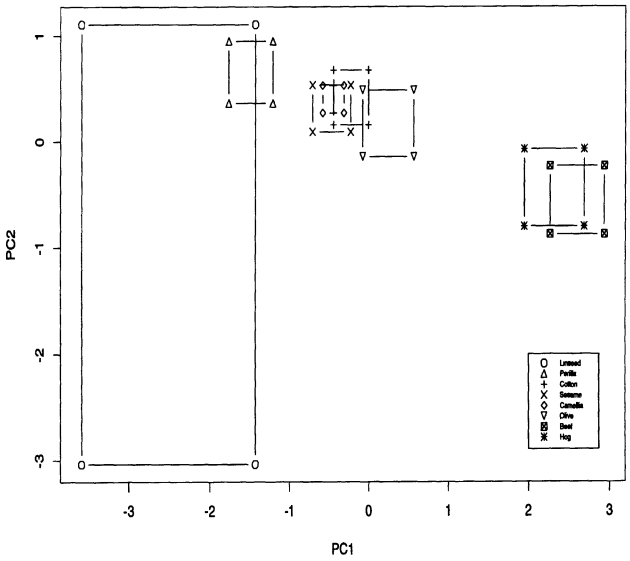

FIGURE 2
Projection des 8 rectangles associés aux 8 classes d'huile (méthode des Sommets)

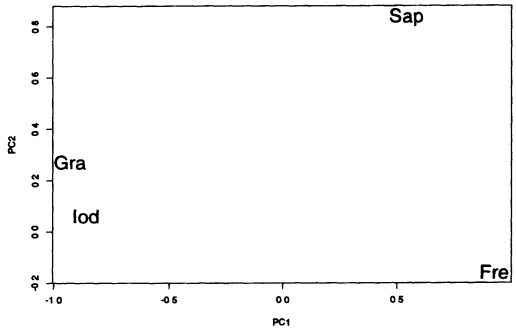

FIGURE 3
Cercle des corrélations (méthode des Sommets)

2. Méthode des Centres

Les valeurs propres et les pourcentages d'inerties sont donnés dans la table 7, les deux premières composantes principales des 8 classes d'huile dans la table 8 tandis que les rectangles associés sont visualisés sur la figure 4. Les corrélations entre les variables descriptives et les composantes principales figurent dans la table 9, et sont représentées sur la figure 5.

TABLEAU 7
Valeurs propres et % d'inertie

	Méthode des Centres		
Numéro	Valeur-propres	% d'inertie	Cumul
1	3.0094	75.24	75.24
2	0.6037	15.09	90.33
3	0.3483	8.71	99.04
4	0.0386	0.96	100

TABLEAU 8
Les deux premières composantes principales de type intervalle des 8 classes d'huile

	Méthode des Centres	
Label	CP1	CP2
L	[−4.80,−1.25]	[−4.46,1.40]
P	[−1.72,−1.03]	[0.32,1.15]
Co	[−0.42,0.18]	[0.26,0.98]
S	[−0.70,−0.13]	[0.15,0.78]
Ca	[−0.55,−0.21]	[0.48,0.85]
O	[−0.09,0.69]	[−0.13,0.77]
B	[2.23,3.04]	[−1.15,−0.23]
H	[1.91,2.85]	[−1.09,−0.07]

TABLEAU 9
Corrélations entre variables descriptives et composantes principales

	Méthode des Centres			
	CP1	CP2	CP3	CP4
GRA	−0.92	0.35	−0.05	0.14
FRE	0.92	−0.2	0.3	0.12
IOD	−0.87	−0.03	0.49	−0.05
SAP	0.74	0.66	0.14	−0.04

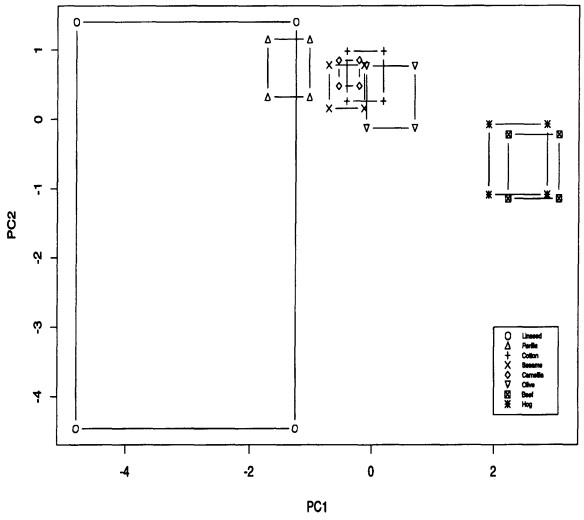

FIGURE 4
Projection des 8 rectangles associés aux 8 classes d'huile (méthode des Centres)

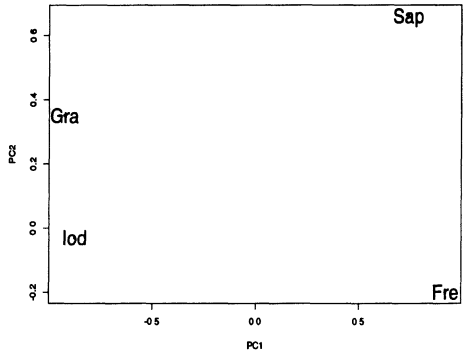

FIGURE 5
Cercle des corrélations (méthode des Centres)

8.3. Comparaison des résultats des deux méthodes

Les résultats obtenus à partir de la Méthode des Sommet et de la Méthode des Centres sont similaires. En effet, Il ressort, dans les deux cas, trois principaux groupements de classe d'huile : {PER, LIN}, {COT, SES, CAM, OLI} et {BEF, HOG}. La représentation simultanée des variables descriptives et des classes d'huile permet de fournir les variables caractérisant tel ou tel groupement d'huile. Notons , par exemple, que des valeurs élevées des variables GRA et IOD caractérisent fortement le groupe {LIN, PER} et l'opposent au groupe {BEF, HOG}. Le groupe {BEF, HOG} est, par contre, caractérisé par des valeurs élevées de la variable FRE ainsi que de SAP. Remarquons que la méthode d'Ichino basée sur la généralisation de la métrique de Minkowsky fournit des résultats similaires (*cf.* figure 6) en donnant les 3 groupes trouvés avec les deux méthodes développées.

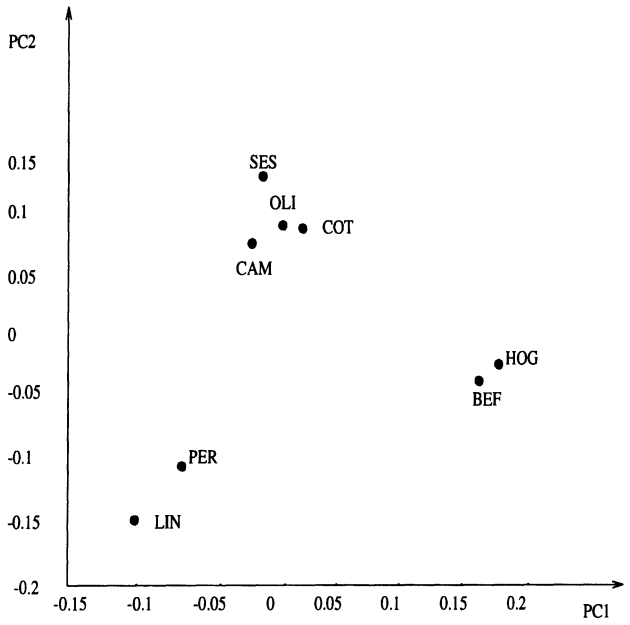

FIGURE 6
Projection des 8 classes d'huile par la méthode de Minkowsky généralisée

On peut noter que le taux d'inertie du premier axe factoriel est légèrement supérieur dans la méthode des centres que dans la méthode des sommets (75% contre 68%), ce qui est logique puisque dans cette dernière méthode le tableau analysé comporte beaucoup plus de lignes. Par contre les deux méthodes donnent pratiquement le même pourcentage d'inertie dans le plan (90.3% pour la méthode des centres contre 88.5% pour celle des sommets). Pour comparer de façon plus précise les résultats fournis par ces deux méthodes, on calcule les corrélations entre les composantes principales issues des deux analyses, ces corrélations étant calculées à partir des sommets (rajoutés en éléments supplémentaires dans la seconde analyse).

Les corrélations entre les composantes principales de même rang sont supérieures ou égales à 0.93, ces corrélations étant même supérieures à 0.97 si on se limite aux 3 premières composantes principales. Par ailleurs les cosinus entre axes factoriels de même rang issus des deux analyses sont toujours supérieurs à 0.96. Les deux méthodes donnent donc ici des résultats équivalents, ce qui semble logique, les intervalles de variations des variables étant (sauf pour la variable SAP et la classe d'huile Linseed) relativement faibles.

9. Conclusion

Les mesures caractérisant les objets traités par les méthodes de l'analyse des données et de la statistique ne sont pas toujours le résultat direct d'une observation unique et précise. Souvent, le résultat d'une observation est un ensemble de valeurs ou un intervalle de valeurs. Les méthodes d'ACP étendues aux intervalles trouvent leur intérêt quand l'expert est confronté à l'analyse d'objets caractérisés par des variables multivaluées de type intervalle. L'expert peut alors, selon l'objectif à atteindre, procéder de deux manières. Si l'objectif de l'analyse est d'estimer et de connaître la tendance globale de la dispersion des objets, il peut alors quantifier chaque intervalle par son centre puis appliquer une ACP classique aux centres des intervalles. Si, par contre, l'objectif de l'analyse consiste, d'une part, à étudier la dispersion globale des objets et d'autre part à savoir comment évolue la position (la dispersion) de chaque objet quand les valeurs des variables observées de départ varient dans leurs intervalles respectifs, il est alors nécessaire de tenir compte des valeurs de type intervalle de départ. Les méthodes d'ACP étendues aux intervalles répondent à un tel objectif. La visualisation à l'aide de rectangles permet d'une part, de localiser le champ de dispersion de chaque objet quand les valeurs observées varient dans leurs intervalles respectifs et d'autre part, de comparer l'amplitude de la dispersion des différents objets traités.

En considérant que chaque objet caractérisé par des variables de type intervalle définit une classe, l'analyse qui a pour objectif d'étudier la tendance globale de la dispersion des objets revient à une analyse interclasses, alors que l'analyse qui s'interesse en plus à la dispersion de chaque objet est une analyse inter et intra classes.

D'autres types d'analyse factorielle actuellement en cours d'étude peuvent être envisagés pour traiter des données de type intervalles, ensemblistes, dotées de structure taxinomique *a priori* etc. Dans le cas des ACP, on peut reporter les problèmes de dispersion non sur les individus mais sur les variables, en remplaçant chaque variable X par 2 variables $X\min$ et $X\max$ respectivement associées à la valeur minimale et à la valeur maximale de l'intervalle caractérisant chaque individu pour la variable X. Dans le cas de l'analyse factorielle des correspondances multiples nous étudions des méthodes de codages pour des variables de type intervalle et des variables qualitatives dotées d'une structure taxinomique *a priori* (Chouakria et al., 1996).

Références bibliographiques

CAZES P. L'analyse de certains tableaux rectangulaires décomposés en blocs : généralisation des propriétés rencontrées dans l'analyse des correspondances multiples. II Questionnaires : variantes de codages et nouveaux calculs de contributions. Les Cahiers de l'analyse des données. Vol. V, n°4, pp. 387- 403, 1980.

CAZES P. L'analyse de certains tableaux rectangulaires décomposés en blocs : généralisation des propriétés rencontrées dans l'analyse des correspondances multiples. IV Cas modèles. Les Cahiers de l'analyse des données. Vol. VI, n°2, pp. 135-143, 1981.

CHOUAKRIA A., DIDAY E., CAZES P. Extension of Principal Components Analysis to interval data. NTTS '95 : New Techniques and Technologies for Statistics, Bonn, novembre 1995.

CHOUAKRIA A., CAZES P., DIDAY E. Extension de l'analyse factorielle des correspondances multiples à des données de type intervalle et de type ensemble. SFC 96 : Actes de la 3ème rencontre de la Société francophone de classification, Namur, septembre 1995.

CHOUAKRIA A., VERDE R., DIDAY E., CAZES P. Généralisation de l'analyse factorielle des correspondances multiples à des objets symboliques. SFC 96 : Actes des la 4ème rencontre de la la Société francophone de classification, Vannes, septembre 1996.

DIDAY E. The symbolic approach in clustering and related methods of Data Analysis : The basic choices. IFCS, Aachen, 1987.

DIDAY E. From Data to Knowledge : Probabilist Objects for a Symbolic Data Analysis. DIMACS : series in discrete mathematics and theoretical computer science, volume 19, 1995.

GALLEGO F.J. Codage flou en analyse des correspondances. Les Cahiers de l'analyse des données. Vol. VII, n° 4, pp. 413-430, 1982.

ICHINO M. Generalized Minkowsky metrics for mixed feature type data analysis. IEEE, transactions on systems, man and cybernetics, vol. 24, n°4, 1994.

LEROY B., CHOUAKRIA A., HERLIN I., DIDAY E. Approche géométrique et classification pour la reconnaissance de visage. RFIA 96 : Reconnaissance de forme et intelligence artificielle, janvier 1996.

NAGABUSHAN P. An efficient method for classifying remotely sensed data, incorporating dimensionality reduction. Ph.d Thesis, Mysore University, India, 1988.

NAGABUSHAN P., CHIDANANDA K., DIDAY E. Dimentionality reduction of symbolic data. Pattern recognition letters. 16 (1995) 219- 223.

SAPORTA G. Probabilités, analyse des données et statistiques, Technip, 1990.