Résumé

Ce document traite des données d'intervalles par l'analyse en composantes principales (ACP).

Des unités statistiques décrites par des données d'intervalles peuvent être considérées comme objets symboliques.

Dans l'analyse des données symboliques ces données sont représentées par des hyper cubes.

Nous proposons quelques prolongements de l'ACP dans le but de la représentation dans un espace à dimension réduite, des images telle que des hyper cubes précisant les différences et les similitudes selon leurs dispositifs structuraux.

Mots-clés : Données d'intervalles ; données symboliques ; Analyse en composantes principale

INTRODUCTION

Des méthodes statistiques ont été principalement développées pour l'analyse de variables monovaluées (la valeur prise par une variable pour un objet est unique). Cependant dans la vie courante il y a beaucoup de situation dans lesquelles ces types de variables peuvent causer la perte grave d'informations. Traitant dans plusieurs domaines, des variables quantitatives, des informations plus complètes sont sûrement obtenues en décrivant un ensemble d'unités statistiques en termes de données d'intervalles. Par exemple les valeurs maximales et minimales des températures quotidiennes enregistrées offrent une vue plus réaliste que les températures moyennes simples. Un autre exemple peut être donné par la série financière: le minimum et le maximum, des prix de transaction quotidiennement enregistrés pour un ensemble de stocks, représentent les informations les plus appropriées pour les experts afin d'évaluer la tendance et la volatilité courantes dans le même jour.

Dans ce mémoire nous ne considérons pas la valeur centrale d'un intervalle mais nous attirons l'attention sur les valeurs minimales et maximales uniquement. Celles-ci sont évaluées comme deux aspects différents relatifs au même phénomène.

Le traitement statistique de données d'intervalles, a été récemment considéré dans le contexte de l' ANALYSE DE DONNEES SYMBOLIQUES (ADS) par [2], dont le but est de prolonger des méthodes statistiques à l'étude des structures de données plus complexes que le modèle tabulaire individusXvariables. Le mémoire présente quelques nouvelles techniques ACP pour visualiser et comparer les structures de données d'intervalles.
Des unités statistiques décrites par des variables d'intervalles peuvent être considérées comme cas spéciaux de données symboliques dans lesquelles seulement des variables quantitatives sont considérées.

D'ailleurs l'ADS, dans le traitement de données d'intervalles, apporte des informations très utiles dans l'interprétation des résultats et dans la représentation des unités statistiques.

Notre travail comprend trois chapitres:

Chapitre 1: ANALYSE EN COMPOSANTES PRINCIPALES (ACP)
Chapitre 2: STRUCTURE DE DONNEES SYMBOLIQUES
Chapitre 3: ANALYSE EN COMPOSANTES PRINCIPALES DES SOMMETS (ACPS)

Dans le premier chapitre nous étudierons le rôle et les techniques de l' ACP sur un tableau de données quantitatives en approfondissant la notion d'inertie expliquée par un axe.
Et dans le chapitre suivant nous parlerons de la caractérisation des objets et données symboliques et de quelques applications des méthodes de statistiques descriptives unidimensionnelles sur des données symboliques.
Par la suite, nous terminerons notre mémoire par le troisième chapitre où nous proposerons quelques prolongements de l 'ACP dans le but de r eprésenter par des hypercubes, dans un e space à dimensions réduites, les unités statiques décrites par des données du type intervalle et considérés comme objets symboliques par Carlo N.Lauro et Francesco Palumbo.

Chapitre 1:

ANALYSE EN COMPOSANTES PRINCIPALES

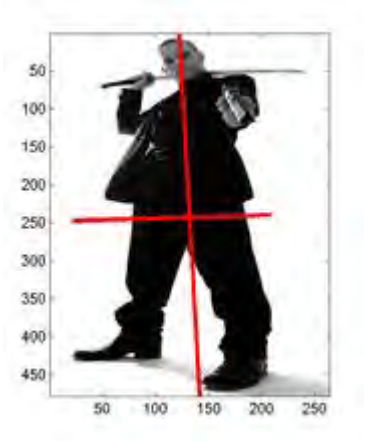
1.1 Introduction [13]

L'Analyse en Composantes Principales (ACP) est une méthode mathématique d'analyse de données qui permet d'analyser tout tableau de données représentant N individus décrits par p -variables quantitatives (où N et p sont des entiers naturels) et de rechercher les directions de l'espace qui représentent le mieux les corrélations entre p variables aléatoires. Elle est aussi connue sous le nom de transformée de KARHUNEN-LOEVE ou transformée de HOTELLING.

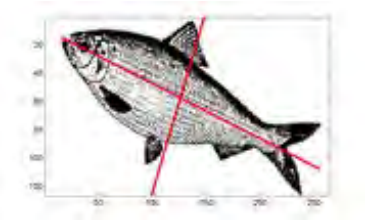
Lorsqu'on veut «compresser» un ensemble de p variables aléatoires, les n premiers axes de l'ACP $n \leq p$ sont un meilleur choix du point de vue de l'inertie expliquée.

L'ACP a pour but de comprendre et de visualiser comment les effets de phénomènes a priori isolés se combinent. Lorsque l'on ne considère que deux effets, il est usuel de caractériser leurs effets conjoints via le coefficient de corrélation (son seul défaut est de ne prendre en compte que des effets conjoints linéaires, ce qui se remarque en regardant les coefficients d'une régression linéaire).Lorsqu'on se place en dimension deux, les points disponibles (l'échantillon de points tirés suivant la loi conjointe de X_1 et X_2) peuvent être

représentés sur un plan. Le résultat d'une ACP sur ce plan est de déterminer les deux axes qui expliquent le mieux la dispersion des points disponibles.
Les figures ci-dessous représentent ces deux axes si on prend comme points ceux d'une photographie.



LES DEUX AXES D'UNE ACP SUR LA PHOTO D'UN HOMME DEBOUT



LES DEUX AXES D'UNE ACP SUR LA PHOTO D'UN POISSON

Lorsqu'il y a plus de deux effets, par exemple trois effets X_1, X_2, X_3 il y a trois coefficients de corrélation à prendre en compte: $\square_{X_1, X_2}$, $\square_{X_1, X_3}$ et $\square_{X_2, X_3}$.

La question qui a donné naissance à l'ACP est: comment avoir une intuition rapide des effets conjoints?
En dimension plus grande que deux, une ACP va toujours déterminer les axes (si on est en dimension 256, il y aura 256 axes à déterminer) qui expliquent le mieux la dispersion du nuage de points disponibles (de la photographie de ces points).
Elle va aussi les ordonner par l'inertie expliquée (dans l'image l'homme au pistolet à gauche, l'axe expliquant le plus d'inertie est l'axe vertical). Si on décide de retenir que les deux premiers axes de l'ACP, on pourra alors projeter notre nuage de dimension 256 sur un plan, et le visualiser.

Même si l'ACP est majoritairement utilisée pour visualiser des données, il ne faut pas oublier que c'est aussi un moyen:

.de décorréler ces données: dans la nouvelle base constituée des nouveaux axes, les points ont une corrélation nulle,

.de débruiter ces données, en considérant que les axes que l'on décide d'oublier sont des axes bruités;

.et même de classifier ces amas (clusters) corrélés.

1.2 Tableau de données

On observe sur chaque individu i , les variables aléatoires quantitatives X_j où $1 \leq i \leq N$ et $1 \leq j \leq p$. Ainsi nous obtiendrons un tableau de données représenté dans la suite où les individus et les variables sont respectivement donnés par les lignes et les colonnes.

VARIABLES / INDIVIDUS	1	2	. . .	j	. . .	p
1						
2						
. . .						
i						
. . .						
N						

Représentons les données dans une matrice appelée X à N lignes et p colonnes.

On applique usuellement une ACP sur une matrice de données quantitatives:

$$X = \begin{pmatrix} X_{1,1} & \cdots & X_{1,p} \\ \vdots & \ddots & \vdots \\ X_{N,1} & \cdots & X_{N,p} \end{pmatrix}$$

Chaque variable aléatoire c'est à dire chaque colonne a une moyenne $\bar{X}_j$ et un écart-type σ_j .

Le vecteur $[\bar{X}_1, \ldots\ldots\ldots, \bar{X}_p]$ est le centre de gravité du nuage de points; on le note souvent **g**.

La matrice X est généralement centrée sur le centre de gravité:

$$M = \begin{pmatrix} X_{1,1} - \bar{X}_1 & \cdots & X_{1,p} - \bar{X}_p \\ \vdots & \ddots & \vdots \\ X_{N,1} - \bar{X}_1 & \cdots & X_{N,p} - \bar{X}_p \end{pmatrix}$$

Elle peut être aussi réduite:

$$\begin{pmatrix} \frac{X_{1,1} - \bar{X}_1}{\sigma[X]_1} & \cdots & \frac{X_{1,p} - \bar{X}_p}{\sigma[X]_p} \end{pmatrix}$$

$$M_1 = \begin{pmatrix} \vdots & & \ddots & & \vdots \\ \frac{X_{N,1} - \bar{X}_1}{\sigma(X_1)} & \cdots & & \cdots & \frac{X_{N,p} - \bar{X}_p}{\sigma(X_p)} \end{pmatrix}$$

Cette transformation de la matrice X permet d'éliminer l'effet arbitraire du au choix des unités de mesure; ainsi dans M_1 les variables sont sans unité de mesure.

Le choix de réduire ou non le nuage de points (c'est à dire les *p*- réalisations de la variable aléatoire $(X_1, \ldots, X_p)$) est un choix de modèle:

.si on ne réduit pas le nuage: une variable à forte variance va «tirer» tout l'effet de l'ACP à elle;

.si on réduit le nuage: une variable qui n'est qu'un bruit va se retrouver avec une variance apparente égale à une variable informative.

1.3 Démarche de l'ACP

Une fois la matrice X transformée en $\bar{M}$ ou M_1 il suffit de la multiplier par sa transposée pour obtenir:

.la matrice de variance-covariance des $X_1, \ldots, X_p$ si X n'est pas réduite;

.la matrice de corrélation des $X_1, \ldots, X_p$ si X est réduite.

La matrice de covariance et de corrélation sont respectivement données par les formules suivantes:

$$Covariances = \frac{1}{N} \bar{M}' \bar{M}$$

$$Corrélations = \frac{1}{N} M_1' M_1$$

Ces deux matrices sont carrées (d'ordre N), symétriques et réelles. Elles sont donc diagonalisables

Dans la suite de ce mémoire, nous considérerons que le nuage est transformé (centré et réduit si besoin est). Chaque X_{ij} est donc remplacé par $X_{ij} - \overline{X_{ij}}$ ou $\frac{X_{ij} - \overline{X_{ij}}}{\sigma_j}$.

Nous utiliserons donc la matrice X pour noter $\overline{M}$ ou M suivant le cas.

1.3.1 Recherche du premier axe factoriel [18]

Définition 1.3.1.1

Soit $N = \{X^i, m_i\}$ un nuage de points dont chacun est muni d'une masse m_i où $i = 1, \ldots, N$.

L'inertie de N par rapport à un point P,quelconque est définie par:

$$I_{np}(N) = \sum_{i=1}^{N} m_i (X^i - P)^2$$

Définition 1.3.1.2

On appelle inertie par rapport à P expliquée par U, l'inertie des points Z^i projection orthogonale de X^i sur U passant par P (à Z^i est associé la masse m_i) où U est le vecteur unitaire de $\mathbb{R}^p$.

Elle s'exprime comme suit : $I_{np}(U) = \sum_{i=1}^{N} m_i (Z^i - P)^2$

Proposition 1

$I_{np}(U) = U'VU$ et $V = M'DM$ où V est la matrice d'inertie du nuage.

Preuve

Supposons que P soit à l'origine des axes :

$$I_{np}(U) = \sum_i m_i (z_i)^2 = \sum_{j,j'} m_i X_{ij} X_{ij'} U_j U_{j'}$$

Or $(z_i)^2 = (x^i U)^2 = \left(\sum_j X_{ij} U_j\right)^2 = U' X^i X^{i\prime} U$

Puisque $X^i X^{i\prime}$ est la matrice d'éléments $X_{ij} X_{ij'}$ et d'ordre $p * p$ donc

$$I_{np}(U) = \sum_i m_i U' X^i X^{i\prime} U = U' \sum_i m_i X^i X^{i\prime} U$$

D'où

$I_{np}(U) = U'VU$ où $V = \sum_i m_i X^i X^{i\prime} = [v_{jj'}]$, V est la matrice d'inertie du nuage.

1.3.1.1 Analyse factorielle du nuage des individus

Considérons le nuage $N(I)$ formé de N points situés dans $\mathbb{R}^p$ en supposant que $m_i = 1$, $X^i = (X_{ij}), j = 1, \ldots, p$. Le principe de l'ACP est de trouver un axe U_1, issu d'une combinaison linéaire des X_{ij}, tel que la variance du nuage autour de cet axe soit maximale. Pour bien comprendre, imaginons que la variance de U_1 soit égale à la variance du nuage; on aurait alors trouvé une combinaison des X_{ij} qui contient toute la diversité du nuage original (en tout cas toute la part de sa diversité captée par la variance).

On veut maximiser la variance expliquée par le vecteur U_1 ; pour les physiciens, cela a plutôt le sens de maximiser l'inertie expliquée par U_1.

Finalement, nous cherchons le vecteur U_1 tel que la projection du nuage sur U_1 ait une variance maximale.

La projection de l'échantillon des $X_1, \ldots, X_p$ sur U_1 s'écrit:

$$\pi_{U_1}(X) = X . U_1$$

La variance empirique de $\pi_{U_1}(X)$ vaut donc:

$\Pi_{U_1}(X)' \frac{1}{N} \Pi_{U_1}(X) = U_1' C U_1$ où $C = X' \frac{1}{N} X$ et correspond à la matrice

de covariance.

Comme nous avons vu plus haut que *C* est diagonalisable, notons *P* le changement de base associé et Δ la matrice diagonale formée de son spectre:

$\Pi_{U_1}(X)' \frac{1}{N} \Pi_{U_1}(X) = U_1' P' \Delta P U_1 = V' \Delta V$ où $V = (P U_1)$

Après cette réécriture, nous cherchons le vecteur unitaire V qui maximise $V' \Delta V$ où $\Delta = Diag(\lambda_1, \ldots, \lambda_N)$ est diagonale (rangeons les valeurs de la diagonale de Δ en ordre décroissant).

On peut rapidement vérifier qu'il suffit de prendre le premier vecteur unitaire; on a alors: $V' \Delta V = \lambda_1$

Plus formellement, on démontre ce résultat en maximisant la variance empirique des données projetées sur U_1 sous la contrainte que U_1 soit de norme 1.

On obtient ainsi les deux résultats suivants:

1. U_1 est vecteur propre de *C* associée à la valeur propre λ_1
2. U_1 est de norme 1

La valeur propre λ_1 est la variance empirique sur le premier axe de l'ACP. On continue la recherche du deuxième axe de projection U_2 sur le même principe en imposant qu'il soit orthogonal à U_1 .

La diagonalisation de la matrice de corrélation (ou de covariance si on se place dans un modèle non réduit), nous a permis d'écrire que le vecteur qui explique le plus d'inertie du nuage est le premier vecteur propre. De même le deuxième vecteur qui explique la plus grande part de l'inertie restante est le deuxième vecteur propre, etc. Si on considère les deux

premières valeurs propres λ_1 et λ_2 associés respectivement aux vecteurs propres U_1 et U_2 alors on visualise le nuage des points par projection sur le plan (U_1 , U_2).

Nous avons vu e n outre que la variance expliquée par le p-ième vecteur propre vaut λ_p Finalement, la question de l'ACP se ramène à un pr oblème de diagonalisation de la matrice de corrélation.

Proposition 2

L'inertie totale du nuage des individus $I_n(N) = Np$.

Preuve

$$I_n(N) = \sum_{i=1}^{N} d^2(i) = \sum_{i=1}^{N}\sum_{j=1}^{p} X_{ij}^2 = \sum_{i=1}^{N}\sum_{j=1}^{p} \frac{(X_{ij} - \bar{X}_j)^2}{\sigma_j^2} = \frac{1}{\sigma_j^2}\sum_{j=1}^{p}\sum_{i=1}^{N} (X_{ij} - \bar{X}_{ij})^2 = \frac{1}{\sigma_j^2}\sum_{j=1}^{p} N * \sigma_j^2$$

$$I_n(N) = \sum_{j=1}^{p} N = N * p$$

1.3.1.2 Analyse factorielle du nuage des variables

Considérons maintenant le nuage $N(J)$ formé de p points situés dans $\mathbb{R}^N$, $Y^j = (y_{ij}) = (x_{ij}); i = 1,\ldots,N; j = 1,\ldots,p$.

Donc $N(J)$ est le nuage dual de $N(I)$.

L'analyse de $N(J)$ est l'analyse factorielle dual de $N(I)$.

La matrice d'inertie de $N(J)$ est:

$\Gamma = XX'$ d'ordre $(N * N)$ où $\Gamma = (y_{ij}, y_{ij}) = \sum_j x_{ij} x_{i'j}$

Théorème

Si W_α est vecteur propre de Γ associé à la valeur propre $\lambda_\alpha \neq 0$, alors: $U_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X'W_\alpha$ est vecteur propre de V associé à la valeur propre $\lambda_\alpha \neq 0$.

Réciproquement si U_α est vecteur propre de V associé à la valeur propre $\lambda_\alpha \neq 0$, alors:

$W_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} XU_\alpha$ est vecteur propre de Γ associé à la valeur propre $\lambda_\alpha \neq 0$.

Preuve

Puisque W_α est vecteur propre normé de V associé à $\lambda_\alpha \neq 0$ donc $\Gamma W_\alpha = \lambda_\alpha W_\alpha$ ce qui équivaut à $XX'W_\alpha = \lambda_\alpha W_\alpha \Rightarrow X'XX'W_\alpha = \lambda_\alpha X'W_\alpha \Rightarrow VX'W_\alpha = \lambda_\alpha X'W_\alpha$.

Si $X'W_\alpha = 0 \Rightarrow XX'W_\alpha = 0 \Rightarrow \Gamma W_\alpha = 0$, ce qui contredit l'hypothèse $\lambda_\alpha \neq 0$ donc $X'W_\alpha \neq 0$.

D'où $X'W_\alpha$ est vecteur propre de V associé à $\lambda_\alpha \neq 0$.

Mais il n'est pas normé car $\|X'W_\alpha\|^2 = \|XW_\alpha\| = \lambda_\alpha$.

Par conséquent,le vecteur propre normé de V associé à $\lambda_\alpha \neq 0$ est:

$U_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X'W_\alpha$.

De même on montre aisément que :

$W_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} XU_\alpha$ est vecteur propre de Γ associé à la valeur propre $\lambda_\alpha \neq 0$.

1.4 Outils d'aide à l'interprétation [18]

Pour classer les points du nuage $[X^i, m_i]$ des individus selon leur rôle plus ou moins grand qu'ils ont joué dans la détermination de U_α , nous allons mesurer la contribution d'un individu i relativement à l'inertie expliquée par U_α qui s'exprime comme suit:

$$CRT_\alpha(i) = m_i \frac{F_\alpha^2(i)}{\lambda_\alpha} \quad \text{où } F_\alpha(i) \text{ est la coordonnée de } X^i \text{ sur } U_\alpha .$$

La qualité de représentation d'un individu par le sous-espace de dimension p, $U_1, U_2, \ldots, U_p$ est définie par:

$$Q_\alpha = \sum_{\alpha=1}^{p} c_\alpha(i) \quad \text{où } c_\alpha(i) = \frac{F_\alpha^2(i)}{\|X^i - P\|^2} = \cos^2(\theta)$$

1.5 Conclusion

L'Analyse en Composantes Principales est usuellement utilisée comme outil de compression linéaire. Le principe est alors de ne retenir que les n premiers vecteurs propres issus de le

diagonalisation de la matrice de corrélation (ou covariance), lorsque l'inertie du nuage projeté sur ces n vecteurs représente q_n pourcents de l'inertie du nuage original, on dit qu'on a un taux de compression de $1-q_n$ pourcents, ou que l'on a compressé à q_n pourcents.

Il est possible d'utiliser le résultat d'une ACP pour construire une classification statistique des variables aléatoires $X_1,\ldots,X_N$, en utilisant la distance suivante:

$d(X_n,X_{n'})=\sqrt{2(1-\rho_{X_nX_{n'}})}$ où $\rho_{X_nX_{n'}}$ est la corrélation entre X_n et $X_{n'}$.

Chapitre 2:

STRUCTURE DE DONNEES SYMBOLIQUES

2.1 Introduction [16]

L'Analyse de Données Symboliques (ADS) est une méthode mathématique d'analyse de données qui permet d'étendre les méthodes d'analyse de donnés classique à des données plus complexes, dites symboliques. Elle est basée sur une modélisation du monde réel supposé constituer d'individus et de concepts. Elle consiste à analyser un ensemble d'individus tout en prenant en compte la statistique propre, les données répétées, la variation interne de chacun d'entre eux, considéré d'abord comme un cas unique.
En principe, il faut utiliser en entrée d'une ADS la définition de données symboliques qui prend en compte la variation interne aux individus et leur complexité.

2.2 Définition de données symboliques et variables symboliques

Définition 1

Les individus sont les unités statistiques de la population à étudier $\square$.

Par exemple si nous étudions une population d'oiseaux alors chacun d'entre eux représente un individu.

Les concepts sont les variables et leurs modélisations mathématiques sont données par les objets symboliques dans un espace L, dit des «objets symboliques».

Exemple de concepts: taille, pays, poids,…

Définition 2

Soit $\square$ l'ensemble des individus et $\{ \omega_i \in \square, i = 1, \ldots, n \ ; \ Y_j, j = 1, \ldots, p \}$ le tableau de données symboliques où ω_i et Y_j sont respectivement l'individu et la variable symbolique ou le concept.

Les données sont symboliques quand dans chaque case du tableau de données symboliques, on ne trouve pas nécessairement une seule valeur qu'elle soit quantitative ou qualitative.

Définition 3

Les variables Y_j d'un tableau de données sont dites symboliques quand les données le sont. Autrement dit, chaque cellule (case) peut contenir des données de types différents: une valeur quantitative unique, une valeur qualitative unique, plusieurs valeurs, un intervalle, plusieurs valeurs avec une pondération (histogramme, fonction d'appartenance ou régression d'une variable...)

La variable symbolique est une fonction $Y_j : \square \square \Gamma_j$ où Γ_j est l'ensemble des observations.

EXEMPLE DE TABLEAU DE DONNEES SYMBOLIQUES

Variables / **Individus**	**POIDS**	**VILLE**	**COULEUR**
Produit 1	3,7	Paris	{rouge; blanc; jaune}
Produit 2	{3 ; 8}	Lyon	

Produit 3	{3,2 ; 4,8 ; 7,4}	{Paris; Lyon}	{0,3 rouge ;0,7 jaune}
Produit 4	{(0,4) [2,3[; (0,6) [3,8]}		

On a: $\Omega = \{\omega_1, \omega_2, \omega_3\} = \{Produit\ 1, Produit\ 2, Produit\ 3\}$ = l'ensemble des individus et $Y = \{Y_1, Y_2, Y_3\} = \{POIDS, VILLE, COULEUR\}$ =l'ensemble des variables symboliques ou concepts.

De plus les données symboliques sont représentées par les données du tableau.

Considérons la variable symbolique Y_1 , elle est définie comme suit:

$$Y_1 : \Omega \to \Gamma_1$$
$$\omega \to Y_1(\omega)$$

Où Γ_1 = l'ensemble des valeurs de le colonne de Y_1 .

2.3 Description des individus

Les individus sont modélisés dans un espace de description D qui exprime leurs propriétés à l'aide des variables.

Une description $d \in D$ est constituée d'un ou plusieurs produits exprimant ces propriétés par leur domaine de variation pour l'individu considéré.

Dans un tableau de données symboliques, les descriptions des individus sont représentées par les lignes.

De plus elles sont données par les connecteurs logiques: disjonction ou conjonction.

EXEMPLE DE DESCRIPTION D'UN INDIVIDU

Considérons le tableau de données précédent, on peut alors décrire le produit 3 comme suit: son poids varie entre les valeurs 3,2 ; 4,8 ; 7,4; sa ville est soit Paris ou Lyon et sa couleur varie entre 0.3 rouge et 0,7 jaune.

Autrement dit,

$d_3 = [\{Y_1 = 3.2\} \vee \{Y_1 = 4.8\} \vee \{Y_1 = 7.4\}] \wedge [\{Y_2 = Paris\} \vee \{Y_2 = Lyon\}] \wedge [\{Y_3 = 0.3\, rouge\} \vee \{Y_3 = 0.7\, jaune\}]$

Notons que:

$\vee$ Lie les valeurs qui se trouvent dans chaque case et

$\wedge$ Lie sur une même ligne les différentes colonnes.

2.4 Objet Symbolique

Définition 4

Un objet symbolique est défini par:

1.une description notée «d» (cellule du concept);

2.une relation binaire «R» sur D appelée opérateur de comparaison de deux descriptions

3.une fonction «a» permettant d'évaluer le résultat de comparaison (à l'aide de R) de la description d'un individu par rapport à la description donnée «d» du concept:

$$a : \omega \rightarrow [Y(\omega) R d_c] \in L$$

(«a» est aussi appelée fonction de reconnaissance).

C'est aussi une description d'un concept munie d'une façon de la comparer à la description d'un individu.

On le note:

$$S = (d, R, d \rightarrow a(\omega) \equiv [Y(\omega) R d_c])$$

Selon que L = {0,1} ou [0,1] nous distinguons deux types d'objets symboliques:

1. objets symboliques booléens: les valeurs de la fonction «a» sont dans {0,1}
2. objets symboliques modaux: les valeurs de la fonction «a» sont dans [0,1]

2.4.1 Objet symbolique Booléen

Un objet symbolique est dit Booléen si L={0 , 1}.

Si on a p-variables dans un tableau de données symboliques, nous notons:

$Y = [Y_1, \ldots, Y_p]$

$D = D_1 \times \ldots \times D_p$; $d = [d_1, \ldots, d_p] \in D$

$R = [R_1, \ldots, R_p]$ avec R_i la relation sur D_i

respectivement les variables, les descriptions liées aux classes d'individus considérés et les opérateurs de comparaison de deux descriptions

Alors on définit l'objet symbolique S = (a, R, d) avec l'assertion:

$$a = \bigwedge_{i=1}^{p} [Y_i R_i d_i]$$

EXEMPLE: OBJETS SYMBOLIQUES BOOLEENS (1)

Pays	Age	Taille	Continent
ω_1	37	1.75	Afrique
ω_2	45	1.82	Afrique
ω_3	42	1.80	Europe

Soit une classe d'individus $C = [\omega_1, \omega_2, \omega_3]$ dont chacun est décrit par les trois variables X_1, X_2, X_3 représentant respectivement âge, taille, continent.

L'assertion booléenne relative à C peut être:

$$a = [age \in [37, 45]] \quad [taille \in [1.75, 1.82]] \quad [\text{continent} \subseteq \{Afrique, Europe\}]$$

Nous posons $d_1 = [37, 45]$, $d_2 = [1.75, 1.82]$, $d_3 = \{Afrique, Europe\}$ et

$R_1 = R_2 = R_3 = \subseteq$.

.

Et l'objet symbolique associé à C est:

$$a(\omega) = \begin{cases} 1 = vrai & si \quad pour \quad tout \quad i \;,\; Y_i(\omega) \in d_i \\ 0 = faux & \sin on \end{cases}$$

EXEMPLE: OBJETS SYMBOLIQUES BOOLEENS (2)

$$a(\omega) = [couleur(\omega) \subseteq \{rouge, vert\}] \quad [taille(\omega) \in [20, 23]]$$

C'est un objet symbolique S= (a, R, d) trivial avec $a(\omega) = [y(\omega) R d]$

où: d = $\{rouge, vert\}$.

Donc $a(\omega)$ = vrai si la couleur de ω est rouge ou verte et la taille appartient à l'intervalle $[20, 23]$; $a(\omega)$ = faux sinon.

2.4.2 Extension d'objets symboliques

Un objet symbolique est la modélisation d'un concept et comme les concepts, admet une extension qui est donnée par l'ensemble des individus satisfaisant aux propriétés du concept suivant les valeurs de L et donc on obtient deux types d'extension:

1.Dans le cas booléen

$$Ext_{\square}(a) = \{w \in \square / a(w) = vrai\} \quad si \quad L = \{0,1\} = \{vrai, faux\}.$$

2.Dans le cas modal

$$Ext_{\square}(a) = \{w \in \square / a(w) \geq \alpha\} \quad si \quad L = [0,1] \;,\; \alpha \in R \;\; un \;\; seuil.$$

2.4.3 Union et intersection d'objets symboliques

Soient $S_1 = (a_1 R d_1)$ et $S_2 = (a_2 R d_2)$ deux objets symboliques. L'union symbolique (resp. l'intersection symbolique) de S_1 et S_2, notée $a_1 \cup a_2$ (resp. $a_1 \cap a_2$) se définit comme la conjonction de tous les objets symboliques dont l'extension sur $\square$ contient:

$$Ext(a_1) \cup Ext(a_2) \quad (resp: Ext(a_1) \cap Ext(a_2))$$

EXEMPLE: EXTENSION UNION ET INTERSECTION

Individus	Tailles	Âges
1	1,85	22
2	1,80	28
3	1,82	32
4	1,88	30

Proposition

Soient les objets symboliques a_1 et a_2 définis comme suit:

$a_1 = [taille = [1.80, 1.90]] \wedge [age = [20,30]]$

$a_2 = [taille = [1.70, 1.85] \wedge [age = 25,35]]$

$Ext(a_1) = \{1,2,4\}, Ext(a_2) = \{2,3\}$

Preuve

Rappelons que $Ext(a) = \{\omega \in \Omega / a(\omega) = vrai\}$

Or $taille(1) = 1.85 \in [1.80, 1.90]$ et $age(1) = 22 \in [20,30]$

Donc $a_1(\omega) = vrai$,par suite l'individu 1 appartient à $Ext(a_1)$.

Par analogie , on montre que les individus 2 et 4 sont les seuls qui répondent aux contraintes d'appartenance à $Ext(a_1)$.

Leur union et intersection:

$a_1 \cup a_2 = [taille = \{1.85, 1,80, 1.82, 1.88\}] \wedge [age = \{22, 28, 30, 32\}]$

$a_1 \cap a_2 = [taille = \{1,80\}] \wedge [age = \{28\}]$

2.5 Méthodes de statistiques descriptives unidimensionnelles appliquées à des données de type intervalle et valeurs multiples

Dans cette partie nous allons étudier comment la statistique descriptive se présente lorsque les valeurs observées sur les unités statistiques sont des données symboliques.
Dans cette étude , deux types de données seront considérés : les variables multi-valuées et les variables de type intervalle.
Ceci n'exclus pas le fait qu'il y ait d'autres types de variables symboliques (variables modales) sauf que nous nous intéressons le plus, dans ce mémoire ,aux données d'intervalles.
Avant de traiter le problème, quelques définitions et remarques utiles dans la suite seront introduites.

Définition 5

Un vecteur de description $d=[D_1, D_2, \ldots, D_p]$ est appelé un i ndividuel (ou vecteur de description individuelle) si pour tout j, D_j est un singleton : $D_j = \{x_j\}$ j=1,...,p.

Remarque

Si d est un vecteur de description individuelle alors il est de la forme :

$$d=[\{x_1\}, \ldots, \{x_p\}]$$

On identifiera $x=[x_1, \ldots, x_p] \in \prod_{j=1}^{p} O_j$ à $d=[\{x_1\}, \ldots, \{x_p\}]$.

Plus généralement on identifiera $d=[D_1, D_2, \ldots, D_p]$ à son ensemble de description cartésienne associée $D = D_1 \times \ldots \times D_p$.

Définition 6

Soit $d=[D_1, D_2, \ldots, D_p]$ un vecteur de description et $D = D_1 \times \ldots \times D_p$ la description cartésienne associée.

On note par V: l'ensemble des règles définies sur $\chi = \prod_{j=1}^{p} O_j$

$v(x) = 1$ si et seulement si x satisfait à la règle $v \in V$

$v(x) = 0$ sinon

On appelle quasi extension de d l'ensemble :

$$vir(d) = \{x \in D / v(x) = 1 \ pourtout v \in V\}$$

Exemple

Considérons le tableau de données symboliques où on obs erve sur $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_3, \omega_5\}$ deux variables : Y_1 et Y_2 .

où

	Y_1	Y_2
ω_1	$\{a,b\}$	$\{3\}$
ω_2	$\{b,c\}$	$\{2,3\}$
ω_3	$\{b\}$	$\{1,2\}$
ω_4	$\{c\}$	$\{1,3\}$
ω_5	$\{a\}$	$\{1,3\}$

$Y_1 : \square\square\square_1$ et $Y_2 : \square\square\square_2$

Soit la règle $v : Y_1 \in \{b,c\} \Rightarrow Y_2 \in \{2\}$.

Après calcul, on prouve les résultats suivants :

$$vir(d)_1 = \{x \in \{a,b\} x \{3\} / v(x) = 1\}$$

$$= \{(a,3)\}$$

De même on montre que :

$$vir(d)_2 = \{(b,2), (c,2)\}$$

$$vird(d)_3 = \{(b,2)\}$$

$$vir(d)_4 = \emptyset$$

$$vir(d)_5 = \{(a,1), (a,3)\}$$

2.5.1 Variables multi-valuées

Supposons que les p-variables symboliques soient à valeur multiples . Notons par Z une de ces variables Y_j $[Z = Y_j]$.

$Z:$ $\square$ $\square$ $\mathbb{Z}$

où $\mathbb{Z} = \Gamma_j$: l'ensemble des observations de Z dont ses éléments seront notés par  .

Définition 7

La fréquence empirique observée de la variable multi-valuée Z au point ε est la fonction

$O_Z : \mathbb{Z} \square \mathbb{R}$

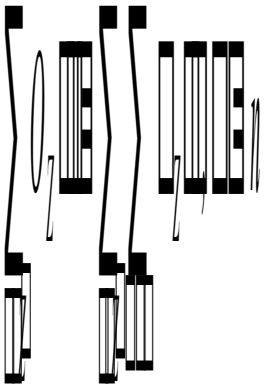
$\varepsilon \square \sum_{\omega \varepsilon \square} \pi_Z [\varepsilon, \omega]$

où

$\pi_Z [\varepsilon, \omega] = \frac{[x_\omega vir\ d_\omega / x_Z = \varepsilon\}[}{vir\ d_\omega [\square]}$ avec $vir\ d_\omega [\neq 0]$ et x_Z étant la valeur prise par Z dans le vecteur x .

Remarque

1.S'il existe l individus ω de $\square$ tels que $vir\ d_\omega [= 0]$ (c'est à dire $vir\ d_\omega = \varnothing$) alors ils seront omis de la sommation et par conséquent cette dernière se fera sur l'ensemble $\square' = \square / \{\omega_j, j = 1, \ldots, l\}$.

2. 

Définition 8

Soit Z une variable multi-valuée numérique , alors la fonction de distribution empirique de Z est la fonction

$$F_Z(\varepsilon) = \frac{1}{n} \sum_{\varepsilon_j \leq \varepsilon} O_Z(\varepsilon_j)$$

Remarque

Pour tout $\varepsilon \in \mathbb{R}$, $F_Z(\varepsilon) = \frac{1}{n} \sum_{\varepsilon_j \leq \varepsilon} O_Z(\varepsilon_j)$

$$= \frac{1}{n} \sum_{\varepsilon_j \leq \varepsilon} \sum_{\omega \in \square} \frac{\{x \in vir(d_\omega) / x_Z = \varepsilon_j\}}{vir(d_\omega)}$$

$$= \frac{1}{n} \sum_{\omega \in \square} \sum_{\varepsilon_j \leq \varepsilon} \frac{\{x \in vir(d_\omega) / x_Z = \varepsilon_j\}}{vir(d_\omega)}$$

$$F_Z(\varepsilon) = \frac{1}{n} \sum_{\omega \in \square} \frac{\{x \in vir(d_\omega) / x_Z \leq \varepsilon\}}{vir(d_\omega)}$$

Ainsi en considérant la distribution suivante:

$$\tau : \varepsilon \to \tau(\varepsilon, \omega)$$

pour tout ω de $\square$ avec $\tau(\varepsilon, \omega) = \frac{\{x \in vir(d_\omega) / x_Z \leq \varepsilon\}}{vir(d_\omega)}$

alors F_Z peut s'écrire sous cette forme $F_Z(\varepsilon) = \frac{1}{n}\sum_{\omega\varepsilon\square} \tau(\varepsilon,\omega)$

Définition 9

Soit Z une variable multi-valuée numérique, alors sa moyenne symbolique et sa variance symbolique sont respectivement définies par :

$$\overline{Z} = \frac{1}{n}\sum_{\varepsilon_j} O_Z(\varepsilon_j)$$

$$S_Z^2 = \frac{1}{n}\sum_{\varepsilon_j} (\varepsilon_j - \overline{Z})^2 O_Z(\varepsilon_j)$$

où O_Z étant la fréquence empirique observée.

2.5.2 Variables de type intervalle

Dans cette partie les p-variables symboliques sont de type intervalle , c'est à dire que si nous posons $Y_j = Z$ alors Z sera définie comme suit:

$$Z : \square \to \square$$
$$\omega \to Z(\omega) = [\underline{Z}_\omega ; \overline{Z}_\omega]$$

Afin d'étudier la variable Z **,** Bertrand et Goupil ont émis des hypothèses :

1. On suppose que la variable soit indépendante des autres variables.
2. Hypothèses d'équidistributions

   (a) Tous les individus sont pris avec la probabilité $\frac{1}{n}$.

   (b) Les valeurs de x_Z pour $x \varepsilon vir(d_\omega)$ sont uniformément distribuées dans l'intervalle $Z(\omega)$.

Si l'hypothèse d'équidistribution est satisfaite alors la densité de Z suivant l'individu ω pour $x \varepsilon vir(d_\omega)$ est :

$$Pr[x_\omega \leq \varepsilon / x \varepsilon vir\ d_\omega] = \int_{-\infty}^{\varepsilon} 1_{Z_\omega}(x)\,dx$$

$$= \begin{cases} 0 & si \quad \varepsilon < \underline{Z_\omega} \\ \dfrac{\varepsilon - \underline{Z_\omega}}{\overline{Z_\omega} - \underline{Z_\omega}} & si \quad \underline{Z_\omega} \leq \varepsilon \leq \overline{Z_\omega} \\ 1 & si \quad \varepsilon > \overline{Z_\omega} \end{cases}$$

Ainsi, on a:

Définition 10

Soit Z une variable de intervalle .La fonction de distribution empirique F_Z de Z est la moyenne des n distributions uniformes .

$$F_Z(\varepsilon) = \frac{1}{n}\sum_{\omega \varepsilon \Omega} Pr[x_\omega \leq \varepsilon / x \varepsilon vir\ d_\omega]$$

$$= \frac{1}{n}\sum_{\omega \varepsilon \Omega} \frac{\varepsilon - \underline{Z_\omega}}{\overline{Z_\omega} - \underline{Z_\omega}}\, 1_{Z_\omega}(\varepsilon) = \frac{1}{n}\sum_{\omega \varepsilon \Omega} 1_{\{\varepsilon \geq \overline{Z_\omega}\}}(\omega)$$

$$= \frac{1}{n}\sum_{\varepsilon \varepsilon Z(\Omega)} \frac{\varepsilon - \underline{Z_\omega}}{\overline{Z_\omega} - \underline{Z_\omega}} = \frac{1}{n}\#\{\omega \varepsilon \Omega / \varepsilon \geq \overline{Z_\omega}\}$$

En dérivant $F_Z(\varepsilon)$ par rapport à ε , on obtient la fonction $f_Z(\varepsilon)$ qui est la densité empirique de Z .

On a:

$$f_Z(\varepsilon) = \frac{1}{n} \sum_{\omega : \varepsilon \in Z(\omega)} \frac{1}{|Z(\omega)|} \text{ pour } \varepsilon \in \mathbb{R}$$

$$= \frac{1}{n} \sum_{\omega \in \Omega} \frac{1_{Z(\omega)}(\varepsilon)}{|Z(\omega)|} \text{ pour } \varepsilon \in \mathbb{R}$$

où $|Z(\omega)| = Z_\omega - Z_\omega$

L'expression de la densité permet de faire l'extension de la moyenne et de la variance sur les données de type intervalle.

Définition 11

On définit la moyenne empirique et la variance empirique d'une variable Z de densité f_Z les réels respectifs :

$$\bar{Z} = \int_{-\infty}^{+\infty} x f_Z(x) dx \text{ et } S_Z^2 = \int_{-\infty}^{+\infty} (x - \bar{Z})^2 f_Z(x) dx$$

Proposition

La moyenne empirique et la variance empirique, de la variable Z de type intervalle définie dans l'ensemble des individus Ω peuvent s'écrire :

$$\bar{Z} = \frac{1}{n} \sum_{\omega \in \Omega} \frac{Z_\omega + Z_\omega}{2} \text{ et } S_Z^2 = \frac{1}{3n} \sum_{\omega \in \Omega} (Z_\omega)^2 + Z_\omega Z_\omega + Z_\omega^2 - \frac{1}{4n^2} \left[\sum_{\omega \in \Omega} (Z_\omega + Z_\omega) \right]^2$$

Preuve

$$\bar{Z} = \int_{-\infty}^{+\infty} x f_Z(x)\,dx$$

$$= \int_{-\infty}^{+\infty} x \frac{1}{n} \sum_{\omega \in \Omega} \cdot \frac{1_{Z(\omega)}(x)}{[Z(\omega)]}\,dx$$

$$= \frac{1}{n} \sum_{\omega \in \Omega} \int_{Z_\omega}^{\bar{Z}_\omega} \cdot \frac{x}{[Z(\omega)]}\,dx$$

$$= \frac{1}{n} \sum_{\omega \in \Omega} \cdot \frac{Z_\omega + \bar{Z}_\omega}{2}$$

$$S_Z^2 = \int_{-\infty}^{+\infty} (x - \bar{Z})^2 f_Z(x)\,dx$$

$$= \int_{-\infty}^{+\infty} (x - \bar{Z})^2 \frac{1}{n} \sum_{\omega \in \Omega} \cdot \frac{1_{Z(\omega)}(x)}{[Z(\omega)]}\,dx$$

$$= \frac{1}{n} \sum_{\omega \in \Omega} \int_{Z_\omega}^{\bar{Z}_\omega} x^2 \frac{1}{[Z(\omega)]}\,dx - \bar{Z}^2$$

$$= \frac{1}{3n} \sum_{\omega \in \Omega} [Z_\omega^2 + Z_\omega \bar{Z}_\omega + \bar{Z}_\omega^2] - \frac{1}{4n^2} \left[\sum_{\omega \varepsilon \Omega} [Z_\omega + \bar{Z}_\omega] \right]^2$$

2.6 Conclusion

L'analyse de données symboliques a pour but en plus de l'extension des méthodes classiques données symboliques (classification automatique, analyse factorielle, discrimination, arbres de décisions, régression, etc.) d'obtenir une description symbolique d'une classe de façon à obtenir des sous classes homogènes et bien discriminantes des autres classique. Par rapport aux approches classiques, l'ADS s'applique à des données plus

complexes, utilise des outils adaptés à la manipulation d'objets symboliques de génération et spécialisation, de calcul d'extension et de mesures de ressemblances. Elle fournit aussi des représentations graphiques exprimant la variation interne des descriptions symboliques.

Chapitre 3:
ANALYSE EN COMPOSANTES PRINCIPALES DES SOMMETS

3.1 Introduction [2]

L'extension des ACP aux données d'intervalles a été proposée par Carlo N Lauro et francesco Palumbo sur «ACP des Sommets» en 2000.

En fait l'Analyse en Composantes Principales des Sommets consiste à ex écuter l'ACP classique sur la matrice normalisée Z.

De cette façon, les sommets seront des éléments du sous espace R^p où les p-descripteurs quantitatifs sont des éléments de R^N .

«ACP.S» recherche un sous-espace approprié pour représenter les objets symboliques et d'un point de vue dual, les p- variables.

Comme dans l'ACP Classique le sous-espace optimal est donné par les axes V_m avec $1 \leq m \leq p$, maximisant la somme des carrées des coordonnées des sommets projetés.

3.2 Construction de la matrice normalisée [2]

Soit l'ensemble des ω_i $(1 \leq i \leq N)$ objets symboliques (décrits par p- variables ou descripteurs: $Y = \{Y_1, \ldots, Y_j, \ldots, Y_p\}$.

De nos jours l'analyse de données symboliques est basée sur des traitements numériques d'objets symboliques convenablement codés suivant une interprétation symbolique des résultats ou des méthodes qui traitent directement les descripteurs symboliques.

Dans cette dynamique, Lauro et Palembo ont pr ésenté le premier cadre d'approche permettant d'analyser les objets symboliques décrits seulement par des variables quantitatives d'intervalles.

Pour eux, la variable générique Y_j ne représente plus une variable évaluée simple (monovaluée) comme dans l'ACP classique mais se rapporte aux $\underline{Y_j}$ bornes inférieures et $\overline{Y_j}$ bornes supérieures de l'intervalle décrit par la j-ème variable donc la matrice symbolique X de données est d'ordre $N \times 2p$.

Par suite, on observe sur chaque individu ω_i (objet symbolique) la variable Y_j .

D'où le tableau suivant:

Descripteurs / Objets symboliques	Y_i	. . .	Y_j	. . .	Y_p
ω_1					

. . .					
ω_i			$[\underline{Y_{ij}}, \overline{Y_{ij}}]$		
. . .					
ω_N					

$\underline{Y_{ij}}$ = plus petite observation de la variable Y_j sur l'objet symbolique ω_i

$\overline{Y_{ij}}$ = plus grande observation de la variable Y_j sur l'objet symbolique ω_i

On a un nuage N points de R^p :

$$X = \begin{pmatrix} . \\ . \\ . \\ [\underline{Y_{ij}}; \overline{Y_{ij}}] \\ . \\ . \\ . \end{pmatrix} \quad i = 1,\ldots,p \quad ; \quad j = 1,\ldots,n$$

La description de l'objet symbolique ω_i est associée à la i-ème ligne de la matrice X^i de données d'intervalles:

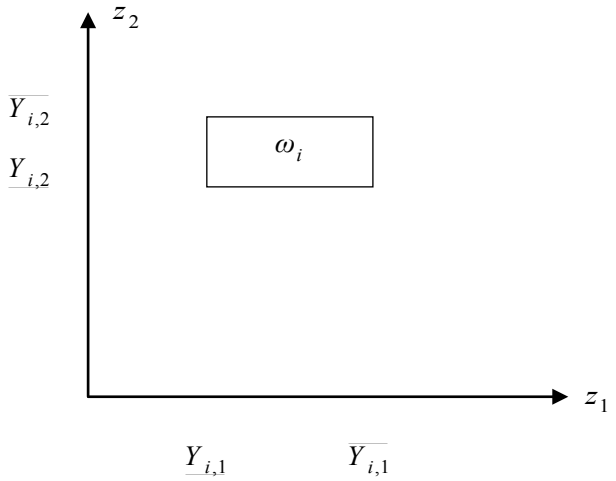
$$X^i = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \underline{Y_{i,1}} & \overline{Y_{i,1}} & \underline{Y_{i,2}} & \overline{Y_{i,2}} & \cdots & \underline{Y_{i,p}} & \overline{Y_{i,p}} \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \end{pmatrix}$$

Exemple 3.2.1

Dans le cas simple $p=2$, la description de l'os générique ω_i est associée à la i-ème ligne de la matrice X de données d'intervalles:

$$X \quad \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \underline{y_{i,1}} & \overline{y_{i,1}} & \underline{y_{i,2}} & \overline{y_{i,2}} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix} \longleftarrow \text{i-ème ligne}$$

Dans une vue géométrique ω_i est représenté par un rectangle ayant $2^p = 4$ sommets correspondant à toutes les combinaisons possibles (min, max).



REPRESENTATION DE L'OS DANS UN ESPACE A DEUX DIMENSIONS

Les coordonnées relatives aux sommets concernant les nouvelles variables z_1 et z_2 répondent aux critères suivant:

1.Avoir les mêmes domaines que Y_1 et Y_2 respectivement

2.Correspondent aux lignes de la matrice Z_i

$$Z_i = \begin{array}{c} \begin{matrix} z_1 & z_2 \end{matrix} \\ \begin{pmatrix} \underline{y_{i,1}} & \underline{y_{i,2}} \\ \underline{y_{i,1}} & \overline{y_{i,2}} \\ \overline{y_{i,1}} & \underline{y_{i,2}} \\ \overline{y_{i,1}} & \overline{y_{i,2}} \end{pmatrix} \end{array}$$

On remplace le nuage X par celui des Z_i et dans le cas général où on a toutes les p-variables, chaque matrice Z_i de codage aura 2^p lignes et p colonnes:

$$Z_i = \begin{array}{c} \begin{matrix} z_1 & z_2 & & & z_p \end{matrix} \\ \begin{pmatrix} \underline{Y_{i,1}} & \underline{Y_{i,2}} & . & . & . & \underline{Y_{i,p}} \\ \underline{Y_{i,1}} & \overline{Y_{i,2}} & . & . & . & \underline{Y_{i,p}} \\ \overline{Y_{i,1}} & \underline{Y_{i,2}} & . & . & . & \underline{Y_{i,p}} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ \overline{Y_{i,1}} & \overline{Y_{i,2}} & . & . & . & \overline{Y_{i,p}} \end{pmatrix} \end{array} \quad i = 1, \ldots, n$$

La matrice de codage Z est obtenue en superposant les N matrices Z_i de codage de l'objet symbolique ω_i , (avec $1 \leq i \leq N$).

Elle a par conséquent $L = Nx2^p$ lignes et p colonnes, et présente le codage numérique de N objets symboliques.

D'où Z est donnée par:

$$Z = \begin{array}{c} \begin{matrix} z_1 & z_2 & & z_p \end{matrix} \\ \begin{pmatrix} \underline{Y_{1,1}} & \overline{Y_{1,1}} & . & . & \underline{Y_{1,p}} \\ . & . & & & . \\ . & . & & & . \\ . & . & & & . \\ \overline{Y_{1,1}} & \underline{Y_{1,2}} & \ldots & & \underline{Y_{1,p}} \end{pmatrix} \leftarrow Z_1 \\ \begin{pmatrix} \underline{Y_{2,1}} & \overline{Y_{2,2}} & . & . & \underline{Y_{2,p}} \\ . & . & & & . \\ . & . & & & . \\ . & . & & & . \\ \overline{Y_{2,1}} & \underline{Y_{2,2}} & \ldots & & \underline{Y_{2,p}} \end{pmatrix} \leftarrow Z_2 \\ \vdots \\ \begin{pmatrix} \quad \\ \quad \\ \quad \end{pmatrix} \end{array}$$

$$\begin{matrix} \underline{Y_{N,1}} & \overline{Y_{N,2}} & \cdots & \overline{Y_{N,p}} \\ \vdots & \vdots & & \\ \vdots & \vdots & & \\ \vdots & \vdots & & \\ \overline{Y_{N,1}} & \underline{Y_{N,2}} & \cdots & \underline{Y_{N,p}} \end{matrix} \longleftarrow Z_N$$

Nous pouvons écrire plus simplement Z comme suit:

$$Z = \begin{matrix} z_1 & z_2 & & z_p \\ \underline{Y_{1,1}} & \overline{Y_{1,1}} & \cdots & \underline{Y_{1,p}} \\ \vdots & \vdots & & \\ \vdots & \vdots & & \\ \vdots & \vdots & & \\ \overline{Y_{1,1}} & \underline{Y_{1,2}} & \cdots & \underline{Y_{1,p}} \\ \\ \underline{Y_{2,1}} & \overline{Y_{2,2}} & \cdots & \underline{Y_{2,p}} \\ \vdots & \vdots & & \\ \vdots & \vdots & & \\ \vdots & \vdots & & \end{matrix}$$

$$
\begin{array}{l}
\overline{Y_{2,1}} \quad Y_{2,2} \ \ldots \ Y_{2,p} \\
\quad\ \ . \\
\quad\ \ . \\
\quad\ \ . \\
\underline{Y_{N,1}} \quad \overline{Y_{N,2}} \ \ldots \ \overline{Y_{N,p}} \\
\ .\quad\ .\quad\ \ . \\
\ .\quad\ .\quad\ \ . \\
\ .\quad\ .\quad\ \ . \\
\overline{Y_{N,1}} \quad \underline{Y_{N,2}} \ \ldots \ \underline{Y_{N,p}}
\end{array}
$$

Z = $[z_{qj}]$ avec $1 \leq q \leq L$ et $1 \leq j \leq p$ où $L = Nx2^p$.

Sans perte de généralité nous pouvons supposer que les variables z_j sont normalisées.

3.3 ACP de la matrice normalisée Z [2]

Dans cette partie nous allons rechercher un sous -espace approprié pour représenter le nuage $N(I)$ des L sommets de R^p des N objets symboliques $(N<L)$ et le nuage $N(J)$ des p-descripteurs de R^L des N objets symboliques.

En tenant compte des conclusions relatives à la démarche de l'ACP sur une matrice de données quantitatives nous allons directement procéder à la diagonalisation de la matrice de corrélation $\Gamma = \frac{1}{N} Z'Z$ pour trouver les axes principaux $V_m (1 \leq m \leq p)$ où sera projeté le nuage $N(I)$.

Par suite V_m en tant que axe principal vérifie les conditions suivantes :

$$V'_m V_{m'} = 0 \; sim \neq m' \quad \text{et} \quad V'_m V_{m'} = 1 \; sim = m'$$

L'équation caractéristique de l'ACP des sommets est donnée par :

$$\Gamma V_m = \lambda_m V_m \Leftrightarrow \frac{1}{N} Z'Z V_m = \lambda_m V_m \qquad 1 \leq m \leq p$$

où V_m est le vecteur propre de Γ associé à la valeur propre non nulle λ_m .

Comme $N(I)$ est le nuage dual de $N(J)$ donc l'analyse de $N(I)$ est l'analyse factorielle duale de $N(J)$.

Par conséquent pour déterminer les axes principaux $W_m (1 \leq m \leq p)$ où sera projeté cette fois-ci le nuage $N(I)$, nous allons procéder à la diagonalisation de

$$\Phi = \frac{1}{N} ZZ' \; .$$

L'équation de l'ACP des descripteurs est donnée par:

$$\Phi W_m = \lambda_m W_m \Leftrightarrow \frac{1}{N} ZZ' W_m = \lambda_m W_m$$

où W_m est la valeur propre de Φ associé à la valeur propre λ_m non nulle.

Proposition 3.3.1 [1]

Si W_m est le vecteur propre de Φ associé à la valeur propre $\lambda_m \neq 0$ alors V_m est vecteur propre de Γ associé à λ_m ; avec $V_m = \lambda_m^{\frac{-1}{2}} Z'W_m$.

Preuve

Montrons que $\Gamma V_m = \lambda_m V_m$ si $V_m = \lambda_m^{\frac{-1}{2}} Z'W_m$ sachant que Γ et Φ ont même valeur propre non nulle λ_m .

$$\Gamma V_m = \Gamma \left[\lambda_m^{\frac{-1}{2}} Z' W_m\right]$$

$$= \lambda_m^{\frac{-1}{2}} \Gamma Z' W_m$$

$$= \lambda_m^{\frac{-1}{2}} \frac{1}{N} Z' Z Z' \quad W_m$$

$$= \lambda_m^{\frac{-1}{2}} Z' \frac{1}{N} Z Z' \quad W_m$$

$$= \lambda_m^{\frac{-1}{2}} Z' \Phi \quad W_m$$

$$= \lambda_m^{\frac{-1}{2}} Z' \lambda_m W_m$$

$$= \lambda_m \lambda_m^{\frac{-1}{2}} Z' W_m$$

$$= \lambda_m V_m$$

Proposition 3.3.2 [18]

Si V_m est le vecteur propre de Γ associé à la valeur propre $\lambda_m \neq 0$, alors $W_m = \lambda_m^{\frac{-1}{2}} Z V_m$ est le vecteur propre de Φ associé à λ_m .

Preuve

Montrons que $\Phi W_m = \lambda_m W_m$ car il est clair que λ_m vecteur propre de Φ .

$$\Phi W_m = \frac{1}{N} Z Z' \lambda_m^{\frac{-1}{2}} Z V_m$$

$$= \lambda_m^{\frac{-1}{2}} \frac{1}{N} Z Z' Z V_m$$

$$= \lambda_m^{\frac{-1}{2}} Z \frac{1}{N} Z' Z V_m$$

$$= \lambda_m^{\frac{-1}{2}} Z \Gamma V_m$$

$$= \lambda_m^{\frac{-1}{2}} Z\lambda_m V_m$$

$$= \lambda_m \lambda_m^{\frac{-1}{2}} ZV_m$$

$$= \lambda_m W_m$$

Définition 3.3.1

On appelle composantes principales de l'os ω_i ,les coordonnées de ses sommets sur les axes principaux V_m .

Elles s'expriment comme suit: $\Psi_{i.m} = Z_i V_m$ avec $1 \leq i \leq n$ et $1 \leq m \leq p$

Définition 3.3.2

On appelle composantes principales du descripteur Y_j ,ses coordonnées sur les axes principaux W_m .

Elles sont définies par : $\rho_{j.m} = Y'_j W_m$ avec $1 \leq j \leq n$ et $1 \leq m \leq p$

Proposition 3.3.3

1.Sur l'axe principal V_m , la coordonnée de Z_i peut s'écrire en fonction des composantes de W_m : $\Psi_{i.m} = \lambda_m^{\frac{1}{2}} w_{i.m}$ où $W_m = [w_{i,m}]$.

2 .Sur l'axe principal W_m , la coordonnée du descripteur Y_j peut être définie en fonction de composantes de V_m par: $\rho_{j.m} = \lambda_m^{\frac{1}{2}} v_{j.m}$ où $V_m = [v_{j,m}]$.

Preuve

Par définition $\Psi_{i,m} = V_m' Z_i = Z_i' V_m$ car c'est la projection orthogonale de $Z_i sur V_m$,

Or

$$Z = \begin{matrix} Z_1' \\ Z_2' \\ . \\ . \\ . \\ Z_N' \end{matrix} \quad \text{Donc} \quad ZV_m = \begin{matrix} Z_1'V_m \\ Z_2'V_m \\ . \\ . \\ . \\ Z_N'V_m \end{matrix}$$

D'où $\Psi_{i,m}$ est la i-ème coordonnée de ZV_m dans R^n .

Par suite $\Psi_{i.m} = Z_i V_m$, comme $ZV_m = \lambda_m^{1/2} W_m$ d'après Proposition 2.2.2 .

Donc $\Psi_{i.m} = \lambda_m^{\frac{1}{2}} w_{i.m}$.

De même sur l'axe principal W_m , $\rho_{j.m} = Y'_j W_m$

$Z' = [Z_1 Z_2 \dots Z_n[$ donc $\rho_{j,m}$ est la j-ème coordonnée de $Z'V_m$ dans $\mathbb{R}^p$.

Or $Z'V_m = \lambda_m^{1/2} V_m$ d'après Proposition 2.2.1 par conséquent $\rho_{j.m}$ est la j-ème coordonnée de $\lambda_m^{1/2} V_m$.

D'où $\rho_{j,m} = \lambda_m^{\frac{1}{2}} v_{j.m}$

Les remarques 1 et 2 nous permettent de construire les projections des nuages N(I) et N (J) sur les axes principaux respectifs V_m et W_m en faisant une seule analyse factorielle.
Cependant nous allons mesurer quelques contributions dans la suite.

3.4 Outils d'aide à l'interprétation [2]

L'interprétation des axes principaux et celle de l'ACP-S sont faites en se référant aux variables z_j ayant des contributions maximales.
Dans ce cas-ci des variables normalisées, des contributions sont calculées comme des corrélations variable / facteur:

Proposition 3.4.1

La contribution $CTA_{j,m}$ classe les points Z_i selon leurs rôles plus ou moins grands qu'ils ont joué dans la détermination de W_m et mesure la contribution relative de Z_i à l'inertie par W_m .

On a: $$CTA_{j,m} = \frac{\lambda_m^{1/2} v_{j,m}}{\lambda_m}$$

Preuve

Par définition [2]: $$CTA_{j,m} = \frac{\square_{j,m}}{\lambda_m} = \frac{\lambda_m^{\frac{1}{2}} v_{j,m}}{\lambda_m}$$

De même on montre que $CTA_{i,m} = w_{i,m}^2$.

3.5 Notion de MCAR= Maximum Covering Area Rectangle

La représentation de ω_i sur l'axe générique m est donné par le segment contenant toutes les projections des sommets. Adoptant le même critère dans un espace bidimensionnel formé par les axes m et m' alors les projections extrêmes des sommets vont définir un rectangle appelé MCAR. Par conséquent si la représentation MCAR des objets symboliques est faite dans le plan alors on aurait des hypercubes associés à chaque objet symbolique mais souvent il se produit un surdimensionnement sur le vrai image de O.S des $\mathbb{R}^p$.
Afin de surmonter inconvénient, Lauro et Palembo ont proposé de réduire les représentations, les sommets ayant une très bonne qualité de représentation.

Proposition 3.5.1 [2]

La qualité de la représentation des sommets Z_i par le sous-espace $\langle V_1, \ldots, V_p \rangle$ est mesurée en termes de critère du carré du cosinus et s'exprime par:

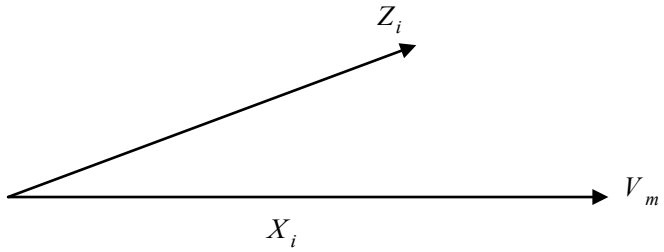
.

$$CRT_{q,m} = \frac{\sum_{j=1}^{p} (z_{q,j} v_{j,m})^2}{\sum_{j=1}^{p} z_{q,j}^2} \quad ; \quad 1 \leq m$$

où $z_{q,j}$ est le sommet de l'objet symbolique ω_i .

Preuve

$$C_{q,m} = \cos^2(\Theta) = \frac{\|\Psi_{q,m}\|^2}{\|z_q\|^2} = \frac{\|z_q v_m\|^2}{\|z_q\|^2} = \frac{\sum_{j=1}^{p} (z_{q,j} v_{j,m})^2}{\sum_{j=1}^{p} z_{q,j}^2}$$



.

$$\cos^2\Theta = \frac{\|\Psi_{i,m}\|^2}{\|Z_i - O\|^2}$$

3.6 Application de la méthodologie de l'ACP des sommets sur un tableau de données d'huiles

Dans cette partie nous traitons une illustration concrète de la méthodologie proposée sur un ensemble réel de données .

Nous prenons un ensemble de données d'huiles (ICHINO,1998) représenté dans le tableau ci-dessous ,en grande partie utilisé dans les applications d'Analyses de Données Symboliques où les caractéristiques sont bien connues par les chercheurs de ce domaine.

L'ensemble de données présente huit différentes classes d'huiles, $\omega_i avec 1 \leq i \leq 8$,décrites par quatre variables quantitatives d'intervalles:

Y_1 = «densité» ; Y_2 = «point de congélation» ; Y_3 = «valeur d'iode» et

Y_4 = «saponification»

Tableau : MATRICE DE DONNEES D'HUILES

Descripteurs / Individus	**Densité**	**Point de congélation**	**Valeur d'iode**	**Saponification**
Linseed	[0,93 ; 0,94]	[-27 ; -18]	[170 ; 204]	[118 ; 196]
Perilla	[0,93 ; 0,94]	[-5 ; -4]	[192 ; 2008]	[188 ; 197]
Cotton	[0,92 ; 0,92]	[-6 ; -4]	[99 ; 113]	[189 ; 198]
Sesame	[0,92 ; 0,93]	[-6 ; -4]	[104 ; 116]	[187 ; 193]
Camellia	[0,92 ; 0,92]	[-21 ; -15]	[80 ; 82]	[189 ; 193]
Olive	[0,91 ; 0,92]	[0 ; 6]	[79 ; 80]	[187 ; 196]
Beef	[0,86 ; 0,87]	[30 ; 38]	[40 ; 48]	[190 ; 199]
Hog	[0,86 ; 0,86]	[22 ; 32]	[53 ; 77]	[190 ; 202]

Nous allons maintenant appliquer à ce tableau de données,l'ACPS.

Notons que dans cet exemple $p=4$ et $N=8$,or à chaque os ω_i est associé une matrice Z_i de codage qui est ici d'ordre(16 x 4) .

Déterminons pour chaque ω_i , Z_i avec $1 \leq i \leq 8$.

Pour $i=1$,ona $\omega_1 = linseed$ et

$$Z_1 = \begin{matrix} 0.93 - 27\ 170\ 118 \\ 0.93 - 18\ 170\ 118 \\ 0.93 - 27\ 204\ 118 \\ 0.93 - 27\ 170\ 196 \\ 0.93 - 18\ 204\ 118 \\ 0.93 - 18\ 170\ 196 \\ 0.93 - 27\ 204\ 196 \\ 0.93 - 18\ 204\ 196 \\ 0.94 - 18\ 204\ 196 \\ 0.94 - 27\ 204\ 196 \\ 0.94 - 18\ 170\ 196 \\ 0.94 - 18\ 204\ 118 \\ 0.94 - 27\ 170\ 196 \\ 0.94 - 27\ 204\ 118 \\ 0.94 - 18\ 170\ 118 \\ 0.94 - 27\ 170\ 118 \end{matrix}$$

On fait le même travail pour les autres objets symbolique et en superposant les matrices Z_i on obtient Z comme suit : $Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \\ \\ Z_8 \\ \\ \\ \end{pmatrix}$

Dans la suite, nous nous limitons à la représentation graphique des résultats de la méthode proposée dans ce mémoire (ACPS).

Dans la figure suivante, nous montrons les résultats réalisés par l'ACPS en considérant les deux premiers axes (premier plan). Notons que 88,4 % de toute l'inertie est expliquée par les deux premiers axes.

Notons que 88.4% de toute l'inertie est expliquée par les deux premiers axes .

Dans la figure, la proximité entre les MCAR est principalement indiquée par les OS influencés par les mêmes descripteurs.

Nous ne pouvons donner aucune interprétation sur la similitude entre les MCAR relativement à la taille et la forme.

Comme points supplémentaires nous avons aussi représenté les variables, même si elles étaient représentées dans l'espace R^N.

Cependant, cette représentation simultanée est très utile pour l'interprétation, à condition que les variables soient seulement évaluées en tenant compte de leurs directions.

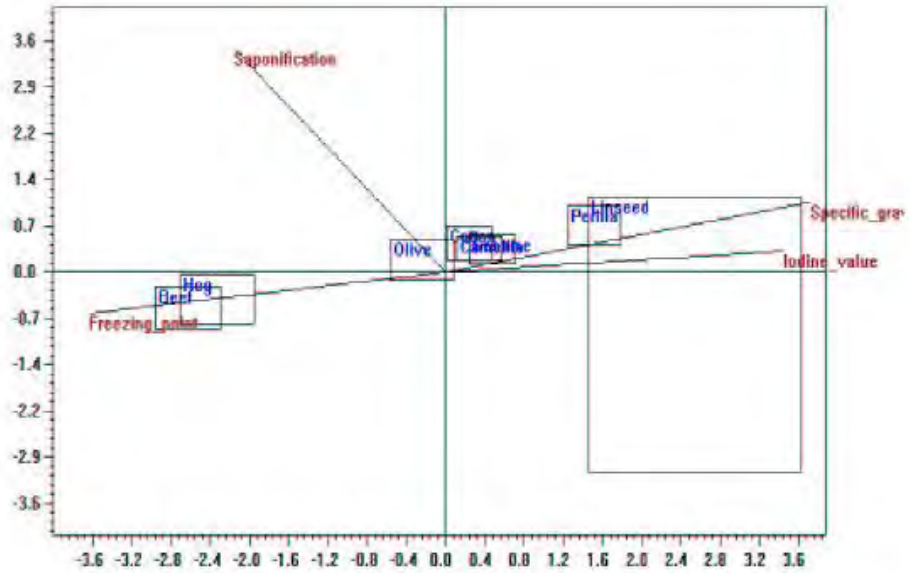


Figure 4: V-PCA: first two axes (88.4%)

Les objets symboliques associés aux représentations des MCAR sur le côté gauche du plan se rapportent aux huiles principalement caractérisées par les plus grandes valeurs de «saponification» et «poi nt de congélation», tandis que, du côté opposé nous avons les représentations d'huiles données par la «densité» et les plus grandes valeurs de la «valeur d'iode».

PERSPECTIVES ET CONCLUSION

Dans ce mémoire nous avons proposé quelques nouvelles techniques de l'ACP des données d'intervalles sous une approche d'analyse de données symboliques..

De ce p oint de vue, nous avons représenté les objets symboliques sur les axes factoriels mettant en évidence d'autres aspects structurels (taille et forme) et leurs positionnement dans le sous-espace factoriel.

La méthodologie proposée représente une première dans l'analyse de données symboliques et doit être généralisée sur n'importe quelle topologie de descripteur symbolique (mode et variables multinominales) et, en même temps, prendre en compte les règles et taxonomies qui peuvent être définies dans la structure de données symboliques (BOCK ET DIDAY, 2000).

L'autre aspect pertinent concerne la représentation graphique qui pourrait être faite avec des outils géométriques plus efficaces que le MCAR. Par exemple, nous indiquons les formes convexes et l'inertie maximale diagonale (VERDE ET DE ANGEUS, 1997)

Enfin, nous indiquons la représentation des axes qui, en ordre de garantir la cohérence nécessaire, devrait être faite en utilisant le langage et les outils de l'ADS.

Aussi dans cette direction il y a quelques contributions (GETTLER- SUMMA, 1997) dans la documentation qui pourraient être étendues sur l'analyse factorielle.

BIBLIOGRAPHIE

[1] .Bock, H. H. and Diday, E. (eds): 2000, Analysis of Symbolic Data, Springer. (in press).

[2] Carlo.N.Lauro and Francesco Palumbo:2000,Principal Component Analysis of Interval Data : a symbolic analysis approach

[3] Cazes, P., Chouakria, A., Diday, E. and Schektman, Y :1997, Extension de l'analyse en composantes principales des données de type intervalle, Revue de Statistique Appliquée XIV(3), 5–24.

[4] Chouakria, A., Diday, E. and Cazes, P.: 1998, An improved factorial representation of symbolic objects, KESDA'98 27-28 April, Luxembourg.

[5] D'Ambra, L. and Lauro, C. N.: 1982, Analisi in componenti principali in rapporto a un sottospazio di riferimento, Rivista di Statistica Applicata 15(1), 51–67.

[6] De Carvalho, F. A. T.: 1992, Méthodes Descriptives en Analyse de Données Symboliques, Thèse de doctorat., Université Paris Dauphine, Paris.

[7] De Carvalho, F. A. T.: 1997, Clustering of constrained symbolic objects based on dissimilarity functions, Indo–French Workshop on Symbolic Data Analysis and its Applications, University of Paris IX.

[8] Diday, E.: 1987, Introduction à l'approche symbolique en analyse des données, Journées Symbolique-Numérique, Université Paris Dauphine.

[9] Diday, E.: 1996, Une introduction à l'analyse des données symboliques, SFC, Vannes, France.

[10] Escofier, B. and Pagés, J.: 1988, Analyse factorielles multiples, Dunod, Paris.

[11] Gettler–Summa, M.: 1997, Symbolic marking application on car accident reports, Applied Stocastic Models and Data Analysis, Vol. Invited and Specialised Session Papers, ASMDA, Anacapri, pp. 299–306.

[12] Ichino, M.: 1988, General metrics for mixed features - the cartesian space theory for pattern recognition, International Conference on Systems, Man and Cybernetics.

[13] Jean Paul Benzécri: Analyse des données. T2 (leçons sur l'analyse factorielle et la reconnaissance des formes et travaux du Laboratoire de statistique de l'Université de Paris 6. T. 2: l'analyse des correspondances), Dunod Paris Bruxelles Montréal, 1973.

[14] Lauro, C. and Palumbo, F.: 1998, New approaches to principal components analysis to interval data, International Seminar on New Techniques & Technologies for Statistics, NTTS'98, 4/6 nov. 1998, Sorrento, Italy.

[15] Lebart, L., Morineau, A. and Piron, M.: 1995, Statistique exploratorie multidimensionelle, Dunod, Paris.

[16] Monzer Boubou : Analyse de données symboliques

[17] Meccariello, G.: 1999, Analisi in componenti principali per dati ad intervallo, Tesi di Laurea in Statistica Universit´ di Napoli ”Federico II”.

[18] Thiam Sada Sory : 2005, Cours d'AEA sur l'analyse de données, Université Cheikh Anta Diop de Dakar.

[19] V erde, R. and De Angelis, P.: 1997, Symbolic objects recognition on a factorial plan, NGUS’97, Bilbao Spain.