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:

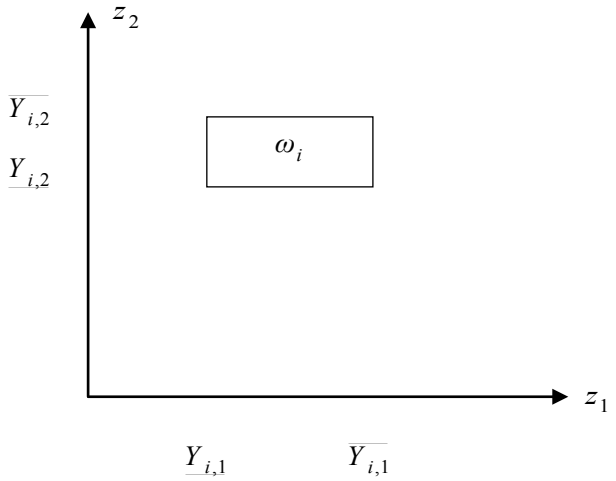
$$X^i = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \underline{Y_{i,1}} & \overline{Y_{i,1}} & \underline{Y_{i,2}} & \overline{Y_{i,2}} & \cdots & \underline{Y_{i,p}} & \overline{Y_{i,p}}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

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 :

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

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:

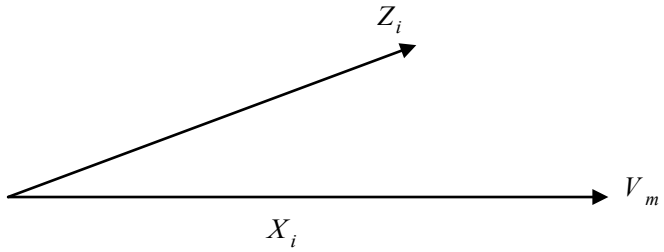
.

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

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.