L'ACP DE DONNEES DE TYPE INTERVALLE

3.1 Introduction

L'ACP classique traite des tableaux de données de la forme $I \times J$ où I représente l'ensemble des objets et J celui des variables. La case du tableau, croisement de la ième ligne et de la jème colonne, contient la valeur observée x_{ij} supposée unique, de la jème variable quantitative pour le ième objet.

Dans ce chapitre, nous étendons l'ACP à des tableaux des données où x_{ij} est un intervalle de valeurs introduisant la variation ou l'imprécision :$x_{ij} = [\underline{x}_{ij}, \bar{x}_{ij}]$ où $\underline{x}_{ij}, \bar{x}_{ij}$ sont respectivement, la plus petite et la pluls grande valeur observée, de la jème variable pour le ième objet.

3.2 Données du Problème et Objectif :

Soient $S_1, S_2, ..., S_m$ m objets décrits par n variables $X_1, ..., X_n$ de type intervalle

$$\begin{pmatrix} S_1 \\ \vdots \\ S_m \end{pmatrix} = \begin{pmatrix} x_{S_{11}} & \dots & x_{S_{1n}} \\ \vdots & & \vdots \\ x_{S_{m1}} & \dots & x_{S_{mn}} \end{pmatrix} = \begin{pmatrix} [\underline{x}_{11}, \bar{x}_{11}] & \dots & [\underline{x}_{1n}, \bar{x}_{1n}] \\ \vdots & & \vdots \\ \underline{x}_{m1}, \bar{x}_{m1} & \dots & \underline{x}_{mn}, \bar{x}_{mn} \end{pmatrix} \quad \textbf{(1)}$$

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

A chaque objet S_i on associe un poids $p_i > 0$, avec $\sum_{i=1}^{n} p_i = 1$.

Classiquement, étant donné un ensemble d'objets décrits chacun par un vecteur $(x_{i1}, ..., x_{in})$, l'objectif de toute méthode de réduction de dimension en particulier, l'ACP es de réduire le nombre de variables descriptives, tout en préservant la "structure de distribution" des objets [chapitre 1].

Soient $Y_1, ..., Y_p \quad (p < n)$ les nouvelles variables descriptives obtenues après réduction : chaque objet S_i sera décrit par un vecteur $(y_{i1}, ..., y_{ip})$ dans un espace de dimension plus faible.

De façon similaire, partant d'un ensemble d'objets S_i caractérisés chacun par un n-uple $([\underline{x}_{i1}, \bar{x}_{i1}], ..., [\underline{x}_{in}, \bar{x}_{in}])$, 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 sagit en fait de décrire la structure de distribution des S_i dans un espace de dimension faible défini par des variables de type intervalle $Y_1, ..., Y_p \quad (p < n)$; chaque objet S_i sera alors décrit par un p-uple $([\underline{y}_{i1}, \bar{y}_{i1}], ..., [\underline{y}_{ip}, \bar{y}_{ip}])$.

On présente ici deux méthodes :

- La méthode des sommets
- La méthode des centres

3.3. Méthodes des Sommets

3.3 a) Introduction

Soit un objet S décrit par le n-uple $([\underline{x}_1, \bar{x}_1], ..., [\underline{x}_n, \bar{x}_n])$, cet objet peut-être visualisé dans l'espace de description, par un hypercube a 2^n sommets. La longueur des côtés de l'hypercube est donnée par l'étendue des intervalles associés à chaque variable de description.

Exemple 1

Pour $n = 2$ l'objet S décrit par

$$S = ([\underline{x}_1, \bar{x}_1], ..., [\underline{x}_2, \bar{x}_2]) \quad \textbf{(2)}$$

et il est représenté par le rectangle ci-dessous

Représentation de l'objet S dans un espace à 2 dimensions.

Exemple 2
Pour $n = 3$ l'objet S décrit par

$$S = ([\underline{x}_1, \bar{x}_1], [\underline{x}_2, \bar{x}_2], [\underline{x}_3, \bar{x}_3]) \qquad \textbf{(3)}$$

et il est représenté par l'hypercube à $(2^3 = 8)$ sommets suivant

Représentation de l'objet S dans un espace à 3 dimensions.
Un hypercube dans un espace de dimension n sera décrit par une matrice à 2^n lignes et n colonnes où la ième ligne correspond aux coordonnées du ième sommet. Ainsi,
- L'objet S défni en (2) sera décrit par la matrice suivante

$$M = \begin{pmatrix} \underline{x}_1 & \underline{x}_2 \\ \underline{x}_1 & \bar{x}_2 \\ \bar{x}_1 & \underline{x}_2 \\ \bar{x}_1 & \bar{x}_2 \end{pmatrix} \qquad \textbf{(4)}$$

- L'objet S défini en (3) sera décrit par la matrice suivante :

$$M = \begin{pmatrix} \underline{x}_1 & \underline{x}_2 & \underline{x}_3 \\ \underline{x}_1 & \underline{x}_2 & \bar{x}_3 \\ \underline{x}_1 & \bar{x}_2 & \underline{x}_3 \\ \underline{x}_1 & \bar{x}_2 & \bar{x}_3 \\ \bar{x}_1 & \underline{x}_2 & \underline{x}_3 \\ \bar{x}_1 & \underline{x}_2 & \bar{x}_3 \\ \bar{x}_1 & \bar{x}_2 & \underline{x}_3 \\ \bar{x}_1 & \bar{x}_2 & \bar{x}_3 \end{pmatrix} \qquad \textbf{(5)}$$

Notons qu'un objet peut-être caractérisé soit par un vecteur de composantes de type intervalle (2), (3) ; soit par une matrice réelle (ou à éléments réels) (4), (5).

3.3. b) Algorithme de la méthode des sommets [2]

1 - Chaque objet S_i est décrit par une matrice de données numériques M_i à 2^n lignes et n colonnes dont les éléments sont les n coordonnées des 2^n sommets des hypercubes associés.

2 - Puis on construit une nouvelle matrice M à $2^n \times m$ lignes et n colonnes en concatenant les m matrices M_i précédentes. Ainsi au tableau $(m \times n)$ suivant :

$$S = \begin{pmatrix} S_1 \\ \vdots \\ S_m \end{pmatrix} = \begin{pmatrix} [\underline{x}_{11}, \bar{x}_{11}] & \dots & [\underline{x}_{1n}, \bar{x}_{1n}] \\ \vdots & \ddots & \vdots \\ \underline{x}_{m1}, \bar{x}_{m1} & \dots & [\underline{x}_{mn}, \bar{x}_{mn}] \end{pmatrix} \qquad \textbf{(6)}$$

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

$$M = \begin{pmatrix} M_1 \\ \vdots \\ M_m \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \underline{x}_{11} & \dots & \underline{x}_{1n} \\ \vdots & \ddots & \vdots \\ \bar{x}_{11} & \dots & \bar{x}_{1n} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \underline{x}_{m1} & \dots & \underline{x}_{mn} \\ \vdots & \ddots & \vdots \\ \bar{x}_{m1} & \dots & \bar{x}_{mn} \end{bmatrix} \end{pmatrix} \qquad \textbf{(7)}$$

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, on donne ainsi la même importance à chacun des 2^n sommets associés à S_i.

3 - Ensuite on applique l'ACP classique à la matrice M de données numériques définie en (7).

Soient $Y_1, Y_2, ..., Y_p \quad (p \leq n)$ les p premières composantes principales (à valeurs numériques) issues de cette ACP et $\lambda_1, ..., \lambda_p$ les valeurs propres associées.

4 - Enfin on détermine les composantes principales à valeurs intervalles $Y_1^I, ..., I_p^I$ à partir des composantes numériques $Y_1, ..., Y_p$.

Soit L_{S_i} l'ensemble des numéros de lignes dans la matrice M associés à l'objet S_i et $y_{kj} \quad k \in L_{S_i}$, la valeur de la jème composante principale numérique Y_j associée au somme de l'objet S_i correspondant à la kème ligne

de M. La valeur de la jème composante principale de type intervalle Y_j^I pour l'objet S_i est alors $y_{S_{ij}}^I = [\underline{y}_{ij}, \bar{y}_{ij}]$ **(8)** avec, $\underline{y}_{ij} = \min_{k \in L_{s_i}}(y_{kj})$ *et* $\bar{y}_{ij} = \max_{k \in L_{s_i}}(y_{kj})$.

• Explication des étapes de l'algorithme précédent

1 - Chaque objet $S_i = ([\underline{x}_{i1}, \bar{x}_{i1}], ..., [\underline{x}_{in}, ..., \bar{x}_{in}])$ est décrit par une matrice de données numériques M_i à 2^n lignes et n colonnes

$$\forall i = 1, 2, ..., m \ ; \ M_i = (X_1, X_2, ..., X_n) \text{ où } X_1 = \begin{pmatrix} \underline{x}_{11} \\ \vdots \\ \underline{x}_{11} \\ \bar{x}_{11} \\ \vdots \\ \bar{x}_{11} \end{pmatrix} \begin{matrix} 2^n/2 = 2^{n-1} \ fois \\ \\ \\ 2^{n-1} \ fois \end{matrix}$$

et $\forall j = 1, 2, ..., n \ ; \ X_j = \begin{pmatrix} \underline{x}_{ij} \\ \vdots \\ \bar{x}_{ij} \end{pmatrix}$ $\underline{x}_{ij}$ apparaît 2^{n-1} fois et $\bar{x}_{ij}$ apparaît 2^{n-1} fois.

2 - On a fait correspondre au tableau

$$S = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_m \end{pmatrix} = \begin{pmatrix} [\underline{x}_{11}, \bar{x}_{11}] & \dots & [\underline{x}_{1n}, \bar{x}_{1n}] \\ \vdots & \ddots & \vdots \\ \underline{x}_{m1}, \bar{x}_{m1} & \dots & [\underline{x}_{mn}, \bar{x}_{mn}] \end{pmatrix}$$

la matrice $M = \begin{pmatrix} M_1 \\ \vdots \\ M_m \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \underline{x}_{11} & \dots & \underline{x}_{1n} \\ \vdots & \ddots & \vdots \\ \bar{x}_{11} & \dots & \bar{x}_{1n} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \underline{x}_{m1} & \dots & \underline{x}_{mn} \\ \vdots & \ddots & \vdots \\ \bar{x}_{m1} & \dots & \bar{x}_{mn} \end{bmatrix} \end{pmatrix}$

M a $m \times 2^n$ lignes et n colonnes, si on pose $M = \begin{pmatrix} X_1^1 & \dots & X_n^1 \\ \vdots & \ddots & \vdots \\ X_1^m & \dots & X_n^m \end{pmatrix}$ on aura : $\forall (i,j) \in \mathbb{R}^m \times \mathbb{R}^n, X_j^i = \begin{pmatrix} \underline{x_{ij}} \\ \vdots \\ \bar{x}_{ij} \end{pmatrix}$ où chacune des $\underline{x}_{ij}$ et $\bar{x}_{ij}$ apparaît

2^{n-1} fois.

3 - La matrice qu'on diagonalise est la matrice variance V_S a n lignes et n colonnes dont le terme général $(v_s)_{jj'}$ est la covariance entre X_j et $X_{j'}$ (où $X_j = \begin{pmatrix} \underline{x_{ij}} \\ \vdots \\ \bar{x}_{ij} \end{pmatrix}$; $\underline{x}_{j'}$ et $\bar{x}_{ij}$ apparaîtrons chacune 2^{n-1} fois et $X_j j' = \begin{pmatrix} \underline{x_{ij'}} \\ \vdots \\ \bar{x}_{ij'} \end{pmatrix}$; $\underline{x}_{ij'}$ *et* $\bar{x}_{ij}$ apparaîtrons chacune 2^{n-1} fois et chacun des X_j et $X_{j'}$ a 2^n coordonnées).

Proposition :

$$V_S = \left[(v_s)_{jj'}\right]_{1\leq j,j'\leq n} \text{ où } (v_s)_{jj'} = \begin{cases} \sum_{i=1}^{m} \frac{p_i}{4} \ (\underline{x}_{ij} + \bar{x}_{ij})(\underline{x}_{ij'} + \bar{x}ij') & si\ j \neq j' \\ \sum_{i=1}^{m} \frac{p_i}{2}(\underline{x}_{ij}^2 + \bar{x}_{ij}^2) & si\ j = j' \end{cases} \quad \textbf{(9)}$$

Démonstration

$(\mathcal{V}_s)_{jj'} Cov(X_j, X_{j'} = E(X_j X_{j'} - E(X_j)E(X_{j'}),$ mais $\forall j = 1, ..., n : E(X_j) = 0$ car les variables $X_j, j = 1, ..., n$ sont centrées par hypothèse.

D'où $(\mathcal{V}_s)_{jj'} = E(X_j X_{j'} = \sum_{i=1}^{m} \frac{p_i}{2^n}(\underline{x}_{ij} + \bar{x}_{ij})(\underline{x}_{ij'} + \bar{x}_{ij'}) \times 2^{n-2}$. Puisque le produit des coordonnées de chacun des quatre sommets du rectangle défini par $(\underline{x}_{ij}, \bar{x}_{ij})$ *et* $(\underline{x}_{ij'}, \bar{x}_{ij'})$ apparaît 2^{n-2} fois.

P_i étant le poids de l'individu ; $i \ \ i = 1, ..., m$ et chaque sommet muni du poids $\frac{p_i}{2^n}$.

Si $j = j'$ on aura $(\mathcal{V}_s)_{jj} = Var(X_j) = E(X_j^2) = \sum_{i=1}^{m} \frac{p_i}{2^n}(2^{n-1}\underline{x}_{ij}^2 + 2^{n-1}\bar{x}_{ij}^2) = \sum_{i=1}^{m} \frac{p_i . 2^{n-1}}{2^n}(\underline{x}_{ij}^2 + \bar{x}_{ij}^2) = \sum_{i=1}^{m} \frac{p_i \cdot}{2}(\underline{x}_{ij}^2 + \bar{x}_{ij}^2)$ (Puisque chacune des $\underline{x}_{ij}^2$ *et* $\bar{x}_{ij}$ apparaît 2^{n-1} fois dans le vecteur X_j).

Après avoir appliqué l'ACP classique à la matrice M on note $Y_1, Y_2, ..., Y_p$ $(p \leq n)$ les p premières composantes principales issues de cette ACP et $\lambda_1, \lambda_2, ..., \lambda_p$ les valeurs propres associées $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.
Chaque composante principale $Y_j; j = 1, ..., p$ à $m2^n$ coordonnées. En effet, si on pose :

$$M = \begin{pmatrix} M_1 \\ \vdots \\ M_i \\ \vdots \\ M_m \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \underline{x}_{11} & \dots \underline{x}_{1n} \\ \vdots & \vdots \\ \bar{x}_{11} & \dots \bar{x}_{1n} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \underline{x}_{i1} & \dots \underline{x}_{in} \\ \vdots & \vdots \\ \bar{x}_{i1} & \dots \bar{x}_{in} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \underline{x}_{m1} & \dots \underline{x}_{mn} \\ \vdots & \vdots \\ \bar{x}_{m1} & \dots \bar{x}_{mn} \end{bmatrix} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} X^1 \\ \vdots \\ X^{2^n} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} X^{(i-1)2^n+1} \\ \vdots \\ X^{i.2^n} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} X^{(m-1)2^n+1} \\ \vdots \\ X^{m.2^n} \end{bmatrix} \end{pmatrix}$$

$$\begin{aligned} &L_{S_1} = \{1, ..., 2^n\} \\ &\vdots \\ &L_{S_i} = \{(i-1)2^n + 1, ..., i.2^n\} \\ &\vdots \\ &L_{S_m} = \{(m-1)2^n + 1, ..., m.2^n\} \end{aligned}$$

On aura $Y_j = \begin{pmatrix} \langle X^1, u_j \rangle \\ \vdots \\ \langle X^k, u_j \rangle \\ \vdots \\ \langle X^{m.2^n}, u_j \rangle \end{pmatrix}$ où u_j est le vecteur propre de V_s associé à λ_j.

4) Maintenant on pose $Y_j = \begin{pmatrix} y_{1j} \\ \vdots \\ y_{kj} \\ \vdots \\ y_{m.2^n j} \end{pmatrix}$ où $y_{kj} = \langle X^k, u_j \rangle; k \in \{1, ..., m.2^n\}$.

La jème composante principale à valeurs intervalles Y_j^I où $j \in \{1, ..., p\}$ s'obtient à partir de Y_j comme suit :

Soient $k \in L_{S_i} = \{(i-1)2^n + 1, ..., i.2^n\}$,

$$\begin{array}{ll} \underline{y}_{ij} & = \min\limits_{k\in L_{S_i}}(y_{kj}) \\ \bar{y}_{ij} & = \max\limits_{k\in L_{S_i}}(y_{kj}) \end{array} \quad \text{et} \quad y^I_{S_ij} = [\underline{y}_{ij}, \bar{y}_{ij}].$$

Dans ce cas-ci on a :$Y^I_j = \begin{pmatrix} y^I_{S_{1j}} \\ \vdots \\ y^I_{S_{mj}} \end{pmatrix} = \begin{pmatrix} [\underline{y}_{1j}, \bar{y}_{1j}] \\ \vdots \\ \underline{y}_{mj}, \bar{y}_{mj} \end{pmatrix}$. **10**

• Qualité de Représentation des individus

Comme nous l'avons vu au 1er chapitre. La représentation du nuage $N(I) = \{(X^k, p_k); k = 1, ..., m.2^n\} \subset \mathbb{R}^n$ dans le sous-espace factoriel, de dimension p, E_p en donne une image approximative. La qualité globale de cette représentation est mesurée par le pourcentage d'inertie pris en compte par $E_p = \dfrac{\lambda_1 + \lambda_2 + ... + \lambda_p}{tr(V_S)} \times 100$.

Il est important de pouvoir juger de la qualité de représentation de chaque point X^k sur les axes factoriels. Les vecteurs unitaires $u_1, ..., u_n$ des axes factoriels constituent une base M-orthonormée de $\mathbb{R}^n$ et on a : $X^k = \sum\limits_{j=1}^{n} y_{kj}u_j$ où $Y_j = (y_{kj}, k = 1, ..., m.2^n)$ étant la jème composante principale. D'où $\|X^k\|_M = \sum\limits_{j=1}^{n}(y_{kj})^2$ et $I_T = \sum\limits_{k=1}^{m.2^n} p_k\|X^k\|^2_M = \sum\limits_{j=1}^{n}\sum\limits_{k=1}^{m.2^n} p_k(y_{kj})^2 = \sum\limits_{j=1}^{n}\lambda_j$.

On en déduit les indices de qualité de représentation d'un point X^k : $y^2_{kj}/_{\|X^k\|^2_M}$ est la contribution relative du jème axe factoriel à l'inertie de X^k c'est la part d'inertie de X^k prise en compte par cet axe. Cette quantité est le cosinus carré de l'angle θ^k_j formé par X^k et u_j.

De plus $\sum\limits_{\ell=1}^{p} y^2_{\ell k}/_{\|X^k\|^2_M} = \sum\limits_{\ell=1}^{p} cos^2\theta^k_\ell$ est la contribution relative de l'espace factoriel E_p engendré par les p premiers axes factoriels à l'inertie de X^k. Par ailleurs, $P_k y^2_{kj}/_{\lambda_j}$ est la contribution relative de X^k à l'inertie du jème axe. C'est la part d'inertie de cet axe prise en compte par le point X^k.

• Paramétres d'aide à l'interprétation

Avant de préciser ces paramètres, nous présentons maintenant les notions qu'ils utilisent :

- Centre de gravité pour l'objet

$S_i = ([\underline{x}_{i1}, \bar{x}_{i1}], ..., [\underline{x}_{in}, \bar{x}_{in}])$ est le point de $\mathbb{R}^n$ défini par $G_i\left(\dfrac{\underline{x}_{i1} + \bar{x}_{1n}}{2}, ..., \dfrac{\underline{x}_{n1} + \bar{x}_{in}}{2}\right)$. **(11)**.

- Centre de gravité pour le système constitué par les objets $S_i, i = 1, ..., m$ est le point de $\mathbb{R}^n$ définie par

$$G = \sum_{i=1}^{m} G_i = \left(\frac{1}{2} \sum_{i=1}^{m} (\underline{x}_{i1} + \bar{x}_{i1}), .., \frac{1}{2} \sum_{i=1}^{m} (\underline{x}_{in} + \bar{x}_{in}) \right) \qquad \textbf{(12)}$$

.

- $d(k, G)$ est la distance entre le sommet k et G.

Maintenant, les paramètrs d'interprétation se généralisent très naturellement : Pour mesurer la qualité de la représentation de l'objet S_i sur l'axe factoriel Δu_j de direction u_j, on peut proposer :

- la formule qui 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 comme suit :

$$COR_I^1(S_i, u_j) = \frac{\sum\limits_{k \in L_{S_i}} P_k y_{kj}^2}{\sum\limits_{k \in L_{S_i}} p_k d^2(k, G)} = \frac{\sum\limits_{k \in L_{S_i}} \frac{p_i}{2^n}\ y_{kj}^2}{\sum\limits_{k \in L_{S_i}} \frac{p_i}{2^n}\ d^2(k, G)} = \frac{\frac{p_i}{2^n} \sum\limits_{k \in L_{S_i}} y_{kj}^2}{\frac{p_i}{2^n} \sum\limits_{k \in L_{S_i}} d^2(k, G)}$$

$$= \frac{\sum\limits_{k \in L_{S_i}} y_{kj}^2}{\sum\limits_{k \in L_{S_i}} d^2(k, G)}$$

D'où

$$COR_I^1(S_i, u_j) = \frac{\sum\limits_{k \in L_{S_i}} y_{kj}^2}{\sum\limits_{k \in L_{S_i}} d^2(k, G)} \qquad \textbf{(13)}$$

- ou bien la formule qui 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. Comme suit :

$$COR_I^2(S_i, u_j) = \frac{1}{2^n} \sum_{k \in L_{S_i}} \frac{y_{kj}^2}{d^2(k, G)} \qquad \textbf{(14)}$$

On mesure de même la contribution de S_i

- à l'inertie λ_j du jème axe factoriel par :

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

D'où

$$CTR_I(S_i, u_j) = \frac{p_i}{(\lambda_j 2^n)} \sum_{k \in L_{S_i}} y_{kj}^2 \qquad \textbf{(15)}$$

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

$$INR_I(S_i) = \sum_{k \in L_{S_i}} p_k \, d^2(k, G) / I_T = \sum_{k \in L_{S_i}} \frac{p_i}{2^n} d^2(k, G) / \sum_{j=1}^{n} \lambda_j = \frac{p_i}{2^n \sum_{j=1}^{n} \lambda_j} \cdot \sum_{k \in L_{S_i}} d^2(k, G)$$

D'où

$$INR_I(S_i) = \sum_{k \in L_{S_i}} \frac{p_i}{2^n} \ d^2(k, G) / \sum_{j=1}^{n} \lambda_j \qquad \textbf{(16)}$$

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

3.4. Méthode des centres

3.4. a) Introduction

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 (17)

$$\begin{pmatrix} x_{11}^c & \dots & x_{1n}^c \\ \vdots & & \vdots \\ x_{m1}^c & \dots & x_{mn}^c \end{pmatrix} \qquad \textbf{(17)}$$

avec

$$x_{ij}^c = \frac{\underline{x}ij + \bar{x}_{ij}}{2} \qquad \textbf{(18)}$$

3.4.b) Algorithme de la méthode des centres [2]

1 - Transformer la matrice donnée en (6) en la matrice donnée en (17). Soient $X_1^c, ..., X_n^c$ 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 kième axe principal du point C_i (centre de l'hypercube associé à l'objet S_i) de coordonnées $(x_{i1}^c, ..., x_{in}^c)$. Cette valeur est obtenue à l'aide de la formule donnée en (19), où $\bar{X}_u^c$ est la moyenne de la variable X_j^c et u_{jk} la jème composante du kème vecteur axial factoriel, $y_{ik}^c = \sum_{j=1}^{n} \left(\underline{x}_{ij}^c - \bar{X}_j^c\right) \cdot u_{jk}$

• **Explication des étapes de l'Algorithme précédent** :

1 - On transforme la matrice $S = \begin{pmatrix} S_1 \\ \vdots \\ S_m \end{pmatrix} = \begin{pmatrix} [\underline{x}_{11}, \bar{x}_{11}] & \dots & [\underline{x}_{1n}, \bar{x}_{1n}] \\ \vdots & & \\ \underline{x}_{m1}, \bar{x}_{m1} & \dots & \underline{x}_{mn}, \bar{x}_{mn} \end{pmatrix}$ en la matrice $S^c = \begin{pmatrix} x_{11}^c & \dots & x_{1n}^c \\ \vdots & & \vdots \\ x_{m1}^c & \dots & x_{mn}^c \end{pmatrix}$

où $x_{ij}^c = \frac{\underline{x}_{ij} + \bar{x}_{ij}}{2} \quad \forall (i,j) \in \mathbb{R}^m \times \mathbb{R}^n$.

Les nouvelles variables sont $X_j^c = \begin{pmatrix} x_{1j}^c \\ \vdots \\ x_{mj}^c \end{pmatrix} \quad j = 1, ..., n.$

2 - On applique l'ACP classique sur la matrice $S^c = \begin{pmatrix} x_{11}^c & \dots & x_{1n}^c \\ \vdots & & \vdots \\ x_{m1}^c & \dots & x_{mn}^c \end{pmatrix}$.

On diagonalise la matrice variance $V_c = [(\upsilon_c)_{jj'}]_{1 \le j,j' \le n}$ où $(\upsilon_c)_{jj'} = COV(X_j^c, X_{j'}^c)$ avec $X_j^c = \begin{pmatrix} x_{1j}^c \\ \vdots \\ x_{mj}^c \end{pmatrix}$ et $X_{j'}^c = \begin{pmatrix} x_{1j'}^c \\ \vdots \\ x_{mj'}^c \end{pmatrix}$ en supposans

que chaque variable $X_j, j \in \{1, ..., n\}$ est centrée c'est-à-dire

$$\bar{X}_j^c = E(X_j) = \sum_{j=1}^{m} p_i x_{ij}^c = 0$$

Proposition :

Le terme général de V_c est $(v_c)_{jj'} = \begin{cases} \sum_{i=1}^{m} p_i x_{ij}^c \; x_{ij'}^c & si \quad j \neq j' \\ \sum_{i=1}^{m} p_i (x_{ij}^c)^2 & si \quad j = j' \end{cases}$

Preuve :

$COV(X_j^c, X_{j'}^c) = (v_c)_{jj'} = E(X_j^c X_{j'}^c) - E(X_j^c)E(X_{j'}^c) = E(X_j^c X_{j'}^c) = \sum_{i=1}^{m} p_i x_{ij}^c x_{ij'}^c$

et $Var(X_j^c) = (v_c)_{jj} = E(X_j^{c^2} - E^2(X_j^c) = E(X_j^{c^2}) = \sum_{i=1}^{m} p_i (x_{ij}^c)^2$.

3 - Le centre C_i de l'hypercube associé à l'objet S_i es défini par :

$$c_i = (x_{i1}^c, ..., x_{in}^c) \in \mathbb{R}^n$$

En utilisant la formule (19) on obtient que les coordonnées du point C_i (centre de l'hypercube associé à l'objet S_i) dans l'espace constitué par les axes factoriels obtenu après avoir appliqué l'ACP classique sur la matrice S^c, sont $(y_{i1}^c, ..., y_{in}^c) = \left(\sum_{j=1}^{n} (\underline{x}_{ij}^c - \bar{X}_j^c) u_{j_1}, ..., \sum_{j=1}^{n} (\underline{x}_{ij}^c - \bar{X}_j^c) u_{jn} \right)$. Pour préciser pour chaque objet S_i les intervalles de variation sur les axes factoriels. On utilise la règle suivante :

règle [2]

Soit un point quelconque $x^r = (x_{i1}^r, ..., x_{in}^r)$ 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 $\bar{y}_{ik}$ prise par y_{ik}^r (coordonnée du point x^r. Sur l'axe factoriel k) quand les variables x_{ij}^r variant dans l'intervalle $[\underline{x}_{ij}, \bar{x}_{ij}]$ pour $j = 1, ..., n$.

Comme y_{ik}^* est une fonction linéaire des n variables $x_{i1}^r, .., x_{1n}$, varient indépendamment dans $[\underline{x}_{ij}, \bar{x}_{ij}]$ pour $j = 1, ..., n$ les valeurs extrêmes de y_{ik}^r

sont données par les formules **(20)** et **(21)**.

$$\underline{y}_{ik} = \sum_{j=1}^{n} \min_{\underline{x}_{ij} \leq x_{ij}^r \leq \bar{x}_{ij}} (x_{ij}^r - \bar{X}_j^c)\ u_{jk} \qquad \textbf{(20)}$$

$$\bar{y}_{ik} = \sum_{j=1}^{n} \max_{\underline{x}_{ij} \leq x_{ij}^r \leq \bar{x}_{ij}} (x_{ij}^r - \bar{X}_j^c)\ u_{jk} \qquad \textbf{(21)}$$

Soit encore :

$$\underline{y}_{ik} = \sum_{j,u_{jk}<0} (\bar{x}_{ij} - \bar{X}_j^c)\ u_{jk} + \sum_{j,u_{jk}>0} (\underline{x}_{ij} - \bar{X}_j^c)\ u_{jk} \qquad \textbf{(22)}$$

$$\bar{y}_{ik} = \sum_{j,u_{jk}<0} (\underline{x}_{ij} - \bar{X}_j^c)\ u_{jk} + \sum_{j,u_{jk}} (\bar{x}_{ij} - \bar{X}_j^c)\ u_{jk} \qquad \textbf{(23)}$$

$\bar{X}_j^c$ et u_{jk} désignant respectivement la moyenne de la variable X_j^c et la jème composante du kème vecteur axial factoriel.

• **Représentation des individus**

Si on considère la matrice $S^c = \begin{pmatrix} x_{11}^c & \dots & x_{1n}^c \\ \vdots & & \vdots \\ x_{m1}^c & \dots & x_{mn}^c \end{pmatrix}$ où chacun de ses éléments est une valeur numérique, on peut refaire sur S^c les mêmes étapes qu'on a fait sur la matrice $X = \begin{pmatrix} x_1^1 & \dots & x_1^p \\ \vdots & & \\ x_n^1 & \dots & x_n^p \end{pmatrix}$ au 1er chapitre.

3.5 - Comparaison des deux méthodes [2]

Nous allons voir que les matrices de variance υ_s et υ_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 :

$$\forall j = 1, ..., n : \bar{X}_j^c = \sum_{i=1}^{m} p_i\ x_{ij}^c = 0 \qquad \textbf{(24)}$$

p_i étant le poids de l'individu i

alors le terme général $(\upsilon_c)_{jj'}$ de la matrice variance dans la méthode des

centres s'écrit :

$$(\upsilon_c)_{jj'} = \sum_{i=1}^{m} p_i \; x_{ij}^c \; x_{ij'}^c \qquad \textbf{(25)}$$

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 $\bar{x}_{ij}$ apparaîtront chacune 2^{n-1} fois. La moyenne $\bar{X}_j$ de X_j s'écrira donc :

$$\bar{X}_j = \sum_{i=1}^{m} \frac{p_i}{2^n}(2^{n-1} \; \underline{x}_{ij} + 2^{n-1}\bar{x}_{ij}) = \sum_{i=1}^{m} p_i x_{ij}^c = \; \bar{X}_j^c = 0 \qquad \textbf{(26)}$$

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

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

$$(\upsilon_s)_{jj} \; = \sum_{i=1}^{m} \frac{p_i}{2^n}(2^{n-1}(\underline{x}_{ij})^2 + 2^{n-1}(\bar{x}_{ij})^2) \qquad \textbf{(27)}$$

$$= \sum_{i=1}^{m} \frac{p_i}{2^n}[(\underline{x}_{ij})^2 + (\bar{x}_{ij})^2] \qquad \textbf{(28)}$$

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

$$(\upsilon_s)_{jj} \; = \sum_{i=1}^{m} p_i \; [(x_{ij}^c)^2 + (\bar{x}_{ij} - \underline{x}_{ij})^2/4] \qquad \textbf{(29)}$$

$$(\upsilon_s)_{jj} \; = \sum_{i=1}^{m} p_i (\bar{x}_{ij} - \underline{x}_{ij})^2/4 \qquad \textbf{(30)}$$

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

$$(\underline{x}_{ij} - \bar{x}_{ij})^2/4 = \frac{1}{2}(\underline{x}_{ij} - x_{ij}^c)^2 + \frac{1}{2}(\bar{x}_{ij} - x_{ij}^c)^2 \qquad \textbf{(31)}$$

Dans le calcul de la covariance entre Xj et X'_j dans la méthode des sommets, vu que le produit des coordonnées de chacun des quatre sommets du rectangle défini par $(\underline{x}_{ij}, \bar{x}_{ij}$ et $(\underline{x}_{ij}, \underline{x}_{ij'}$ apparaît 2^{n-2} fois, on a, après

mise en facteurs :

$$(\upsilon_s)_{jj'} = \sum_{i=1}^{m} \frac{p_i}{2^n} 2^{n-2} (\underline{x}_{ij} + \bar{x}_{ij})(\underline{x}_{ij'} + \bar{x}_{ij'}) \quad \textbf{(32)}$$

$$= \sum_{i=1}^{m} p_i \; x^c_{ij} \; x^c_{ij'} = (\upsilon_c)_{jj'} \quad \textbf{(33)}$$

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 3.1. b. iv) 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 $\bar{y}_{ik}$ 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 (22) et (23) , où $\bar{X}^c_j = 0$ par hypothèse, et où u_{jk} est la jème composante du kème vecteur propre normé de la matrice υ_s).

• **Exemple des Huiles - Description des données**

Afin d'illustrer les méthodes proposées, nous utilisons les données d'Ichino (1994) de la table 1.

Chaque ligne du tableau représente une classe d'huile décrite par 4 variables quantitatives : "Specific gravity", "Freezing point", "Iodine value", "Saponification". L'intervalle $[\underline{x}_{ij}, \bar{x}_{ij}]$, croisement de la ième ligne et de la jème colonne signifie que la valeur de la jème variable pour toute huile appartenant à la ième classe d'huile, appartient à l'intervalle $[\underline{x}_{ij}, \bar{x}_{ij}]$

Tableau 1

La description des 8 classes d'huiles par 4 variables de type intervalle

Nom	Label	GRA	FRE	IOD	SAP
Linsead	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]
Cameltia	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.

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

1 - Méthode des sommets

D'abord on fait la description de chaque classe d'huile S_i
"objet S_i" $i = 1, ..., 8$ par une matrice de données numériques M_i à 2^4 lignes et 4 colonnes dont les éléments sont les 4 coordonnées des 2^4 sommets des hypercubes associés comme suit :

$$M_1 = \begin{pmatrix} 0,93 & -27,00 & 170,00 & 118,00 \\ 0,93 & -27,00 & 170,00 & 196,00 \\ 0,93 & -27,00 & 204,00 & 118,00 \\ 0,93 & -27,00 & 204,00 & 196,00 \\ 0,93 & -18,00 & 170,00 & 118,00 \\ 0,93 & -18,00 & 170,00 & 196,00 \\ 0,93 & -18,00 & 204,00 & 118,00 \\ 0,93 & -18,00 & 204,00 & 196,00 \\ 0,94 & -27,00 & 170,00 & 118,00 \\ 0,94 & -27,00 & 170,00 & 196,00 \\ 0,94 & -27,00 & 204,00 & 118,00 \\ 0,94 & -27,00 & 204,00 & 196,00 \\ 0,94 & -18,00 & 170,00 & 118,00 \\ 0,94 & -18,00 & 170,00 & 196,00 \\ 0,94 & -18,00 & 204,00 & 118,00 \\ 0,94 & -18,00 & 204,00 & 196,00 \end{pmatrix} ; M_2 = \begin{pmatrix} 0,93 & -5,00 & 192,00 & 188,00 \\ 0,93 & -5,00 & 192,00 & 197,00 \\ 0,93 & -5,00 & 208,00 & 188,00 \\ 0,93 & -5,00 & 208,00 & 197,00 \\ 0,93 & -4,00 & 192,00 & 188,00 \\ 0,93 & -4,00 & 192,00 & 197,00 \\ 0,93 & -4,00 & 208,00 & 188,00 \\ 0,93 & -4,00 & 208,00 & 197,00 \\ 0,94 & -5,00 & 192,00 & 188,00 \\ 0,94 & -5,00 & 192,00 & 197,00 \\ 0,94 & -5,00 & 208,00 & 188,00 \\ 0,94 & -5,00 & 208,00 & 197,00 \\ 0,94 & -4,00 & 192,00 & 188,00 \\ 0,94 & -4,00 & 192,00 & 197,00 \\ 0,94 & -4,00 & 208,00 & 188,00 \\ 0,94 & -4,00 & 208,00 & 197,00 \end{pmatrix}$$

Pour les classes $S_3, S_4, S_5, S_6, S_7, etS_8$ qui sont respectivement : $CO, S, Ca, O, BetH$ on construit leurs matrices $M_3, M_4, M_5, M_6, M_7etM_8$ de la même manière que celles des $S_1 = L$ dont la matrice est M_1 et $S_2 = P$ dont la matrice est M_2.

Puis on construit une nouvelle matrice M à $2^4 \times 8 = 128$ lignes et 4 colonnes en concatenant les 8 matrices précédentes comme suit :

$$M = \begin{pmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \\ M_7 \\ M_8 \end{pmatrix}$$

De plus, à chacune des lignes de M (i.e à chaque sommet), on attribue un poids, à savoir $p_i/_{2^4} = p_i/_{16}$ où $p_i = \frac{1}{8}, i =, ..., 8$ c'est-à-dire $\frac{p_i}{24} = \frac{1}{128}$.
Ensuite on applique l'ACP classique à la matrice de données numériques M.

On aura, les valeurs propres et les pourcentages d'inerties comme ceux qui figurent dans la table 2, tandis que les deux premières composantes principales de type intervalle sont données dans la table 3.

Chaque classe d'huile caractérisée par les deux composantes principales de type intervalle es visualisée dans le plans factoriel à 2 dimensions par un rectangle (figure 1).

Les corrélations entre les variables initiales et les composantes principales sont données dan la table 4, et visualisées sur la figure 2.

Tableau 2

Valeurs propres et % d'inertie

Numéro	Valeurs 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 3

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

Méthode des sommets

Label	CP_1	CP_2
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 4

Corrélation 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 : Cercle des corrélations (méthodes des Sommets)

2 - Méthode des centres

D'abord on construit la matrice M' à 8 lignes et 4 colonnes à valeurs numériques sont précisément les centres d'intervalles du tableau 1 :

$$M' = \begin{pmatrix} X_1^c & X_2^c & X_3^c & X_4^c \\ 0,935 & -22,50 & 187,00 & 157,00 \\ 0,935 & -4,50 & 200,00 & 192,50 \\ 0,920 & -3,50 & 106,00 & 193,50 \\ 0,925 & -5,00 & 110,00 & 190,00 \\ 0,920 & -18,00 & 81,00 & 191,00 \\ 0,915 & 3,00 & 84,50 & 191,50 \\ 0,865 & 34,00 & 44,00 & 194,50 \\ 0,860 & 27,00 & 65,00 & 196,00 \end{pmatrix} ; \text{ on pose } X_j^c = \begin{pmatrix} x_{1j} \\ x_{2j} \\ x_{3j} \\ x_{4j} \\ x_{5j} \\ x_{6j} \\ x_{7j} \\ x_{8j} \end{pmatrix} ; j = 1, ..., 4$$

Puis on applique l'ACP classique sur la matrice M'.

On centre les 4 variables X_1^c, X_2^c, X_3^c *et* X_4^c sachant que :

$$\bar{X}_1^c = \quad E(X_1^c) = \tfrac{1}{8}(0,935 + 0,935 + 0,92 + 0,925 + 0,92 + 0,915 + 0,865 + 0,86) = 0,909375$$

$$\bar{X}_2^c = \quad E(X_2^c) = \tfrac{1}{8}(-22,5 - 4,5 - 3,5 - 5 - 18 + 3 + 34 + 27) = 1,3125$$

$$\bar{X}_3^c = \quad E(X_3^c) = \tfrac{1}{8}(187 + 200 + 106 + 110 + 81 + 84,5 + 44 + 65) = 109,6875$$

$$\bar{X}_4^c = \quad E(X_4^c) = \tfrac{1}{8}(157 + 192,5 + 193,5 + 190 + 191 + 191,5 + 194,5 + 196) = 188,25$$

On calcule la matrice M'' dont les colonnes sont $\dfrac{X_j^c - E(X_j^c)}{\sigma(X_j^c)}; j = 1, ..., 4\,''E(X_j^c - E(X_j^c)) = 0; j = 1, ..., 4''$ et on construit la matrice variance-covariance

$V_c = [(v_c)_{jj'}]_{1\leq j,j'\leq 4}$ où $(v_c)_{jj'} = \begin{cases} \frac{1}{8}\sum_{i=1}^{8} x_{ij}x_{ij'} = (v_s)_{jj'} & j \neq j' \\ \frac{1}{8}\sum_{i=1}^{8} x_{ij}^2 & j = j' \end{cases}$

et on diagonalise V_c.

On aura les valeurs propres et les pourcentages d'inerties comme ce qui indiqué dans la table 5, le sdeux premières composantes principales des 8 classes d'huile dans la table 6 tandis que les rectangles associés sont visualisés sur la figure 3.

Les corrélations entre les variables descriptives et les composantes principales figurent dans la table 7 et sont représentées sur la figure 4.

Tableau 5

Valeurs propres et % d'inertie

Méthode des centres

Numéros	Valeurs 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 6

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

Label	CP_1	CP_2
L	[-4,80, -1,25]	[-4,64, 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]
0	[-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 7

Méthode des centres

Corrélation entre variables descriptives et composantes principales

	CP_1	CP_2	CP_3	CP_4
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 3

Projection des 8 rectangles associés aux 8 classes d'huile

(méthode des Centres) :

Figure 4

Cercle des corrélations (méthode des Centres) :

Conclusion

Les mesures caractérisent 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 exemple 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 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 intevalle 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.

Perspectives
D'autres types d'analyse factorielle actuellement en cours d'étude peuvent être envisagés pour traiter les données de type intervalles, ensemblistes, dotées de structure taxinomique à priori, etc. Dans le cas de l'ACP, on peut rapporter les problèmes de dispersion non sur les individus mais sur les variables, en remplaçant chaque variable X par deux variables X_{min} et X_{max} respectivement associés à la valeur minimale et à la valeur maximale de l'eintervalle caractérisant chaque individu pour la variable X.
On peut aussi envisager l'application de l'AFC à des données symboliques, plus précisément du type intervalle.
D'autres études sont menées pour appliquer l'analyse factorielle à des données qui sont des lois de probabilités ou des histogrammes. (voir Pierre Cazes [Ceremade - UMR 7534 - Université Paris Dauphine - Cahier n° 0114 du 4 - Juillet 2001].