Introduction

Les bases de données statatistiques manipulent des objets conventionnels décrits par des variables monovaluées (la valeur prise par une variable pour un objet est une valeur unique). Les évolutions récentes dans les systèmes de base de données permettent de stocker de nouveaux types de données (intervalles, ensembles, ...) introduisant de l'imprécision ou de la variation. Des contraintes de domaines peuvent être exprimées et des liens de hiérarchie et de composition peuvent être stockés.
Ces évolutions dans les systèmes de base de données ont donné lieu à de nombreuses applications manipulant des objets décrits de façon plus proches de la réalité et donc plus complexes qui ceux habituellement traités.
En gestion des stocks, par exemple, on décrit une situation de rupture de stock comme suit "Niveau - de - stock = $[100, 150]$, quantité - en - cours de - commande = $[50, 100]$, Durée - de - livraison = $[30, 45]$, Etat - fournisseur $\{Critique, Mauvais\}$, Etat - écoulement - produit $\{Moyen, Rapide\}$".
On peut décrire des contraintes entre des variables, par exemple, si "Etat - fournisseur) $\{critique\}$" alors "Durée - de - livraison ≥ 0 (On peut avoir des taxinomies, par exemple, dans la variable couleur les modalités blanc et jaune sont remplaçables par la modalité claire.
Des objets incluant dans leurs descriptions de telles informations sont dits symboliques (Diday, 1987, 1995)) car dans chaque case du tableau de données peuvent apparaître des valeurs multiples, parfois pondérées et liées entre elles par des règles). L'extension des méthodes d'analyse des données à de tels objets est appelé "Analyse de données symboliques". Plusieurs auteurs se sont intéressés à l'extension des méthodes de réduction de dimension et de transformation de variables à des données complexes. Nagabushan (1988) a présenté une méthode de réduction à deux dimensions s'appliquant à des objets décrits par des variables à valeurs intervalles ; cette méthode est basée sur le développement en séries de Taylor. Ichino (1994) s'est également intéressé aux problèmes de réduction de dimension ; il propose une extension de la méthode d'Analyse en composantes principales "ACP" à des objets décrits

par des variables de type intervalle, de type ensemble et même structurées. Ichino se base, pour étendre la méthode d'ACP classique à des données complexes sur la généralisation de la distance Minkowsky.
On présentera dans ce mémoire trois chapitres :
- Dans le premier chapitre, on présente l'ACP classique, en indiquant son domaine d'application, son cadre et son but.Puis nous passons à la résolution du problème, en définissant les axes factoriels ainsi que leurs facteurs associés, et les composantes principales. Nous décrivons la représentation des individus et celle des variables ainsi que la qualité de ces représentations.

- Dans le deuxième chapitre, nous présentons les données symboliques en rappelant leur définition. Nous indiquons leurs différentes descriptions, puis le modèle de base et enfin nous définissons les objets symboliques et leurs différents types.
Ce chapitre se termine par la notion de l'analyse des données symboliques.
- Dans le troisième chapitre, on présente l'ACP des données de type intervalle. Deux méthodes sont présentées : L'une dite **méthode des sommets** et l'autre dite la **méthode des centres** qui est plus adaptée à première vue, aux données décrites par un nombre élevé de variables et une comparaison de ces deux méthodes est présentée ainsi qu'un exemple d'application.

Chapitre 1

L'ANALYSE EN COMPOSANTES PRINCIPALES (ACP) " CLASSIQUE "

1.1. Domaine d'application

L'ACP permet d'analyser tout tableau de données statistiques $X(n,p)$ (n lignes, p colonnes) représentant n individus décrits par p variables quantitatives. Son domaine d'application est donc très vaste. Ainsi si l'ensembe des individus doit être homogène (ensemble d'entreprises ou ensemble de personnes par exemple), l'ensemble des variables peut être hétérogène (chiffre d'affaire, nombre d'employés pour une entreprise ou taille, poids d'un individu par exemple).

1.2. Cadre de l'ACP

1.2.a) Nuages de points associés au tableau des données

Soit $X = \{x_i^j, i = 1, ..., n; j = 1, ..., p\}$ le tableau des données, l'individu i est décrit par le vecteur de $\mathbb{R}^p$ $X_i = (x_i^j, j = 1, ..., p)$. De plus, chaque individu i est muni d'un poids $p_i : \forall i = 1, ..., n \;\; p_i > 0 \;\; ; \;\; \sum_{i=1}^{n} p_i = 1$ (en général, on a $p_i = \frac{1}{n}, \;\; \text{pour} \;\; i = 1, ..., n$).

La variable j est décrite par le vecteur de $\mathbb{R}^n : X^j = (x_i^j, i = 1, ..., n)$.

au tableau des données X sont donc associés deux nuages de points

- Le nuage de points pesants $N(I) = \{(X_i, p_i), i = 1, ..., n\} \subset \mathbb{R}^p$ dit nuage des individus.
- Le nuage $N(J) = \{X^j, j = 1, ..., p\} \subset \mathbb{R}^n$ dit nuage des variables.

Avant de préciser le but de l'ACP, nous présentons maintenant les notions qu'elle utilise et quelques résultats préliminaires.

1.2. b) Centre de gravité du nuage $N(I)$
C'est un vecteur de $\mathbb{R}^p$ que l'on notera g qui s'écrit :

$$g = \left(g^j = \sum_{i=1}^{n} p_i \; x_i^j, j = 1, ..., p \right)$$

Dans la suite, on supposera pour simplifier les calculs que $g = 0$, on peut toujours se ramener à ce cas en centrant les p variables. Matriciellement, l'égalité $g = 0$ s'écrit : ${}^tXD_p\mathbb{I} = 0$. où ${}^tX$ est la matrice transposée de X. et avec $\mathbb{I} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, Dp = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix}$ matrice des poids des individus.
Afin de définir des distances entre individus et des distances entre variables, on munit les espace $\mathbb{R}^p$ et $\mathbb{R}^n$ de métriques euclidiennes (c'est à dire associées à des matrices symétriques définies positives).

1.2.c) Métrique D_p **sur l'espace des variables** $\mathbb{R}^n$
La métrique dont on munit $\mathbb{R}^n$ est la métrique Dp dite métrique des poids. Le choix de cette métrique est naturel. En effet :

$$\langle X^j, X^{j'} \rangle_{D_p} = \; {}^tX^j D_p X^{j'} = \sum_{i=1}^{n} p_i \; x_i^j \; x_i^{j'} = COV(X^j, X^{j'})$$

car les variables sont centrées. De manière analogue :

$$||X^j||^2_{D_p} = \; {}^tX^j \; D_p \; X^j = \sum_{i=1}^{n} p_i (x_i^j)^2 = Var(X^j)$$

1.2. d) Métrique M **sur l'espace des individus** $\mathbb{R}^p$
Le choix de la matrice M dépend des caractéristiques des données. Rappelons que les plus usuelles sont la matrice identité et la matrice diagonale D_{1/σ^2}, dont le terme général de la diagonale est l'inverse de la variance des variables. Nous reviendrons sur ce choix important au paragraphe 1) 11)

1.2. e) Matrice variance - covariance du nuage $N(I)$

Le terme général de la matrice variance $V_{(p,p)}$ du nuage $N(I)$ s'écrit

$$\forall j, j' = 1, ..., p \;\; COV(X^j, X^{j'}) = \sum_{i=1}^{n} p_i \; x_i^j \; x_i^{j'} = \; {}^t X^j D_p X^{j'}$$

Matriciellement, on a donc : $V = \sum_{i=1}^{n} p_i \; X_i \; {}^t X_i$, d'où $V = \; {}^t X D_p \; X$.

1.3. Inerties

1.3.a) Inertie par rapport à un point

On rappelle que l'inertie du nuage $N(I) \subset \mathbb{R}^p$ muni de la métrique M, par rapport à un point a de $\mathbb{R}^p$ s'écrit :

$$I_a = \sum_{i=1}^{n} p_i \; d_M^2(X_i, a) = \sum_{i=1}^{n} p_i \; {}^t(X_i - a) \; M(X_i - a)$$

L'inertie par rapport au centr de gravité O de $N(I)$ s'écrit :

$$I_O = \sum_{i=1}^{n} p_i \; ||X_i||_M^2 = \sum_{i=1}^{n} p_i \; {}^t X_i \; M X_i.$$

D'après le théorème de Huygens, l'inertie de $N(I)$ par rapport à son centre de gravité est minimum. Cette quantité est désignée comme l'inertie totale du nuage $N(I)$, on notera désormais $I_O = I_T$.

Théorème de Huygens [1]

Si $g = \sum_{i=1}^{n} p_i \; X_i$ est le centre de gravité du nuage $N(I)$ on a :

$$\forall a \in \mathbb{R}^p, \;\; I_a = I_g + d_M^2(a, g)$$

Démonstration

$$\forall a \in \mathbb{R}^p, \;\; I_a = \sum_{i=1}^{n} p_i \; {}^t(X_i - a) \; M(X_i - a)$$

$$I_a = \sum_{i=1}^{n} p_i \; {}^t(X_i - g + g - a) \; M(X_i - g + g - a)$$

$$I_a = \sum_{i=1}^{n} p_i \; {}^t(X_i - g) \; M(X_i - g) + 2 \sum_{i=1}^{n} p_i \; {}^t(g - a) \; M(X_i - g)$$

$$+\sum_{i=1}^{n} p_i \ {}^t(g-a)\ M(g-a) \quad \text{(car } M \text{ est symétrique)}$$

$$I_a = I_g + d^2_M(g,a) + 2{}^t(g-a)\ M \sum_{i=1}^{n} p_i\ (X_i - g) \ \ \text{(car } \sum_{i=1}^{n} p_i = 1)$$

Or $\sum_{i=1}^{n} p_i\ (X_i - g) = 0$ par définition de g, d'où le résultat annoncé.

Remarques

- Le centre de gravité g est le point par rapport auquel l'inertie du nuage est minimum.
- I_g est l'inertie totale du nuage $N(\Omega)$ et sera souvent notée I ou T.

1.3.b) Inertie par rapport à un sous - espace affine

Soit E_1 un sous - espace vectoriel de $E = \mathbb{R}^p$. Considérons la décomposition en somme directe de $E : E = E_1 \oplus E_1^{\perp}$, $E_1^{\perp}$ étant le sous - espace vectoriel orthogonal à E_1 pour la métrique M ; $\forall\ X_i \in E$, on a $X_i = \alpha_i + \beta_i$, $\alpha_i \in E_1$, $\beta_i \in E_1^{\perp}$; cette décompositioni étant unique.

Soient F_1 et $F_1^{\perp}$ les sous - espaces affines associés à E_1 et $E_1^{\perp}$ passant par un point a de E, dont la décomposition suivant E_1 et $E_1^{\perp}$ s'écrit : $a = a_1 + a_2$. On appelle inertie du nuage $N(I)$ par rapport à F_1 la quantité :

$$I_{F_1} = \sum_{i=1}^{n} p_i\ d^2_M(\beta_i, a_2) = \sum_{i=1}^{n} p_i\ d^2_M(X_i, F_1)$$

De même on définit l'inertie $N(I)$ par rappport à $F_1^{\perp}$:

$$I_{F_1^{\perp}} = \sum_{i=1}^{n} p_i\ d^2_M(\alpha_i, a_1) = \sum_{i=1}^{n} p_i\ d^2_M(X_i, F_1^{\perp})$$

Par le théorème de Pythagore, on a :

$$d^2_M(X_i, a) = d^2_M(\alpha_i, a_1) + d^2_M(\beta_i, a_2).$$

Puisque $(\alpha_i - a_1) \in E_1$ est orthogonal à $(\beta_i - a_2) \in E_1^{\perp}$, on peut donc en déduire $I_a = I_{F_1} + I_{F_1^{\perp}}$. En particulier, si $a = 0$, on a $I_T = I_{E_1} + I_{E_1^{\perp}}$
Par le théorème de Huygens, on a :

$$\forall a \in \mathbb{R}^p,\ I_{F_1} = I_{E_1} + d^2_M(a_2, 0).$$

Autrement dit, le sous - espace affine associé à E_1 minimisant I_{F_1} est celui qui contient le centre de gravité de $N(I)$.

Remarque
Les relations précédentes permettent de voir que $I_{E_1^\perp}$ peut aussi s'interpréter comme l'inertie de la projection du nuage $N(I)$ sur E_1. On désignera $I_{E_1^\perp}$ comme l'inertie portée par E_1.

1.4 But de l'ACP

Le but de l'analyse en composantes principales est d'obtenir une réprésentation du nuage $N(I)$ de $\mathbb{R}^p$ dans un espace de dimension réduirte de telle manière que l'inertie portée par cet espace soit la plus grande possible.
La principale opération de l'ACP est de déterminer les axes principaux d'inertie du nuage autour de son centre de gravité. Ce sont les axes qui prennent le mieux en compte la dispersion du nuage au sens de la distance d_M définie sur $\mathbb{R}^p$. Ces axes principaux d'inertie appelés axes factoriels permettent de représenter les points du nuage sur des espaces de dimension réduite. Par exemple, on obtiendra une représentation plane du nuage en projetant orthogonalement au sens de la métrique M tous les points sur le plan principal d'inertie, c'est-à-dire sur l'espace de dimension 2 qui porte le plus d'inertie.

1.5. Formulation du problème de l'ACP

Mathématiquement, le problème s'énonce ainsi : Trouver le sous - espace affine E_k de dimension k $(k < p)$ tel que l'inertie du nuage $N(I)$ par rapport à E_k : $I_{E_k} = \sum_{i=1}^{n} p_i \; d_M^2(X_i, E_k)$ soit minimum.
E_k est l'espace tel que l'inertie $I_{E_k^\perp}$ du nuage projeté sur E_k soit maximum.
D'après le théorème de Huygens, E_k contient nécessairement le centre de gravité O du nuage.
Nous donnons maintenant deux théorèmes qui vont permettre de traiter le problème en plusieurs étapes.
Théorème 1 d'inclusion
Si E_{k-1} est un sous - espace vectoriel optimal de dimension $k-1$, alors la recherche d'un sous - espace vectoriel optimal de dimension k peut se faire parmi l'ensemble des sous - espaces vectoriels de dimension k contenant E_{k-1}.

Démonstration
Soit F_k un sous - espace vectoriel de dimension k et $H = F_k + E_{k-1}^\perp$.

$F_k \cap E_{k-1}^{\perp}$ ne peut être réduit au vecteur nul. Sinon on aurait

$$H = F_k \oplus E_{k-1}^{\perp} \text{ et } dim(H) = k + (p - (k-1)) = p + 1.$$

Ce qui est absurde puisque $H \subset \mathbb{R}^p$. Il existe donc $v \neq 0 \in F_k \cap E_{k-1}^{\perp}$. Soit Δv l'axe engendré par v.

Soit G l'espace supplémentaire M- orthogonal à Δv dans F_k : $F_k = G \oplus \Delta V$ et soit $E_k = E_{k-1} \oplus \Delta v$. On a $I_{F_k} = I_G + I_{\Delta v}$ car G est orthogonal à Δv, mais par hypothèse E_{k-1} est optimal, donc $I_{k-1} \leq I_G$ d'où $I_{E_k} \leq I_{F_k}$.
On peut donc restreindre la recherche d'un sous - espace optimal aux sous - espaces contenant E_{k-1}.

Théorème 2
La recherche d'un sous - espace vectoriel E de dimension k contenant un espace vectoriel F de dimension $k-1$ minimisant I_E est équivalente à la recherche d'un axe Δv , M - orthogonal à F et mnimisant $I_{\Delta v}$

Démonstration
Quel que soit l'espace E contenant F, on a une décomposition $E = F \oplus \Delta v$ avec $\Delta v \perp F$ donc $I_E = I_F + I_{\Delta v}$: I_F étant constant, minimiser I_E revient à minimiser $I_{\Delta v}$.
A partir de ces théorèmes, on ramène donc le problème de l'ACP au problème suivant :

1. Rechercher un axe $E_1 = \Delta u_1$ à inertie minimum, u_1 étant le vecteur unitaire engendrant E_1.

2. Rechercher un axe Δu_2, M - orthogonal à Δu_1 et à inertie minimum. Soit $E_2 = \Delta u_1 \oplus \Delta u_2$; E_2 est un sous - espace optimal de dimension 2.

3. Rechercher un axe Δu_k, M - orthogonal à E_{k-1} et à inertie minimum. Soit $E_k = E_{k-1} \oplus \Delta u_k$, E_k est alors une solution du problème.
   On a $E_k = \Delta u_1 \oplus \Delta u_2 \oplus ... \oplus \Delta u_k$. Les axes $\Delta u_1, \Delta u_2, ..., \Delta u_k$ sont appelés les axes factoriels.

Remarque
On a du même coup obtenu toutes les solutions pour $h < k$.

1.6. Résolution du problème

1.6. a) Résultats Préliminaires

• **Expression de l'Inertie totale I_T**

$$I_T = \sum_{i=1}^{n} p_i \ {}^t X_i \ M \ X_i.$$

Notons $tr(A)$ la trace d'une matrice A. Comme l'expression ${}^t X_i \ M \ X_i$ est un réel, on a : ${}^t X_i \ MX_i = tr({}^t X_i \ MX_i) = tr(X_i \ {}^t X_i \ M)$ car
$tr(AB) = tr(BA)$. d'où :

$$I_T = tr\left(\left(\sum_{i=1}^{n} p_i \ X_i \ {}^t X_i\right) M\right)$$

On a donc : $I_T = tr \ VM$

• **Expression de l'inertie portée par un axe**
Soit Δu l'axe engendré par le vecteur unitaire u : on a $\forall X_i \in E, \ X_i = \alpha_i + \beta_i$, avec $\alpha_i \in \Delta u$ et $\beta_i \in \Delta u^{\perp}$. Or α_i s'écrit $a_i = \alpha_i u, a_i \in \mathbb{R}$.
Donc $I_{\Delta u^{\perp}} = \sum_{i=1}^{n} p_i \ d^2_M(\alpha_i, 0)$.
α_i est la projection orthogonale de X_i sur Δu, donc $a_i = \langle u, X_i \rangle = {}^t u \ M \ X_i$ et $d^2(\alpha_i, 0) = ||a_i \ u||^2 = a_i^2 = {}^t u \ M \ X_i \ {}^t X_i \ Mu$ d'où

$$I_{\Delta u^{\perp}} = {}^t uM \left[\sum_{i=1}^{n} p_i \ X_i \ {}^t X_i\right] Mu$$

Soit $I_{\Delta u^{\perp}} = {}^t uMVMu.$

On en déduit que $I_{\Delta u} = tr(VM) - {}^t uMVMu$

• **Etude de VM :**
Rappelons tout d'abord que M est symétrique définie positive et que V est symétrique positive. Par ailleurs, VM est M symétrique : ${}^t(VM)M = M(VM)$.
On en déduit les propriétés suivantes :
- Les valeurs propres de VM sont réelles, positives ou nulles.
Il existe une base M - orthonormée de $E = \mathbb{R}^p$ constituée de vecteurs propres de VM.

1.6. b) Détermination des axes factoriels

On commence par chercher le premier axe factoriel Δu_1.
Le problème s'écrit maintenant :

Problème 1 :
Maximiser ${}^t uMVMu$ sous la contrainte ${}^t uMu = 1$.
Munissons E de la base M - orthonormée constituée des vecteurs propres $e_1, ..., e_p$ de VM associés aux valeurs propres $\lambda_1, ..., \lambda_p$ les valeurs propres étant rangées par ordre décroissant $(\lambda_1 > \lambda_2 > ... > \lambda_p \geq 0)$.
Dans cette base, le vecteur u_1 cherché s'écrit :

$$u_1 = \sum_{i=1}^{p} \alpha_j \ e_j \quad \text{avec} \quad \sum_{i=j}^{p} \alpha_j^2 = 1$$

et on a :

$${}^t u_1 \ MVMu_1 = \langle VM\left(\sum_{j=1}^{p} \alpha_j \ e_j\right), \ \sum_{j'=1}^{p} \alpha_{j'} e_{j'} \rangle = \langle \sum_{j=1}^{p} \lambda_j \alpha_j \ e_j, \ \sum_{j'=1}^{p} \alpha_{j'} e_{j'} \rangle$$

d'où ${}^t u_1 MVMu_1 = \sum_{j=1}^{p} \lambda_j \ \alpha_j^2$.

On doit donc maximiser $\sum_{i=j}^{p} \lambda_j \ \alpha_j^2$ sous la contrainte $\sum_{j=1}^{p} \alpha_j^2 = 1$.

Or

$$\sum_{j=1}^{p} \lambda_j \ \alpha_j^2 \leq \lambda_1 \sum_{j=1}^{p} \alpha_j^2 = \lambda_1.$$

Il suffit donc de prendre $\alpha_1 = 1$ et $\alpha_j = 0$, pour $j > 1$.
Finalement :
u_1 est le vecteur propre de VM associé à la plus grande valeur propre λ_1.
$\lambda_1 = I_{\Delta u_1^{\perp}}$ représente l'inertie du nuage $N(I)$ projeté sur le premier axe factoriel Δu_1.
$\dfrac{\lambda_1}{tr(VM)}$ est la part d'inertie du nuage porté par le premier axe factoriel.
Le deuxième axe factoriel Δu_2 est engendré par le vecteur u_2 , qui est solution du problème.

Problème 2 :
Maximiser ${}^t u_2 MVMu_2$ sous les contraintes : ${}^t u_2 Mu_2 = 1$ et ${}^t u_2 Mu_1 = 0$.

Partant de l'écriture $u_2 = \sum_{j=1}^{p} \alpha_j \, e_j$ avec $\sum_{j=1}^{p} \alpha_j^2 = 1$ dans la base des vecteurs propres de VM. On montre de manière analogue à l'étape précédente que u_2 est le vecteur propre de VM associé à la deuxième plus grande valeur propre λ_2 qui est l'inertie portée par l'axe Δu_2.

$\frac{\lambda_1 + \lambda_2}{tr(VM)}$ est la part d'inertie du nuage $N(I)$ porté par le premier plan factoriel engendré par (u_1, u_2).
La recherche du k^{ime} axe factoriel Δu_k engendré par u_k se mène de manière analogue.
u_k est le vecteur propre unitaire de VM associé à la k^{ime} plus grande valeur propre λ_k qui est l'inertie portée par l'axe u_k.
$\frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{tr(VM)}$ est la part d'inertie du nuage $N(I)$ portée par l'espace factoriel E_k de dimension k, avec :

$$E_k = \Delta u_1 \oplus \Delta u_2 \oplus \Delta u_3 \oplus ... \oplus \Delta u_k$$

1.7. Facteurs associés aux axes factoriels

A tout vecteur unitaire u de $E = \mathbb{R}^p$ est canoniquement associé la forme linéaire b sur $\mathbb{R}^p$ définie par l'opérateur de la projection sur l'axe Δu. On a donc $b(X) = {}^tXMu$ que l'on notera ${}^tX.b$ en identifiant le vecteur Mu à la forme linéaire b,

$$ {}^tXMu = {}^tXb = b(X)$$

Ainsi aux axes factoriels de vecteurs unitaires $u_1, u_2, ..., u_r$ (r étant le rang de X) sont associées les formes linéaires $b_1, b_2, ..., b_r$ appelées facteurs de l'analyse en composantes principales.
Il est facile de voir que le premier facteur b_1 est vecteur propre de MV associé à la valeur propre λ_1, que le deuxième facteur b_2 est vecteur propre de MV associé à la valeur propre λ_2. etc...
Les facteurs caractérisent les axes factoriels aussi bien que les valeurs $u_1, ..., u_r$.
Ainsi, on montre de mnaière immédiate que la recherche du premier axe factoriel qui est de maximiser ${}^tuMVMu$ sous la contrainte ${}^tuMu = 1$ revient à la recherche de la forme linéaire $b = Mu$ qui maximise ${}^tbVb$ sous la contrainte ${}^tbM^{-1}b = 1$; $b_1 = Mu_1$ est la solution de ce problème. Plus généralement la recherche du k^{ime} axe factoriel Δu_k revient à rechercher la forme linéaire $b = Mu$ qui maximise ${}^tbVb$ sous les contraintes ${}^tbM^{-1}b = 1$ et ${}^tb_\ell M^{-1}b = 0$ pour $\ell = 1, ..., k-1; b_k = Mu_k$ est solution de ce problème.

1.8. Composantes principales

1.8.a) Définition
Pour tout $i = 1, .., n$ la projection de X_i sur le premier axe factoriel Δu_1 s'écrit : $C_1^i u_1$ avec $C_1^i = \langle X_i, u_1 \rangle = {}^t X_i M u_1 = {}^t X_i b_1$.
Le vecteur $C_1 = (C_1^i, i = 1, ..., n)$ de $\mathbb{R}^n$ s'appelle la première composante principale et s'écrit : $C_1 = XMu_1 = Xb_1$.
On définit de manière analogue les autres composantes principales. On notera C_k la k^{ime} composante principale.

1.8.b) Propriétés des composantes principales

Proposition 1
Les composantes principales C_k sont centrées, de variance λ_k et non corrélées deux à deux.

Démonstration
La moyenne de la k^{ime} composante C_k s'écrit :

$$\sum_{i=1}^{n} p_i \, C_k^i = \sum_{i=1}^{n} p_i \, {}^t u_k \, MX_i = {}^t u_k \, M \left(\sum_{i=j}^{n} p_i \, X_i \right) = 0$$

Calculons la covariance de C_k et $C_{k'}$.

$$COV(C_k, C_{k'}) = {}^t C_k \, D_p \, C_{k'} = {}^t b_k \, {}^t X \, D_p \, X b_{k'} = {}^t b_k V b_{k'} = {}^t u_k M V M u_{k'}$$

$$COV(C_k, C_{k'}) = \lambda_k \langle u_k, u_{k'} \rangle$$

D'où $COV(C_k, C_{k'} = 0$, Si $k \neq k'$ et on obtient $Var(C_k) = \lambda_k$.
proposition 2
Les composantes principales C_k sont vecteurs propres de $XM{}^t XD_p$ associées aux valeurs propres λ_k.

Démonstration
De $MVb_k = \lambda_k \, b_k$ on tire $XMVb_k = \lambda_k X b_k$ Ce qui implique $XM{}^t XD_p X b_k = \lambda_k X b_k$ d'où le résultat puisque $C_k = Xb_k$.

1.9. Représentation des individus
La projection du nuage $N(I)$ dans les sous - espace vectoriel, de dimension

k, E_k en donne une image approximative. La qualité globale de cette représentation est mesurée par le pourcentage d'inertie pris en compte par E_k :

$$\frac{\lambda_1 + ... + \lambda_k}{tr(VM)} \times 100.$$

Il est important de pouvoir juger de la qualité de représentation de chaque point X_i sur les axes factoriels. Les vecteurs unitaires $u_1, ..., u_p$ des axes factoriels constituent une base M - orthonormée de $\mathbb{R}^p$ et on a :

$$X_i = \sum_{k=1}^{p} c_k^i \ u_k \ , \ c_k = (c_k^i, \ i = 1, ..., n)$$

étant la k^{ime} composante principale. D'où

$$||X_i||_M^2 = \sum_{k=1}^{p} (C_k^i)^2 \text{ et } I_T = \sum_{i=1}^{n} p_i ||X_i||_M^2 = \sum_{k=1}^{p} \sum_{i=1}^{n} p_i (c_k^i)^2 = \sum_{k=1}^{p} \lambda_k$$

On en déduit les indices de qualité de représentation d'un point X_i.

• $(C_k^i)^2 \Big/ _{||X_i||_M^2}$ est la contribution relative du k^{ime} axe factoriel à l'inertie de X_i : c'est la part d'inertie de X_i prise en compte par cet axe. Cette quantité est le cosinus carré de l'angle θ_k^i formé par X_i et u_k.

• De plus, $\sum_{\ell=1}^{k} (C_\ell^i)^2 \Big/ _{||X_i||_M^2} = \sum_{\ell=1}^{k} cos^2 \ \theta_\ell^i$ est la contribution relative de l'espace factoriel E_k engendré par les k premiers axes factoriels à l'inertie de X_i.

• Par ailleurs, $\dfrac{p_i \ (C_k^i)^2}{\lambda_k}$ est la contribution relative de X_i à l'inertie du kième axe.
C'est la part d'inertie de cet axe prise en compte ("expliquée") par le point X_i.

1.10. Représentation des variables

A chaque vecteur propre unitaire u_k de VM correspond une composante principale $C_k \in \mathbb{R}^n$: $C_k = (C_k^i = \ ^t u_k M X_i, \ i = 1, ...n)$.
$\mathbb{R}^n$ étant muni de la métrique D_p, on a vu au paragraphe 1) 8) que les p

composantes principales $\left(\frac{C_k}{\sqrt{\lambda_k}}\ ,\ k\ =\ 1,...,p\right)$ forment un système D_p - orthonormé de $\mathbb{R}^n$. (Si X est de rang $r < p$, on se restreint aux r composantes principales associées aux valeurs propres non nulles, on suppose ici que $r = p$).
On peut donc représenter les variables X^j dans la base constituée par les vecteurs $\left(\frac{C_k}{\sqrt{\lambda_k}}\ , k = 1,...,p\right)$ du nuage formé par les p points $(X^1, X^2, ..., X^p)$ de $\mathbb{R}^n$.
La k^{ime} coordonnée de X^j dans cette base s'écrit :

$$d_j^k = \frac{{}^tC_k}{\sqrt{\lambda_k}}\ D_p\ X^j \quad \text{et on a } ||X^j||^2_{D_p} = \sum_{k=1}^{p} (d_j^k)^2,$$

d_j^k est la covariance entre la variable X^j et la composante principale normée $\frac{C_k}{\sqrt{\lambda_k}}$.
La qualité de représentation d'une variable X^j sur l'axe engendré par $\frac{C_k}{\sqrt{\lambda_k}}$ se mesure, comme pour les individus, par le Cosinus carré de l'angle formé par X^j et C_k :
$(d_j^k)^2 \Big/ _{||X^j||^2_{D_p}}$ est le carré du coefficient de corrélation entre X^j et C_k.
De même, la qualité de représentation d'une variable X^j sur l'espace engendré par les k premières composantes principales normées $\left(\frac{C_1}{\sqrt{\lambda_1}}, ...,\ \frac{C_k}{\sqrt{\lambda_k}}\right)$ sera mesurée par :

$$\sum_{\ell=1}^{k} \frac{(d_j^\ell)^2}{||X^j||^2_{D_p}}$$

qui est le carré du coefficient de corrélation multiple de X^j avec $(C_1, ..., C_k)$: une variable sera d'autant mieux représentée que cette quantité sera proche de un.
Par ailleurs, dans le cas très général où la métrique M est diagonale et si on note M_j le terme diagonal générique de M, on montre [Cazes, 1985] que l'inertie totale peut s'écrire :

$$I_T = \sum_{k=1}^{p} \sum_{j=1}^{p} M_j\ (d_j^k)^2 \quad \text{avec} \quad \sum_{j=1}^{p} M_j\ (d_j^k)^2 = \lambda_k.$$

Le rapport $\frac{M_j\ (d_j^k)^2}{\lambda_k}$ est alors la contribution relative de la variable X^j à l'inertie portée par le k^{ime} axe factoriel.

1.11. Choix de la Métrique M

L'analyse en composantes principales impose au départ la définition d'une métrique euclidienne M sur $\mathbb{R}^p$. Le choix de cette métrique est très important car il influe beaucoup sur les résultats de l'analyse. On a vu que les deux métriques les plus usitées sont les suivantes :
$M = I_p$: Cette métrique s'emploie lorsque les variables sont mesurées dans des unités identiques et ont des variances de même ordre.
La matrice à diagonaliser $VM = VI_p = V$ est alors la matrice de variance - covariance des variables.
$M = D_{1/\sigma^2}$: Cette matrice s'emploie lorsque les variables sont mesurées dans des unités différentes et plus généralement lorsqu'elles ont des variance notablement différentes.
L'emploie de cette métrique revient à analyser le nuage centré réduit avec la métrique I_p. Le choix de cette métrique rend les résultats de l'analyse indépendants des unités de mesure choisies pour les variables. Par ailleurs, il permet de rédurie considérablement l'effet taille de l'analyse en composantes principales sur les variables.

1.12. Les éléments illustratifs

Dans toute analyse factorielle, il est possible de projeter dans le système d'axes trouvé des individus ou des variables n'ayant pas participé en analyse. On parlera d'éléments illustratifs ou supplémentaires. Les formules de projection de ces points sur un axe factoriel sont les suivantes :
• Pour un individu illustratif X_s, l'abscisse de sa projection sur le k^{ime} axe Δu_k est ${}^tX_sMu_k = {}^tX_sb_k$.
• Pour une variable illustrative X^s, l'abscisse de sa projection sur le k^{ime} axe Δ_{C_k} est ${}^t(X^s)\ D_p\ \frac{C_k}{\sqrt{\lambda_k}}$.

Chapitre 2

DONNEES SYMBOLIQUES

2.1 Définition de Données symboliques

Nous allons traiter dans cette partie les données symboliques et les variables symboliques.

On rappelle dans un premier temps la définition même de données dites "symboliques" qui prennent en compte la variation interne aux individus et leur complexité.

Grossièrement les données "symboliques" sont aux antipodes des données "purement numériques".

Ainsi, un tableau de données symboliques autorise plusieurs valeurs par case. Ces valeurs étant parfois pondérées et liées entre elles par des règles et des taxinomies.

Les variables d'un tableau de données sont dites "symboliques" quand les données le sont.

Quatre notions fondamentales caractérisent les données symboliques
a) La variation :
Dans le tableau 1 ci- dessous, elle est caractérisée par des variables telles que Y_3 pouvant prendre un intervalle
b) Les pondérations :
Y_5 constitue une belle illustration de pondération qui permet dans ce cas de donner la fréquence d'apparition d'une maladie
c) Les règles
Par soucis de conserver une information, on peut utiliser des règles telles que "Si $Y_1 \geq 4$ alors $Y_5 =$ diabète".
d) Les taxinomies
C'est le cas où l'ensemble d'observations des individus est ordonné (total ou

partiel)

Par définition, une variable Y est une fonction $Y : \Omega \longrightarrow O$ où Ω est l'ensemble des individus et O est l'ensemble des observations. Le type de variable dépend de la structure algébrique de O.

Si O est un intervalle, nous dirons que Y est quantitative

Si O est fini ou dénombrable alors Y est qualitative

Y est qualitative ordinal ou nominale suivant que O est ordonné ou non.

Exemple

Soit $\Omega = \{w_1, w_2, w_3, w_4, w_5\}$ cinq individus sur lesquels on observe les variables classe d'âge, taille, poids, nationalité et la maladie atteinte par $w_i \;\; i = 1, ..., 5$ respectivement notées Y_1, Y_2, Y_3, Y_4, Y_5 définissent le tableau suivant

	Y_1	Y_2	Y_3	Y_4	Y_5
w_1	1	$1,65$	$[50, 58]$	Française	$\{0,7$ cancer ; $0,3$ diabète$\}$
w_2	3	$1,80$	$[70, 90]$	Sénégalaise	$\{0,9$ cancer ; $0,1$ tuberculose$\}$
w_3	2	$1,29$	$[60, 70]$	Sénégalaise	$\{0,7$ paludisme ; $0,3$ cancer$\}$
w_4	1	$1,30$	$\{40, 41, 42\}$	U.S.A	$\{0,8$ paludisme ; $0,2$ cancer$\}$
w_5	4	$1,58$	$[60, 65]$	Française	$\{0,8$ diabète ; $0,1$ cancer ; 0,1autres $\}$

Tableau 1 : Exemple de tableau de données symboliques

Y_1 : variable qualitative ordinale

Y_2 : variable nominale

Y_3 : variable quantitative

Y_4 : variable pouvant prendre des intervalles et des valeurs multiples

Y_5 : variable prenant des valeurs pondérées

2.2. Description de données symboliques

Dans un tableau de données, les descriptions des individus représentent les lignes.

Elles expriment les propriétés des individus lesquelles sont exprimées à l'aide des variables.

A titre d'exemple, dans le tableau 1

- On peut décrire l'individu w_4 en disant qu'il est de classe d'âge 1, de nationalité américaine, mesure $1,30$; de poids variant entre les valeurs $40, 41$ et 42 et a le paludisme avec 80% de possibilité et le cancer avec 20%

De même pour w_1, il est de classe d'âge 1 ; de nationalité française ; mesure $1,65$; de poids borné par 50 et 58 et a le cancer avec 70% de possibilité et diabète avec 30%.

La description d'un individu peut se faire de deux manières.

1 - A l'aide de deux opérateurs

a) Celui lié à la description des valeurs qui se trouvent dans chacune des

cases du tableau.
Il est considéré comme une disjonction ($\vee$).
b) La conjonction ($\wedge$) liant sur une même ligne les différentes colonnes.
La description notée d_4 de w_4 du tableau 1 s'écrira :

$$d_4 = [Y_1 = 1] \wedge [Y_2 = 1,30] \wedge [\{Y_3 = 40\} \vee \{Y_3 = 41\} \vee \{Y_3 = 42\}] \wedge [Y_4 = U.S.A.]$$

$$\wedge [Y_5 = \{0,8 \text{ paludisme } ; \quad 0,2 \text{ cancer}\}].$$

2 - Sous forme de produit cartésien d'ensembles

Pour illustrer cela mettons d_5 description de w_5, sous cette forme

$d_5 = D_1 \times D_2 \times D_3 \times D_4 \times D_5$

$= \{4\} \times \{158\} \times [60,65] \times \{\text{française}\} \times \{0,8 \text{ diabète } ; 0,1 \text{ cancer } ; 0,1 \text{ autres}\}$

2.3. Modèle de base

2.3.a) Remarque

Nous allons utiliser les symboles suivants et donc afin de se comprendre. On préfère donner une définition concise de ces derniers. Soit :

$-\Omega$ un ensemble dénombrable d'individus

$-D$ un ensemble de descriptions

$-Y$ une application de Ω dans D qui à chaque élément de Ω associe une description

$-T$ une application, dite de "généralisation " ou de "fusion" de $D \times D$ dans un ensemble de généralisation des descriptions noté G contenant D et permettant d'associer une description à un couple de descriptions.

$-R$ une relation de D dans G. Elle est définie par la donnée d'une partie de l'ensemble produit $D \times G$. d'une partie de l'ensemble produit $D \times G$.
Si $(x, y) \in D \times G$, on dit que x et y sont connectés par R noté xRy. On peut alors définir la fonction caractéristique de la relation R

$$\begin{aligned} H_R : D \times G &\longrightarrow L \\ (x, y) &\longmapsto H_R(x, y) = [yRx] \end{aligned}$$

L peut être $L = \{\text{vrai, faux}\}$ *ou* $[0, 1]$.

H_R est dite de " comparaison " ou d' "apperiement" car elle permet de comparer la description d'un individu à celle d'une classe.
Si $L = \{\text{vrai, faux}\}$, il y a une connection (interprétée comme une adéquation) par R entre x et y si et seulement si $[yRx] = \text{vrai}$.

Si $L = [0, 1]$, le résultat de la comparaison est imprécis alors $[yRx]$ exprime le degré de la connection de x à y par R.
Relation de comparaison R induite par un ordre sur D.

Définition 1
On dit qu'une relation de comparaison R est induite par un ordre r, si r est un ordre sur D tel que $[yRx]$ = vrai et si et seulement si yrx.

• **Exemple**
Supposons que $x \in D$ décrive le poids et la taille et compare y et x uniquement par la taille de façon que $[yRx]$ = vrai si et seulement si la taille de l'individu décrite par y est inférieur à celle décrite par x, alors R est induite par un ordre r.
L'application Y permet de décrire un "seul" individu.
Par contre la description d'une classe d'individus nécessite l'utilisation de T.
T peut être une t - norme ou une t - conorme.

Définition 2
Une t - norme est une application $T : [0, 1] \longrightarrow [0, 1]$ satisfaisant à :
$i)$ $T(u, 1) = u$
$ii)$ si $u_1 \leq u_2$ alors $T(u_1, v) \leq T(u_2, v)$
$iii)$ $T(u, v) = T(v, u)$
$iv)$ $T(u, T(v, w)) = T(T(u, v), w)$
Une t - conorme a les propriétés qu'une t - norme avec la modification suivante de la première condition

$$T(u, 0) = u.$$

• **Conséquence**
Les propriétés iii) (commutativité) et iv) (associativité) permettent de définir l'application

$$T^* : \mathcal{P}(\Omega) \longrightarrow D$$

tel que pour tout $A = \{a_1, a_2, ..., a_n\}$ de $\mathcal{P}(\Omega)$ on ait $T^*(A) = T(a_n\,, T(a_{n-1}\,, T(...(T(a_2\,, a_1)...)$.
Ainsi pour $Y(A) = \{Y(w)/w \in A\}$ on a $T^*(Y(A)) = T_{w \in A}(Y(w))$.
C'est sous cette forme que T sera appelé opérateur de génératlisation.

• **Remarque**
On peut définir une t -norme sur un espace de description en considérant que D est muni d'une relation d'ordre et possède un plus grand et un plus petit élément respectivement notés 1_D et 0_D.
Dans ce cas pour une t - norme, 1 sera remplacé par 1_D et 0 par 0_D pour une t - conorme.

• **Exemple**
Si D est l'ensemble des intervalles contenus dans $[a, b]$ et l'ordre est l'inclusion $\subset$ alors $1_D = [a, b]$ et $0_D = \varnothing$.
Dans ce cas, on peut donner comme exemple d'opérateur de généralisation l'intersection $\cap$.
On vérifie aisément que l'intersection est une t-norme.
Puisqu'un individu peut être considéré comme une classe d'individus ayant un seul élément.
On considérera dans toute la suite que D et G sont confondus en un seul ensemble que nous noterons D et appelé espace des descriptions.

2.3. b) Concepts

Les concepts entités de " second espèce " sont décrits par les variables dont les valeurs pouvant être histogrammes, des intervalles, des valeurs multiples...
Un concept se définit par son intension et son extension,
- L'intension ou " définition en compréhension " est un ensemble de propriétés caractéristiques du concept.
- L'extension est la classe des individus qui satisfont ces propriétés.

Exemple de concepts : taille, régions, poids, etc..
Contrairement aux individus (entités de première espèce), les concepts considérés comme une entité de niveau supérieur n'entre pas dans la formulation mathématique.
A sa place, viendra ce que nous appelerons objet symbolique qui est sa modélisation mathématique.

2.4. - Objets Symboliques

2.4. a) Commentaire

Les objets symboliques peuvent être vus comme des sortes d'atomes de connaissances aptes à propager des concepts donnés ou découverts d'une base de données à l'autre.
En effet, les objets symboliques fournissent un résumé explicatif de la classe d'individus qu'ils représentent et permettent de les connaître sans les conserver en mémoire.

2.4. b) Définition d'un objet symbolique

Un objet symbolique $S = (a, R, d)$ est défini par :
- Une description notée d.
- Une relation binaire R sur D permettant de comparer d à une autre description d' de D.
- Une fonction

$$a : \Omega \longrightarrow L \text{ où } L = [0, 1] \text{ ou } \{\text{vrai, faux}\}$$
$$w \longrightarrow a(w) = [Y(w)Rd]$$

a : permet d'évaluer le résultat de la comparaison (à l'aide de R) de la description $Y(w)$ d'un individu w de Ω par rapport à la description notée d'.
a : est également dite fonction de reconnaissance car elle exprime un degré d'adéquation d'un individu w à la description d.

2.4. c) Extension d'un objet symbolique

Un objet symbolique est la modélisation mathématique d'un concept. Comme les concepts, il admet une extension. Son extension se définit comme étant l'ensemble des individus satisfaisant aux propriétés du concept suivant les valeurs de L, on obtient deux types d'objets symboliques dont leur extension sont définies comme suit :
Si $L = \{\text{vrai, faux}\}$ alors l'objet symbolique S est dit booléen et a pour extension :

$$Ext(S) = \{w \in \Omega / a(w) = \text{ vrai}\}$$

Si $L = [0, 1]$ alors S est appelé objet symbolique modal l'extension se calcule à partir d'un seuil α.

$$Ext(S) = \{w \in \Omega / a(w) \geq \alpha\}.$$

Dans ce cas les données sont souvent de type plusieurs valeurs avec pondérations.

2.4. d) Objet symbolique booléen

• Assertions booléennes

Les assertions booléennes constituent un cas particulier d'objets symboliques booléens.

$$D = D_1 \times D_2 \times ... \times D_p \ , \quad Y = (Y_1, Y_2, ..., Y_p)$$

Dans le cas standard, c'est à dire où la seule donnée de R et d permet de définir la fonction de reconnaissance, une assertion booléenne est définie par $S = (a, R, d)$,
- $d = (d_1 \ , d_2 \ , ..., \ d_p)$
-$R = (R_1 \ , \ R_2 \ , \ ..., \ R_p)$ est définie par $[d' R d] = \bigwedge_{i=1,...,p} [d'_i R_i d_i]$
- $a(w) = [Y(w) \ Rd] = \bigwedge_{i=1,...,p} [d'_i R d_i]$.

Son extension est

$$\begin{aligned} Ext(S) &= \{w \in \Omega \ / \ a(w) = \text{ vrai}\} \\ &= \{w \in \Omega \ / \ a_i(w) = \text{vrai}, i = 1, ..., p\} \\ &= \{w \in \Omega \ / \ [Y_i(w) \ R_i \ d_i], i = 1, ..., p\} \end{aligned}$$

• Objet assertion simplifié

Si seulement quelques $[Y_i(w) \ R_i \ d_i]$ suffisent pour calculer l'extension, on

peut utiliser la règle syntaxique suivante :

$$a(w) = \bigwedge_{i+1,\ldots,k} [Y_i(w)\ R_i\ d_i] \bigwedge_{j=k+1,\ldots,p} [Y_j(w)\ R_j\ d_j]$$

en mettant en exergue les éléments qui permettent de calculer $a(w)$.
En supposant que ce sont les k premiers, on a

$$a(w) = \bigwedge_{i=1,\ldots,k} [Y_i(w)\ R_i\ d_i]$$

L'objet défini par $a(w) = \bigwedge_{i=1,\ldots,k} [Y_i(w)\ R_i\ d_i]$ est appelé objet simplifié.

Exemple
Exemple tiré (1) :
Supposons que Ω soit un ensemble de champignons décrits par des variables :
Y_1 : exprime la taille du pied
Y_2 : exprime la couleur du chapeau.
Un objet tel que $Y_1 = 1$ et $Y_2 =$ blanc peut s'exprimer sous la forme de l'assertion $a(w) = [Y_1 = 1] \wedge [Y_2 =$ blanc].
Supposons que Y_2 ne puisse prendre que les valeurs blanc et noir. Alors l'assertion booléenne $a(w) = [Y_1 = 1] \wedge [Y_2 = \{\text{blanc, noir}\}]$ se simplifie en $a(w) = [Y_1 = 1]$.

Les deux objets définis respectivement par $a(w) = [Y_1 = 1]$ et $a(w) = [Y_1 = 1] \wedge [Y_2 = \{\text{blanc, noir}\}]$ ont la même extension.

Cependant, il pourra se produire dans certains contextes que ce type de simplification ne soit pas possible.

Dans le tableau 1 de données symboliques. Considérons l'objet défini par :

$$a(w) = [Y_1 = [1,3]] \wedge [Y_4 = \{\text{sénégalaise, française, U.S.A}\}] \wedge [Y_4 = \text{sénégalaise}]$$

qui sera simplifié en $S(w) = [Y_1 = [2,3]] \wedge [Y_4 =$ sénégalaise].
Par contre $S(w)$ ne sera pas simplifié bien que logiquement équivalent à $b(w) = [Y_1 = 1]$ puisque $Ext(S) = Ext(b) = \{w_1, w_2\}$.

Ces deux objets ont même extension mais sémantiquement différents.

2.4. e) Objet symbolique individuel et objet symbolique de classe

• Définition

Un objet symbolique $S = (a, R, d)$ est dit individuel si d est l'image par Y d'un individu (c'est à dire si d est description individuelle) et si R est l'égalité.

Il est dit de classe si d est la description d'une classe non réduite et un seul élément (individu) de Ω.

• **Objet assertion individuel**

Un objet assertion individuel $S = (a, R, d)$ où $R = (R_1, R_2, ..., R_p)$ tel que tous les R_i sont l'égalité alors R aussi est l'égalité.

Ainsi un objet assertion booléen individuel s'écrit $S = (a, =, d)$ où $a(w) = \bigwedge_{i=1,...,p} [Y_i(w) = d_i]$.

• **Proposition**

L'ensemble des descriptions des individus est en bijection avec l'ensemble des objets symboliques individuels où R est l'égalité et ces deux ensembles sont en bijection avec l'ensemble des individus si et seulement si les individus ont tous des descriptions différentes.

2.4. f) Objets symboliques de différents types

• **Objet histogramme**

Les objets histogramme constituent un cas particulier de ce qu'on appelle les objets symboliques modaux et qui sont décrits par des variables à valeurs multiples pondérées.

Pour ces premiers les valeurs multiples sont des intervalles. Autrement, dans un tableau de données histogrammes se trouve dans chaque case des valeurs multiples de type intervalle pondérées.

La description de tels objets peut s'exprimer à l'aide d'un produit ou d'une union de produits cartésiens de fonctions.

Notons q_{jw} la fonction qui associe un poids à chaque valeur prise par la variable Y_j pour l'individu w.

La description d'un individu w peut s'écrire sous la forme :

$$Y(w) = q_{1w} \times .... \times q_{pw}.$$

L'objet symbolique individuel $S = (a, =, Y(w))$ aura pour fonction de reconnaissance $a(w) = \bigwedge_{j=1,...,p} [Y_j(w) = q_{jw}]$.

L'extension d'un objet histogramme se définit comme en 2.4. d. i).

2.5. Notion de l'Analyse de Données symboliques

Le but majeur de l'analyse de données symboliques est d'étendre l'analyse de données traditionnel aux tableaux de données symboliques pour en extraire des objets symboliques.

Toute analyse de données symboliques porte sur un ensemble d'individus qu'il faut se donner au départ.

L'analyse de données symboliques accroît encore l'importance des individus, par rapport à l'analyse de données traditionnelle. Car elle fournit un cadre où ils peuvent être représentés, puis analysés, en prenant en compte de façon plus proche de la réalité, leur variation interne et leur complexité.

On peut ensuite associer une description cohérente à une classe d'individus représentatifs d'un concept pour obtenir un " objet " dit "intensionnel".

Un objet, ainsi défini, constitue la modélisation mathématique et statistique d'un tel individu.

En munissant un objet d'un opérateur de comparaison à la description de tout individu on obtient un " objet symbolique ".

Chapitre 3

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].

Bibliographie

[1] Edwin Diday (1992) ; Analyse des données et classification automatique numérique et symbolique.

[2] P. Cazes, A. Chouakria (I), E. Diday, Y. Schekuman ; Entension de l'Analyse en composantes principales à des données de type intervalle.

[3] Chouakria A., Diday E. , Cazes P. Extension of principal components Analysis to interval data NTTS 95 : New Techniques and Technologies for Statistics, Bonn, novembre 1995.

[4] Chouakria A., Cazes P. , Diday E. Extension de l'analyse factorielle des correspondances multiples à des données de type intervalle et de type ensemble.
SFC 96 : Actes de le 3ème rencontre de la Société francophone de classification. Namur, septembre 1995.

[5] Chouakria A. , Verde R., Diday E. , Cazes P. Généralisation de l'analyse factorielle des correspondances multiples à des objets symboliques. SFC 96 : Actes de la 4ème rencontre de la Société francophone de classification, Vannes, septembre 1996.

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

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

[8] Ichino M. Generalized Minkowskky metrics for mixed feature type data analysis. IEEE, transactions on systems, man and cyberneticss, vol. 24, n°4, 1994.

[9] Nagabushan P. An efficient method for classifying remotely sensed data, incorporating dimensionnality reduction. Ph.d Thesis, Mysore University, India, 1988.