Chapitre 6

Analyse factorielle en composantes

I - *INTRODUCTION*

Cet ouvrage est consacré à la condensation de l'information sur une structure finie ; toutefois il est important de reconnaître que les méthodes d'analyse factorielle sont particulièrement utilisées en Analyse des Données. Ces méthodes proposent une représentation euclidienne de condensation qui n'est par conséquent pas finie. Le but de ce chapitre est d'abord de présenter les deux principales méthodes d'analyse factorielle : l'Analyse en Composantes Principales (Hotelling, 1933) et l'Analyse des Correspondances (J.P. Benzecri, 1963). Il s'agira ensuite de situer l'approche factorielle par rapport à l'approche classificatoire, laquelle, domine notre travail.

Pour présenter l'analyse factorielle au paragraphe II, nous profiterons de l'expérience d'un cours que nous avons assuré à l'Université de Rennes I, que nous résumerons en cherchant à préserver son caractère pédagogique. Etant donné un nuage fini (i.e. ensemble fini de points affectés de masses finies) dans un espace euclidien (espace vectoriel muni d'une forme quadratique définie positive donnant lieu à une distance), le problème de l'analyse factorielle se ramène à la recherche d'un sous-espace affin S de dimension donnée qui rend minimum la somme pondérée des carrés des distances à S des différents sommets du nuage. La représentation consiste alors en la projection du nuage, au sens de la métrique définie par la forme quadratique, sur le sous-espace S trouvé. La nature mathématique du problème est donc toujours la même ; ce qui est en jeu c'est d'une part l'espace de représentation et de condensation de cette dernière ; d'autre part, la nature du critère qui est ici directement lié à la représentation, puisqu'il s'agit d'un moment d'inertie.

II - *LES DEUX PRINCIPALES METHODES D'ANALYSE FACTORIELLE*

1. REDUCTION D'UN ENDOMORPHISME SYMETRIQUE

Un *espace euclidien* est un espace vectoriel de dimension finie muni d'une forme quadratique définie positive. Nous noterons $\mathcal{E}$ un tel espace et q la forme quadratique dont il est muni. A q correspond un isomorphisme de $\mathcal{E}$ sur son dual $\mathcal{E}^* = \mathcal{L}(\mathcal{E}, R)$: ensemble des formes linéaires sur $\mathcal{E}$; cet

isomorphisme noté également q est défini de la façon suivante : pour $x \in \mathfrak{E}$, $q(x)$ est la forme linéaire de $\mathfrak{E}^*$ qui associe à tout y de $\mathfrak{E}$, le nombre réel $q(x,y)$. L'isomorphisme réciproque q^{-1} permet de "transporter" *dans* $\mathfrak{E}$ les éléments et sous espaces de $\mathfrak{E}^*$. La notion d'orthogonalité initialement définie entre un élément de $\mathfrak{E}$ et un élément de $\mathfrak{E}^*$ ($x \in \mathfrak{E}$ et $y^* \in \mathfrak{E}^*$ sont orthogonaux si l'on a $y^*(x) = 0$), peut être définie dans $\mathfrak{E}$, pour q ; ainsi, deux vecteurs x et y de $\mathfrak{E}$ sont orthogonaux pour q si x et $q(y)$ le sont ; c'est-à-dire si $q(x,y) = 0$, relation symétrique en x et y.

Si $\mathfrak{E}$ est un espace euclidien pour la forme quadratique q, dont la dimension est notée m, on démontre qu'il existe dans $\mathfrak{E}$ une base $\{e_j / 1 \leqslant j \leqslant m\}$ telle que l'on ait $q(e_j, e_k) = o$ pour $j \neq k$ et $q(e_j, e_j) = 1$ pour $1 \leqslant j \leqslant m$. Une telle base est dite q-orthonormale.

Si $\mathfrak{E}$ est un espace vectoriel de dimension finie, rappelons l'existence d'un isomorphisme canonique entre $\mathfrak{E}$ et $\mathfrak{E}^{**}$ défini de la façon suivante : à $x \in \mathfrak{E}$ on fait correspondre $\tilde{x} \in \mathfrak{E}^{**}$ défini tel que

$$(\forall y^* \in \mathfrak{E}^*), \tilde{x}(y^*) = y^*(x).$$

Cet isomorphisme permet l'identifiaction de $\mathfrak{E}$ et de $\mathfrak{E}^{**}$

Soit $\mathfrak{F}$ un second espace vectoriel de dimension finie ; rappelons également pour toute application linéaire ℓ de $\mathfrak{E}$ dans $\mathfrak{F}$: $\ell \in \mathfrak{L}(\mathfrak{E}, \mathfrak{F})$, l'application de $\mathfrak{F}^*$ dans $\mathfrak{E}^*$: $z^* \to z^* \circ \ell$, est une application linéaire que l'on appelle transposée de ℓ et que l'on note ${}^t\ell$; ${}^t\ell \in \mathfrak{L}(\mathfrak{F}^*, \mathfrak{E}^*)$. On a $(\forall x \in \mathfrak{E})$, ${}^t\ell(z^*)(x) = z^*[\ell(x)]$.

Soit maintenant $\mathfrak{E}$ un espace euclidien pour une forme quadratique q et ℓ un endomorphisme de $\mathfrak{E}$: $\ell \in \mathfrak{L}(\mathfrak{E}, \mathfrak{E})$; le transport au moyen de q de la structure de $\mathfrak{E}^*$ dans $\mathfrak{E}$ permet de regarder ℓ comme une application linéaire de $\mathfrak{E}$ dans $\mathfrak{E}^*$: pour tout x de $\mathfrak{E}$, $\ell(x)$ sera la forme linéaire sur $\mathfrak{E}$ qui à tout $y \in \mathfrak{E}$, associe le nombre $q(\ell(x), y)$. ${}^t\ell$ est alors également une application de $\mathfrak{E} = \mathfrak{E}^{**}$ dans $\mathfrak{E}^*$ et on peut parler de symétrie pour ℓ : ${}^t\ell = \ell$, qui s'exprime par la relation

$$(\forall (x,y) \in \mathfrak{E} \times \mathfrak{E}),\ q(\ell(x), y) = q(x, \ell(y))$$

ℓ étant un endomorphisme d'un espace vectoriel $\mathfrak{E}$, le sous espace propre de ℓ relatif à la valeur propre λ est défini par

$$\mathfrak{E}(\lambda) = \{x / \ell(x) = \lambda x\}$$

λ est une des racines de l'équation dite caractéristique

$$\det(\ell - \xi 1) = 0,$$

où 1 désigne l'application identique ; cette équation fournissant l'ensemble de toutes les valeurs propres de ℓ.

L'endomorphisme ℓ est dit *trigonalisable* s'il existe une base $\{f_j / 1 \leqslant j \leqslant m\}$ par rapport à laquelle la matrice de l'endomorphisme est tri-

diagonale ; en d'autres termes : $\ell(f_1)=\alpha_{11}f_1, \ell(f_2)=\alpha_{21}f_1+\alpha_{22}f_2, \ldots, \ell(f_j)=\alpha_{j1}f_1+\alpha_{j2}f_2+\ldots+\alpha_{jj}f_j, \ldots, \ell(f_m)=\alpha_{m1}f_1+\alpha_{m2}f_2+\ldots+\alpha_{mm}f_m$.
ℓ est *diagonalisable* s'il existe une base $\{f_j/1\leqslant j\leqslant m\}$ par rapport à laquelle la matrice de l'endomorphisme est diagonale ; en d'autres termes : $\ell(f_j)=\beta_j f_j$ pour tout j, $1\leqslant j\leqslant m$.

On démontre que l'endomorphisme ℓ est trigonalisable si et seulement si toutes ses valeurs propres sont *réelles*, il est diagonalisable si *de plus*, pour toute valeur propre λ, la dimension du sous espace propre $\&(\lambda)$ est égale à la multiplicité de λ comme racine de l'équation caractéristique. (cf.[4] § Th. 5 et 6)

Si $\&$ est un espace euclidien pour q, le caractère symétrique d'un endomorphisme ℓ de $\&$, permet précisément de montrer que les valeurs propres de ℓ sont réelles et que, pour toute valeur propre λ, la dimension du sous espace propre $\&(\lambda)$ est égale à la multiplicité de λ comme racine de l'équation caractéristique. Par conséquent ℓ *est diagonalisable*. Soit $\{\lambda_1,\lambda_2,\ldots,\lambda_h\}$ l'ensemble des valeurs propres de ℓ, y compris, éventuellement, la valeur propre 0 ; et soient $r_1,r_2,\ldots,r_h$ les ordres de multiplicité respectifs de $\lambda_1,\lambda_2,\ldots,\lambda_h$ comme racines de l'équation caractéristique. On a, $\dim \&(\lambda_i) = r_i$ et $r_1+r_2+\ldots+r_h=m$.

On peut par conséquent construire dans chaque $\&(\lambda_i)$ une base q-orthonormale de r_i vecteurs propres associés à la valeur propre λ_i ; deux vecteurs de deux sous espaces propres distincts étant q-orthogonaux, on obtient ainsi une base $\{f_j/1\leqslant j\leqslant m\}$ q-orthonormale formée de vecteurs propres.

Dans la mesure où ℓ dérive d'une forme quadratique Q sur $\&x\&$:

$$(\forall(x,y)\in \&x\&),\ q(\ell(x),y)=q(x,\ell(y)=Q(x,y)\ ;$$

la base $\{f_j/1\leqslant j\leqslant m\}$ formée de vecteurs propres de ℓ est *à la fois* q *orthonormale et* Q*-orthogonale* ; de plus, si f est un des vecteurs de la base associé à la valeur propre λ, on a

$$Q(f,f) = \lambda q(f,f) = \lambda\ ;$$

par conséquent, toutes les valeurs propres de ℓ sont positives ou nulles si la forme quadratique Q est positive.

A la décomposition de $\&$ en somme directe de sous espaces propres

$$\& = \&(\lambda_1) \oplus \&(\lambda_2) \oplus \ldots \oplus \&(\lambda_h) \qquad (1)$$

on peut associer, au moyen de l'isomorphisme q, une décomposition duale de $\&^*$:

$$\&^* = \&^*(\lambda_1) \oplus \&^*(\lambda_2) \oplus \ldots \oplus \&^*(\lambda_h) \qquad (1^*)$$

où on a noté $\&^*(\lambda_i) = q(\&(\lambda_i))$ pour $1\leqslant i\leqslant h$; $\&^*(\lambda_i)$ peut d'ailleurs être défini comme suit

$$\&^*(\lambda_i) = \{u\in \&^*/u(x) = 0 \text{ pour tout } x\in \&(\lambda_j) \text{ et } j\neq i\}.$$

ℓ étant un endomorphisme de $\&$, ${}^t\ell$ est un endomorphisme de $\&^*$. On peut ai-

sément montrer que la ci-dessus décomposition de $\mathcal{E}^*$ est celle en sous espace propres relativement à ${}^t\ell$ dont l'ensemble des valeurs propres est le même que celui de ℓ. Réciproquement si $u \in \mathcal{E}^*(\lambda_i)$: sous espace propre de ${}^t\ell$ relatif à la valeur propre λ_i ; le vecteur x de $\mathcal{E}$ tel que $u=q(x)$ est vecteur propre de ℓ relatif à la valeur propre λ_i. notons que si $x \in \mathcal{E}(\lambda_i)$ est de norme unité ($q(x,x)=1$) ; $q(x) \in \mathcal{E}^*(\lambda_i)$ est tel que

$$ {}^t\ell\,[q(x)](x) = q(x,\ell(x)) = \lambda $$

Remarquons enfin que si ℓ dérive de la forme quadratique Q, on a

$$ \ell = q^{-1} \circ Q \quad \text{et} \quad {}^t\ell = Q \circ q^{-1} $$

où Q désigne l'application de $\mathcal{E}$ dans $\mathcal{E}^*$: $x \to Q(x,.)$.

2. CONDENSATION DE LA DIMENSION DE L'ESPACE DE REPRESENTATION D'UN NUAGE

2.1. Caractéristiques liées à un nuage de points dans un espace euclidien

Soit $\mathcal{E}$ un espace vectoriel de dimension finie m sur $\mathbb{R}$. On appelle "nuage" un ensemble fini de points de $\mathcal{E}$ dont chacun est affecté d'une masse ponctuelle finie, positive. Le nuage peut être représenté par

$$ \mathcal{N} = \{(M_i,\mu_i)/i \in I\} \qquad (1) $$

où $I = \{1,2,\ldots,n\}$ est l'ensemble fini d'indexation de l'ensemble des points du nuage. La *masse totale* du nuage est

$$ \mu = \sum_{i \in I} \mu_i \qquad (2) $$

Le *centre de gravité* G est le point de $\mathcal{E}$ défini par l'équation ponctuelle

$$ G = \frac{1}{\mu} \sum_{i \in I} \mu_i M_i \qquad (3) $$

qui est le reflet de l'équation vectorielle.

$\vec{OG} = \frac{1}{\mu} \sum_{i \in I} \mu_i \vec{OM_i}$ qui s'écrit également $(G-O) = \frac{1}{\mu} \sum_{i \in I} \mu_i (M_i - O)$.

où O désigne l'origine.

Le support affin A du nuage est la plus petite variété linéaire contenant tous les points M_i. On peut aisément voir que A qui ne contient pas en général l'origine contient nécessairement le centre de gravité G. A se déduit par translation du sous espace vectoriel S engendré par l'ensemble des vecteurs

$$ \{(M_i-G)/i \in I\} \qquad (4) $$

Supposons à partir de maintenant que $\mathcal{E}$ est un espace euclidien pour une forme quadratique q. Soient s et t deux vecteurs unitaires ($q(s,s)=q(t,t)=1$) et soient H et K les deux hyperplans, passant par G, d'équations respectives.

$$ q(s,M-G)=0 \quad \text{et} \quad q(t,M-G)=0, \qquad (5) $$

qui sont respectivement q-orthogonaux à s et à t.

La distance (pour q) d'un point M_i à H est $q(s,M_i-G)$. *Le moment d'inertie du nuage par rapport à* H est défini par

$$\sum_{i\in I} \mu_i\{q(s,M_i-G)\}^2 \qquad (6)$$

Signalons, en vertu d'un théorème de Huygens, que si H' est un sous espace affin parallèle à H, le moment d'inertie du nuage par rapport à H' est égal à la somme du moment d'inertie par rapport à H et du produit par la masse totale μ du carré de la distance de H à H'.

Le produit d'inertie du nuage par rapport aux deux hyperplans H *et* K est donné par la formule

$$\sigma(u,v)=\sum_{i\in I} \mu_i q(s,M_i-G)q(t,M_i-G) \qquad (7)$$

où on a noté $u = q(s)$ et $v = q(t)$.
On peut, comme ci-dessus signaler, que si H' et K' sont deux sous espaces affins respectivement parallèles à H et K ; le produit d'inertie par rapport à H' et à K' est égal à la somme du produit d'inertie par rapport à H et K et du produit par la masse totale μ, du produit des distances de H à H' et de K à K'.

Dans la formule (7), lorsque (s,t) décrit &x&, (q(s),q(t)) décrit $\&^*x\&^*$, puisque q est définie. La forme bilinéaire symétrique et possible sur $\&^*x\&^*$ définie par

$$(\forall(u,v) \in \&^*x\&^*),\sigma(u,v) = \sum_{i\in I} \mu_i u(M_i-G)v(M_i-G) \qquad (8)$$

est appelée *forme quadratique d'inertie du nuage.*

L'application σ de $\&^*$ dans & est définie par

$$u \to \sigma(u) = \sum_{i\in I} \mu_i u(M_i-G)(M_i-G) \qquad (9)$$

L'image par σ de $\&^*$ peut se mettre sous la forme

$$\sigma(\&^*) = \{x \in \&/x = \sum_{i\in I} \mu_i q(s,M_i-G)(M_i-G), \text{ pour } s \in \&\} \qquad (10)$$

On démontre précisément que $\sigma(\&^*)$ est le sous espace vectoriel S engendré par (4).

2.2. Axes factoriels et facteurs

q réalise un isomorphisme d'espaces vectoriels entre & et $\&^*$ et $\sigma \in \mathcal{L}(\&^*,\&)$; de sorte que $\sigma \circ q$ est un endomorphisme de & défini par

$$x \to \sigma \circ q(x) = \sum_{i\in I} \mu_i q(x,M_i-G)(M_i-G), \qquad (11)$$

cet endomorphisme est *symétrique* et dérive de la forme quadratique sur &x&, *positive* mais non définie :

$$Q(x,y)=q[\sigma \circ q(x),y] = \sum_{i\in I} \mu_i q(x,M_i-G)q(y,M_i-G), \qquad (12)$$

Considérée comme une application de $\mathcal{E}$ dans $\mathcal{E}^*$, Q se met sous la forme

$$Q = q \circ \sigma \circ q \tag{13}$$

On se trouve par conséquent dans la situation du paragraphe 2.1 précédent où le rôle joué par ℓ est ici joué par $\sigma \circ q$;

$$\ell = \sigma \circ q \text{ et } {}^t\ell = q \circ \sigma \tag{14}$$

Par conséquent on peut associer à $\sigma \circ q$ et à $q \circ \sigma$ les deux décompositions duales (1) et (1^*) du paragraphe précédent. Q étant positive, les valeurs propres réelles $\lambda_1, \lambda_2, \ldots, \lambda_h$ sont ici positives ou nulles. Dans le cas, qui est celui des applications, où on s'intéresse à la restriction σ' de σ à S^* ; $\sigma' \in \mathcal{L}(S^*, S)$ et $q \in \mathcal{L}(S, S^*)$ réalisent des isomorphismes d'espaces vectoriels ; de sorte que les valeurs propres de $\sigma' \circ q$ sont toutes strictement positives.

On appelle *axe factoriel* relatif au moment principal $\lambda (\lambda > 0)$, un vecteur propre associé à la valeur propre λ ; il s'agit par conséquent d'un vecteur s appartenant à $\mathcal{E}(\lambda)$; c'est-à-dire, tel que

$$\sigma \circ q(s) = \lambda s, \tag{15}$$

On appelle *facteur* relatif à λ, une forme linéaire u appartenant à $\mathcal{E}^*(\lambda)$; c'est-à-dire, telle que

$$q \circ \sigma (u) = \lambda u, \tag{16}$$

Dans les applications on ne rencontre pour ainsi dire jamais de valeurs propres multiples de sorte que ces définitions conduisent à une détermination unique, à une homothétie près.

Si s est un axe factoriel unitaire : $q(s,s) = 1$, $u = q(s)$ est un facteur de norme $\sqrt{\lambda}$ pour σ : $\sigma(u,u) = \lambda$.

Réciproquement si u est facteur de norme 1 pour σ ; $\sigma(u,u)=1$, $s=\sigma(u)$ est un axe factoriel de norme $\sqrt{\lambda}$ pour q : $q(s,s)=\lambda$

Deux axes factoriels relatifs à deux moments distincts sont q-orthogonaux ; deux facteurs relatifs à deux moments distincts sont σ-orthogonaux. Dans chaque $\mathcal{E}(\lambda_h)$ (resp. $\mathcal{E}^*(\lambda_h)$), $\lambda_h \neq 0$, on peut choisir une base d'axes factoriels (resp. de facteurs) q-orthogonaux (resp. σ-orthogonaux) de telle sorte que les deux bases de $\mathcal{E}(\lambda_h)$ et de $\mathcal{E}^*(\lambda_h)$ se correspondent.

Un axe factoriel relatif à $\lambda > 0$ appartient au sous-espace S ; un facteur relatif à $\lambda > 0$ est déterminé par sa restriction à S, un tel facteur s'annulle sur le sous espace orthogonal (pour q) à S.

2.3. Caractère optimal de la représentation

Soit R un sous espace vectoriel de l'espace euclidien S engendré par l'ensemble des vecteurs $\{(M_i - G)/i \in I\}$. En notant $d(M_i, R)$ la q-distance d'un point M_i du nuage à R (i.e. $\min\{[q(M_i - N, M_i - N)]^{1/2} / N \in R\}$), le moment d'inertie du nuage par rapport à R s'écrit

$$\mathcal{M}(R) = \sum_{i \in I} \mu_i [d(M_i, R)]^2, \tag{17}$$

Le problème de l'*ajustement linéaire des moindres carrés* consiste, pour un entier k quelconque strictement inférieur à la dimension p de S, à trouver dans la classe des sous espaces affins de dimension k, un sous espace R par rapport auquel le moment d'inertie soit minimum. L'objet de ce paragraphe est de montrer que la recherche des facteurs ou des axes factoriels donne précisément toutes les solutions du problème de l'ajustement linéaire. Commençons par remarquer qu'en vertu de la relation de Huygens, on peut se restreindre à la classe des sous espaces affins passant par le centre de gravité G ; c'est-à-dire aux sous espaces vectoriels de S dont l'origine est le point G.

Donnons alors l'expression explicite du moment d'inertie par rapport à un sous espace R de S, de dimension k. Soit $\{e_j/1\leq j\leq p\}$ une base q-orthonormale de S, telle que $\{e_j/1\leq j\leq k\}$ soit une base de R. Désignons par T le sous espace supplémentaire q-orthogonal à R, dont une base est nécessairement $\{e_j/k<j\leq p\}$. Les projections q-orthogonales sur R et sur T d'un vecteur x de S sont respectivement

$$x_R = \sum_{1\leq j\leq k} q(x,e_j)e_j,\ x_T = \sum_{k<j\leq p} q(x,e_j)e_j. \qquad (18)$$

Il en résulte que les moments d'inertie du nuage par rapport aux sous espaces R et T sont respectivement

$$\mathcal{M}(R) = \sum_{i\in I} \mu_i q(x_{i_T},x_{i_T}),\ \mathcal{M}(T) = \sum_{i\in I} \mu_i q(x_{i_R}, x_{i_R})\ ; \qquad (19)$$

où on a noté x_i le vecteur (M_i-G) et où, rappelons le, G est à l'origine (G=0).

En décomposant par rapport à la base $\{e_j/1\leq j\leq k\}$ et en inversant les signes sommes, on obtient les formules

$$\mathcal{M}(R) = \sum_{k<j\leq p} \sigma(q(e_j),q(e_j)),\ \mathcal{M}(T) = \sum_{1\leq j\leq k} \sigma(q(e_j),q(e_j)), \qquad (20)\ ;$$

et on a bien entendu

$$\mathcal{M}(R) + \mathcal{M}(T) = \sum_{i\in I} \mu_i q(M_i-G,M_i-G) = \sum_{1\leq j\leq k} \sigma(q(e_j),q(e_j)), \qquad (21)\ .$$

La dernière expression est la *trace* (i.e somme des éléments diagonaux) de la *matrice d'inertie* de terme général $(\sigma(q(e_j),q(e_h)))$, par rapport à la base choisie. Si cette base est formée de vecteurs propres de l'endomorphisme $\sigma \circ q$; cette expression apparait comme *la somme pondérée des valeurs propres, chacune comptée avec son ordre de multiplicité.*

Soit $\lambda(1) > \lambda(2)>\ldots>\lambda(h)>0$, la suite décroissante des valeurs propres de l'endomorphisme $\sigma \circ q$ de S. Désignons par r(c) l'ordre de multiplicité de la valeur propre $\lambda(c)$; il s'agit de la dimension du sous espace propre S(c) relatif à la valeur propre $\lambda(c)$. Si k est un entier inférieur à p=dim(S), posons s(k) l'entier, indice supérieur de la somme, inférieure ou égale à k, de la plus longue section commençante de la suite $(r(1), r(2),\ldots,r(c),\ldots,r(h))$; en d'autres termes, s(k) se trouve défini par la double inégalité

$$\sum_{1\leqslant c\leqslant s(k)} r(c) \leqslant k < \sum_{1\leqslant c\leqslant s(k)+1} r(c) \qquad (22)$$

Avec ces notations on démontre le théorème suivant (cf. [1])

THEOREME. *Parmi les sous espaces* R *de dimension* k *; ceux qui réalisent le minimum de* $\mathcal{M}_G(R)$, *sont caractérisés par la double inclusion*

$$S(1) \oplus \ldots \oplus S(s(k)) \subseteq R \subset S(1) \oplus \ldots \oplus S(s(k)+1) \qquad (23)$$

Ce théorème établit le caractère optimal de la représentation du nuage par sa projection q-orthogonale sur un sous espace R satisfaisant la condition (23). Cette représentation fournit la meilleure approximation de dimension k du nuage au sens suivant : pour tout ensemble de points $\{N_i/i\in I\}$ qui soutend une sous variété linéaire affine de dimension k de S, on a

$$\sum_{i\in I} \mu_i\{d(M_i(k),M_i)\}^2 \leqslant \sum_{i\in I} \mu_i\{d(N_i,M_i)\}^2 \qquad (24)$$

où $M_i(k)$ désigne la projection q-orthogonale de M_i sur R et où la distance correspond à la métrique q. En effet, le premier membre de (24) n'est autre que $\mathcal{M}_G(R) = \sum_{k<r\leqslant p} \lambda(r)$.

Comme nous l'avons déjà signalé, dans les applications numériques, l'existence de valeurs propres multiples est tout à fait exceptionnelle ; nous allons par conséquent préciser la représentation du nuage dans le cas, que nous retiendrons désormais, où toutes les valeurs propres sont simples.

Nous avons vu qu'il était équivalent de rechercher les axes factoriels ou les facteurs ; pour des raisons de simplicité technique qui apparaîtront ci-dessous et qui sont liées au fait que dans les cas pratiques la matrice de q est diagonale par rapport à la base canonique, on procède d'abord à la recherche des facteurs $u^1,u^2,\ldots,u^k,\ldots,u^p$; dans l'ordre décroissant des valeurs propres $\lambda(1),\lambda(2),\ldots,\lambda(p)$:

$$q \circ \sigma\,(u^r) = \lambda(r)u^r, 1\leqslant r\leqslant p \qquad (25)$$

qu'on normalise au moyen de la condition

$$\sigma(u^r,u^r) = \sum_{i\in I} \mu_i\{u^r(M_i-G)\}^2 = 1 \qquad (26)$$

Le vecteur axial factoriel de norme 1 pour q, associé à u^r, est

$$a_r = \sigma(u^r)/\sqrt{\lambda(r)} \qquad (27)$$

En rapportant S au système des axes factoriels $\{a_r/1\leqslant r\leqslant p\}$; on a pour le point M_i la décomposition suivante

$$M_i = G + \sum_{1\leqslant r\leqslant p} q(a_r,M_i-G)a_r \qquad (28)$$

formule qu'on peut écrire

$$M_i = G + \sum_{1\leqslant r\leqslant p} u^r(M_i-G)\sigma(u^r) \qquad (29)$$

La représentation condensée du nuage sur l'espace R des k premiers axes factoriels est définie par la relation très importante

$$(\forall i \in I),\ M_i(k) = G + \sum_{1 \leq r \leq k} u^r (M_i - G)\sigma(u^r) \qquad (30)$$

où $M_i(k)$ est la projection q-orthogonale de M_i sur R.

Lorsqu'on détermine la suite $u^1, u^2, \ldots$; la représentation (30) permet, pour k de plus en plus grand, de préciser de plus en plus la structure métrique qui se trouve d'une certaine façon extraite, d'où le nom d'"extraction de facteurs".

2.4. Application aux tableaux de données.

On s'intéressera seulement aux tableaux rectangulaires de description. Soit T un tel tableau dont on suppose ici que l'ensemble des lignes est indexé par l'ensemble I des objets et que celui des colonnes par un ensemble V de variables numériques. On note n=card(I) et m=card(V). Une même ligne i représente la suite des valeurs des différentes variables sur l'objet i ; une même colonne j représente la suite, sur les différents objets, des valeurs de la j-ème variable de V. A l'intersection de la ligne i et de la colonne j se trouve le nombre ξ_{ij} : mesure de la j-ème variable sur le i-ème objet.

Dans ces conditions, l'espace & de représentation est identifié à l'espace géométrique R^m qu'on rapporte à la base canonique $\{e_j / 1 \leq j \leq m\}$ où e_j est le vecteur dont toutes les composantes sont nulles sauf la j-ème qui vaut 1. Si $\{e^*_j / 1 \leq j \leq m\}$ désigne la base duale de cette dernière : $e^*_j(e_h) = 0$ (resp. 1) si $j \neq h$ (resp. $j = h$) ; e^*_j qu'on appelle la j-ème forme coordonnée, représente la j-ème variable v^j de V. L'objet i de I sera représenté par le point de R^m de coordonnées $(\xi_{i1}, \ldots, \xi_{ij}, \ldots, \xi_{im})$; c'est-à-dire, par le point M_i défini par

$$M_i - O = \sum_{1 \leq j \leq m} \xi_{ij} e_j \qquad (31)$$

lequel est affecté de la masse μ_i qu'on suppose donnée.

Le centre de gravité G du nuage est le point de coordonnées $(g_1, \ldots, g_j, \ldots, g_m)$ où

$$g_j = \sum_{i \in I} \frac{\mu_i}{\mu} \xi_{ij}, \qquad (32)$$

où μ est la masse totale.

La métrique q est supposée fournie par un tableau carrée m×m, symétrique (q_{jh}), $1 \leq j \leq m$ et $1 \leq h \leq m$; où

$$q_{jh} = q(e_j, e_h). \qquad (33)$$

Dans ces conditions, la q-distance entre deux points M_i et $M_{i'}$ ($i \in I$ et $i' \in I$) est donnée par

$$(d(M_i, M_{i'}))^2 = \sum_{(j,h)} q_{jh} (\xi_{ij} - \xi_{i'j})(\xi_{ih} - \xi_{i'h}) \qquad (34)$$

La forme quadratique d'inertie σ est supposée précisée par sa matrice par rapport à la base canonique $\{e^*_j / 1 \leq j \leq m\}$; le terme général de cette matrice symétrique est

$$\sigma_{jh} = \sigma(e^*_j, e^*_h) = \sum_{i \in I} \mu_i (\xi_{ij} - g_j)(\xi_{ih} - g_h), \qquad (35)$$

qu'on peut mettre sous la forme

$$\sigma_{jh} = \sum_{i \in I} \mu_i \xi_{ij} \xi_{ih} - \mu g_j g_h$$

en vertu des propriétés du centre de gravité.

Le facteur u, vecteur propre de l'endomorphisme $q \circ \sigma$ de $\mathfrak{E}^*$, relatif à la valeur propre λ, sera déterminé par la suite de ses valeurs sur la base canonique ; en d'autres termes par les équations

$$q \circ \sigma(u)(e_j) = \lambda u(e_j) \; ; \; 1 \leq j \leq m \qquad (36)$$

soit, en détaillant

$$\sum_{i \in I} \mu_i u(M_i - G) q(M_i - G, e_j) = \lambda u(e_j) \; ; \; 1 \leq j \leq m \qquad (37)$$

En exprimant $(M_i - G)$ par rapport à la base canonique et en notant $u_j = u(e_j)$, la formule (37) peut être amenée à la forme suivante

$$\sum_{1 \leq h \leq m} \; \sum_{1 \leq k \leq m} \sigma_{hk} q_{jk} u_h = \lambda u_j \; ; \; 1 \leq j \leq m \qquad (38)$$

Pratiquement (Analyse en composantes principales et des correspondances), la matrice (q_{jk}) est diagonale ; de sorte que la relation (38) se simplifie et devient

$$\sum_{1 \leq h \leq m} \sigma_{hj} q_{jj} u_h = \lambda u_j, \; 1 \leq j \leq m, \qquad (39)$$

Sur la forme linéaire w qui se déduit de u par

$$w(e_j) = w_j = u_j / \sqrt{q_{jj}} \text{ pour } 1 \leq j \leq m, \qquad (40)$$

l'équation (39) se traduit par

$$\sum_{1 \leq h \leq m} \sigma_{hj} (q_{jj} q_{hh})^{1/2} w_h = \lambda w_j, \; 1 \leq j \leq m, \qquad (41)$$

On se ramène ainsi, pour la recherche des facteurs, à la diagonalisation de la matrice symétrique m×m *de terme général*

$$\alpha_{hj} = \sigma_{hj} (q_{jj} q_{hh})^{1/2} \; ;$$

le vecteur $(u_1, \ldots, u_j, \ldots, u_m)$ définit la suite des composantes du facteur relatif à la valeur propre λ :

$$u = \sum_{1 \leq j \leq m} u_j e^*_j \qquad (42)$$

u apparait ainsi comme une variable de synthèse obtenue à partir des di-

verses variables v^j de V par la combinaison linéaire de coefficients les u_j.

A chacun des facteurs de la suite $(u^r/1\leq r\leq k)$ des k premiers facteurs on impose la condition de normalisation (26) qui se met ici sous la forme

$$\sigma(u^r,u^r) = \sum_{(j,h)} \sigma_{jh}u^r{}_j u^r{}_h = 1, \text{ pour } 1\leq r\leq k, \qquad (43)$$

rappelons que les différents facteurs sont σ-orthogonaux comme relatifs à des valeurs propres distinctes.

L'axe factoriel σ(u), où u est le facteur relatif à la valeur propre λ, se met sous la forme

$$\sigma(u) = \sum_{i\in I} \mu_i u(M_i-G)(M_i-G) = \sum_{(j,h)} \sigma_{jh}u_h e_j, \qquad (44)$$

σ(u) est de norme $\sqrt{\lambda}$ pour q.

La formule de représentation condensée du nuage devient

$$(\forall i\in I),\ M_i(k) = G + \sum_{1\leq r\leq k} \eta_{ir}a_r, \qquad (45)$$

où $\{a_r/1\leq r\leq k\}$ qui représente $\{\sigma(u^r)/\sqrt{\lambda(r)}/1\leq r\leq k\}$ est un système q-orthonormal de vecteurs et où

$$(\forall r,\ 1\leq r\leq k),\ \eta_{ir} = \sqrt{\lambda(r)}u^r(M_i-G) = \sqrt{\lambda(r)} \sum_{1\leq j\leq m} u^r{}_j(\xi_{ij}-g_j), \qquad (46)$$

Le nuage sera dans ces conditions représenté dans $\mathbb{R}^k$ dont l'origine définit le centre de gravité G et la base canonique, $\{a_r/1\leq r\leq k\}$; de la sorte la métrique q est représentée par la métrique euclidienne ordinaire de l'espace géométrique $\mathbb{R}^k$. Dans ce système, les coordonnées du point représentant le point M_i, sont $(\eta_{i1},\ldots,\eta_{ir},\ldots,\eta_{ik})$ où η_{ir}, $1\leq r\leq k$, est défini dans la formule (46).

3. ANALYSE EN COMPOSANTES PRINCIPALES NORMEE

3.1. Introduction ; définition de la métrique.

Le problème posé est l'étude du comportement d'une population dont on dispose d'un échantillon formant un ensemble fini I, vis à vis d'un ensemble fini V de variables numériques, par exemple, pour fixer les idées, il peut s'agir de la manière qu'a une population d'une région donnée de répartir ses dépenses vis à vis de divers besoins : un élément de I peut alors être un ménage et un élément de V, un type de dépense.

Les notations sont les mêmes que celles du paragraphe 2.4 précédent. L'ensemble des lignes du tableau T des données est indexé par I de cardinal n et celui des colonnes par V de cardinal m ; ξ_{ij} est la mesure de la j-ème variable v^j sur le i-ème élément i de I qui, dans l'exemple représente la dépense du ménage i pour le besoin j.

Comme ci-dessus (§2.4) à I, qu'on notera aussi $\{1,2,\ldots,i,\ldots,n\}$, on fait correspondre un nuage de points de $\mathbb{R}^m$ dont chaque sommet portera ici

la même masse ponctuelle 1/n. En munissant $\mathbb{R}^m$ d'une métrique adéquate et en condensant au mieux la dimension de l'espace de représentation, on espère que les facteurs définiront des variables de synthèse non directement observables qui nous font "comprendre" les tendances dominantes du comportement de la population étudiée en disposant les diverses variables introduites par rapport à ces tendances.

La métrique la plus simple à adopter compte tenu de la structure de $\mathbb{R}$ et de notre prerception de l'espace géométrique, est défini par $q(e_j,e_h)=0$ (resp.1) si $h\neq j$ (resp. $h=j$) ; où, bien entendu, $\{e_j/1\leqslant j\leqslant m\}$ est la base canonique de $\mathbb{R}^m$. Cette métrique, qui donne lieu à la distance suivante entre deux éléments i et i' de I

$$d^2(i,i') = \sum_{1\leqslant j\leqslant m} (\xi_{ij}-\xi_{i'j})^2, \qquad (1)$$

conduit à l'analyse en composantes principales non normée. Cependant la contribution à la distance (1) d'une variable dépend intimement de la dispersion de ses valeurs et cette dispersion est liée à la nature intrinsèque de la variable ; ainsi, dans l'exemple cité, une variable telle que "achat de timbres postes" aura une variance sensiblement plus faible que celle "dépense liée à l'entretien de la voiture". Comme on souhaite dans l'étude du comportement de la population, accorder a priori la même "importance" aux diverses variables mises en jeu ; la métrique q sera définie de la façon suivante :

$$q(e_j,e_h) = \begin{cases} 0 \text{ si } j \neq h \\ \dfrac{1}{s_j^2} \text{ si } j = h \end{cases} \qquad (2)$$

où $s_j^2 = \frac{1}{n} \sum_{1\leqslant i\leqslant n} (\xi_{ij}-\bar{\xi}_j)^2$ avec $\bar{\xi}_j = \frac{1}{n} \sum_{1\leqslant i\leqslant n} \xi_{ij}$. s_j^2 est la variance de la distribution de la j-ème variable v^j sur I.

Avec une telle métrique, la distance d(i,i') sera définie par

$$d^2(i,i') = \sum_{1\leqslant j\leqslant m} \left(\frac{\xi_{ij}-\xi_{i'j}}{s_j}\right)^2 \qquad (3)$$

De la sorte, nous verrons que la recherche des k premiers facteurs correspond à déterminer un système $\{u^1,u^2,\ldots,u^r,\ldots,u^k\}$ de k variables de synthèse , dont le coefficient de corrélation de deux quelconques d'entre elles est nul, tel que soit maximale la somme pour ces variables de la somme des carrés des coefficients de corrélation de chacune d'elle avec les différentes variables initiales de V ; soit

$$\sum_{1\leqslant r\leqslant k,\ 1\leqslant j\leqslant m} (\rho(v_j,u^r))^2 \qquad (4)$$

où $\rho(v^j, u^r)$ est le coefficient de corrélation entre les deux variables v^j et u^r.

3.2. Facteurs et axes factoriels ; représentation condensée.

Nous allons maintenant appliquer les considérations du paragraphe 2.4 précédent au cas particulier de l'analyse en composantes principales normée. Le centre de gravité du nuage $\mathcal{N}(I)$ de $\mathbb{R}^m$ est le point de coordonnées

$$(\bar{\xi}_1,\ldots,\bar{\xi}_j,\ldots,\bar{\xi}_m) \text{ où } \bar{\xi}_j = \frac{1}{n} \sum_{1\leq i\leq n} \xi_{ij} \qquad (5)$$

Le moment total d'inertie du nuage se met sous la forme

$$\sum_{1\leq i\leq n} \frac{1}{n} \sum_{1\leq j\leq m} \frac{1}{s_j^2} (\xi_{ij}-\bar{\xi}_j)^2 = m, \qquad (6)$$

Le terme (j,h) de la matrice de la forme quadratique d'inertie σ est

$$\sigma_{jh} = \sigma(e^*_j,e^*_h) = \sum_{1\leq i\leq n} \frac{1}{n}(\xi_{ij}-\bar{\xi}_j)(\xi_{ih}-\bar{\xi}_h) = \mathrm{Cov}(v^j,v^h), \qquad (7)$$

où $\mathrm{Cov}(v^j,v^h)$ désigne la covariance entre les deux variables v^j et v^h de V.

Le facteur u relatif à la valeur propre λ s'obtient à partir de la forme w solution de l'équation

$$\sum_{1\leq h\leq m} \frac{\sigma_{jh}}{s_j s_h} w_h = \lambda w_j \text{ pour } j = 1,2,\ldots,m, \qquad (8)$$

où $w_j = w(e_j)$, au moyen des relations

$$u_j = w_j/s_j, \text{ où } u_j = u(e_j), \text{ pour } j = 1,2,\ldots,m, \qquad (9)$$

La condition de normalisation du facteur u se met sous l'une des deux formes

$$\sum_{(j,h)} \mathrm{Cov}(v^j,v^h)u_j u_h = \sum_{(j,h)} \rho_{jh} w_j w_h = 1, \qquad (10)$$

La matrice symétrique de terme général $\rho_{jh}=\sigma_{jh}/s_j s_h$, $1\leq j\leq m$ et $1\leq h\leq m$; n'est autre que celle $\mathcal{R}$ des coefficients de corrélation entre variables de V. En introduisant le tableau T' des mesures $(\xi_{ij}-\bar{\xi}_j)/s_j$ *centrées réduites* des variables, $1\leq i\leq m$ et $1\leq j\leq m$, on a

$$\mathcal{R} = \frac{1}{n} {}^tT'.T' \qquad (11)$$

où ${}^tT'$ est la transposée de la matrice T'.

L'équation (8) peut alors se mettre sous la forme matricielle

$$\mathcal{R}\bar{w} = \lambda\bar{w}, \qquad (12)$$

où $\bar{w}$ désigne ici le vecteur *colonne* dont la suite des composantes est $w_1,w_2,\ldots,w_m$.

L'axe factoriel $\sigma(u)$ associé au facteur u s'exprime ici par

$$\sigma(u) = \sum_{(j,h)} \rho_{jh} w_h s_j e_j \qquad (13)$$

La formule de la représentation condensée du nuage sur l'espace des k permiers axes factoriels (cf. formule (45) § 2.4), devient ici

$$(\forall i \in I),\ M_i(k) = G + \sum_{1\leq r\leq k,\ 1\leq j\leq m} \{ \sum \left(\frac{\xi_{ij}-\overline{\xi}_j}{s_j}\right) w_j^r \} \sqrt{\lambda(r)}\, a_r, \qquad (14)$$

où w^r est la forme linéaire d'où résulte par les équations (9) le r-ème facteur u^r relatif à la r-ème valeur propre $\lambda(r)$ de la suite décroissante des k plus grandes valeurs propres.

En plaçant l'origine de l'espace $\mathbb{R}^m$ de représentation au centre de gravité G du nuage ; *un même facteur* u *apparait comme une variable centrée réduite* (i.e. de moyenne 0 et de variance 1, en vertu de la condition de normalisation). Le coefficient de corrélation entre la j-ème variable v^j et le facteur u relatif à la valeur propre λ se met dans ces conditions sous la forme

$$\rho(v^j,u) = \frac{\mathrm{Cov}(v^j,u)}{s_j}, \qquad (15)$$

En explicitant, on obtient

$$\rho(v^j,u) = \sum_h \left(\sum_i \frac{(\xi_{ij}-\overline{\xi}_j)}{s_j} \frac{(\xi_{ih}-\overline{\xi}_h)}{s_h} \right) w_h, \qquad (16)$$

qui n'est autre que la j-ème composante du vecteur colonne $\mathfrak{R}\overline{w}$; par conséquent λw_j (cf. formule (12)) et

$$\rho(v^j,u) = \lambda w_j, \qquad (17)$$

La formule de représentation (14) peut alors s'écrire aussi

$$(\forall i\in I),\ M_i(k) = G + \sum_{1\leq r\leq k} \frac{1}{\sqrt{\lambda(r)}} \{ \sum_{1\leq j\leq m} \frac{\xi_{ij}-\overline{\xi}_j}{s_j} \rho(v^j,u^r) \} a_r, \qquad (18)$$

Ainsi dans $\mathbb{R}^k$ dont l'origine représentera le centre de gravité et la base canonique, la base q-orthonormale $\{a_r / 1\leq r\leq k\}$; la r-ème coordonnée du point représentatif $M_i(k)$ se met sous l'une des deux formes

$$\sqrt{\lambda(r)} \sum_{1\leq j\leq m} \xi'_{ij} w_j^r = \frac{1}{\sqrt{\lambda(r)}} \sum_{1\leq j\leq m} \xi'_{ij} \rho(v^j,u^r) \qquad (19)$$

où $\xi'_{ij} = (\xi_{ij} - \overline{\xi}_j)/s_j$ est la mesure centrée réduite de la j-ème variable v^j sur le i-ème objet. La première expression sert au calcul et la seconde à l'interprétation. Cette r-ème coordonnée du i-ème point se présente comme une somme, sur l'ensemble des variables, pondérée par les coefficients de corrélation de ces variables avec le facteur en question, des mesures centrées réduites des variables sur le i-ème objet. On se rend par conséquent compte de l'intérêt à faire figurer sur le même espace de représentation des points-objets, chaque variable v^j par un point dont la suite des coordonnées $(\rho(v^j,u^1),\ldots,\rho(v^j,u^k))$ est celle des coefficients de corrélation de la variable avec les différents facteurs. Avec une telle représentation, on "voit" quelles sont les variables qui interviennent dans la définition du facteur qu'on interprétera alors. Une proximité d'autant plus grande entre deux points-objets représentés signifiera un comportement d'autant plus semblable des deux objets vis à vis des diverses tendances dégagés à travers les facteurs retenus. Une proximité donnée entre deux points-variables exprimera que les deux variables sont d'autant plus corél-

lés que leurs points représentatifs sont plus éloignés de l'origine. Enfin, une proximité sensible entre un point objet et un point variable, exprimera que cette variable, notons la v^j, a un rôle prépondérant par rapport aux autres chez le sujet i en question : ξ'_{ij} est relativement grand par rapport aux mesures $\xi'_{ih}, h \neq j$, des autres variables.

Compte tenu de la formule (6), on a

$$\lambda(1) + \lambda(2) + \ldots + \lambda(k) < m \qquad (20)$$

et la part de l'inertie expliquée par la projection du nuage est

$$(\lambda(1) + \lambda(2) + \ldots + \lambda(k)/m. \qquad (21)$$

Pratiquement, à partir de la valeur du rapport (21), on retient un petit nombre de facteurs (trois à quatre généralement) et on considère la projection du nuage sur les différents plans factoriels respectivement portés par les différents couples d'axes factoriels ; ainsi, si k=4, on considérera les six plans factoriels (1,2), (1,3), (1,4), (2,3), (2,4), (3,4).

Le plus souvent, le cardinal de I est très grand, pouvant aisément atteindre quelques milliers de sorte que le graphique des points-objets peut très vite devenir indéchiffrable. On introduit dans ces conditions une variable exogène ($\notin V$) ayant un très fort degré de discrimination qui définit une classification de I facilement perceptible et on représente chacune des classes par son centre de gravité. Ainsi, dans l'exemple qui illustre ce paragraphe, cette variable peut être le "revenu du ménage".

3.3. Représentation des variables

Nous allons montrer ici que la représentation des variables que nous avons ci-dessus considérée sur l'espace factoriel, est, à une homothétie près, celle qu'on obtient en regardant chaque variable v^j comme le point de $\mathbb{R}^n$ dont la suite $(\xi'_{1j},\ldots,\xi'_{ij},\ldots,\xi'_{nj})$ des coordonnées est la suite des mesures sur I, centrées réduites de cette variable ; et, en analysant le nuage des m points ainsi obtenus dans $\mathbb{R}^n$ muni de la métrique définie par le produit scalaire ordinaire. Chacun de ces points étant affecté de la masse unité ; le moment d'inertie de ce nuage par *rapport à l'origine* est

$$\sum_{1\leq j\leq m} \Big(\sum_{1\leq i\leq n} (\xi'_{ij})^2 \Big) = \sum_{1\leq j\leq m} \Big(\sum_{1\leq i\leq n} [(\xi_{ij}-\bar{\xi}_j)/s_j]^2 \Big) = mn, \qquad (22)$$

L'analyse du nuage des points-variables se fera au moyen de sa projection sur le *sous espace vectoriel de* $\mathbb{R}^n$, de dimension k, par rapport auquel le moment du nuage est minimum. Ce sous espace sera engendré par un système orthonormal de k vecteurs de $\mathbb{R}^n$ que nous noterons $\{b_r / 1\leq r\leq k\}$ où

$$b_r = \beta r_1 e_1 + \ldots + \beta r_i e_i + \ldots + \beta_{rn} e_n, \quad 1 \leq r \leq k, \qquad (23)$$

où $\{e_i / 1\leq i\leq n\}$ est la base canonique et où la condition d'orthonormalité s'écrit

$$(\forall(r,s)), 1\leq r\leq k \text{ et } 1\leq s\leq k; \sum_{1\leq i\leq n} \beta_{r_i}\beta_{s_i} = 1 (\text{resp.} 0) \text{ si } r=s \ (\text{resp.} r\neq s). \qquad (24)$$

Ce sous espace doit être déterminé tel que soit maximale la somme des normes au carré des projections des différents vecteurs

$$x'_j = \sum_{1\leq i\leq n} \xi'_{ij} e_i \text{ où } \xi'_{ij}=(\xi_{ij}-\bar{\xi}_j)/s_j,\ 1\leq j\leq m\ ; \tag{25}$$

représentant les divers sommets du nuage étudié ; soit

$$\sum_{1\leq r\leq k} \{ \sum_{1\leq j\leq m} [\sum_{1\leq i\leq n} \xi'_{ij}\beta_{ri}]^2\} \tag{26}$$

$\{b_r/1\leq r\leq k\}$ sera dans ces conditions formé d'un système orthonormal des k vecteurs propres correspondants aux k plus grandes valeurs propres de l'endomorphisme de $\mathbb{R}^n$ dont la matrice par rapport à la base canonique est $(T'{}^tT')$ où, rappelons le, tableau n×m, T', a pour j-ème vecteur colonne

$$x'_j = \begin{bmatrix} \xi'_{1j} \\ \vdots \\ \xi'_{nj} \end{bmatrix} ;$$

en effet, l'expression entre accolades de (26) peut se mettre sous la forme

$${}^tb_r(\sum_{1\leq j\leq m} x'_j {}^tx'_j)b_r = {}^tb_r(T'{}^tT')b_r\ ; \tag{27}$$

où nous avons noté également le vecteur de $\mathbb{R}^n$ et sa forme matricielle colonne. Ainsi, le r-ème vecteur propre b_r est solution de l'équation matricielle

$$(T'{}^tT')b_r = \nu(r)b_r, \tag{28}$$

auquel on imposera la condition de normalisation.

Reprenons l'équation (12) du paragraphe 3.2 précédent. En multipliant à gauche les deux membres de cette équation relative au r-ème facteur du nuage de I, par T' ; on a

$$(T'{}^tT')(T'\overline{w}) = n\lambda(r)\overline{w}. \tag{29}$$

A l'axe factoriel b_r, associons le facteur β^r qui est la forme linéaire définie par la projection orthogonale sur b_r :

$$\beta^r = \sum_{1\leq i\leq n} \beta_{ri} e_i^* \tag{30}$$

où $\{e_i^*/1\leq i\leq n\}$ est la base duale de la base canonique de $\mathbb{R}^n$ et où les β_{ri} ont été définis dans la formule (23). L'équation (29) montre que la suite des facteurs $(\beta^1,\ldots,\beta^r,\ldots,\beta^k)$ attachée à la suite des k plus grandes valeurs propre s'obtient, *à un facteur de normalisation près*, à partir de la suite, déjà définie (cf.§ 3.2), de formes linéaires sur $\mathbb{R}^n(w^1,\ldots,w^r,\ldots,w^k)$ par les relations

$$\beta^r = T'w^r\ ;\ 1\leq r\leq k\ ; \tag{31}$$

de plus, on a pour les moments d'inertie

$$\nu(r) = n\lambda(r), \quad 1 \leqslant r \leqslant k \; ; \tag{32}$$

La suite des composantes du vecteur colonne $T'w^r$ n'est autre que la suite des valeurs du facteur normalisé u^r sur la suite des points du nuage $\mathcal{N}(I)$, rapporté à son centre de gravité. Comme

$$\sum_{1 \leqslant i \leqslant n} \beta^2_{r_i} = n \; \mathrm{var}(T'w^r) = n, \tag{33}$$

le facteur de normalisation, auquel nous venons de faire allusion, est donc $\frac{1}{\sqrt{n}}$; il en résulte que

$$(\forall i \in I), \quad \beta_{ri} = \frac{1}{\sqrt{n}} u^r(M_i - G). \tag{34}$$

Par conséquent, *la suite des composantes du r-ème facteur normalisé* β^r *du nuage associé à* V, *est, au facteur multiplicatif* $1/\sqrt{\lambda(r)n}$, *la valeur de la* r*-ème coordonnée factorielle sur la suite des points de* I.

D'autre part, la projection orthogonale du point variable x'_j sur le r-ème axe factoriel b_r, se met sous la forme

$$^t x'_j . b_r = \sum_{1 \leqslant i \leqslant n} \xi'_{ij} \beta_{ri} = \frac{1}{\sqrt{n}} (^t x'_j T' w^r) \tag{35}$$

qui vaut (voir formule (16)) $\sqrt{n}\rho(v^j, u^r)$; $\rho(v^j, u^r)$ étant le coefficient de corrélation entre la j-ème variable et le r-ème facteur du nuage de I.

Quant au critère maximisé (26) ; il s'agit bien, au facteur n près, de la somme des carrés des coefficients de corrélation des différentes variables v^j de V avec chacune des variables de synthèse définies respectivement par chacun des facteurs $u^1, u^2, \ldots, u^r, \ldots, u^k$.

4. ANALYSE DES CORRESPONDANCES

4.1. Introduction

L'analyse en composantes principales normée apparait, on l'a vu, comme une analyse des corrélations ; et, on a pu se rendre compte dans cette analyse qu'il y avait une dissymétrie dans le traitement de chacun des côtés du tableau rectangulaire de description. Cette dissymétrie est finalement liée à la différence de nature entre une variable, représentée par une colonne, et un objet représenté par une ligne ; il y a en effet autant de différence entre une variable et un objet qu'entre un appareil de mesure et l'objet sur lequel on effectue la mesure (e.g. entre une balance et l'individu pesé). Dans les enquêtes ou études expérimentales, l'ensemble I des individus ou objets s'impose pratiquement ; il s'agit, on l'espère, d'un échantillon "représentatif" de la population étudiée ; tandis que l'ensemble V des variables, chacun de ses éléments est déterminé avec minutie par le Spécialiste et derrière une même variable, il y a toute la connaissance et l'appréhension d'un domaine scientifique donné.

L'analyse factorielle des correspondances est essentiellement une analyse métrique des lignes ou colonnes d'un "grand" tableau de contingence

croisant deux variables dont chacune définit une partition sur la population étudiée. I de cardinal n (resp. J de cardinal m) représentant l'ensemble des modalités de l'une des variables (resp. de l'autre) , nous avons déjà présenté (cf. chapitre 5 § IV 2) le support de l'information, qui est le tableau des proportions

$$\{f_{ij}/(i,j) \in I\times J\}, \tag{1}$$

obtenues à partir d'un échantillon. Il s'agit d'un système de masses positives ou nulles, de somme 1, affectées aux couples $(i,j) \in I\times J$, sensé refléter celui de la loi de probabilité

$$\pi_{I\times J} = \{\pi_{ij}/(i,j) \in I\times J\}, \tag{2}$$

sur l'ensemble produit $I\times J$; laquelle, rarement connue, est définie au niveau de la population globale. Dans la mesure où les éléments de l'échantillon sont extraits indépendamment l'un de l'autre, (1) constitue, en vertu des lois des grands nombres, une estimation possédant de bonnes qualités statistiques de (2).

Les deux côtés du tableau des données étant de même nature, il est manifeste que le rôle de I par rapport à J est de même nature que le rôle de J par rapport à I ; une analyse métrique de J par rapport à I sera tout à fait symétrique d'une analyse métrique de I par rapport à J ; d'ailleurs on verra comment les facteurs de l'un des nuages $\mathcal{N}(I)$ ou $\mathcal{N}(J)$ (cf. chap. 5 § IV. 2) se déduisent aisément par des formules de "transition" des facteurs de l'autre nuage.

Rappelons que dans l'introduction du nuage

$$\mathcal{N}(I) = \{(f^i_J, p_{i.})/i\in I\}, \tag{3}$$

on retient pour la description de $i\in I$, son "profil" à travers J ; c'est à dire, la suite des parts des différents $j\in I$ qui entrent dans la composition de la catégorie i ; ainsi i est représenté par le point $\mathbb{R}^m$

$$f^i_J = (f^i_1,\ldots,f^i_j,\ldots,f^i_m) \tag{4}$$

où $f^i_j = f_{ij}/p_{i.}$. En effet, pour j donné, ce qui importe dans la comparaison des divers éléments de I, ce n'est pas tant les proportions absolues f_{ij} que celles relatives f^i_j qui permettent la mesure de la part intrinsèque de j, en ne tenant pas compte de la variabilité de l'importance numérique $p_{i.}$ sur I. Ainsi, pour une étude sur la répartition géographique des professions, où l'une des variables est "profession exercée" et l'autre variable "lieu de l'exercice" ; relativement à deux professions i et i' et à un lieu de travail j, ce qu'il y a lieu de comparer *dans* chacune des deux professions, c'est bien *la part* de ceux qui pratiquent au lieu j. On préserve l'importance de présence $p_{i.}$ de i en affectant le sommet f^i_J du poids $p_{i.}$.

Rappelons également que pour analyser le nuage $\mathcal{N}(I)$ on munit l'espace ambiant $\mathbb{R}^m$ de la métrique q suivante

$$q(e_j,e_k) = \begin{cases} 0 & \text{si } j\neq k \\ \dfrac{1}{p_{.j}} & \text{si } j=k \end{cases} \tag{5}$$

où $\{e_j/1\leq j\leq m\}$ est la base canonique. Nous avons déjà signalé que l'introduction de cette métrique se justifiait pour deux raisons ; la première, algébrique, est définie par la condition de l'équivalence distributionnelle (cf. chapitre 5 § IV.2) qui permet de comprendre la stabilité des résultats lorsqu'on réunit deux modalités voisines quant à leurs profis de I (resp. de J) ; ainsi, dans l'exemple mentionné, relativement au problème de la délimitation des régions $j\in J$, dans la mesure où, pour deux régions contigües j_1 et j_2, la distribution de la part de chaque profession exercée ne varie pas sensiblement de j_1 à j_2 ; on ne change pratiquement pas les résultats de l'analyse en remplaçant les modalités j_1 et j_2 par une seule j_o représentant la réunion de j_1 et j_2. La seconde raison de l'introduction de (5) est statistique ; en effet, avec une telle métrique, le moment total d'inertie du nuage $\mathcal{N}(I)$ n'est autre que la statistique du χ^2 attachée au tableau de contingence $I\times J$; de plus, la distance entre deux éléments i et i':

$$d^2(i,i') = \sum_{1\leq j\leq m} \frac{1}{p_{\cdot j}} (f_j^i - f_j^{i'})^2 \qquad (6)$$

est exactement la distance du χ^2 associée à la loi de probabilité $\{p_{\cdot j}/j\in J\}$ entre les deux distributions $\{f_j^i/j\in J\}$ et $\{f_j^{i'}/j\in J\}$.

On peut certes très aisément ramener formellement tout tableau rectangulaire de données à un tableau de nombres positifs ou nuls de somme 1 ; c'est d'ailleurs ce que le praticien chevronné de l'Analyse des Correspondances a tôt fait de faire ; mais il est clair que cette analyse ne se justifie que dans la mesure où la représentation retenue du tableau respecte la nature mathématique des objets qui indexent l'un ou l'autre des côtés du tableau et où la métrique du χ^2 semble indiquée pour exprimer les proximités.

4.2. Forme quadratique d'inertie, Facteurs et Axes factoriels

La forme quadratique $\sigma(u,v)$ d'inertie du nuage $\mathcal{N}(I)$ (cf.(3)), est

$$\sigma(u,v) = \sum_{i\in I} p_{i\cdot} u(f_J^i - g_J) v(f_J^i - g_J) \qquad (7)$$

où, rappelons le, le centre de gravité g_J du nuage, a pour coordonnées $(p_{\cdot 1},\ldots,p_{\cdot m})$ d'où

$$\sigma_{jk} = \sigma(e_j^*, e_k^*) = \sum_{i\in I} p_{i\cdot}(f_j^i - p_{\cdot j})(f_k^i - p_{\cdot k}) = \sum_{i\in I} \frac{f_{ij} f_{ik}}{p_{i\cdot}} - p_{\cdot j} p_{\cdot k} \qquad (8)$$

Le facteur u relatif à la valeur propre λ s'obtient à partir de la forme w solution de l'équation

$$\sum_{1\leq h\leq m} \Big\{ \sum_{1\leq i\leq n} p_{i\cdot} \frac{f_j^i - p_{\cdot j}}{\sqrt{p_{\cdot j}}} \times \frac{f_h^i - p_{\cdot h}}{\sqrt{p_{\cdot h}}} \Big\} w_h = \lambda w_j \qquad (9)$$

pour $1\leqslant j\leqslant m$; où $w_j=w(e_j)$, au moyen des relations

$$u_j=w_j/\sqrt{p_{.j}}, \text{ où } u_j=u(e_j), \text{ pour } j=1,2,\ldots,m.$$

En utilisant la seconde expression de la formule (8) pour σ_{hj} ; le terme général de la matrice à diagonaliser devient

$$\sum_{1\leqslant i\leqslant n} \frac{f_{ij}f_{ih}}{p_{i.}\sqrt{p_{.j}p_{.h}}} - \sqrt{p_{.j}p_{.h}} \qquad (10)$$

Le nuage $\mathcal{N}(I)$ étant situé dans un même hyperplan orthogonal au vecteur dont toutes les composantes sont égales à 1 ; le facteur v, défini par

$$v(e_j)=v_j=1 \text{ pour tout } j=1,\ldots,m \qquad (11)$$

est relatif à la valeur propre 0, comme d'ailleurs le lecteur peut le vérifier.

En posant $v=q(t)$ et $u=q(s)$; la condition de q-orthogonalité des axes factoriels s et t, montre que tout facteur u relatif à une valeur propre λ non nulle, satisfait la relation

$$\sum_{1\leqslant j\leqslant m} u_j p_{.j}=0 \ ; \qquad (12)$$

Cette dernière relation avec (10) permettent de ramener la recherche des facteurs non triviaux (i.e. relatifs à des valeurs propres non nulles) à la solution des équations

$$\sum_{1\leqslant h\leqslant m} \{ \sum_{1\leqslant i\leqslant n} \frac{f_{ij}f_{ih}}{p_{i.}\sqrt{p_{.j}p_{.h}}}\}w_h=\lambda w_j, \text{ pour } j=1,\ldots, m \ ; \qquad (13)$$

lesquelles résultent de la simplification des équations (9).

On peut donner de (13) la forme suivante faisant directement intervenir le facteur u

$$\sum_{1\leqslant h\leqslant m} \{ \sum_{1\leqslant i\leqslant n} f_i^j f_h^i\}u_h=\lambda u_j \ ; \ j=1,\ldots,m. \qquad (14)$$

Si l'expression (13) sert au calcul, celle (14) fait mieux apparaître que, ce dont il est question ici, est bien l'analyse des distributions. La matrice carré $n\times n$ de terme général $\sum_{1\leqslant i\leqslant n} f_i^j f_h^i$ résulte du produit de deux matrices qu'on peut noter f_I^{*J} et f_J^{*I} où alors f_I^{*J} est le tableau des proportions conditionnelles f_i^j à $m=\text{card}(J)$ lignes et à $n=\text{card}(I)$ colonnes ; cette matrice définit une transition probabiliste de J vers I : $p_{i.}=\sum_{i\in I} f_i^j p_{.j}$. De même f_J^{*I} est la matrice des proportions conditionnelles f_j^i à $n=\text{card}(I)$ lignes et à $m=\text{card}(J)$ colonnes ; cette matrice définit une transition probabiliste de I vers J.

La condition de normalisation du facteur u se met sous l'une des deux formes

$$\sum_{i\in I} p_{i.}\{u(f_J^i-g_J)\}^2=\sum_{(j,h)} \{ \sum_{i\in I} \frac{(f_{ij}-p_{i.}p_{.j})}{\sqrt{p_{i.}p_{.j}}}\times\frac{(f_{ih}-p_{i.}p_{.j})}{\sqrt{p_{i.}p_{.h}}}\}w_j w_h=1, \qquad (15)$$

Si on veut exprimer $\sigma(u)$, de norme $\sqrt{\lambda}$, par rapport à la base canonique ; on a

$$\sigma(u)= \sum_{(h,j)} \sum_{i\in I} p_{i.}(f^i_j-p_{.j})(f^i_h-p_{.h})u_h e_j, \qquad (16)$$

En considérant la base q-orthonormale des k premiers axes factoriels $\{\sigma(u^r)/\sqrt{\lambda(r)} \,/\, 1\leqslant r\leqslant k\}$, on a la représentation condensée du point f^i_J du nuage $\mathcal{N}(I)$ définie par

$$f^i_J-g_I= \sum_{1\leqslant r\leqslant k} \{\sqrt{\lambda(r)} \sum_{1\leqslant j\leqslant m} (f^i_j-p_{.j})u^r_j\} \frac{\sigma(u^r)}{\sqrt{\lambda(r)}}, \qquad (17)$$

relation qui peut aussi se mettre sous les formes plus synthétiques

$$f^i_J-g_J= \sum_{1\leqslant r\leqslant k} u^r(f^i_J-g_J)\sigma(u^r)= \sum_{1\leqslant r\leqslant k} q(a_r,f^i_J-g_J)a_r ; \qquad (18)$$

où q est la métrique et où, on a posé pour a_r, le vecteur unitaire $\sigma(u^r)/\sqrt{\lambda(r)}$.

En plaçant l'origine au centre de gravité g_J du nuage $\mathcal{N}(I)$, la *moyenne* et la *variance* de la suite des q-projections du nuage sur le r-ème axe factoriel, sont respectivement 0 et $\lambda(r)$; ce qui se traduit par les relations

$$\sum_{i\in I} p_{i.}F(i,r)=0 \text{ et } \sum_{i\in I} p_{i.}\{F(i,r)\}^2=\lambda(r), \qquad (19)$$

où on a posé $F(i,r)=q(a_r,f^i_J-g_I)$: coordonnée du i-ème point se le r-ème axe factoriel. On peut également écrire les relations (19), en faisant appel au facteur u^r, sous la forme

$$\sum_{i\in I} p_{i.}u^r(f^i_J-g_J)=0 \text{ et } \sum_{i\in I} p_{i.}\{u^r(f^i_J-g_J)\}^2=1, \qquad (20)$$

4.3. Formules de transition de $\mathcal{N}(I)$ à ceux de $\mathcal{N}(J)$.

Nous avons annoncé dans l'introduction qu'on va pouvoir déterminer facilement les facteurs du nuage $\mathcal{N}(J)=\{(f^j_J,p_{.j})/j\in J\}$ à partir de ceux de $\mathcal{N}(I)$ par des formules de transition. Ceci est pratiquement d'une importance cruciale lorsque, comme c'est le cas le plus fréquent, l'un des deux ensembles J est de cardinal m sensiblement plus petit que celui n de I ; car de la sorte, on se ramène à la diagonalisation d'une matrice m×m. Par ailleurs, la recherche simultanée des facteurs de chacun des deux nuages permet une interprétation plus symétrique de l'Analyse des Correspondances.

Si on se conformait à [6], en posant

$$t_{ij}= \frac{f_{ij}-p_{i.}p_{.j}}{\sqrt{p_{i.}p_{.j}}}, \qquad (21)$$

l'équation (9) se met sous la forme

$$t_{TT}\,\overline{w} = \lambda\overline{w} \qquad (22)$$

où T est la matrice n×m des nombres t_{ij} et où $\overline{w}$ est le vecteur colonne dont la suite des composantes est $(w_1,\ldots,w_j,\ldots,w_m)$. Le caractère symétrique en i et j de t_{ij} permet de voir que l'équation relative aux facteurs du nuage $\mathcal{N}(J)$ se met sous la forme

$$T^tT\,\overline{v} = \lambda'\overline{v} \qquad (23)$$

où $\overline{v}$ est le vecteur colonne dont la suite des composantes $(v_1,\ldots,v_i,\ldots,v_n)$ définit la suite des valeurs de la forme linéaire v sur la base canonique $\{e_i/1\leq i\leq n\}$. Dans ces conditions, les valeurs propres non nulles de T^tT et de ^tTT coïncident et les vecteurs propres respectifs $\overline{w}$ et $\overline{v}$ relatifs à une même valeur propre λ se correspondent, à une homothétie près, au moyen de la relation $\overline{v}=T\overline{w}$. Alors les deux facteurs normalisés ϕ et Ψ, respectifs aux deux nuages $\mathcal{N}(I)$ et $\mathcal{N}(J)$, relatifs à une même valeur propre λ, dont les composantes sont définies par

$$\phi_j=\phi(e_j)=w_j/\sqrt{p_{.j}} \text{ et } \Psi_i=\Psi(e_i)=v_i/\sqrt{p_{i.}} \;;\; 1\leq j\leq m \text{ et } 1\leq i\leq n \;;$$

se déduisent l'un de l'autre par les relations, que nous allons reprendre,

$$\phi_j=\frac{1}{\sqrt{\lambda}}\Psi(f_I^j-g_I) \text{ et } \Psi_i=\frac{1}{\sqrt{\lambda}}\phi(f_J^i-g_J), \text{ pour tout } (i,j)\in I\times J. \qquad (24)$$

qui, en vertu de la relation (12), se simplifient

$$\phi_j=\frac{1}{\sqrt{\lambda}}\sum_{1\leq i\leq n} f_i^j\Psi_i=\frac{1}{\sqrt{\lambda}}\Psi(f_I^j) \text{ et } \Psi_i=\frac{1}{\sqrt{\lambda}}\sum_{1\leq j\leq m} f_j^i\phi_j=\frac{1}{\sqrt{\lambda}}\phi(f_j^i), \qquad (25)$$

Les formules (25), dites de "transition", montrent que la suite des valeurs du facteur normalisé Ψ sur la suite des points $(f_I^j/j\in J)$ définit, au facteur $\sqrt{\lambda}$ près, la suite des composantes du facteur normalisé ϕ, relatif à la même valeur propre. De la même façon, en intervertissant les rôles de I et J; la suite des valeurs du facteur normalisé ϕ sur la suite des points $(f_J^i/i\in I)$ définit, au facteur $\sqrt{\lambda}$ près, la suite des composantes du facteur normalisé Ψ.

Pour établir les formules de transition nous allons entreprendre une démarche équivalente à celle figurée ci-dessus, mais qui paraîtrera plus "naturelle", car plus liée à la nature des objets manipulés qui sont des distributions ; et ce, en partant directement de l'équation (14) de définition d'un facteur. Cette équation se met sous la forme matricielle

$$f_I^{*J}.f_J^{*I}\,\overline{\phi} = \lambda\overline{\phi}, \qquad (26)$$

où, rappelons le, f_I^{*J} est la matrice des proportions conditionnelle f_i^j, indexée par J×I et où $\overline{\phi}$ est le vecteur colonne dont la suite des composantes est celle du facteur ϕ.

Par symétrie, un même facteur Ψ du nuage $\mathcal{N}(J)$, associé à la valeur propre μ, est défini par

$$f_J^{*I}.f_I^{*J}\,\overline{\Psi} = \mu\,\overline{\Psi}, \qquad (27)$$

où $\overline{\Psi}$ est le vecteur colonne dont la suite des composantes est celle du facteur.

En mutlipliant à gauche les deux membres de (26) (resp. de (27)) par f_J^{*I}(resp. par f_I^{*J}), on voit que si $\overline{\phi}$ (resp.$\overline{\Psi}$) est vecteur propre de f_I^{*J}. f_J^{*I}(resp. $f_J^{*I}.f_I^{*J}$) relatif à la valeur propre λ (resp.μ), $f_J^{*I}\overline{\phi}$ (resp.$f_I^{*J}\overline{\Psi}$) est vecteur propre relatif à la valeur propre λ (resp.μ) de $f_J^{*I}.f_I^{*J}$(resp. de $f_I^{*J}.f_J^{*I}$).

Ainsi les valeurs propres associées à l'un des nuages sont identiques à celles associées à l'autre nuage ; de plus, si ϕ est facteur associé au nuage $\mathcal{N}(I)$ et relatif à la valeur propre λ, $f_J^{*I}\overline{\phi}$ est un vecteur dont la suite des composantes représente celle du facteur Ψ associé au nuage $\mathcal{N}(J)$ et relatif à la même valeur propre λ ; d'autre part, la suite des composantes de $f_J^{*I}\overline{\phi}$ est exactement la suite des valeurs du facteur ϕ sur la suite des points $(f_J^i/i\in I)$. De façon précise, la correspondance entre facteurs *normalisés* ϕ et Ψ , relatifs à la même valeur propre λ, est définie par les formules (25) ; en effet si ϕ est un facteur normalisé, le facteur Ψ', dont la suite des composantes est définie par le vecteur $f_J^{*I}\overline{\phi}$, est, en vertu de l'équation (26), de norme λ.

En désignant par F(i,r) (resp. G(j,r) la coordonnée du i-ème (resp.j-ème) point du nuage $\mathcal{N}(I)$ (resp.$\mathcal{N}(J)$) sur le r-ème axe factoriel, les formules de transition (25) peuvent se mettre sous la forme

$$G(j,r)=\frac{1}{\sqrt{\lambda(r)}}\sum_{1\leqslant i\leqslant n} F(i,r)f_i^j \text{ et } F(i,r)=\frac{1}{\sqrt{\lambda(r)}}\sum_{1\leqslant j\leqslant m} G(j,r)f_j^i, \qquad (28)$$

Les formules (28) montrent l'intérêt d'une représentation simultanée des deux nuages $\mathcal{N}(I)$ et $\mathcal{N}(J)$ dans le même sous espace factoriel. En considérant, pour fixer les idées, le plan des deux premiers axes factoriels correspondants aux deux plus grandes valeurs propres $\lambda(1)$ et $\lambda(2)$; le point i de I (resp. j de J) sera représenté par le point de coordonnées (F(i,1),F(i,2)) (resp.(G(j,1),G(j,2))). L'interprétation de la représentation simultanée repose sur le fait que la r-ème coordonnée du point i de I (resp. j de J) apparait comme la moyenne, pondérée par les proportions conditionnelles f_j^i (resp. f_i^j), des coordonnées sur le r-ème auxe factoriel des différents points de J (resp. de I) au coefficient de dilatation $1/\sqrt{\lambda(r)}$ près. Ainsi, pour l'exemple qui illustre ce paragraphe, le lieu j est d'autant plus proche de la profession i que $f_j^i=f_{ij}/p_{i.}$ est plus grand ; c'est-à-dire que, parmi ceux qui pratiquent la profession i, la part qui exerce au lieu j, est plus grande. Il est donc instructif de regarder l'éloignement d'un même i des différents $j\in J$. Toutefois l'interprétation la plus féconde est relative à la définition des facteurs d'un même nuage et à l'analyse des positions relatives, sur différents plans factoriels, entre les différents points représentatifs de ce nuage.

Notons que dans une même solution (28), la somme des pondérations positives étant égale à 1, si $1/\sqrt{\lambda(r)}$ était strictement inférieur à 1 ; on aurait pour tout couple (i,j) les inégalités strictes suivantes

$\min_j G(j,r)<F(i,r)<\max_j G(j,r)$ et $\min_i F(i,r)<G(j,r)<\max_i F(i,r)$
qui sont impossibles à réaliser simultanément. Donc, on a $\lambda(1)\leq 1$.

Terminons ce paragraphe par une question de notations. Nous avons vu que l'équation (14) de définition du facteur ϕ, du nuage $\mathcal{N}(I)$, relatif à la valeur propre λ, peut se mettre sous la forme

$$f_I^{*J} \circ f_J^{*I} \circ \phi_J = \lambda \phi_J \tag{29}$$

où, rappelons le, f_I^{*J} est la matrice des proportions conditionnelles f_i^j ; laquelle est ici indexée par $J\times I$; c'est à dire, f_i^j est le terme de la j-ème ligne et de la i-ème colonne de cette matrice. ϕ_J est le vecteur colonne à m composantes dont la j-ème est définie par $\phi_j=\phi(e_j)$. Compte tenu des formules de transition, on peut par transposition mettre l'équation (29) sous la forme

$$\Psi^J \circ f_J^I \circ f_I^J = \lambda \Psi^J \tag{30}$$

où, cette fois-ci, la matrice f_I^J est indexée par $I\times J$; f_i^j est le terme de la i-ème ligne et de la j-ème colonne. Ψ^J est le vecteur ligne à m composantes dont la j-ème définit la valeur du facteur Ψ, du nuage $\mathcal{N}(J)$ et relatif à la valeur propre λ, sur le j-ème sommet f_I^j. C'est cette dernière forme (29), où un facteur est défini comme une fonction, par la suite de ses valeurs sur les différents points du nuage auquel il est relatif, qui est adoptée dans [1] .

4.4. Reconstitution approchée du tableau $f_{I\times J}$.

A partir de la formule (18) on obtient une reconstitution approchée, à partir des k premiers facteurs, du tableau des fréquences $\{f_{ij}/(i,j)\in I\times J\}$. En effet, reprenons cette formule

$$f_J^i - g_j \tilde{=} \sum_{1\leq r\leq k} \phi^r (f_J^i - g_J)\sigma(\phi^r), \tag{31}$$

et considérons la j-ème composante, par rapport à la base canonique, des deux membres. En approximant composante par composante, on obtient,

$$(\forall (i,j), f_{ij} \tilde{=} p_{i.}p_{.j}\{1+ \sum_{1\leq r\leq k} (\lambda(r))^{3/2}\phi_j^r\psi_i^r\} \tag{32}$$

compte tenu de ce que la j-ème composante de $\sigma(\phi^r)=\lambda q^{-1}(\phi^r)$ est $\lambda p_{.j}\phi_j^r$ et des formules de transition (25) qui permettent d'ailleurs de donner à la formule (32) la forme

$$(\forall (i,j), f_{ij} \tilde{=} p_{i.}p_{.j}\{1+ \sum_{1\leq r\leq k} \sqrt{\lambda(r)}\psi^{rj}\phi^{ri}\} \tag{33}$$

où $\psi^{rj}=\psi^r(f_I^j)$ et $\phi^{ri}=\phi(f_J^i)$. On peut enfin écrire

$$(\forall (i,j), f_{ij} \tilde{=} p_{i.}p_{.j}\{1+ \sum_{1\leq r\leq k} (\lambda(r))^{-1/2}F(i,r)G(j,r)\} \tag{34}$$

4.5. Aides à l'interprétation

Relativement au nuage auquel on s'intéresse, soit par exemple $\mathcal{N}(I)$, pour faciliter et rendre plus sûre l'interprétation de ses facteurs ; on définit pour chaque élément i de I, deux coefficients $\alpha_r(i)$ et $\rho_r(i)$. $\alpha_r(i)$, appelée *contribution absolue* de l'élément i de I au r-ème facteur, représentera la part prise par cet élément dans l'inertie expliquée par ce facteur. $\rho_r(i)$, appelé *contribution relative* de l'élément i de I au r-ème facteur, représentera la part prise par ce facteur dans la dispersion de l'élément i ; il s'agira du carré du coefficient de corrélation entre le r-ème facteur et la variable. q-projection sur $(f_J^i - g_J)$; c'est à dire, encore, le carré du cosinus de l'angle formé avec le r-ème axe factoriel par la droite orientée reliant le centre de gravité au point i.

Compte tenu de la seconde formule (20), la part de l'inertie $\lambda(r)$ imputable au point i est

$$\alpha_r(i) = p_{i.}\{\phi^r(f_J^i - g_J)\}^2 = p_{i.}\{F(i,r)\}^2/\lambda(r). \qquad (35)$$

D'autre part, le carré de la distance au sens de q, de la projection du point i sur le r-ème axe factoriel, vaut $\lambda(r)\{\phi^r(f_J^i - g_J)\}^2$; tandis que la distance du point i au centre de gravité est définie par

$$d^2(f_J^i, g_J) = \sum_{1\leqslant j\leqslant m} \frac{1}{p_{.j}}(f_j^i - p_{.j})^2 = \sum_{r\geqslant 1} \lambda(r)\{\phi^t(f_J^i - g_J)\}^2$$

de sorte qu'on a

$$\rho_r(i) = \lambda(r)\{\phi^r(f_J^i - g_J)\}^2 / \sum_{1\leqslant j\leqslant m} \frac{1}{p_{.j}}(f_j^i - p_{.j})^2 \qquad (36)$$

De la même façon, on définit, relativement au nuage $\mathcal{N}(J)$, les coefficients $\alpha_r(j)$ et $\rho_r(j)$ représentant les contributions absolue et relative de j appartenant à J, au r-ème facteur.

Il est clair que ces deux coefficients peuvent être définis dans le cas général de l'analyse d'un nuage de points dans un espace euclidien.

BIBLIOGRAPHIE

[1] J.P. BENZECRI, *"L'Analyse des Données"*, tome II, Dunod, Paris, 1973.

[2] J. DIEUDONNE, *"Algèbre Linéaire et Géométrie élémentaire"*, Annexe II : *Géométrie d'une forme bilinéaire symétrique. Les Langages "projectifs" et "non euclidien"* ; 3ème édition, Hermann, Paris.

[3] Mme ESCOFFIER née B. CORDIER, *"Analyse factorielle des Correspondances"*, Cahier du B.U.R.O. n° et thèse de 3è cycle, Rennes, 1965.

[4] R. GODEMENT, *"Cours d'Algèbre"*, Hermann, Paris.

[5] L. LEBART, A. MOINEAU, N. TABARD, *"Techniques de la description statistique"*, Dunod, Paris 1977.

[6] Mme NORA, *"Analyse factorielle d'un nuage de points"*, publication interne (L.S.M. - I.S.U.P.), Université Paris VI.