Statistique et analyse des données

Yves Le Foll

Pondération des distances en analyse factorielle

Statistique et analyse des données, tome 7, n° 1 (1982), p. 13-31.

<http://www.numdam.org/item?id=SAD_1982__7_1_13_0>

Statistique et Analyse des données
1982 - 1 - 13-31

PONDERATION DES DISTANCES EN ANALYSE FACTORIELLE

Yves LE FOLL

Département d'Informatique de l'E.N.S.T.
46, rue Barrault - 75634 PARIS CEDEX 13

Résumé : *De nombreuses méthodes d'analyse factorielle se rattachent au formalisme de l'Analyse en Composantes Principales (ACP). Celui-ci ne dépend que de l'ensemble des points représentés, de ses pondérations et de ses distances.*

Pour donner aux composantes factorielles des propriétés quasi-ordinales, locales ou statistiques supplémentaires, il est fructueux de pondérer les distances, c'est à dire les couples de points.

Après avoir précisé les problèmes extrêmaux associés et leurs applications (§.1), le problème de base est directement résolu et discuté en terme d'analyse factorielle sur tableau de distances (§.2). Sa solution est explicitée dans le cadre de l'ACP et les classiques indicateurs d'aide à l'interprétation sont repréci-sés (§.3).

Comme application, on montre que l'analyse des correspondances peut être retrouvée, caractérisée et généralisée en pondérant des distances très élémentaires par les valeurs du tableau analysé (§.4).

Abstract : *Many factorial analysis methods can be considered as examples of Principal Component Analysis (PCA). The latter is characterized by the weights and distances of the points shown.*

When the weights are associated with the distances, the factorial components can own quasi-ordinal, local or statistical properties.

First, associated extremal problems and their applications are mentioned (§.1). Afterwards the basic problem is resolved and discussed as a problem of proximity analysis (§.2). Its solution is also translated in the PCA language and usual helping interpretation indicators are specified (§.3).

As an application, Correspondence Analysis can be demonstrated when one wants and characterized and generalised starting from very elementary distances. We use the values of the analysed contingency-table for the weighting of these distances (§.4).

Mots-clés : *Pondération des distances, Analyse en Composantes Principales, Analyse des Proximités, Analyse des Correspondances.*

0 - INTRODUCTION

De nombreuses méthodes d'analyse factorielle conduisent à la démarche suivante : Chercher à caractériser les axes et les sous-espaces principaux d'inertie d'une représentation euclidienne et à interpréter leurs composantes principales. Cette représentation est alors complètement déterminée par un triple (E, P_E, d_{EE}) où

- E est l'ensemble des éléments $\{e | e \in E\}$ que l'on représente par des points ;
- P_E est l'ensemble de pondérations $\{p_e | e \in E\}$ respectivement affectées aux éléments de E ;
- d_{EE} est un ensemble de distances $\{d_{ee'} | (e,e') \in E \times E\}$ euclidiennes, entre les éléments de E pris 2 à 2.

Or dans un certain nombre de situations pratiques, les axes factoriels extraits n'ont pas toutes les propriétés souhaitables, qu'elles soient ORDINALES (pour des points associés à des modalités ordonnées), LOCALES (quant à la restitution de certains types de distances) ou STATISTIQUES (élimination de certains effets, stabilité des formes identifiées). Ces problèmes sont usuellement traités en adaptant à la situation observée p_E ou d_{EE}, ou plus spécifiquement en imposant des contraintes de structure aux facteurs (cf [4], [5], [12], [19], [20] et [21]).

Pour répondre de façon plus souple ou plus pratique à ces problèmes, nous introduisons des pondérations généralisées $P_{EE} = \{p_{ee'} | (e,e') \in E \times E\}$ s'appliquant, non plus aux éléments de E, mais à leurs distances. Il s'agit tout d'abord de définir quels types de problèmes extrêmaux cela permet de résoudre et les techniques ou applications qui s'y rattachent (§.1). Le problème de base est ensuite directement résolu et sa solution qui ne dépend que du triple (E,P_{EE},d_{EE}), apparaît comme une généralisation de l'Analyse Factorielle sur Tableau des Distances (§.2). Des considérations de dualité permettent de compléter cette solution et de replacer celle-ci dans le cadre classique (cf [3]) de l'Analyse en Composantes Principales (A.C.P.). Pour l'interprétation des résultats, les contributions absolues et relatives usuelles sont adaptées au critère optimisé (§.3). Comme application, on montre que l'on peut retrouver, caractériser et généraliser l'Analyse Factorielle des Correspondances (cf [15]), en pondérant des distances très élémentaires (§.4).

1 - PROBLEMES EXTREMAUX ASSOCIES AU TRIPLE (E, P_{EE}, d_{EE})

Avant de définir le sens de la pondération des distances en analyse factorielle, c'est à dire les propriétés extrêmales que cela permet de donner aux axes factoriels, il convient de préciser quelque peu les notations utilisées.

1.1 Notations Utilisées

Dans les applications l'ensemble E, caractérisant le triple (E,P_{EE},d_{EE}), pourra concerner, soit les individus (ou observations), soit les variables (ou caractères), soit les deux simultanément. Bien que les généralisations en dimension infinie soient tout à fait envisageables (cf [7], [12] et [19]), E sera supposé de cardinal q fixé dans la suite.

Soit D la matrice symétrique dont l'élément courant est $d^2_{ee'}$. Celui-ci représente le carré d'une distance euclidienne entre les éléments e et e' de E.

Soit V_p la matrice dont l'élément courant $p_{ee'}$ viendra pondérer $d^2_{ee'}$. En fait $d^2_{ee'}$ est pondéré par $p_{ee'} + p_{e'e}$, ce qui permet de supposer V_p symétrique. On peut lui associer une matrice diagonale D_p, dont l'élément diagonal courant est :

$$P_e = \sum_{e'} p_{ee'} \qquad \forall\, e \in E$$

On ne suppose seulement que $p_e \geqslant 0 \quad \forall \quad e \in E$ et que $\sum_e p_e = 1$
Enfin on sera amené à introduire la matrice

$$A = \begin{bmatrix} p_1 & \cdots & p_e & \cdots & p_q \\ \vdots & & & & \\ p_1 & \cdots & p_e & \cdots & p_q \end{bmatrix}$$

1.2 Le problème de base (pb)

Les distances d_{EE} étant euclidiennes, il est facile de construire une représentation de E dans un espace R^q. Les points $P^1 \ldots P^e \ldots P^q$ représentatifs de E sont placés aux extrémités de la base canonique de R^q (d'origine O). Pour retrouver les distances d_{EE}, il suffit en effet de munir R^q d'une métrique m telle que :

$$m(P^eP^{e'}, P^eP^{e'}) = d^2_{ee'} \qquad \forall \quad e,e' \in E$$

Cela conduit à prendre pour Matrice représentative de m la matrice M^* d'élément courant :

$$M^*_{ee'} = m(OP^e, OP^{e'}) = -\frac{1}{2} d^2_{ee'} \qquad \forall \quad e,e' \in E$$

car $m(P^eP^{e'}, P^eP^{e'}) = m(OP^e, OP^e) + m(OP^{e'}, OP^{e'}) - 2m(OP^e, OP^{e'})$

En pratique pour avoir une vraie métrique, on doit restreindre M^* au support du nuage $\{(P^e,p_e) | e \in E\}$ en lui substituant :

$$M = (Iq-A)\, M^*\, {}^t(Iq-A)$$

où Iq note la matrice identité de dimension q.

Le problème de base de la représentation du triple (E,P_{EE},d_{EE}) est de rechercher un vecteur unitaire $\vec{\varphi}_E$ de R^q rendant le résidu pondéré.

$$(Pb) \quad \begin{cases} R = \sum_e \sum_{e'} p_{ee'} (C^e - C^{e'})^2 & \underline{\text{extrêmal}} \\ C^e \text{ étant la coordonnée de } P^e \text{ sur } \vec{\varphi}_E \\ \text{avec itérations (sur } \vec{\varphi}_E) \text{ sous contraintes d'orthogonalité} \end{cases}$$

1.3 Les problèmes associés et leurs applications

Trois types de problèmes, notés (P_1), (P_2) et (P_3), peuvent conduire à(Pb).

$$(P_1)\quad \begin{cases} \text{chercher le sous-espace de dimension } r \text{ rendant} \\ R_r = \sum_e \sum_{e'} p_{ee'}\, {}_r d^2_{ee'} \qquad \underline{\text{Maximal}} \\ {}_r d_{ee'} \text{ étant la distance projetée associée à } d_{ee'} \end{cases}$$

Dans cette formulation, R_r apparaît comme une généralisation de la notion usuelle d'inertie projetée, les masses p_{EE} étant relatives aux couples de E x E.

a_1)- On retrouve l'ANALYSE EN COMPOSANTES PRINCIPALES usuelle lorsque

$$p_{ee'} = p_e p_{e'} \qquad \forall\ e \in E \qquad \forall\ e' \in E$$

Vérifions le sur la première composante principale :

$$\begin{aligned} R_1 &= \sum_e \sum_{e'} p_e\, p_{e'}\, (c^e - c^{e'})^2 \\ &= 2 \sum_e p_e\, (c^e)^2 - 2(\sum_e p_e\, c^e)^2 \\ &= 2 \sum_e p_e\, (c^e - \sum_e p_e\, c^e)^2 \end{aligned}$$

On rend donc maximal le double de l'inertie projetée au sens usuel du terme.

b_1)- Lorsque E représente des observations spatio-temporelles, un choix judicieux de p_{EE} permet de faire une ANALYSE DES EVOLUTIONS (cf [14] et [16])

$$\begin{aligned} p_{ee'} &= p_e\, p_{e'} \quad \text{si } e \text{ et } e' \text{ sont associés au même site} \\ &= 0 \qquad \text{sinon} \qquad \forall\ e \in E\ ,\ e' \in E. \end{aligned}$$

Dans ce cas, R_r ne prend en compte que l'inertie d'origine temporelle (c'est à dire celle des trajectoires des sites).

c_1)- Plus généralement, si E repère des unités géographiques contigues ou non (par exemple les arrondissements d'une ville), on peut limiter la somme R_1 aux seuls couples d'unités géographiques contigues. On obtient alors ce que L. Lebart appelle une ANALYSE LOCALE (cf [13]).

$$(P_2)\quad \left[\begin{array}{l} \text{chercher le sous-espace de dimension r rendant} \\ \\ R_r = \sum_e \sum_{e'} p_{ee'}\ {}_r d^2_{ee'} \qquad \underline{\text{Minimal}} \end{array}\right.$$

a_2)- Il est clair que l'on retrouve ainsi les dimensions éliminées par l'ANALYSE FACTORIELLE SUR TABLEAU DE DISTANCES lorsque

$$p_{ee'} = p_e\, p_{e'} \qquad \forall\ e \in E,\ e' \in E \qquad \text{(voir } a_1\text{)}$$

On peut alors neutraliser l'effet de certaines distances excessives ou erronées en annulant certains poids p_e ou $p_{ee'}$ dans le résidu R_r.

b_2)- Comme on le démontre au §.4, l'A.F.C. peut être retrouvée, lorsque p_{EE} est proportionnel à un tableau de Burt (tables de contingence juxtaposées) et que d_{EE} s'identife à la distance atomique du χ^2 du centre p_E (cf [15]). Les facteurs extraits tendent alors à rapprocher les couples de E x E ou les k-uplets de ExEx...xE les plus fréquemment associés.

c_2)- Si E contient des modalités qu'il serait bon de trouver ORDONNEES (ou quasi-ordonnées) sur les axes factoriels, il est possible d'approcher la solution rigoureuse (cf [19]) par pondération des distances. Comme on recherche les dimensions rapprochant les couples (e,e') les plus lourds, il suffit de construire une table de contingence p_{EE} représentative du graphe des combinaisons permises de cette structure d'ordre.

$$(P_3)\quad \left[\begin{array}{l} \text{chercher le sous-espace de dimension r rendant} \\ R_r = \sum_e \sum_{e'} p_{ee'}\ {}_r d^2_{ee'} \qquad \underline{\text{Minimal}} \\ I_r = \sum_e \sum_{e'} p_e p_{e'}\ {}_r d^2_{ee'} \qquad \underline{\text{Maximal}} \end{array}\right.$$

Ce problème extrêmal s'identifie au précédent (P_2) et donc à (Pb), lorsque l'on rend la distance d_{EE} à inertie constante pour toutes les directions de R^q (I_r = Cste). Celà est toujours possible en effectuant une AFTD du triple (E,p_E,d°_{EE}) de départ et en substituant à d°_{EE} la distance euclidienne associée à ses facteurs réduits.

a_3)- La distance atomique du χ^2 de centre p_E possédant justement la propriété (I_r = Cste), l'A.F.C. est encore solution du problème (P_3). D'autre part cela suggère certains aménagements de l'A.F.C..lorsque celle-ci est perturbée par des absences intempestives de données (non-réponses à des questions, cases vides). Le centre p_E calculé est estimé de façon à minimiser la perturbation (par exemple sa valeur "théorique" cf [18]). Mais dans certains cas, il faut également estimer les valeurs manquantes (cf [17] par exemple).

b_3)- En partant de la distance Euclidienne classique, on est amené par le procédé évoqué précédemment à lui associer sa distance de Mahalanobis. La solution au problème (P_3) généralise l'ANALYSE DISCRIMINANTE au cas de relations binaires symétriques quelconques (cf [13]).

c_3)- Certains auteurs (cf [1]) proposent d'identifier p_{EE} pour faire une A.C.P. sur des UNITES STATISTIQUES CORRELEES. Dans le cas des processus stochastiques multidimensionnels du second ordre, des modèles probabilistes plus ou moins généraux les conduisent à des solutions assez satisfaisantes mais susceptibles de faire intervenir des valeurs quelconques de p_{EE}.

D'autres formulations et d'autres applications conduisent au problème de base (Pb). Notre propos n'est pas d'être exhaustif en cette matière mais de résoudre ce problème, de replacer sa solution dans un cadre plus classique (cf [3]) et de montrer ses liens avec les représentations graphiques de l'analyse factorielle de correspondances (cf [2]).

2 - RESOLUTION DU PROBLEME DE BASE

Le problème (Pb) est résolu ici dans le cadre des notations générales introduites au § 1.1. Le formalisme de l'analyse factorielle sur tableau de distance est simplement retrouvé par ce biais. L'existence et l'unicité de la solution font l'objet du dernier paragraphe.

2.1 Résolution formelle de (Pb)

Soient $\{\varphi_e | e \in E\}$ les composantes de $\vec{\varphi}_E$. Comme C^e est la projection de $\vec{OP}_e$ sur $\vec{\varphi}_E$ au sens de la métrique m, il est clair que

$$C^e = m(OP^e, \vec{\varphi}_E) = m(OP^e, \sum_{e'} \varphi_{e'}, OP^{e'})$$

$$C^e = \sum_{e'} M_{ee'} \varphi_{e'} \quad \forall \ e \in E \quad \Leftrightarrow C = M\varphi \qquad (1)$$

où C et φ sont les vecteurs colonnes associés à C^E et φ_E.

Il s'agit de rechercher φ unitaire $\quad {}^t\varphi M\varphi = 1 \qquad (2)$

rendant $R = \sum_e \sum_{e'} p_{ee'} (C^e - C^{e'})^2$ extrêmal. Avec l'expression matricielle

$$R = 2\ {}^tC\ (Dp - Vp)\ C = 2\,{}^t\varphi M(Dp - Vp)M\varphi$$

on reconnaît un problème extrêmal classique en ACP : rendre $R = {}^t\varphi M(Dp-Vp)M\varphi$ extrêmal avec ${}^t\varphi M\varphi = 1$. On sait qu'alors $\exists\ \mu$ extrêmal, tel que :

$$(Dp - Vp)M\varphi = \mu\varphi$$

$$\Leftrightarrow \quad M(Dp - Vp)C = \mu C \qquad (3)$$

La matrice M, qui peut être supposée semi-définie positive, admet une décomposition de la forme :

$$M = {}^tYY$$

Par exemple la décomposition de Cholewski (si M est définie positive) ou la décomposition propre suivante. Soit $(U_1 \ \ldots \ U_q)$ les vecteurs propres unitaires de M respectivement associés aux valeurs propres $(\lambda_1, \ldots, \lambda_q)$. Comme $M = {}^tUD_\lambda U$, où D_λ est la forme diagonalisée de M, on peut identifier

$$Y_i = \sqrt{\lambda_i}\ U_i \qquad \forall\ i = 1, q$$

et limiter Y aux seuls vecteurs propres de valeurs propres strictement positives (c'est à dire aux s premiers).

(1) et (3) $\Rightarrow$ ${}^tY\ Y\ (D_p - V_p)\ {}^tY\ Y\ \varphi = \mu\ {}^tY\ Y\ \varphi$

Posons $v = Y\varphi$ et simplifions par ${}^tY$

$$\Rightarrow \begin{cases} Y(D_p - V_p)\,{}^tY\ v = \mu\ v & (4) \\ {}^tv\ v = 1 & (2') \end{cases}$$

$$(1) \quad \Rightarrow \quad C = {}^tY\ v \qquad (1')$$

Ces relations montrent que la solution C cherchée s'obtient à partir des éléments propres de la matrice $Y(D_p - V_p)\,{}^tY$. Il s'agit des plus grandes valeurs propres pour (P_1) et des plus petites pour (P_2) et (P_3).

Remarque : Les composantes principales C ainsi trouvées sont D_p-centrées, $(D_p - V_p)$-orthogonales et de $(D_p - V_p)$-norme $\sqrt{\mu}$. N'étant pas en général D_p-orthogonales, elles peuvent être correlées comme cela est préconisé par certains auteurs (cf[20] et [21]).

2.2 Cas de l'analyse factorielle sur tableau de distances

Ce cas correspond à : $p_{ee'} = p_e p_{e'} \qquad \forall\ e \in E,\ e' \in E$

Cette forme particulière de V_p permet d'éviter la décomposition de M

Comme $D_p - V_p = (I_q - A)D_p$, l'équation (3) s'écrit

$$(I_q - A)\ (M)\ D_p\ C = \mu C$$

$$(I_q - A)(I_q - A)(-\tfrac{1}{2} D)\,{}^t(I_q - A)D_p C = \mu\ C$$

Comme $(I_q - A)(I_q - A) = I_q - A$ puisque $AA = A$

$$(3) \iff M\ D_p\ C = \mu C$$

(3) prend alors la forme classique :

$$W\, D_p\, C = \mu\, C \qquad (3')$$

où W est déduit de D par le double recentrage habituel.

2.3 Existence et Unicité de la solution

L'existence est assurée dès que M est semi-définie positive (distances euclidiennes). En fait il suffit que $M(D_p - V_p)$ soit semi-définie positive, ce qui est moins contraignant. En effet la positivité au sens large de μ est requise, puisque :

$$R_1 = \sum_e \sum_{e'} p_{ee'} (C^e - C^{e'})^2 \geqslant 0$$

$$= \sum_e \sum_{e'} p_{ee'} (C^e - C^{e'}) \sum_{e''} (M_{ee''} - M_{e'e''})\, \varphi_{e''}$$

$$= \sum_e \mu\, C^e\, \varphi_e = \mu \geqslant 0$$

Si d_{EE} était issue d'un tableau de dissimilarité, il apparaittrait, en suivant un raisonnement parallèle à celui de [3] ou [17], que le formalisme reste acceptable pour les valeurs propres positives. Cependant cette façon de procéder n'exclut pas les anomalies (cf [2] p.86-88).

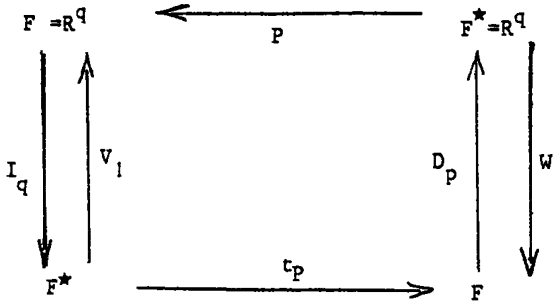
L'unicité de la solution est issue du fait que les équations (4), (2') et (1'), qui permettent de la calculer, ne dépendent que du triple (E, p_{EE}, d_{EE}). Il est clair que cette unicité reste relative au signe de C ou φ et au choix de C ou φ dans les sous-espaces de dimension t, lorsque les valeurs propres associées sont de multiplicité t.

Remarque : Pour le problème (P_2), il y a lieu d'extraire dans certains cas (comme l'A.F.C.) les seules dimensions associées à des valeurs propres strictement positives.

3 - SCHEMAS DE DUALITE ET MISE EN OEUVRE PRATIQUE

Le formalisme précédent s'interprète avantageusement en terme d'Analyse en Composantes Principales au sens usuel. Cette identification et la construction des schémas de dualité (cf [3]) associés précisent complètement cette parenté, y compris lorsqu'on part d'un tableau de données. La mise en oeuvre de l'A.C.P. PONDEREE est donc conforme aux pratiques habituelles. Toutefois le rôle actif des pondérations peut être précisé grâce à de nouveaux indicateurs d'aide à l'interprétation adapté au critère optimisé.

3.1 Le schéma de dualité de l'A.F.T.D.



Le formalisme de l'Analyse Factorielle sur Tableau de Distances consiste à calculer les éléments propres de WD_p. Si P est la matrice des composantes principales on doit avoir :

$V_1 = P\, D_p\, {}^tP$.

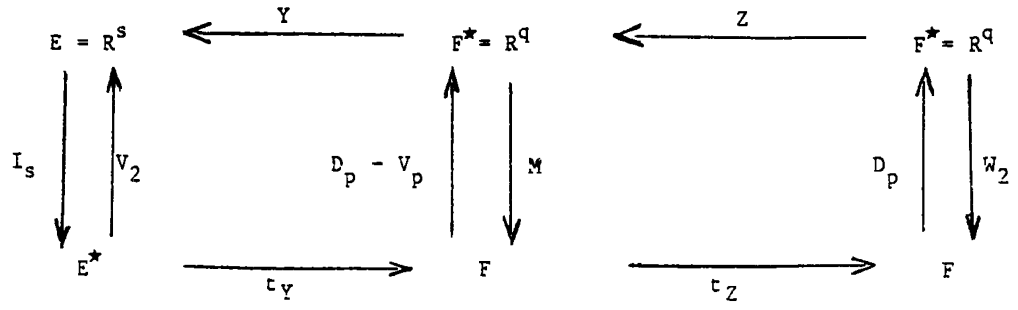
L'ensemble de ces relations peut être résumé sous la forme du schéma de dualité ci-dessus. Pour plus de détails, on peut consulter [3].

3.2 Le schéma de dualité de l'A.F.T.D. Pondérée

$(I_q - V_p\, D_p^{-1})$ étant semi-définie positive (cf la relation (5) du §.4), il existe une matrice Z telle que :

$$D_p - V_p = Z\, D_p\, {}^tZ \qquad \Rightarrow \qquad V_2 = YZ\, D_p\, {}^tZ\, {}^tY$$

cela permet de reconnaître le formalisme de l'A.C.P.

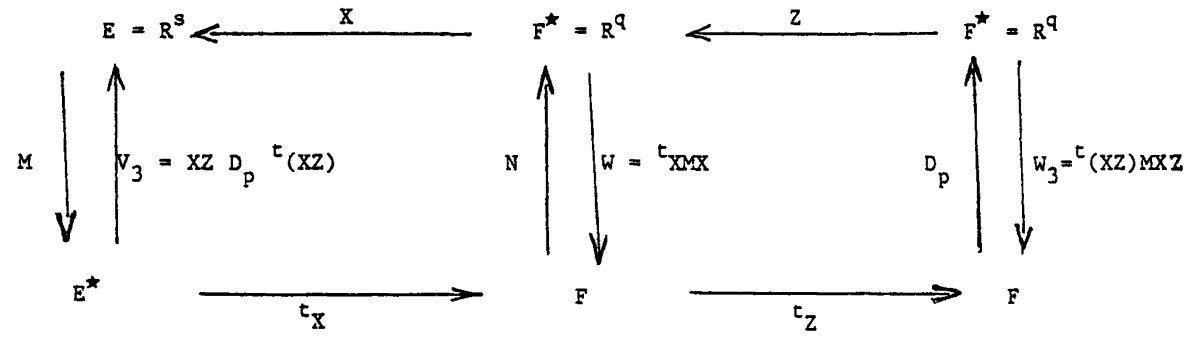


où $W_2 = {}^t(YZ)\ YZ$ et I_s est une matrice identité de dimension s.

Remarque : L'espace F est donc muni d'une métrique $N = D_p - V_p$ telle que les distances associées respectent la correspondance V_p. On verra au §.4 que les facteurs de l'AFC de V_p constituent le tableau Z.

3.3 Le schéma de dualité de l'A.C.P. Pondérée

Lorsque les distances d_{EE} sont calculées à partir d'un tableau de données X (q lignes, s colonnes), R^s est muni d'une métrique M. En suivant une démarche parallèle à celle du §. 2.1, on montre que pondérer les distances par V_p (et non par AD_p) revient à munir R^q de la métrique $N = D_p - V_p$ (au lieu de D_p). En posant $N = Z\ D_p\ {}^tZ$, on retrouve le schéma de dualité classique de l'A.C.P.



3.4 Mise en oeuvre pratique

En pondérant les distances, on reste donc dans le cadre de l'A.C.P., mais avec une métrique $N = D_p - V_p$. Les facteurs obtenus sont tout à fait comparables à ceux que l'on a avec la métrique D_p usuelle (cf remarques des § 2.1 et 3.2). Pour les interpréter, on doit adapter les classiques contributions absolues et relatives à la forme quadratique, que l'on rend extrêmale. On suggère d'appeler

- Pourcentage d'inertie pondérée de l'axe a :

$$PR_a = \frac{\sum_e \sum_{e'} p_{ee'} (C_a^e - C_a^{e'})^2}{\sum_e \sum_{e'} p_{ee'} d^2_{ee'}} = \frac{\mu a}{\sum_a \mu a}$$

- Contribution Absolue du couple (e,e') à la détermination de l'axe a :

$$CA_a\ (e,e')\ \% = 100\ p_{ee'}\ (c_a^e - c_a^{e'})^2 \mu_a^{-1}$$

$$\Rightarrow \qquad CA_a(e)\ \% = \sum_{e'} CA_a\ (e,e')$$

- Contribution Relative du couple (e,e') à la détermination de l'axe a :

$$CR_a\ (e,e')\ \% = 100\ (c_a^e - c_a^{e'})^2 d_{ee'}^{-2}$$

$$\Rightarrow \qquad CR_a\ (e)\ \% = \sum_{e'} CR_a\ (e,e')$$

Ces éléments remplacent ou complètent les éléments de même nom que l'on pourrait calculer avec les formules classiques en substituant à X le tableau XZ.

4 - APPLICATION A L'ANALYSE DES CORRESPONDANCES

Lorsque p_{EE} est, à un facteur multiplicatif près, un tableau de Burt, il est possible de retrouver tout le formalisme de l'A.F.C. à partir d'un triple (E, p_{EE}, d_{EE}) très simple. Cette approche, outre qu'elle ouvre la voie à des généralisations de l'A.F.C. (cf § 1.3 a3), attire l'attention sur une propriété graphique remarquable de l'analyse des correspondances multiples. L'intérêt pratique d'une telle propriété apparaît notamment, lorsqu'on cherche à construire des représentations graphiques optimales des matrices d'importations-exportations, éventuellement ventilées par produits (cf [17]).

4.1 Le triple associé à l'A.F.C.

Il s'agit de choisir d_{EE}, de telle sorte que l'équation (3) prenne la forme usuelle suivante :

$$(D_p)^{-1}\ V_p\ C = \lambda\ C \qquad (5)$$

Par identification, on est conduit à avoir :

$$M = (D_p)^{-1} \quad \text{et} \quad \mu = 1 - \lambda \qquad (R)$$

or $\qquad d_{ee'}^2 = m(P^e P^{e'},\ P^e P^{e'}) \qquad \forall\ e', e' \in E$

$$= M_{ee} + M_{e'e'} - 2\ M_{ee'}$$

$V_p\ C = \lambda D_p C = (\lambda - 1 + 1) D_p C$

$(D_p - V_p) C = (1-\lambda) D_p C \Rightarrow (D_p)^{-1} (D_p - V_p) C = (1-\lambda) C$

$$\Rightarrow \quad d^2_{ee'} = \begin{cases} 0 & \text{si } e = e' \in E \\ (p_e)^{-1} + (p_{e'})^{-1} & \text{si } e \neq e' \in E \end{cases}$$

Cette distance peut être appelée "distance atomique du χ^2 ", car elle correspond à la distance du χ^2 pour des éléments de E supposés disjoints comme des atomes. On peut le vérifier en appliquant la distance du χ^2 au tableau D_p (forme diagonale du tableau V_p. cf § 1.1).

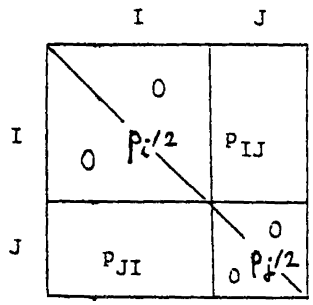
En pondérant ces distances atomiques par les pondérations V_p, on est donc conduit aux facteurs de l'A.F.C. Il est aisé de démontrer que cette relation (5) permet de retrouver tout le formalisme de l'analyse des correspondances et d'extraire les facteurs dans l'ordre usuel (cf [15]).

4.2 Interprétation pour une correspondance binaire

Soit k_{IJ} une correspondance binaire entre deux ensembles I et J ; k_{ij} repère par exemple le nombre d'observations associées aux éléments i et j.

Si E est défini comme la réunion de I et J, on construit p_{EE} comme ci-contre avec $\forall$ i et j

$$p_{ij} = p_{ji} = k_{ij} \Big/ 4 \sum_i \sum_j k_{ij}$$



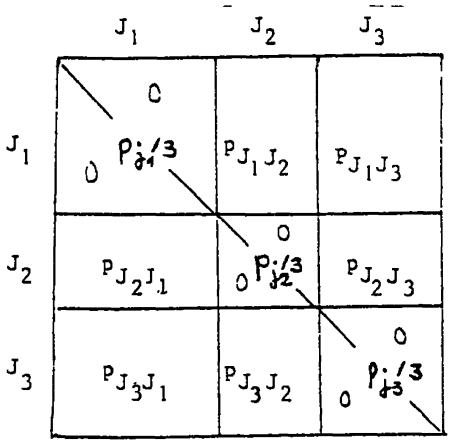
Il est simple de vérifier que (5) équivaut aux équations de transition de l'A.F.C.. Essayons d'interpréter la propriété extrêmale du premier facteur :

$$\text{Min } R_l = \sum_e \sum_{e'} p_{ee'} \, (c_l^e - c_l^{e'})^2$$

$$= \sum_i \sum_j k_{ij} \, (c_l^i - c_l^j)^2 \Big/ 2 \sum_i \sum_j k_{ij}$$

Ainsi le premier facteur tend à rapprocher c'est à dire à METTRE EN CORRESPONDANCE GRAPHIQUE les éléments i et j les plus fréquemment associés.

4.3 Interprétation pour une correspondance multiple



La propriété de mise en correspondance graphique se généralise dans le cas d'une correspondance multiple entre plusieurs ensembles :
$E = J_1 + J_2 + J_3$ par exemple. p_{EE} a la forme ci-contre où $p_{j_1j_2}$ est proportionnel au nombre d'observations associées à j_1 et j_2.

La véritable propriété graphique de l'A.F.C. du tableau p_{EE} se généralise de la façon suivante. Soit $p_{J_1J_2J_3}$ la matrice à trois entrées, dont l'élément courant $p_{j_1j_2j_3}$ est proportionnel au nombre d'observations associées à j_1, j_2 et j_3.

Il est aisé de montrer : $R_1 = \sum_e \sum_{e'} p_{ee'} (C_1^e - C_1^{e'})^2$ Minimal

$$\Leftrightarrow \quad R_1' = \sum_{j_1} \sum_{j_2} \sum_{j_3} p_{j_1j_2j_3} \; D^2(C_1^{j_1}, C_1^{j_2}, C_1^{j_3}) \qquad \text{Minimal}$$

où D^2 note l'écart quadratique moyen associé.

Cela est lié au fait que celui-ci ne dépend que des couples $(C_1^e - C_1^{e'})^2$ et que p_{EE} est déduit de $p_{J_1J_2J_3}$ par sommation.

Autrement dit les facteurs de l'Analyse des Correspondances multiples tendent donc à rapprocher les éléments e, e', e" ... les plus associés ($p_{ee'e''\cdots}$ grands).

5 - CONCLUSION

La pondération des distances en analyse factorielle conduit donc à une A.C.P. dans laquelle la métrique $N = D_p - V_p$ remplace la classique métrique D_p. Les solutions associées ne dépendent que du triple (E, d_{EE}, p_{EE}).

- Si $p_{ee'} = p_e p_{e'}$ $\forall (e,e') \in E \times E$, on retrouve la forme usuelle de l'A.C.P. ou de l'A.F.T.D.
- Si $d^2_{ee'} = (p_e)^{-1} + (p_{e'})^{-1}$ $\forall e \neq e'$ et 0 sinon, on obtient l'A.F.C. du tableau V_p représentatif de p_{EE}.

Dans le cas général, cette technique conduit à des méthodes intermédiaires permettant de rendre les facteurs peu SENSIBLES aux données manquantes ou erronées (distances excessives par exemple), COMPATIBLES avec certaines contraintes d'ordre ou encore ADAPTES aux conditions d'observation (élimination d'effets statistiques, renforcement de certaines structures ...etc). Cette approche opère un filtrage de l'A.C.P. associée au triple (E, d_{EE}, p_E) par les facteurs Z de la correspondance V_p. Aussi est-il souhaitable de réaliser cette A.C.P. et/ou cette A.F.C. de base à titre préliminaire. Les calculs de l'A.C.P. pondérée en seront simplifiés (cf § 3.3) et leurs résultats seront plus aisément contrôlables et interprétables(cf § 3.4).

Formellement le choix de la métrique N, auquel on aboutit, apparaît comme une solution duale de celles habituellement associées à l'A.C.P. sous contraintes (cf [12],[19]et [21]) ou aux visualisations de variances résiduelles (cf [6],[9] et [11]). Quant à la notion de mise en correspondance graphique, elle résume bien la propriété graphique de l'A.F.C. et de ses généralisations.

REFERENCES BIBLIOGRAPHIQUES

[1] ARAGON Y. et CAUSSINUS H. 1978
"Une analyse en composantes principales pour des unités statistiques correlées". Data Analysis and Informatics.
IRIA - VERSAILLES.

[2] BENZECRI J.P. 1973
"L'analyse des données" (2 tomes). Dunod - PARIS.

[3] CAILLIEZ F. et PAGES J.P. 1976
"Introduction à l'analyse des données" SMASH, PARIS.

[4] CROQUETTE A., GRAU D., HAIT J.R, SCHEKTMAN Y. 1980
"Proposition d'une métrique à effet relationnel : Propriétés et application au cas des analyses en composantes principales sous contraintes linéaires".
Les cahiers du CERO, Vol 22 n° 2, p. 193-199 - BRUXELLES

[5] CHEN C.W. 1974
"An optimal property of principal component"
Communications in statistics 3, p. 979-983.

[6] DAUDIN J.J. 1981
"Analyse factorielle des dépendances partielles"
Revue de Statistique Appliquée (A paraître) PARIS.

[7] DAUXOIS J. et POUSSE A. 1976
"Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique".
Thèse d'Etat - Université de TOULOUSE.

[8] ESCOFIER B. 1981
"Analyse des questionnaires avec des non-réponses"
Journées de Statistiques de l'A.S.U., NANCY.

[9] ESCOUFIER Y. et L'HERMIER DES PLANTES H. 1978
"A propos de la comparaison graphique des matrices de variance"
Biom. J 20 (5) p. 491-497.

[10] HOLLMAN E.W. 1972
"The relation between hierarchical and euclidean models for psychological distances"
Psychometrika, Vol 37, n° 4, p. 472-486.

[11] KOBILINSKI A. 1980
"Décomposition de formes quadratiques en analyse des données"
Thèse de 3è Cycle. Université de PARIS-SUD.

[12] LAFAYE DE MICHEAUX D. 1978
"Approximation d'analyses canoniques non linéaires de variables aléatoires et analyses factorielles privilégiantes".
Thèse de Docteur-Ingénieur, NICE.

[13] LEBART L. 1973
"Recherche sur la description automatique des données socio-économiques"
Rapport CORDES - CREDOC.

[14] LE FOLL Y. 1972
"L'analyse factorielle des évolutions"
Actes du petit séminaire sur l'analyse factorielle, n° 2, p. 23-27
Université des Sciences Sociales de GRENOBLE.

[15] LE FOLL Y. 1979
"Sur les propriétés de l'analyse des correspondances pour diverses formes complètes de données"
Thèse 3è Cycle. Université P, et M. Curie. PARIS.

[16] LE FOLL Y. 1980
"The use of multidimensional analysis in comparaison of the 1976 VS 1971 river Seine basin survey". Prog. Wat. Tech. Vol 12, p.1011-1033, LONDRES.

[17] LE FOLL Y. et BURTSCHY B. 1982
"Représentations optimales des matrices imports-exports"
Revue de Statistique Appliquée (à paraître)

[18] MAILLES J.P. 1971
"Analyse factorielle des tableaux de dissimilarités"
Thèse de 3è Cycle. Université P. et M. Curie.

[19] MASSON M. 1980
"Méthodologies générales de traitement statistique de l'information de masse". Ed. Cedic - F. Nathan - PARIS.

[20] RAO C.R. 1964
"The use and interpretation of principal component analysis in applied research". Sankhia, Ser. A, 26, p. 329-359.

[21] SCHEKTMAN Y. 1978
"Contribution à la mesure en facteurs dans les sciences expérimentales et à la mise en oeuvre des calculs statistiques"
Thèse d'Etat - Université Paul Sabatier - TOULOUSE.