Fontainebleau N-292

INTRODUCTION A L'ANALYSE FACTORIELLE

J.P. ORFEUIL

Juin 1972

Centre de Morphologie Mathématique
Ecole des Mines de Paris

S O M M A I R E

I - PRESENTATION GENERALE

La situation la plus classique en Analyse Factorielle est la suivante : on dispose d'un grand nombre d'individus, qu'on peut répartir en n classes ; on dispose par ailleurs d'un ensemble J de propriété (Card. J = p) ; on définit alors un tableau de correspondance K(n × p), dont le terme général k(i,j) représente le nombre d'individus de la classe i qui ont la propriété j : par exemple si l'ensemble des individus est l'ensemble des français, si les classes sont les départements, les propriétés les votes au dernier référendum, k(i,j) désignera par exemple le nombre de corrèziens qui ont voté non au dernier référendum.

Posons $k(i) = \sum_{j=i}^{j=p} k(i,j)$. Alors la suite $\{f_j^i = \frac{k(i,j)}{k(i)}, j=1,..p\}$ peut être considérée comme une estimation de la loi de probabilité de j, conditionnelle à i ; Le test du χ^2 permet en principe de confirmer ou d'infirmer l'hypothèse que la loi de j est indépendante de i. L'objet de l'Analyse factorielle est, en gros, de préciser la nature de cette dépendance, lorsqu'elle existe, de préciser les proximités entre éléments de I (ou de J), et même, dans certains cas, de mettre en évidence des proximités entre éléments de I et de J.

L'idée est la suivante : on considère les individus comme des éléments d'un espace vectoriel R^p, la coordonnée de l'individu i sur l'élément e_j de la base orthonormée canonique de R^p étant f_j^i. On munit R^p d'une métrique adaptée au caractère probabiliste du problème, et les classes d'individus de poids significatifs de leur importance. Il est alors naturel de chercher le centre de gravité et les directions propres d'inertie du nuage de points ainsi construit. La variété linéaire d'ordre k qui approxime au mieux le nuage est la variété engendrée par les k directions propres dont l'inertie est la plus grande. Si l'inertie du nuage par

rapport au plan des deux premiers axes représente une part importante de l'inertie totale, on estime que la projection du nuage sur ce plan est une bonne approximation du nuage lui-même. Sinon, il conviendra de considérer les plans des axes 1-3, 1-4, 1-5, ...2-3, 2-4 ... pour disposer d'une information plus grande.

II - LE CHOIX DE LA DISTANCE ET DES POIDS

1/ LE CHOIX DE LA DISTANCE

Commençons par préciser quelques notations ; soit $\{k(i,j),\ i = 1,\ n,\ j = 1,\ p\}$ un tableau de correspondance sur $I \times J$. On définit :

$$k(i) = \sum_{j=i}^{j=p} k(i,j)$$

$$k(j) = \sum_{i=1}^{i=n} k(i,j)$$

$$k = \sum_{j=1}^{j=p} \sum_{i=1}^{n} k(i,j) = \sum_{i=1}^{n} k(i) = \sum_{j=1}^{j=p} k(j)$$

On définit alors $\{f_{ij}\}$, loi de probabilité sur $I \times J$, $\{f_i^j\}$ loi de probabilité sur I, conditionnelle et j, $\{f_j^i\}$, loi sur J conditionnelle à i, par :

$$\{f_{ij} = \frac{k(i,j)}{k},\ i = 1,..n,\ j = 1,..p\}$$

$$\{f_i^j = \frac{k(i,j)}{k(j)},\ i = 1,..,n\}$$

$$\{f_j^i = \frac{k(i,j)}{k(i)},\ j = 1,..,p\}$$

On définit enfin les lois a priori de i et de j par :

$$\{f_i = \frac{k(i)}{k},\ i = 1,...n\}$$

$$\{f_j = \frac{k(j)}{k},\ j = 1,...p\}$$

Pour tester si la loi f_i^j diffère significativement de la loi f_i, on définit la quantité :

$$\gamma^2 = \sum_{i=1}^{i=n} \frac{1}{f_i} (f_i^j - f_i)^2$$

Cette quantité peut être interprétée comme le carré d'une distance entre les 2 lois de probabilité. On définira de même :

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_i} (f_i^j - f_i^{j'})^2$$

$$d^2(i,i') = \sum_{j=1}^{j=p} \frac{1}{f_j} (f_j^i - f_j^{i'})^2$$

C'est cette distance qui est utilisée en Analyse des correspondances, en raison de son caractère probabiliste d'une part, et d'autre part parce qu'elle satisfait au principe d'équivalence distributionnelle, que nous préciserons ultérieurement.

B/ LE CHOIX DES POIDS

En Analyse des correspondances, le choix des poids est parfois imposé par la structure même du tableau. Par exemple, il est naturel d'affecter à chaque département un poids proportionnel à sa population. (C'est-à-dire à la somme des termes de la ligne correspondante), et à chaque option un poids proportionnel au nombre de voix qu'elle a obtenu (poids proportionnel à sa colonne). Il est clair qu'il en est en fait ainsi chaque fois qu'on a effectivement affaire à un véritable tableau de fréquence sur $I \times J$.

En fait les protagonistes de l'Analyse Factorielle ont des ambitions beaucoup plus grandes, et ils entendent extraîre la "substantifique moelle" de tout tableau de nombre positifs, pourvu que la signification des individus, des objets et des nombres $k(i,j)$ leur soit donnée.

Une première extension de la méthode a concerné les tableaux de description logique, c'est-à-dire tels que $k(i,j) \in \{0,1\}$; L'analyse des questionnaires est à l'origine de cette préoccupation. On attribue la valeur 1 à $k(i,j)$ si l'individu i a répondu oui à la question j, et la valeur 0 dans le cas contraire. Il est clair qu'un individu qui répond plus souvent "oui" que "non" aura, si l'on s'en tient à la pondération précédente, un poids plus important qu'un individu qui répond plus souvent "non" que "oui". On peut tourner la difficulté en dédoublant le tableau des questions, c'est-à-dire en adjoignant à toute question j une question $\overline{j}$ égale à "non-j" (tout individu qui répond oui à j répond non à $\overline{j}$). On peut de même dédoubler les individus si on souhaite analyser les questions en leur accordant un poids commun :

Il est clair qu'en effet, après cette transformation du tableau qui ne créé pas d'information, les lignes (respectivement les colonnes) ont même poids.

Une situation un peu plus générale est la suivante : l'ensemble J est **partitionné en classes q_1, $q_2 \ldots, q_n$ telles que** $"(\forall i \; \forall j \; \exists ! \; k \in qj : k(i,k)=1) \wedge (\forall l ; \; l \neq k \;\; l \in qj \Rightarrow k(i,l) = 0)$. En clair, $k(i,j) = 0$, sauf pour un j et un seul dans chaque classe ; on peut montrer qu'alors, les facteurs sur $J \cup \overline{J}$ (tableau dédoublé) issus de cette correspondance sont les mêmes que ceux issus de l'analyse du tableau sur $J \times J$ défini par :

$$k(j,j') = \text{Card}\{i | k(i,j) = k(i,j') = 1\}$$

Un exercice de style a été réalisé par L. LEBART, à propos des tableaux logiques, sur les départements français. Il a posé $k(i,i') = 1$ si les départements i et i' ont une frontière commune, $k(i,i') = 0$ dans le cas contraire (en regroupant les départements de la région parisienne en un seul, pour des raisons d'homogénéité). La projection du nuage sur le plan des deux axes principaux reconstitue bien la carte de la France, à une affinité près.

Une autre extension de l'Analyse répond à des besoins plus "scientifiques". Un ensemble I d'individus est caractérisé par un ensemble J de mesures. Soit k(i,j) le résultat de la mesure J sur i (on suppose toujours qu'il s'agit de nombres positifs). On suppose qu'il existe une partition Q de J tel que, pour tout j appartenant à un élément q donné de la partition, la mesure s'exprime dans la même unité. Par exemple, on caractérise une lame mince successivement par sa covariance, sa surface spécifique, son étoile, ses caractéristiques pétrophysiques, etc... Il est naturel de laisser invariant les profils des individus i à l'intérieur d'un élément q de la partition, mais le profil global dépend largement du choix des unités de mesure J.P. BENZECRI propose la solution suivante : on se donne appriori un système de masses sur la partition, soit : $\{a(q), q \in Q\}$ qui définit l'importance qu'on souhaite attribuer à chaque type de mesure. On cherche s'il existe un système de masses $\{\lambda(q), q \in Q\}$, tel que si $\lambda(j) = \lambda(q)$ pour tout j appartenant à q, la contribution de q

à l'analyse du tableau $\{k(i,j)\lambda(j)\}$ soit proportionnelle à a(q). Le terme de contribution sera explicité plus tard. Disons simplement qu'il s'agit du produit de la masse par le carré de la distance au centre de gravité). On se reportera avec profit à la notr de J.P. BENZECRI "Sur le choix des unités et des poids dans un tableau en vue d'une Analyse de Correspondance" pour plus de détails sur cette question.

C/ L'EQUIVALENCE DISTRIBUTIONNELLE

Cette propriété peut s'énoncer ainsi : Supposons que 2 points, i_1 et i_2, de masses f_{i_1} et f_{i_2} coïncident dans R^p. Alors on ne change pas l'Analyse en remplaçant ces points par un point i_o confondu avec les 2 premiers, de masse $(f_{i_1} + f_{i_2}) = f_{i_o}$

Les hypothèses sont les suivantes :

$$\forall j \in J \quad \frac{f_{i_1}{}^j}{f_{i_1}} = \frac{f_{i_2}{}^j}{f_{i_2}}$$

$$\forall j \in J \quad f_{i_o}{}^j = f_{i_1}{}^j + f_{i_2}{}^j$$

$$f_{i_o} = f_{i_1} + f_{i_2}$$

On note que $\{f_j, j \in J\}$ reste inchangée.

1/ Invariance dans R^p

$$d^2(i,i') = \sum_{j=1}^{j=p} \frac{1}{f_j} (f_j^i - f_j^{i'})^2$$

Si $i \neq i_o$ et $i' \neq i_o$, la distance est bien entendu conservée

Si $i \neq i_o$ et $i' = i_o$, $f_j^{i'} = f_j^{i_o} = f_j^{i_1} = f_j^{i_2}$, la distance est conservée.

2/ Invariance dans R^n

$$d^2(j,j') = \sum_{\substack{i=1 \\ i\neq i_1 \\ i\neq i_2}}^{i=n} \frac{1}{f_i} (f_i^j - f_i^{j'})^2 + \frac{1}{f_{i_1}} (f_{i_1}^j - f_{i_2}^{j'}) + \frac{1}{f_{i_2}} (f_{i_2}^j - f_{i_2}^{j'})^2$$

Posons $A_k = \frac{1}{f_{i_k}} \left(\frac{f_{i_k j}}{f_j} - \frac{f_{i_k j'}}{f_{j'}}\right)^2$ $\quad k = 0, 1, 2$

$$A_k = f_{i_k} \left(\frac{f_{i_k j}}{f_{i_k} f_j} - \frac{f_{i_k j'}}{f_{i_k} f_{j'}}\right)^2 \qquad k = 0, 1, 2$$

La quantité entre parenthèses ne dépend pas de k et comme $f_{i_o} = f_{i_1} + f_{i_2}$, on déduit $A_o = A_1 + A_2$. Donc les distances dans R^n sont également inchangées. Cette propriété de la distance du χ^2 confère à l'analyse des correspondances une certaine stabilité : regrouper les votes en faveur de Rocard et Krivine ne changera pas notablement la physionomie générale du scrutin, de même regrouper cyrrhose et alcoolisme dans une analyse des causes de décès.

Inversement, deux individus qui, à l'Analyse s'avèrent très proches apportent sensiblement la même information. On pourra n'en retenir qu'un dans une étude ultérieure. C'est ce que suggère Monsieur ROUX dans son étude sur la Piezométrie ("Incidence de l'Analyse des Correspondances sur l'optimisation d'un réseau piezométrique". Laboratoire d'Hydrogéologie Mathématique Fontainebleau).

III - LA RESOLUTION DU PROBLEME DANS R^p

La distance entre 2 éléments i et i' est : 0

$$d^2(i,i') = \sum_{j=1}^{j=p} \frac{1}{f_j} \left(f_j^i - f_j^{i'}\right)^2$$

Cette distance n'est malheureusement pas une somme de carrés de telle sorte que le formalisme habituel de la mécanique ne s'applique pas immédiatement. Il suffit, pour se ramener au cas classique, de changer l'échelle de chacun des axes, en prenant pour nouvelles coordonnées de i :

$$\left\{\frac{f_{ij}}{f_i\sqrt{f_j}}, \quad j = 1,p\right\}$$

Alors la distance entre i et i' est à nouveau mise sous forme d'une somme de carrés :

$$d'(i,i') = \sum_{j=1}^{j=p} \left(\frac{f_{ij}}{f_i\sqrt{f_j}} - \frac{f_{ij}}{f_i\sqrt{f_j}}\right)^2$$

Il faudra bien entendu se souvenir le moment venu de cette transformation. Le nuage, qui était précédemment situé dans l'hyperplan des lois de probabilité $\left(\sum_{j=1}^{j=p} f_j^i = 1\right)$, se trouve maintenant dans l'hyperplan $\sum_{j=1}^{j=p} \sqrt{f_j}\, x_j = 1$. Ce plan est orthogonal au vecteur u de composant $u_j = \sqrt{f_j}$, qui constitue une direction propre du nuage, dont l'inertie est nulle. Le centre de gravité du nuage a pour coordonnée $\left\{x_j = \sum_{i=1}^{i=n} f_i \frac{f_{ij}}{f_i\sqrt{f_j}} = \sqrt{f_j}\right.$

La matrice d'inertie du nuage, rapporté à son centre de gravité et à la base définie plus haut, a pour terme général :

$$v_{jj'} = \sum_{i=1}^{i=n} f_i \left(\frac{f_{ij}}{f_i\sqrt{f_j}} - \sqrt{f_j}\right)\left(\frac{f_{ij'}}{f_i\sqrt{f_{j'}}} - \sqrt{f_{j'}}\right)$$

Développons ce terme, il vient (développement classique d'Huyghens)

$$v_{jj'} = \sum_{i=1}^{i=n} \frac{f_{ij}\, f_{ij'}}{f_i \sqrt{f_j\, f_{j'}}} - \sum_{i=1}^{i=n} \frac{f_{ij}\sqrt{f_{j'}}}{\sqrt{f_j}} - \sum_{i=1}^{i=n} f_{ij'} \frac{\sqrt{f_j}}{\sqrt{f_{j'}}} + \sum_{i=1}^{i=n} f_i \sqrt{f_j\, f_{j'}}$$

Comme $\sum_{i=1}^{i=n} f_{ij} = f_i$, $\Sigma f_i = 1$

$$v_{jj'} = \sum_{i=1}^{i=n} \frac{f_{ij}\, f_{ij'}}{f_i\, f_j\, f_{j'}} - \sqrt{f_j \times f_{j'}}$$

Les vecteurs u_α cherchés sont solution du système :

$$\sum_{k=1}^{k=p} v_{jk}\, u_\alpha^k = \lambda\, u_\alpha^j$$

On remarque que, comme prévu, le vecteur $u_o(\sqrt{f_j},\ j=1,\ p)$ est solution du système pour $\lambda = 0$.

En effet :

$$\sum_{k=1}^{k=p} v_{jk} \sqrt{f_k} = \sum_{k=1}^{k=p} \sum_{i=1}^{i=n} \frac{f_{ij}\, f_{ik}}{f_i \sqrt{f_j\, f_k}} \sqrt{f_k} - \sum_{k=1}^{k=p} \sqrt{f_j}\, f_k$$

$$= \sum_{i=1}^{i=n} \frac{f_{ij}}{f_i \sqrt{f_j}} \sum_{k=1}^{k=p} f_{ik} - \sqrt{f_j}$$

$$= \sum_{i=1}^{i=n} \frac{f_{ij}}{\sqrt{f_j}} - \sqrt{f_j} = 0.$$

Montrons que tout vecteur propre u_α de V (différent de u_o) est vecteur propre de W, définie par :

$$W_{jj'} = \sum_{i=1}^{i=n} \frac{f_{ij}\, f_{ij'}}{f_i \sqrt{f_j\, f_{j'}}}$$

En effet, les vecteurs propres d'une matrice symétrique forment un système orthogonal. On a donc :

$$\sum_{j=1}^{j=p} u_\alpha^j \sqrt{f_j} = 0$$

Il vérifie :

$$\forall j : \sum_{j'=1}^{j'=p} W_{jj'} u_\alpha^{j'} - \sqrt{f}_j \sum_{j'=1}^{j'=p} \sqrt{f}_{j'} u_\alpha^{j'} = \lambda u_\alpha^j$$

Le deuxième terme étant nul, il reste :

$$Wu = \lambda u$$

La recherche des facteurs est donc simplifiée puisqu'il est inutile de centrer le nuage par rapport à son centre de gravité.

Soit $u^j (j = 1,p)$ un vecteur propre de W. La projection du point i sur cette direction est :

$$f(i) = \sum_{j=1}^{j=p} u^j \frac{f_{ij}}{f_i \sqrt{f_j}}$$

Comme on s'intéresse aux points i de coordonnées $\{\frac{f_{ij}}{f_i}, j=1, p\}$ les facteurs véritables sont :

$$\varphi^j = \frac{u^j}{\sqrt{f_j}}$$

IV - LA RESOLUTION DU PROBLEME DANS R^N

Il est bien clair que dans les paragraphes précédents, nous avons particularisé R^p de manière arbitraire. On peut tout aussi bien résoudre le problème dans R^n, et diagonalisant la matrice

$$W_{ii'} = \sum_{j=1}^{j=p} \frac{f_{ij}\, f_{ij}}{f_j \sqrt{f_i\, f_{i'}}}$$

En fait, dans la pratique, on résoud toujours le problème dans l'espace de plus petite dimension, qui est en général celui des caractères, car on obtient les facteurs sur un espace en fonction des facteurs sur l'autre part des calculs purement linéaires (sans inversion ni diagonalisation). La raison en est la suivante : la matrice W, carrée d'ordre p peut s'écrire comme le produit d'une matrice R(p,n) par sa transposée R'(n,p). Soit en effet $r_j = \{\frac{f_{ij}}{\sqrt{f_i f_{j'}}}$, i=1..p} le terme général cjj' de RR' s'écrit :

$$cjj' = < rj,\ rj' > = \sum_{i=1}^{i=n} \frac{f_{ij}\, f_{ij'}}{f_i \sqrt{f_j\, f_{j'}}} = Wjj'$$

Donc W(p,p) = RR'

De même W(n,n) = R'R

Soit donc $u = (u^1 \ldots u^p)$ un vecteur propre de RR', pour la valeur propre λ. Alors :

$$RR'u = \lambda u$$

Prémultiplions par R', il vient :

$$R'RR'\ u = \lambda\ R'\ u$$

R'u est donc vecteur propre de R'R, pour la même valeur propre. Il est souvent commode de normer à 1 les facteurs. Si u est de norme 1 (pour la métrique usuelle), V = R'u est de norme :

$$\| V \|^2 = < u'R,\ R'u > = <u',\ \lambda u > = \lambda$$

On prendra donc pour vecteur v le vecteur de composantes :

$$\forall\, i \qquad v_i = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{j=p} \frac{f_{ij}}{\sqrt{f_i\, f_j}}\, u_j$$

Le facteur proprement dit φ^i s'obtient pour une transformation analogue à celle de R^p :

$$\varphi^i = \frac{v_i}{\sqrt{f_i}}$$

Il est de norme 1 pour le métrique des poids :

$$\sum_{i=1}^{i=n} f_i\, \varphi^{i2} = \sum_{i=1}^{i=n} f_i\, \frac{v_i^2}{f_i} = 1$$

On a donc finalement la relation suivante, entre les facteur φ_α^i et Ψ_α^j (sur R^n et R^p) relatifs à la même valeur propre λ_α

$$\sqrt{f_i}\, \varphi_\alpha^i = \frac{1}{\sqrt{\lambda}} \sum_j \frac{f_{ij}}{\sqrt{f_i\, f_j}}\, \phi_\alpha^j \quad f_j - \frac{1}{\sqrt{\lambda_\alpha}}$$

Soit encore :

$$\sqrt{\lambda_\alpha}\, f_i\, \varphi_\alpha^i = \sum_j f_{ij}\, \phi_\alpha^j \qquad (1)$$

De même :

$$\sqrt{\lambda_\alpha}\, f_j\, \phi_\alpha^j = \sum_{i=1}^{i=n} f_{ij}\, \varphi_\alpha^i \qquad (2)$$

Ces formules, dites de transition d'un espace à l'autre, sont d'une grande importance pratique : quel que soit l'intérêt que l'on porte à l'un ou l'autre espace, il est souvent préférable de faire l'analyse dans l'espace de plus petite dimension, et d'appliquer la formule de transition pour obtenir les directions propres de l'autre espace.

L'interprétation "physique" de ces formules est relativement simple. Dans la formule 1 par exemple, $\sum_{j=1}^{j=p} \frac{f_{ij}}{f_i}\, \varphi_\alpha^j$ représente la projection du point i sur l'axe factoriel de rang α. Cette projection est, au coefficient $\sqrt{\lambda_\alpha}$ près, la $i^{ème}$ composante de l'axe factoriel de même rang dans R^n.

N.B. Les présentations classiques de l'Analyse des Correspondances (cf.[1], [3], [6]) sont plus rigoureuses et cohérentes. Nous avons cependant préféré ne pas nous référer trop directement à leurs travaux, qui utilisent des notions algébriques assez abstraites (Produits tensoriels, Résolution de diagrammes commutatifs, Passage au dual) qui auraient pu alourdir un exposé dirigé vers les applications.

V - QUELQUES PROPRIETES DES FACTEURS

1/ Les facteurs sont de moyenne nulle :

$$③ \quad \sum_{i=1}^{i=n} f_i \, \varphi_\alpha^i = 0$$

On sait en effet que les directions propres d'une matrice symétrique sont orthogonaux. Donc $V = \{V_i = \sqrt{f_i}\, \varphi_\alpha^i\}$ est orthogonal au facteur trivial $t = \{\sqrt{f_i}\}$, soit :

$$\sum_{i=1}^{i=n} f_i \, \varphi_\alpha^i = 0$$

Plus simplement la suite $\{\sqrt{\lambda_\alpha}\, \varphi_\alpha^i,\ i=1,\ n\}$ représente la suite des projections $\{G\ M_\alpha^i\}$ des points du nuage sur le $\alpha^{\text{ème}}$ axe factoriel. La condition exprime donc que le centre de gravité est bien le barycentre du système de points, affectés des masses convenables.

2/ Les facteurs sont orthogonaux deux à deux et ont pour variance 1

$$\sum_i f_i \, \varphi_\alpha^i \, \varphi_{\alpha'}^i = \delta_{\alpha\alpha'}$$

3/ On peut reconstituer le nuage lorsqu'on a extrait tous les facteurs. Un point i a pour coordonnées $f_J^i = \{f_{ij}/f_i, j=1,p\}$ sur la base canonique de R^p. Les axes factoriels constituent un repère orthonormé, pour la métrique $f_J = \{f_j,\ j=1,\ p\}$, d'origine $G = \{f_j,\ j=1,\ p\}$. La projection de i sur l'axe α est $F_\alpha(i) = \sqrt{\lambda_\alpha}\, \varphi_\alpha^i$. Donc :

$$f_J^i = f_J + \sum_\alpha f_J \, F_\alpha(i) \, \varphi_\alpha^j$$

Soit en mettant en évidence la symétrie en i et J :

$$④ \quad f_{ij} = f_i \, f_j \, [1 + \sum_\alpha \sqrt{\lambda_\alpha}\, \varphi_\alpha^i \, \varphi_\alpha^j]$$

On montre que réciproquement, des fonctions sur I et J vérifiant 4 , orthogonales et de variance 1, sont les facteurs de la correspondance $\{f_{ij}\}$

4/ Une double application de la formule de transition montre que les facteurs (sur I par exemple) vérifient :

⑤ $$\varphi^I \circ f_I^J \circ f_J^I = \lambda \varphi^I$$

Ces notations sont utilisées couramment par J.P. BENZECRI on somme sur les indices répétés, pourvu que le 1er soit en haut et le deuxième en bas ;

en effet : $\forall_j \sum_{i=1}^{i=n} \varphi^i f_i^j = \sqrt{\lambda} \Psi^j$

$$\forall_i \sum_{j=1}^{j=p} \Psi^j f_j^i = \sqrt{\lambda} \varphi^i$$

Donc : $\forall i_1 \sum_j (\sum_i \varphi^i \frac{f_i^j}{\sqrt{\lambda}}) f_j^{i_1} = \sqrt{\lambda} \varphi^{i_1}$

Soit : $\forall i_1 \sum_i \sum_j \varphi^i f_i^j f_j^{i_1} = \lambda \varphi^{i_1}$

Cette formule est utile dans les exercices d'école et pour l'Analyse en continu, où elle se transpose bien.

5/ La formule de transition suggère une autre interprétation des facteurs. Considérons $\{\varphi_\alpha(i), i=1, n\}$ et $\{\varphi_\alpha(j), j=1, p\}$ facteurs associés à la valeur propre λ_α, comme des fonctions sur IxJ. On définit pour cela :

$$\forall j : \varphi_\alpha(i,j) = \varphi_\alpha(i)$$

$$\forall i : \Psi_\alpha(i,j) = \Psi_\alpha(j)$$

Ces fonctions sont de moyenne nulle et de variance 1 (par rapport à $\{f_{ij}, i=1,n, j=1,p\}$. Leur coefficient de corrélation est donc :

$$\text{Corr}(\varphi_\alpha, \Psi_\alpha) = \sum_{i,j} f_{ij}\, \varphi_\alpha^i\, \Psi_\alpha^j$$

$$= \sum_i \varphi_\alpha^i\, f_i \sum_i \varphi_\alpha^i\, f_i \left(\sum_j f_j^i\, \psi_\alpha^j\right)$$

$$= \sum \varphi_\alpha^i\, f_i\, \lambda_\alpha^{1/2}\, \varphi_\alpha^i \quad \text{(par application de la formule de transition}$$

$$= \lambda_\alpha^{1/2}$$

Ainsi, en Analyse des Correspondances, le couple de facteurs relatifs à la plus grande valeur propre est le couple de fonctions, de moyenne nulle et de variance 1 le plus corrélé possible. Le $k^{\text{ième}}$ couple de facteurs a la même propriété, dans les espaces orthogonaux aux (k-1) premiers couples mis en évidence.

VI - DEFINITIONS UTILES

Nous nous plaçons dans R^p. Nous introduisons quelques définitions utiles, suggérées par les décompositions possibles de l'inertie totale.

. La distance de i au centre de gravité est :

$$p(i) = \left(\sum_{j=1}^{j=p} \frac{1}{f_j} (f_j^i - f_j)^2 \right)^{1/2}$$

. L'inertie totale est la somme des valeurs propres, c'est aussi la somme des produits mr^2 de tous les éléments i :

$$\sum_{\alpha} \lambda_\alpha = \sum_i f_i [p(i)]^2$$

. L'inertie portée par l'axe α est l'inertie de la projection du nuage sur cet axe, soit :

$$\lambda_\alpha = \sum_i f_i \, F_\alpha^2(i)$$

. On a évidemment (théorème de Pythogore)

$$p^2(i) = \sum_\alpha F_\alpha^2(i)$$

. On peut donc définir les contributions suivantes :

$p^2(i) f_i$: contribution absolue de i à l'inertie totale

$F_\alpha^2(i) f_i$: contribution absolue de i à l'inertie λ_α

$F_\alpha^2(i)$: contribution absolue de α à i

$f_i \frac{F_\alpha^2(i)}{\lambda_\alpha}$: contribution relative de i à l'inertie λ_α

$\frac{F_\alpha^2(i)}{p^2(i)}$: contribution relative à α à i

Notons qu'il est possible que $\frac{F_\alpha^2(i)}{p^2(i)}$ soit égal à 1 (le point i est alors sur l'axe α. Par contre il n'est pas possible que $\frac{f_{i_1} F_\alpha^2(i_1)}{\lambda\alpha}$ soit égal à 1.

On aurait alors en effet i

$\forall i \neq i_1$, $f_i F_\alpha^2(i) = 0$, donc $F_\alpha(i) = 0$. La fonction F_α ne serait pas de moyenne nulle.

Il se peut fort bien que cette contribution soit élevée il y a alors intérêt à reprendre l'analyse en éliminant le point i (quitte à le projeter par la suite en élément supplémentaire, cf. infra.). En effet, l'axe factoriel ainsi apparu explique en général une opposition entre l'élément i_1 et les autres éléments du tableau, et masque les rapports internes entre les éléments $i \neq i_1$.

Ces contributions permettent également d'éprouver la validité de la représentation obtenue : si l'inertie portée par les 2 premiers axes représente 90 % de l'inertie totale, il est généralement inutile de poursuivre l'analyse plus à fond. Mais des contributions faibles ne sont pas nécessairement le fait de structures mal mises en évidence par les 2 ou 3 premiers facteurs. On sait que les différents facteurs sur un même espace sont non corrélés deux à deux, mais il peut exister une relation fonctionnelle (approchée dans les cas réels, exacte dans certains exercices d'école), respectant la non-corrélation linéaire ; par exemple, les facteurs d'ordre supérieur à 2 s'expriment en fonction des 2 premiers facteurs. Ce phénomène a reçu, dans la littérature, le nom d'Effet GUTTMAN.

VII - L'ADJONCTION D'ELEMENTS SUPPLEMENTAIRES

Comme on l'a vu précédemment, il peut être souhaitable d'éliminer un individu (ou un caractère) s'il s'oppose trop à tous les autres, au point de fausser l'analyse. Il n'en est pas moins souhaitable, une fois cette différenciation acquise, de préciser ses rapports avec les autres éléments, définis indépendamment de lui. Il suffit pour celà de faire l'analyse en éliminant l'individu i_S, puis de le replacer sur la carte au moyen de la formule de transition. Soit en effet $\{f_j^{i_S}, j=1, p\}$ son profil ; la formule de transition donne :

$$\forall\ \alpha \quad \sum_j \Psi_\alpha^j f_j^{i_S} = \lambda_\alpha^{1/2} \varphi_\alpha^{i_S} = F_\alpha(is)$$

Il est donc aisé de calculer les projections de i_S sur les différents axes, et par suite de placer l'élément supplémentaire sur la carte.

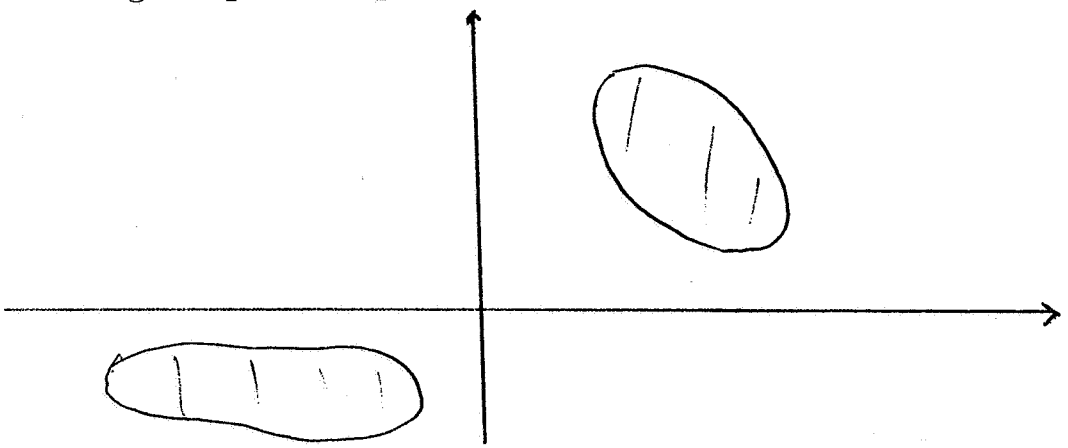
Il faut néanmoins prendre garde qu'en toute généralité $\sum_\alpha F_\alpha^2(iS) < p^2(iS)$, car le point i_S n'appartient pas nécessairement au support du nuage. Il sera utile d'évaluer la composante de $f_J^{i_S} - f_J$ orthogonale au support du nuage.

L'intérêt des éléments supplémentaires est à la fois théorique et pratique.

On peut d'une part essayer de dégager une interprétation de "cause à effet". Supposons qu'on veuille interpréter les résultats du dernier référendum en fonction de ceux des dernières présidentielles et on projettera les éléments "référendum" en éléments supplémentaires de préférence à une analyse globale.

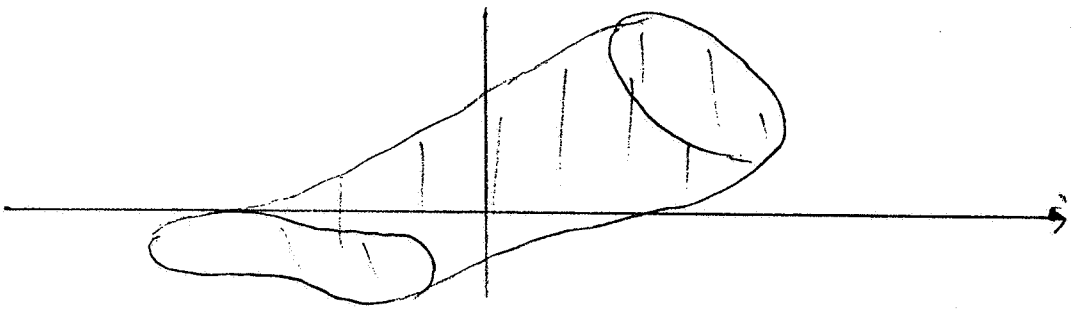
On peut parfois tester la validité d'une analyse. Supposons que l'analyse d'un tableau IxJ ait fait apparaître une dichotomie de l'ensemble I dans le plan des 2 premiers axes factoriels.

Le nuage a par exemple l'aspect suivant :



L'ensemble I n'est, bien souvent, qu'un échantillon d'un ensemble I' beaucoup plus vaste et potentiellement infini (ex. i les rongeurs, les fumeurs, les français ...]

La dichotomie n'aura une signification intrinsèque que si tout $i_s \in I'$ se projette au voisinage de l'un des deux blocs. Si par contre, après projection de quelques éléments supplémentaires, le nuage prend une forme "connexe" telle que celle-ci :



il faudra conclure que la dichotomie mise en évidence révélait en fait un échantillonnage imparfait, et reprendre l'analyse sur des bases plus saines.

Du point de vue pratique, l'adjonction d'éléments supplémentaires permet des analyses "en temps réel" au sens suivant : supposons qu'un certain nombre d'analyses aient été effectuées sur un grand nombre de patients, et que ces analyses permettent d'établir une carte factorielle sur laquelle différentes maladies soient fortement individualisées. Un nouveau patient arrive et subit les analyses. Il est possible de le placer sur la carte factorielle, et d'avoir une présomption de diagnostic, présomption car les méthodes statistiques comportent toujours des risques, et d'autre part parce que la possibilité que le profil du patient ait une forte composante sur l'orthogonal du nuage n'est jamais exclue a priori.

VII - LA PROGRAMMATION

Les programmes d'Analyse Factorielle sont assez nombreux. Le plus classique est celui de ROUX et FRIANT, publié au Laboratoire de STATISTIQUE MATHEMATIQUE. Un des derniers parus est celui de J.P. BORDET, publié au Centre d'Informatique Géologique (ANAXDI LHM/N72/17), P. ROUX l'a adapté pour Fontainebleau, et seuls le programme principal, les subroutines Prepar et Lyre sont à compiler. Les autres sous-programmes sont enregistrés en bibliothèque. L'intérêt de ce programme est le suivant : on peut traiter un nombre d'individus aussi grand qu'on veut (moyennant bien entendu un temps de lecture croissant). En effet, au lieu de mettre en mémoire le tableau $\{k(i,j),\ i \in I,\ j \in J\}$ on place successivement toutes les lignes au même endroit mémoire, et on constitue pas-à-pas la matrice à diagonaliser.

Un programme complémentaire (interface BENSON) permet de tracer les ellipses d'inertie de sous-ensembles de points du nuage. Il est également dû à J.P. BORDET.

BIBLIOGRAPHIE

1/ THEORIE

J.P. BENZECRI [1] - Distance distributionnelle et métrique du χ^2 en Analyse des Correspondances. LSM 1970

[2] - Pratique de l'Analyse des Correspondance. LSM

[3] - Représentation Euclidienne d'un ensemble fini muni de masses et de distances. LSM 1970

[4] - Sur le choix des unités et des poids dans un tableau en vue d'une analyse de Correspondances. LSM 1972

P. CAZES [5] - Applications linéaires. LSM

[6] - Analyse Factorielle d'un nuage de Points.LSM

LEBART ET FENELON - Statistique et Informatique Appliquées DUNOD 1971

2/ ETUDES PRATIQUES.

J.P. BORDET - DENSITES

P. CAZES - Exemple de Traitement statistique de données hydrochimiques Bulletin du B.R.G.M. n° 1970

J.F. MARCOTORCHINO - Les marques de cigarettes. LSM

P. ROUX - Incidence de l'Analyse des Correspondances sur l'optimisation d'un réseau piézométrique. Traitement visuel de l'information géologique. NANCY 1971.

3/ PROGRAMMES

ROUX-FRIANT	Analyse Factorielle	LSM
	ANAXDI	LHM
	ELLIPSE	LSM