ANNALES DE L'INSEE — N° 35-1979

Analyse factorielle sphérique : une exploration

Dominique DOMENGES
Michel VOLLE *

Ce travail a pour origine les deux constatations suivantes : les distributions sur un ensemble fini I peuvent être repérées sur un orthant de sphère ; en effet, si $\Sigma\, p_i = 1$ avec $p_i \geqslant 0$, la distribution p peut être représentée par un point de coordonnées $\sqrt{p_i}$. Il est en outre possible de définir sur la sphère une distance entre distributions qui corresponde à la métrique du χ^2 que l'on utilise habituellement sur le simplexe des distributions.

Cette représentation des distributions ouvre la voie d'une nouvelle méthode d'analyse factorielle, qui a des rapports étroits avec l'analyse factorielle des correspondances. Nous explorerons rapidement le domaine de ses applications, et donnerons enfin quelques exemples concrets des résultats qu'elle fournit.

* Dominique DOMENGES est allocataire DGRST, laboratoire de statistique de l'Université Pierre et Marie CURIE, 4 place Jussieu, PARIS 5e.
Michel VOLLE est chef de la division « Comptes trimestriels » de la Direction des Synthèses économiques de l'INSEE.

Introduction *

A l'origine de la méthode que nous présentons ici se trouvent deux interrogations. D'une part la métrique du χ^2 utilisée en analyse des correspondances donne au simplexe des distributions, comme l'a remarqué J.-P. BENZECRI, *une structure d'espace de* RIEMANN ; *que se passe-t-il si l'on essaie de tirer parti de cette structure, c'est-à-dire de rechercher dans le simplexe la géodésique qui joint deux points, et de calculer la distance géodésique entre deux distributions ? Quelle utilisation peut-on faire des résultats de ce calcul ?*

D'autre part, il nous était apparu que l'analyse des correspondances pouvait être d'un usage difficile lorsque le tableau analysé est très éloigné du produit de ses marges : par exemple lorsque la structure du tableau est quasi diagonale (c'est le cas couramment lorsqu'il s'agit de tableaux de transition), ou lorsque la diagonale n'est pas définie (tableaux d'échanges). Toute l'analyse des correspondances peut en effet être construite en partant de l'expression :

$$\|f-\varphi\|^2_\varphi = \sum_{ij} \frac{(f_{ij}-f_i f_j)^2}{f_i f_j}$$

qui mesure la distance entre les tableaux f_{IJ} *et* $\varphi_{IJ} = f_I \otimes f_J$, *calculée selon la métrique du* χ^2 *centrée sur* φ_{IJ}. *On peut donc dire que, dans un certain sens, l'analyse des correspondances revient à l'étude de la différence entre un « tableau concret »* f *et un « tableau de référence »* φ, *ici égal au produit des marges de* f. *En utilisant une métrique centrée sur* φ, *on confère à l'analyse un caractère local — d'ailleurs justifié par de puissantes raisons. Mais n'est-ce pas ce caractère local qui est à l'origine des difficultés que l'on rencontre en analyse des correspondances lorsque* f *est très éloigné de* φ *? Ces difficultés peuvent-elles être levées si l'on supprime le caractère local de la métrique, par exemple en la remplaçant par une expression dérivée de la distance géodésique ?*

Telles sont les deux intuitions qui ont provoqué cette recherche. En se développant, celle-ci nous a procuré des résultats en partie inattendus, et d'une portée qui paraît générale.

Nous montrerons d'abord comment, en considérant le simplexe des distributions comme un espace de Riemann, *on est conduit à utiliser entre distributions la* métrique de Hellinger (*ce qui revient, géométriquement, à représenter la distribution* p_i *par un point de coordonnée* $x_i = \sqrt{p_i}$, *situé sur la sphère* $\Sigma x_i^2 = 1$; *puis à prendre comme distance entre deux distributions* p *et* q *la longueur de la « corde » qui joint les deux points, de coordonnées* $x_i = \sqrt{p_i}$ et $y_i = \sqrt{q_i}$). *Cette métrique a, entre autres propriétés, de permettre un repérage commode des « distributions » comportant des valeurs négatives.*

On peut, à l'aide de la métrique de Hellinger, calculer la distance entre un « tableau concret » (*de contingence*) f *donné et un « tableau de référence »* φ :

$$d^2(f,\varphi) = \sum_{ij} (\sqrt{f_{ij}} - \sqrt{\varphi_{ij}})^2$$

Il est aisé d'établir le formulaire de l'analyse factorielle sphérique, en suivant une démarche analogue à celle qui permet d'établir le formulaire de l'analyse des correspondances en partant de la métrique du χ^2, $\| f - \varphi \|_{\varphi}^2$.

On remarque que, si l'on utilise la métrique du χ^2, *il est indispensable qu'il n'existe aucun* (i, j) *tel que* $\varphi_{ij} = 0$ et $f_{ij} \neq 0$: *le calcul de la distance devient alors en effet impossible. Par contre, cette restriction n'existe pas si l'on utilise la métrique de Hellinger : elle permet par exemple de calculer la distance entre* f *et un tableau de référence diagonal, et de réaliser l'analyse factorielle associée à cette distance; cette analyse semble particulièrement bien adaptée pour l'étude des tableaux de transition ou d'échanges.*

A priori, le champ des applications possibles de l'analyse factorielle sphérique est large. Nous avons réalisé plusieurs essais; chacun sait qu'en Analyse des données le moment de vérité se situe dans les applications concrètes, et que des formulaires d'une grande élégance logique peuvent fort bien procurer des résultats décevants: ici nous n'avons pas été déçus. Les résultats ont confirmé, d'une façon tout empirique, les espoirs qu'un raisonnement analogique nous avait fait placer dans la métrique de Hellinger. Nous en sommes restés à cette situation exploratoire, qui nous paraît prometteuse. Notre travail n'avait pas pour objet d'étudier l'ensemble des propriétés que la métrique de Hellinger peut avoir en statistique mathématique, ni encore moins de nier l'intérêt que peuvent avoir d'autres métriques.

C'est dans le travail que nous avions d'abord en vue, l'étude des tableaux de transition et d'échanges, que nous avons rencontré la seule difficulté réelle : le premier axe fourni par l'analyse factorielle nécessite dans cette application un effort d'interprétation particulier. Mais d'autres applications furent possibles : analyse d'un tableau de variations de stocks (comportant donc des valeurs négatives); comparaison de deux tableaux « hommes » et « femmes » issus d'une enquête de type démographique; enfin analyse d'un nuage à partir d'un point de ce nuage, particulièrement intéressante pour l'étude de données chronologiques. D'autres applications sont en cours de réalisation.

---

* Ce travail résulte d'une recherche effectuée à l'unité de Recherche de l'INSEE.
Nous remercions les personnes qui ont bien voulu s'y intéresser, l'encourager, et faire des suggestions qui ont permis certains développements, notamment MM. Benzecri et Jambu, de l'ISUP, et MM. Deville, Froment, Grandjean, Kaminski, Laroque, Oudiz, de l'INSEE.

1 Représentation des distributions sur la sphère

Nous allons suivre dans cette partie l'itinéraire suivant : nous introduirons d'abord la métrique du χ^2, dont l'usage se justifie logiquement et pratiquement lorsqu'on doit calculer des distances entre distributions. Puis nous introduirons à partir de la métrique du χ^2 une autre métrique, la métrique de Hellinger, dont nous verrons qu'elle est simultanément équivalente à la « distance géodésique » entre distributions (calculée en intégrant la métrique du χ^2 considérée comme une métrique de Riemann), et à une expression particulièrement intéressante du gain d'information. Ainsi placée au point de rencontre de propriétés logiques variées, la métrique de Hellinger semble devoir être un outil mathématique fécond; de plus, cet outil permet des calculs particulièrement simples, ce qui est un gage supplémentaire de fécondité : si l'on représente les distributions sur la sphère, la métrique d'Hellinger devient la métrique euclidienne canonique. Entre autres conséquences, cette représentation permet de traiter des distributions comportant des valeurs négatives.

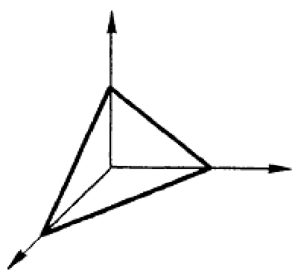
Considérons d'abord les distributions positives sur un ensemble fini I : une telle distribution p obéit aux relations :

$$\sum_{i \in I} p_i = 1 \text{ et } p_i \geqslant 0.$$

Chaque distribution peut donc être représentée, dans l'espace $R^{\text{Card } I}$ par un point du simplexe défini par les relations $\Sigma x_i = 1$ et $x_i \geqslant 0$. Ce simplexe est un segment de droite si Card I $= 2$, un triangle équilatéral si Card I $= 3$, un tétraèdre si Card I $= 4$, etc. (graphique 1).

GRAPHIQUE 1

Simplexe des distributions si Card I = 3



Supposons maintenant que l'on observe la distribution d'une population concrète d'effectif k selon le caractère I, et notons f cette distribution. Peut-on considérer la population en question comme un échantillon provenant d'une population dans laquelle la distribution selon I serait p ? Pour répondre à cette question, on calcule la quantité :

$$k \sum \frac{(f_i - p_i)^2}{p_i}$$

et on la compare à un χ^2 à Card I $- 1$ degrés de liberté. Si cette quantité prend une valeur qui n'a qu'une faible probabilité d'être dépassée par ce χ^2, on devra conclure que f s'écarte trop de p pour que l'on puisse conserver l'hypothèse que l'échantillon considéré provient d'une population répartie selon p. Ce test est bien connu sous le nom de « test du χ^2 ».

Ce résultat, ainsi que certaines considérations provenant de la théorie de l'information sur lesquelles nous reviendrons dans la partie 1.2, conduit à définir une distance entre distributions à l'aide d'une métrique analogue à celle que l'on utilise pour le test du χ^2. Si l'on considère trois distributions p, q et r, le carré de la distance entre p et q, calculée avec la « métrique du χ^2 centrée sur r », est donné par :

$$\| p - q \|_r^2 = \sum \frac{(p_i - q_i)^2}{r_i}$$

Avec cette notation, la quantité calculée lors du test du χ^2 s'écrit :

$$k \| f - p \|_p^2$$

on voit qu'une métrique du χ^2 dépend de la distribution sur laquelle elle est centrée : à chaque point du simplexe peut être ainsi associée une métrique différente.

1.1. Le simplexe des distributions considéré comme un espace de Riemann

Supposons que nous définissions dans le simplexe une métrique locale, en associant à chaque point la métrique du χ^2 centrée sur ce point. Cette métrique dote le simplexe d'une structure d'espace de Riemann ([6, p. 101] et [1, vol. 2, p. 136]); la forme différentielle quadratique définissant le carré de l'élément de distance dans cet espace au point r est :

$$(ds)^2 = \sum \frac{(dr_i)^2}{r_i}$$

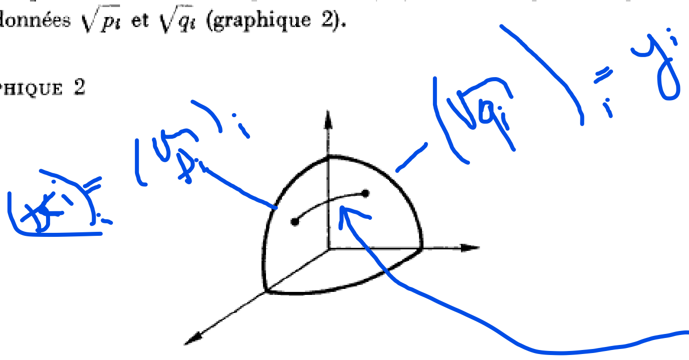
On peut alors se poser deux questions, qui se présentent très naturellement lorsqu'on a affaire à un espace de Riemann : étant données deux distributions p et q, quelle est la géodésique qui joint ces deux distributions, c'est-à-dire la courbe G telle que, le long de cette courbe, l'intégrale de ds de p à q soit minimum ? Et quelle est la valeur de ce minimum, que nous appellerons « distance géodésique entre p et q » ?

Un changement de coordonnées permet de trouver très facilement la réponse à ces questions. Posons $x_i = \sqrt{r_i}$; $(dx_i)^2 = (dr_i)^2/4r_i$, et $ds^2 = 4\ \Sigma\ (dx_i)^2$.

Donc, si l'on associe à la distribution r le point de coordonnées $\sqrt{r_i}$, qui se trouve sur la sphère $\Sigma x_i^2 = 1$, la métrique locale du simplexe devient sur la sphère la métrique euclidienne canonique (à une constante multiplicative près).

Sur la sphère, la géodésique joignant deux points est le plus petit des deux arcs du grand cercle passant par ces deux points : la géodésique entre p et q sur le simplexe sera donc la courbe transformée, par la relation $r_i = x_i^2$, du plus petit des deux arcs de grand cercle joignant sur la sphère les points de coordonnées $\sqrt{p_i}$ et $\sqrt{q_i}$ (graphique 2).

Graphique 2



La distance géodésique entre deux points de la sphère de coordonnées x_i et y_i est la longueur de l'arc de grand cercle qui les joint, soit $d_g\,(x, y) = \text{Arc cos}\ (\Sigma x_i y_i)$. La distance géodésique entre p et q sur le simplexe est donc :

$$\boxed{d_g\,(p, q) = 2\ \text{Arc cos}\ (\Sigma \sqrt{p_i\, q_i})}$$

En passant du simplexe à la sphère, l'expression de l'élément de distance s'est simplifiée, ce qui a permis de trouver les géodésiques et la distance géodésique dans le simplexe sans recourir au calcul. Cette simplicité nous suggère de travailler directement sur la sphère. Cependant, l'expression de la distance géodésique telle que nous l'avons donnée se prêterait mal au calcul; nous verrons en 1.3 que l'on peut utiliser une autre distance, équivalente à la distance géodésique, et qui permet des calculs très simples.

1.2. Une approche par la théorie de l'information

Rényi a établi [7, p. 527], à partir de six postulats donnant les propriétés que l'on attend du gain d'information résultant du remplacement d'une distribution par une autre, une expression plus générale que celle de Shannon. La « mesure d'ordre α du gain d'information » définie par Rényi est :

$$I_\alpha\,(q/p) = \frac{1}{\alpha - 1} \log_2 \left[\sum \left(\frac{q_i}{p_i}\right)^\alpha p_i \right]$$

Rényi démontre que, si $\alpha \to 1$, $I_\alpha(q/p) \to \Sigma q_i \log_2 (q_i/p_i)$;
on retrouve donc comme cas particulier de la mesure d'ordre α le gain d'information de Shannon.

Remarquons en passant que la métrique du χ^2 centrée sur q donne, à une constante multiplicative près, une mesure approchée du gain d'information de Shannon lorsque q et p sont proches. Posons en effet $p_i = q_i (1 + \varepsilon_i)$;
on trouve :

$$\log_2 \frac{q_i}{p_i} = \frac{1}{\operatorname{Log} 2} \left[-\varepsilon_i + \frac{\varepsilon_i^2}{2} + 0\,(\varepsilon_i^2) \right]$$

d'où, si $\varepsilon_i \to 0$,

$$\Sigma\, q_i \log_2 \frac{q_i}{p_i} \approx \frac{1}{2 \operatorname{Log} 2} \| p - q \|_q^2$$

Ce résultat, qui permet d'établir une relation entre la métrique du χ^2 et la théorie de l'information, est au fondement de certaines présentations des classifications sur tableau de contingence [1, vol. 1, p. 216].

Parmi toutes les mesures d'ordres α du gain d'information, celle qui correspond à $\alpha = 1/2$ possède une propriété de symétrie très intéressante : il est aisé de vérifier que :

$$I_{\frac{1}{2}}(q/p) = I_{\frac{1}{2}}(p/q)$$

on trouve :

$$I_{\frac{1}{2}}(q/p) = -2 \log_2 (\Sigma \sqrt{p_i q_i})$$

1.3. La métrique de Hellinger

Nous voyons apparaître, dans $I_{\frac{1}{2}}(q/p)$, l'expression :

$$\Sigma \sqrt{p_i q_i}$$

que nous avions déjà rencontrée dans la distance géodésique de p et q. Ceci nous conduit à examiner de plus près les propriétés du produit scalaire :

$$< p, q > = \Sigma \sqrt{p_i q_i}$$

et de la distance entre distributions qui lui est associée, dite « métrique de Hellinger ».

$$\boxed{d^2(p,q) = \Sigma (\sqrt{p_i} - \sqrt{q_i})^2}$$

C'est le carré de la longueur de la corde qui joint les deux points de la sphère de coordonnées $\sqrt{p_i}$ et $\sqrt{q_i}$. Il est évident que :

$$< p, q > = 1 - \frac{1}{2} d^2(p,q)$$

Si $d^2(p, q)$ est petit, on a donc :

$$I_{\frac{1}{2}}(q/p) \approx \frac{1}{\mathrm{Log}\,2} d^2(p, q)$$

Par ailleurs, lorsqu'un arc est petit, la longueur de la corde est équivalente à celle de l'arc; d'où :

$$d_g^2(p, q) \approx 4d^2(p, q)$$

$I_{\frac{1}{2}}(q/p)$ et $d^2_g(p, q)$ sont toutes deux des fonctions monotones croissantes de $d^2(p, q)$, prenant la valeur zéro pour $p = q$, équivalentes à $d^2(p, q)$ pour p proche de q [1]. Le choix de la métrique de Hellinger pour exprimer la distance entre distributions apparaît donc comme naturel : d'une part cette métrique donne à la distance une expression très simple; d'autre part elle donne, à des constantes multiplicatives près, des approximations satisfaisantes de la distance géodésique et de l'information discriminante d'ordre ½. Il est facile de voir qu'en outre la métrique de Hellinger est équivalente à la métrique du χ^2 centrée sur r (avec $r_i \neq 0$ pour tout i) si p et q sont proches de r, puisqu'elle est équivalente à la distance géodésique, elle-même équivalente à la métrique du χ^2 (on retrouve la distance du χ^2 en dérivant la distance géodésique). On vérifie aisément que :

$$d^2(p,q) \approx \frac{1}{4}\|p-q\|_r^2$$

En raison de l'ensemble de ses propriétés, nous allons utiliser dans la suite de ce travail la métrique de Hellinger pour mesurer la distance entre distributions. On trouvera dans un article de Le Cam [5] une description de cette distance du point de vue de la statistique mathématique.

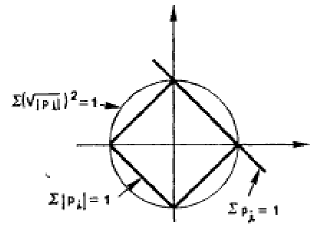
1.4. Représentation des distributions comportant des valeurs négatives

Considérons maintenant une distribution p telle que $\Sigma|p_i| = 1$, chaque p_i pouvant avoir un signe positif ou négatif (on rencontrera de telles distributions lorsque, par exemple, on considérera pour une région donnée des soldes migratoires par classes d'âge, ou, pour une entreprise, un ensemble de soldes comptables etc.). On remarque que la contrainte de normalisation que nous retenons est $\Sigma|p_i| = 1$ et non $\Sigma p_i = 1$: en effet, cette dernière formulation ne permettrait pas de représenter des distributions parfaitement plausibles, et pour lesquelles $\Sigma k_i = 0$ (k_i étant la valeur du poste i).

1. Deux métriques 1 et 2 sont équivalentes s'il existe deux réels positifs s et r tels que :

$$s < \frac{\|x-y\|_1}{\|x-y\|_2} < r$$

GRAPHIQUE 3



$\Sigma |p_i| = 1$ définit un carré si Card I $= 2$, un octaèdre si Card I $= 3$, etc. (graphique 3). Il est difficile de définir sur l'octaèdre une distance entre points appartenant à des faces différentes, en raison des discontinuités causées par les arêtes de l'octaèdre. Par contre, si l'on représente une distribution par un point de coordonnées :

$$[\text{signe de } p_i] \sqrt{|p_i|}$$

et si l'on dote la sphère du produit scalaire :

$$< p, q > = \Sigma \, [\text{signe de } p_i \, q_i] \sqrt{|p_i| \cdot |q_i|}$$

il est possible de définir une distance entre distributions valable sur toute la sphère et non plus seulement sur son orthant positif : contrairement à l'octaèdre la sphère ne comporte qu'une face.

Ceci nous permettra, lors de certains des calculs qui suivront, d'admettre la possibilité de distributions comportant des termes négatifs — et donc de ne pas tenir compte de la contrainte $p_i \geqslant 0$. Ceci nous permettra également de procéder à l'analyse factorielle de tableaux de fréquences comportant des cases négatives, que l'on ne sait pas traiter commodément par l'analyse des correspondances (voir ci-dessous en 4.1.).

2 Comparaison de deux tableaux de fréquences

Considérons deux caractères qualitatifs I et J possédant chacun un ensemble fini de modalités; nous noterons ces deux ensembles également I et J, et nous repérerons les modalités par les indices i et j.

Soit une population concrète d'effectif k répartie conjointement selon I et J; nous noterons k_{ij} le nombre des individus possédant à la fois les modalités i de I et j de J. La fréquence du couple (i, j) dans cette population est $f_{ij} = k_{ij}/k$. La pratique de l'analyse des correspondances montre que l'on peut traiter aussi des tableaux donnant non la répartition d'une population d'individus, mais la ventilation d'une quantité (somme d'argent par exemple) selon deux caractères [1, vol. 2, p. 20].

Nous supposons ici $f_{ij} \geqslant 0$. Nous noterons f_{IJ} (ou plus simplement f, lorsqu'il n'y a pas de risque de confusion) le tableau des f_{ij}; ce tableau peut être représenté par un point du simplexe des distributions, dans l'espace à Card I $\times$ Card J dimensions.

On peut considérer aussi les distributions marginales f_{I}, de terme courant $f_i = \sum_j f_{ij}$, et f_{J} de terme courant $f_j = \sum_i f_{ij}$. Nous noterons $f_{\mathrm{I}} f_{\mathrm{J}}$ le tableau de terme courant $f_i f_j$: ce tableau s'obtient en faisant le produit des marges de f_{IJ} [2].

Nous verrons en 2.2 que l'on peut reconstituer le formulaire de l'analyse des correspondances à partir de l'expression suivante :

$$\begin{aligned} \text{Lien}\,(\mathrm{I}, \mathrm{J}) &= \sum_{ij} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} \\ &= \| f_{\mathrm{IJ}} - f_{\mathrm{I}} f_{\mathrm{J}} \|^2_{f_{\mathrm{I}} f_{\mathrm{J}}} \end{aligned}$$

Lien (I, J) est le carré de la distance entre f_{IJ} et $f_{\mathrm{I}} f_{\mathrm{J}}$, calculée selon la métrique du χ^2 centrée sur $f_{\mathrm{I}} f_{\mathrm{J}}$.

On démontre que, si les deux caractères I et J sont indépendants sur la population considérée et si k est assez grand, la quantité k Lien (I, J) suit la loi du χ^2 à (Card I $-$ 1) (Card J $-$ 1) degrés de liberté. Cette propriété peut servir à tester l'indépendance des deux caractères.

---

2. Nous supposerons par la suite que $f_i f_j \neq 0$ pour tous les couples (i, j). S'il en était autrement, on se ramènerait aisément au cas que nous étudions en modifiant I et J par suppression des modalités pour lesquelles $f_i = 0$ ou $f_j = 0$.

L'analyse des correspondances va beaucoup plus loin que le test du χ^2 dans l'utilisation de l'expression Lien (I, J) : elle lui associe des nuages de points représentant les ensembles I et J; elle procède à l'analyse factorielle de ces nuages, dont elle fournit des règles d'interprétation; au total, elle permet non seulement de décider si f_{IJ} est proche ou non de $f_I f_J$, mais aussi d'étudier en détail la différence entre ces deux tableaux — et donc de mettre à jour, si elle existe, une « correspondance » entre I et J sur la population étudiée.

Nous allons construire la méthode de l'analyse factorielle sphérique en utilisant une démarche analogue à celle de l'analyse des correspondances. Alors que l'analyse des correspondances compare le tableau f au tableau $\varphi = f_I f_J$, et utilise comme expression de la distance :

$$\| f - \varphi \|_{\varphi}^{2} = \sum_{ij} \frac{(f_{ij} - \varphi_{ij})^2}{\varphi_{ij}}$$

nous partirons de la métrique de Hellinger :

$$d^2 (f, \varphi) = \sum_{ij} (\sqrt{f_{ij}} - \sqrt{\varphi_{ij}})^2$$

On peut remarquer des différences importantes entre ces deux distances : d'abord, la métrique de Hellinger fait jouer aux deux distributions f et φ un rôle symétrique, alors que la distribution φ sert à centrer la métrique du χ^2; ensuite — et c'est là sans doute le point le plus important — il est impératif qu'aucun des termes de la distribution qui sert à centrer la métrique du χ^2 ne soit nul (sans cela la distance n'a plus de sens) : si l'on utilise la métrique du χ^2, le choix du tableau de référence auquel on compare un tableau concret donné f est donc pratiquement limité à des tableaux qui par construction ne contiennent pas de case nulle, comme le tableau « produit des marges ». Par contre, l'expression de la métrique de Hellinger permet de prendre comme tableau de référence n'importe quel tableau, y compris un tableau diagonal ou tout autre type de tableau présentant des cases nulles. Le champ des utilisations possibles de l'analyse factorielle sphérique apparaît donc, *a priori*, comme très vaste.

Nous allons procéder comme suit : nous montrerons d'abord comment on peut associer une analyse factorielle à une somme quelconque de carrés sur I $\times$ J; nous vérifierons que cette méthode permet de reconstituer le formulaire de l'analyse des correspondances; puis nous établirons le formulaire de l'analyse sphérique.

2.1. Analyse factorielle à partir d'une somme de carrés sur I $\times$ J [2, p. 493]

Le problème que nous allons traiter ici est, en raison même de sa généralité, tout à fait formel, et peut donc sembler assez artificiel : mais il se justifie par le cadre commun qu'il fournit aux méthodes d'analyse factorielle, dont il donne une approche particulière.

Supposons donnée l'expression :

$$t^2 = \sum_{ij} t_{ij}^2 \ ;$$

nous avons eu affaire à des expressions de ce type lors du calcul de la distance entre deux tableaux, mais les développements qui suivent sont indépendants de l'origine des quantités t_{ij}^2, et on peut donc les interpréter comme la construction d'une analyse factorielle à partir d'une somme quelconque de carrés. Supposons aussi données une mesure m_I sur I et une mesure m_J sur J. Certains des résultats que nous obtiendrons sont indépendants de m_I et m_J, que l'on peut considérer comme des intermédiaires de calcul.

On peut poser :

$$t^2 = \sum_i \left[m_i \sum_j \left(\frac{t_{ij}}{\sqrt{m_i}} \right)^2 \right]$$

Posons :

$$x_j^i = t_{ij} / \sqrt{m_i} \ ;$$

t^2 apparaît alors comme l'inertie par rapport à l'origine, dans l'espace à Card J dimension, du nuage des points X^i de coordonnée courante x_j^i et munis chacun de la masse m_i. Nous noterons ce nuage X^I ou, plus simplement, X. La métrique utilisée pour calculer l'inertie est la métrique euclidienne canonique. La matrice d'inertie de X est V, de terme général.

$$v_{jj'} = \sum_i m_i \, x_{ij} \, x_{ij'} = \sum_i t_{ij} \, t_{ij'}$$

Notons T le tableau des t_{ij} et T' son transposé; on a :

$$V = T'T$$

On peut construire un autre nuage de points, en écrivant

$$t^2 = \sum_j \left[m_j \sum_i \left(\frac{t_{ij}}{\sqrt{m_j}} \right)^2 \right]$$

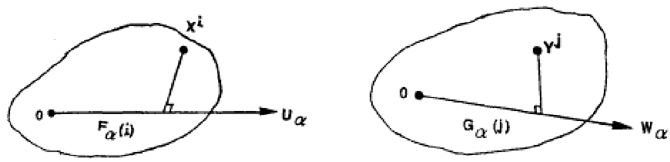
et en posant $y_i^j = t_{ij} / \sqrt{mj}$. Nous noterons Y^J (ou simplement Y) le nuage des Y^j munis chacun de la masse m_j. Il est situé dans l'espace à Card I dimensions et son inertie par rapport à l'origine, calculée en utilisant la métrique euclidienne canonique, est t^2. Sa matrice d'inertie est TT'.

Les matrices T'T et TT' sont symétriques : rappelons que les valeurs propres d'une matrice symétrique sont réelles, et que l'on peut former une base orthonormée avec ses vecteurs propres. De plus, il s'agit ici de matrices d'inertie : les valeurs propres ne peuvent pas être négatives, car elles mesurent l'inertie du nuage projeté sur le vecteur propre qui leur est associé (graphique 4).

GRAPHIQUE 4

*a. **nuage X dans l'espace à Card J dimensions***

*b. **nuage Y dans l'espace à Card I dimensions***



On montre aisément que, si U est vecteur propre unitaire de T'T associé à la valeur propre non nulle et non dégénérée λ, on peut trouver W, vecteur propre unitaire de TT' associé à la même valeur propre, par la relation :

$$U = \frac{1}{\sqrt{\lambda}} T' W;$$

réciproquement,

$$W = \frac{1}{\sqrt{\lambda}} TU$$

Dans le cas où λ serait multiple, l'énoncé devrait être un peu compliqué, mais sans que le fond de notre exposé en soit modifié : nous négligerons donc ce cas qui a été étudié ailleurs [11, p. 90].

On obtient donc les relations suivantes entre les coordonnées des vecteurs propres, dont on remarquera qu'elles sont indépendantes de m_I et m_J:

$$(1) \qquad \boxed{u_j = \frac{1}{\sqrt{\lambda}} \sum_i t_{ij}\, w_i \quad \text{et} \quad w_i = \frac{1}{\sqrt{\lambda}} \sum_j t_{ij}\, u_j}$$

Notons F (i) la coordonnée de X^i sur U, et G (j) la coordonnée de Y^j sur W; comme

$$F(i) = \sum_j x^i_j\, u_j \quad \text{et} \quad G(j) = \sum_i y^j_i\, w_i$$

on trouve :

$$(2) \qquad \boxed{u_j = G(j)\sqrt{\frac{m_j}{\lambda}} \quad \text{et} \quad w_i = F(i)\sqrt{\frac{m_i}{\lambda}}}$$

en combinant (1) et (2), on trouve les « formules de transition » :

$$(3)\qquad \boxed{\begin{aligned} F(i) &= \frac{1}{\sqrt{\lambda}} \sum_j t_{ij}\, G(j) \sqrt{\frac{m_j}{m_i}} \\ G(j) &= \frac{1}{\sqrt{\lambda}} \sum_i t_{ij}\, F(i) \sqrt{\frac{m_i}{m_j}} \end{aligned}}$$

Enfin, on peut classer les axes factoriels dans l'ordre des valeurs propres décroissantes, et les repérer par un indice α. Notons $F_\alpha(i)$ la coordonnée de X^i sur U_α; les axes factoriels forment une base de l'espace à Card J dimensions, donc :

$$X^i = \sum_\alpha U_\alpha F_\alpha(i)$$

$$x^i_j = \sum_\alpha \sqrt{\frac{\lambda_\alpha}{m_i}}\, w_{\alpha_j} u_{\alpha_j}$$

$$(4)\qquad \boxed{t_{ij} = \sum_{\alpha i} w_{\alpha i}\, u_{\alpha j} \sqrt{\lambda_\alpha}}$$

C'est la « formule de reconstitution » du tableau T; elle est indépendante de m_I et m_J. En utilisant les relations (2), on peut lui donner la forme suivante:

$$(5)\qquad \boxed{t_{ij} = \sqrt{m_i m_j} \sum_\alpha \frac{1}{\sqrt{\lambda_\alpha}} F_\alpha(i)\, G_\alpha(j)}$$

L'expression des aides classiques à l'interprétation d'une analyse factorielle se déduit immédiatement de ce formulaire; voici leurs valeurs :

$$\text{POIDS}\ (i) = m_i$$

$$\text{CONTR}\ (i) = \sum_j t^2_{ij}/t^2$$

$$\text{CTR}_\alpha\ (i) = m_i F^2_\alpha(i)/\lambda_\alpha$$

c'est la contribution du point i à l'inertie expliquée par l'axe α;

$$\text{CO}\,2_\alpha\,(i) = \frac{F^2_\alpha(i)}{\|X^i\|^2}$$

c'est le carré du cosinus de l'angle de X^i avec l'axe α.

On remarque que CONTR (i) ne dépend pas des distributions m_I et m_J; il est de même de CTR_α (i) et de CO2_α (i) : on démontre en effet facilement, en utilisant les relations (2), que :

$$\text{CTR}_\alpha\,(i) = w^2_{\alpha i}$$

et :

$$CO\,2_{\alpha}(i) = \frac{\lambda_{\alpha}\, w_{\alpha i}^2}{\sum_j t_{ij}^2}$$

A l'exception de POIDS (i), toutes les aides à l'interprétation de l'analyse factorielle ne dépendent donc que du tableau T. Les distributions m_I et m_J servent uniquement à construire les nuages X et Y, et à calculer les coordonnées $F_{\alpha}(i)$ et $G_{\alpha}(j)$ qui permettent de visualiser les résultats de l'analyse factorielle en éditant des graphes présentant la projection des nuages sur des couples d'axes factoriels. Dans les applications, nous prendrons en général $m_i = m_j = 1$ (ce qui a l'avantage de simplifier le formulaire ci-dessus), sauf lorsqu'un autre choix pour m_I et m_J permet de donner une signification géométrique plus intéressante à X et à Y : c'est notamment le cas en analyse des correspondances.

2.2. Application à l'analyse des correspondances

En analyse des correspondances,

$$t^2 = \sum_{ij} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \sum_{ij} \left(\frac{f_{ij}}{\sqrt{f_i f_j}} - \sqrt{f_i f_j} \right)^2$$

prenons :

$$m_I = f_I \text{ et } m_J = f_J$$

on trouve :

$$x_j^i = \frac{f_{ij}}{f_i \sqrt{f_j}} - \sqrt{f_j} \quad \text{et} \quad y_i^j = \frac{f_{ij}}{f_j \sqrt{f_i}} - \sqrt{f_i}$$

En utilisant la métrique euclidienne canonique, la matrice d'inertie du nuage X est V = T'T avec :

$$t_{ij} = \frac{f_{ij}}{\sqrt{f_i f_j}} - \sqrt{f_i f_j}$$

Les points X^i appartiennent à l'hyperplan d'équation $\Sigma\, x_j \sqrt{f_j} = 0$; le vecteur G, de coordonnées $g_j = \sqrt{f_j}$, est donc vecteur propre de V associé à la valeur propre zéro. Les autres vecteurs propres de V lui sont orthogonaux : il en découle que, si $\lambda > 0$, $\sum_j u_j \sqrt{f_j} = 0$ $\left(\text{et de même } \sum_i w_i \sqrt{f_i} = 0\right)$, ce qui permet de simplifier l'écriture des relations (1) et (3) qui deviennent :

(1') $$u_j = \frac{1}{\sqrt{\lambda}} \sum_i \frac{f_{ij}}{\sqrt{f_i f_j}} w_i \quad \text{et} \quad w_i = \frac{1}{\sqrt{\lambda}} \sum_j \frac{f_{ij}}{\sqrt{f_i f_j}} u_j$$

(3') $$F(i) = \frac{1}{\sqrt{\lambda}} \sum_j \frac{f_{ij}}{f_i} G(j) \quad \text{et} \quad G(j) = \frac{1}{\sqrt{\lambda}} \sum_i \frac{f_{ij}}{f_j} F(i)$$

La relation (2) est inchangée; (4) et (5) deviennent :

$$(4') \qquad \frac{f_{ij}}{\sqrt{f_i f_j}} = \sqrt{f_i f_j} + \sum_\alpha w_{\alpha i} u_{\alpha j} \sqrt{\lambda_\alpha}$$

$$(5') \qquad f_{ij} = f_i f_j \left[1 + \sum_\alpha \frac{1}{\sqrt{\lambda_\alpha}} \mathrm{F}_\alpha(i)\, \mathrm{G}_\alpha(j) \right]$$

L'expression des aides à l'interprétation est inchangée. Nous avons donc reconstitué, à partir de l'expression de t^2 particulière à cette méthode, l'essentiel du formulaire de l'analyse des correspondances.

On peut aisément reconstruire par le même procédé le formulaire de l'analyse en composantes principales.

2.3. Formulaire de l'analyse factorielle sphérique

Nous sommes maintenant en mesure d'écrire facilement le formulaire de l'analyse factorielle sphérique. Supposons que nous désirions comparer deux tableaux f_{IJ} et φ_{IJ}, et que nous choisissions deux distributions auxiliaires m_I et m_J. Les matrices à diagonaliser sont TT' et T'T, avec :

$$t_{ij} = \sqrt{f_{ij}} - \sqrt{\varphi_{ij}}$$

les coordonnées de X^i sont :

$$x^i_j = \sqrt{f_{ij}/m_i} - \sqrt{\varphi_{ij}/m_i}$$

celles de Y^j sont :

$$y^j_i = \sqrt{f_{ij}/m_j} - \sqrt{\varphi_{ij}/m_j}$$

Le formulaire devient celui de l'encadré ci-dessous.

Formules de transition entre vecteurs propres :

(1) $u_j = \frac{1}{\sqrt{\lambda}} \sum_i (\sqrt{f_{ij}} - \sqrt{\varphi_{ij}})\, w_i$ et $w_i = \frac{1}{\sqrt{\lambda}} \sum_j (\sqrt{f_{ij}} - \sqrt{\varphi_{ij}})\, u_j$

Relations entre facteurs et vecteurs propres :

(2) $u_j = \mathrm{G}\,(j) \sqrt{\frac{m_j}{\lambda}}$ et $w_i = \mathrm{F}\,(i) \sqrt{\frac{m_i}{\lambda}}$

Formules de transition entre facteurs :

(3) $\mathrm{F}\,(i) = \frac{1}{\sqrt{\lambda}} \sum_j (\sqrt{f_{ij}} - \sqrt{\varphi_{ij}})\, \mathrm{G}\,(j) \sqrt{\frac{m_j}{m_i}}$

$\mathrm{G}\,(j) = \frac{1}{\sqrt{\lambda}} \sum_i (\sqrt{f_{ij}} - \sqrt{\varphi_{ij}})\, \mathrm{F}\,(i) \sqrt{\frac{m_i}{m_j}}$

Formules de reconstitution :

(4) $\sqrt{f_{ij}} = \sqrt{\varphi_{ij}} + \sum_\alpha w_{\alpha i}\, u_{\alpha j} \sqrt{\lambda_\alpha}$

(5) $\sqrt{f_{ij}} = \sqrt{\varphi_{ij}} + \sqrt{m_i m_j} \sum_\alpha \frac{1}{\sqrt{\lambda_\alpha}} \mathrm{F}_\alpha(i)\, \mathrm{G}_\alpha(j)$

2.4. Comparaison d'un tableau au produit de ses marges par l'analyse sphérique

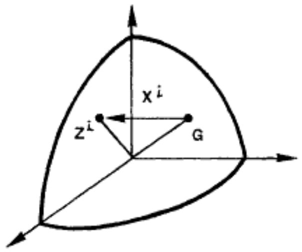
Posons :

$$\varphi_{\mathrm{IJ}} = f_{\mathrm{I}} f_{\mathrm{J}},\ m_{\mathrm{I}} = f_{\mathrm{I}},\ m_{\mathrm{J}} = f_{\mathrm{J}}$$
$$d^2(f, \varphi) = \sum_{ij} (\sqrt{f_{ij}} - \sqrt{f_i f_j})^2$$

Le formulaire de l'analyse factorielle associée à cette expression s'obtient aisément à partir des résultats du paragraphe précédent. On peut retrouver par développement limité le formulaire de l'analyse des correspondances, en posant $f_{ij} = f_i f_j\,(1 + \varepsilon_{ij})$ et en supposant ε_{ij} petit : ceci découle de l'équivalence entre la métrique de Hellinger et la métrique du χ^2, que nous avons établie en 1.3.

Il est intéressant d'examiner à titre d'exemple la signification géométrique des nuages de points construits lors de cette application. Les coordonnées des points X^i sont $x_j^i = \sqrt{f_{ij}/f_i} - \sqrt{f_j}$; le point G, de coordonnées $g_j = \sqrt{f_j}$, représente sur la sphère la distribution f_{J}; le point Z^i, de coordonnées $z_j^i = \sqrt{f_j^i}$, représente la distribution conditionnelle de J sur la ligne i du tableau f (rappelons que $f_j^i = f_{ij}/f_i$). Le vecteur X^i représente donc la différence entre les distributions f_J et f_J^i (graphique 5).

GRAPHIQUE 5



Il est possible de construire de façon analogue le nuage Y.

Les nuages X et Y *ne sont pas centrés*; notons L le centre de gravité des points Z^i; la coordonnée courante de L est :

$$l_j = \sum_i f_i z_j^i = \sum_i \sqrt{f_i f_{ij}}$$

On remarquera que $d^2(f, \varphi)$ peut s'écrire :

$$d^2(f, \varphi) = 2\left(1 - \sum_{ij} \sqrt{f_{ij} f_i f_j}\right)$$
$$= 2\,(1 - \langle G, L \rangle)$$

en notant $\langle G, L \rangle$ le produit scalaire ordinaire des vecteurs G et L.

Chacun des nuages X et Y est muni de la distance euclidienne ordinaire

$$\| X^i - X^{i'} \|^2 = \sum_j \left(\sqrt{f_j^i} - \sqrt{f_j^{i'}}\right)^2$$

La distance entre les points X^i et $X^{i'}$ est donc la longueur de la corde qui joint les deux points Z^i et $Z^{i'}$ sur la sphère.

Nous allons établir le résultat suivant (équivalence distributionnelle) : si deux points X^i et $X^{i'}$ du nuage X ont les mêmes coordonnées (c'est-à-dire si, pour tout j, $f_j^i = f_j^{i'}$), le fait d'agréger les deux lignes i et i' du tableau f en une ligne unique i'' telle que $f_{i''j} = f_{ij} + f_{i'j}$ (c'est-à-dire le fait de remplacer X^i et $X^{i'}$ par un point $X^{i''}$ de mêmes coordonnées et muni de la masse $f_{i''} = f_i + f_{i'}$) ne modifie pas les distances entre deux points quelconques Y^j et $Y^{j'}$ du nuage Y.

Cette propriété est vérifiée par l'analyse des correspondances; elle est considérée comme l'une des caractéristiques les plus importantes de cette méthode. Elle est également vérifiée par l'analyse sphérique dans le cas où le tableau de référence φ est égal à $f_I\, f_J$, et aussi, selon un résultat établi par P. KAMINSKI [4], par toutes les analyses factorielles réalisées à partir de

t_{ij} homogènes et de degré 1/2 en f_{ij} et en $f_i\, f_j$; voici la démonstration :

Supposons que $t_{ij} = t\,(f_{ij},\, f_i\, f_j)$ soit une fonction homogène de degré r en f_{ij} et en $f_i\, f_j$. Le carré de la distance entre Y^j et $Y^{j'}$ est :

$$\| Y^j - Y^{j'} \|^2 = \sum_i \left(\frac{t_{ij}}{\sqrt{m_j}} - \frac{t_{ij'}}{\sqrt{m_{j'}}}\right)^2$$

Dans cette expression, les modalités i et i' interviennent dans deux termes dont la somme est :

$$\left(\frac{t_{ij}}{\sqrt{m_j}} - \frac{t_{ij'}}{\sqrt{m_{j'}}}\right)^2 + \left(\frac{t_{i'j}}{\sqrt{m_j}} - \frac{t_{i'j'}}{\sqrt{m_{j'}}}\right)^2$$

Lorsqu'on remplace les lignes i et i' par la ligne i'', on obtient un nouveau nuage Y et, dans l'expression de la distance entre Y^j et $Y^{j'}$, les deux termes ci-dessus sont remplacés par un terme unique :

$$\left(\frac{t_{i''j}}{\sqrt{m_j}} - \frac{t_{i''j'}}{\sqrt{m_{j'}}}\right)^2$$

Pour que $\|Y^j - Y^{j'}\|^2$ ne soit pas modifié par l'agrégation de i et i' en i'', il faut et il suffit que ce terme soit égal à la somme précédente. Montrons quelle est la condition de cette égalité : comme t_{ij} est homogène de degré r en f_{ij} et en $f_i\, f_j$,

$$t_{ij} = (f_i)^r\, t\,(f^i_j, f_j)$$

Comme :

$$f^i_j = f^{i'}_j = f^{i''}_j$$

pour tout j, la condition devient simplement :

$$f^{2r}_i + f^{2r}_{i'} = (f_i + f_{i'})^{2r},$$

qui n'est une identité que si $r = 1/2$.

On peut se demander si l'analyse sphérique respecte l'équivalence distributionnelle, y compris lorsque le tableau de référence φ est quelconque : il est relativement aisé d'établir que ce n'est pas le cas. Pour que l'agrégation de deux points X^i et $X^{i'}$ ne modifie pas les distances dans le nuage Y, il faut non seulement que ces deux points soient confondus, mais aussi que :

$$\frac{f_{ij}}{\sqrt{m_i}} = \frac{f_{i'j}}{\sqrt{m_{i'}}} \text{ et } \frac{\varphi_{ij}}{\sqrt{m_i}} = \frac{\varphi_{i'j}}{\sqrt{m_{i'}}} \text{ pour tout } j$$

2.5. Classification sur I ou J associée à l'analyse factorielle sphérique

Il est souvent utile d'associer à une analyse factorielle une classification ascendante hiérarchique. Le principe en est le suivant : si l'on considère le nuage X, plongé dans un espace doté d'une métrique euclidienne $\|X^i - X^{i'}\|^2$, et auquel est associée une distribution de masses m_I, on calcule pour chaque couple (i, i') l'inertie de l'« haltère » formé par les deux points X^i et $X^{i'}$; cette inertie est $\delta^2(i, i')$, avec :

$$\delta^2(i, i') = \frac{m_i m_{i'}}{m_i + m_{i'}} \| X^i - X^{i'} \|^2$$

Si l'on agrège les deux points i et i' en un point i'' placé au centre de gravité de l'haltère, et si l'on dote i'' de la masse $m_{i''} = m_i + m_{i'}$, on provoque dans le nuage X une perte d'inertie égale à $\delta^2(i, i')$. La démarche de la classification ascendante hiérarchique est de procéder à une succession d'agrégations binaires, portant à chaque étape sur le couple de points pour lequel $\delta^2(i, i')$ est minimum.

Par analogie, nous noterons $\delta^2(i, i')$ la diminution de $d^2(f, \varphi)$ entraînée par l'agrégation de i et i' en une modalité unique i'' — c'est-à-dire par la suppression des lignes f_{ij} et $f_{i'j}$ de f et leur remplacement par une ligne $f_{i''j} = f_{ij} + f_{i'j}$, associés à une transformation analogue de φ.

Comme :

$$d^2(f, \varphi) = 2\left(1 - \sum_{ij} \sqrt{f_{ij}\,\varphi_{ij}}\right)$$

$$\boxed{\delta^2(i, i') = 2 \sum_{j} \left[\sqrt{(f_{ij} + f_{i'j})(\varphi_{ij} + \varphi_{i'j})} - \sqrt{f_{ij}\,\varphi_{ij}} - \sqrt{f_{i'j}\,\varphi_{i'j}}\right]}$$

On vérifie aisément que $\delta^2(i, i') \geqslant 0$; on voit aussi que $\delta^2(i, i') = 0$ si, pour tout j, $f_{i'j}/f_{ij} = \varphi_{i'j}/\varphi_{ij}$: l'agrégation de i et i' ne fait perdre aucune inertie si les lignes i et i' sont dans les mêmes rapports dans le tableau f et dans le tableau φ. D'après 2.4, ceci sera le cas si l'analyse sphérique respecte l'équivalence distributionnelle pour le couple (i, i').

2.6. Premières applications de l'analyse factorielle sphérique

Il arrive fréquemment qu'un statisticien ait à comparer deux tableaux de fréquences : par exemple, une enquête réalisée selon une périodicité régulière sur une population déterminée donne lieu à des exploitations croisant les caractères I et J, et il s'agit de comparer entre elles deux exploitations successives. Si l'on note $f^{(1)}_{IJ}$ et $f^{(2)}_{IJ}$ les tableaux obtenus en faisant les exploitations relatives aux dates 1 et 2, on pourra les comparer en faisant l'analyse sphérique de leur différence

$$\sum_{ij} \left(\sqrt{f^{(1)}_{ij}} - \sqrt{f^{(2)}_{ij}}\right)^2$$

De même, si lors de l'exploitation on produit deux tableaux différents relatifs chacun à une sous-population (par exemple tableau « hommes » et tableau « femmes » dans une enquête démographique), on pourra comparer ces deux tableaux. Un exemple d'une telle comparaison est donné en 4.2.

D'une manière générale, l'analyse sphérique semble un bon instrument pour comparer entre eux deux tableaux définis sur le même ensemble $I \times J$. On se rappellera que le choix des distributions auxiliaires m_I et m_J comporte une part d'arbitraire, et qu'il n'influe pas l'interprétation des axes. On prendra sauf exception $m_i = m_j = 1$.

2.7. Comparaison d'un tableau avec le « Tableau nul »

Nous avons toujours jusqu'ici supposé :

$$\sum_{ij} | \varphi_{ij} | = 1$$

Il peut être cependant intéressant, dans certaines applications, de poser $\varphi_{ij} = 0$ pour tous les couples (i, j) : nous dirons que l'on compare alors f au « tableau nul ».

Si l'on prend $m_i = m_j = 1$, les points ont pour coordonnées :

$$x_j^i = y_j^i = \sqrt{f_{ij}}$$

la matrice d'inertie de X a pour terme courant :

$$v_{jj'} = \sum_i \sqrt{f_{ij} f_{ij'}}$$

Si l'on prend $m_i = f_i$ et $m_j = f_j$, on a :

$$x_j^i = \sqrt{f_j^i} \text{ et } y_i^j = \sqrt{f_i^j}$$

les points X^i et X^j représentent sur la sphère les distributions f_J^i et f_I^j. Le premier axe factoriel est forcément dirigé dans l'orthant positif si l'on considère une distribution dont tous les termes sont non négatifs. Comme $t_{ij} = \sqrt{f_{ij}}$, l'équivalence distributionnelle est respectée d'après le résultat établi en 2.4.

Les formules de transition s'écrivent :

$$F(i) = \frac{1}{\sqrt{\lambda}} \sum_j G(j) \sqrt{f_j f_j^i}$$

$$G(j) = \frac{1}{\sqrt{\lambda}} \sum_i F(i) \sqrt{f_i f_i^j}$$

et la formule de reconstitution devient :

$$\sqrt{f_{ij}} = \sqrt{f_i f_j} \left[\sum_\alpha \frac{1}{\sqrt{\lambda_\alpha}} F_\alpha(i) \, G_\alpha(j) \right]$$

Il est facile de démontrer que, si f est très proche de $f_I f_J$, l'analyse factorielle du nuage avec :

$$x_j^i = \sqrt{f_j^i} \text{ et } m_i = f_i$$

donnera, comme premier axe factoriel, un axe très voisin de G (de coordonnées $\sqrt{f_j}$) et correspondant à une valeur propre très proche de 1 : on retrouve l'axe trivial de l'analyse des correspondances. Le plan (2, 3) de cette analyse sphérique donnera une image semblable à celle du plan (1, 2) de l'analyse des correspondances. Nous verrons en 3.2 que l'on peut donner encore une autre interprétation à cette analyse.

2.8. Analyse d'un nuage vu d'un point

Supposons que, dans le nuage X, un point du nuage (que nous noterons X^0) joue un rôle particulièrement important. Ce sera notamment le cas lorsque I est la date d'une observation : la dernière observation disponible joue souvent un rôle privilégié dans les commentaires du statisticien. Considérons par exemple une succession de comptes trimestriels repérés par leurs dates i, le caractère J servant à repérer des postes comptables. Il peut être intéressant pour un économiste de regarder les comptes du passé en se situant dans l'instant présent, c'est-à-dire en prenant comme point de vue le dernier point observé. On remarquera qu'il n'est pas possible de réaliser une analyse factorielle du nuage en utilisant la métrique du χ^2 centrée sur ce point si celui-ci est situé sur l'un des bords du simplexe (la distribution qu'il représente comporte alors une fréquence nulle). Par contre en utilisant la métrique de Hellinger cette limitation disparaît.

Notons f_J^0 la structure de ce point. On comparera f_{IJ} au tableau :

$$\varphi_{IJ} = f_J^0 \times f_I$$

l'expression de $d^2(f, \varphi)$ est :

$$d^2(f, \varphi) = \sum_{ij} \left(\sqrt{f_{ij}} - \sqrt{f_i f_{0j} / f_0}\right)^2$$

Choisissons $m_I = f_I$ et $m_J = f_J$; les points X^i ont pour coordonnées :

$$x_j^i = \sqrt{f_j^i} - \sqrt{f_j^0}$$

X^i est la différence entre les distributions f_J^i et f_J^0 repérées sur la sphère. Les points Y^j ont pour coordonnées :

$$y_j^i = \sqrt{f_j^i} - \sqrt{f_j^0 f_i / f_0}$$

Si l'on veut reproduire de façon plus fidèle encore le comportement de l'économiste lorsqu'il situe une observation récente par rapport au passé, il faut ajouter encore ceci : le passé pris en compte s'estompe d'autant plus qu'il est plus lointain, à une vitesse qui dépend d'ailleurs de l'horizon temporel dans lequel se situe le raisonnement. Pour tenir compte de ce phénomène, nous introduisons dans l'expression de la distance entre f et φ un facteur qui diminue exponentiellement l'influence des périodes anciennes (ici l'indice i est remplacé par un indice t, croissant avec l'ancienneté).

On remplace alors le tableau f_{tj} par un tableau :

$$f'_{tj} = Ce^{-at} f_{tj}$$

le terme C étant tel que :

$$\sum_{ij} f'_{ij} = 1$$

Cela revient à prendre comme expression de $d^2(f', \varphi)$:

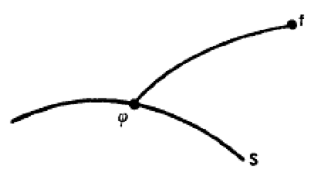
$$d^2(f', \varphi) = C \sum_{ij} e^{-at} \left(\sqrt{f_{tj}} - \sqrt{f_t f_{0j} / f_0}\right)^2$$

On obtient des mises en perspective « courte » ou « longue » selon la valeur accordée au coefficient a. Cet exemple montre à quel point l'analyse sphérique peut s'adapter, de façon très souple, aux besoins d'un utilisateur. Nous décrivons une application de ce procédé en 4.3.

3 Comparaison d'une distribution concrète et d'une structure de distributions

Supposons définies des contraintes qui délimitent sur la sphère un sous-ensemble S, que nous désignerons comme une structure de distributions. On peut se poser le problème suivant : étant donnée une distribution concrète f, quelle est la distribution $\varphi \in S$ telle que $d^2(f, \varphi)$ soit minimum (graphique 6)?

GRAPHIQUE 6



(On remarque que la distance de Hellinger, la distance géodésique et l'information discriminante d'ordre 1/2 étant toutes trois des fonctions décroissantes du produit scalaire $\Sigma \sqrt{f_i \varphi_i}$, la distribution qui rend maximal ce produit scalaire peut être interprétée comme « la plus proche de f » de ces trois points de vue simultanément).

26

Dans le cas où les distributions considérées sont définies sur un couple de caractères I $\times$ J, le problème devient celui de l'ajustement d'un tableau sous contraintes. De plus, une fois φ identifié, on peut procéder à l'étude factorielle de la distance $d^2(f, \varphi)$; elle permet d'étudier, de façon détaillée, l'écart entre le tableau concret f et la structure S. Cette démarche est utilisable de façon très générale, en partant d'expressions de la distance entre tableaux qui peuvent être très variées; nous explorerons ici ses résultats lorsque cette distance est calculée à l'aide de la métrique de Hellinger.

3.1. Solution générale du problème

Supposons que l'on impose à la distribution les M contraintes[3] $F_m(\varphi) = 0$. Nous devons rechercher les extrema de :

$$L = \sum_i \sqrt{f_i \varphi_i} + \sum_m \lambda_m F_m(\varphi) \;;$$

on obtient :

$$\frac{\partial L}{\partial \varphi_i} = \frac{1}{2}\sqrt{\frac{f_i}{\varphi_i}} + \sum_m \lambda_m \frac{\partial F_m}{\partial \varphi_i} = 0$$

D'où un système de Card I + M équations à Card I + M inconnues. Il faudra s'assurer dans chaque cas particulier que l'extremum correspond bien à un maximum de $\Sigma \sqrt{f_i \varphi_i}$.

3.2. Ajustement à la structure « Produit de deux distributions »

Supposons que S soit l'ensemble des distributions ψ telles que $\psi_{ij} = \psi_i \psi_j$; quelle est la plus proche de f au sens de $d^2(f, \varphi)$?

Contrairement à ce que pourrait suggérer un raisonnement hâtif, la solution n'est pas $\varphi_{ij} = f_i f_j$. Ce problème va nous permettre de donner une nouvelle interprétation à la comparaison de f au tableau nul.

Il s'agit de trouver φ_i et φ_j qui rendent maximum

$$\sum_{ij} \sqrt{f_{ij} \varphi_i \varphi_j}$$

sous les contraintes

$$\sum_i \varphi_i = \sum_j \varphi_j = 1$$

3. Nous ne nous soucierons pas de la contrainte $\varphi_i \geqslant 0$, puisque, comme nous l'avons vu en 1.4, on peut représenter sur la sphère des distributions comportant des éléments négatifs.

Posons :

$$u_j = \sqrt{\varphi_j}\,,\ \ w_i = \sqrt{\varphi_i}\,,\ \ t_{ij} = \sqrt{f_{ij}}\ ;$$

en usant de notations évidentes, le problème peut s'écrire ainsi : maximiser W' TU, sous les contraintes U' U = 1 et W' W = 1.

Nous devons rechercher les extrema de

$$L = W'\,TU + \alpha\, U'U + \beta\, W'W$$

on obtient :

$$\frac{\partial L}{\partial U} = T'W + 2\alpha\, U = 0$$

$$\frac{\partial L}{\partial W} = TU + 2\beta\, W = 0$$

donc

$$W'TU = -2\alpha = -2\beta\ ;\ \text{posons}\ W'TU = \sqrt{\lambda}$$

On trouve $U = \frac{1}{\sqrt{\lambda}}\,T'\,W$ et $W = \frac{1}{\sqrt{\lambda}}\,TU$, d'où : $T'\,TU = \lambda U$ et $TT'\,W = \lambda W$.

Les solutions sont les vecteurs propres respectifs de T' T et de TT' associés à la plus grande valeur propre. On reconnaît en TT' et T' T les matrices d'inertie rencontrées en analyse sphérique lorsqu'on compare f au tableau nul; $\sqrt{\varphi_i}$ et $\sqrt{\varphi_j}$ sont alors les coordonnées des premiers axes factoriels des nuages X et Y.

Pour comparer un tableau à la structure « produit de deux distributions », il suffit donc de faire son analyse sphérique en le comparant au tableau nul; les premiers axes obtenus lors de cette analyse permettent d'obtenir la distribution $\varphi = \varphi_i\, \varphi_j$ la plus proche de f au sens de la métrique de Hellinger. Nous appellerons « distributions centrales » les distributions φ_i et φ_j : cette appellation rappelle que les premiers axes passent « au centre » des nuages de points, et souligne que φ_i et φ_j sont différents des distributions marginales f_i et f_j. Les axes suivants sont les mêmes que ceux que l'on obtiendrait en faisant l'analyse factorielle à partir de $d^2\,(f, \varphi)$.

3.3. Ajustement sur une distribution comportant des fréquences nulles

Supposons S définie ainsi : $\varphi_i = 0$ pour $i \in K$, K étant un sous-ensemble de I. Le système d'équations devient, en remarquant que $\Sigma\varphi_i = 1$ pour $i \in I - K$:

$$\begin{cases} \varphi_i = 0 \text{ pour } i \in K \\ \sqrt{f_i} + 2\lambda\,\sqrt{\varphi_i} = 0 \text{ pour } i \in I - K \end{cases}$$

Il en découle que, pour $i \in \mathrm{I}-\mathrm{K}$, $\varphi_i = \frac{f_i}{\sum\limits_{k \in \mathrm{I}-\mathrm{K}} f_k}$.

A titre d'exemple, parmi les distributions sur $\mathrm{I} \times \mathrm{I}$, l'ensemble des tableaux diagonaux est défini par les contraintes $\varphi_{ii'} = 0$ si $i \neq i'$. Le tableau diagonal le plus proche d'un tableau concret donné a pour terme courant de sa diagonale $\varphi_{ii} = f_{ii}/\Sigma f_{kk}$: c'est le tableau diagonal dans lequel les termes diagonaux sont dans les mêmes proportions que ceux de f. Ce cas particulier donne occasion à des applications qui nous paraissent parmi les plus intéressantes de l'analyse sphérique : le traitement des tableaux de transition et d'échanges.

a. *Les tableaux de transition*

Nous entendons sous cette appellation des tableaux donnant la ventilation d'une population selon deux caractères I_1 et I_2 représentant chacun l'observation d'un même caractère I effectuée dans des conditions différentes.

Exemples :

a. Un ensemble de ménages classés selon la CSP du père du mari (I_1) et la CSP du père de la femme (I_2);

b. Un ensemble de salariés ventilés selon leur métier à la date 1 et leur métier à la date 2;

c. Une « matrice de confusion » classant un ensemble d'expériences réalisées sur la reconnaissance des sons selon le son émis et le son identifié.

Les tableaux de transition sont habituellement carrés, car on utilise la même nomenclature pour I_1 et pour I_2. Ils comportent presque toujours une diagonale très chargée : en reprenant chacun des exemples énumérés ci-dessus, on verra que les cases de la diagonale correspondent :

— à l'endogamie;

— à la stabilité professionnelle;

— à une reconnaissance correcte des sons...

La comparaison d'un tableau de transition concret avec la structure diagonale permet de bien voir « ce qui se passe en-dehors de la diagonale », et donc d'étudier la mobilité sociale ou professionnelle, les confusions, etc., et de façon générale tous les phénomènes qui comportent un certain « brouillage » autour d'une structure globalement stable.

Il est facile de voir que, dans le cas d'un tableau de transition :

$$d^2(f, \varphi) = 2\left(1 - \sum_{ij} \sqrt{f_{ij}\varphi_{ij}}\right)$$

$$= 2\left(1 - \sum_i \frac{f_{ii}}{(\sum\limits_k f_{kk})^{1/2}}\right) = 2\left(1 - \sqrt{\Sigma f_{ii}}\right)$$

Soit en posant $\mathrm{T} = \Sigma f_{ii}$:

$$\boxed{d^2(f, \varphi) = 2\left(1 - \sqrt{\mathrm{T}}\right)}$$

Supposons que l'on agrège deux rubriques i et i' de I (s'agissant d'un tableau de transition, nous allons supposer que cette agrégation porte *à la* *fois* sur I_1 et I_2). La trace T de f va être augmentée de $f_{ii'} + f_{i'i}$, ce qui nous permet de définir aisément l'indice de distance à retenir pour la classification ascendante hiérarchique :

$$\delta^2 (i, i') = 2 \left(\sqrt{T + f_{ii'} + f_{i'i}} - \sqrt{T}\right)$$

Nous rencontrons cependant un paradoxe; la démarche usuelle de l'analyse arborescente nous conduirait à rechercher d'abord le couple tel que (i, i') soit le plus petit : c'est celui pour lequel $f_{ii'} + f_{i'i}$ est minimum. Nous serions ainsi amenés à agréger d'abord les rubriques qui ont le moins d'échanges.

En réalité, deux stratégies sont possibles : chercher à minimiser la perte d'inertie, ou au contraire à la maximiser. Chacune de ces stratégies correspond à un objectif différent :

- Minimiser la perte d'inertie : on regroupe donc ensemble celles des rubriques qui ont le moins d'échanges; par agrégations successives on fabrique des tableaux de plus en plus réduits mais dans lesquels les cases d'échanges restent aussi remplies que possible. On peut comprendre cette stratégie par une image : imaginons qu'un pays soit traversé en deux par une rivière, et que l'essentiel de ses ressources fiscales proviennent de droits de péage payés par les entreprises lorsqu'un transport passe la rivière. Si le tableau f_{ij} représente les échanges entre entreprises, la direction des impôts essaiera de les obliger à se répartir de part et d'autre de la rivière conformément à la répartition en deux classes obtenue en minimisant la perte d'inertie;
- Maximiser la perte d'inertie : cette stratégie conduit à constituer des ensembles ayant le maximum d'échanges internes, et à réduire les échanges externes. Dans l'exemple précédent, ce serait la stratégie du patronat.

On verra en 4.4 l'exemple du traitement d'un tableau de transition.

b. *Tableaux d'échanges*

Nous appellerons « tableau d'échanges » un tableau repérant des flux entre classes de I. I_1 est l'ensemble des classes « émettrices de flux », I_2 l'ensemble des classes « réceptrices de flux ».

Contrairement aux tableaux de transition, les tableaux d'échanges ne donnent pas la répartition d'une population (dont l'effectif est fixé *a priori*) selon I_1 et I_2 : le volume des flux considéré peut dépendre, nous allons le voir, de la nomenclature choisie pour I.

Par exemple, un tableau d'échange interindustriel décrit les échanges entre branches (achats et ventes d'une branche à l'autre) à l'intérieur d'une certaine économie et pendant une période déterminée. Habituellement, on utilise la même nomenclature pour les ventes et les achats. Le contenu n_{ij} de la case (i, j) du tableau représente les ventes de la branche i à la branche j.

En toute rigueur, les nombres n_{ii} situés sur la diagonale du tableau n'ont pas de signification. Si les échanges sont repérés avec la nomenclature utilisée pour construire le tableau, les échanges internes sont par construction nuls et la diagonale du tableau est donc nulle. Si les échanges sont repérés avec une nomenclature plus fine que celle du tableau, qui serait construit ensuite par agrégation, les « échanges internes » d'une branche sont la somme des échanges qui, dans la nomenclature plus fine, existent entre les sous-branches qui composent cette branche. A la limite, en augmentant indéfiniment la finesse de la nomenclature dans laquelle on observe les échanges, on fait croître jusqu'à l'infini les échanges internes d'une branche.

Entre zéro et l'infini, les échanges internes sont donc une quantité arbitraire dépendant de la nomenclature d'observation, et d'autant plus grande que cette nomenclature est plus fine.

Nous allons aborder le problème du tableau d'échange de la façon suivante : en partant des résultats obtenus pour l'étude des tableaux de transition, nous ferons tendre vers l'infini les termes de la diagonale de ce tableau. Nous mettrons alors en évidence des propriétés limites, qui nous permettront de réaliser l'analyse d'un tableau d'échange.

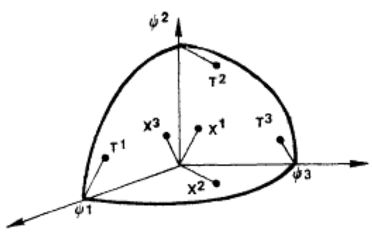
Prenons comme tableau de référence pour l'étude d'un tableau de transition un tableau diagonal de terme courant f_i.

Les lignes sont représentées par le nuage X, dont le point courant est X^i :

$$\begin{cases} x_j^i = \sqrt{f_j^i} \text{ si } i \neq j \\ x_i^i = \sqrt{f_i^i} - 1 \end{cases}$$

Notons ψ^i le vecteur dont toutes les composantes sont nulles, sauf celle de rang i qui vaut 1. Notons T^i le point qui représente sur la sphère la distribution f_j^i; on a $X^i = T^i - \psi^i$ (graphique 7).

Graphique 7



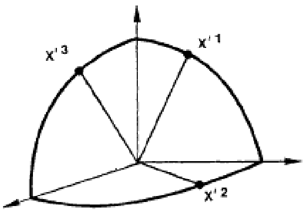
Faisons tendre n_{ii} vers l'infini. Les points X^i vont tendre vers zéro, les coordonnées de X^i étant dans des rapports du type :

$$x_1^i / x_2^i = \sqrt{n_{i1}/n_{i2}} \quad \text{si } i \neq 1 \text{ ou } 2$$

La direction limite de X^i est donc le vecteur unitaire X'^i de coordonnées :

$$\begin{cases} x'^i_j = \sqrt{n_{ij}/n'_i} & \text{si } i \neq j \quad \text{avec } n'_i = \sum\limits_{j \neq i} n_{ij} \\ x'^i_i = 0 \end{cases}$$

GRAPHIQUE 8



Considérons le tableau d'échange où la diagonale est nulle : en analysant ce tableau par comparaison avec le tableau de référence nul, on obtient le nuage de points que nous venons de construire. Il paraît logique d'associer à chacun de ces points une masse n'_i/n', si n' représente la somme des échanges en dehors de la diagonale.

L'analyse d'un tableau d'échange se fera donc de la façon suivante : *après avoir annulé la diagonale,* on procèdera à l'analyse sphérique de f_{IJ} comparé au tableau nul, en choisissant $m_I = f_I$ et $m_J = f_J$.

On trouvera en 4.5 un exemple d'analyse d'un tableau d'échange.

4 Exemples d'applications pratiques

Nous avons regroupé dans cette partie quelques exemples qui nous ont semblé bien illustrer les propriétés de l'analyse factorielle sphérique. Les données traitées ont été choisies non en raison de leur intérêt propre, mais parce qu'elles fournissaient la matière nécessaire pour des applications. On voudra donc bien ne pas nous tenir rigueur du caractère hétéroclite des exemples traités, de l'absence d'une description précise des conditions dans lesquelles les données analysées ont été recueillies, et du caractère très limité des esquisses d'interprétation que nous proposons : une exploitation complète des données que nous avons analysées aurait nécessité un texte beaucoup plus long.

Les séries de variations de stocks par produits issues des comptes trimestriels établis par l'INSEE nous ont fourni un tableau comportant des valeurs négatives (4.1) ; l'exercice de comparaison de deux tableaux de contingence a été réalisé sur deux tableaux (hommes et femmes) issus de l'exploitation de l'enquête « Emploi 1973 » (INSEE), et répartissant les personnes sorties du système éducatif selon le niveau de diplôme atteint et l'emploi occupé (4.2); l'analyse d'un nuage vu d'un point est faite à partir de données provenant des comptes trimestriels sur la valeur ajoutée par branche entre 1970 et 1978 (4.3) ; l'analyse d'un tableau de transition est faite à partir de données recueillies par l'ISUP au cours d'expériences sur la reconnaissance des sons (4.4) ; enfin, nous avons analysé un tableau d'échanges en 40 branches, extrait des comptes nationaux 1976 (4.5).

Nous rappelons que la description des aides à l'interprétation est donnée en 2.1.

4.1. Tableaux comportant des valeurs négatives

a. *Données et signification du modèle mathématique*

Nous analyserons ici un tableau de séries trimestrielles donnant la variation des stocks de divers produits de 1970 à 1978. Le tableau contient en ligne les produits, en colonne les trimestres. Le nombre k_{ij} désigne ici la variation des stocks du produit i entre le début et la fin du trimestre j. Cette variation peut évidemment être positive ou négative. Elle est mesurée en francs constants, aux prix de 1970.

Le tableau des variations de stocks, présenté en annexe A, va être comparé au tableau nul. Les nuages X (produits) et Y (trimestres) sont formés de points X^i et Y^j dotés chacun de la masse unité. La coordonnée courante de X^i est :

$$x_j^i = [\text{signe de } k_{ij}] \sqrt{|f_{ij}|}$$

avec :

$$f_{ij} = \frac{k_{ij}}{k}, \text{ où } k = \sum_{ij} |k_{ij}|$$

La matrice d'inertie de X est T'T, où T a pour terme général :

$$t_{ij} = [\text{signe de } k_{ij}] \sqrt{|f_{ij}|}$$

Il est facile de voir que l'inertie totale est égale à l'unité, en remarquant qu'elle est égale à $\Sigma \| X_i \|^2$.

L'examen de la formule de transition nous permet de comprendre les relations entre les projections des deux nuages :

$$G(j) = \frac{1}{\sqrt{\lambda}} \sum_i [\text{signe de } k_{ij}] \sqrt{|f_{ij}|}\, F(i)$$

Si la variation des stocks du produit i durant le trimestre j est négative, $k_{ij} < 0$ et il y aura un effet de rejet entre les points représentant i et j. Si la variation est positive, il y a au contraire un effet d'attraction.

La valeur absolue de la variation de stocks a elle aussi une influence. Supposons :

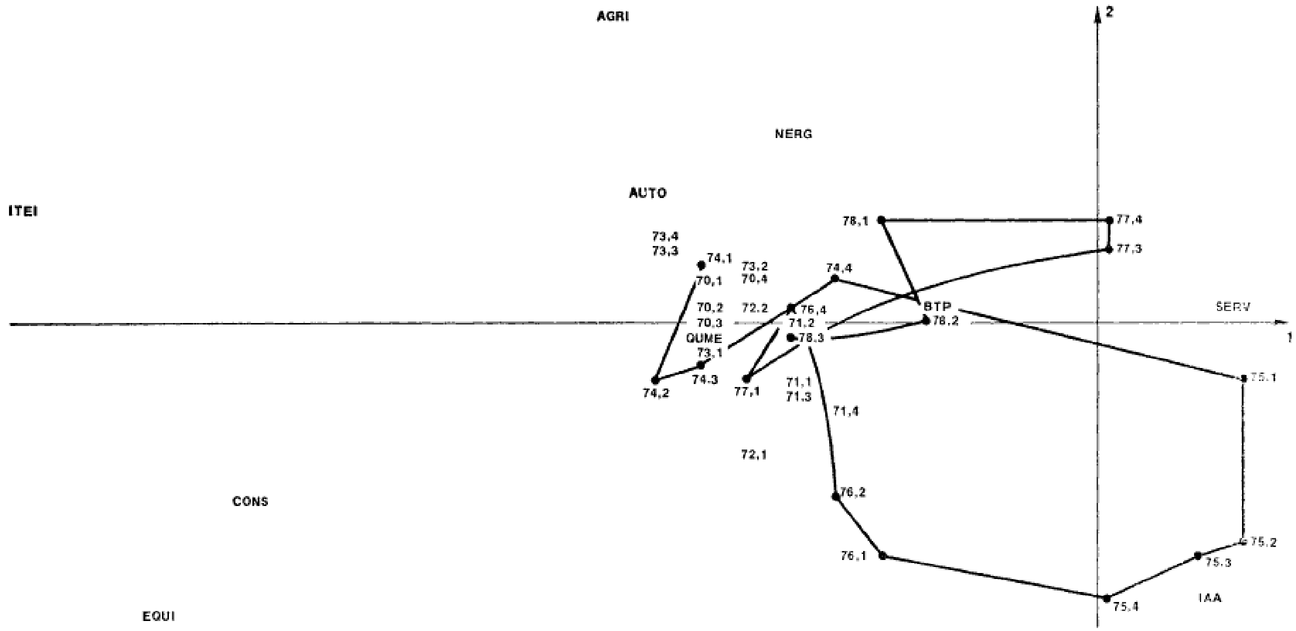
$$k_{ij_1} > k_{ij_2} > 0$$

le point représentant j_1 sera plus attiré par i que le point représentant j_2.

$$\text{Si } k_{ij} < k_{ij_2} < 0$$

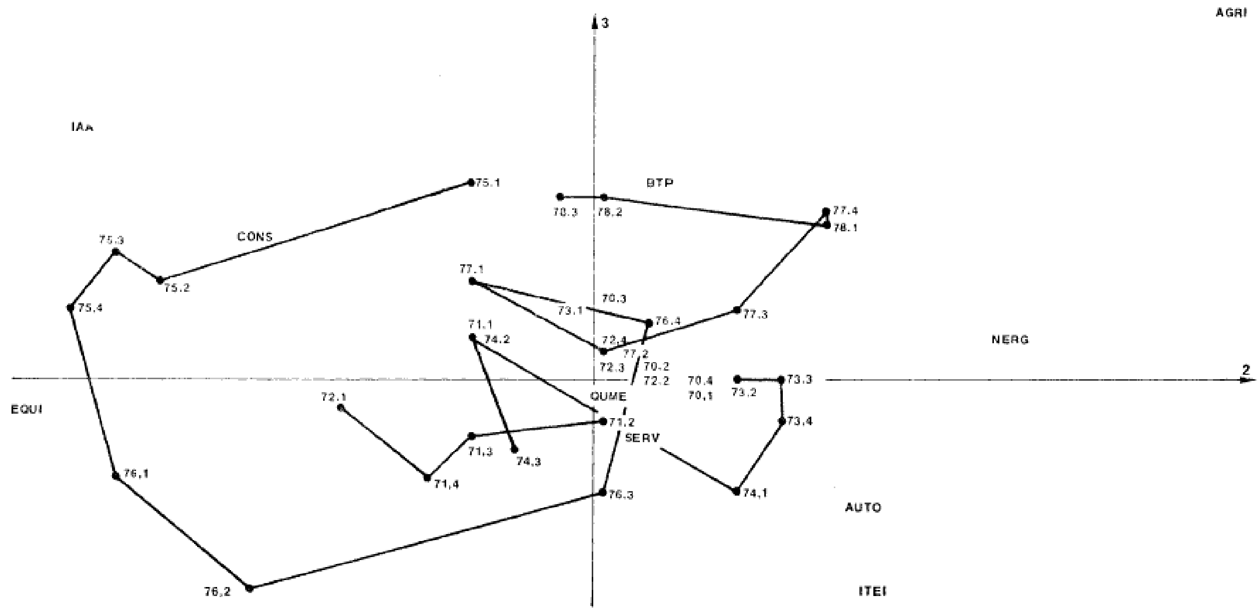
l'effet est inverse.

GRAPHIQUE 9. ***Évolution des stocks. Plan (1, 2)***



AGRI	: agriculture	QUME	: biens d'équipement des ménages
NERG	: énergie	BTP	: bâtiment et travaux publics
ITEI	: biens intermédiaires	SERV	: services
AUTO	: automobile	EQUI	: biens d'équipement professionnel
CONS	: biens de consommation	IAA	: industries agricoles et alimentaires

Graphique 10. *Évolution des stocks. Plan (2, 3)*



Pour les symboles, voir la nomenclature au bas du graphique 9.

b. ***Résultats de l'analyse***

Le tableau 1 donne les valeurs propres; le tableau des aides à l'interprétation des résultats de l'analyse est présenté en annexe A.

TABLEAU 1

Valeurs propres

Rang des axes	1	2	3	4	5	6	7
Valeurs propres	0,62	0,10	0,08	0,07	0,05	0,03	0,02
Pourcentages...............	62,4	10,0	7,6	7,1	5,0	3,4	2,3
Pourcentages cumulés.......	62,4	72,4	80,0	87,0	92,0	95,4	97,7

L'axe 1 (graphique 9) apparaît comme un axe qui oppose les périodes de stockage aux périodes de déstockage : presque tous les produits apparaissent sur cet axe avec une forte coordonnée négative, ce qui permet de repérer les trimestres où le déstockage a été important (année 1975, 1977.3 et 1977.4). On remarque que les services, qui contribuent très faiblement à l'inertie totale, ont une corrélation importante avec cet axe et une coordonnée positive, ce qui est exceptionnel. Mais on se rappellera que la notion économique de stock du produit « service » est assez conventionnelle, et l'on se gardera donc d'interpréter ce fait. La coordonnée du produit « IAA » est, elle aussi, négative; mais la corrélation de ce point avec l'axe 1 est très faible, il ne convient donc pas de lui accorder beaucoup d'importance. On peut seulement noter que l'évolution des stocks dans les IAA est non corrélée avec l'évolution d'ensemble.

Dans le plan (2,3), le produit « agriculture » joue un rôle très important et détermine l'essentiel des mouvements (graphique 10); le fort déstockage de ce produit en 1975 et début 1976 est particulièrement visible. On note sur l'axe 2 que l'évolution des stocks du produit « IAA » va à contre-courant de celle du produit « agriculture » en 1975 et 1977. On note aussi l'influence des produits « biens d'équipement » (axe 2) et « biens intermédiaires » et « BTP » (axe 3).

4.2. Comparaison de deux tableaux de contingence

a. ***Les données***

Nous avons utilisé pour cet exemple deux petits tableaux de contingence (tableaux 2 et 3) issus de l'exploitation de l'enquête « Emploi 1973 ». Ces tableaux répartissent la population des élèves scolarisés en 1972-1973, sortis du système éducatif en 1973 et ayant trouvé un emploi, selon les deux caractères I : emploi occupé, J : niveau de diplôme atteint. Le tableau 2 est relatif aux hommes, le tableau 3 aux femmes. Le nombre $k_{ij}^{(1)}$, par exemple, désigne le nombre d'hommes sortis du système éducatif en 1973, ayant atteint le niveau de diplôme j et ayant trouvé un emploi du type i.

On remarque que les distributions marginales sont différentes d'un tableau à l'autre, surtout en ce qui concerne les emplois : les ouvriers et les agriculteurs sont plus nombreux chez les hommes, les employés chez les femmes. En ce qui concerne les diplômes, le baccalauréat général et « DUT/BTS/santé » sont plus fréquents chez les femmes.

b. *Signification du modèle mathématique*

Pour interpréter l'analyse, on peut se reporter au formulaire de l'AFS fourni en 2.3; dans ce cas particulier, nous posons $m_i = m_j = 1$, car aucun autre choix ne semble s'imposer pour les distributions auxiliaires de masses. L'analyse factorielle porte donc sur deux nuages de points X et Y; à titre d'exemple, un point X^i de X est doté de la masse unité et a pour coordonnées courante dans l'espace à Card J dimensions :

$$x_j^i = \sqrt{f_{ij}^{(1)}} - \sqrt{f_{ij}^{(2)}} \quad , \quad \text{avec } f_{ij} = \frac{k_{ij}}{k}$$

x_j^i et y_i^j dépendent de la différence des fréquences du couple (i, j) dans chacun des deux tableaux. X^i représente l'ensemble de ces différences pour un emploi donné, Y^j les représente pour un diplôme donné.

La matrice d'inertie du nuage X est T'T, où T a pour terme général :

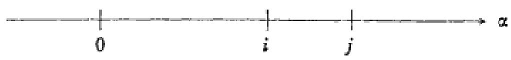
$$t_{ij} = \sqrt{f_{ij}^{(1)}} - \sqrt{f_{ij}^{(2)}}$$

Pour interpréter les résultats de l'analyse factorielle, on utilise la formule de transition qui lie les coordonnées des points X^i et Y^j sur l'axe factoriel de rang α :

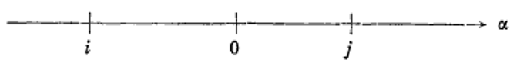
$$F_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_j \left(\sqrt{f_{ij}^{(1)}} - \sqrt{f_{ij}^{(2)}}\right) G_\alpha(j)$$

La position d'un élément i sera essentiellement liée à celles des éléments j pour lesquels la différence des fréquences du couple (i, j) est importante entre les deux populations considérées.

On remarque en outre le phénomène suivant : étant donné l'ordre dans lequel nous comparons les deux tableaux (« hommes *moins* femmes »), les proximités ou oppositions entre points sur un axe factoriel ont un sens précis :



Si $F_\alpha(i)$ et $G_\alpha(j)$ ont le même signe, l'axe indique que la fréquence du couple (i, j) est plus forte dans la population « hommes » que dans la population « femmes »; dans le cas contraire :



la fréquence de (i, j) est plus forte dans la population « femmes » que dans la population « hommes ».

Voici les deux tableaux que nous avons comparés :

TABLEAU 2

Élèves scolarisés en 1972-1973, sortis du système éducatif en 1973 et ayant trouvé un emploi — sexe masculin*

Emplois occupés*	Niveaux de diplôme								
	Sans diplôme	BEPC	BEP/CAP	BAC général	BAC technique	DEUG/ENT	DUT/BTS/Santé	SUP	Total
1. Agriculteurs (AGR)	15 068	2 701	5 709	297	1 242	–	322	–	25 339
2. Ingénieurs (ING)	–	337	309	917	–	308	–	4 383	6.254
3. Techniciens (TEC)	302	1 697	2 242	1 969	1 399	357	1 943	381	10 290
4. Ouvriers qualifiés (OQ)	10 143	3 702	30 926	314	1 861	–	–	337	47 283
5. Ouvriers non qualifiés (ONQ)	59 394	8 087	17 862	2 887	1 696	–	–	323	90 249
6. Cadres supérieurs (CSP)	596	298	892	1 227	298	2 362	318	6 781	12 772
7. Cadres moyens (CMO)	2 142	2 801	672	6 495	924	2 807	2 301	4 030	22 172
8. Employés qualifiés (EMQ)	5 445	7 348	4 719	4 353	1 280	614	982	–	24 741
9. Employés non qualifiés (EMNQ)	4 879	4 987	1 514	3 478	886	1 326	–	661	17 431

* *Source :* « Bilan formation-emploi 1973 », CEREQ, INSEE, SEIS, volume D 59 des *Collections de l'INSEE*, p. 102 et 103.

Tableau 3

Élèves scolarisés en 1972-1973, sortis du système éducatif en 1973 et ayant trouvé un emploi — sexe féminin*

Emplois occupés*	Niveaux de diplôme								
	Sans diplôme	BEPC	BEP/CAP	BAC général	BAC technique	DEUG/ENT	DUT/BTS/Santé	SUP	Total
1. Agriculteurs (AGR)	5 089	1 212	1 166	–	–	–	–	–	7 467
2. Ingénieurs (ING)	–	–	–	316	–	–	304	1 033	1 653
3. Techniciens (TEC)	281	–	320	320	283	–	683	–	1 887
4. Ouvriers qualifiés (OQ)	7 470	1 859	4 017	1 752	657	–	285	–	16 040
5. Ouvriers non qualifiés (ONQ)	29 997	4 334	4 538	1 882	–	–	–	–	40 751
6. Cadres supérieurs (CSP)	–	–	–	2 236	595	911	569	6 788	11 099
7. Cadres moyens (CMO)	1 577	1 806	4 549	17 063	875	4 152	15 731	3 991	49 744
8. Employés qualifiés (EMQ)	21 616	19 915	32 452	16 137	5 865	1 256	3 332	1 286	101 859
9. Employés non qualifiés (EMNQ)	19 849	7 325	6 484	5 111	898	294	635	–	40 596

* Voir la source dans la note du tableau 2.

Le phénomène aurait été inversé si l'on avait comparé les tableaux dans l'ordre « femmes *moins* hommes ».

La démarche de l'analyse apparaît dès lors clairement : nous allons rechercher, pour chaque axe factoriel, les points *i* et *j* dont les contributions relatives sont les plus fortes : ce sont ceux qui ont joué le plus grand rôle dans la détermination de l'axe; nous examinerons aussi ceux qui sont bien représentés sur l'axe, et dont le CO2 est fort. Ensuite nous interpréterons les proximités et les oppositions comme il a été indiqué ci-dessus.

c. *Résultats de l'analyse*

Le listage des valeurs propres nous indique que 70 % des différences sont représentés sur le premier axe, 12 % sur le deuxième et 8 % sur le troisième.

Le tableau des aides à l'interprétation des résultats de l'analyse est présenté dans l'annexe B.

L'axe 1 (graphique 11) est caractérisé par les diplômes BECA (signe —) [4] et SSD (—) et par les emplois EMQ (+), OQ (—) et ONQ (—) : le phénomène le plus important est donc que la proportion du couple (BEP/CAP, employé qualifié) est nettement plus importante chez les femmes (12 %) que chez les hommes (1,8 %); les différences notables repérées par cet axe sont les suivantes :

Tableau 4

Proportions dans chaque sexe

En %

Formation :	Sans diplôme			BEP/CAP		
Emploi :	ONQ (Ouvriers non qualifiés)	OQ (Ouvriers qualifiés)	EMQ (Employés qualifiés)	ONQ	OQ	EMQ
Hommes	4	7	2	12	7	1,8
Femmes	2,8	1,7	8	1,5	1,7	12

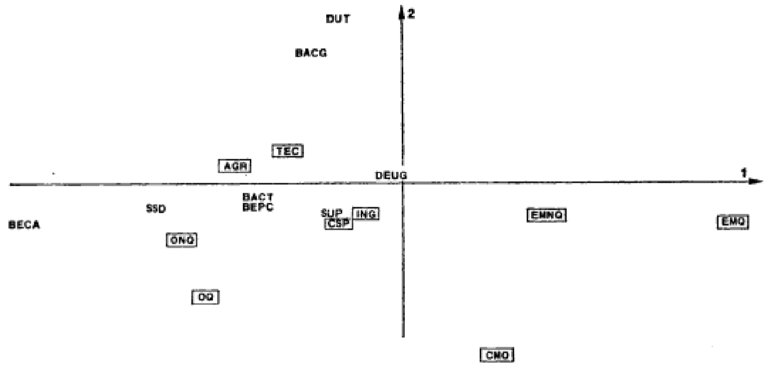
Dans les deux catégories « sans diplômes » et « BEP ou CAP », la proportion des femmes est nettement plus réduite dans les emplois ouvriers, et plus forte dans les emplois d'employé qualifié.

L'axe 2 est caractérisé par les diplômes DUT (+) et BAC G (+), et par les emplois CMO (—) et OQ (—). En fait, la proportion d'ouvriers qualifiés titulaires d'un DUT est faible dans les deux populations (0 chez les hommes, 0,1 % chez les femmes); les ouvriers qualifiés ayant un baccalauréat général sont rares, mais plus nombreux chez les femmes que chez les hommes (0,6 % contre 0,1 %). L'essentiel de la différence expliquée par cet axe porte sur les cadres moyens (tableau 5).

4. Voir la signification des codes dans les tableaux 2 ou 3.

GRAPHIQUE 11

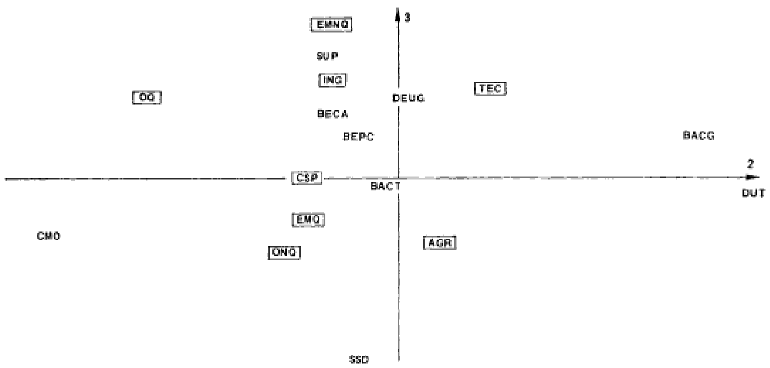
Comparaison de deux tableaux
Plan (1, 2)



On encercle les professions
Pour les abréviations des emplois, voir les tableaux 2 et 3.
En ce qui concerne les diplômes atteints : BECA = BEP ou CAP ; BACG = baccalauréat général ; BACT = baccalauréat technique ; DEUG = DEUG ou ENI ; DUT = DUT ou BTS ou Santé.

GRAPHIQUE 12

Comparaison de deux tableaux
Plan (2, 3)



Pour les symboles et la légende, voir les notes au bas du graphique 11.

TABLEAU 5

Proportion dans chaque sexe

En %

Emploi :	Cadres moyens	
Formation :	BAC G	DUT
Hommes	2,5	0,9
Femmes	6,3	5,8

Sur *l'axe 3* (graphique 12), on remarque surtout la formation SSD (—) et l'emploi EMNQ (+) : le couple « sans diplôme, employé sans qualification » est plus fréquent chez les femmes (7,3 %) que chez les hommes (1,9 %).

Nous n'irons pas plus loin dans l'étude de cet exemple, cité ici à titre d'exercice plutôt que par son intérêt propre. Il nous semble clair que la comparaison de deux tableaux doit partir d'une bonne connaissance de chacun d'entre eux : l'AFS ne peut intervenir ici, à notre avis, qu'après une AFC sur chacun des tableaux (que nous n'avons pas reproduite pour des raisons de place).

4.3. Analyse d'un nuage « vu d'un point »

a. *Données et signification du modèle mathématique*

Nous étudierons ici des séries trimestrielles de valeurs ajoutées par branche (annexe C), mesurées à prix constants (millions de F 1970) sur la période 1970-1978 (*Source* : INSEE, comptes trimestriels).

On se place au point de vue du dernier trimestre connu, 78.3. On prend comme tableau de référence le tableau qui représente ce qu'aurait été l'évolution de la valeur ajoutée des branches si les proportions entre branches avaient été constamment identiques à ce qu'elles sont en 78.3, et si le total de la valeur ajoutée avait évolué comme dans la réalité historique.

En codant par 0 la date 78.3, on a :

$$\varphi_{it} = f_{i0} \cdot \frac{f_t}{f_0}$$

La formule de transition devient :

$$\mathrm{F}(i) = \frac{1}{\sqrt{\lambda}} \sum_t \left(\sqrt{\frac{f_{it}}{f_t}} - \sqrt{\frac{f_{i0}}{f_0}}\right) \frac{f_t}{\sqrt{f_i}}\, \mathrm{G}(t)$$

Le point de vue va évidemment se trouver à l'origine.

Si $f_{it}/f_t > f_{i0}/f_0$, l'importance relative de la branche i est supérieure dans le trimestre t à ce qu'elle est en 78.3; i est alors attiré par le point t sur les graphes d'analyse factorielle. Il est par contre repoussé si $f_{it}/f_t < f_{i0}/f_0$. En outre, si $f_{it_1}/f_{t_1} > f_{it_2}/f_{t_2}$, le point i sera attiré davantage par t_1 que par t_2.

Ces résultats vont nous permettre d'interpréter très simplement la « trajectoire » des trimestres parmi les points représentant les branches : globalement, cette trajectoire va s'éloigner des branches dont l'importance relative décroît et se diriger vers celles dont l'importance relative croît. Cette règle d'interprétation est identique à celle que l'on utiliserait si l'on avait appliqué l'AFC à ce tableau de mesures.

Nous avons fait trois analyses différentes à partir du même point de vue : d'abord dans une optique « long terme », où nous avons attribué à chaque trimestre une masse proportionnelle à la valeur ajoutée historique du trimestre (à prix constants); puis une optique « moyen terme », dans laquelle nous avons attribué à chaque trimestre la masse historique multipliée par un coefficient exponentiel décroissant avec l'ancienneté, la « période » (durée au bout de laquelle le coefficient est de ½) étant de 3 ans; enfin une optique « court terme », avec une période de un an. Les expressions de long, moyen et court terme sont utilisées ici de façon intuitive.

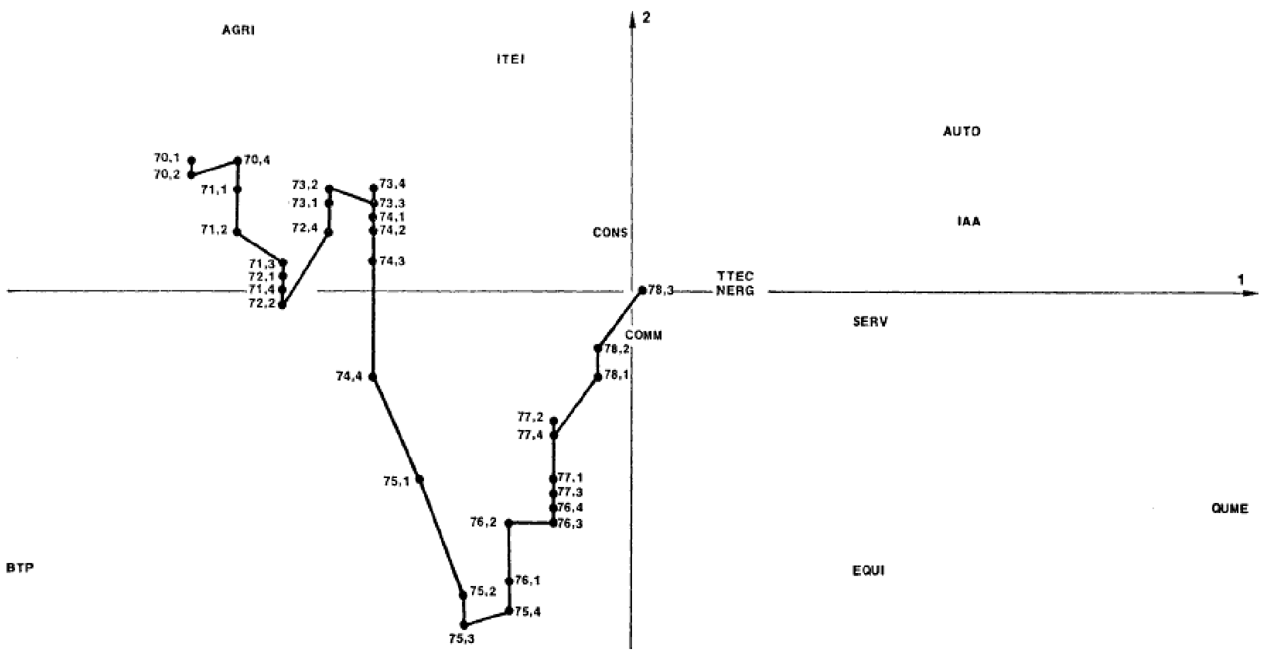
b. *Résultats de l'analyse « long terme »*

Sur l'axe 1 (graphique 13), tous les trimestres ont des coordonnées négatives, et ces coordonnées croissent avec le temps : on trouvera donc à droite les branches dont l'importance relative a crû sur l'ensemble de la période, à gauche celles dont l'importance relative a décrû. L'axe est surtout influencé par les branches BTP et agriculture (décroissantes) et par les services (croissants). La baisse tendancielle de l'importance relative du BTP et de l'agriculture a été interrompue par des périodes de croissance, apparaissant sur l'axe 2. Le point « équipement des ménages », en raison de son faible poids, contribue faiblement à l'inertie des deux premiers axes; on remarque cependant la forte croissance de l'importance relative de ce poste.

c. *Résultats des analyses « moyen terme » et « court terme »*

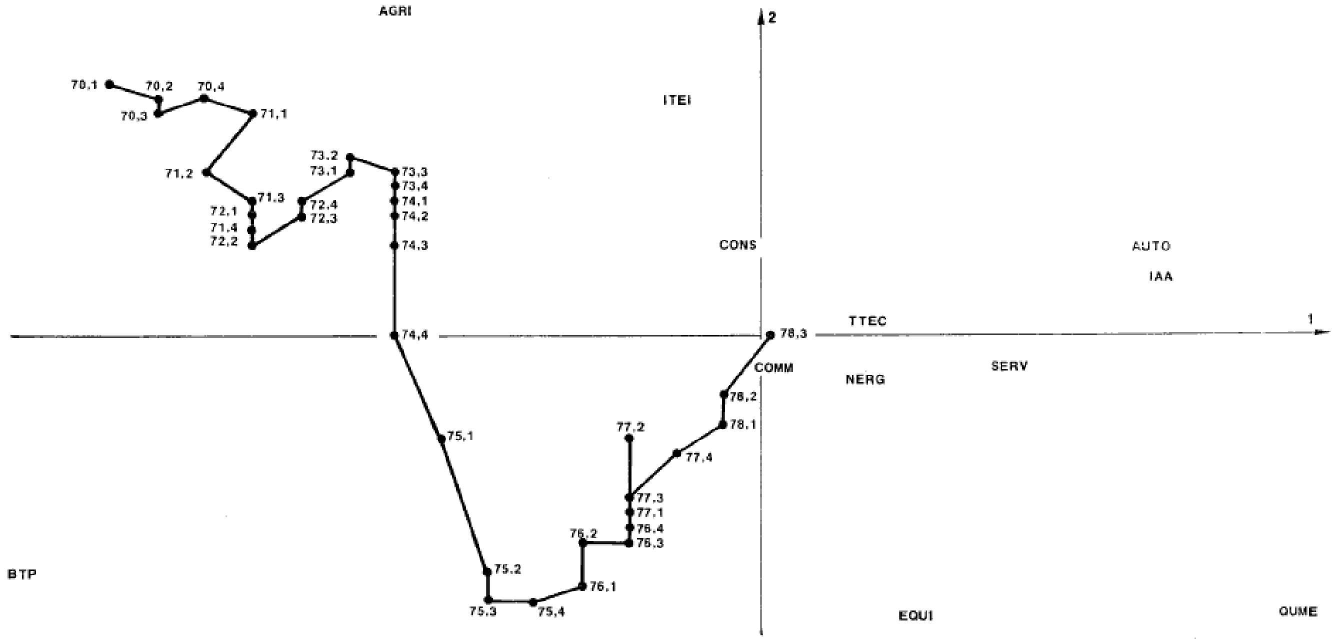
Nous ne nous attarderons pas sur l'analyse « moyen terme » (graphique 14), qui donne ici des résultats intermédiaires entre les analyses sur long et court terme. L'analyse « court terme » fait apparaître des résultats intéressants (graphiques 15 et 16) : le tableau 6 donne les valeurs propres; le tableau des aides à l'interprétation des résultats de l'analyse « court terme » est présenté en annexe C. Dans l'ensemble, la logique du court terme est reflétée par l'axe 1, caractérisé par la décroissance du BTP, la croissance des IAA et des services. La part relative de l'agriculture apparaît comme stable depuis 75.3. L'importance des biens d'équipement des ménages est croissante, mais à une allure nettement moins rapide que dans l'optique « long terme ». L'examen du plan (2,3) fait apparaître l'importance de la branche énergie en 78.1 et en 78.2.

GRAPHIQUE 13. ***Évolution de la valeur ajoutée Optique « long terme ». Plan (1, 2)***



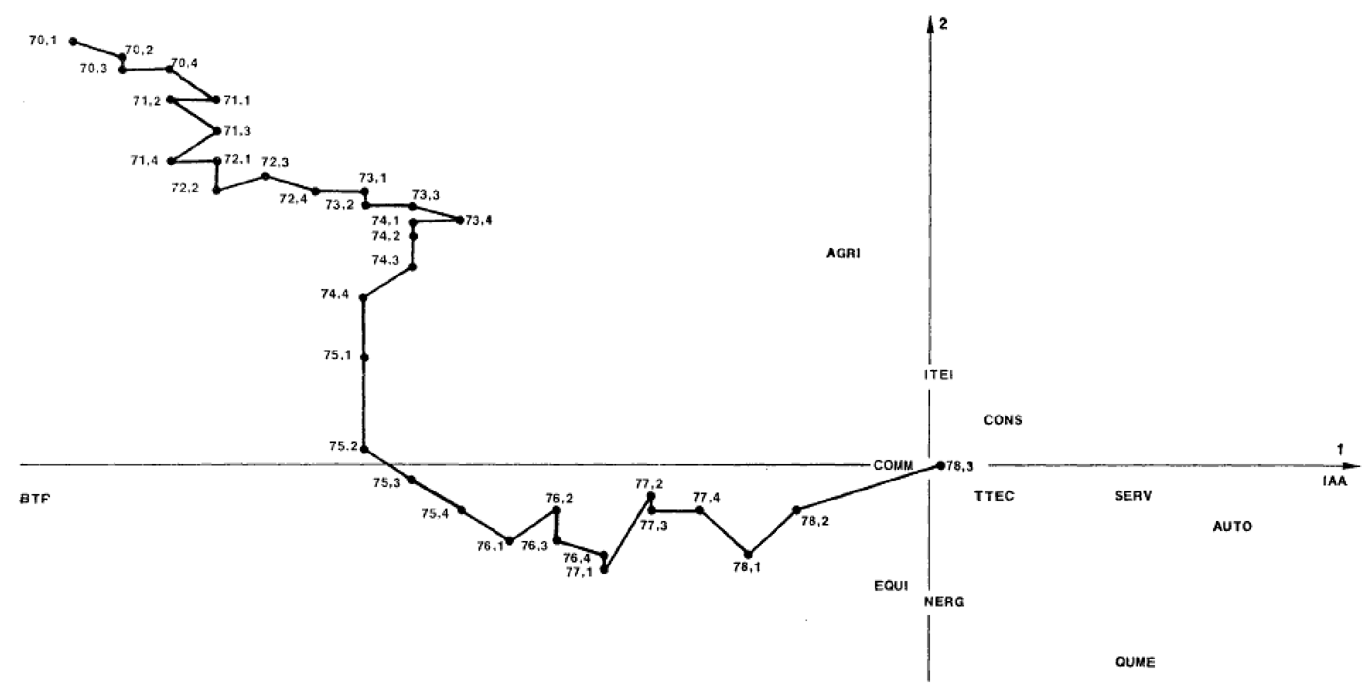
Codification des trimestres et des branches : la même que celle du premier tableau de l'annexe A avec en outre : TTEC = transports et télécommunications et COMM = commerces.

GRAPHIQUE 14. ***Évolution de la valeur ajoutée Optique « moyen terme ». Plan (1, 2)***



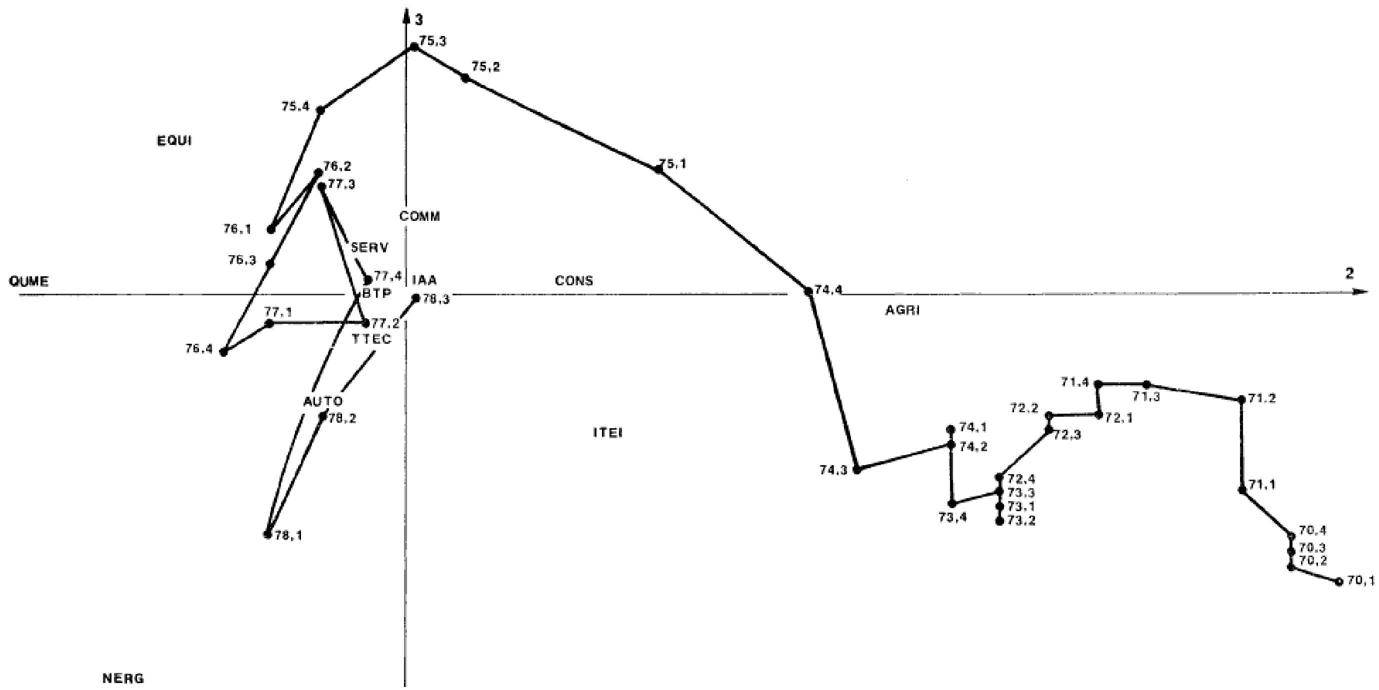
Pour les symboles, voir la note au bas du graphique 13.

GRAPHIQUE 15. *Évolution de la valeur ajoutée Optique « court terme ». Plan (1, 2)*



Pour les symboles, voir la note au bas du graphique 13.

Graphique 16. ***Évolution de la valeur ajoutée Optique « court terme ». Plan (2, 3)***



Pour les symboles, voir la note au bas du graphique 13.

Tableau 6

Valeurs propres de l'analyse « court terme »

Rang des axes	1	2	3	4	5	6	7
Valeurs propres (en 10^{-4})....	4,28	0,67	0,44	0,26	0,08	0,04	0,02
Pourcentages................	73,6	11,6	7,6	4,4	1,4	0,7	0,3
Pourcentages cumulés........	73,6	85,2	92,8	97,2	98,6	99,3	99,6

4.4. Tableaux de transition

a. *Les données*

Nous utiliserons pour cet exemple le résultat d'une expérience de reconnaissance des sons faite sur une population de 416 personnes. Ce tableau nous a été communiqué par l'ISUP.

Six sons de fréquence croissante doivent être associés à des lettres. A la sortie, chaque son est classé deux fois : selon le stimulus (S_i) et selon la réponse (R_j). Si la reconnaissance des sons était parfaite, seule la diagonale du tableau croisant les S_i et les R_j serait remplie. Les réponses comptées hors de la diagonale traduisent des confusions (tableau 7).

Tableau 7

Matrice de confusion

Stimulus	Réponses					
	R_1	R_2	R_3	R_4	R_5	R_6
S_1................	233	96	55	26	6	0
S_2................	94	142	118	40	16	6
S_3................	20	62	122	110	71	31
S_4................	11	22	76	139	130	38
S_5................	3	4	15	54	174	166
S_6................	4	0	3	18	98	293

On remarque que la diagonale principale est chargée; de plus, les diagonales situées au-dessus et en dessous de la diagonale principale sont fortes, surtout la diagonale qui est au-dessus : les confusions se sont donc faites le plus souvent avec les sons les plus proches, et l'erreur la plus fréquente a été de surévaluer la fréquence du son.

Traité par l'analyse des correspondances, ce tableau donne un effet Guttman; la confusion entre deux sons n'apparaît pas bien clairement sur le plan des deux premiers axes factoriels.

b. *Le modèle mathématique*

Nous avons comparé le tableau de fréquence f_{ij}, associé au tableau de confusion, avec le tableau diagonal le plus proche de f_{ij}, au sens de la métrique d'Hellinger :

$$\varphi_{ij} = 0 \quad \text{si} \quad i \neq j \; ; \; \varphi_{ii} = \frac{f_{ii}}{\sum_k f_{kk}}$$

Les points X^i sont munis de la masse unité, et leur coordonnée courante est :

$$x^i_j = \sqrt{f_{ij}} \quad \text{si} \quad i \neq j, \quad x^i_i = \sqrt{f_{ii}} - \sqrt{\frac{f_{ij}}{\Sigma f_{kk}}}$$

on remarque que x^i_i est toujours négative, car $\Sigma f_{kk} < 1$.

L'inertie des nuages X et Y est égale à $2\left(1 - \sqrt{\Sigma f_{kk}}\right)$: elle croît donc avec la fréquence des confusions. La proximité d'un stimulus et d'une réponse s'interprète à l'aide de la formule de transition :

$$F(S_i) = \frac{1}{\sqrt{\lambda}} \sum_j \left(\sqrt{f_{ij}} - \sqrt{\frac{\delta_{ij} f_{ij}}{\sum f_{kk}}}\right) G(R_j)$$

S_i est attiré par le R_j pour lequel f_{ij} est le plus fort, c'est-à-dire que S_i et R_j seront d'autant plus proches que la confusion entre eux est plus forte. S_i est par contre toujours repoussé par R_i, et d'autant plus que f_{ii} est plus fort.

c. *Résultats de l'analyse*

Le tableau 8 donne les valeurs propres; le tableau des aides à l'interprétation des résultats de l'analyse figure en annexe D.

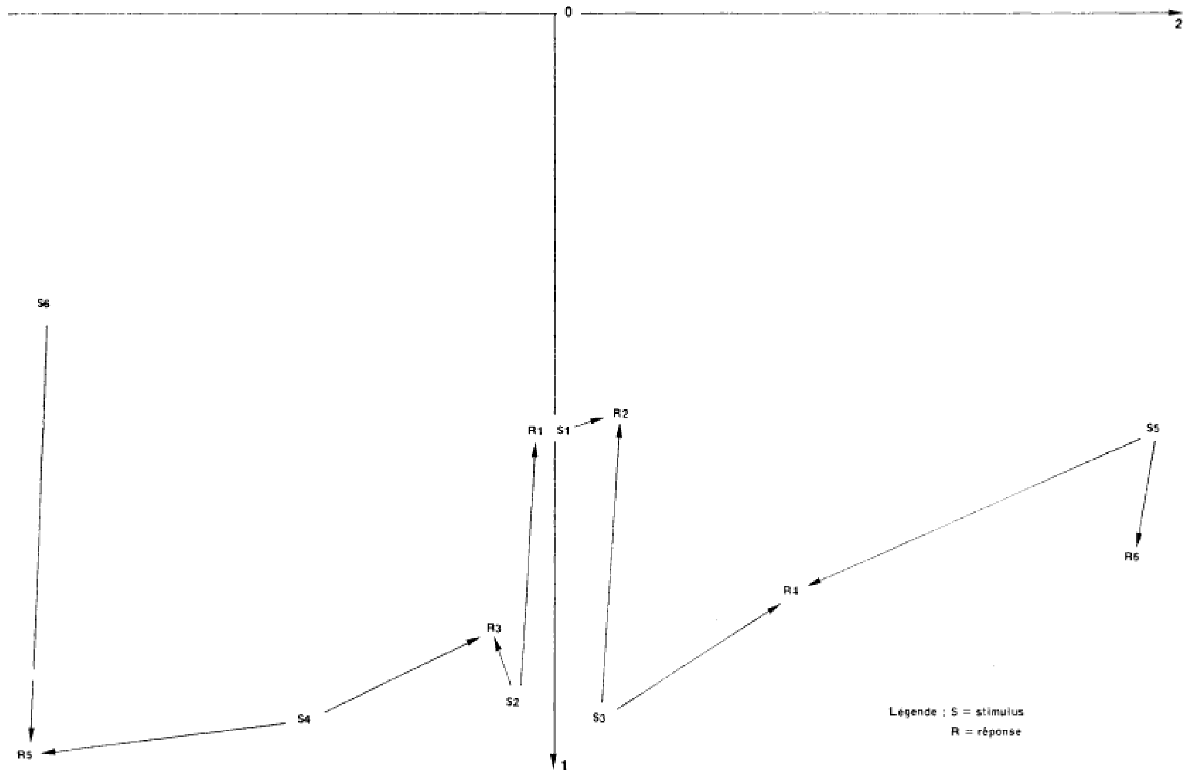
TABLEAU 8

Valeurs propres

Rang des axes	1	2	3	4	5
Valeurs propres................	0,23	0,16	0,12	0,10	0,06
Pourcentages..................	34,6	23,6	18,5	14,3	9,1
Pourcentages cumulés...........	34,6	58,1	76,7	90,9	100,0

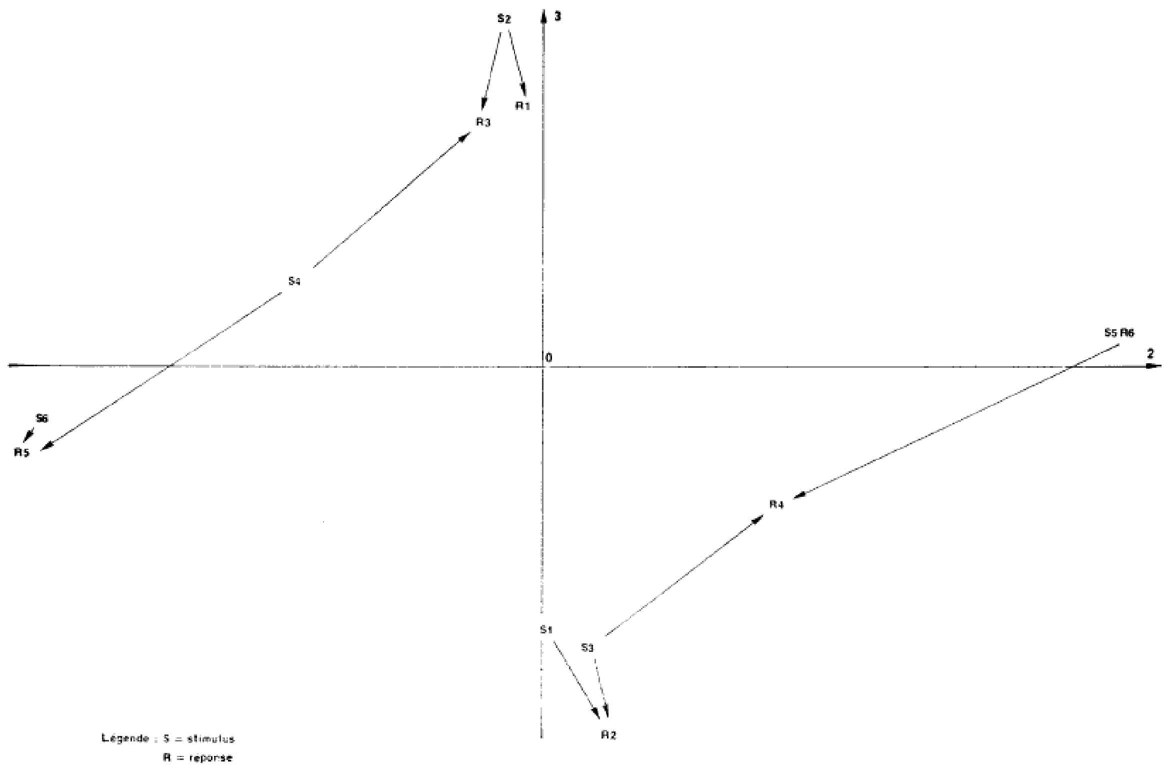
GRAPHIQUE 18

Confusion des sons Plan (1, 2)



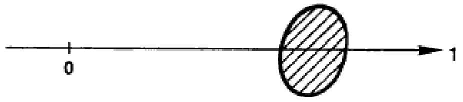
GRAPHIQUE 19

Confusion des sons Plan (2, 3)



Tous les points ont des coordonnées positives sur le premier axe. Cette situation se rencontre chaque fois que le nuage est tout entier « du même côté » de l'origine. C'est ce qui se passe en analyse des correspondances, si

GRAPHIQUE 17



l'on fait l'analyse factorielle à partir de l'origine : le premier axe est alors l'axe trivial qui joint l'origine au centre de gravité du nuage (graphique 17). Dans le cas que nous examinons ici, l'axe 1 ne présente qu'un intérêt limité (graphique 18). La configuration la plus intéressante est celle que l'on obtient dans le plan (2, 3) [graphique 19]; en traçant sur cette figure les arcs qui représentent les confusions les plus fréquentes, on voit clairement se dégager deux familles de points. La lecture de l'axe 1 permet de préciser l'image donnée par le plan (2, 3), et notamment de corriger un « effet de perspective » sur le couple (S_6, R_5).

Chaque axe signale des couples particuliers : l'axe 1 signale (S_4, R_5). (S_2, R_3) et (S_3, R_4); l'axe 2 signale (S_5, R_6), l'axe 3 signale (S_1, R_2). Certaines proximités soulignent l'importance des confusions (S_4, R_3), (S_2, R_1) et (S_6, R_5). Au total, on obtient une visualisation très « parlante » des confusions.

4.5. Tableaux d'échange

Nous terminons par l'étude d'un tableau d'échanges interindustriel selon la méthode décrite en 3.4.2. : la diagonale de ce tableau est donc posée égale à zéro, puis le tableau de fréquences qui lui est associé est étudié par l'analyse factorielle sphérique en le comparant au tableau nul, et en posant $m_I = f_I$ et $m_J = f_J$.

a. *Les données*

Elles sont tirées d'un TES établi sur l'année 1976 (*Source* : Comptes nationaux de l'année 1976, TES provisoire 1976 aux prix courants). Ce tableau croise 34 branches et 34 produits. L'ensemble des produits sera noté I, l'ensemble des branches J. On différencie les deux ensembles en codant un produit par un numéro, et en codant la branche par le numéro du produit correspondant précédé du signe +. On trouvera la nomenclature des produits en annexe E, ainsi que le tableau analysé.

La quantité figurant dans la case (i, j) du tableau représente la valeur de la consommation en produit i réalisée par la branche j durant l'année 1976, mesurée en millions de francs 1976.

b. ***Le modèle mathématique***

Nous avons vu que la formule de transition pouvait s'écrire :

$$F(i) = \frac{1}{\sqrt{\lambda}} \sum_j G(j) \sqrt{f_j f_j^i}$$

ou encore :

$$F(i)\sqrt{f_i} = \frac{1}{\sqrt{\lambda}} \sum_j [G(j)\sqrt{f_j}]\sqrt{f_{ij}}$$

On voit donc que l'abscisse du point représentant le produit i sera d'autant plus influencée par celle du point représentant la branche j que la consommation de i par j sera plus forte.

c. ***Résultats de la première analyse***

On remarque la grandeur de la première valeur propre (53,6 %). L'axe 1 apparaît ici comme un axe trivial, ce qui n'est pas surprenant car toutes les coordonnées sont non-négatives (cf. 2.7.). On va donc procéder surtout à l'analyse selon les autres axes, l'axe 1 n'étant utilisé que pour corriger d'éventuels effets de perspective.

On trouve sur l'axe 2 (graphique 20) un résultat qui apparaît de façon très classique dans tous les travaux portant sur les tableaux d'échange : les produits agricoles sont fortement consommés par les industries agricoles et alimentaires et par les hôtels, cafés et restaurants. Ces produits expliquent à eux seuls 82 % de l'inertie du premier axe. Ceci ne peut qu'avoir une influence négative sur la visualisation des autres échanges : nous avons donc recommencé l'analyse en supprimant les branches et les produits 1, 2 et 3.

d. ***Résultats de la deuxième analyse***

Le tableau 9 présente les valeurs propres; le tableau des aides à l'interprétation des résultats de l'analyse figure à l'annexe E.

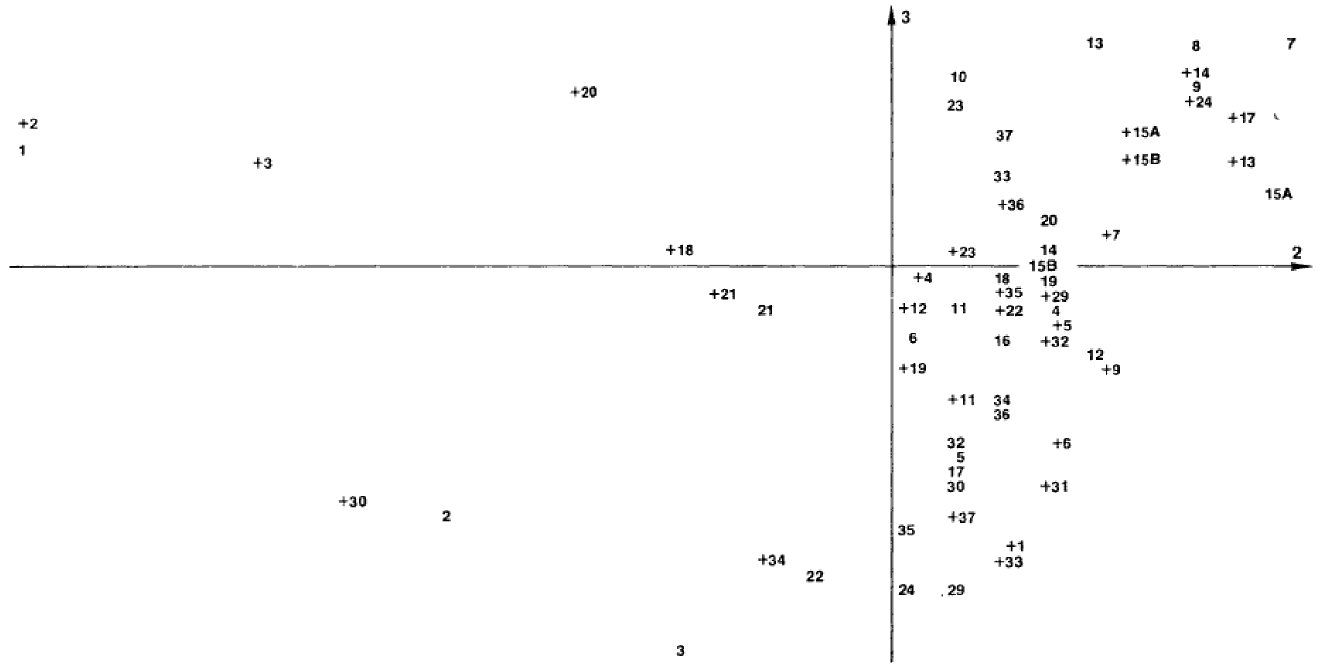
TABLEAU 9

Valeurs propres de la seconde analyse

Rang des axes	1	2	3	4	5	6	7
Valeurs propres	0,64	0,08	0,05	0,05	0,03	0,03	0,02
Pourcentages	63,5	8,0	5,2	4,6	3,3	2,8	2,2
Pourcentages cumulés	63,5	71,6	76,8	81,4	84,7	87,5	89,7

GRAPHIQUE 20

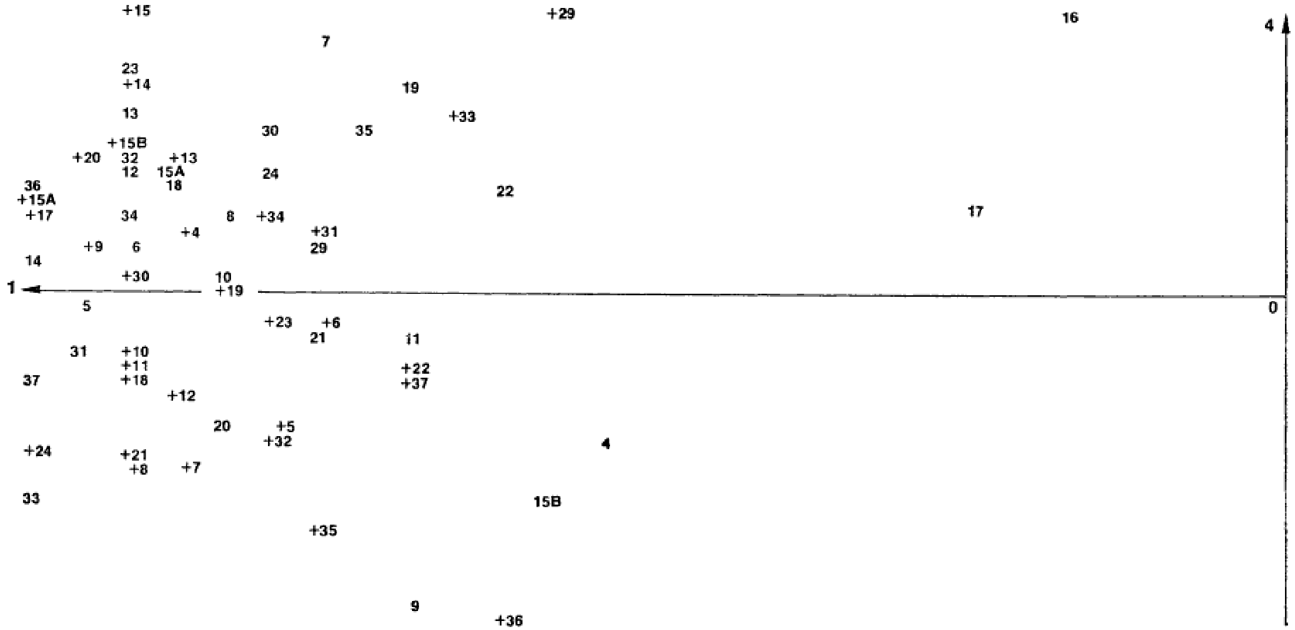
Tableau d'échanges interindustriels provisoire de 1976
Première analyse. Plan (2, 3)



La nomenclature des produits est présentée au début de l'annexe E, page 77. Les branches sont notées par le code du produit correspondant précédé du signe +.

Graphique 21

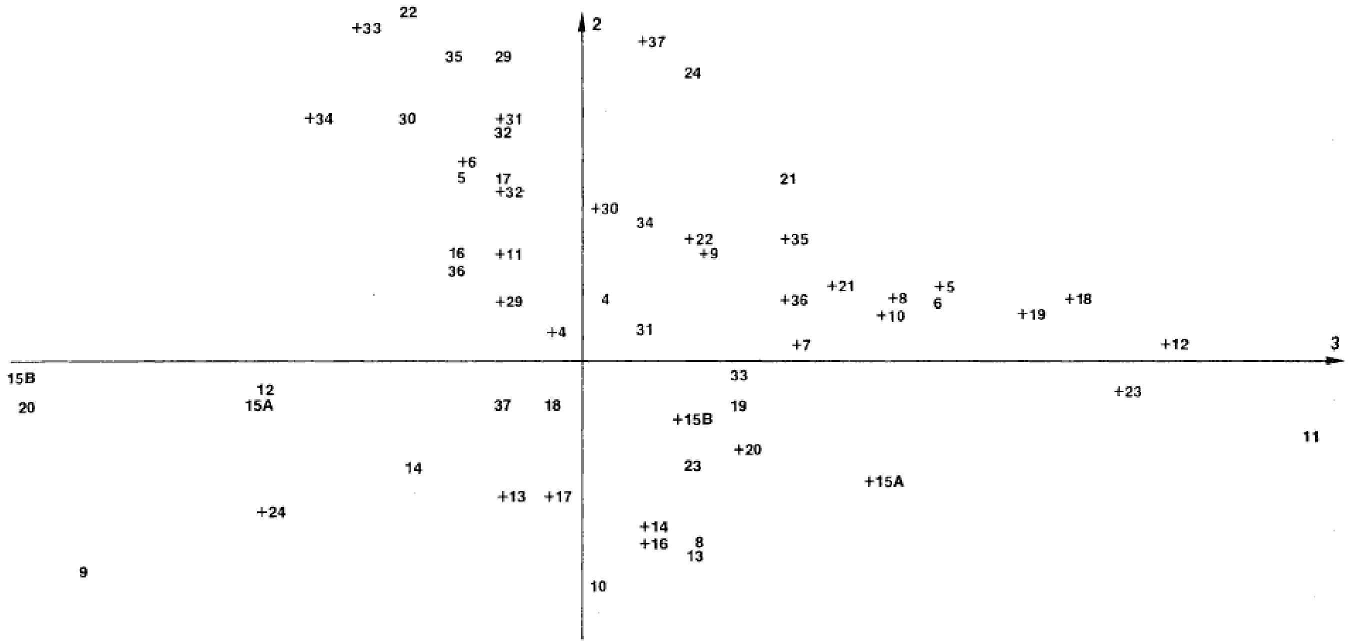
Tableau d'échanges interindustriels provisoire de 1976
Seconde analyse. Plan (1, 4)



Voir note du graphique 20.

GRAPHIQUE 22

Tableau d'échanges interindustriels provisoire de 1976
Seconde analyse. Plan (2, 3)



La nomenclature des produits est présentée au début de l'annexe E, page 77. Les branches sont notées par le code du produit correspondant précédé du signe +.

Les graphiques 21 et 22 donnent respectivement les projections sur les plans (1, 4) et (2, 3).

L'axe 1 est encore un axe parasite.

La branche « Bâtiment et génie civil » (+ 24) joue un rôle important sur les axes 2,3 et 4; elle est associée aux produits « Matériaux de construction » (09), « Équipement des ménages » (15 B), « Bois, ameublement » (20).

L'axe 3 isole nettement le produit « Chimie de base, fibres synthétiques » (11) qui est associé aux branches « Parachimie, pharmacie » (+ 12), « Textiles » (+ 18), « Cuirs et chaussures » (+ 19), « Caoutchouc et matières plastiques » (+ 23).

En bas de l'axe 2, on trouve le produit « Minerais et métaux ferreux » (07) associé aux branches « Fonderie et travail des métaux » (+ 13) et « Construction navale, aéronautique, armement » (+ 17). Dans le même zone, mais un peu à part, on trouve les branches « Construction mécanique » (+ 14) et « Automobile » (+ 16) associées surtout aux produits « Minerais et métaux non ferreux » (07) et « Fonderie et travail des métaux » (13). Au total, le bas de l'axe 2 représente bien les achats des industries mécaniques.

Le haut de l'axe 2 représente un mélange d'associations variées. Les « Services marchands » (+ 33) consomment les produits « Presse et édition » (22), « Hôtels, cafés, restaurants » (30), « Location et crédit bail immobilier » (35); la branche « Transports » (+ 31) consomme des « Produits pétroliers » (05) et des « Réparations automobiles » (29); la branche « Services et organismes financiers » (+ 37) consomme des « Produits du bâtiment et du génie civil » (24).

L'analyse du tableau d'échanges peut être poursuivie et complétée de diverses façons, mais nous ne le ferons pas ici car notre propos est seulement de donner des exemples d'applications possibles et non de procéder à des analyses approfondies. On pourrait ainsi refaire l'analyse après suppression du produit et de la branche « Bâtiment et génie civil », qui jouent un grand rôle sur les premiers axes. On pourrait aussi comparer des tableaux d'échanges successifs (par la méthode de comparaison de deux tableaux), de façon à voir en quoi la structure des échanges a pu varier d'une période à l'autre.

Conclusion

Cette exploration rapide des propriétés de l'analyse factorielle sphérique aura permis de mesurer le champ de ses applications possibles, qui paraît relativement large. D'une façon très générale, elle semble pouvoir être utilisée dans les occasions où l'on a à comparer deux tableaux de contingence, ou bien à comparer un tableau de contingence à une structure de tableaux. Il est aisé de voir qu'un assez grand nombre de problèmes concrets peuvent être considérés comme relevant de l'un ou l'autre de ces deux cas.

Les travaux sur les applications de l'analyse factorielle sphérique se poursuivent, notamment en ce qui concerne l'étude des séries chronologiques dont nous avons déjà vu deux cas (étude des variations de stocks, étude d'un nuage vu d'un point). Les possibilités ouvertes semblent intéressantes; l'analyse factorielle sphérique va sans doute avoir bientôt une place parmi les outils dont dispose l'analyste des données.

9 671335 6 17 3

Variations de stock par produit

Tableau des données*

Produits*** / Trimestres**	AGRI	NERG	ITEI	AUTO	CONS	QUME	BTP	SERV	EQUI	IAA
701	286	391	1 472	302	955	175	358	— 24	1 586	— 519
702	412	128	1 646	416	679	128	391	— 26	1 768	— 53
703	855	349	1 836	205	473	100	347	— 26	1 583	143
704	550	175	1 247	818	343	76	224	— 40	1 448	— 127
711	326	— 31	600	583	640	51	23	— 62	1 248	261
712	112	413	567	350	575	37	— 83	— 43	1 388	— 124
713	32	— 21	1 351	— 262	843	148	— 94	— 52	1 503	— 201
714	— 128	44	836	79	747	137	— 10	— 53	1 285	— 64
721	— 249	150	1 471	456	1 239	119	169	— 36	1 096	576
722	219	137	1 154	24	1 072	137	275	— 30	1 025	— 404
723	470	— 124	958	299	729	35	308	— 31	1 084	— 204
724	573	— 17	1 792	3	1 570	169	268	— 23	1 067	— 116
731	1 081	— 376	2 233	294	645	182	155	— 40	1 188	208

732	1 168	— 74	1 969	309	209	118	78	— 27	1 283	— 206
733	1 005	622	2 193	28	790	223	36	— 40	1 423	— 421
734	923	127	2 080	981	408	231	29	15	1 396	— 396
741	129	325	2 825	463	312	215	57	— 24	1 515	— 603
742	423	181	2 222	712	1 338	261	92	— 15	1 958	573
743	— 61	1 216	2 609	737	533	238	134	— 1	1 564	230
744	305	413	396	330	188	40	182	— 76	1 190	— 314
751	586	— 99	— 924	— 341	— 380	— 80	236	— 23	241	557
752	— 229	— 350	— 1 218	— 719	— 52	— 70	227	— 2	719	586
753	— 257	— 352	— 2 155	— 243	472	— 70	154	1	510	451
754	— 486	— 330	— 908	— 389	409	45	17	— 11	1 258	566
761	— 1 116	— 73	554	79	1 187	199	— 184	— 16	1 100	654
762	— 1 127	— 396	774	52	1 511	443	— 301	— 7	1 179	— 342
763	7	163	1 311	422	790	495	— 333	— 8	687	— 207
764	521	991	— 18	602	1 744	549	— 281	— 32	820	— 93
771	544	652	375	124	498	693	144	— 19	140	630
772	1 114	— 362	641	348	1 551	408	— 32	— 16	563	— 114
773	1 258	— 162	— 334	— 36	— 411	339	56	— 6	557	— 738
774	718	149	— 1 210	165	1 486	207	120	— 23	— 437	— 389
781	694	1 314	714	535	1 188	297	159	— 60	— 460	820
782	827	168	990	— 473	1 594	— 92	189	— 63	— 38	574
783	1 343	162	807	— 230	1 778	314	209	— 104	452	182

* *Unité :* millions de francs 1970. — *Source :* Comptes trimestriels INSEE.

** Les trimestres sont repérés à l'aide d'un code à trois chiffres : les deux premiers chiffres désignent l'année, le troisième chiffre désigne le trimestre (713 : troisième trimestre 1971).

*** Nomenclature des produits :

AGRI : Agriculture;
NERG : Energie;
ITEI : Biens intermédiaires;
AUTO : Automobile;
CONS : Biens de consommation;
QUME : Biens d'équipement des ménages;
BTP : Bâtiment et travaux publics;
SERV : Services;
EQUI : Biens d'équipement professionnel;
IAA : Industries agricoles et alimentaires.

Tableau des aides à l'interprétation des résultats de l'analyse

Produits *

I 1	QLT	POID	INR	1 ≠ F	COR	CTR	2 ≠ F	COR	CTR	3 ≠ F	COR	CTR	4 ≠ F	COR	CTR
AGRI	987	1 000	110	— 204	382	67	166	251	277	178	289	420	76	53	82
NERG	991	1 000	60	— 126	267	26	99	162	98	17	5	4	— 122	249	213
ITEI	1 000	1 000	243	— 451	843	328	63	17	40	— 103	45	143	— 66	18	63
AUTO	963	1 000	68	— 192	551	60	69	71	48	— 61	56	50	— 41	27	25
CONS	990	1 000	165	— 356	767	204	— 89	49	82	66	26	58	— 63	25	59
QUME	877	1 000	39	— 169	739	46	— 4	1	0	— 3	0	0	4	0	0
BTP	692	1 000	33	— 70	153	8	12	4	1	94	273	117	30	28	13
SERV	804	1 000	6	62	658	6	9	15	1	— 24	105	8	— 2	1	0
EQUI	997	1 000	206	— 395	761	252	— 156	119	245	— 13	1	2	131	83	242
IAA	990	1 000	69	43	27	3	— 142	293	206	122	217	197	— 145	309	303
	4 290,0		1 000			1 000			1 000			1 000			1 000

* Voir note ***, page 67.

Trimestres *

J 1	QLT	POID	INR	1 # F	COR	CTR	2 # F	COR	CTR	3 # F	COR	CTR	4 # F	COR	CTR
701	979	1 000	33	— 172	900	48	20	13	4	— 9	3	1	24	18	8
702	985	1 000	31	— 170	948	47	5	1	0	4	1	0	19	12	5
703	1 000	1 000	32	— 167	876	45	5	1	0	37	43	18	— 6	2	1
704	983	1 000	28	— 156	895	40	26	24	7	— 1	0	0	22	18	7
711	983	1 000	21	— 127	779	26	— 32	53	11	24	28	8	2	0	0
712	940	1 000	20	— 128	830	27	0	0	0	— 19	21	6	6	2	0
713	938	1 000	25	— 124	634	25	— 38	60	15	— 25	27	9	38	58	20
714	995	1 000	19	— 117	752	22	— 43	102	19	— 44	109	27	1	0	0
721	968	1 000	30	— 140	652	32	— 68	155	47	— 14	8	3	— 62	130	56
722	980	1 000	24	— 146	879	35	8	3	1	0	0	0	25	25	9
723	940	1 000	23	— 136	809	30	2	0	0	6	2	1	48	99	33
724	989	1 000	31	— 162	868	43	— 4	1	0	18	11	4	29	28	12
731	999	1 000	35	— 161	747	42	— 14	6	2	32	30	14	20	11	5
732	1 000	1 000	30	— 153	801	38	30	30	9	0	0	0	54	98	41
733	986	1 000	37	— 180	882	52	41	45	17	3	0	0	21	12	6
734	989	1 000	36	— 176	874	50	48	63	23	— 19	11	5	29	23	12
741	998	1 000	35	— 173	861	49	32	29	10	— 51	77	36	18	9	5
742	998	1 000	43	— 193	887	60	— 29	22	9	26	15	9	— 35	30	18
743	992	1 000	40	— 171	734	47	— 22	13	5	— 32	27	14	— 65	110	62
744	979	1 000	19	— 119	768	23	25	33	6	0	0	0	30	48	13
751	974	1 000	19	46	112	3	— 29	47	9	97	498	125	54	154	41
752	998	1 000	23	58	149	5	— 112	559	128	50	109	33	48	101	33
753	943	1 000	26	43	72	3	— 124	613	157	64	160	54	32	40	14
754	996	1 000	24	7	2	0	— 144	867	210	37	56	18	30	37	13
761	997	1 000	28	— 89	288	13	— 121	531	150	— 45	74	28	— 47	80	32
762	995	1 000	34	— 104	327	18	— 87	230	77	— 97	294	126	17	9	4
763	971	1 000	24	— 136	774	30	4	1	0	— 53	121	33	— 11	6	2
764	967	1 000	31	— 121	484	24	8	2	1	27	24	10	— 6	2	1
771	896	1 000	32	— 144	659	34	— 31	32	10	50	79	33	— 44	63	28
772	999	1 000	28	— 137	678	81	5	1	0	19	13	5	39	55	22
773	916	1 000	21	— 9	5	0	38	66	14	35	56	16	127	754	227
774	910	1 000	27	— 10	4	0	52	100	27	79	234	83	3	0	0
781	980	1 000	34	— 98	287	16	55	88	30	72	152	69	— 122	440	212
782	978	1 000	27	— 82	254	11	5	1	0	90	296	107	— 61	141	54
783	975	1 000	31	— 137	623	30	— 6	2	1	88	256	103	— 11	5	2
	4 290,0		1 000			1 000			1 000			1 000			1 000

* Nomenclature des produits et codification des trimestres présentés à la fin du premier tableau de l'annexe A, page 67.

Élèves scolarisés en 1972-1973, sortis du système scolaire en 1973, et ayant trouvé un emploi, par emploi occupé et par diplôme atteint

Tableau des aides à l'interprétation des résultats de l'analyse

Emplois occupés *

1 1*	QLT	POID	INR	1 ≠ F	COR	CTR	2 ≠ F	COR	CTR	3 ≠ F	COR	CTR	4 ≠ F	COR	CTR
AGR.	997	1 000	78	− 151	862	95	16	10	6	− 46	82	74	20	14	21
ING.	1 000	1 000	30	− 47	226	10	− 26	69	16	56	428	148	43	180	101
TEC.	999	1 000	55	− 101	543	43	36	70	31	63	209	134	44	100	103
OQ.	1 000	1 000	173	− 193	636	156	− 101	175	238	56	52	104	− 89	135	436
ONQ.	999	1 000	149	− 211	880	186	− 48	47	55	− 47	44	76	22	10	27
CSP.	1 000	1 000	27	− 69	529	20	− 36	150	32	0	0	0	− 4	3	1
CMO.	999	1 000	108	82	180	27	− 157	674	572	− 36	37	46	61	101	202
EMQ.	999	1 000	289	309	960	393	− 35	13	30	− 25	7	22	− 37	14	78
EMNQ.	1 000	1 000	91	130	538	70	− 29	28	20	109	376	396	24	19	32
	271 096,0		1 000			1 000			1 000			1 000			1 000

* Pour les abréviations, voir tableaux 2 ou 3, pages 38 et 39.

Diplômes atteints **

J 1**	QLT	POID	INR	1 # F	COR	CTR	2 # F	COR	CTR	3 # F	COR	CTR	4 # F	COR	CTR
SSD	1 001	1 000	214	— 234	756	229	— 19	6	10	— 124	213	525	38	20	79
BEPC	1 000	1 000	75	— 139	753	81	— 19	16	10	26	27	23	60	137	193
BECA	1 000	1 000	410	— 362	933	543	— 33	8	26	47	16	74	— 77	43	327
BACG	1 000	1 000	85	— 96	326	39	127	550	368	25	22	21	47	77	122
BACT	999	1 000	60	— 124	757	64	— 10	6	3	— 2	0	0	36	64	72
DEUG	996	1 000	18	— 20	72	2	7	9	1	58	540	114	25	102	35
DUT	999	1 000	90	— 68	152	19	157	796	566	— 10	4	4	— 32	34	57
SUP	998	1 000	47	— 73	339	23	— 26	47	17	84	435	237	46	132	116
	271 096,0		1 000			1 000			1 000			1 000			1 000

** Pour les abréviations, voir la note du graphique 11, page 41.

Séries trimestrielles des valeurs ajoutées par branche

Tableau des données*

Trimestres \ Branches	AGRI	IAA	NERG	ITEI	EQUI	AUTO	CONS	QUME	BTP	TTEC	COMM	SERV
701	12 378	7 523	8 717	17 135	10 914	3 534	11 198	722	14 339	10 346	19 426	35 376
702	12 575	7 903	8 859	17 467	11 441	3 847	11 211	733	14 586	10 510	19 738	35 885
703	12 728	8 128	8 943	17 364	11 405	3 835	11 273	734	14 626	10 639	19 898	36 388
704	12 862	8 211	9 050	17 359	11 448	4 240	11 582	747	14 764	10 776	20 564	37 014
711	12 848	8 403	9 199	17 469	11 944	4 150	11 758	761	14 560	10 730	20 911	37 655
712	12 924	8 514	9 128	17 493	12 339	3 807	12 053	763	14 806	10 840	21 368	38 350
713	12 905	8 597	9 238	17 910	12 857	4 058	12 317	829	15 081	11 123	21 666	39 127
714	12 810	8 678	9 414	18 113	12 989	4 147	12 544	847	15 382	11 472	22 017	39 831
721	12 789	8 978	9 532	18 429	12 792	4 457	12 897	851	15 661	11 755	22 364	40 615
722	12 777	8 779	9 655	18 550	13 169	4 537	12 961	870	15 831	11 885	22 297	41 294
723	12 983	9 008	9 787	18 759	13 099	4 701	12 912	884	15 794	12 117	22 812	41 886
724	13 120	9 145	10 053	19 443	13 478	4 926	13 133	980	15 805	12 294	22 770	42 654
731	13 391	9 350	10 408	20 318	13 887	4 982	13 082	1 032	15 620	12 526	23 345	43 349
732	13 636	9 368	10 559	20 539	14 169	4 952	12 899	1 014	15 455	12 837	23 594	43 871
733	13 735	9 347	10 664	20 466	14 448	4 927	13 002	1 060	15 281	12 713	23 477	44 516

734	13 785	9 489	10 869	21 119	14 764	5 310	13 168	1 135	15 401	12 993	24 129	45 554
741	13 750	9 445	10 613	21 627	15 111	4 932	13 512	1 203	15 725	13 181	24 594	45 881
742	13 741	9 751	10 796	21 582	15 110	5 169	13 699	1 245	16 013	13 162	24 606	46 549
743	13 525	9 874	11 150	21 532	15 157	5 137	13 679	1 214	16 174	13 141	24 538	47 103
744	13 404	9 918	10 609	19 787	15 217	4 704	13 322	1 120	16 172	12 775	24 014	47 015
751	13 152	10 242	10 451	18 590	15 583	4 593	13 172	1 062	16 380	12 833	23 937	47 422
752	12 779	10 362	10 644	18 089	16 530	4 504	13 158	1 097	16 470	12 808	24 238	47 993
753	12 558	10 481	10 665	17 718	16 691	4 717	13 549	1 139	16 445	13 016	24 735	48 321
754	12 356	10 625	11 112	18 548	17 034	4 971	13 575	1 246	16 334	13 344	25 629	48 590
761	12 080	10 641	11 651	19 446	17 296	5 221	13 723	1 299	16 099	13 546	25 698	49 080
762	12 013	10 577	11 094	20 086	17 178	5 560	13 893	1 393	15 974	13 803	26 082	49 689
763	12 054	10 546	11 717	20 457	17 444	5 556	13 664	1 449	15 836	13 908	26 283	49 989
764	12 102	10 677	12 376	20 639	17 375	5 650	14 235	1 496	15 883	14 150	26 586	50 566
771	12 591	11 211	12 394	21 087	18 129	5 881	14 214	1 503	15 975	14 320	26 360	51 126
772	12 706	11 124	11 950	21 040	17 432	5 956	14 129	1 377	15 792	14 290	25 844	51 486
773	12 822	11 139	11 573	20 717	18 003	5 910	13 674	1 400	15849	14 371	26 233	52 217
774	12 856	11 486	12 019	20 303	17 225	5 941	14 252	1 352	15 734	14 495	26 052	52 611
781	12 938	11 966	13 823	21 056	17 310	5 998	14 363	1 407	15 468	14 793	26 189	53 141
782	13 008	11 949	12 987	21 680	17 631	6 051	14 550	1 390	15 182	14 863	26 662	54 041
783	13 145	12 123	12 234	21 258	17 476	6 061	14 599	1 422	14 173	14 666	26 670	53 923

* Codification des trimestres et des branches : la même que celle du premier tableau de l'annexe A, page 67 avec, en outre : TTEC = transports et télécommunications et COMM = commerces.

Analyse « court terme »

Tableau des aides à l'interprétation des résultats de l'analyse

Branches

I 1	QLT	POID	INR	1 ≠ F	COR	CTR	2 ≠ F	COR	CTR	3 ≠ F	COR	CTR	4 ≠ F	COR	CTR
AGRI	995	63	61	— 5	69	6	22	857	449	0	0	0	5	49	68
IAA	995	56	87	29	915	108	0	0	0	1	2	2	8	74	146
NERG	999	60	65	0	0	0	— 12	260	147	— 20	678	586	6	48	71
ITEI	995	102	41	1	2	0	9	353	124	— 6	232	124	— 9	391	361
EQUI	996	85	38	— 3	74	4	— 11	521	173	9	301	152	— 2	38	33
AUTO	999	28	36	22	654	32	— 4	40	12	— 4	35	17	— 8	121	100
CONS	935	70	10	3	84	1	6	376	33	1	7	1	1	12	3
QUME	930	7	12	15	202	3	— 19	362	38	2	2	0	— 18	333	93
BTP	999	76	555	— 64	995	750	— 2	2	9	1	0	1	2	1	18
TTEC	947	70	4	5	551	3	— 1	132	5	— 1	95	6	0	11	1
COMM	942	128	12	— 1	71	1	0	0	0	5	373	60	— 2	157	43
SERV	998	254	77	12	864	91	— 1	15	10	3	50	51	3	36	63
			1 000			1 000			1 000			1 000			1 000

Trimestres

J 1	QLT	POID	INR	1 # F	COR	CTR	2 # F	COR	CTR	3 # F	COR	CTR	4 # F	COR	CTR
701	999	0	3	— 58	619	3	43	332	9	— 14	40	2	5	5	0
702	998	0	4	— 57	634	3	41	318	10	— 13	39	2	4	3	0
703	995	0	4	— 55	616	4	41	329	12	— 13	37	2	7	10	1
704	992	1	5	— 53	603	4	40	331	14	— 12	35	2	6	7	1
711	994	1	5	— 49	612	4	37	332	14	— 10	27	2	6	9	1
712	995	1	6	— 51	639	5	37	317	17	— 4	6	1	9	17	2
713	999	1	7	— 50	686	6	33	288	17	— 3	5	0	5	6	1
714	997	1	8	— 51	718	8	31	256	17	— 4	6	1	4	5	1
721	995	1	9	— 49	685	9	31	266	21	— 5	12	1	4	4	1
722	996	2	11	— 49	736	11	28	230	21	— 5	10	1	3	2	0
723	995	2	12	— 46	698	11	29	257	26	— 6	14	2	3	2	1
724	998	3	13	— 43	687	12	27	262	29	— 8	29	5	0	0	0
731	998	3	14	— 40	644	12	27	291	35	— 11	51	9	— 3	7	2
732	990	4	16	— 38	620	13	27	307	42	— 11	54	11	— 3	8	3
733	997	5	17	— 36	621	14	26	313	45	— 9	50	11	— 2	6	2
734	992	5	18	— 33	594	15	24	299	47	— 10	62	15	— 7	32	13
741	996	7	25	— 36	627	21	25	290	63	— 6	21	7	— 9	49	28
742	996	8	27	— 35	659	24	23	272	64	— 7	29	10	— 7	36	22
743	995	9	30	— 35	710	29	20	214	55	— 9	49	20	— 5	18	12
744	996	11	34	— 37	809	37	17	167	49	0	0	0	5	11	9
751	997	13	41	— 37	815	45	11	69	24	7	27	15	12	82	76
752	999	16	49	— 37	803	54	2	3	1	12	75	49	14	104	116
753	999	19	55	— 35	769	57	0	1	0	14	116	84	14	109	135
754	996	23	51	— 32	852	59	— 3	15	7	10	79	53	7	33	38
761	997	27	45	— 29	902	55	— 6	58	23	4	17	10	1	1	1
762	998	33	46	— 25	·849	53	— 4	25	10	7	57	35	— 5	47	50
763	997	39	51	— 24	834	58	— 6	73	32	2	6	4	— 6	71	82
764	996	47	58	— 23	780	61	— 8	106	53	— 2	10	7	— 5	48	63
771	995	57	57	— 21	804	62	— 7	103	50	— 1	4	3	— 5	53	68
772	991	67	49	— 19	906	60	— 2	16	7	0	3	2	— 3	35	39
773	994	80	68	— 19	775	72	— 3	40	24	6	72	65	— 3	28	43
774	999	96	52	— 15	838	59	— 2	36	16	1	6	4	3	33	39
781	999	116	81	— 11	336	37	— 7	162	113	— 12	419	449	6	80	149
782	985	139	31	— 7	542	23	— 3	119	32	— 5	285	116	1	2	1
783	99	163	0	0	12	0	0	0	0	0	5	0	0	0	0
			1 000			1 000			1 000			1 000			1 000

Confusion des sons

Tableau des aides à l'interprétation des résultats de l'analyse

Stimulus

I 1	QLT	POID	INR	1 ≠ F	COR	CTR	2 ≠ F	COR	CTR	3 ≠ F	COR	CTR	4 ≠ F	COR	CTR
S 1	1 000	1 000	144	145	219	91	5	0	0	— 172	310	241	151	235	238
S 2	999	1 000	185	237	454	243	— 12	1	1	236	446	447	— 72	42	55
S 3	1 001	1 000	194	247	468	263	35	9	8	— 183	261	273	— 164	209	285
S 4	1 000	1 000	187	247	486	263	— 108	95	75	54	24	24	179	256	335
S 5	999	1 000	171	147	188	93	303	798	580	31	8	8	9	1	1
S 6	1 000	1 000	118	104	137	47	— 229	670	336	— 30	12	8	— 90	104	86
	1 103,0		1 000			1 000			1 000			1 000			1 000

Réponses

J 1	QLT	POID	INR	1 ≠ F	COR	CTR	2 ≠ F	COR	CTR	3 ≠ F	COR	CTR	4 ≠ F	COR	CTR
R 1	999	1 000	114	148	287	95	— 15	3	2	168	369	228	— 140	259	208
R 2	1 000	1 000	132	141	226	86	25	7	4	— 240	659	468	95	103	95
R 3	1 000	1 000	178	215	389	200	— 23	5	4	162	220	211	173	252	314
R 4	1 000	1 000	169	203	364	178	111	109	78	— 87	69	63	— 181	290	345
R 5	1 000	1 000	218	260	462	292	— 265	483	448	— 52	19	22	— 14	1	2
R 6	1 000	1 000	189	186	274	149	271	580	465	31	8	8	58	27	36
	1 103,0		1 000			1 000			1 000			1 000			1 000

Tableau d'échange interindustriels provisoire de 1976

Code	Nomenclature des produits
01	Agriculture, sylviculture, pêche.
02	Viande et produits laitiers.
03	Autres produits agricoles et alimentaires.
04	Combustibles minéraux solides, coke.
05	Produits pétroliers, gaz naturel.
06	Électricité, gaz et eau.
07	Minerais et métaux ferreux.
08	Minerais et métaux non ferreux.
09	Matériaux de construction.
10	Verre.
11	Chimie de base, fibres synthétiques.
12	Parachimie, pharmacie.
13	Fonderie et travail des métaux.
14	Construction mécanique.
15 A	Matériel électrique professionnel.
15 B	Bien d'équipement ménager.
16	Automobile transport terrestre.
17	Construction navale, aéronautique, armement.
18	Textiles, habillement.
19	Cuirs et chaussures.
20	Bois, meubles, industries diverses.
21	Papier, carton.
22	Presse et édition.
23	Caoutchouc, matières plastiques.
24	Bâtiment et génie civil.
29	Réparations et commerce automobiles.
30	Hôtels, cafés, restaurants.
31	Transports.
32	Télécommunications et postes.
33	Services marchands rendus aux entreprises.
34	Services marchands rendus aux particuliers.
35	Location et crédit-bail immobilier.
36	Assurances.
37	Services et organismes financiers.

*Tableau des données**

Produits \ Branches	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11
1	0	64 244	28 186	57	0	0	26	14	0	0	139
2	442	0	3 346	0	0	0	0	0	0	0	0
3	17 045	132	0	5	24	0	0	0	0	0	979
4	54	12	38	0	0	3 205	5 439	89	202	2	157
5	3 030	554	1 722	125	0	8 549	1 404	406	2 340	497	3 245
6	1 095	217	540	358	490	0	2 208	1 596	1 181	370	1 558
7	1 910	0	0	53	0	0	0	76	633	10	0
8	0	0	180	11	0	484	1 994	0	123	30	819
9	1 374	33	70	0	0	38	574	51	0	172	1 312
10	240	25	722	0	0	0	0	0	0	0	37
11	8 850	38	463	0	645	62	711	241	328	575	0
12	4 199	10	112	96	115	69	66	0	156	0	269
13	251	81	1 299	72	106	398	884	44	104	186	169
14	3 597	84	353	127	230	308	307	140	411	47	153
15 A	30	0	23	20	47	919	401	64	75	56	205
15 B	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0
17	101	0	0	0	0	0	0	0	0	0	0
18	405	0	51	6	0	21	15	0	56	0	62
19	69	0	0	1	0	1	2	0	19	0	0
20	368	168	567	82	0	314	72	10	31	18	0
21	8	398	2 667	2	2	149	15	46	241	495	699
22	31	99	343	5	69	250	6	4	7	36	6
23	382	559	1 578	0	92	8	23	62	417	146	360
24	1 491	83	196	51	288	2 686	146	59	199	42	220
29	1 627	110	166	11	122	94	92	36	918	24	88
30	80	51	83	9	68	43	45	16	37	11	46
31	1 051	224	1 637	12	4 638	788	2 086	454	2 980	245	1 870
32	18	98	388	15	157	273	240	64	47	1	316
33	694	905	5 112	166	2 503	1 096	4 377	2 410	720	271	804
34	824	124	166	19	99	915	203	83	240	43	279
35	0	36	98	0	97	0	27	18	23	64	164
36	624	47	102	1	36	14	32	13	111	20	112
37	0	89	153	0	94	0	76	30	59	14	96

Produits \ Branches	+12	+13	+14	+15 A	+15 B	+16	+17	+18	+19	+20	+21
1.	265	0	0	0	0	0	0	2 574	1	6 287	724
2.	0	0	0	0	0	0	1	0	1 627	0	0
3.	1 132	0	0	0	0	0	1	52	0	0	48
4.	17	55	49	9	7	28	24	43	24	17	32
5.	1 440	540	599	691	167	833	307	611	92	289	393
6.	1 395	1 246	854	551	136	865	243	654	66	659	816
7.	0	9 496	6 525	946	377	7 035	1 594	0	0	509	0
8.	357	2 219	1 409	3 001	0	630	539	0	0	998	14
9.	112	496	125	14	0	0	136	0	0	95	71
10.	646	3	195	546	37	925	50	0	0	238	0
11.	9 412	705	447	1 602	355	465	342	3 228	774	453	546
12.	0	1 145	441	360	64	936	461	93	13	1 191	514
13.	1 322	0	8 614	4 734	1 471	9 895	2 534	818	183	1 807	0
14.	98	1 032	0	568	72	1 651	2 082	819	14	783	45
15 A.	0	888	2 461	0	1 849	671	2 194	0	0	29	73
15 B.	0	0	0	0	0	0	4	0	0	0	0
16.	0	0	0	0	0	0	24	0	0	0	0
17.	0	0	0	0	0	0	0	0	0	94	0
18.	217	67	158	243	0	809	490	0	605	1 673	44
19.	0	35	55	142	0	260	25	308	0	25	15
20.	0	686	87	196	131	54	711	31	21	0	0
21.	1 559	194	349	135	136	9	23	936	252	405	0
22.	282	35	27	46	64	36	2	49	38	75	19
23.	1 607	477	1 183	1 508	479	4 228	151	693	387	1 785	70
24.	207	187	245	270	100	408	379	208	32	87	69
29.	85	143	168	138	30	164	75	167	24	81	32
30.	43	73	82	70	8	86	29	89	11	48	19
31.	676	1 316	1 250	1 549	320	1 455	458	902	178	946	625
32.	214	581	381	415	167	356	124	290	57	399	34
33.	6 566	1 174	3 264	4 315	856	2 129	3 617	3 569	506	1 610	1 406
34.	217	329	331	240	64	366	172	331	37	168	54
35.	197	36	190	148	62	128	51	62	27	58	23
36.	82	100	181	177	46	95	135	175	30	171	38
37.	90	112	163	143	21	154	49	146	11	58	41

* Voir nomenclature au début de l'annexe E page 77. Les branches sont notées par le code du produit correspon'ant précédé du signe +.

*Tableau des données** (suite et fin)

Branches / Produits	+22	+23	+24	+29	+30	+31	+32	+33	+34	+35	+36	+37
1	0	292	0	0	8 614	0	2	0	1 325	0	0	0
2	0	0	0	0	8 671	0	0	1	2 072	0	0	0
3	1	0	0	0	13 660	0	0	4	1 555	0	0	0
4	1	23	270	0	26	39	0	27	21	0	0	0
5	224	665	6 409	698	1 173	9 835	543	2 460	3 312	392	23	250
6	55	866	465	470	1 079	1 006	23	414	310	331	60	158
7	12	582	5 357	183	0	346	0	0	0	0	0	0
8	103	59	1 264	0	0	0	0	0	16	0	0	0
9	0	32	24 280	0	0	0	0	0	0	0	0	0
10	0	295	2 005	0	133	0	0	25	52	0	0	0
11	46	6 376	352	0	0	0	0	0	0	0	0	0
12	649	0	3 417	392	74	248	0	231	3 407	0	0	0
13	13	315	6 577	267	950	0	0	370	296	151	0	0
14	93	245	5 101	545	193	232	0	90	787	0	0	0
15 A	1	110	3 882	763	35	538	1 695	1 883	75	1	26	0
15 B	337	0	1 090	0	19	0	0	0	675	0	0	0
16	0	0	0	13 666	0	750	0	85	0	0	0	0
17	0	0	0	0	0	3 328	0	0	0	0	0	0
18	45	779	1 388	328	231	329	56	0	1 677	0	0	1
19	75	0	20	0	0	2	0	0	294	0	0	0
20	4	16	9 209	0	267	129	138	385	1 671	0	0	321
21	4 965	189	763	0	21	555	69	2 363	66	21	113	794
22	0	28	91	22	193	116	228	6 296	3 923	132	397	680
23	142	0	2 650	2 608	4	1 117	73	20	217	0	0	0
24	34	139	0	132	403	2 020	135	414	991	5	0	2 659
29	40	58	469	0	96	2 941	587	1 009	2 434	175	2	3 880
30	20	30	241	71	0	128	52	2 014	910	75	0	5
31	4 959	219	6 307	161	77	0	736	1 908	1 163	15	101	823
32	493	184	926	990	512	738	0	3 956	1 546	96	259	1 886
33	1 089	1 671	22 186	889	1 428	3 690	2 656	0	1 617	4 758	9 706	3 625
34	33	88	688	71	212	163	0	2 143	0	414	0	681
35	107	14	212	45	275	146	107	3 999	2 086	0	0	899
36	40	70	836	291	83	1 067	0	346	292	70	0	23
37	42	40	1 429	108	42	134	0	98	262	5	591	0

* Voir nomenclature au début de l'anenxe E, page 77. Les branches sont notées par le code du produit correspondant précédé du signe +.

Tableau des aides à l'interprétation des résultats de la seconde analyse

Produits

I 1	QLT	POID	INR	1 # F	COR	CTR	2 # F	COR	CTR	3 # F	COR	CTR	4 # F	COR	CTR
4.	776	17	17	— 520	272	7	116	14	3	23	1	0	— 193	38	14
5.	960	83	83	— 862	745	98	317	100	104	— 126	16	26	— 13	0	0
6.	851	35	35	— 853	730	40	101	10	4	289	84	56	50	2	2
7.	888	58	58	— 715	513	47	— 478	229	166	— 109	12	13	338	114	144
8.	825	24	24	— 768	591	23	— 317	101	31	85	7	3	89	8	4
9.	947	47	47	— 628	396	29	— 370	138	81	— 448	202	182	— 421	178	182
10.	779	9	9	— 774	600	8	— 388	151	17	— 6	0	0	21	0	0
11.	859	48	48	— 637	406	30	— 126	16	10	636	405	368	— 68	5	5
12.	859	25	25	— 845	716	28	— 51	3	1	— 281	79	38	154	24	13
13.	902	73	73	— 830	691	79	— 344	119	108	91	8	11	240	57	90
14.	866	28	28	— 892	797	35	— 177	32	11	— 139	19	10	34	1	1
15 A.	768	33	33	— 802	645	33	— 57	3	1	— 284	81	50	156	24	17
15 B.	740	4	4	— 532	284	2	— 36	1	0	— 491	242	17	— 281	80	6
16.	909	25	25	— 195	38	2	171	29	9	— 103	11	5	380	144	78
17.	445	6	6	— 230	54	0	317	101	7	— 73	5	1	110	12	2
18.	695	16	16	— 795	633	16	— 80	7	1	— 42	2	1	165	27	9
19.	569	2	2	— 646	419	1	— 74	6	0	111	12	1	276	76	4
20.	920	25	25	— 777	605	24	— 86	8	2	— 499	250	120	— 188	36	19
21.	780	27	27	— 716	514	22	320	102	34	154	24	12	— 50	3	2
22.	879	23	23	— 564	319	11	611	373	105	— 134	18	8	139	19	9
23.	921	35	35	— 845	715	40	— 191	37	16	99	10	7	294	87	66
24.	891	22	22	— 725	527	18	490	240	66	76	6	2	167	28	13
29.	859	24	24	— 717	515	20	519	269	82	— 92	9	4	49	2	1
30.	865	8	8	— 739	548	6	410	168	16	— 163	27	4	216	47	8
31.	912	67	67	— 887	788	84	59	3	3	44	2	3	— 75	6	8
32.	936	27	27	— 839	705	30	400	160	54	— 71	5	3	179	32	19
33.	991	163	163	— 920	848	217	— 21	0	1	125	16	49	— 269	73	258
34.	830	15	15	— 849	723	17	245	60	11	29	1	0	100	10	3
35.	942	16	16	— 676	458	12	510	260	52	— 122	15	5	215	46	16
36.	914	8	8	— 905	821	10	166	28	3	— 102	11	2	135	18	3
37.	897	7	7	— 903	818	9	— 75	6	1	— 89	8	1	— 129	17	3
	0,0		1 000			1 000			1 000			1 000			1 000

Tableau des aides à l'interprétation des résultats de la seconde analyse **(suite et fin)**

Branches

J 1	QLT	POID	INR	1 # F	COR	CTR	2 # F	COR	CTR	3 # F	COR	CTR	4 # F	COR	CTR
+ 4.	700	2	2	— 809	656	2	49	2	0	— 42	2	0	71	5	0
+ 5.	756	17	17	— 746	557	15	122	15	3	293	86	28	— 171	30	11
+ 6.	899	36	36	— 710	506	28	329	108	48	— 104	11	7	— 49	3	2
+ 7.	855	37	37	— 800	642	37	22	0	0	154	24	17	— 238	57	46
+ 8.	861	10	10	— 837	702	11	94	9	1	262	69	14	— 245	60	14
+ 9.	860	20	20	— 860	741	23	176	31	8	102	10	4	60	4	2
+ 10.	854	6	6	— 842	711	6	101	10	1	245	60	7	— 88	8	1
+ 11.	889	22	22	— 843	712	25	174	30	9	— 85	7	3	— 106	11	6
+ 12.	921	46	46	— 793	630	46	27	1	0	510	260	229	— 137	19	19
+ 13.	748	40	40	— 791	628	40	— 222	50	25	— 87	8	6	182	33	29
+ 14.	940	51	51	— 854	731	59	— 291	85	55	63	4	4	284	81	90
+ 15 A.	925	39	39	— 897	807	50	— 193	38	18	247	61	45	110	12	10
+ 15 B.	805	12	12	— 850	725	14	— 106	11	2	73	5	1	205	42	11
+ 16.	969	60	60	— 835	699	66	— 318	101	75	73	5	6	385	149	192
+ 17.	904	29	29	— 898	808	37	— 241	58	21	— 45	2	1	94	9	6
+ 18.	920	24	24	— 839	705	27	103	11	3	421	177	83	— 106	11	6
+ 19.	804	6	6	— 779	608	6	80	6	0	391	153	17	— 4	0	0
+ 20.	878	25	25	— 887	788	31	— 141	20	6	137	19	9	187	35	19
+ 21.	834	9	9	— 833	696	9	135	18	2	227	52	8	— 217	47	9
+ 22.	736	23	23	— 656	432	16	209	44	13	101	10	5	— 99	10	5
+ 23.	774	23	23	— 748	561	20	— 55	3	1	451	204	87	— 49	2	1
+ 24.	994	186	186	— 892	797	233	— 246	61	141	— 288	84	296	— 216	47	190
+ 29.	941	39	39	— 531	283	17	98	10	5	— 79	6	5	377	142	120
+ 30.	822	13	13	— 842	711	15	252	64	10	7	0	0	13	0	0
+ 31.	896	51	51	— 722	523	42	421	177	112	— 91	8	8	84	7	8
+ 32.	722	12	12	— 730	534	10	283	80	12	— 60	4	1	— 200	40	11
+ 33.	913	53	53	— 617	383	32	564	319	209	— 188	36	36	245	60	68
+ 34.	855	48	48	— 750	564	43	420	177	106	— 237	56	52	94	9	9
+ 35.	857	11	11	— 697	488	9	201	40	6	188	35	8	— 316	101	25
+ 36.	830	19	19	— 581	339	10	107	11	3	187	35	13	— 446	200	84
+ 37.	831	29	29	— 651	425	19	539	290	104	32	1	1	— 108	12	7
	0,0		1 000			1 000			1 000			1 000			1 000

● Références bibliographiques

[1] BENZÉCRI J.-P. — *L'analyse des données* (vol. 1 : *La taxinomie:* vol. 2 : *Correspondances*), Dunod, 1973.

[2] BENZÉCRI J.-P. — « Mémoire reçu : analyse des correspondances sur la sphère », *Les cahiers de l'analyse des données,* vol. III, 1978, n° 4.

[3] ESCOFIER B. — « Analyse factorielle et distances répondant au principe d'équivalence distributionnelle », *Revue de statistique appliquée,* vol. XXVI, n° 4, 1978.

[4] KAMINSKI P. — « Généralisations de l'analyse des correspondances des tableaux de contingence : propriétés de l'invariance distributionnelle », note ronéotée INSEE, service des Programmes, février 1979.

[5] LE CAM L. — « On the Assumptions used to Prove Asymptotic Normality of Maximum Likelihood Estimates », *the Annals of Mathematical Statistics,* vol. 41, n° 3, 1970.

[6] LICHNEROWICZ A. — *Éléments de calcul tensoriel,* Armand-Colin, 1964.

[7] RENYI A. — *Calcul des probabilités,* Dunod, 1966.

[8] TABET N. — *Programme d'analyse factorielle des correspondances,* polycopié de laboratoire du professeur Benzécri, tour 45-55, 2e étage, 4, place Jussieu, 75230 Paris Cedex 05, mars 1973.

[9] VOLLE M. — « Analyse arborescente des tableaux d'échange ou de transition », note ronéotée INSEE, *Unité de Recherche,* février 1976.

[10] VOLLE M. — « Analyse des correspondances sur la sphère », note ronéotée, *Unité de Recherche,* août 1978.

[11] VOLLE M. — « *Analyse des données* », Economica, 1978.

Summary

Spherical factor analysis : an exploration

Dominique DOMENGES and Michel VOLLE

This study originated in a remark by J.-P. BENZECRI : the distributions on a finite set I can be plotted on a spherical orthant; that is, if $\sum p_i = 1$ where $p_i \geqslant 0$, the distribution p can be represented by a point with coordinates $\sqrt{p_i}$. It is further possible to define on the spherical surface a distance between distributions that corresponds to the metric of χ^2 that is habitually used in the simplex of distributions. This representation of distributions paves the way for a new method of factor analysis, with close ties to the factor analysis of correspondances. We shall rapidly explore the domain of applications, and finally provide several examples of concrete results.

Reseña

Análisis factorial esférico : una exploración

Dominique DOMENGES y Michel VOLLE

Este trabajo está originado por una observación de J.-P. BENZECRI : las distribuciones en un conjunto finito I se pueden localizar en uno de los cuadrantes de esfera (en francés : orthant); en efecto, si $\sum p_i = 1$ con $p_i \geqslant 0$, la distribución p se puede representar por medio de un punto de coordenadas $\sqrt{p_i}$. Es factible, además, definir sobre la esfera una distancia entre distribuciones, la que corresponda a la métrica del χ^2 que se suele utilizar en el simplex de las distribuciones. Esta representación de las distribuciones dá paso a un nuevo método de análisis factorial, el que tiene rigorosos vínculos con el análisis factorial de las correspondencias. Exploraremos brevemente el terreno de sus aplicaciones y, por último, presentaremos unos cuantos ejemplos concretos de los resultados que suministra.