REVUE DE STATISTIQUE APPLIQUÉE

B. ESCOFIER

Analyse factorielle et distances répondant au principe d'équivalence distributionnelle

Revue de statistique appliquée, tome 26, n° 4 (1978), p. 29-37
<http://www.numdam.org/item?id=RSA_1978__26_4_29_0>

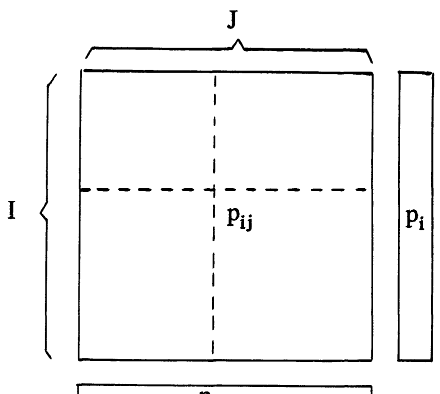
ANALYSE FACTORIELLE ET DISTANCES RÉPONDANT AU PRINCIPE D'ÉQUIVALENCE DISTRIBUTIONNELLE

B. ESCOFIER*

RESUME

On donne ici quelques exemples de distances définies sur un ensemble à partir d'un tableau de correspondance. Ces distances, comme celle du χ^2 utilisée en analyse des correspondances, sont des distances entre profils, satisfont au "principe d'équivalence distributionnelle", et sont représentables dans un espace euclidien, donc utilisables pour une analyse factorielle.

1. NOTATIONS ET RAPPELS



Soit p_{ij} un tableau de correspondance, p_i et p_j les sommes des lignes et des colonnes. On *appelle profil associé à* $j \in J$ l'ensemble $\left\{ \frac{p_{ij}}{p_j} \text{ noté } p(i/j) \; ; i \in I \right\}$, et profil associé à $i \in I$ l'ensemble $\left\{ \frac{p_{ij}}{p_i} \text{ noté } p(j/i) \; ; j \in J \right\}$;

*I.N.S.A. Rennes, 20, avenue des Buttes de Coësmes – B.P. 14A, 35031 Rennes Cedex.

Mots Clés : **Analyse factorielle, analyse des correspondances, distance du χ^2, équivalence distributionnelle.**

On dit qu'une distance définie sur I est *une distance entre profils*, si la distance entre i et i' est nulle quand les profils qui leur sont associés sont égaux, et petite quand ces profils sont proches, (quelles que soient les valeurs de p_i et de $p_{i'}$). Et le *principe d'équivalence distributionnelle* est le suivant : si deux éléments j_1 et j_2 de J, de même profil (ou équivalents distributionnellement) sont remplacés par un seul élément j_o avec $p_{ij_0} = p_{ij_1} + p_{ij_2}$, la distance définie sur I reste inchangée.

Rappelons aussi la distance du χ^2, de centre la loi marginale p_j :

$$D^2(i,i') = \sum_{j \in J} \frac{1}{p_j} (p(j/i) - p(j/i'))^2$$

2. ETUDE DE LA DISTANCE $D^2(i,i') = \sum_{j \in J} (\sqrt{p(j/i)} - \sqrt{p(j/i')})^2$

Il est clair que cette distance est une distance entre profils. On vérifie facilement qu'elle satisfait au principe d'équivalence distributionnelle en l'écrivant :

$$D^2(i,i') = \sum_{j \in J} p(j) (\sqrt{p(i/j)/p(i)} - \sqrt{p(i'/j)/p(i')})^2 \tag{2}$$

En effet, soient j_1 et j_2 deux éléments de J de même profil : $p(i/j_1) = p(i/j_2)$ pour tout i. Si l'on remplace j_1 et j_2 par un seul élément j_0 tel que $p_{ij_0} = p_{ij_1} + p_{ij_2}$, alors $p_{j_0} = p_{j_1} + p_{j_2}$; et pour tout i, les égalités $\frac{p_{ij_0}}{p_{j_0}} = \frac{p_{ij_1}}{p_{j_1}} = \frac{p_{ij_2}}{p_{j_2}}$ sont vérifiées.

L'élément j_0 a donc le même profil que j_1 et j_2. Et dans l'expression de la distance, les deux termes correspondant à j_1 et j_2 sont remplacés par leur somme.

Une *représentation euclidienne* évidente de cette distance est de placer dans $\mathbf{R}_J$, muni du produit scalaire identité, l'élément i au point de coordonnées $\sqrt{p(j/i)}$. On remarque que *tous les points sont situés sur la sphère unité* puisque : $\sum_j (\sqrt{p(j/i)})^2 = 1$.

Pour la recherche des *axes principaux d'inertie du nuage* associé à cette distance, il faut affecter à chaque point un poids. On peut, comme en correspondance, affecter à i le poids p_i, ou bien, en liaison avec la formule de définition de la distance, $\sqrt{p_i}$, ou même affecter le même poids à tous les éléments. On peut remarquer que dans la seconde solution, pour des tableaux sous forme disjonctive complète, le centre de gravité est sur la première bissectrice, comme en correspondance.

En affectant à i le poids p_i, le centre de gravité a pour coordonnées :

$$G_j = \sum_i \sqrt{p_i p_{ij}} / \sum_i p_i$$

Les axes principaux d'inertie sont les vecteurs propres V_λ de la matrice de la forme quadratique principale d'inertie :

$$M_{jj'} = \quad p_i (\sqrt{p(j/i)} - G_j)(\sqrt{p(j'/i)} - G_{j'})$$

Et les facteurs, projection du nuage sur les axes d'inertie s'écrivent, si V_λ est normé :

$$F_\lambda(i) = \sum_j \sqrt{p(j/i)\, V_\lambda}\,(j)$$

On peut représenter simultanément les deux ensembles I et J, de plusieurs manières : en représentant simultanément V_λ et F_λ, ce qui revient à projeter sur V_λ les vecteurs unitaires e_j de la base canonique de $\mathbf{R}_J$ et le nuage étudié ; $V_\lambda(j)$ est alors le cosinus de l'angle entre V_λ et e_j. Ou bien représentant un élément j de J au centre de gravité des éléments i affectés du poids p_{ij}. Il n'y a évidemment pas comme en correspondance de symétrie des représentations de I et J.

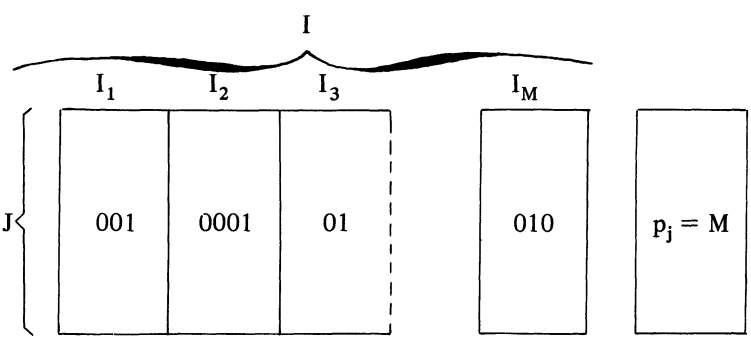
Etudions maintenant les propriétés de cette distance. Remarquons d'abord qu'en effectuant les calculs dans sa deuxième expression, elle s'écrit :

$$D^2(i,i') = 2\left(1 - \sum \frac{\sqrt{p_{ij}\, p_{i'j}}}{\sqrt{p_i\, p_{i'}}}\right) \qquad (3)$$

Ce qui montre qu'elle est égale à 2 si les domaines de i et i' sont disjoints, quelles que soient leurs marginales p_i et $p_{i'}$ (le domaine de i est l'ensemble des j de J tels que p_{ij} soit différent de zéro).

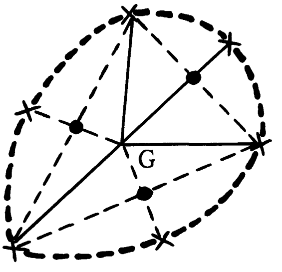
Mais nous allons étudier spécialement le cas où le tableau de données est constitué de 0 et de 1 et plus particulièrement celui où les données sont mises *sous forme disjonctive complète.*

Un tableau est mis sous forme disjonctive complète si il existe une partition de I en M sous ensemble $I_1, \ldots, I_M$ telle que pour j de J et tout q compris entre 1 et M, p_{ij} vaut 1 pour un et seul des éléments de I_q. [1] et [3]. On code ainsi les questionnaires où, pour chaque question un individu j doit choisir une réponse i et une seule. Ou bien, des ensembles de mesures où le domaine de valeurs de chaque mesure q est partitionné : on pose alors $p_{ij} = 1$ si la mesure effectuée sur j appartient à l'élément i de la partition des valeurs de q.



En traitant ce type de tableaux par l'analyse des correspondances, les éléments i de I de marginale p_i petite, (réponse peu choisie dans un questionnaire par exemple), sont très éloignés du centre de gravité, alors que ceux de marginale grande en sont très proches. Ceci ne se justifie pas toujours pour les données traitées, et de tels éléments perturbent souvent les analyses. Dans le cas de mesures ainsi codées, le nombre d'éléments de la partition des valeurs d'une mesure q est généralement sans relation avec son importance relative. Elle joue cependant un grand rôle en correspondance dans les résultats de l'analyse : plus le nombre d'éléments de la partition est grand, plus les effectifs sont petits et plus les éléments associés sont loin du centre de gravité et plus d'inertie totale de la mesure est grande. L'inertie de l'ensemble des éléments de I_q par rapport au centre de gravité vaut Card $(I_q) - 1$. Comme de plus, la symétrie des analyses des 2 ensembles, caractéristique de l'analyse des correspondances ne s'appuie pas ici sur une symétrie des données, comme pour un tableau de fréquence. il peut être intéressant d'introduire une distance différente de celle du χ^2.

Pour des données sous forme disjonctive complète, il est possible de comparer le nuage de points associé à la distance étudiée et celui construit en analyse des correspondances. En effet, dans ce cas, en correspondance, la métrique induite sur $\mathbf{R}_J$, (où l'élément i a pour coordonnées p (i/j)) est une métrique uniforme puisque la marginale p_j est égale à M pour tout j. Il suffit d'une simple homothétie pour placer ces deux nuages dans le même espace. Le graphique ci-dessous représente ces deux nuages dans $\mathbf{R}_J$.



• Nuage associé aux correspondances. Les points ont leurs coordonnées non nulles toutes égales à $1/p_i$ et une norme égale à $\sqrt{1/p_i}$.

* Nuage déduit. Les points ont pour norme 1, pour coordonnées non nulles $1/\sqrt{p_i}$. Si on affecte à chaque élément i les poids p_i, les inerties totales de tous les sous-ensemble I_q par rapport à l'origine sont égales. Elles valent, puisque tous les points ont pour norme 1,

$$\sum_{i \in I_q} p_i = \text{card } J$$

Le point représentant l'élément i a pour coordonnées dans $\mathbf{R}_J$, p_{ij}/p_i pour l'analyse des correspondances et $\sqrt{p_{ij}/p_i}$ pour cette nouvelle distance. Comme p_{ij} vaut 0 ou 1, $\sqrt{p_{ij}}$ est égal à p_{ij}, le deuxième point se déduit du premier par une homothétie de rapport $\sqrt{p_i}$. Ce deuxième point se trouvant sur la sphère unité, c'est en quelque sorte une "normalisation" du nuage.

Il peut être intéressant de faire conjointement à une classification automatique, une analyse factorielle basée sur la même notion de proximité, les résultats obtenus pouvant s'éclairer réciproquement. Or, la distance étudiée ici est, pour les tableaux en 0-1, (non nécessairement sous forme disjonctive complète) liée au coefficient de similarité S. d'Ochiai [4], par la relation :

$$S(i,i') = 1 - \frac{1}{2} D^2(i,i').$$

En effet, on a :

$$S(i,i') = \frac{oo}{\sqrt{(oo + No)(oo + oN)}} = \sum_j \frac{p_{ij}\, p_{i'j}}{\sqrt{p_i\, p_{i'}}}$$

où oo, oN, No, NN désignent respectivement le nombre de colonnes comportant aux ligne i et i' : 1 et 1, 1 et o, o et 1, o et o. Et, puisque p_{ij} valant o ou 1, $p_{ij} \cdot p_{i'j} = \sqrt{p_{ij}} \cdot p_{i'j}$, on retrouve l'expression (3) de la distance.

Une distance peut être *définie symétriquement sur* J. Pour des données sous forme disjonctive complète, la distance se simplifie, les p_j étant tous égaux à M, elle devient :

$$D^2(j,j') = \frac{1}{p_j} \sum_i (p_{ij} - p_{ij'})^2 = \frac{1}{M} \sum_i (p_{ij} - p_{ij'})^2$$

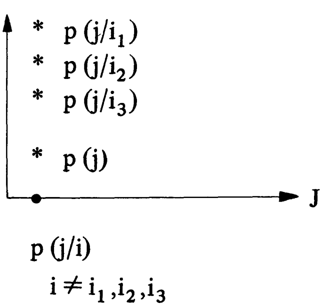
(pour un questionnaire, ceci représente le rapport entre le nombre de réponses distinctes des individus j et j' au nombre total de réponses). Dans ce cas le coefficient de similarité auquel est liée la distance est confondu avec S_{SM}, celui introduit en taxinomie numérique par Sokal et Michener :

$$S_{SM}(j,j') = 1 - \frac{oN + No}{oo + oN + No + oo} = 1 - \frac{1}{2} D^2(j,j')$$

3. ETUDE DE LA DISTANCE

$$D^2(i,i') = \sum_j (p(j/i) - p(j/i'))^2 \frac{r_j}{p_j} \quad \text{avec} \quad r_j = \sum_{i/p_{ij} \neq 0} p_i$$

C'est *une distance du* χ^2, mais *de centre la loi* $n_j = p_j/r_j$ et non la loi marginale p_j comme en correspondance. Pour des tableaux ne comportant pas de zéro, ces deux distances sont évidemment égales. Pondérer en chaque point l'écart $(p(j/i) - p(j/i'))^2$ par la moyenne p_j de ces lois affectées des poids p_i paraît logique : une petite différence entre des valeurs en moyenne petites est aussi importante qu'une grande sur des valeurs en moyenne grandes. Mais que se passe-t-il si pour un j donné, beaucoup de p(j/i) sont égaux à zéro ? La moyenne sera très petite relativement aux p(j/i) non nuls, l'importance donnée à l'écart sur cette coordonnée sera grande.



Au lieu de quotienter par la moyenne des p(j/i), on peut alors *quotienter par la moyenne n_j des* p(j/i) *non nuls en* j :

$$p_j = \frac{\sum_i p_i \, p(j/i)}{\sum_i p_i} \qquad n_j = \frac{\sum_{i/p_{ij} \neq 0} p_i p(j/i)}{\sum_{i/p_{ij} \neq 0} p_i} = \frac{p_j}{r_j}$$

Il est visible que la distance ainsi définie est une distance entre profils. Remarquons que si les profils associés à j_1 et j_2 sont égaux, alors $r_{j1} = r_{j2}$. On vérifie alors facilement le principe d'équivalence distributionnelle en écrivant comme ci-dessus :

$$D^2(i,i') = \sum_j (p(i/j)/p_i - p(i'/j)/p_{i'})^2 \, r_j \, p_j$$

Il suffit de munir $\mathbf{R}_J$ du produit scalaire $\delta_j^{j'}$ (r_j/p_j) et de placer comme en correspondance l'élément i au point de coordonnées p(j/i) pour obtenir un *nuage de points* représentant cette distance dans un espace euclidien. Le centre de gravité G est, comme en correspondance, le point de coordonnées p_j si on affecte à i le poids p_i. Les projections sur les axes principaux d'inertie V_λ sont alors les vecteurs propres de :

$$M_{jj'} = \sum_i \frac{r_j}{p_j} \left(\frac{p_{ij}}{p_j} - p_j\right)\left(\frac{p_{ij'}}{p_i} - p_{j'}\right) p_i$$

(On ne peut comme en correspondance se ramener à l'étude de la forme quadratique d'inertie à l'origine, le nuage n'étant généralement plus orthogonal à OG pour le produit scalaire $\delta_j^{j'}$ (r_j/p_j)).

Les facteurs $\mathcal{F}_\lambda$ sont égaux à : $\mathcal{F}_\lambda(i) = \sum_j p(j/i) \, V_\lambda(j)$ et les remarques du § précédent concernant la représentation simultanée des ensembles I et J sont toujours valables.

Comparons les résultats obtenus avec ceux de l'analyse des correspondances.

La matrice M dont les vecteurs propres sont ici les axes principaux d'inertie, se déduit de son homologue N de l'analyse des correspondances par multiplication par la matrice diagonale D de "changement de métrique" de terme général $\delta_j^{j'}$ $(1/r_j)$. On a M = DN. On sait [2] comparer les vecteurs propres de M et de N :

Si θ est l'angle dans $\mathbf{R}_J$ muni de la métrique des correspondances, entre les axes principaux d'inertie de rang s de ces deux analyses, si γ_s est la valeur propre de rang s des correspondances et ϵ son écart avec les autres valeurs propres, on a :

si $\gamma_s (\sup_j r_j / \inf_j r_j - 1) < \epsilon$ avec $\epsilon = \inf(\lambda_{s-1} - \lambda_s, \lambda_s - \lambda_{s+1})$

Alors

$$\sin 2\theta < \frac{\gamma_s}{\epsilon}\left(\frac{\sup r_j}{\inf r_j} - 1\right) \quad \text{et} \quad \theta < \frac{\pi}{4}$$

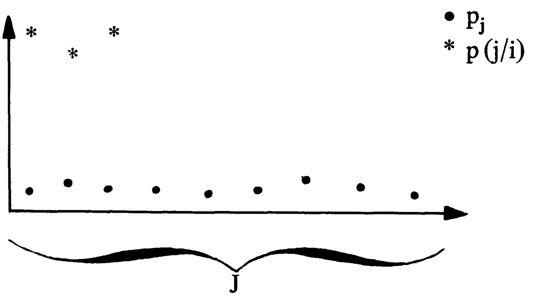
Ceci implique en particulier que cet angle est nul, i.e. que les résultats des 2 analyses sont identique si les r_j sont tous égaux. C'est le cas notamment si aucun des p_{ij} n'est nul, les distances étant alors les mêmes. Sinon les résultats différeront surtout si le nombre de p_{ij} nuls par colonne est très différent suivant les colonnes.

Pour des tableaux mis sous forme disjonctive complète, et pour la distance définie par cette formule sur l'ensemble J, on obtient le même résultat qu'avec la distance précédente : $D^2(j,j') = \frac{1}{p_j} \sum_j (p_{ij} - p_{ij'})^2$.

4. ETUDE DE LA DISTANCE

$$D^2(i,i') = \sum_j (p(j/i)\ \mathcal{C}_i - p(j/i')\ \mathcal{C}_{i'})^2 \ \frac{1}{p_j} \qquad \text{avec } \mathcal{C}_i = \sum_{j/p_{ij} \neq 0} p_j$$

Cette distance, comme la précédente, est égale à celle utilisée en correspondance si aucun des p_{ij} n'est nul. Sinon, elle rapproche de la moyenne, et des éléments i', les i de domaine très restreint. En correspondance, ces éléments bien que d'effectifs généralement très faibles, sont situés très loin du centre de gravité. Ils ont dans la détermination des facteurs une importance qu'ils usurpent quelquefois : on est souvent amené pour obtenir des résultats interprétables à les supprimer de l'analyse.



Les sommes sur J de la moyenne p_j, et de $p(j/i)$ sont toutes deux égales à 1. Si le domaine de i est très restreint les valeurs de $p(j/i)$ non nulles sont donc très grandes relativement à $p(j)$..

En multipliant $p(j/i)$ par $\mathcal{C}_i$ somme sur le domaine de i des $p(j)$, on le rapproche de la marginale $p(j)$: $p(j)$ et $\mathcal{C}_i\, p(j/i)$ ont toutes deux sur le domaine de i une somme égale à $\mathcal{C}_i$.

Par rapport à la distance du χ^2 classique, *la distance définie au paragraphe précédent* diminuait l'importance des *colonnes contenant beaucoup de zéros. Celle définie dans ce paragraphe* diminue directement les distances à la moyenne et aux autres éléments des *lignes contenant beaucoup de zéros.*

Pour des *tableaux mis sous forme disjonctive complète*, l'expression de cette distance se simplifie. On a, au facteur constant $p_j = M$ près, $D^2(i,i') = \sum_j (p_{ij} - p_{i'j})^2$.
En effet p_j est alors constant et $\mathcal{C}_i = p_i p_j$. Pour un questionnaire, cette distance représente le nombre d'individus ayant choisi l'une et non l'autre des réponses i et i',

c'est la distance liée au coefficient de similarité S_{Sn} de Sokal et Michener que nous avions déjà obtenue dans les 2 premiers paragraphes pour des tableaux sous forme disjonctive complète, mais c'était alors sur l'ensemble J.

Pour un tableau quelconque, vérifions que la distance ainsi définie est une distance entre profils. Si $p(j/i) = p(j/i')$ pour tout j, alors $p(i,j) = 0 \Leftrightarrow p(i',j) = 0$, donc $\mathcal{C}_i = \mathcal{C}_{i'}$ et $D^2(i,i')$ est nulle. Si les profils sont proches, alors $\mathcal{C}_i$ et $\mathcal{C}_{i'}$ sont peu différents et la distance $D^2(i,i')$ est faible.

Pour vérifier le *principe d'équivalence distributionnelle*, il suffit de remarquer que si les profils associés à j_1 et j_2 sont égaux, comme $p(i,j_1) = 0 \Leftrightarrow p(i,j_2) = 0$, le regroupement de j_1 et de j_2 ne modifie pas les $\mathcal{C}_i$. Et d'autre part, que la distance s'écrit :

$$D^2(i,i') = \sum_j p(j) \left(\frac{p(i/j)}{p_i} \mathcal{C}_i - \frac{p(i'/j)}{p_{i'}} \mathcal{C}_{i'}\right)^2$$

Dans $\mathbf{R}_J$ muni comme en correspondance du produit scalaire $\delta_j^{j'}(1/p_j)$, on peut construire un *nuage de points* en plaçant l'élément i au point de coordonnées $\mathcal{C}_i\, p(j/i)$. Ce point se déduit du point construit en correspondance par une homothétie de centre l'origine et de rapport $\mathcal{C}_i$. La recherche des axes principaux d'inertie et des facteurs se fait de façon analogue aux cas précédents : en affectant à i le poids p_i, le centre de gravité a pour coordonnées $\mathcal{G}_j = \sum_i \mathcal{C}_i\, p_{ij}$. Les projections sur les axes principaux d'inertie sont les vecteurs propres de la matrice :

$$M_{jj'} = \sum_i \frac{p_i}{p_j} \left(\mathcal{C}_i \frac{p_{ij}}{p_i} - \mathcal{G}_j\right) \left(\mathcal{C}_i \frac{p_{ij'}}{p_i} - \mathcal{G}_{j'}\right)$$

CONCLUSION

Ces 3 distances ne sont pas les seules distances euclidiennes, avec celle utilisée en analyse des correspondances à satisfaire à l'axiome d'équivalence distributionnelle.

Mais elles permettent chacune de résoudre des problèmes particuliers ; la première liée pour des tableaux en 0-1 au coefficient de similarité d'Ochiai, permet d'obtenir un nuage "normalisé" et pour des tableaux sous forme disjonctive complète d'égaliser l'inertie de chaque sous-ensemble q. Relativement à la distance du χ^2 la deuxième diminue dans la distance entre les lignes, l'importance des colonnes contenant beaucoup de zéros et la troisième diminue la distance des lignes contenant beaucoup de zéros. Pour des tableaux sous forme disjonctive complète, ces distances sont liées pour l'un des deux ensembles au coefficient de similarité de Sokal et Michener.

A titre de curiosité, on peut montrer qu'il existe une infinité dénombrable de distances dont les facteurs extraits par l'analyse factorielle sont identiques à ceux obtenus en correspondance. On les construit par récurrence de la manière suivante : si p_{ij} est un tableau de correspondance quelconque du produit I x J, on peut construire sur le produit I x I le tableau de correspondance symétrique $t_{ii'} = \sum_j \frac{p_{ij}\, p_{i'j}}{p_j}$

Les facteurs sur I déduits du tableau p_{ij} et du tableau $t_{ii'}$ sont identiques et les valeurs propres du second sont les carrés de celles du premier. Le tableau $t_{ii'}$ induit sur I la distance du χ^2 :

$$d^2(i\,,i') = \sum_{i''} \frac{1}{p_{i''}} \left[\sum_j \frac{p_{ij}\, p_{i''j}}{p_j\, p_i} - \sum_j \frac{p_{i'j}\, p_{i''j}}{p_j\, p_{i'}} \right]^2$$

$$= \sum_{i''} \frac{1}{p_{i''}} \left[\sum_j p(j/i)\, p(i''/j) - \sum_j p(j/i')\, p\,(i''/j) \right]^2$$

Cette distance entre profils satisfait au principe d'équivalence distributionnelle puisque le tableau $t_{ii'}$ reste inchangé si deux éléments j et j' équivalents sont remplacés par un seul élément. En répétant l'opération sur le tableau $t_{ii'}$ on obtient une nouvelle distance ayant les mêmes propriétés, etc. . .

BIBLIOGRAPHIE

[1] BENZECRI. – *L'analyse des données.* Dunod, 1973.

[2] ESCOFIER, LE ROUX. – *Etude de trois problèmes de stabilité en analyse factorielle.* Publication de l'ISUP, 1976.

[3] LEBART, MORINEAU, TABARD. – *Technique de la description statistique.* Dunod, 1977.

[4] LERMAN. – *Les bases de classification automatique.* Gauthier-Villars, 1970.