A. LECLERC

L'analyse des correspondances sur juxtaposition de tableaux de contingence

Revue de statistique appliquée, tome 23, nº 3 (1975), p. 5-16
<http://www.numdam.org/item?id=RSA_1975__23_3_5_0>

L'ANALYSE DES CORRESPONDANCES SUR JUXTAPOSITION DE TABLEAUX DE CONTINGENCE (1)

A. LECLERC
INSERM, Division de la Recherche Médico-Sociale

On sait que l'analyse des correspondances est particulièrement adaptée à l'étude des tableaux de contingence ; elle permet alors d'étudier les relations entre deux variables qualitatives. Si l'on cherche à étudier les relations entre deux *groupes* de variables qualitatives, et que l'on dispose des valeurs prises par chacune de ces variables sur les mêmes sujets, on peut former un tableau "juxtaposition de tableaux de contingence" croisant l'ensemble des modalités des variables du premier groupe avec l'ensemble des modalités des variables du second groupe.

Nous nous demandons ici quelles propriétés possède l'analyse des correspondances effectuée sur un tableau de ce genre, et comment on peut interpréter au mieux ses résultats.

I – ANALYSE D'UN TABLEAU DE CONTINGENCE

1/ Rappel sur l'analyse des correspondances

Pour un exposé complet sur l'analyse des correspondances, il faut se référer à (1) ou (2). Mais on peut ici rappeler les grandes lignes de la méthode, en adoptant une présentation assez simple.

Soit $I \times J$ un tableau que l'on suppose être ici un tableau de contingence, et que l'on cherche à analyser. On notera :

* n_{ij} l'effectif de la case (i , j)
* n_i et n_j les effectifs marginaux : somme des éléments d'une ligne ou d'une colonne.
* n la somme des éléments du tableau : c'est ici le nombre de sujets
* $f_{ij} = n_{ij}/n$ la fréquence associée à la case (i , j)
* f_i et f_j les fréquences marginales : n_i/n et n_j/n
* $f^i_j = f_{ij}/f_i$ et $f^j_i = f_{ij}/f_j$ les fréquences conditionnelles

Pour simplifier, on appellera aussi I et J le nombre de lignes et le nombre de colonnes.

---

(1) Article remis en Décembre 1973, révisé en Octobre 1974.

On imagine dans un espace R_J (espace à J dimensions), muni de la métrique $1/f_j$, le nuage des points "i" représentés par des vecteurs de coordonnées f_j^i, auxquels sont associés des masses f_i. L'analyse des correspondances sera l'analyse de ce nuage par rapport à son centre de gravité.

Il revient absolument au même de munir R_J de la métrique euclidienne classique et d'analyser alors, par rapport à leur centre de gravité, des points "i" de coordonnées $f_{ij}/f_i\sqrt{f_j}$, toujours munis de masses f_i, (analyse en composantes principales classique).

L'analyse fournira des fonctions projection sur les axes factoriels d'inertie (directions d'allongement du nuage). On réservera le nom de "facteurs" aux fonctions s'appliquant, non pas aux vecteurs de composantes $f_{ij}/f_i\sqrt{f_j}$, mais aux vecteurs f_J^i de composantes $f_{ij}/f_i = f_j^i$, représentant une loi de probabilité sur J, connaissant i. On appellera φ_q^J une fonction de ce type, correspondant qu q-ème axe factoriel, et φ_q^j une composante de cette fonction :

$$\varphi_q^J (f_J^i) = \sum_{j \in J} \varphi_q^j \frac{f_{ij}}{f_i}$$

Une démarche parallèle peut être appliquée à l'analyse de points "j" dans un espace R_I. On appellera de façon analogue φ_q^I une fonction s'appliquant aux vecteurs "j" de composantes f_{ij}/f_j, et φ_q^i une composante de cette fonction.

Les valeurs propres des deux analyses sont les mêmes : on les notera λ_q. On appellera Q l'ensemble des indices q caractérisant des couples de facteurs $(\varphi_q^I , \varphi_q^J)$ correspondant aux valeurs propres λ_q rangées dans l'ordre décroissant.

2/ Quelques propriétés de l'analyse des correspondances

Les caractéristiques de la méthode, rappelées ici, serviront de référence pour l'analyse d'une juxtaposition de tableaux de contingence. Toutes les formules présentées sont vérifiées en effet quelque soit la nature du tableau analysé. Seule leur utilisation en vue de l'interprétation peut différer.

Signification de l'inertie totale

Les deux nuages de points évoqués plus haut ont même inertie par rapport à leur centre de gravité. Pour les points "i" de coordonnées f_j^i, de masses f_i, dans l'espace R_J muni de la métrique $1/f_j$, cette inertie s'écrit :

$$\sum_{i \in I} f_i \sum_{j \in J} \frac{1}{f_j} [f_j^i - f_j]^2$$

(le centre de gravité est le point de coordonnées fj)

L'inertie s'écrit sous une forme plus symétrique :

$$\sum_{j \in J} \sum_{i \in I} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \frac{1}{n} \sum_{j \in J} \sum_{i \in I} \frac{(n\, f_{ij} - n\, f_i f_j)^2}{n\, f_i f_j}$$

La quantité de droite rappelle une expression connue. C'est ce que l'on calcule habituellement sur un tableau de contingence, et qui suit un Khi – 2 sous certaines hypothèses, d'où le résultat :

Dans l'analyse des correspondances appliquée à un tableau de contingence, n fois l'inertie suit, sous l'hypothèse d'indépendance entre lignes et colonnes, et à condition que les effectifs ne soient pas trop petits, un Khi – 2 à (I – 1) (J – 1) degrés de liberté.

Propriétés des facteurs

$$* \qquad \sum_{i' \in I} \sum_{j \in J} \varphi_q^{i'} \frac{f_{i'j}}{f_j} \frac{f_{ij}}{f_i} = \lambda_q \varphi_q^i$$

$$\sum_{j' \in J} \sum_{i \in I} \varphi_q^{j'} \frac{f_{ij'}}{f_i} \frac{f_{ij}}{f_j} = \lambda_q \varphi_q^j \qquad (1)$$

Ce qui peut s'écrire en notations condensées :

$$\varphi_q^I \circ f_I^J \circ f_J^I = \lambda_q \varphi_q^I$$

$$\varphi_q^J \circ f_J^I \circ f_I^J = \lambda_q \varphi_q^J$$

$$* \qquad \sum_{i \in I} \varphi_q^i \frac{f_{ij}}{f_j} = \sqrt{\lambda_q}\, \varphi_q^j$$

$$\sum_{j \in J} \varphi_q^j \frac{f_{ij}}{f_i} = \sqrt{\lambda_q}\, \varphi_q^i \qquad (2)$$

ou : $\varphi_q^I \circ f_I^j = \sqrt{\lambda_q}\, \varphi_q^j$ et $\varphi_q^J \circ f_J^i = \sqrt{\lambda_q}\, \varphi_q^i$

$$* \qquad \sum_{i \in I} \varphi_q^i f_i = 0$$

$$\sum_{j \in J} \varphi_q^j f_j = 0 \qquad (3)$$

ou : $\varphi_q^I \circ f_I = \varphi_q^J \circ f_J = 0$

Ceci entraîne que dans une représentation graphique, l'origine est centre de gravité des points représentant les variables "i" affectés des masses f_i, comme des points représentant les variables "j" affectés des masses f_j.

$$* \qquad \sum_{i \in I} f_i \varphi_q^i \varphi_{q'}^i = 0 \quad \text{si} \quad q \neq q'$$

$$= 1 \quad \text{si} \quad q = q'$$

$$\sum_{j \in J} f_j \varphi_q^i \varphi_{q'}^i = 0 \quad \text{si} \quad q \neq q' \qquad (4)$$

$$= 1 \quad \text{si} \quad q = q'$$

* La formule de décomposition :

$$f_{ij} = f_i f_j \left[1 + \sum_{q \in Q} \sqrt{\lambda_q}\, \varphi_q^i\, \varphi_q^j \right] \tag{5}$$

associée aux conditions ci-dessus, permet de démontrer facilement :

* que toute reconstitution du tableau par un sous ensemble Q' de facteurs respectera les marges et l'effectif total du tableau : (on caractérise ce sous ensemble par le sous ensemble Q' des indices correspondants)

$$\sum_{i \in I} g_{ij} = \sum_{i \in I} f_{ij} \quad \text{et} \quad \sum_{j \in J} g_{ij} = \sum_{j \in J} f_{ij}$$

en notant :

$$g_{ij} = f_i f_j \left[1 + \sum_{q \in Q'} \sqrt{\lambda_q}\, \varphi_q^i\, \varphi_q^j \right] \tag{6}$$

* que l'inertie du tableau s'exprime comme somme de termes dépendant chacun d'un facteur :

$$\sum_{j \in J} \sum_{i \in I} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \sum_{q \in Q} \lambda_q \tag{7}$$

ceci permettra de tester l'hypothèse : "un sous ensemble Q' de facteurs est suffisant pour expliquer le tableau de départ".

* On pourra étudier l'apport d'une variable j (ou d'une variable i) au niveau de l'explication du tableau par un facteur :

$$\lambda_q\, f_j\, (\varphi_q^j)^2$$

Puisque

$$\lambda_q = \sum_{j \in J} \lambda_q\, f_j\, (\varphi_q^j)^2 \tag{8}$$

cette quantité pourra être comparée à λ_q (c'est ce que l'on appelle parfois la contribution absolue d'une variable à un facteur)

* D'autre part, on peut démontrer facilement :

$$\sum_{i \in I} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \sum_{q \in Q} \lambda_q\, f_j\, (\varphi_q^j)^2 \tag{9}$$

On pourra donc comparer l'apport d'une variable j au niveau d'un facteur q à l'apport de cette variable dans l'inertie du tableau (contribution relative).

II – ANALYSE D'UNE JUXTAPOSITION DE TABLEAUX DE CONTINGENCE

Nous allons voir si les propriétés précédentes de l'analyse des correspondances sont transposables à l'analyse d'une juxtaposition de tableaux de contingence, et si certaines notions fructueuses pour l'interprétation (part d'inertie expliquée, contributions absolues et relatives) ont des équivalents pour le type de tableau qui nous intéresse.

1/ Notation et exemple

I x J sera le tableau global, avec :

$$I = \bigcup_{l=1}^{L} I_l \qquad \text{et} \qquad J = \bigcup_{m=1}^{M} J_m$$

I x J sera juxtaposition de L. M tableaux de contingence sur les mêmes sujets. Nous appellerons encore n_{ij} l'effectif en (i , j) ; n_i et n_j seront

$$\sum_{j \in J} n_{ij} \qquad \text{et} \qquad \sum_{i \in I} n_{ij} \ ;$$

nous noterons K la somme des éléments du tableau.

Nous noterons :

$$f_{ij} = \frac{n_{ij}}{K} \ ; \ f_i = \sum_{j \in J} f_{ij} \ ; \ f_j = \sum_{i \in I} f_{ij}$$

$$f_j^i = \frac{f_{ij}}{f_i} \qquad \text{et} \qquad f_i^j = \frac{f_{ij}}{f_j}$$

Il faut remarquer que ces quantités, qui avaient un sens très simple quand il s'agissait d'un tableau de contingence (effectifs, fréquences), n'ont plus ici une signification évidente.

Pour reprendre les notations habituelles, nous parlerons de variable i ou j ; un ensemble I_1 ou J_m sera appelé groupe de variables bien qu'il représente en fait l'ensemble des modalités d'une variable qualitative m ou l.

Exemple de tableau de ce type

		J1 Age du chef de ménage				J2 Profession du chef de ménage								J3 Ancienneté dans le logement			
		1	2	3	4	1	2	3	4	5	6	7	8	1	2	3	4
I1 Existence d'un médecin de famille	oui	93	379	291	233	76	200	172	122	150	160	59	57	122	169	188	517
	non	145	226	114	77	31	125	105	59	55	99	46	42	160	131	93	178
I2 Existence d'un dentiste habituel	oui	128	417	296	209	79	262	199	130	129	133	52	66	156	192	201	501
	non	110	188	109	101	28	63	78	51	76	126	53	33	126	108	80	194

Le tableau précédent est la juxtaposition de 6 tableaux de contingence réalisés sur le même échantillon de 1 558 ménages. Ici L = 2, M = 3. L'intersection d'une ligne et d'une colonne représente un effectif. Par exemple, la case (3, 6) contient le nombre de ménages (262) qui à la fois :

– ont un dentiste habituel,

– appartiennent à la profession 2 (cadres supérieurs).

2/ Inertie du tableau global

L'analyse du tableau global sera basée sur la décomposition d'une inertie calculée formellement comme si I x J était un tableau de contingence, mais dont la signification ici reste à préciser.

L'inertie du tableau global s'écrit :

$$\sum_{i \in I} \sum_{j \in J} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{i \in I_l} \sum_{j \in J_m} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

On va montrer que la quantité :

$$\sum_{i \in I_l} \sum_{j \in J_m} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

est $\frac{1}{L \cdot M}$ fois l'inertie telle qu'elle serait calculée si on considérait le tableau $I_l \times J_m$ seul.

Il faut pour cela exprimer f'_i ; f'_j ; f'_{ij} ; fréquences calculées sur le tableau $I_l \times J_m$, et K_{lm}, somme des termes de ce tableau, en fonction des quantités f_i, f_j, f_{ij}, K :

K_{lm} est l'effectif du tableau $I_l \times J_m$. C'est le nombre de sujets concernés. Or ces sujets ont été comptés L. M. fois dans le tableau I x J, d'où :

$$K_{lm} = \frac{K}{L \cdot M}$$

$$f'_{ij} = \frac{n_{ij}}{K_{lm}} = L \cdot M \cdot \frac{n_{ij}}{K}$$

donc

* $$f'_{ij} = L.M.\ f_{ij}$$

f'_i est le pourcentage de sujets concernés par la modalité i.

Le nombre de sujets concernés par la modalité i est n_i/M ; en effet dans la ligne i de I x J, de somme n_i, ces personnes ont été comptées M fois. D'où :

$$f'_i = \frac{n_i}{M \cdot K_{lm}} = L \frac{n_i}{K}$$

* $$f'_i = L\ f_i$$

De même

$$* \qquad f'_j = M\, f_j$$

Il s'ensuit :

$$\frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \frac{1}{L\,.\,M}\,\frac{(f'_{ij} - f'_i f'_j)^2}{f'_i f'_j}$$

Soit le résultat :

$$\sum_{i\in I}\sum_{j\in J}\frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \frac{1}{L\,.\,M}\sum_{l\in L}\sum_{m\in M}\sum_{i\in I_l}\sum_{j\in J_m}\frac{(f'_{ij} - f'_i f'_j)^2}{f'_i f'_j}$$

L'inertie du tableau global est la moyenne des inerties des tableaux de contingence qui le composent.

Ce premier point est encourageant quant à l'étude par l'analyse des correspondances d'une juxtaposition de tableaux de contingence : la quantité qui sera décomposée par les facteurs –l'inertie globale– a aussi une signification précise pour notre tableau.

3/ Test sur la reconstitution du tableau global

Non seulement l'inertie du tableau global est la moyenne des inerties des tableaux qui le composent, mais encore, sous l'hypothèse d'indépendance entre les groupes de variables I_l et J_m , K fois cette quantité suit une loi du khi-2 à (I – L) (J – M) degrés de liberté.

K fois l'inertie s'écrit, d'après la formule précédente :

$$\sum_{l\in L}\sum_{m\in M}\sum_{i\in I_l}\sum_{j\in J_m}\frac{(K_{lm}\, f'_{ij} - K_{lm}\, f'_i f'_j)^2}{K_{lm}\, f'_i f'_j}$$

Soit la somme de I.J termes de la forme :

$$\frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}}$$

les effectifs observés étant liés aux effectifs théoriques par un certain nombre de relation linéaires indépendantes.

La propriété (7) des facteurs : expression de l'inertie du tableau global comme somme des valeurs propres est encore valable ici. On pourra donc tester, comme on le fait habituellement, et avec les mêmes réserves, l'apport d'un facteur dans l'explication du tableau global : on pourra se demander si, après explication par un ou plusieurs facteurs, K fois l'inertie restante peut être considérée comme la réalisation d'un Khi-2 à (I – L) (J – M) degrés de liberté.

On conçoit que ce test n'est pas suffisant, ni toujours très intéressant : le choix des variables qualitatives introduites dans le tableau peut être assez arbitraire, et on aimerait en savoir plus sur l'explication par les facteurs des relations pouvant exister au niveau de groupes de variables I_l ou J_m prises séparément.

4/ Moyenne des facteurs sur I_1 ou J_m

La plupart des propriétés vérifiées sur I ou J par les facteurs φ_q^I ou φ_q^J ne sont pas vérifiées sur I_1 ou J_m par des restrictions $\varphi_q^{I_1}$ ou $\varphi_q^{J_m}$ des facteurs à I_1 ou J_m. En particulier, les facteurs $\varphi_q'^{I_1}$ que l'on obtiendrait par l'analyse de $I_1 \times J$, ne sont pas des restrictions $\varphi_q^{I_1}$ des facteurs de l'analyse globale.

Parmi les propriétés (1) à (4) des facteurs, seule la propriété (3) a un équivalent pour les restrictions des facteurs :

Une fonction $\varphi_q^{I_1}$ *est de moyenne nulle pour la métrique* f_I

Une fonction $\varphi_q^{J_m}$ *est de moyenne nulle pour la métrique* f_J

soit :

$$\sum_{i \in I_1} \varphi_q^i f_i = 0$$

$$\sum_{j \in J_m} \varphi_q^j f_j = 0$$

(on aurait aussi bien pu remplacer f_i et f_j par f_i' et f_j' fréquences calculées sur $I_1 \times J_m$).

Démonstration (Pour une fonction $\varphi_q^{I_1}$)

Soit R_J l'espace des lois de probabilité sur J. Les éléments f_J^i sont des points de cet espace et φ_q^J, facteur de l'analyse des f_J^i, est un élément de R^J.

Comme élément de R^J, φ_q^J est combinaison linéaire des δ_j^J, fonctions qui à un point de R_J font correspondre la coordonnée d'ordre j.

Pour toutes les fonctions δ_j^J on a :

$$\sum_{i \in I_1} \delta_j^J (f_J^i - f_J) f_i = 0$$

Cette expression vaut en effet :

$$\sum_{i \in I_1} (f_{ij} - f_i f_j)$$

j appartient au groupe de variables J_m et dans le tableau de contingence $I_1 \times J_m$, on a :

$$\sum_{i \in I_1} (f_{ij}' - f_i' f_j') = 0$$

Ce qui s'écrit en exprimant f_{ij}', f_i', f_j' en fonction de f_{ij}, f_i, f_j :

$$\sum_{i \in I_1} (f_{ij} - f_i f_j) = 0$$

On a donc :

$$\sum_{i \in I_1} \delta_j^J (f_J^i - f_J) f_i = 0$$

qui entraîne :

$$\sum_{i \in I_1} \varphi_q^J (f_J^i - f_J) f_i = 0$$

Or $\varphi_q^J (f_J) = 0$ d'après la propriété (3) des facteurs

Donc

$$\sum_{i \in I_1} \varphi_q^J (f_J^i) f_i = 0$$

Ce qui s'écrit aussi :

$$\sum_{i \in I_1} \varphi_q^i f_i = 0$$

. D'après la propriété (2) des facteurs

Conséquence sur la représentation des résultats

Les variables sont en général représentées par des points de coordonnée

$$F(i) = \sqrt{\lambda q}\, \varphi_q^i \qquad \text{ou} \qquad G(j) = \sqrt{\lambda q}\, \varphi_q^j$$

La représentation de résultats d'analyse des correspondances sur une juxtaposition de tableaux de contingence possède donc la qualité suivante :

L'origine est centre de gravité pondéré non seulement de chaque ensemble de variables, I ou J, mais encore de chaque sous-ensemble I_1 ou J_m. Il revient au même de pondérer une variable par son poids dans l'ensemble du tableau, ou par son poids par rapport aux autres variables du groupe auquel elle appartient.

5/ Apport d'un groupe de variables I_1 ou J_m

On sait que dans un tableau de correspondances, on peut isoler l'apport d'une variable j (ou i) dans l'explication du tableau par un facteur.

$$\lambda_q\, f_j\, (\varphi_q^j)^2$$

Ici on isolera la quantité :

$$\sum_{j \in J_m} \lambda q\, f_j\, (\varphi_q^j)^2 \qquad \text{(ou son équivalent pour } I_1\text{)}$$

On pourra comparer cette quantité à λ_q, puisque la propriété (8) est vérifiée sur le tableau global :

$$\lambda q = \sum_{m=1}^{M} \sum_{j \in J_m} \lambda q\, f_j\, (\varphi_q^j)^2$$

On parlera de contribution absolue d'un groupe de variables à un facteur.

On pourra aussi comparer l'apport d'un groupe J_m de variables à l'expression :

$$\sum_{i \in I} \sum_{j \in J_m} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

et l'on parlera alors de contribution relative d'un facteur à un groupe de variables.

On a en effet (propriété (9) des facteurs)

$$\sum_{i \in I} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \sum_{q \in Q} \lambda_q \, f_j (\varphi_q^j)^2$$

Donc aussi :

$$\sum_{i \in I} \sum_{j \in J_m} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} = \sum_{q \in Q} \sum_{j \in J_m} \lambda_q \, f_j \, (\varphi_q^j)^2$$

La contribution relative présente l'avantage de pouvoir donner lieu à un test :

la quantité $$K \sum_{i \in I} \sum_{j \in J_m} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

est en effet, dans l'hypothèse d'indépendance entre J_m et chaque groupe de variables I_1, une variable qui suit une loi du Khi $-$ 2 à $(I - L)\,(J_m - 1)$ degrés de liberté. La démonstration est analogue à celle évoquée en II-3.

On pourra donc comparer la quantité :

$$K \sum_{i \in I} \sum_{j \in J_m} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j} - K \sum_{q \in Q'} \sum_{j \in J_m} \lambda_q \, f_i \, (\varphi_q^j)^2$$

(où Q' représente un sous-ensemble de facteurs)

à un Khi-2 à $(I - L)\,(J_m - 1)$ ddl

Ceci pourra permettre de juger de l'apport d'un sous-ensemble Q' de facteurs à l'explication des relations entre J_m et tous les groupes de variables I_1 prises dans leur ensemble.

Ce procédé présente des défauts analogues au test global sur $I \times J$: le choix arbitraire de groupes de variables I_1 peut enlever de la signification à l'inertie de $I \times J_m$; le test ne renseigne pas sur la façon dont un sous-ensemble Q' de facteurs explique chacun des tableaux $I_1 \times J_m$. De plus, on sait qu'en analyse de correspondances, la comparaison de l'"inertie restante" à un Khi--2 dont le nombre de ddl est celui du tableau de départ, prête à la critique.

Néanmoins, les tests au niveau de groupes de variables I_1 ou J_m pourront préciser la structure du tableau :

Des groupes de variables J_m et $J_{m'}$ expliqués par des facteurs différents pourront être considérés comme dissemblables vis-à-vis des variables I.

Un tableau $I \times J_m$ pourra rester inexpliqué par les premiers facteurs, alors même que ces facteurs expliquent bien le tableau $I \times J$ dans son ensemble.

6/ Reconstitution d'un tableau $I_1 \times J_m$ par un ou plusieurs facteurs

Venons-en maintenant à un tableau de contingence $I_1 \times J_m$ et démontrons la propriété suivante :

Toute reconstitution de $I \times J$ par un sous-ensemble Q' de facteurs conserve les marges des tableaux $I_1 \times J_m$

Démonstration

soit
$$g_{ij} = f_i f_j \left(1 + \sum_{q \in Q'} \sqrt{\lambda_q}\, \varphi_q^i\, \varphi_q^j\right)$$

et
$$f_{ij} = f_i f_j \left(1 + \sum_{q \in Q} \sqrt{\lambda_q}\, \varphi_q^i\, \varphi_q^j\right)$$

(formule de décomposition)

On aura :

$$\sum_{i \in I_1} g_{ij} = \sum_{i \in I_1} f_{ij} = \sum_{i \in I_1} f_i f_j$$

car la sommation sur I_1 de termes de la forme :

$$f_i f_j \sqrt{\lambda_q}\, \varphi_q^i\, \varphi_q^j$$

est nulle puisque φ_q^I restreinte à I_1 est de moyenne nulle.

On a évidemment un résultat analogue pour une sommation sur J_m.

Ainsi la reconstitution par un ou plusieurs facteurs redonne des tableaux $I_1 \times J_m$ qui ne diffèrent des tableaux de départ que parce qu'ils expriment des relations différentes entre I_1 et J_m, avec les mêmes effectifs marginaux, et en conservant évidemment le nombre de sujets initial. Ceci nous permet de comparer un tableau reconstitué au tableau de départ correspondant.

Nous pouvons considérer le tableau de départ $(I_1 \times J_m)$ comme notre hypothèse nulle, tableau de référence indiquant un certain type de relation entre I_1 et J_m, éloigné de l'indépendance : la reconstitution d'un tableau qui n'exprimerait pas de relation entre I_1 et J_m ne nous intéresse pas.

Nous calculerons alors :

$$K \sum_{i \in I_1} \sum_{j \in J_m} \frac{(g_{ij} - f_{ij})^2}{f_{ij}}$$

g_{ij} étant le résultat de la reconstitution par un sous ensemble Q' de facteurs et comparerons cette quantité à un Khi-2 à $(I_1 - 1)(J_m - 1)$ ddl

Ceci permet d'apprécier, pour chaque tableau $(I_1 \times J_m)$, sa reconstitution par un ou plusieurs facteurs. On pourra baser l'interprétation des facteurs sur les tableaux qu'ils expliquent le mieux. On pourra parler de tableaux "semblables" s'ils sont bien reconstitués par les mêmes facteurs ou groupes de facteurs, et isoler les tableaux mal expliqués par les premiers facteurs de l'analyse. (Le procédé ne se justifie pas si l'effectif est trop faible, le tableau $I_l \times J_m$ pouvant être difficilement considéré comme tableau de référence).

Il faut noter que, l'inertie de $(I_1 \times J_m)$ n'étant pas décomposable en une somme de termes positifs dépendant chacun d'un facteur, il peut se faire que l'introduction d'un facteur supplémentaire n'améloiore pas la reconstitution du tableau.

Remarques sur les tests proposés

Les tests visant à juger de l'apport des facteurs à l'explication du tableau prètent tous à la même critique : l'hypothèse nulle qu'ils testent s'exprime en fonction des facteurs (sous la forme : tel sous-ensemble de facteurs est suffisant pour expliquer les relations observées), les facteurs étant eux-mêmes directement construits à partir des observations. Or une hypothèse nulle devrait toujours être définie avant examen des données.

Dans l'explication d'un sous-ensemble du tableau, du type $I_1 \times J$, ou $I_1 \times J_m$, on peut nuancer la critique : les facteurs peuvent être assez peu dépendants du sous-ensemble considéré ; il peut se faire qu'une analyse refaite en enlevant les sous-ensemble I_1 ou J_m donnerait des facteurs ayant à peu près même signification. C'est ce qui se passe quand on analyse un tableau de grandes dimensions avec des variables redondantes.

Dans tous les cas, il convient de donner aux seuils de signification observés une valeur plutôt indicatrice de distance entre les données et certaines hypothèses.

BIBLIOGRAPHIE

(1) BENZECRI (J.P.) – L'analyse des données (II : l'analyse des correspondances) Dunod, Paris, 1973.

(2) LEBART (L.), FENELON (J.P.) – Statistique et informatique appliquées Dunod, Paris, 1973.

(3) LECLERC (A.) – Etude de certains types de tableaux par l'analyse des correspondances Thèse 3ème cycle, Université Paris VI, 1973.

(4) DAVIDSON (F.), CHOQUET (M.), DEPAGNE (M.) – Les lycéens devant la drogue et les autres produits psychotropes, monographie de l'INSERM. 1974
(exemple d'application)