

A. EL MOUSSAOUI

Sur la reconstitution approchée d'un tableau de correspondance à partir du tableau cumulé par blocs suivant deux partitions des ensembles I et J

Les cahiers de l'analyse des données, tome 12, n° 3 (1987), p. 365-370

http://www.numdam.org/item?id=CAD_1987__12_3_365_0

© Les cahiers de l'analyse des données, Dunod, 1987, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

SUR LA RECONSTITUTION APPROCHÉE D'UN TABLEAU DE CORRESPONDANCE À PARTIR DU TABLEAU CUMULÉ PAR BLOCS SUIVANT DEUX PARTITIONS DES ENSEMBLES I ET J

[REC. COR. BLOCS]

A. EL MOUSSAOUI ()*

1 . Reconstitution et approximation

1.1 . Principe: vision d'un tableau au travers de deux partitions

Partant d'un tableau de correspondance sur $I \times J$, il est usuel d'effectuer des classifications ascendantes hiérarchiques sur les deux ensembles I et J. De chacune de ces CAH on retient après interprétation des partitions respectivement $C(I)$ et $C(J)$. Dans une certaine mesure la réalité dont rend compte le tableau initial k_{IJ} , est maintenant saisie au travers de ces partitions, en ce sens qu'on assimile deux individus i et i' (ou j et j') qui appartiennent à une même classe; en gardant présents à l'esprit les masses respectives individuelles des i et j , mais assimilant les affinités qui peuvent exister entre individus, aux affinités entre leurs classes respectives $C(i)$ et $C(j)$. En termes mathématiques, cela revient à réduire un tableau $m \times n$ ($m = \text{card } I$; $n = \text{card } J$) à deux vecteurs de longueur m et n (les $k(i)$ et les $k(j)$) et un tableau rectangulaire $p \times q$ (où $p = \text{Card } C(I)$ et $q = \text{card } C(J)$ sont les nombres de classes des partitions retenues). La reconstitution approchée du tableau $k(i,j)$ se faisant par une formule précise fondée sur l'hypothèse de l'indépendance intrabloc entre i et j .

1.2. La formule de reconstitution

On a le tableau de correspondance cumulé sur le produit $C(I) \times C(J)$ des deux partitions:

(*) Docteur en Statistique, Laboratoire de Statistique, Université Pierre et Marie Curie.

$$\forall c \in C(I), \forall g \in C(J) : k(c,g) = \Sigma\{k(i,j) \mid i \in c ; j \in g\}.$$

Dans la reconstitution on pose:

$$kr(i,j) = k(i) k(j)(k(c(i),g(j))/k(c(i))(g(j))) ;$$

où $c(i)$ est la classe de la partition $C(I)$, contenant i , et de même $g(j)$ la classe de la partition $C(J)$ contenant j .

En sommant par rapport à j (resp i) on constate que les deux tableaux kr_{IJ} et k_{IJ} ont même marge sur I (resp J). Par exemple on a:

$$\begin{aligned} kr(i) &= \Sigma\{kr(i,j) \mid j \in J\} \\ &= k(i) \Sigma\{k(j) k(c(i), c(j))/k(g(j)) \mid j \in J\}/(c(i)). \end{aligned}$$

Dans cette formule la somme Σ intermédiaire peut se calculer en utilisant la partition de J en classes $g \in C(J)$. Il vient:

$$\begin{aligned} &\Sigma\{\Sigma\{k(j) k(c(i), g)/k(g) \mid j \in g\} \mid g \in C(J)\} \\ &= \Sigma\{k(g) k(c(i), g)/k(g) \mid g \in C(J)\} \\ &= \Sigma\{k(c(i), g) \mid g \in C(J)\} = k(c(i)); \end{aligned}$$

en revenant à l'expression de $kr(i)$ on trouve bien:

$$kr(i) = k(i) * k(c(i))/k(c(i)) = k(i).$$

La technique des sommations suivant les classes des partitions considérées nous servira constamment dans la présente note.

1.3. La notion d'indépendance par blocs

Puisque les 3 tableaux k_{IJ} , $k_{C(I)C(J)}$, kr_{IJ} ont le même total k , la formule de reconstitution peut s'écrire en termes de loi de fréquence f ou fr .

$$fr_{ij} = (f_i/f_{C(i)})(f_j/f_{C(j)})f_{C(i)C(j)};$$

on reconnaît dans cette formule que les quotients $(f_i/f_{C(i)})$ (resp $f_j/f_{C(j)}$) représentent la part de i (resp j) dans la classe $C(i)$ (resp $C(j)$) à laquelle il appartient; le coefficient $f_{C(i)C(j)}$ étant la masse du bloc $C(i) \times C(j)$ du tableau $I \times J$, où se trouve la paire (i,j) . On voit bien qu'au sein du bloc on a une loi produit d'où le terme d'indépendance par bloc utilisé dans le titre de la précédente note.

2 . Comparaison entre tableau initial et tableau reconstitué

2.1 . Principe de la comparaison

Une fois le tableau reconstitué, il est impératif de le comparer au tableau initial, pour voir dans quelle mesure il est proche de celui-ci. On utilise pour cela la métrique du χ^2 de centre $f_I f_J$ (produit des marges, communes aux deux tableaux). On pose donc:

$$\|fr_{IJ} - f_{IJ}\|^2 = \sum\{(fr_{ij} - f_{ij})^2 / (f_i f_j) \mid i \in I; j \in J\};$$

Pour apprécier l'erreur de reconstitution, il faut ensuite rapporter l'écart (au carré) entre les deux distributions à celui de f_{IJ} à l'indépendance soit:

$$E = \|fr_{IJ} - f_{IJ}\|^2 / \|f_{IJ} - f_I f_J\|^2.$$

Dans la suite nous montrerons que l'on a pour E une autre formule qui a l'intérêt de calculer la proximité entre le tableau des fréquences "initiales" f_{IJ} et celui des fréquences "reconstituées" fr_{IJ} sans se référer à ce dernier.

$$E = 1 - (\|f_{C(I),C(J)} - f_{C(I)} f_{C(J)}\|^2 / \|f_{IJ} - f_I f_J\|^2);$$

Le dénominateur de la fraction qui intervient dans cette formule n'est rien d'autre que l'inertie totale du nuage $N_J(I) = \{(f_j^i, f_i) \mid i \in I\}$; inertie qui elle-même est égale à la somme des valeurs propres issues de l'analyse factorielle du tableau k_{IJ} . Il en est de même pour le numérateur $\|f_{C(I),C(J)} - f_{C(I)} f_{C(J)}\|^2$ relativement à la correspondance cumulée $k_{C(I),C(J)}$. Ainsi la quantité E désigne la perte relative d'inertie subie en substituant au nuage $N_J(I)$, le nuage $N_{C(J)}(C(I))$; (ou en substituant $N_{C(I)}(C(J))$ à $N_I(J)$).

Il est clair que plus $C(I)$ et/ou $C(J)$ sont fines plus la reconstitution est "pertinente" en ce sens que les deux tableaux de correspondance $K_{r_{IJ}}$ et K_{IJ} (ou les deux lois conjointes fr_{IJ} et f_{IJ} qui leurs sont afférentes) sont distributionnellement proches au sens du chi-deux de centre $f_I f_J$.

Ce résultat peut être démontré rigoureusement, en considérant qu'à des suites de partitions de plus en plus grossières, correspond une suite de tableaux reconstitués de k_{IJ} , dont chacun peut être considéré comme le reconstitué du précédent; la reconstitution se faisant chaque fois avec une perte d'inertie. Dans la cas limite où:

$$C(I) = \{\{i\}; i \in I\} \text{ et } C(J) = \{\{j\}; j \in J\}$$

on a pour tout couple $(i,j) \in IJ$:

$$C(i) = \{i\} \text{ et } C(j) = \{j\}; kr(i,j) = k(i,j);$$

le tableau initial et le tableau reconstitué coïncident, et il n'y a pas de perte d'inertie!

Reste à démontrer la formule annoncée pour l'écart \bar{E} .

2.2. Calcul de l'écart entre tableau initial et tableau reconstitué

Nous effectuerons ce calcul en tenant compte de la bilinéarité du produit scalaire, au sein de l'espace euclidien des mesures sur $I \times J$, muni de la métrique du χ^2 de centre $f_I f_J$. On a:

$$\begin{aligned} \|fr_{IJ} - f_I f_J\|^2 &= \|(fr_{IJ} - f_I f_J) - (f_{IJ} - f_I f_J)\|^2 \\ &= \|fr_{IJ} - f_I f_J\|^2 + \|f_{IJ} - f_I f_J\|^2 - 2 \langle (fr_{IJ} - f_I f_J), (f_{IJ} - f_I f_J) \rangle ; \end{aligned}$$

Dans cette formule, le deuxième terme carré $\|f_{IJ} - f_I f_J\|^2$, n'est autre que la trace de la correspondance donnée: il n'y a rien de plus à en dire. En revanche nous transformerons le premier terme carré, ainsi que le produit scalaire.

2.2.1. Calcul de $\|fr_{IJ} - f_I f_J\|^2$:

On a classiquement:

$$\begin{aligned} \|fr_{IJ} - f_I f_J\|^2 &= \text{trace}(fr_J^I \circ fr_I^J) - 1 ; \\ \text{trace}(fr_J^I \circ fr_I^J) &= \sum\{fr_j^i fr_i^j \mid i \in I; j \in J\} \\ &= \sum\{(fr_{ij})^2 / (fr_i fr_j) \mid i \in I; j \in J\}. \end{aligned}$$

En remplaçant fr_{ij} par sa définition donnée au § 1, et tenant compte de ce que $fr_i = f_i$, $fr_j = f_j$, on obtient:

$$\text{trace}(fr_J^I \circ fr_I^J) = \sum\{f_i f_j (f^2_{C(i);C(j)}) / (f^2_{C(i)} f^2_{C(j)}) \mid i \in I, j \in J\};$$

Le second membre de cette égalité peut être transformé en considérant d'abord la somme partielle par rapport à j .

$$\begin{aligned} & \Sigma\{f_i f_j f^2_{C(i);C(j)} / (f^2_{C(i)} f^2_{C(j)}) \mid j \in J\} \\ &= (f_i / f^2_{C(i)}) \Sigma_{g \in C(J)} \Sigma_{j \in g} \{f_j f^2_{C(i);g} / f^2_g\} \\ &= (f_i / f^2_{C(i)}) \Sigma_{g \in C(J)} \{f^2_{C(i);g} / f_g\}; \end{aligned}$$

(où on a tenu compte de ce que $\Sigma\{f_j \mid j \in g\} = f_g$).

Sachant que $I = \cup \{c \mid c \in C(I)\}$, on a pour trace $(fr_I^I \circ fr_I^J)$ l'évaluation:

$$\begin{aligned} & \Sigma_{c \in C(I)} \Sigma_{i \in c} \{ (f_i / f^2_c) \Sigma_{g \in C(J)} \{ f^2_{cg} / f_g \} \} \\ &= \Sigma_{g \in C(J)} (1 / f_g) \Sigma_{c \in C(I)} \{ f^2_{cg} / f_c \}; \end{aligned}$$

(où on a tenu compte de ce que $\Sigma\{f_i \mid i \in c\} = f_c$); et finalement, pour la trace de $(fr_I^I \circ fr_I^J)$:

$$\Sigma\{f^2_{cg} / (f_c f_g) \mid c \in C(I); g \in C(J)\};$$

expression qui se réfère à la correspondance cumulée. D'où il résulte que l'on a d'après déduction de 1)

$$\|fr_{IJ} - f_I f_J\|^2 = \|f_{C(I)C(J)} - f_{C(I)} f_{C(J)}\|^2.$$

L'inertie de la correspondance reconstituée n'est autre que celle de la correspondance cumulée.

2.2.2. Calcul du terme produit scalaire

Celui-ci peut être décomposé en quatre termes:

$$\begin{aligned} & \langle fr_{IJ} - f_I f_J, f_{IJ} - f_I f_J \rangle = \\ & \langle fr_{IJ}, f_{IJ} \rangle + \|f_I f_J\|^2 - \langle fr_{IJ}, f_I f_J \rangle - \langle f_{IJ}, f_I f_J \rangle; \end{aligned}$$

Les derniers termes de cette somme valent 1: on peut le vérifier pour l'un d'entre-eux:

$$\langle f_{IJ}, f_I f_J \rangle = \Sigma_{ij} \{ f_{ij} f_i f_j / (f_i f_j) \} = \Sigma_{ij} \{ f_{ij} \} = 1.$$

Pour le 1-er terme, on procédera comme au § 2.2.1, par sommations partielles suivant les classes de partitions $C(I)$ et $C(J)$.

$$\begin{aligned}
\langle f_{IJ}, f_{IJ} \rangle &= \Sigma \{ f_{ij} f_{ij} / (f_i f_j) \mid i \in I; j \in J \} \\
&= \Sigma \{ f_{ij} f_{C(i)C(j)} / (f_{C(i)} f_{C(j)}) \mid i \in I; j \in J \} \\
&= \Sigma \{ \Sigma \{ f_{ij} f_{C(i)C(j)} / (f_{C(i)} f_{C(j)}) \mid i \in c; j \in g \} \mid c \in C(I); g \in C(J) \} \\
&= \Sigma \{ (f_{cg})^2 / (f_c f_g) \mid c \in C(I); g \in C(J) \} \\
&= \text{trace}(f_{C(J)}^{C(I)} \circ f_{C(I)}^{C(J)}).
\end{aligned}$$

En rassemblant tous les termes que nous avons successivement calculés on a :

$$\begin{aligned}
\langle f_{IJ} - f_{I^c J}, f_{IJ} - f_{I^c J} \rangle &= \text{trace}(f_{C(J)}^{C(I)} \circ f_{C(I)}^{C(J)}) - 1 \\
&= \|f_{C(I)C(J)} - f_{C(I)} f_{C(J)}\|^2;
\end{aligned}$$

$$\|f_{IJ} - f_{I^c J}\|^2 = \|f_{IJ} - f_{I^c J}\|^2 - \|f_{C(I)C(J)} - f_{C(I)} f_{C(J)}\|^2;$$

formule annoncée au § 2.2.1.

2.3. Remarque : interprétation en terme de projecteurs

Les résultats du § 2 résultent immédiatement de ce que le tableau f_{IJ} reconstitué à partir de $f_{I^c J}$ pour satisfaire à la condition d'indépendance intrabloc relativement aux partitions $C(I)$ et $C(J)$, n'est autre que la projection orthogonale de f_{IJ} sur le sous-espace des lois satisfaisant à cette condition, (cette projection étant comprise au sens de la métrique du χ^2 de centre $f_i f_j$ sur l'espace des mesures sur IJ). C'est ce que nous avons démontré ailleurs [6]. Mais dans la présente note, nous avons voulu donner une démonstration qui en dépit de la longueur des calculs soit accessible aux lecteurs peu habitués aux projections orthogonales en géométrie euclidienne multidimensionnelle

Références

- [1] BENZECRI J. P.: [AJUST. MARGES INTER.], CAD Vol VIII, 1983, n° 2, pp. 227-233.
- [2] BENZECRI J. P.: [GEN. AJUS. MARGES] CAD Vol VIII, 1983, pp. 359-370.
- [3] BENZECRI J. P., BOURGARIT Ch., MADRE J. L.: [AJUS. MARGES FAC.] CAD, Vol V, 1980, n°1, pp. 87-99.
- [4] CHOLAKIAN V.: [AJUST. MARGES CRIT.] CAD Vol IX, 1984, n°1, pp. 113-117.
- [5] CHOLAKIAN V.: [EXEMP. AJUST. MARGES], CAD Vol V, 1980, n° 2, pp. 173-176.
- [6] EL MOUSSAOUI A. [CAH. VAR. CUM.], CAD Vol XII, 1987, n° 2, pp. 237-242 .