

V. CHOLAKIAN

Méthodes et critères pour l'ajustement d'un tableau à des marges imposées

Les cahiers de l'analyse des données, tome 9, n° 1 (1984),
p. 113-117

http://www.numdam.org/item?id=CAD_1984__9_1_113_0

© Les cahiers de l'analyse des données, Dunod, 1984, 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/>

MÉTHODES ET CRITÈRES POUR L'AJUSTEMENT D'UN TABLEAU A DES MARGES IMPOSÉES

[AJUST. MARGE CRIT.]

par V. Cholakian

0 Le présent article fait suite aux travaux publiés dans cette revue (cf. Références). Dans le § 1, on rappelle le problème en citant quelques résultats. Au § 2, on démontre que la méthode d'ajustement d'après la formule de reconstitution de l'a. des correspondances, équivaut à minimiser une certaine distance quadratique. Au § 3 on reprend le problème de l'ajustement suivant la distance du χ^2 , en tirant parti des propriétés mises en évidence par l'analyse des correspondances.

1 Rappel des problèmes de l'ajustement

Nous reprenons les notations usuelles en analyse des correspondances pour une mesure f_{IJ} (ou g_{IJ}) sur un produit de deux ensembles finis I et J et ses marges f_I, f_J (ou g_I, g_J). De plus nous notons f^{IJ} (ou g^{IJ}) la fonction densité par rapport au produit des marges :

$$f^{ij} = f_{ij}/(f_i f_j) ; g^{ij} = g_{ij}/(g_i g_j) .$$

Le problème de l'ajustement est le suivant : étant donnés deux ensembles finis I et J, une loi de probabilité f_{IJ} sur $I \times J$, et deux lois g_I, g_J sur I et J, trouver une loi g_{IJ} ayant pour marges g_I et g_J et "ressemblant le plus possible" à f_{IJ} .

Ici, ce qui nous intéresse est le critère de distance susceptible de donner un sens précis à cette ressemblance maxima. Dans le § 2 de son article J.L. Madre [6], introduit ou rappelle plusieurs formules de distance :

a) minimiser la distance :

$$d^2(f_{IJ}, g_{IJ}) = \sum \{(f_{ij} - g_{ij})^2 / (b_i c_j) \mid i \in I, j \in J\}$$

(où les b_i, c_j sont des coefficients qui peuvent dépendre des données f_i, f_j, g_i, g_j) conduit au résultat :

$$g_{ij} = f_{ij} + ((g_i - f_i) c_j / c_t) + ((g_j - f_j) b_i / b_t), \text{ où } b_t = \sum_i \{b_i\}; \\ c_t = \sum_j \{c_j\}. \text{ En particulier, lorsque } b_i = (f_i + g_i) / 2, \text{ il vient :}$$

$$g_{ij} = f_{ij} + g_i g_j - f_i f_j .$$

Le même résultat est obtenu par J.P. Benzécri dans [1], par une autre voie.

(1) Docteur 3° cycle en statistique.

b) minimiser le critère d'entropie relative :

$$H(g_{IJ}; f_{IJ}) = \sum \{g_{ij} \log(g_{ij}/f_{ij}) \mid i \in I, j \in J\},$$

conduit à l'équation suivante que résout l'algorithme R.A.S. :

$$(1) \forall i, j : \log(g_{ij}/f_{ij}) = \alpha^i + \beta^j$$

où les α, β sont des coefficients à déterminer (pour une interprétation géométrique, cf. [6], § 2.3).

c) une autre solution est donnée dans [2], en partant de la formule de reconstitution, mais sans proposer de critère.

$$(2) g^{ij} = 1 + \sum_{\alpha} \{\lambda_{\alpha}^{-1/2} (F_{\alpha}(i) - A_{\alpha})(G_{\alpha}(j) - B_{\alpha})\}; \text{ où}$$

$$A_{\alpha} = \sum_i \{g_i F_{\alpha}(i)\}; B_{\alpha} = \sum_j \{g_j G_{\alpha}(j)\},$$

et où les λ, F, G sont les v.p. et facteurs issus de l'analyse de f_{IJ} .

2 Formule de reconstitution modifiée et critère de distance entre densités de f et g

On démontrera qu'on obtient (2) en minimisant :

$$D(f_{IJ}, g_{IJ}) = \sum \{g_i g_j (f^{ij} - g^{ij})^2 \mid i \in I, j \in J\},$$

sous les contraintes de marges (g_I, g_J) imposées à g_{IJ} . En effet on a :

$$\partial D / \partial g_{ij} = 2(g^{ij} - f^{ij}).$$

Les équations constantes sont :

$$(3) \begin{aligned} \forall i : \sum \{g_{ij} \mid j \in J\} &= g_i \Rightarrow \sum_j dg_{ij} = 0; \\ \forall j : \sum \{g_{ij} \mid i \in I\} &= g_j \Rightarrow \sum_i dg_{ij} = 0; \end{aligned}$$

la condition $dD = 0$, doit être équivalente à une combinaison linéaire de ces contraintes, avec des coefficients μ^i, μ^j convenables ; on doit donc avoir :

$$(4) \forall i, j : g^{ij} - f^{ij} = \mu^i + \mu^j;$$

En sommant l'équation (4) d'abord par rapport à i et j en pondérant par $g_i g_j$; puis par rapport à une variable seulement, il vient après simplifications :

$$(5) \forall \sum_i \{g_i \mu^i\} + \sum_j \{g_j \mu^j\} = 1 - \sum_{ij} \{g_i g_j f^{ij}\} = 1 - \beta$$

$$(6) \sum_i \{g_i \mu^i\} + \mu^j = 1 - \sum_i \{g_i f^{ij}\} = 1 - \beta^j;$$

$$(7) \sum_j \{g_j \mu^j\} + \mu^i = 1 - \sum_j \{g_j f^{ij}\} = 1 - \beta^i.$$

Où il est possible de récrire les définitions de β, β^i, β^j dans les notations usuelles du calcul des transitions :

$$\beta = f^{IJ} \circ (g_I g_J); (\beta^I = (f_I^J \circ g_J / f_I)^I; \beta^J = (f_J^I \circ g_I / f_J)^J.$$

En additionnant les équations (6) et (7) et retranchant (5) il vient :

$$(8) \mu^i + \mu^j = 1 - (\beta^i + \beta^j) + \beta = g^{ij} - f^{ij} ;$$

reste à démontrer l'équivalence entre (2) et (8). La formule de reconstitution s'écrit :

$$f^{ij} = 1 + \sum_{\alpha} \lambda_{\alpha}^{-1/2} F_{\alpha}(i) G_{\alpha}(j) ;$$

la formule (2) peut être réécrite :

$$g^{ij} = 1 + \sum_{\alpha} \lambda_{\alpha}^{-1/2} (F_{\alpha}(i) - \sum_{i'} g_{i'} F_{\alpha}(i')) (G_{\alpha}(j) - \sum_{j'} g_{j'} G_{\alpha}(j')) ;$$

à l'intérieur des sommations en i' , j' que comporte cette formule, on voit apparaître les termes de (8). En effet, e.g. :

$$\begin{aligned} & \sum_{\alpha} \lambda_{\alpha}^{-1/2} (\sum_{i'} g_{i'} F_{\alpha}(i')) G_{\alpha}(j) \\ &= \sum_{i'} g_{i'} \sum_{\alpha} \{ \lambda_{\alpha}^{-1/2} (F_{\alpha}(i')) G_{\alpha}(j) \} \\ &= \sum_{i'} g_{i'} (f^{i'j} - 1) = -1 + \sum_{i'} g_{i'} f^{i'j} = \beta^j - 1 ; \end{aligned}$$

de même la sommation en i' , j' donne $(1 - \beta)$. D'où finalement pour g^{ij} :

$$g^{ij} = f^{ij} + 1 - (\beta^i + \beta^j) + \beta,$$

qui n'est autre que la formule (8) ci-dessus.

Ce résultat peut d'ailleurs être affirmé sans faire dans le détail la vérification de l'égalité ; car il est clair que d'une part la formule (2) donne pour $g^{ij} - f^{ij}$ une expression en somme de deux fonctions l'une de i l'autre de j et que d'autre part cette formule satisfait aux contraintes marginales ; ce qui nous ramène à (8).

3 Ajustement suivant la distance du χ^2 de centre f_{IJ}

On considère le critère :

$$S = \sum \{ (f_{ij} - g_{ij})^2 / f_{ij} \mid i \in I, j \in J \} = \|f_{IJ} - g_{IJ}\|_{f_{IJ}}^2.$$

pour que ce critère soit applicable il faut que les f_{ij} soient strictement positifs quels que soient i et j . Le critère S a été considéré par Deming et Stéphan [4] : et ce sont ces deux auteurs qui ont pour la première fois proposé l'algorithme R.A.S. en supposant (indûment : cf. § 1.b) que RAS minimise S . On a en fait :

$$\partial S / \partial g_{ij} = 2(g_{ij} - f_{ij}) / f_{ij},$$

avec les équations de contrainte (3) du § 2 ; et comme précédemment $dS = 0$ doit résulter des équations (3) ; d'où ici, $\forall i, j$:

$$(9) (g_{ij} / f_{ij}) = 1 + \mu^i + \mu^j ;$$

où on remarquera l'analogie avec (1). En sommant (9) par rapport à i ou j en pondérant par f_{ij} , on a le système :

$$(10) \begin{aligned} \forall i \in I : f_i \mu^i + \sum_j \{ f_{ij} \mu^j \} &= g_i - f_i \\ \forall j \in J : f_j \mu^j + \sum_i \{ f_{ij} \mu^i \} &= g_j - f_j. \end{aligned}$$

On peut reporter dans les équations indicées par j (2-ème ligne de (10)) les valeurs de μ^i issues des équations indicées par i (1-ère ligne) ; il vient :

$$(11) \quad f_j \mu^j = \sum_{ij} \{f_{ij} f_{ij}, \mu^{j'} / f_i\} + g_j - \sum_i \{f_{ij} g_i / f_i\}.$$

Le système de ces équations (que Deming et Stéphan ont proposé de résoudre par des méthodes classiques) prend une forme particulièrement claire si on divise chaque équation par f_j et utilise les notations classiques du calcul des transitions ; il vient :

$$(12) \quad \mu^J = \mu^J \circ f_J^I \circ f_I^J + F^J$$

où $F^J = (g_j / f_j) - \sum_i \{f^{ij} g_i\}$; c'est-à-dire :

$$\mu^J \circ (1 - f_J^I \circ f_I^J) = F^J ; \mu^J = F^J \circ (1 - f_J^I \circ f_I^J)^{-1}$$

Or il est bien connu en analyse des correspondances que la transition composée $f_J^I \circ f_I^J$ a toutes ses valeurs propres comprises entre 0 et 1 ; il en est donc de même de $(1 - f \circ f)$: ce qui permet de résoudre l'équation proposée. Plus précisément, à condition que le tableau f_{IJ} ne soit pas décomposé en blocs diagonaux, la transition $f \circ f$ n'admet comme vecteur propre relatif à la v.p. 1, que le facteur trivial $\delta^J = 1$; par conséquent l'application :

$$\mu^J \circ (1 - f_J^I \circ f_I^J)$$

est une surjective sur l'ensemble des fonctions de moyenne nulle pour la loi f_J , fonctions qui sont les combinaisons linéaires des facteurs relatifs à une v.p. éventuellement nulle mais non égale à 1. Or on vérifie que F^J est de moyenne nulle pour la loi f_J . On pourra, si on le désire, obtenir itérativement une solution approchée :

$$\mu^J = F^J + F^J \circ f_J^I \circ f_I^J + \dots + F^J \circ (f_J^I \circ f_I^J)^{n+1} \dots$$

(suivant la formule usuelle $(1 - t)^{-1} = 1 + t + \dots + t^n$; qui s'applique ici parce que dans le sous-espace des fonctions de moyenne nulle, $f \circ f$ a toutes ses valeurs propres strictement inférieures à 1).

Remarque : On sait (cf. [INF. TAB.] TI B n° 5), que pour g_{IJ} voisin de f_{IJ} (c'est-à-dire ici si $g_I \approx f_I$ et $g_J \approx f_J$) on a l'équivalence entre les critères du § 1.b et du § 3 :

$$2H(g_{IJ}; f_{IJ}) \approx \|f_{IJ} - g_{IJ}\|_{f_{IJ}}^2.$$

Ce qui a attiré notre attention, c'est l'analogie entre les équations (9) et (1) qui peuvent s'écrire respectivement selon les modèles additifs et multiplicatifs.

$$(g_{ij} / f_{ij}) = 1 + \mu^i + \mu^j ; (g_{ij} / f_{ij}) = M^i M^j.$$

Et nous nous proposons d'utiliser ces modèles pour l'étude des interactions multiples,

REFERENCES

- [1] BENZECRI J.P. : [AJUST. MARGES INTER.], CAD Vol VIII, 1983, n°2, pp 227-233.
- [2] BENZECRI J.P., BOURGARIT Ch., MADRE J.L. : [AJUS. MARGES FAC.], CAD Vol V, 1980, n° 1, pp 87-99.
- [3] CHOLAKIAN V. : [EXEMP. AJUST. MARGES], CAD Vol V, 1980, n° 2, pp 173-176.
- [4] DEMING W.E. and STEPHEN F.F. : "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known", Ann. Math. Stat., II, 1940, pp 427-444.
- [5] FADDEEVA V.N. : Computational Methods of Linear Algebra, 1959, Dover publication.
- [6] MADRE J.L. : [METH. AJUST. MARGES], CAD Vol V, 1980, n° 1, pp 87-99.