

J. L. MADRE

Méthodes d'ajustement d'un tableau à des marges

Les cahiers de l'analyse des données, tome 5, n° 1 (1980),
p. 87-99

http://www.numdam.org/item?id=CAD_1980__5_1_87_0

© Les cahiers de l'analyse des données, Dunod, 1980, 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 D'AJUSTEMENT
D'UN TABLEAU A DES MARGES
[MÉTH. AJUS. MARGES]

par J. L. Madre (1)

1 Le problème et ses variantes : Au départ (§ 1.1) l'énoncé du problème est simple, mais imprécis. On peut le préciser selon divers critères (§§ 1.2 & 1.3) ; qui aboutissent à des méthodes (§§ 2, 3, 4) ; dont il faudra se demander si elles possèdent toutes les propriétés souhaitables, notamment si l'on doit construire un tableau de nombres positifs (§ 1.4).

Pour exposer les résultats acquis depuis quelques années déjà, on a utilisé les travaux publiés dans les Annales de l'INSEE (n°s 22-23 ; 1976 : la table de ce n° double est parue dans les Cahiers Vol II n° 1). Mais notre but est ici d'introduire les méthodes d'analyse factorielle (§ 3), qui font d'autre part l'objet d'un article séparé (cf [AJUS. MARGES FAC.]*) ; et dont l'intérêt apparaît surtout dans un problème voisin de celui traité ici, l'extrapolation de tableau, dont on traitera ensuite.

1.1 Problème et notations : Soit I, J deux ensembles finis ; un tableau rectangulaire de nombres sur $I \times J$, ou ensemble de nombres réels indicés par $(i, j) \in I \times J$, sera noté :

$$f_{IJ} = \{f_{ij} | i \in I, j \in J\} ;$$

on définit pour un tel tableau des marges notées $f_{I.}$ et $f_{.J}$ (ou plus brièvement f_I, f_J) :

$$f_{I.} = \{f_{i.} | i \in I\} ; f_{i.} = \sum \{f_{ij} | j \in J\} ;$$

$$f_{.J} = \{f_{.j} | j \in J\} ; f_{.j} = \sum \{f_{ij} | i \in I\} .$$

Ceci posé le problème considéré est le suivant. Etant donné un tableau f_{IJ} sur $I \times J$, et des systèmes de nombres g_I et g_J (indicés respectivement par I et J), trouver un tableau g_{IJ} ayant pour marges g_I et g_J et ressemblant le plus possible à f_{IJ} : on dira que g_{IJ} est obtenu en ajustant le tableau f_{IJ} aux marges imposées g_I et g_J ; ce qu'on pourra noter :

$$g_{IJ} = \text{Ajust}(f_{IJ} ; g_I, g_J) .$$

Ainsi qu'on l'a annoncé l'énoncé est vague : il comporte le terme de "ressemblance", dont on doit chercher à préciser le sens. Auparavant faisons quelques remarques.

(1) Attaché de Recherches CNRS - CREDOC 140 rue du Chevaleret.
Le présent article est issu d'une thèse de 3° cycle (univ. P. & M. Curie) Paris 1980.

(*) à paraître.

1) En reprenant ici les notations de l'a. des correspondances, nous suggérons que le cas qui nous intéresse le plus est celui où f_{IJ} est un tableau de nombres positifs (ou nuls), les g_i, g_j étant de même tous positifs : il est naturel alors de demander que g_{IJ} soit de même un tableau de nombres positifs. Nous reviendrons sur cette exigence au § 1.4.

2) Plus précisément le total des nombres inscrits dans le tableau donné f_{IJ} et aussi dans le tableau inconnu g_{IJ} , sera en général 1. En tous cas, il est indispensable que :

$$\Sigma\{g_i | i \in I\} = \Sigma\{g_j | j \in J\} = \text{tot.}$$

car l'un et l'autre total devra être celui des nombres inscrits dans g_{IJ} . Presque toujours nous supposons que le total est 1 : car l'ajustement d'un tableau (de total non-nul) à un total donné s'obtient immédiatement par multiplication par une constante ; poser tot = 1 ne restreint donc pas la généralité des raisonnements.

3) Nous parlons ici de l'ajustement à des marges pour un tableau rectangulaire $I \times J$: plus généralement le problème se pose pour un tableau multiple : e.g. un tableau ternaire f_{IJK} , avec des marges imposées g_I, g_J, g_K . (Les méthodes proposées aux §§ 2 et 3 se généralisent à de tels tableaux).

1.2 Critères de distance : Un premier type de critère d'ajustement est de minimiser une distance entre g_{IJ} (inconnu) et f_{IJ} (donné) sous la contrainte que g_{IJ} ait les marges g_I, \bar{g}_J (données). Le plus simple est de définir la distance entre deux tableaux par une formule quadratique telle que :

$$d^2(f_{IJ}, g_{IJ}) = \|f_{IJ} - g_{IJ}\|^2 = \Sigma\{(f_{ij} - g_{ij})^2 | i \in I, j \in J\};$$

formule qui par l'introduction de coefficients, offre de nombreuses variantes (cf § 2.2).

Or à côté de cette distance euclidienne, une autre définition, issue de la théorie de l'information (cf § 3.3) joue dans l'ajustement de tableau un rôle essentiel. L'écart entre f et g peut être défini par :

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

nous reviendrons sur cette formule : (sous la réserve que g et f soient des tableaux de nombres strictement positifs de total 1) on peut montrer que $H(g; f)$ est un nombre positif, qui ne s'annule que si $g=f$; mais il est clair qu'il ne s'agit pas d'une distance à strictement parler ; car l'axiome de symétrie n'est pas satisfait ; on a $H(g; f) \neq H(f; g)$. Ceci n'interdit certes pas de définir g comme étant la loi qui sous la contrainte d'avoir les marges imposées, s'écarte le moins de f : mais il faut justifier l'introduction de H .

1.3 Critères de forme : Modifier le moins possible un tableau f_{IJ} , c'est n'en pas changer la forme. Que faut-il entendre par là ? Ce pourrait être de ne pas changer les profils des lignes ni ceux des colonnes. En fait cela n'est possible que si l'une des marges est déjà juste : par exemple, si $f_I = g_I$, il suffit de poser $g_{ij} = (g_j/f_j)f_{ij}$; i. e. de multiplier chaque colonne j par un coefficient (g_j/f_j) ce

qui conserve les profils des colonnes en perturbant toutefois ceux des lignes. Mais en général cette transformation est le premier pas d'un remarquable algorithme, l'algorithme RAS, objet du § 3. Et il se trouve (cf § 3.3) que la solution à laquelle on aboutit peut être caractérisée par la propriété de minimiser l'écart $H(g_{IJ}; f_{IJ})$ dont la formule complexe a été rappelée au § 1.2.

Si plus généralement on cherche à perturber le moins possible la structure de fonction $(f_{ij}/f_i f_j) \approx (g_{ij}/g_i g_j)$, ou densité du tableau par rapport au produit de ses marges, on est conduit à fonder sur une analyse factorielle l'ajustement du tableau f_{IJ} à des marges imposées (cf § 4).

1.4 Propriétés demandées : Aux critères de formes, on peut rattacher la propriété d'agrégation dont plusieurs auteurs ont signalé l'importance. En bref on dit qu'une méthode M possède la propriété d'agrégation si on obtient ce même résultat en agrégeant des colonnes ou des lignes du tableau g_{IJ} (obtenu en ajustant f_{IJ} à des marges données g_i, g_j) ; ou en agrégeant ces lignes ou colonnes sur le tableau f_{IJ} avant d'appliquer la méthode M, (en prenant soin, cela s'entend, de modifier les marges imposées : e.g. si on cumule les lignes i_1 et i_2 en une seule ligne i_s à celle-ci est imposé un total $g_{i_s} = g_{i_1} + g_{i_2}$). Pratiquement on vérifiera surtout la propriété d'agrégation *restreinte* selon laquelle l'agrégation n'est permise qu'entre lignes ou colonnes de même profil.

Dans la pratique, une propriété essentielle est la *stabilité* du résultat g_{IJ} , vis-à-vis des perturbations éventuelles des données. L'exigence de la stabilité est à la base de la méthode statistique ASAM (cf Payen et Durieux : INSEE et *infra* § 2.2) ; et nous l'avons eu en vue, en recourant à l'analyse factorielle (§ 4).

Enfin dans la plupart des cas le tableau g_{IJ} doit être exclusivement constitué de nombres positifs. Or, particulièrement si les marges imposées sont très différentes de celles du tableau initial f_{IJ} , ou que ce dernier contient des cases faibles ou nulles, la plupart des méthodes (à l'exception de RAS : § 3) introduisent des termes négatifs. On peut remplacer ceux-ci par des zéros ; et s'en tenir là si les marges sont peu perturbées par ces corrections. Sinon on applique une ou deux itérations de RAS afin de revenir à des marges acceptables.

Eventuellement, on peut imposer qu'à toute case nulle f_{ij} , il corresponde dans le tableau ajusté une case nulle g_{ij} .

Plus généralement, on pourra s'inquiéter chaque fois que le rapport (g_{ij}/f_{ij}) prend des valeurs trop écartées de 1. De ce point de vue on a appliqué le traitement suivant qui généralise celui des cas négatifs.

- a) faire l'histogramme des Card I \times Card J nombres (g_{ij}/f_{ij})
- b) fixer un seuil min correspondant à la valeur au-dessous de laquelle le quotient ne descend par exemple que dans 5% des cas (notamment les cas où g_{ij} est négatif).
- c) dans ces cas, substituer à g_{ij} la valeur $g'_{ij} = f_{ij} \times \text{min}$.
- d) on peut traiter de même les valeurs très élevées de (g_{ij}/f_{ij})

en introduisant un seuil MAX, qui n'est dépassé (e. g.) que dans 5 % des cas.

2 Méthodes de minimisation d'une distance quadratique

2.1 Distance euclidienne à coefficients constants : Rappelons la formule du § 1.2 :

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

l'ajustement suivant cette distance conduit à la formule :

$$g_{ij} = f_{ij} + (g_i - f_i)/\text{Card}J + (g_j - f_j)/\text{Card}I ;$$

en d'autres termes, pour modifier de $(g_i - f_i)$ le total de la ligne i , on ajoute à chaque terme de cette ligne une même quantité ; et de même pour ajuster le total des colonnes. Il est remarquable que (à la différence de ce qu'on verra au § 3.1 pour la méthode RAS) l'ajustement des colonnes ne perturbe pas celui des lignes et réciproquement : cela est dû à ce que (selon notre hypothèse) $\sum\{g_i - f_i\} = \sum g_i - \sum f_i = 0$ et de même $\sum\{g_j - f_j\} = 0$. Si toutefois on suppose que les tableaux considérés ont un total quelconque il convient d'ajouter à chacune des cases la quantité $(t_g - t_f)/(\text{Card}I \text{ Card}J)$, où $t_f = \sum f_i = \sum f_j$ et $t_g = \sum g_i = \sum g_j$ sont respectivement le total général du tableau donné et celui du tableau demandé.

On notera que la méthode conserve la forme des lignes et des colonnes (cf § 1.3) non en respectant les profils, mais en effectuant des décalages...

2.2 Introduction de coefficients dans la formule quadratique : Considérons la formule de distance suivante où interviennent des coefficients b_i et c_j (qui peuvent dépendre des données f_i, f_j, g_i, g_j) :

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

avec des tableaux de correspondance (tableaux de total 1), le résultat de l'ajustement est :

$$g_{ij} = f_{ij} + (g_i - f_i)(c_j/ct) + (g_j - f_j)(b_i/bt)$$

(où $ct = \sum_j c_j$; $bt = \sum_i b_i$).

Froment et Lenclud (*INSEE op. cit.*) ont montré que cette méthode est la seule parmi celles qui reposent sur la minimisation d'une distance quadratique à posséder la propriété d'agrégation (pourvu qu'on prenne soin d'agréger les b_i et c_j). Elle conserve la structure d'indépendance (i.e. si $f_{ij} = f_i f_j$ alors $g_{ij} = g_i g_j$) dans le cas où les b_i, c_j sont déterminés en fonction des marges de f et g , et d'une constante k ($k \in (0,1)$) suivant les formules :

$$b_i = k f_i + (1 - k) g_i ; c_j = k g_j + (1 - k) f_j.$$

Si en particulier $k = 1/2$ il vient :

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

Aux formules quadratiques quelconques, se rattache la méthode ASAM qui consiste en bref, à minimiser une distance quadratique pondérée à coefficients de pondération indéterminés ; puis à choisir ces

coefficients de telle sorte que soit minimale la variance des estimations obtenues. D'où le sigle ASAM : *Ajustement Statistique et Algèbre d'une Matrice à des marges*.

2.3 Projection et fibration : En géométrie euclidienne, minimiser une distance, c'est effectuer une projection orthogonale. Voyons ce qu'il en est ici dans le cas d'une distance à *coefficients constants*. De façon précise, la contrainte des marges g_I, g_J définit au sein de l'ensemble R_{IJ} de tous les tableaux sur $I \times J$, un sous-ensemble L de tableaux x :

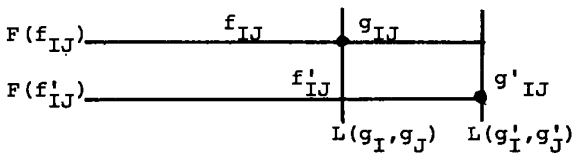
$$L(g_I, g_J) = \{x_{IJ} \mid x_{IJ} \in R_{IJ} ; x_I = g_I ; y_J = g_J\}$$

(plus précisément on peut se restreindre à l'hyperplan H_{IJ} formé des tableaux dont le total est 1) ; $L(g_I, g_J)$ est un sous-espace linéaire (on dit encore sous-espace affín : analogue d'un plan etc.) ; et quand varie la contrainte de marge, (i.e. les g_I, g_J), le sous-espace $L(g_I, g_J)$ varie en restant parallèle à lui-même.

Le tableau $g_{IJ} = \text{Ajus}(f_{IJ}; g_I, g_J)$ n'est donc autre que la projection orthogonale de f_{IJ} (considéré comme un point de H_{IJ}), sur le sous-espace linéaire $L(g_I, g_J)$. Ou encore en notant F la direction d'un sous-espace de dimension maximale orthogonal à L (i.e. tout vecteur de F est orthogonal à tout vecteur de L : par exemple dans R^n si L a pour équation $x^1 = x^2 = \dots = x^p = 0$; F aura pour équation $x^{p+1} = \dots = x^n = 0$) ; on peut dire que g_{IJ} est à l'intersection du sous-espace $F(f_{IJ})$ de direction F passant par f_{IJ} (sous-espace qu'on peut appeler la *fib*re de f_{IJ}), et du sous-espace $L(g_I, g_J)$:

$$\text{Ajus}(f_{IJ}; g_I, g_J) = F(f_{IJ}) \cap L(g_I, g_J).$$

C'est ce que symbolise le dessin ci-joint ; où on a figuré également, l'ajustement d'un autre tableau f'_{IJ} , à d'autres marges g'_I, g'_J .



De ce point de vue, il apparaît qu'il y a en quelque sorte dans l'espace H_{IJ} de tous les tableaux un double système de coordonnées. D'une part la condition de marge impose au tableau d'être sur une variété $L(g_I, g_J)$; d'autre part, l'ajustement du tableau initial f_{IJ} ne peut aboutir qu'à un tableau g_{IJ} qui appartienne à la variété (ou *fib*re) $F(f_{IJ})$, et cela quelles que soient les marges g_I et g_J qu'on lui imposera.

3 La méthode R.A.S. : A l'exception du § 3.5, nous considérerons exclusivement ici des tableaux de correspondance binaires (ou tableaux rectangulaires) remplis de nombres *strictement positifs* de total 1.

3.1 Algorithme RAS et critère de forme : L'idée est simple : pour donner au tableau f_{IJ} la marge g_I imposée, il suffit de multiplier chacune de ses lignes i par le rapport (g_i/f_i) correspondant : ainsi sans modification de profil le total de la ligne ne sera plus f_i mais $(g_i/f_i) f_i = g_i$. De même pour donner au tableau f_{IJ} la marge g_J on multipliera chaque ligne j par le rapport (g_j/f_j) . La difficulté est que ces deux opérations d'ajustement ne sont pas compatibles entre elles : en ajustant à la marge g_J on détruit l'ajustement qu'on vient de faire à la marge g_I ; en ajustant alors une deuxième fois à g_I , on perturbe l'ajustement fait à g_J ... Pourtant en répétant les opérations d'ajustement, l'écart aux marges imposées diminue graduellement (cf § 3.4) ; il y a convergence vers un tableau g_{IJ} ajusté aux marges imposées. De façon précise on peut concevoir les trois algorithmes suivants, qui tous trois convergent, et convergent vers la même solution.

a) *Ajustement alternatif en ajustant d'abord à g_J* : On construit une suite de tableaux de correspondances $x_{IJ}^{(m)}$, convergeant vers la solution g_{IJ} , et définie comme suit :

$$\begin{aligned} x_{IJ}^{(0)} &= f_{IJ} ; x_{ij}^{(1)} = f_{ij} (g_j/f_j) ; x_{ij}^{(2)} = x_{ij}^{(1)} (g_i/x_i^{(1)}) ; \dots \\ x_{ij}^{(2n)} &= x_{ij}^{(2n-1)} (g_i/x_i^{(2n-1)}) ; \\ x_{ij}^{(2n+1)} &= x_{ij}^{(2n)} (g_j/x_j^{(2n)}) ; \dots \end{aligned}$$

dans ces formules $x_I^{(p)}$, $x_J^{(p)}$ sont les marges du p -ème tableau $x_{IJ}^{(p)}$.

b) *Ajustement alternatif en ajustant d'abord à g_I* : La construction est analogue :

$$\begin{aligned} y_{IJ}^{(0)} &= f_{IJ} ; y_{ij}^{(1)} = f_{ij} (g_i/f_i) ; y_{ij}^{(2)} = y_{ij}^{(1)} (g_j/y_j^{(1)}) ; \dots \\ y_{ij}^{(2n)} &= y_{ij}^{(2n-1)} (g_j/y_j^{(2n-1)}) ; \\ y_{ij}^{(2n+1)} &= y_{ij}^{(2n)} (g_i/y_i^{(2n)}) ; \dots \end{aligned}$$

c) *Ajustement simultané aux deux marges* : On construit la suite des $z_{IJ}^{(p)}$:

$$\begin{aligned} z_{IJ}^{(0)} &= f_{IJ} ; z_{ij}^{(1)} = c^{(1)} f_{ij} (g_i/f_i) (g_j/f_j) ; \dots \\ z_{ij}^{(p+1)} &= c^{(p+1)} z_{ij}^{(p)} (g_i/z_i^{(p)}) (g_j/z_j^{(p)}) ; \dots \end{aligned}$$

En fait, comme on l'a dit, il est impossible d'ajuster à la fois aux deux marges g_I et g_J par simple multiplication des lignes et colonnes par les quotients des marges. Telle quelle cette simple opération d'ajustement ne respecte même pas le total du tableau (qui, on l'a dit, doit être 1) : c'est pourquoi on introduit une constante $c^{(p)}$ choisie justement pour que $z_{IJ}^{(p)}$ ait bien pour total 1.

3.2 Fibration associée à l'algorithme : Pour comprendre comment ces algorithmes conduits différemment, aboutissent à un même g_{IJ} (alors qu'il y a tant de tableaux possibles ayant les marges g_I et g_J !), le plus simple est de noter que les suites $\{x_{IJ}^{(p)}\}$, $\{y_{IJ}^{(p)}\}$, $\{z_{IJ}^{(p)}\}$ sont toutes sur une même fibre $F(f_{IJ})$ et cela quelles que soient d'ailleurs les marges imposées g_I et g_J . En effet tous les tableaux x , y ou z construits à partir de f par l'un ou l'autre des trois algorithmes sont compris dans la formule :

$$t_{ij} = f_{ij} A_i B_j ;$$

dans cette formule les A_i , B_j sont des nombres positifs quelconques (qu'il faut seulement choisir tels que le tableau t_{IJ} ait pour total 1). Nous pouvons noter $F(f_{IJ})$ l'ensemble (ou fibre) des tableaux qu'on peut ainsi obtenir à partir de f_{IJ} . On peut encore dire que dans l'espace des coordonnées logarithmiques (on se souvient que nous avons imposé que les f_{ij} soient tous strictement positifs) les fibres $F(f_{IJ})$ sont définies par des équations linéaires : de façon précise ; $F(f_{IJ})$ est l'ensemble des lois t_{IJ} telles que, quel que soit le tableau α_{IJ} ayant marges nulles (i.e. $\forall i : \alpha_i = 0$; $\forall j : \alpha_j = 0$) on ait :

$$\Sigma \{ \alpha_{ij} (\log f_{ij} - \log t_{ij}) \mid i \in I, j \in J \} = 0 .$$

Mais si l'on considère le complexe P_{IJ} des lois de probabilité portées par $I \times J$, les fibres $F(f_{IJ})$ sont des variétés courbes (à la différence des $F(f_{IJ})$ introduites au § 2.3) ; tandis que les $L(g_I; g_J)$ sont les mêmes variétés affines que précédemment (au § 2.3). Ceci nous conduit à réduire l'étude de la méthode RAS à deux problèmes :

a) *Existence et unicité* : quelles que soient les lois f_{IJ} , g_I , g_J (assujetties à être des mesures strictement positives de total 1), la fibre $F(f_{IJ})$ et la variété linéaire $L(g_I; g_J)$ se coupent en un point et un seul (au sein du simplexe P_{IJ}).

b) *Convergence* : ce point unique est en effet atteint par l'une ou l'autre des variantes de l'algorithme (cf § 3.1, a, b, c).

La démonstration de l'existence et de l'unicité est clairement donnée dans la littérature : nous la rappelons au § 3.3. Quant à la convergence il est plus délicat de l'établir en toute rigueur : nous donnons au § 3.4 pour l'alg. (a) (ou b qui lui est équivalent ; mais non pour c) une démonstration qui nous satisfait...

3.3 Existence et unicité de la solution définie par un écart : Classiquement s'introduit dans la comparaison de deux lois de probabilité p_I et q_I sur un même ensemble I la quantité $H(p_I; q_I)$ appelée entropie relative (ou parfois information de Kullback) ; cf e.g. Traité TI B n° 5 § 1.3) :

$$H(p_I; q_I) = \Sigma \{ p_i \log_2(p_i/q_i) \mid i \in I \} .$$

On peut montrer ce qui suit : étant donné f_{IJ} , g_I , g_J , il existe sur $L(g_I, g_J)$ une loi de probabilité g_{IJ} réalisant le minimum de

l'écart $H(g_{IJ}; f_{IJ})$ à la loi f_{IJ} à ajuster ; cette loi g_{IJ} appartient à la fibre $F(f_{IJ})$ et elle en est l'unique intersection avec $L(g_I, g_J)$.

La démonstration résulte des propriétés de $H(p_I; q_I)$: (cf *Traité Loc. cit.*).

a) pour q_I fixé, $H(p_I; q_I)$ est une fonction de p_I continue sur le simplexe P_I des lois de probabilités sur I (plus exactement sur le simplexe fermé : i.e. avec les lois p_I comportant des masses p_i nulles ; en effet $x \log x$ est une fonction continue tendant vers zéro par valeur positive) ;

b) la fonction $H(p_I; q_I)$ est positive et ne s'annule que si $p_I = q_I$.

c) la fonction est strictement convexe : en ce sens que si l'on se restreint à faire varier p_I sur un segment de droite : (e.g. $p_I^{(t)} = p_I^{(0)} + t(p_I^{(1)} - p_I^{(0)})$; où $p_I^{(0)}$ et $p_I^{(1)}$ sont deux lois sur I) $H(p_I^{(t)}; q_I)$ est une fonction strictement convexe de t (pour $t \in (0,1)$) ; ou encore une fonction dont la courbe représentative est située au dessus de chacune de ses tangentes ne la touchant qu'en un point.

Ceci rappelé, de la continuité (a) résulte l'existence d'un g_{IJ} réalisant le minimum absolu de $H(g_{IJ}; f_{IJ})$ sur $L(g_I, g_J)$ (plus exactement sur $L(g_I; g_J) \cap P_{IJ}$: au sein du simplexe). Ce g_{IJ} est intérieur à P_{IJ} (i.e. les g_{ij} sont tous non-nuls), parce que en bref, à l'origine la dérivée de $x \log x$ vaut $-\infty$. En écrivant que les dérivées partielles de H par rapport aux g_{ij} s'annulent sous les contraintes de marge on exprime exactement que g_{IJ} est un point de la fibre $F(f_{IJ})$. En effet :

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

$$\partial H / \partial g_{ij} = 1 - \log(g_{ij}/f_{ij}) ;$$

les équations de contrainte sont :

$$\forall i : \sum_j g_{ij} = g_i \Rightarrow \sum_j dg_{ij} = 0 ;$$

$$\forall j : \sum_i g_{ij} = g_j \Rightarrow \sum_i dg_{ij} = 0 ;$$

une combinaison linéaire de ces contraintes avec coefficients α_i, β_j doit donner $dH = 0$; il faut donc que

$$\forall i, j : 1 - \log(g_{ij}/f_{ij}) = \alpha_i + \beta_j ;$$

ce qui équivaut à la condition $g_{IJ} \in F(f_{IJ})$: i.e. il existe des A_i, B_j tels que $\forall i, j : g_{ij} = A_i B_j f_{ij}$. Enfin l'unicité de l'extremum de H (donc de l'intersection $L(g_I, g_J) \cap F(f_{IJ})$) résulte de la convexité stricte.

Reste à justifier l'introduction de l'écart H : en bref, on remonte les calculs qu'on vient d'effectuer : pour que le point d'écart minimum soit de la forme $g_{ij} = A_i B_j f_{ij}$, il faut que $\log(g_{ij}/f_{ij})$ soit de la forme $\alpha_i + \beta_j$; et ceci résulte de la condition d'extrémalité si

$\partial H / \partial g_{ij}$ est une fonction linéaire de $\log(g_{ij}/f_{ij})$: d'où H par intégration (il importe peu qu'on pose $\partial H / \partial g_{ij} = a \log(g_{ij}/f_{ij}) + b$, avec a et b autres que $a = -1$ et $b = 1$: car vu que $\Sigma g_{ij} = 1$, b ne joue aucun rôle ; quant à a son choix n'est qu'un choix d'échelle : il faut seulement prendre a négatif ; afin que le signe de H soit +).

3.4 Convergence de l'algorithme : Nous considérons l'algorithme a) du § 3.1 :

$$x_{IJ}^{(0)} = f_{IJ} ; x_{ij}^{(1)} = f_{ij}(g_j/f_j) ; x_{ij}^{(2)} = x_{ij}^{(1)}(g_i/x_i^{(1)}) ; \dots$$

$$x_{ij}^{(2n)} = x_{ij}^{(2n-1)}(g_i/x_i^{(2n-1)}) ; \text{ d'où : } x_i^{(2n)} = g_i ;$$

$$x_{ij}^{(2n+1)} = x_{ij}^{(2n)}(g_j/x_j^{(2n)}) ; \text{ d'où : } x_j^{(2n+1)} = g_j .$$

L'essentiel est de vérifier la convergence des marges vers g_I et g_J . Pour cela on note :

$$E^{(m)} = \Sigma_i |x_i^{(m)} - g_i| + \Sigma_j |x_j^{(m)} - g_j| ;$$

plus précisément selon la parité du rang l'une ou l'autre des deux sommes dont se compose $E^{(m)}$ est nulle (car alternativement l'ajustement est exact, soit à g_I soit à g_J) ; on a :

$$E^{(2n-1)} = \Sigma_i |x_i^{(2n-1)} - g_i| ; E^{(2n)} = \Sigma_j |x_j^{(2n)} - g_j| .$$

On va montrer que la suite des $E^{(m)}$ tend vers zéro comme une progression géométrique de raison $(1-\epsilon)$ inférieure à 1 (ϵ sera fixé ci-dessous par le lemme 1 : plus précisément on a la

$$\text{proposition : } \forall m : E^{(m+1)} \leq (1-\epsilon)E^{(m)} .$$

De la proposition, il résulte facilement que chacune des suites $x_{ij}^{(m)}$ converge quand $m \rightarrow \infty$: car en bref la variation $|x_{ij}^{(m)} - x_{ij}^{(m+1)}|$ est majorée par $|(g_i/x_i^{(m)}) - 1|$ ou $|(g_j/x_j^{(m)}) - 1|$, eux-mêmes majorés par le terme d'une série géométrique de raison $(1-\epsilon)$.

Reste à établir la proposition : pour cela, on fait usage du

Lemme 1 : il existe un nombre ϵ positif auquel à tout rang m , tous les termes $x_{ij}^{(m)}$ de la matrice $x_{IJ}^{(m)}$ sont supérieurs.

Nous démontrons d'abord la proposition, puis le lemme.

démonstration de la proposition : considérons par exemple la parité $m = 2n-1$ (le cas m pair se traite de façon analogue en échangeant les rôles de I et J). En calculant $E^{(2n)}$ en fonction du tableau $x^{(2n-1)}$, il apparaît que $E^{(2n)}$ est une somme de termes dont les valeurs absolues ont pour somme $E^{(2n-1)}$; avec dans $E^{(2n)}$ certaines destructions de termes de signes opposés qui nous assurent l'inégalité $E^{(2n)} < (1-\epsilon) E^{(2n-1)}$. On a :

$$\begin{aligned}
 E^{(2n)} &= \sum_j |x_j^{(2n)} - g_j| \\
 &= \sum_j | \sum_i \{x_{ij}^{(2n-1)} (g_i/x_i^{(2n-1)})\} - g_j | \\
 &= \sum_j | \sum_i \{ (x_{ij}^{(2n-1)}/x_i^{(2n-1)}) (g_i - x_i^{(2n-1)}) \} |
 \end{aligned}$$

(où on a tenu compte de ce que $x_j^{(2n-1)} = \sum_i x_{ij}^{(2n-1)} = g_j$) ; de plus parce que $\sum_i g_i = \sum_i x_i^{(2n-1)} = 1$, on peut écrire :

$$\begin{aligned}
 E^{(2n)} &= \sum_j | \sum_i \{ ((x_{ij}^{(2n-1)}/x_i^{(2n-1)}) - \epsilon) (g_i - x_i^{(2n-1)}) \} | \\
 &\leq \sum_j |g_i - x_i^{(2n-1)}| | ((x_{ij}^{(2n-1)}/x_i^{(2n-1)}) - \epsilon) | ;
 \end{aligned}$$

où on a tenu compte de ce que la norme \sum_j des modules des sommes partielles \sum_i est majorée par la somme générale des modules des termes ; et de ce que $(x_{ij}^{(m)}/x_i^{(m)}) - \epsilon = (x_{ij}^{(m)}/x_i^{(m)}) - \epsilon$, parce que le quotient $x_{ij}^{(m)}/x_i^{(m)}$ est d'après le lemme supérieur à ϵ . Donc :

$$\begin{aligned}
 E^{(2n)} &\leq \sum_i |g_i - x_i^{(2n-1)}| | (\sum_j \{x_{ij}^{(2n-1)}/x_i^{(2n-1)}\}) - \epsilon \text{ Card } J | \\
 &\leq \sum_i |g_i - x_i^{(2n-1)}| (1 - \epsilon \text{ Card } J) \\
 &\leq E^{(2n-1)} (1 - \epsilon \text{ Card } J) < E^{(2n-1)} (1 - \epsilon) .
 \end{aligned}$$

Ce qui achève de démontrer la proposition.

Reste le lemme 1, qu'on déduira des lemmes 2 et 3. De façon précise on notera (cf § 3.2) :

$$x_{ij}^{(m)} = A_i^{(m)} B_j^{(m)} f_{ij}$$

compte-tenu de ce qu'on a ajusté alternativement aux marges g_I et g_J , on a :

$$\forall j \forall n : B_j^{(2n-1)} = B_j^{(2n)} ; \quad \forall i \forall n : A_i^{(2n-1)} = A_i^{(2n)} .$$

Définissons maintenant deux constantes K et k qui permettent de définir ϵ et interviennent dans l'énoncé des lemmes 2 et 3 :

$$K = 1/\inf\{f_{ij} \mid i \in I, j \in J\} ;$$

$$k = \inf(\inf\{g_i \mid i \in I\}, \inf\{g_j \mid j \in J\}) ; \quad \epsilon = (k/K)^2 .$$

Ceci posé on a (en notant simplement $\sup A^{(m)}$ pour $\sup\{A_i^{(m)} \mid i \in I\}$ et de même pour B , pour $\inf \dots$)

$$\text{Lemme 2 : } \forall m : \sup A^{(m)} \sup B^{(m)} < K .$$

$$\text{preuve du lemme 2 : } x_{ij}^{(m)} = A_i^{(m)} B_j^{(m)} f_{ij} < 1 ;$$

$$\text{Lemme 3 : } \forall n : \inf A^{(2n)} \sup B^{(2n-1)} > k ;$$

$$\inf B^{(2n+1)} \sup A^{(2n)} > k .$$

preuve du lemme 3 : démontrons par exemple la première inégalité.

Soit i tel que $A_i^{(2n)} = \inf A^{(2n)}$; on a (en se souvenant de ce que

$$B^{(2n)} = B^{(2n-1)} :$$

$$k < g_i = x_i^{(2n)} = \sum_j A_i^{(2n)} B_j^{(2n-1)} f_{ij} < \inf A^{(2n)} \sup B^{(2n-1)} \sum_j f_{ij} < \inf A^{(2n)} \sup B^{(2n-1)}.$$

preuve du lemme 1 : e.g. au niveau $m = 2n+1$; (avec $A^{(2n+1)} = A^{(2n)} \dots$) :

$$\begin{aligned} \inf B^{(2n+1)} \inf A^{(2n+1)} &= \inf B^{(2n+1)} \inf A^{(2n)} \\ &= (\inf B^{(2n+1)} \sup A^{(2n)}) / (\inf A^{(2n)} \sup B^{(2n-1)}) / (\sup A^{(2n)} \sup B^{(2n-1)}) \\ &\geq k^2 / K ; \text{ d'où :} \end{aligned}$$

$$\begin{aligned} x_{ij}^{(2n+1)} &= f_{ij} A_i^{(2n+1)} B_j^{(2n+1)} \\ &\geq \inf\{f_{ij}\} \inf A^{(2n+1)} \inf B^{(2n+1)} = k^2 / K^2 = \epsilon. \end{aligned}$$

Ce qui achève la démonstration du lemme 1, de la proposition ; et donc de la convergence de l'algorithme.

3.5 Propriétés de la méthode R.A.S. et généralisation : La propriété d'agrégation est vérifiée par la méthode RAS, sous sa forme restreinte, i.e. à condition qu'on se borne à agréger des lignes ou des colonnes de même profil. Et il faut souligner que dans le traitement d'un tableau de nombres positifs la méthode RAS n'introduit pas de nombres négatifs. La méthode se généralise aisément à une correspondance multiple : par exemple à l'ajustement d'un tableau ternaire f_{IJK} à des marges imposées g_I, g_J, g_K . En revanche le fait que tous les f_{ij} soient strictement positifs joue un rôle essentiel. Certes les algorithmes du § 3.1 peuvent s'écrire pour un tableau f_{IJ} quelconque : mais la convergence n'est pas assurée ni même l'existence d'une solution. On en peut juger dès le cas le plus simple d'un tableau 2×2 (cf Ph. L'Hardy in Ann. de l'INSEE N°s 22-23, 1976). Mais à supposer que l'algorithme aboutisse, le résultat obtenu g_{IJ} comportera nécessairement des zéros, là où on a le tableau f_{IJ} .

4 Ajustement d'après une analyse factorielle : Dans ce § nous considérerons exclusivement des tableaux de correspondance binaires, ou tableaux rectangulaires, de nombres positifs ou nuls de total 1. Deux des formules considérées ci-dessus, évoquent la formule de reconstitution de l'analyse des correspondances.

a) Par minimisation d'une distance quadratique comportant des coefficients convenables (cf § 2.2) on a obtenu :

$$g_{ij} = (f_{ij} - f_i f_j) + g_i g_j ;$$

Dans cette formule $g_i g_j$ représente le 1-er terme d'une loi g_{IJ} de marges g_I, g_J : le terme qui correspond à l'hypothèse d'indépendance. Tandis que $f_{ij} - f_i f_j$ est dans la décomposition de f_{IJ} , l'écart à l'indépendance ; ou encore le terme dépendant des facteurs non tri-axiaux :

$$(f_{ij} - f_i f_j) = f_i f_j \left(\sum_{\alpha} \{\lambda_{\alpha}^{-1/2} F_{\alpha}(i) G_{\alpha}(j)\} \right).$$

En somme g_{IJ} est construit en prenant le terme principal fourni par les marges imposées g_I et g_J ; et l'écart qui provient du tableau f_{IJ} à ajuster.

b) Une voie pour aboutir itérativement à la solution RAS est de poser (cf 3.1 (c)) :

$$g_{ij} \approx c f_{ij} (g_i/f_j) = c (f_{ij}/(f_i f_j)) g_i g_j ;$$

(où c est une constante introduite pour que le tableau construit ait bien pour total 1). Dans cette formule (qui n'est qu'approchée ; mais est analogue à la formule exacte), on rapporte la loi cherchée g_{IJ} au produit $g_I g_J$ des marges imposées ; et on accepte pour densité $(g_{ij}/(g_i g_j))$, la densité même $(f_{ij}/(f_i f_j))$ du tableau initial par rapport au produit de ses marges.

Il semble donc opportun de construire d'après l'analyse factorielle du tableau f_{IJ} donné, un reste r^{ij} tel que la formule :

$$g_{ij} = g_i g_j (1 + r^{ij})$$

représente effectivement une loi g_{IJ} ajustée aux marges imposées.

Prendre tel quel le reste r^{ij} correspondant à la loi f_{IJ} donne la formule :

$$g_{ij} \approx g_i g_j (1 + \sum_{\alpha} \{\lambda_{\alpha}^{i,j} F_{\alpha}(i) G_{\alpha}(j)\}),$$

(où les λ_{α} , F_{α} , G_{α} sont ceux issus de l'analyse de f_{IJ}) : et l'on sait (cf b) que cette formule ne représente pas une loi ayant exactement les marges imposées g_I et g_J : en fait il n'est même pas assuré que le total du tableau ainsi construit soit 1. Il faut donc de quelque manière modifier les F_{α} , G_{α} (voire les λ_{α}) issus de f_{IJ} .

Une première voie est de substituer aux F_{α} des fonctions F_{α}^g de moyenne nulle, deux à deux non corrélées et de variance λ_{α} pour la loi marginale g_I (différente de f_I). Le plus simple est de raisonner plutôt sur les φ_{α}^I (facteurs de variance 1), en commençant par le facteur trivial φ_0^I (constant et égal à 1) ; et de construire la suite des φ_{α}^{gI} par orthonormalisation de la suite des φ_{α}^I relativement à la loi marginale imposée g_I ; les nouveaux facteurs φ_{α}^{gI} sont combinaisons linéaires des facteurs originels φ_{α}^I ; (le terme en φ_0 ou terme constant sur I , fournissant en particulier le décalage nécessaire pour que les φ_{α}^{gI} soient de moyenne nulle). L'on procède de même pour les facteurs φ_{α}^{gJ} sur J , qu'on orthogonalise relativement à g_J . Les calculs reposent sur la procédure classique d'orthogonalisation de Schmidt et Hilbert (cf TII B n° 12 § 1.3) dont L. Lebart a écrit un programme efficace. L'intérêt de cette voie est que par retouche des résultats de l'analyse factorielle du tableau donné f_{IJ} , on obtient un tableau g_{IJ} ajusté aux marges imposées g_I, g_J , et dont les $F_{\alpha}^g = \lambda_{\alpha}^{i,j} \varphi_{\alpha}^{gI} \varphi_{\alpha}^{gJ}$ et $G_{\alpha}^g = \lambda_{\alpha}^{i,j} \varphi_{\alpha}^{gI}$ sont précisément les facteurs (les valeurs propres restant λ_{α}) : cette dernière propriété est évidente sur la définition même des g_{ij} :

$$g_{ij} = g_i g_j (1 + \sum_{\alpha} \{\lambda_{\alpha}^{1/2} F_{\alpha}^g(i) G_{\alpha}^g(j)\}),$$

compte-tenu de l'orthonormalisation.

Toutefois, on peut craindre que la rotation des facteurs qu'implique l'orthogonalisation n'aboutisse finalement à un g_{IJ} dont la forme soit moins proche de celle de f_{IJ} , qu'on ne le désirerait. C'est pourquoi on a également tenté une deuxième voie. Soit la formule :

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

ici on a conservé les facteurs F_{α} , G_{α} issus de l'analyse du tableau initial f_{IJ} , en faisant seulement subir à chacun d'eux un décalage constant, A_{α} , B_{α} . On peut montrer, et c'est l'objet de l'article [AJUS. MARGES FAC.] que pour des valeurs faciles à déterminer des constantes A_{α} , B_{α} , le tableau g_{IJ} défini ci-dessus a en effet les marges imposées g_I, g_J . Cette deuxième méthode plus simple que la précédente, a toutefois relativement à celle-ci un inconvénient : les $(F_{\alpha}(i) - A_{\alpha})$ et $(G_{\alpha}(j) - B_{\alpha})$ ne sont pas les facteurs issus de l'analyse du tableau g_{IJ} que l'on construit ; en fait ces fonctions n'ont même pas moyenne nulle pour les lois marginales g_I et g_J ; pour ne rien dire de l'orthogonalité. Mais il est en revanche facile de situer les profils des lignes et colonnes du tableau g_{IJ} par rapport à ceux relatifs au tableau initial f_{IJ} , en adjoignant g_{IJ} en lignes et colonnes supplémentaires à f_{IJ} ; ainsi qu'on l'explique dans l'article cité. Ainsi on ne peut apprécier *de visu*, les déformations du tableau qu'entraîne l'ajustement à des marges.

Dans l'une et l'autre voie il faut éventuellement éliminer les g_{ij} négatifs, comme on l'a dit au § 1.4. Un avantage de l'analyse factorielle est que la formule de reconstitution peut être arrêtée au nombre de facteurs qu'on désire : ainsi le tableau ajusté est construit en quelque sorte après filtrage du tableau donné (en ne retenant de celui-ci que les facteurs qui représentent une structure réelle ; et non des fluctuations d'échantillonnage). Enfin l'ajustement par l'analyse factorielle est compatible avec l'agrégation de lignes ou de colonnes à condition que les éléments agrégés aient même profil cela est visible sur la formule de reconstitution (qu'on ait procédé par orthonormalisation ; ou par simple décalage des facteurs issus de f_{IJ}).

Nous terminons ici cet article, en renvoyant le lecteur à [AJUS. MARGES FAC.] ; et en annonçant des applications ultérieures à l'extrapolation des suites de tableaux.