Statistique et analyse des données

Bernard Fichet

Note sur la métrique de l'analyse des correspondances

Statistique et analyse des données, tome 3, n° 2 (1978), p. 87-93

<http://www.numdam.org/item?id=SAD_1978__3_2_87_0>

Statistique et Analyse des Données
2 - 1978

NOTE SUR LA METRIQUE DE L'ANALYSE DES CORRESPONDANCES

FICHET Bernard

Laboratoire de Physique. Faculté de Médecine

Université de Marseille II

Nous montrons que la métrique de l'analyse des correspondances est, à une constante positive multiplicative près, la seule métrique de la forme

$d^2(a_i,a_{i'}) = \sum_j f[p(.,j)]\left[\frac{p(i,j)}{p(i,.)} - \frac{p(i',j)}{p(i',.)}\right]^2$ (où f est une fonction numérique strictement positive), satisfaisant au principe d'équivalence distributionnelle, lorsque l'une des trois conditions suivantes est vérifiée :

- les p(i,j) sont tous rationnels
- f est continue
- f est monotone non croissante

1 -INTRODUCTION

Rappelons les données de l'analyse des correspondances. Deux ensembles finis $A=\{a_1,..,a_n\}$ et $B=\{b_1,..,b_p\}$ ($n \geqslant 2$, $p \geqslant 2$) sont mis en association - on dit en correspondance - par une matrice réelle C d'ordre (n,p), de terme général p(i,j),i=1,..,n ; j=1,..,p, que nous appellerons matrice des correspondances, et qui vérifie les relations :

$$\forall i = 1,..,n \quad \forall j = 1,..,p \quad p(i,j) \geqslant 0$$

$$\forall i = 1,..,n \quad \sum_{j=1}^{p} p(i,j) > 0$$

$$\forall j = 1,..,p \quad \sum_{i=1}^{n} p(i,j) > 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{p} p(i,j) = 1$$

A titre d'exemples,une telle matrice peut être, soit donnée par une probabilité sur $(A \times B, \mathcal{P}(A \times B))$ - correspondance probabiliste -,soit déduite d'un tableau de contingence - correspondance statistique -

Rappelons maintenant quelques notations classiques.

$$(1)\begin{cases} \forall i = 1,..,n \quad p(i,.) = \sum_{j=1}^{p} p(i,j) > 0 \ ; \quad \forall j = 1,..,p \quad p(.,j) = \sum_{i=1}^{n} p(i,j) > 0 \\ \forall i = 1,..,n,\ \forall j = 1,..,p \quad p_A(i|j) = \dfrac{p(i,j)}{p(.,j)} \ , \ p_B(j|i) = \dfrac{p(i,j)}{p(i,.)} \end{cases}$$

Pour une partition donnée $\{b'_1,..,\ b'_m\}$ (d'éléments non vides), de B, notons encore :

$$(2)\begin{cases} \forall i=1,..,n \quad \forall r=1,...,m \quad p'(i,r) = \sum_{\substack{j \\ b_j \in b'_r}} p(i,j) \\ \forall r = 1,...,m \quad p'(.,r) = \sum_{i=1}^{n} p'(i,r) = \sum_{\substack{j \\ b_j \in b'_r}} p(.,j) \\ \forall i=1,..,n \quad \forall r = 1,..,m \quad p'_A(i|r) = \dfrac{p'(i,r)}{p'(.,r)} \ , \ p'_B(r|i) = \dfrac{p'(i,r)}{p(i,.)} \end{cases}$$

Considérons alors une famille d'écarts sur A, soit $\{d_f\}$, définis par :

$$(3) \quad \forall i, \forall i' = 1,..,n \quad d_f^2(a_i,a_{i'}) = \sum_{j=1}^{p} f[p(.,j)]\,[p_B(j|i) - p_B(j|i')]^2$$

où f appartient à une classe de fonctions numériques strictement positives.

La propriété connue sous le nom d'équivalence distributionnelle -[1], tome 1, p.24, tome 2, p.152 ; [2] p.229 - peut alors être présentée comme suit.
Similairement on introduit une famille d'écarts $\{\delta_f\}$ sur B. De façon classique, soit $\sim$ la relation d'équivalence sur B définie par :

$$b_j \sim b_{j'} \iff \delta_f(b_j,b_{j'}) = 0 \iff \forall i = 1,..,n \quad p_A(i|j) = p_A(i|j'), \text{ et soit}$$

$B^* = B/\sim = \{b_1',..,b'_m\}$ l'ensemble des classes d'équivalence. Alors, avec les notations de (2), on peut introduire une nouvelle famille d'écarts sur A, soit $\{d'_f\}$, par :

$$(4) \quad \forall i, \forall i' = 1,...,n \quad {d'_f}^2(a_i,a_{i'}) = \sum_{r=1}^{m} f[p'(.,r)]\,[p'_B(r|i) - p'_B(r|i')]^2$$

Alors pour n et p donnés, on dit que d_f satisfait à la propriété d'équivalence distributionnelle, si $d_f = d'_f$, quelle que soit la matrice des correspondances appartenant à une classe donnée.
Il est clair que s'il en est ainsi pour d_f, il en est encore ainsi pour $d_{\alpha f}$, où α est une

constante positive ; et il est classique - voir [1], [2] - que la métrique de l'analyse des correspondances (telle que $f(x) = 1/x$) satisfait au principe d'équivalence distributionnelle.

Nous nous interessons par cette note à la condition nécessaire, envisageant deux cas : celui où les éléments de la matrice des correspondances sont rationnels, comme dans le cas d'une correspondance statistique ; celui, plus général, où ces éléments sont réels, comme dans le cas d'une correspondance probabiliste. Dans ce but, nous notons :
- pour n et p fixés, $\mathcal{C}$ (resp. $\mathcal{C}_Q$) l'ensemble des matrices des correspondances d'ordre (n,p) à éléments réels (resp. rationnels).
- F (resp. F_Q) l'ensemble des fonctions numériques strictement positives, définies sur $]0,1[$ (resp. $]0,1[\cap Q$).
Il peut être souhaitable également d'imposer à f certaines conditions de régularité, comme d'être continue ou monotone non croissante. Cette dernière condition - discutée dans [2], p.228 - permet, en général, de limiter, au niveau de la métrique, l'influence d'un évènement de forte fréquence d'apparition. Nous notons encore :
- F^C (resp. F^D) l'ensemble des fonctions numériques positives continues (resp. monotones non croissantes) définies sur $]0,1[$
Remarquons enfin que si $f \in F$ (resp. F_Q, F^C, F^D), d_f est toujours définie, alors que ce n'est pas le cas pour d'_f s'il n'existe qu'une seule classe d'équivalence dans B ; mais comme dans ce cas $\left(\frac{p'(i,1)}{p(i,.)} - \frac{p'(i',1)}{p(i',.)} \right)^2 = (1-1)^2 = 0$, $\forall i, \forall i' = 1,..,n$, pour éviter de définir f au point 1, nous poserons alors $d'_f = 0$.

2 - RESULTATS

Proposition 1

Pour $n \geq 2$, $p > 2$ fixés, si $f \in F_Q$ est telle que $d_f = d'_f$, $\forall C \in \mathcal{C}_Q$, alors nécessairement il existe $\alpha \in \mathbb{R}_+^*$ tel que :

$$\forall x \in]0,1[\cap Q, \quad f(x) = \alpha/x$$

Proposition 2

Pour $n \geq 2$, $p > 2$ fixés, si $f \in F^C$ (resp. $f \in F^D$) est telle que $d_f = d'_f$ $\forall C \in \mathcal{C}$, alors nécessairement il existe $\alpha \in \mathbb{R}_+^*$ tel que :

$$\forall x \in]0,1[, \quad f(x) = \alpha/x$$

3 - DEMONSTRATION DES RESULTATS

A toute fonction numérique f définie sur un sous-ensemble de $]0,1[$, nous associons la fonction g_f telle que $g_f(x) = x^2 f(x)$, pour tout x appartenant au domaine de définition de f. La démonstration des propositions repose alors sur le :

Lemme

Pour $n \geq 2$, $p > 2$ fixés, pour que $f \in F_Q$ (resp. $f \in F$) soit telle que $d_f = d'_f$ $\forall C \in \mathcal{C}_Q$ (resp. $\forall C \in \mathcal{C}$), il est nécessaire et suffisant que :

$\forall x$, $\forall y \in]0,1[\cap Q$ tels que $x + y < 1$
(resp. $\forall x$, $\forall y \in]0,1[$ tels que $x + y < 1$) :
$g_f(x + y) = g_f(x) + g_f(y)$

Démonstration :

Si $b_{j_1}, .., b_{j_\ell}$ sont des éléments de B, on a :

$$(5)\quad b_{j_1} \sim b_{j_2} \sim \ldots \sim b_{j_\ell} \iff \forall i = 1,..,n \quad p_A(i|j_1) = .. = p_A(i|j_\ell)$$

$$\iff \forall i = 1,..,n \quad \frac{p(i,j_1)}{p(.,j_1)} = \ldots = \frac{p(i,j_\ell)}{p(.,j_\ell)} = \frac{p(i,j_1)+..+p(i,j_\ell)}{p(.,j_1)+..+p(.,j_\ell)}$$

Donc, pour que des éléments de B soient à un écart nul, il faut et il suffit que les vecteurs colonne correspondants de la matrice des correspondances, soient proportionnels.
Prouvons alors la condition nécessaire.
Soit donc $f \in F_Q$ (resp. $f \in F$) telle que $d_f = d'_f$, $\forall C \in \mathcal{C}_Q$ (resp. $\forall C \in \mathcal{C}$), et soient x et y appartenant à $]0,1[\cap Q$ (resp. appartenant à $]0,1[$) tels que : $x + y < 1$.

Soient j_1 et j_2 ($j_1 \neq j_2$) deux indices compris entre 1 et p. Construisons $C \in \mathcal{C}_Q$ (resp. $C \in \mathcal{C}$) telle que (b_{j_1}, b_{j_2}) soit le seul couple d'éléments de B, distincts et situés à un écart nul, et telle que : $p(.,j_1) = x$, $p(.,j_2) = y$. Il suffit de considérer C telle que :

$$(6)\quad \begin{cases} \forall i = 1,\ldots,n \quad p(i,j_1) = \frac{x}{n},\ p(i,j_2) = \frac{y}{n} \\ \forall j = 1,\ldots,p \quad j \neq j_1,\ j \neq j_2 \quad p(1,j) = a \\ \forall j = 1,\ldots,p \quad j \neq j_1,\ j \neq j_2 \quad p(2,j) \text{ tous distincts et différents de a.} \\ (\text{si } n > 2)\ \forall j = 1,\ldots,p,\ j \neq j_1,\ j \neq j_2,\ \forall i > 2, p(i,j) = 0 \end{cases}$$

avec : $a \in Q_+^*$ (resp. $a \in \mathbb{R}_+^*$) tel que : $x + y + (p-2)\,a < 1$

$p(2,j) \in Q_+$ (resp. $\in \mathbb{R}_+$) pour $j \neq j_1$, $j \neq j_2$
(nous préciserons ces nombres plus loin).

Notant b'_r la classe d'équivalence de b_{j_1} (ou b_{j_2}) :

(5) donne : $\forall i = 1,..,n \quad \frac{p(i,j_1)}{p(.,j_1)} = \frac{p(i,j_2)}{p(.,j_2)} = \frac{p'(i,r)}{p'(.,r)} = \alpha_i$ (noté).

Alors l'égalité $d_f = d'_f$ entraîne :

$\forall i, \forall i' = 1,..,n : f\,[p'(.,r)]\,[p'_B(r|i) - p'_B(r|i')]^2$

$= f\,[p(.,j_1)]\,[p_B(j_1|i) - p_B(j_1|i')]^2 + f\,[p(.,j_2)]\,[p_B(j_2|i) - p_B(j_2|i')]^2,$

soit : $\forall i, \forall i' = 1,..,n : f\,[p'(.,r)]\,p'(.,r)^2\,\left[\frac{\alpha_i}{p(i,.)} - \frac{\alpha_{i'}}{p(i',.)}\right]^2$

$= f\,[p(.,j_1)]\,p(.,j_1)^2\,\left[\frac{\alpha_i}{p(i,.)} - \frac{\alpha_{i'}}{p(i',.)}\right]^2 + f\,[p(.,j_2)]\,p(.,j_2)^2\,\left[\frac{\alpha_i}{p(i,.)} - \frac{\alpha_{i'}}{p(i',.)}\right]^2$

Or, il est aisé de préciser C de telle sorte qu'il existe i et i' vérifiant :

$$\frac{\alpha_i}{p(i,.)} \neq \frac{\alpha_{i'}}{p(i',.)} .$$

Sur notre exemple $\alpha_1 = \alpha_2 = 1/n$. Si nous choisissons $a \in Q_+^*$ (resp. $\in \mathbb{R}_+^*$) tel que :

$(p-2)a < [1 - (x+y)]/2$ (par exemple $a = [1 - (x+y)]/[3(p-2)]$), on peut choisir

$p(2,j)$, $j \neq j_1$, $j \neq j_2$, tous distincts, appartenant à Q_+ (resp. à $\mathbb{R}_+$) et supérieurs à a

(par exemple égaux à : a+b,..., a+(p-2)b avec $\frac{(p-2)(p-1)b}{2} = 1 - (x+y) - 2\,(p-2)a$

$\in Q_+^*$ (resp. $\in \mathbb{R}_+^*$)) ;

et alors $p(1,.) < p(2,.)$

D'où : $f\,[p'(.,r)]\,p'(.,r)^2 = f[p(.,j_1)]\,p(.,j_1)^2 + f\,[p(.,j_2)]\,p(.,j_2)^2,$

soit $g_f\,[p'(.,r)] = g_f\,[p(.,j_1)] + g_f\,[p(.,j_2)]$

et donc : $g_f(x+y) = g_f(x) + g_f(y)$

Pour démontrer la condition suffisante, soit donc $f \in F_Q$ (resp. $f \in F$) telle que g_f vérifie la condition donnée dans le lemme. Un raisonnement par récurrence montre que g_f vérifie :

$\forall x_1,..x_\ell \in]0,1[\cap Q$ (resp.$]0,1[$), tels que $x_1 +...+ x_\ell < 1$,

$g_f(x_1 +..+ x_\ell) = g_f(x_1)+..+ g_f(x_\ell)$

Pour $C \in \mathcal{C}_Q$ (resp. $C \in \mathcal{C}$) soient $b'_1,..,b'_m$ les classes d'équivalence de B.

Si m = 1, tous les vecteurs colonne de B sont proportionnels deux à deux, et alors une simple vérification montre que : $\forall i, \forall i' = 1,..,n\ d_f^2(a_i, a_{i'}) = 0$. D'où $d_f = d'_f = 0$. (résultat vrai $\forall f \in F_Q$ (resp. $f \in F$)).

Si m > 1, soit r un indice compris entre 1 et m et soit $b'_r = \{b_{j_1}, .., b_{j_r}\}$

Alors par (5) : $\forall i = 1,..,n \quad \frac{p(i,j_1)}{p(.,j_1)} = \ldots = \frac{p(i,j_r)}{p(.,j_r)} = \frac{p'(i,r)}{p'(.,r)}$

Et donc : $\forall i, \forall i' = 1,..,n \quad \sum_{\substack{j \\ b_j \in b'_r}} f[p(.,j)] [p_B(j|i) - p_B(j|i')]^2$

$$= \sum_{\substack{j \\ b_j \in b'_r}} f[p(.,j)] \left[\frac{p'(i,r)}{p'(.,r)} \frac{p(.,j)}{p(i,.)} - \frac{p'(i',r)}{p'(.,r)} \frac{p(.,j)}{p(i',.)} \right]^2$$

$$= \frac{1}{p'(.,r)^2} [p'_B(r|i) - p'_B(r|i')]^2 \sum_{\substack{j \\ b_j \in b'_r}} g_f [p(.,j)]$$

$$= f [p'(.,r)] [p'_B(r|i) - p'_B(r|i')]^2 \quad \text{puisque} \quad \sum_{\substack{j \\ b_j \in b'_r}} p(.,j) < 1.$$

D'où $d_f = d'_f$.

Démonstration des propositions

Soit $f \in F_Q$ (resp. $f \in F^C$ ou $f \in F^D$) telle que $d_f = d'_f$ $\forall C \in \mathcal{C}_Q$ (resp. $\forall C \in \mathcal{C}$).

Utilisant le lemme, il est classique que g_f vérifie alors :

$\forall r \in Q_+^*$, $\forall x \in]0,1[\cap Q$ (resp. $\forall x \in]0,1[$) tels que $r x < 1$,

(7) $g_f(rx) = r g_f(x)$.

Dès lors pour la proposition 1, il suffit de choisir $x_0 \in]0,1[\cap Q$ et de poser $\alpha = \frac{g_f(x_0)}{x_0} > 0$ pour prouver le résultat.

Pour la proposition 2, si $f \in F^C$, utilisant (7) et la continuité de g_f on a :

(8) $\forall \lambda \in R_+^*$, $\forall x \in]0,1[$ tels que $\lambda x < 1$, $g_f(\lambda x) = \lambda g_f(x)$

Et le résultat en découle, comme dans la proposition 1.

Enfin, si $f \in F^D$, on montre que f vérifie :

(9) $\forall \lambda \in R_+^*$, $\forall x \in]0,1[$ tels que $\lambda x < 1$, $f(\lambda x) = f(x)/\lambda$.

En effet, $\forall \varepsilon > 0$, choisissant deux rationnels positifs r_1 et r_2 tels que :

(10) $x < 1/r_2 \leqslant 1/\lambda \leqslant 1/r_1$ et $1/r_1 - 1/r_2 \leqslant \varepsilon / f(x)$

On a, puisque f est non croissante : $f(r_2 x) \leqslant f(\lambda x) \leqslant f(r_1 x)$,

et donc par (7) : $f(x)/r_2 \leqslant f(\lambda x) \leqslant f(x)/r_1$

D'où avec (10) $\frac{f(x)}{\lambda} - \varepsilon \leqslant f(\lambda x) \leqslant \frac{f(x)}{\lambda} + \varepsilon$ et (9) en découle.

Dès lors il suffit de choisir $x_0 \in]0,1[$ et de poser $\alpha = x_0 f(x_0) > 0$ pour avoir le résultat annoncé.

REMARQUES

Remarque 1

Si $f \in F$ et si $\forall C \in \mathcal{C}$, $d_f = d'_f$, il est faux que f soit nécessairement de la forme $f(x) = \alpha/x$, $\forall x \in]0,1[$, $\alpha \in \mathbb{R}_+^*$. Pour un contre-exemple, il suffit de considérer une fonction numérique positive g définie sur $]0,1[$, satisfaisant à la condition du lemme, et n'étant pas de la forme $g(x) = \alpha x$, $x \in]0,1[$, $\alpha \in \mathbb{R}_+^*$.
Or la restriction à $]0,1[$, d'une forme linéaire h, définie sur $\mathbb{R}$, espace vectoriel de dimension infinie sur le corps Q, telle que h(x) ne soit pas de la forme $h(x) = \alpha x$, $\forall x \in]0,1[$, $\alpha \in \mathbb{R}_+^*$, est une telle fonction. Et l'existence de h est assurée par l'axiome du choix (si e_1 et e_2 sont deux éléments d'une base algébrique de l'espace considéré, que l'on peut toujours supposer appartenir à $]0,1[$ - quitte à les changer de signe et à les diviser par un rationnel qui leur est supérieur en module -, il suffit de définir h, par $h(e_1) = e_1$, $h(e_2) \neq e_2$ - h nulle, par exemple, sur les autres vecteurs de base - ; notons que cela revient à construire une forme linéaire non continue sur l'espace cité).

Remarque 2

Pour $p = 2$, on a : $\forall f \in F_Q$ (resp. $f \in F$), $\forall C \in \mathcal{C}_Q$ (resp. $\forall C \in \mathcal{C}$), $d_f = d'_f$.
En effet, nous avons constaté, dans la démonstration de la réciproque du lemme, que tel était le cas, s'il n'existait dans B qu'une seule classe d'équivalence.
Ainsi, pour $p = 2$, il existe une infinité de façons de choisir f, même si on lui impose d'être continue et monotone non croissante ; l'intérêt de la métrique d_f s'en trouve donc restreint. Certes, sur le plan concret, pour p=2, c'est l'analyse des correspondances elle-même qui présente peu d'intérêts ; ce n'est toutefois pas le cas si on considère le problème plus général d'une correspondance entre un ensemble A et plusieurs ensembles, chacun de cardinalité 2 - réponses OUI-NON dans ces ensembles - et que l'on munit A de la métrique dont le carré est la somme des carrés des métriques des correspondances entre A et chacun de ces ensembles - ce qui revient numériquement, à une homothétie près dans les résultats, à juxtaposer les tableaux de contingence -.

BIBLIOGRAPHIE

1 - BENZECRI *L'analyse des données*
Dunod 1973

2 - LEBART et FENELON *Statistique et informatique appliquées*
Dunod 1971