LES CAHIERS DE L'ANALYSE DES DONNÉES

B. ESCOFIER

Analyse de la différence entre deux mesures définies sur le produit de deux mêmes ensembles

Les cahiers de l'analyse des données, tome 8, n° 3 (1983), p. 325-329

<http://www.numdam.org/item?id=CAD_1983__8_3_325_0>

Les Cahiers de l'Analyse des Données
Vol. VIII - 1983 - n° 3 - p. 325-329

ANALYSE DE LA DIFFÉRENCE ENTRE DEUX MESURES DÉFINIES SUR LE PRODUIT DE DEUX MÊMES ENSEMBLES

[ANA. DIFF. PROD.]

par B. Escofier (1)

0 *Position du problème* : Un tableau de fréquence k_{IJ} croisant deux ensembles I et J définit sur le produit IxJ une mesure de probabilité notée f_{IJ}. L'analyse des correspondances de k_{IJ} permet d'analyser, ligne par ligne,et, colonne par colonne, la différence entre f_{IJ} et le produit $f_I \otimes f_J$ des mesures marginales sur I et J. Le produit $f_I \otimes f_J$ est en quelque sorte un modèle qui correspond à l'hypothèse d'indépendance.

Nous généralisons cette analyse à la comparaison de deux mesures de probabilité quelconques f_{IJ} et f'_{IJ} définies sur IxJ. En correspondance, les marges f_I et f_J sont les centres des métriques du χ^2 définies sur R_J et sur R_I ; elles servent aussi de pondération. Nous prenons comme centre des métriques et comme pondération deux mesures de probabilité strictement positives quelconques g_I et g_J (définies respectivement sur I et sur J).

1 *Les nuages N(I) et N(J)*

1.1. *Le nuage N(I)*

Le nuage N(I) est situé dans l'espace des mesures sur J, noté R_J. L'élément i du nuage N(I) est le point de coordonnée $(f_{iJ}/g_i - f'_{iJ}/g_i)$. Il représente la différence entre les deux lignes de même indice i de chacun des tableaux f_{IJ} et f'_{IJ}, quotientées par le terme g_i. Le poids g_i est affecté au point i :

$$N(I) = \{(((f_{iJ}-f'_{iJ})/g_i) \ ; \ g_i) \ | i \in I\}$$

L'espace R_J est muni de la métrique du χ^2 de centre g_J ; le carré de la distance du point i à l'origine des axes vaut donc $\sum\{(f_{ij}-f'_{ij})/g_i)^2/g_j \ | j \in J\}$ et son inertie s'écrit $\sum\{(f_{ij}-f'_{ij})^2/(g_i g_j) \ | j \in J\}$

Cas où $\delta_I = \delta'_I = g_I$

Si les marges sur I de f_{IJ} et de f'_{IJ} sont égales et que l'on choisisse pour g_I cette valeur commune, le point i représente la différence entre les profils des lignes des deux tableaux i.e. les deux mesures de probabilités conditionnelles induites par le même élément i sur f_{IJ} d'une part et f'_{IJ} d'autre part. Sinon, l'une des mesures au moins (f_{iJ}/g_i ou f'_{iJ}/g_i) a une valeur différente de 1 sur l'ensemble J et le vecteur $\vec{oi}$ ne représente pas la différence entre les profils des lignes des deux tableaux, mais entre les lignes elles-mêmes (à l'homothétie $1/g_i$ près).

(1) *Chercheur CNRS - IRISA RENNES*

Cas où $\delta_J = \delta'_J = g_J$

Si les marges sur J, f_J et f'_J des mesures f_{IJ} et f'_{IJ} sont égales, le nuage N(I) est centré : $\sum\{g_i(f_{iJ}/g_i - f'_{iJ}/g_i) \mid i \in I\} = f_J - f'_J$. D'autre part, l'importance relative de la différence $f_{ij}/g_i - f'_{ij}/g_i$, (j-ème coordonnée de i) dépend de la valeur moyenne sur J de f_{ij}/g_i et de f'_{ij}/g_i. Or, les marges des tableaux, f_J et f'_J sont les moyennes des mesures f_{iJ}/g_i définies sur J par l'ensemble de leurs lignes : $\sum\{(f_{iJ}/g_i)g_i \mid i \in I\} = f_J$. Si ces deux marges sont égales, on pourra choisir g_J (qui définit la métrique de R_J) égale à cette valeur commune. Ainsi, dans l'expression de la distance du χ^2, entre deux points du nuage N(I), le terme correspondant à la coordonnée j sera quotienté par cette valeur moyenne.

Cas où $\delta'_{IJ} = \delta_I \otimes \delta_J$

Si f'_{IJ} est la mesure produit $f_I \otimes f_J$, alors $f'_I = f_I$ et $f'_J = f_J$. Si de plus, on pose $g_I = f_I$ et $g_J = f_J$, le nuage N(I) est exactement celui qui est considéré dans l'analyse des correspondances du tableau f_{IJ}. Les mesures définies par les lignes de f'_{IJ} sont toutes égales à leur moyenne f_J, et soustraire cette mesure du profil f_{iJ}/f_i ne fait que placer l'origine des axes au centre de gravité du nuage de ces profils.

1.2. *Le nuage N(J)*

La définition et les propriétés de N(J) se déduisent de celle de N(I) en échangeant les indices i et j.

2 *Analyse des nuages N(I) et N(J)*

Le nuage N(I) représente la différence entre les deux mesures f_{IJ} et f'_{IJ} ligne par ligne. Un point i est d'autant plus éloigné de l'origine que les deux mesures f_{iJ}/g_i et f'_{iJ}/g_i diffèrent ; deux points i et i' sont proches (resp. opposés) si la différence entre f_{IJ} et f'_{IJ} est analogue (resp. opposée) sur les deux lignes. La projection de ce nuage sur ses premiers axes d'inertie (calculés à l'origine des axes) en donne une image approchée qui permet de comparer facilement le comportement des deux mesures sur l'ensemble des lignes. Dans le cas où $f_I = f'_I = g_I$ les mesures définies par chaque ligne sont des probabilités conditionnelles et leur interprétation est claire.

De même, la projection du nuage N(J) sur ses premiers axes d'inertie, permet de comparer les mesures f_{IJ} et f'_{IJ} colonne par colonne.

Pour montrer la dualité de ces deux analyses, considérons, au lieu de N(I) et de N(J) leurs images $N(I)$ et $N(J)$ par les isométries définies par les métriques de R_J et de R_I dans leurs duaux R^J et R^I. Ces espaces sont munis respectivement des métriques diagonales $\delta_j{}^{j'} g_j$ et $\delta_i{}^{i'} g_i$. La j-ème coordonnée du point i de $N(I)$ est égale à la i-ème coordonnée du point j de $N(J)$ et vaut :

$$x_{ij} = (f_{ij} - f'_{ij})/(g_i g_j)$$

Notons X la matrice (IxJ) de terme général x_{ij}, X' sa transposée, D_I la matrice diagonale de terme général $\delta_i^{i'} g_i$ et D_J la matrice diagonale de terme général $\delta_j{}^{j'} g_j$. Les coordonnées des projections de $N(I)$

(ou de N(I)) sur ses axes d'inertie à l'origine, que nous notons F_s et que nous appelons facteurs sur I sont les vecteurs propres du produit

$$X\ D_J\ X'\ D_I$$

De même, les facteurs sur J, notés G_s, sont les vecteurs propres de :

$$X'D_I\ X\ D_J$$

Ces deux matrices se décomposent sous la forme du produit (à gauche ou à droite) de $X'D_I$ par $X\ D_J$. Leurs valeurs propres sont donc égales et leurs vecteurs propres se déduisent les uns des autres en appliquant $X'D_I$ ou $X'D_J$. Notons λs la valeur propre d'ordre s de ces matrices, i.e. l'inertie de la projection de N(I) ou de N(J) sur leur s-ième axe d'inertie à l'origine : $\lambda s = \sum\{f_i(F_s(i))^2 \mid i\in I\} = \sum\{g_j(G_s(j))^2 \mid j\in J\}$. Les formules de transition qui permettent de déduire les facteurs F_s et G_s les uns des autres s'écrivent :

$$F_s(i) = \lambda_s^{-1/2} \sum\{((f_{ij}-f'_{ij})/g_i)\ G_s(j) \mid j\in J\}$$

$$G_s(j) = \lambda_s^{-1/2} \sum\{((f_{ij}-f'_{ij})/g_j)\ F_s(i) \mid i\in I\}$$

La dualité des deux analyses permet de représenter simultanément l'ensemble des lignes et l'ensemble des colonnes. Un élément i est situé (en moyenne) du côté des éléments j de J auxquels il s'associe plus dans f_{IJ} que dans f'_{IJ} et à l'opposé des éléments j auxquels il s'associe moins dans f_{IJ} que dans f'_{IJ}. Si $f_J=f'_J$, le nuage N(I) et les facteurs F_s sont centrés. Si $f_I=f'_I$ le nuage N(J) et les facteurs G_s sont centrés.

3 *Reconstitution des données*

Les facteurs G_s sont, dans l'espace R^J, des vecteurs directeurs des axes d'inertie du nuage $N(I)$. Ecrivons les coordonnées d'un point i de $N(I)$ dans la base orthonormée de R^J définie par ces axes.

$$x_{ij} = \sum_s \{\lambda s^{-1/2}\ F_s(i)G_s(j)\}$$

$$\text{D'où } f_{ij}-f'_{ij} = g_i g_j \sum_s \{\lambda s^{-1/2}\ F_s(i)G_s(j)\}$$

Les facteurs F_s et G_s et les mesures g_I et g_J permettent de reconstituer la différence entre les deux mesures f_{IJ} et f'_{IJ}.

4 *Cas particulier* $f_I=f'_I=g_I$ *et* $f_J=f'_J=g_J$

Le cas où les deux égalités $f_I=f'_I$ et $f_J=f'_J$ sont vérifiées est particulièrement intéressant, si l'on pose, ce qui est naturel, $g_I=f_I=f'_I$ et $g_J=f_J=f'_J$. En effet, les points des nuages N_I(resp.N_J) représentent alors les différences entre les profils des lignes (resp.colonnes), le centre de la distance du χ^2 est la valeur moyenne de ces profils et les deux nuages sont centrés. (Si f'_{IJ} est le produit $f_I \otimes f_J$ des marginales, on obtient exactement l'analyse des correspondances). De plus, les résultats peuvent être obtenus avec un programme classique d'analyse des correspondances en traitant le tableau $h_{IJ} = f_{IJ} - f'_{IJ} + f_I \otimes f_J$.

Démonstration

Certains termes du tableau h_{IJ} peuvent être négatifs, mais ses marges, qui sont égales respectivement à f_I et f_J, sont positives et les programmes d'analyse des correspondances s'appliquent. Le "profil" d'une ligne i de h_{IJ} est : $(f_{iJ}-f'_{iJ})/f_i+f_J$. Le nuage de ces profils affectés des poids f_i a pour centre de gravité le point f_J. Le nuage centré, considéré en analyse des correspondances est donc confondu avec le nuage N(I) défini au § 2. La métrique de R_J est égale à f_J dans les deux analyses. De même, le nuage des profils des colonnes est confondu avec N(J). L'analyse des correspondances du tableau h_{IJ} donne donc les résultats de l'analyse de la différence entre les mesures f_{IJ} et f'_{IJ}.

5 *Eléments supplémentaires*

Des lignes ou des colonnes supplémentaires peuvent être prises en compte et projetées sur les axes d'inertie des nuages N(I) et N(J). Dans le cas général, les facteurs sur J, G_s sont calculés en diagonalisant la matrice $X'D_I X D_J$ (cf.§2) ; les facteurs sur l'ensemble I, et sur l'ensemble des lignes supplémentaires s'en déduisent par la première formule de transition. Les valeurs des facteurs pour les colonnes supplémentaires sont déduites de F_s par la deuxième formule. Naturellement, la mesure g_I (resp.g_J) doit être définie sur les lignes(resp. les colonnes) supplémentaires. Dans le cas particulier du §4, le programme d'analyse des correspondances classiques permet de calculer aussi les valeurs des facteurs pour les éléments supplémentaires, à condition qu'ils vérifient la même propriété que les éléments principaux : pour chaque ligne supplémentaire, les sommes f_i et f'_i doivent être égales et on pose $k_{ij}=(f_{ij}-f'_{ij})+ f_i f_j$. On met souvent en ligne supplémentaire la somme de plusieurs lignes pour avoir les valeurs des facteurs et des contributions de leur centre de gravité, cette propriété est bien entendu vérifiée encore pour le tableau h_{IJ}. Pour les colonnes, la situation est identique.

6 *Interprétation de l'analyse dans* R_{IJ}

En plus de la comparaison ligne par ligne et colonne par colonne des deux mesures f_{IJ} et f'_{IJ}, les résultats obtenus permettent de décomposer la distance du χ^2 entre ces deux mesures.

Plaçons-nous dans l'espace R_{IJ} des mesures définies sur le produit IxJ et munissons cet espace de la métrique diagonale $\delta^{i'j'}_{ij}/g_i g_j$. La distance entre deux mesures est alors la distance du χ^2 de centre $g_I \otimes g_J$ et la norme du vecteur $f_{IJ}-f'_{IJ}$ a pour valeur cette distance :

$$D^2_{g_I \otimes g_J}(f_{IJ}, f'_{IJ}) = \sum\{(f_{ij}-f'_{ij})^2/g_i g_j \mid i\in I,\ j\in J\}$$

On retrouve dans cette expression l'inertie du nuage N(I)(resp. N(J)), qui a été décomposé sur ses axes d'inertie.

L'application de R^I dans son dual R_I permet d'associer à chaque facteur F_s, une mesure sur I notée $g_I F_s$ qui vaut $g_i F_s(i)$ pour l'élément i. De même les facteurs G_s définissent des mesures sur J notées $g_I G_s$. Les facteurs F_s(resp. G_s) étant orthogonaux pour la métrique g_I (resp.g_J), les produits $g_I F_s \otimes g_J G_s$ de ces mesures forment un système orthogonal dans R_{IJ} :

$$\sum\{(1/g_i g_j)(g_i F_s(i) g_j G_s(j))(g_i F_{s'}(i) g_j G_{s'}(i) \mid i\in I,\ j\in J\} = \delta^{s'}_{s} \lambda_s^2$$

Le vecteur $f_{IJ}-f'_{IJ}$ se decompose sur ce système orthonormé des $\{g_I F_s \otimes g_J G_s/\lambda_s\}$ avec les coordonnées $(\lambda_s)^{1/2}$. (On le déduit facilement de la formule de reconstitution des données). Donc $D^2_{g_I \otimes g_J}(f_{IJ}, f'_{IJ})$ se décompose en une somme de carrés de distances entre des éléments de R_{IxJ} qui sont des produits d'une mesure sur I et d'une mesure sur J.

7 *Applications*

Nous avons déjà souligné l'intérêt du cas où les marges des deux mesures f_{IJ} et f'_{IJ} sont égales. Lorsque ces marges diffèrent, le choix des mesures g_I et g_J ne s'impose pas. On peut choisir, par exemple, g_I égal à f_I, à f'_I ou à leur demi-somme. La multiplicité de ces possibilités ne facilite pas l'interprétation des résultats. De plus, si les marges sur I sont très différentes, le centre de gravité du nuage N(J) qui a pour coordonnées $f_I - f'_I$ est très éloigné de l'origine. Les axes d'inertie de N(J) étant calculés à l'origine, on risque alors de mettre en évidence seulement l'écart entre les marges des deux tableaux et non les écarts entre chacune de leurs colonnes. L'application de la méthode à la comparaison de deux mesures définit par deux tableaux de fréquence risque donc d'être assez difficile si les marges des deux tableaux diffèrent beaucoup. Par contre, la méthode s'est montrée efficace en prenant pour f'_{IJ} un "modèle" -différent du modèle simple d'indépendance- dont les marges sont égales à celle de f_{IJ}. Citons deux exemples issus d'une correspondance ternaire, f_{IJT} définie sur le produit de trois ensembles IxJxT. Dans le premier exemple détaillé et illustré dans |ANA.PLUS.TAB|, nous comparons f_{IJT}, considérée comme une mesure sur le produit de (IxT) par J et notée, alors $f_{(IxT)J}$ au modèle $f'_{(IxT)J}$ défini par :

$f'_{itj} = f_{ij} f_{it}/f_i$ où f_{ij}, f_{it}, et f_i désignent les marges de f_{IJT}.

Ce modèle est celui de l'indépendance entre J et T pour chaque élément i de I : chaque sous-tableau $f'_{(ixT)J}$ est le produit des marges du sous tableau $f_{(ixT)J}$. Les éléments (i,t) du nuage des lignes représentent la différence entre le profil f_{iJt}/f_{it} et la moyenne sur T de ces profils f_{iJ}/f_i. Dans le second exemple, on compare la marge f_{JT} de f_{IJT} au modèle défini par $f'_{jt} = \{\sum f_i (f_{ij}/f_i)(f_{it}/f_i) \mid i \in I\}$. La mesure f'_{JT} est la moyenne sur i des produits des marginales f_{iJ}/f_i et f_{iT}/f_i ; c'est la marge de la mesure $f'_{(IxT)J}$ précédente. C'est un modèle d'indépendance "en moyenne sur I" entre J et T. Cette technique permet donc d'étudier des tableaux de fréquences en référence à toutes sortes de modèles adaptés aux problèmes posés.