

STATISTIQUE ET ANALYSE DES DONNÉES

Y. ESCOUFIER

Objectifs et procédures de l'analyse conjointe de plusieurs tableaux de données

Statistique et analyse des données, tome 10, n° 1 (1985), p. 1-10

http://www.numdam.org/item?id=SAD_1985__10_1_1_0

© Association pour la statistique et ses utilisations, 1985, tous droits réservés.

L'accès aux archives de la revue « Statistique et 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/>

OBJECTIFS ET PROCEDURES DE L'ANALYSE CONJOINTE
DE PLUSIEURS TABLEAUX DE DONNEES

Y. ESCOUFIER

Laboratoire de Biométrie
INRA - ENSA - USTL
34060 MONTPELLIER CEDEX

Les travaux du groupe de travail "Analyse Conjointe de Plusieurs Tableaux de Données : données évolutives" ont conduit à la rédaction de plusieurs textes présentés par leurs auteurs dans les pages suivantes de ce volume. Chaque texte décrit une méthode et les résultats qu'elle fournit sur le même ensemble de données dont la présentation est faite après cette introduction.

Il a paru nécessaire d'aider les lecteurs à situer les différents points de vue les uns par rapport aux autres et également par rapport à d'autres points de vue qui, pour n'avoir pas été présentés dans le groupe de travail, n'en méritent pas moins d'être considérés.

Pour ce faire nous tentons de dresser la liste des objectifs généraux que les différentes méthodes poursuivent et de reconnaître des ressemblances dans les procédures qu'elles utilisent.

Ces objectifs généraux et ces groupes de procédures seront autant de repères pour une compréhension comparative des différentes méthodes.

Les deux paragraphes qui suivent sont construits de façon homologue. A l'objectif énoncé dans le sous-paragraphe 1.i correspondent les procédures décrites dans le sous-paragraphe 2.i. Le lecteur pourra ainsi préférer une lecture transversale des deux paragraphes.

1 - LES OBJECTIFS GENERAUX

Le but de ce paragraphe est de dresser la liste des objectifs qu'une méthode d'Analyse Conjointe de Plusieurs Tableaux peut chercher à satisfaire. Il paraît nécessaire de faire une place spéciale aux situations dans lesquelles un ordre naturel, le plus souvent lié au temps, s'impose dans l'énumération des tableaux : dans ce cas, en effet, il est possible de parler de l'évolution des tableaux successifs ce qui enrichit la liste des objectifs à envisager.

1.1

A un tableau de données le statisticien sait associer un ensemble de représentations graphiques qui lui permettra de reconnaître et de montrer les tendances essentielles des données : ce sera un ensemble de plans factoriels, un ensemble de partitions imbriquées ou tout autre description.

Lorsqu'il dispose de plusieurs tableaux de données, le statisticien peut envisager de les étudier un à un. Il risque alors d'être submergé par le nombre des ensembles de représentations à analyser. Il peut également penser que chacun des tableaux ne donne qu'une vision partielle de la réalité. Ces deux raisons le conduiront à chercher un ensemble unique de représentation dont nous dirons qu'il constitue une synthèse, ou un résumé global, ou mieux un compromis de l'ensemble des tableaux.

1.2

Des individus peuvent être comparés sur la base des valeurs prises sur un ensemble de variables ; des variables peuvent être comparées sur la base des valeurs prises par un ensemble d'individus ; de même des situations expérimentales différentes, des conditions d'observation différentes pourront être comparées sur la base des tableaux de données qu'elles ont permis d'acquérir.

A condition de disposer d'une notion de proximité entre deux tableaux, le statisticien pourra envisager de rechercher un ensemble de représentations graphiques aptes à montrer les tableaux voisins.

Dans cette démarche, chacun des tableaux est assimilé à un objet. Le but est de comparer ces objets entre eux, d'y reconnaître des groupes. Nous dirons que l'objectif poursuivi est celui d'une analyse globale : On reconnaît des tableaux différents sans pouvoir donner une description fine des éléments qui les différencient.

1.3

Il va de soi, que l'analyse globale laisse le statisticien sur sa faim et que son souhait est de pouvoir mettre en évidence les raisons des différences entre tableaux.

Cet objectif est satisfait par des procédures qu'on peut appeler d'analyse fine qui permettront d'étudier les ressemblances entre les éléments correspondants des différents tableaux : on étudiera les modifications de la ressemblance entre deux lignes d'un tableau à l'autre ; les modifications de la ressemblance de deux colonnes.

1.4

Le problème que nous évoquons prend une signification nouvelle quand le temps indique les situations expérimentales ou les conditions d'observation qui ont permis d'acquérir les tableaux de données.

L'objectif naturel est alors de chercher à décrire l'évolution du phénomène. Il pourra être limité à une description de l'évolution globale du phénomène ou au contraire avoir l'ambition d'étudier les évolutions de chacun des éléments composant les tableaux. On parlera alors de l'analyse des trajectoires de ces éléments.

1.5

Dans ce même contexte temporel, l'objectif peut être davantage de reconnaître les évolutions. L'accent sera alors mis sur l'analyse des transitions, c'est-à-dire sur la reconnaissance des modifications que subit le tableau disponible à un instant pour engendrer son successeur.

2 - LES PROCEDURES

Une méthode d'Analyse Conjointe de Plusieurs Tableaux est un ensemble de procédures mathématiques, informatiques et graphiques agrégées pour satisfaire, pour un type de données particulier, un ou plusieurs des objectifs énumérés précédemment. La reconnaissance de similitude entre les procédures est intéressante car les raisons qui ont présidé à leurs choix peuvent parfois être mises en question : on voit alors naître des possibilités d'enrichissement des méthodes par emprunt mutuel.

Afin de ne pas alourdir cette introduction, nous avons fait le choix d'énumérer des procédures envisageables sans renvoyer aux autres textes de ce volume ou à d'autres travaux cités en bibliographie.

2.1 - Les procédures de compromis

Soit $\{X_k ; k = 1, \dots, K\}$ l'ensemble des tableaux disponibles. Nous noterons $\{T(X_k) ; k = 1, \dots, K\}$ un ensemble de tableaux déduits des précédents et tels que tous les $T(X_k)$ aient les mêmes dimensions, ce qui peut ne pas être le cas des X_k .

Soit $\{\alpha_k ; k = 1, \dots, K\}$ un système quelconque de pondérations positives. Alors $\sum_{k=1}^K \alpha_k T(X_k)$ est un compromis possible dont la signification dépend des valeurs des α_k et de la nature des $T(X_k)$.

Selon les cas, les $T(X_k)$ pourront être égaux :

- Aux X_k quand ceux-ci sont des tableaux de contingence concernant les deux mêmes variables qualitatives.
- Aux tableaux X_k eux-mêmes quand ceux-ci sont des tableaux de similarité concernant les mêmes individus ou des tableaux de dissimilarité .
- Aux X_k ou aux tableaux \tilde{X}_k centrés associés quand les X_k sont des tableaux individus x caractères de mêmes dimensions.
- Aux tableaux des produits scalaires $X_k M_k {}^t X_k$ si les X_k concernent les mêmes individus (M_k est une métrique permettant de calculer les distances entre les individus).

En posant alors $\mathcal{X}_1 = (\sqrt{\alpha_1} X_1 ; \dots ; \sqrt{\alpha_K} X_K)$

$$\text{et } M = \left(\begin{array}{c|ccc|c} M_1 & & & & 0 \\ \hline & M_2 & & & \\ & & \dots & & \\ \hline 0 & & & & M_k \end{array} \right)$$

$$\text{On a : } \sum_{k=1}^K \alpha_k X_k M_k {}^t X_k = \mathcal{X}_1 M {}^t \mathcal{X}_1$$

- Aux matrices de variance ${}^t \tilde{X}_k D_k \tilde{X}_k$ si les X_k concernent les mêmes variables (D_k matrice diagonale de pondération des individus nécessaire au calcul des covariances).

En posant alors

$${}^t X_2 = (\sqrt{\alpha_1} \quad {}^t \tilde{x}_1 \quad \vdots \quad \vdots \quad \sqrt{\alpha_K} \quad {}^t \tilde{x}_K)$$

$$\text{et } D = \begin{pmatrix} D_1 & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & D_K \end{pmatrix}$$

$$\text{On a : } \sum_{k=1}^K \alpha_k {}^t \tilde{x}_k D_k \tilde{x}_k = {}^t X_2 D X_2$$

Considérons alors le produit scalaire usuel sur l'ensemble des matrices $n \times p$: $\langle A, B \rangle = \text{Tr} (A^t B)$ et la norme $\| \cdot \|$ qui en découle.

Lorsque les α_k sont égaux entre eux, le compromis a la propriété usuelle d'une moyenne et minimise $\sum_{k=1}^K \| T(X_k) - \sum_{\ell=1}^K \alpha_\ell T(X_\ell) \|^2$.

On peut considérer aussi les α_k qui maximisent :

$$\sum_{k=1}^K \left[\langle T(X_k), \sum_{\ell=1}^K \alpha_\ell T(X_\ell) \rangle^2 \right]$$

Dans le cas où les $T(X_k)$ sont des tableaux de similarité entre les mêmes objets, on peut chercher un tableau individus \times caractères X et des métriques diagonales Δ_k tels que

$$\sum_{k=1}^K \| T(X_k) - X \Delta_k {}^t X \|^2 \text{ soit minimum.}$$

On peut remplacer les Δ_k par des métriques M_k quelconques et utiliser le même critère. On voit le lien avec les choix précédents.

Quand les $T(X_k)$ sont des tableaux de dissimilarité entre les mêmes objets, on peut envisager de rechercher l'ultramétrie ou la partition de distances associée D telle que

$$\sum_{k=1}^K \| T(X_k) - D \|^2 \text{ soit minimum.}$$

2.2 - Les procédures d'analyse globale

Nous voulons parler ici des procédures visant à reconnaître les ressemblances globales entre tableaux.

Une première famille de possibilités consiste à associer à chaque tableau un élément descriptif global, puis à étudier la ressemblance de ces éléments. Quand les tableaux sont des tableaux individus x caractères, on peut par exemple leur associer un individu moyen.

L'analyse globale sera l'analyse des proximités de ces individus moyens.

Quand les tableaux sont des tableaux de contingence, il est assez naturel de leur associer leurs marges. L'analyse des marges fournit alors une information globale sur la ressemblance des tableaux.

L'analyse des moyennes ou des marges peut être menée de plusieurs manières : on peut envisager d'en faire une étude spécifique ; il est commode d'en faire l'étude en les projetant comme points supplémentaires dans les représentations fournies par le compromis.

Une autre approche devient nécessaire quand on choisit de construire la comparaison globale sur les tableaux eux-mêmes et non sur les résumés limités que sont les moyennes et les marges. Cette approche s'impose d'ailleurs quand les tableaux à comparer sont des tableaux pour lesquels les notions de moyennes et de marges non pas de sens : tableaux de similarité, matrices de variance par exemple.

La démarche usuelle de l'Analyse des Données conduit alors à bâtir une matrice \mathcal{C} , $K \times K$, dont l'élément \mathcal{C}_{kl} est la mesure d'une similarité ou d'un produit scalaire entre les tableaux $T(X_k)$ et $T(X_l)$. \mathcal{C} pourra être alors étudiée soit par des techniques de classification, soit par des techniques de positionnement multidimensionnel qui permettront de reconnaître les tableaux proches au sens de la similarité choisie.

Les méthodes qui substituent $X \Delta_k {}^t X$ à $T(X_k)$ représentent en général $T(X_k)$ par un point dont les coordonnées sont les éléments diagonaux de Δ_k . On remarquera que ceci revient à comparer $T(X_k)$ et $T(X_l)$ par leurs approximations $X \Delta_k {}^t X$ et $X \Delta_l {}^t X$ pour le produit scalaire

$$\text{Tr} (X \Delta_k {}^t X X \Delta_l {}^t X) = \text{Tr} (\Delta_k \Delta_l) \quad \text{si} \quad {}^t X X = I.$$

2.3 - Les procédures d'analyse fine

L'idée est ici d'étudier le comportement d'éléments particuliers des tableaux soit en mettant en évidence leurs stabilités sur l'ensemble des tableaux soit en décrivant leurs instabilités. Ces éléments pourront être les lignes ou les colonnes des tableaux originaux ou des descripteurs déduits des tableaux initiaux.

Lorsque par exemple les tableaux X_k sont des tableaux individus \times caractères, on souhaitera étudier les lignes et les colonnes des tableaux eux-mêmes mais aussi les composantes principales qu'on a pu déduire de chaque tableau.

Quand les tableaux X_k sont des tableaux de contingence, lignes et colonnes sont intéressantes par leurs profils. Il est souhaitable de pouvoir les comparer.

Penchons-nous pour le moment sur le cas où les tableaux initiaux sont des tableaux individus \times variables. Lorsque les choix ont conduit à une solution compromis du type $X_1 M^t X_1$, une représentation compromis des lignes (individus) des différents tableaux est classiquement fournie par les premiers vecteurs propres de $X_1 M^t X_1$. Les colonnes des tableaux initiaux (les variables) peuvent alors être représentées par le moyen de leurs corrélations avec ces vecteurs propres. On peut faire la même chose pour les composantes principales des tableaux initiaux.

La représentation des lignes des tableaux est moins immédiate : on peut exploiter toutefois que ce sont les mêmes individus qui sont décrits par les lignes des tableaux et par les vecteurs propres de $X_1 M^t X_1$.

Des procédures inspirées de la régression sont alors envisageables.

Pour ce type de données, la même approche pourrait être suivie dans le cas où on a substitué $X \Delta_k^t X$ ou $X M_k^t X$ à $T(X_k)$. X prendra la place des vecteurs propres de $X_1 M^t X_1$. On obtiendra ainsi, entre autres, une analyse fine des écarts au modèle utilisé.

Cette approche est transposable lorsque le compromis choisi est ${}^t X_2 D X_2$. Seule la représentation des colonnes des tableaux initiaux pose un problème.

Quand les tableaux initiaux sont des tableaux de contingence, un compromis fourni par une moyenne de ces tableaux permet de représenter les lignes et les colonnes des tableaux initiaux en éléments supplémentaires.

2.4 - Les procédures d'analyses d'évolutions globales et de trajectoires

Quand les tableaux sont indicés par le temps, les méthodes doivent être interrogées sur la part qu'elles font à cet indice.

Pour certaines méthodes, les résultats ne dépendent pas de l'ordre dans lequel les tableaux sont soumis à l'étude. C'est dire que le temps n'intervient que comme élément d'interprétation. Il est absent des calculs.

Lorsque l'étude globale se fait par l'intermédiaire d'un élément descriptif synthétique (moyenne en particulier), on peut envisager d'ajuster un modèle temporel à l'évolution de ce descripteur. On étudie ensuite les écarts aux modèles.

Lorsque les données s'y prêtent, on peut déduire des K tableaux initiaux, $(K-1)$ tableaux d'accroissements successifs qui seront soumis aux procédures d'analyses globales et fines décrites précédemment. Plus radicalement encore, on pourra substituer à la recherche du compromis celle d'une représentation qui sauvegarde au mieux la reconstitution de l'évolution de chacun des individus.

Suivant les travaux de L. LEBART [5], ces approches assimilent le fait que des données concernent un même individu à une contrainte de contiguïté temporelle. Pratiquement dans l'approche de compromis qui a été schématisée par ${}^t\mathcal{X}_2 \text{ D } \mathcal{X}_2$, elles reviennent à substituer à D une matrice non diagonale qui traduit les liens d'une donnée avec la donnée suivante observée sur le même individu.

2.5 - Les procédures d'analyse des transitions

Le passage d'un tableau X_k à son successeur X_{k+1} est ici assimilé à une déformation subie par X_k . Le problème est de décrire cette déformation.

Lorsque les tableaux initiaux sont des tableaux individus x variables, la méthode Procruste permet de rechercher les transformations orthogonales

H_k telles que $\sum_{k=1}^K ||X_k H_k - X_{k+1}||^2$ soit minimum.

On peut envisager des variantes à ce problème. On pourrait vouloir trouver H orthogonale telle que $\sum_{k=1}^K ||X_k H - X_{k+1}||^2$ soit minimum ou bien des M_k , inversibles et non nécessairement orthogonales telles que

$\sum_{k=1}^K ||X_k M_k - X_{k+1}||^2$ soit minimum.

Quand les tableaux sont des tableaux de contingence, la méthode RAS (FROMENT, [3]) inspire une procédure qui consiste à chercher des matrices R_k et S_k telles que $\sum_{k=1}^K ||R_k X_k S_k - X_{k+1}||^2$ soit minimum.

Les transformations H_k , R_k , S_k peuvent ensuite devenir objets d'étude pour décrire la stabilité ou l'évolution des transitions successives.

Références bibliographiques

BOUROCHE, J.M., "Analyse des données ternaires : la double analyse en composantes principales", Thèse 3ème cycle, Univ. Paris VI, 1975.

ESCOFIER, B. et PAGES, J., "L'analyse factorielle multiple", cahiers du BUR0, 1984, n° 42.

- [3] FROMENT, R., "Optimisation d'un tableau rectangulaire dont les marges sont connues", Annales de l'INSEE, 1972, n° 9.

JAFFRENOU, P.A., "Sur l'Analyse des familles finies de variables vectorielles", Thèse de 3ème cycle, Univ. Claude Bernard, Lyon I, 1978.

- [5] LEBART L., "Analyse statistique de la contiguité", Thèse de 3ème cycle, Paris, 1966.

LE FOLL, Y., "L'analyse factorielle des évolutions", Annales de l'Univ. des Sciences Sociales de Grenoble, 1972, n° 2, pp. 23-27.

LE FOLL, Y., "Pondération des distances en analyse factorielle," Statistique et Analyse des Données, 1982, Vol. 7, n° 1, pp. 13-31.

LE FOLL, Y. et BURTSCHY, B., "Représentations optimales des matrices imports-exports", RSA, 1983, Vol. XXXI, n° 3, pp. 41-56.

LE NOUVEL, J., "Etude d'une famille de courbes par des méthodes d'analyse des données - Application à l'analyse morphologique de courbes provenant de données médicales", Thèse de 3ème cycle, Univ. de Rennes, 1981.

SAPORTA, G., "Méthodes exploratoires d'analyse de données temporelles", Thèse d'Etat, Univ. Pierre et Marie Curie, Paris VI, 1981.

STEMMELEN, E., "Tableaux d'échanges : description et prévision", Cahiers du BUR0, 1977, n° 28, Paris.

TENENHAUS M. et PRIEURET, B., "Analyse des séries chronologiques multidimensionnelles", R.A.I.R.O., 1974, Vol. 5, n° 2, pp. 5-16.

THORSSON, H., "Une contribution à l'analyse statistique des tableaux d'échanges", Thèse de 3ème cycle, Montpellier, 1981.