Statistique et analyse des données

Gilbert Saporta

Quelques applications des opérateurs d'Escoufier au traitement des variables qualitatives

Statistique et analyse des données, tome 1, n° 1 (1976), p. 38-46

<http://www.numdam.org/item?id=SAD_1976__1_1_38_0>

QUELQUES APPLICATIONS DES OPERATEURS D'ESCOUFIER AU TRAITEMENT DES VARIABLES QUALITATIVES.

Gilbert SAPORTA *

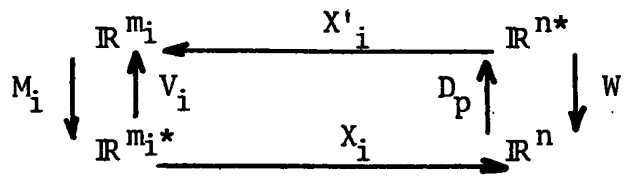
Les opérateurs introduits par Y. ESCOUFIER permettent de représenter par un être mathématique unique un ensemble de variables. En munissant l'espace des opérateurs d'un produit scalaire et d'une norme, on peut alors mesurer des dépendances globales entre groupes de variables. Ces opérateurs peuvent donc être utilisés pour décrire des proximités entre variables qualitatives car toute variable qualitative est équivalent à l'ensemble des variables indicatrices de ses modalités.

I - GENERALITES SUR LES OPERATEURS D'ESCOUFIER

Soit X_i un tableau de données à n lignes et m_i colonnes contenant les valeurs de m_i variables sur n individus. Si $\mathbb{R}^{m_i}$, espace des individus, est muni de la métrique M_i et $\mathbb{R}^n$, espace des variables, de la métrique habituelle du poids D_p, l'opérateur d'ESCOUFIER O_i associé au tableau X_i est :

$$O_i = X_i M_i X'_i D_p.$$

Cet opérateur n'est autre que le WD_p du schéma de dualité suivant associé à l'analyse en composantes principales de X_i dans la métrique M_i :



* Maître assistant à l'Université de PARIS V (IUT)

O_i a pour vecteurs propres les composantes principales de X_i associées aux valeurs propres λ_{ik} de $M_i V_i$.

L'ensemble des opérateurs associés à des tableaux de données X_i à n lignes et aux métriques M_i est un sous-ensemble du sous-espace vectoriel des matrices Dp-symétrique. Ce sous-espace peut être muni du produit scalaire :

$$\langle O_i ; O_j \rangle = \text{Trace } (O_i O_j)$$

et de la norme :

$$\| O_i \|^2 = \text{Trace } (O_i^2) = \sum_k \lambda_{ik}^2$$

A cette norme correspond la notion d'équivalence suivante entre deux tableaux de données X_i et X_j pour les métriques M_i et M_j.

$$X_i \sim X_j \iff \| O_i - O_j \|^2 = 0$$

c'est-à-dire que X_i et X_j ont mêmes systèmes de composantes principales associées aux mêmes valeurs propres.

On définit alors l'angle θ_{ij} entre opérateurs ou entre tableaux de données par :

$$\text{Cos } \theta_{ij} = \frac{\text{Trace } (O_i O_j)}{\sqrt{\text{Trace } O_i^2 \text{ Trace } O_j^2}}$$

Nous supposerons pour la suite que $M_i = (X'_i D_p X_i)^{-1} = V_i^{-1}$ autrement dit que les opérateurs étudiés sont les projecteurs Dp-orthogonaux A_i sur les espaces W_i engendrés par les colonnes des tableaux X_i.

Dans ces conditions Trace $A_i A_j$ est la somme des valeurs propres de l'analyse canonique de X_i et X_j et

$$\text{Trace } A_i^2 = \text{Trace } A_i = \dim W_i$$

En particulier si les variables sont centrées, la trace de $A_i A_j$ est égale à la somme des carrés des coefficients de corrélation canoniques et si $\dim W_i = m_i$ on trouve :

$$\text{Cos } \theta_{ij} = \frac{\sum r_k^2}{\sqrt{m_i m_j}}$$

Deux tableaux de données sont équivalents si leurs colonnes respectives engendrent le même espace vectoriel ce qui entraîne que $r_k = 1 \quad \forall_k$ et $m_i = m_j$.

On a donc :

$$\text{Cos } \theta_{ij} = 1 \iff W_i = W_j$$
$$\text{Cos } \theta_{ij} = 0 \iff W_i \perp W_j$$

On notera que cos θ_{ij} est toujours positif et que si $m_i = m_j = 1$ il est égal au carré du coefficient de corrélation linéaire entre les deux variables.

II - UNE METHODE DE DESCRIPTION DES RELATIONS DEUX A DEUX ENTRE p VARIABLES QUALITATIVES.

Associons à une variable qualitative à m_i modalités le tableau logique X_i (dit tableau disjonctif) de présence ou d'absence des diverses modalités pour les n individus.

$$X_i = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & & & \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

L'espace W_i engendré par les colonnes de X_i est l'ensemble des variables numériques moins fines que la variable qualitative qui réalisent donc ses différents codages.

Pour éviter des solutions parasites nous considererons en fait le sous-espace W_{io} de W_i correspondant aux codages centrés, en d'autres termes W_{io} est la partie de W_i Dp-orthogonale au vecteur $\underline{1}$ de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1.

A_i désignera le projecteur Dp-orthogonal sur N_{io} et on a donc :

$$\text{Trace } A_i^2 = \text{Trace } A_i = \dim W_{io} = m_i - 1$$

Si X_i et X_j sont deux tableaux associés à deux variables qualitatives, on sait que l'analyse spectrale de A_iA_j (ou de A_jA_i) n'est autre que l'analyse des correspondances du tableau de contingence associé dans laquelle la solution triviale $\underline{1}$ a été éliminée.

La somme des valeurs propres, autres que la valeur triviale 1, est alors égale au ϕ^2 de contingence :

$$\text{Trace } A_iA_j = \phi^2_{ij} = \sum_i \sum_j \frac{(p_{ij} - p_{i.} - p_{.j})^2}{p_{i.}\ p_{.j}}$$

et le cosinus d'angle entre opérateurs n'est autre que le coefficient de dépendance de TSCHUPROW.

$$T_{ij} = \frac{\phi^2_{ij}}{\sqrt{(m_i-1)(m_j-1)}}$$

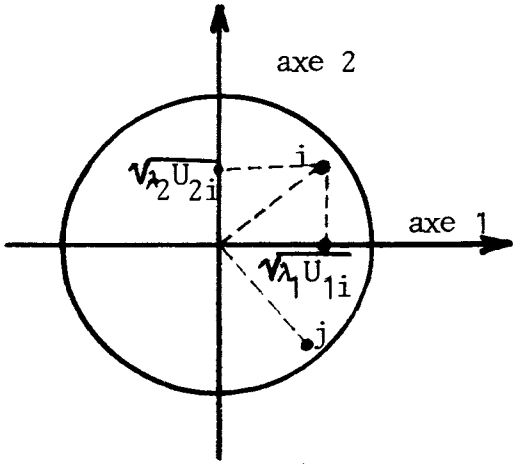
Ce coefficient possède la propriété de n'être égal à zéro que si les variables sont statistiquement indépendantes et de prendre la valeur 1 uniquement dans le cas de la dépendance totale : à une modalité d'une variable ne correspond qu'une modalité de l'autre et réciproquement.

Etant donné p variables qualitatives, construisons la matrice symétrique T des coefficients de TSCHUPROW des variables prises deux à deux ; il est alors très simple de représenter géométriquement les proximités entre les variables en effectuant une analyse en composantes principales sur le opérateurs normés. En effet, le coefficient de TSCHUPROW qui est un cosinus d'angle possède les propriétés d'un coefficient de corrélation et l'extraction des vecteurs propres et des valeurs propres de T, qui est alors l'analogue d'une matrice de corrélation, permet de dégager des facteurs, orthogonaux deux à deux au sein des opérateurs, qui résument le mieux les p variables qualitatives.

Les proximités entre variables peuvent alors être représentées selon la figure usuelle du cercle des correlations : le point représentatif de la $i^{ème}$ variable a pour coordonnée sur l'axe n° k ,la $i^{ème}$ composante du $k^{ième}$ vecteur propre $\underline{U}_k$ de T multipliée par la racine carrée de la valeur propre correspondante :

$\sqrt{\lambda_k}\ U_{ki}$ où $\underline{u}_k$ est tel que $\sum_{i=1}^{\emptyset} U^2_{ki} = 1$

Ainsi sur le premier plan on a la figure suivante :



Si deux variables sont représentées par des points proches de la circonférence et faisant avec l'origine un angle de $\frac{\pi}{2}$, cela veut dire qu'elles sont indépendantes.

On remarquera que l'ensemble des points-variables se trouve dans le demi-plan d'abscisse positive car T a tous ses éléments positifs.

Il est possible de projeter en élément supplémentaire une variable qualitative ne figurant pas parmi les p variables initiales. Il suffit pour cela de connaître le vecteur $\underline{t}$ de ses p coefficients de TSCHUPROW avec les variables de départ. La nouvelle variable sera représentée dans le système des p axes factoriels par un point dont les coordonnées sont les composantes du vecteur :

$$D_{1/\sqrt{\lambda}}\ U'\ \underline{t}$$

où U est la matrice dont les colonnes sont les vecteurs propres normés à 1 de T et $D_{1/\sqrt{\lambda}}$ la matrice diagonale des inverses des racines carrées des valeurs propres rangées dans le même ordre.

Ceci permet, en particulier, de faire figurer dans le cercle des corrélations les diverses modalités d'une variable : chaque modalité, qui est une variable qualitative dichotomique, étant alors projetée en élément supplémentaire. On trouve aisément que le point représentatif d'une variable est entouré par les points représentatifs de ses modalités.

L'utilisation des opérateurs permet aussi de traiter le cas d'un mélange de variables qualitatives et quantitatives.

Ainsi, on obtient sans difficulté que le cosinus d'angle entre l'opérateur associé à une variable à m_i modalités et celui associé à une variable numérique centrée vont :

$$\text{Cos } \theta = \frac{\eta^2}{\sqrt{m_i - 1}} \qquad \text{où} \quad \eta^2 \text{ est le rapport de corrélation}$$

De même le cosinus d'angle entre l'opérateur associé à une variable qualitative à m_i modalités et l'opérateur associé à un groupe de q variables numériques centrées est :

$$\text{Cos } \theta = \frac{\sum_{k=1}^{m_i - 1} \lambda_k}{\sqrt{q(m_i - 1)}} = \frac{\text{Trace } (V^{-1}B)}{\sqrt{q(m_i - 1)}}$$

où V est la matrice de variance-covariance totale du q variable et B la matrice de variance interclasse (ou matrice d'inertie des m_i centres de gravités) ; les λ_k sont alors les valeurs propres de l'analyse discriminante associée.

Les cosinus d'angle entre opérateurs définissent donc des indices de proximité comparables pour des variables qualitatives comme quantitatives. Quelques précautions sont cependant nécessaires pour effectuer ces comparaisons car il ne nous semble pas recommandé d'utiliser des variables qualitatives dont les nombres de modalités seraient trop différents. En effet deux χ^2 de contingence de même valeur numérique n'ont pas la même signification si les degrés de liberté sont différents ; le fait de diviser par la racine du nombre de degrés de liberté dans le coefficient de TSCHUPROW atténue cet inconvénient mais ne l'élimine pas totalement. Si les nombres de modalités sont trop différents il peut être conseillé de compléter la donnée d'un coefficient de TSCHUPROW T_{ij} par la probabilité qu'une variable de χ^2 à $(m_i - 1)(m_j - 1)$ degrés de liberté soit inférieure au χ^2 de contingence trouvé. Cette probabilité est une excellente mesure de la dépendance entre variables qualitatives mais n'a évidemment pas les propriétés d'un cosinus d'angle.

III - SELECTION PROGRESSIVE DE VARIABLES EXPLICATIVES DANS UNE ANALYSE DISCRIMINANTE SUR VARIABLES QUALITATIVES *.

La prévision d'une variable qualitative par p autres a souvent été traitée par la technique de segmentation. On peut ainsi l'aborder sous l'angle de l'analyse discriminante râce au codage, ce qui aboutit alors à affecter une modalité de la variable à expliquer à un individu selon la valeur d'une fonction numérique additive des diverses modalités des variables explicatives.

Le problème peut se formaliser ainsi : chercher un codage simultané de toutes les variables maximisant le coefficient de correlation multiple entre la variable à expliquer codée et les p variables explicatives codées. La solution est alors donnée par l'analyse canonique, moyennant quelques contr ntes sur les codages afin d'éviter des matrices singulières.

Si le choix d'un nombre limité de prédicteurs afin de réaliser une discrimination pas à pas est classique pour des variables numériques, il n'en est pas de même pour des variables qualitatives en raison de la difficulté de définir une mesure de dépendance entre deux variables qualitatives conditionnellement à une ou plusieurs autres. La seule méthode de discrimination pas à pas que nous connaissions étant celle de M. MASSON mais elle ne définit pas une telle mesure de dépendance partielle. Il est certes possible de définir des χ^2 conditionnels ou des quantités d'informations conditionnelles mais le volume des calculs devient vite prohibitif car il faut manier des tables de contingence à plusieurs dimensions.

La méthode que nous proposons s'inspire de la régression progressive et consiste à définir un indice de liaison partielle entre variables qualitatives analogue et la corrélation partielle grâce aux propriétés du coefficient de TSCHUPROW.

Le coefficient de TSCHUPROW étant pour les variables qualitatives l'analogue d'un coefficient de corrélation, nous définissons formellement le coefficient de TSCHUPROW partiel au moyen de la formule classique donnant le coefficient de corrélation partielle.

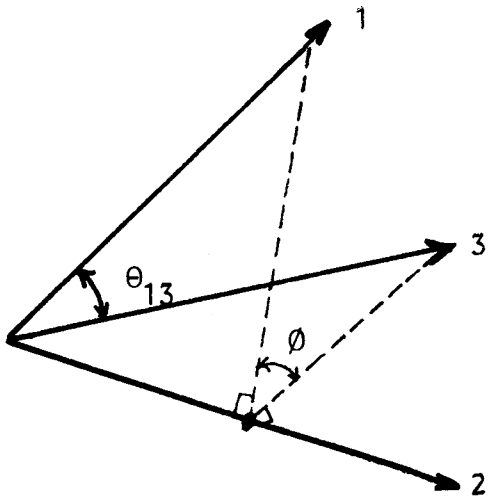
Avec trois variables on trouve ainsi :

$$T_{13.2} = \frac{T_{13} - T_{12} \cdot T_{32}}{\sqrt{(1-T_{12}^2)(1-T_{32}^2)}}$$

---

* cette application a été développée dans le cadre du contrat DGRST n° 75-7-0230

Dans l'espace des opérateurs, le coefficient $T_{13.2}$ est le cosinus de la projection de l'angle θ_{13} sur un plan orthogonal à l'opération n°2.



On voit sans difficulté que ce coefficient jouit de propriétés interessantes :

Si les variables 2 et 3 sont très liées, l'angle ϕ est alors voisin de $\frac{\pi}{2}$ et $T_{13.2}$ est proche de zéro : la prise en compte de la variable 3, une fois connue la variable 2, n'apporte pas d'information utile sur la variable 1.

d'autre part, à T_{12} et T_{13} fixés le coefficient est maximal si $T_{23} = 0$ c'est-à-dire si les variables 2 et 3 sont indépendantes.

On définit alors de proche en proche les coefficients de TSCHUPROW partiels d'ordres supérieurs :

$$T_{14.23} = \frac{T_{14.2} - T_{13.2}\ T_{43.2}}{\sqrt{(1-T^2_{13.2})(1 - T^2_{43.2})}} \qquad \text{etc}$$

L'algorithme de sélection progressive des variables explicatives est alors immédiat :

- au premier pas on cherche, pour expliquer la variable 1, la variable i qui maximise T_{1i}.
- au deuxième pas on introduit la variable j qui maximise $T_{1j.i}$
- au troisième pas on introduit la variable k qui maximise $T_{1k.ij}$
- ...

On peut songer à définir un coefficient de TSCHUPROW multiple $T_{1.23}$ par la formule usuelle

$$(1 - T^2_{1.23}) = (1 - T^2_{12})\ (1 - T^2_{13.2}) \qquad \text{etc}$$

mais ce coefficient ne semble pas posséder de propriétés aisément interprétables sauf dans le cas où les variables explicatives sont indépendantes ($T^2_{1.23} = T^2_{12} + T^2_{13}$) et ont même nombre de modalités : on montre alors que $T^2_{1.23}$ est à un coefficient près la somme des valeurs propres de l'analyse discriminante globale de 1 contre 2 et 3.

Références :

Y. ESCOUFIER : "Echantillonnage dans une population de variables aléatoires réelles" Thèse de Doctorat ès Sciences Montpellier (1970).

M. MASSON : "Processus linéaire et analyse de données non linéaires" Thèse de Doctorat ès Sciences Université de PARIS VI (1974)

J. PAGES : "A propos des opérateurs d'Y. ESCOUFIER" Séminaires de l'IRIA en classification automatique (1974)

G. SAPORTA : "Liaison entre plusieurs ensembles de variables et codage de données qualitatives" Thèse de 3e cycle Université de PARIS VI (1975)