STATISTIQUE ET ANALYSE DES DONNÉES

GILBERT SAPORTA

Pondération optimale de variables qualitatives en analyse des données

Statistique et analyse des données, tome 4, n° 3 (1979), p. 19-31

<http://www.numdam.org/item?id=SAD_1979__4_3_19_0>

Statistique et Analyse des Données
3 - 1979 pp. 19-31

PONDERATION OPTIMALE DE VARIABLES QUALITATIVES EN ANALYSE DES DONNEES

Gilbert SAPORTA
IUT DE PARIS UNIVERSITE RENE DESCARTES
143, Avenue de Versailles 75016 PARIS

La méthode proposée notamment par Y. ESCOUFIER ; J.P. PAGES ; P. CAZES et S. BONNEFOUS [1], [3], [6], pour étudier les relations entre p variables qualitatives sur n individus diffère de l'analyse des correspondances multiples usuelle en ce qu'elle pondère les sous-tableaux du tableau disjonctif complet par des coefficients issus d'une analyse globale des dépendances entre variables qualitatives. En recherchant directement des coefficients satisfaisant à certains critères nous montrons dans ce papier en quoi un tel type d'analyse est optimal et nous l'étendons à des analyses explicatives.

I - RAPPEL CONCERNANT L'ANALYSE EN DEUX TEMPS DE p TABLEAUX DE DONNEES

Soit n objets ou individus munis de poids et D la matrice diagonale des poids. A chaque tableau X_i à n lignes et p_i colonnes on associe son opérateur d'ESCOUFIER O_i :
$O_i = W_i D = X_i M_i X_i D$ qui résume les proximités entre objets (produits scalaires avec la métrique M_i dans l'espace des individus).

On munit l'ensemble des opérateurs du produit scalaire de la trace :

$$< O_i ; O_j > = \text{Trace } (O_i O_j) = t_{ij}$$

Le tableau carré pxp symétrique T d'éléments t_{ij} est alors l'analogue d'une matrice de covariance (ou d'une matrice de corrélation si les O_i sont normés).

Premier temps :

Pour étudier globalement les ressemblances entre les p tableaux de données on effectue alors une analyse en composantes principales sur les opérateurs associés ce qui revient à diagonaliser T. Le premier vecteur propre $\underline{t}$ de T définit une combinaison linéaire à coefficients de même signe (théorème de FROBENIUS car $t_{ij} \geq 0 \; \forall i,j$) que l'on prendra positifs, des p opérateurs. Cette combinaison $\sum_{i=i}^{p} t_i \, O_i$ est donc encore un opérateur d'ESCOUFIER car elle est semi-définie-positive : elle représente un compromis entre les O_i tenant compte des ressemblances majoritaires puisque c'est la première composante principale.

Deuxième temps :

On diagonalise maintenant $\sum_{i=i}^{p} t_i \, O_i$ pour obtenir une représentation des individus et des variables de tous les tableaux tenant compte des proximités globales entre tableaux. Lorsque l'on ne dispose que de tableaux de distances ou de similarité cette méthode est voisine de l'algorithme INDSCAL de J.D. CARROLL.

Si les X_i sont des tableaux d'indicatrices associées aux modalités de p variables qualitatives, on utilise comme opérateur O_i le projecteur A_{io} sur l'espace W_{io} des variables numériques centrées combinaisons linéaires des indicatrices de la variable i

$$A_{io} = X_i \, (X_i' \, D \, X_i) \, X_i' \, D - \underline{1} \, \underline{1}' \, D$$

Le tableau T est alors :

- soit le tableau des ϕ^2 de K. PEARSON
- soit le tableau des T^2 de TSCHUPROW [6], [8]

selon que l'on normalise ou non les opérateurs.

La méthode d'analyse en deux temps consiste ensuite à diagonaliser $\sum_{i=1}^{p} t_i \, A_{io}$ ou, ce qui revient ou même à la solution triviale près, $\sum^{p} t_i \, A_i$ où $A_i = X_i' (X_i' \, DX_i)^{-1} \, X_i' \, D$ avec $\underline{t}$ premier vecteur propre de T.

L'analyse des correspondances multiples du tableau disjonctif complet $(X_1 \mid X_2 \mid .. \mid X_p)$ étant équivalente* à la diagonalisation de $\sum_{i=i}^{p} A_i$ on voit immédiatement que la diagonalisation de $\sum_{i=1}^{p} t_i \, A_i$ revient à effectuer une analyse des correspondances sur le tableau disjonctif

---

* Les valeurs propres et vecteurs propres de $\frac{1}{p} \Sigma A_i$ sont ceux de l'analyse des correspondances de $X = (X_1 \mid X_2 \mid \ldots \mid X_p)$ considérée comme tableau de contingence. Les valeurs propres de l'AFC du tableau de BURT sont les carrés des valeurs propres de l'AFC précédente.

modifié $(t_1 X_1 \mid t_2 X_2 \mid .. \mid t_p X_p)$ ou encore une A.F.C. sur le tableau de BURT modifié d'éléments blocs $t_i\ t_j\ C_{ij}$ où C_{ij} est le tableau de contingence croisant les variables i et j.

Si on a travaillé sur le tableau des coefficients de TSCHUPROW il faut considérer alors le tableau $\left(\frac{t_1}{\sqrt{m_1 - 1}} X_1 \mid \ldots \mid \frac{t_p}{\sqrt{m_p - 1}} X_p \right)$ où m_i est le nombre de modalités de la ième variable.

Les pourcentages d'inertie des premières valeurs propres sont en général plus élevés que dans une analyse des correspondances usuelle [1] car les variables peu liées aux autres se trouvent affectées d'un coefficient voisin de zéro et donc éliminées.

Peut-on parler pour cela d'analyse optimale ?

II - PONDERATION OPTIMALE A BUT DESCRIPTIF

Nous allons chercher directement les poids $\alpha_1, \alpha_2 .. \alpha_p$ à attribuer aux diverses variables ou aux opérateurs associés.

Le but de ce genre d'analyse étant d'avoir une représentation des données sur un espace de faible dimension il semble logique de choisir les α_i de manière à optimiser un critère lié à la qualité de la représentation.

1 - Critères liés à l'inertie

Une multiplication par une constante k des α_i ne fait que multiplier par k les valeurs propres de $\Sigma \alpha_i A_i$ sans modifier les vecteurs propres ; on imposera donc $\Sigma \alpha_i^2 = 1$.

Abordons dans un premier temps l'optimisation de l'inertie du premier axe.

a) Maximisation de la plus grande valeur propre de $\Sigma \alpha_i A_i$

Soit $\underline{u}$ et λ le 1er vecteur propre et la première valeur propre de $\Sigma \alpha_i A_i$. Il s'agit alors de maximiser $\underline{u}' (\Sigma \alpha_i A_i) \underline{u}$ sous les contraintes $\underline{u}' \underline{u} = 1$ et $\Sigma \alpha_i^2 = 1$.

La méthode des multiplicateurs de LAGRANGE donne en dérivant $\left[\Sigma \alpha_i \underline{u}' A_i \underline{u} - \frac{\lambda}{2} \underline{u}' \underline{u} - \frac{\mu}{2} \Sigma \alpha_i^2 \right]$ par rapport à α_i :

$\alpha_i = \underline{u}' A_i \underline{u}$ ce qui prouve que les α_i sont tous positifs.

A l'optimum $\underline{u}$ sera donc 1er vecteur propre de $\Sigma \alpha_i A_i$ et α_i est proportionnel à $\underline{u}' A_i \underline{u}$. Le vecteur $\underline{u}$ est donc tel que $\sum_{i=1}^{p} (\underline{u}' A_i \underline{u}) A_i \underline{u} = \lambda \underline{u}$

Les A_i étant les projecteurs sur les sous-espaces W_i engendrés par les tableaux de données X_i, si $\underline{u}$ est de norme 1, $\underline{u}' A_i \underline{u}$ est le carré du cosinus de l'angle θ_i formé par $\underline{u}$ et W_i.

Le premier vecteur propre de $\Sigma\alpha_i A_i$ a pour propriété de rendre maximal $\Sigma\alpha_i \cos^2 \theta_i$ [7.] Les poids optimaux α_i étant proportionnels aux $\cos^2 \theta_i$, le vecteur $\underline{u}$ rend donc maximal $\Sigma \cos^4 \theta_i$.

L'analyse optimale revient donc à maximiser $\Sigma \cos^4 \theta_i$ alors que la pratique usuelle (diagonaliser ΣA_i) consiste à maximiser $\Sigma \cos^2 \theta_i$.

Lorsque p = 2 on retrouve évidemment l'analyse canonique ordinaire.

Le vecteur $\underline{u}$ rendant maximal $\sum_{i=1}^{p} (\underline{u}' A_i \underline{u})^2$, on peut envisager sa recherche par l'utilisation de divers algorithmes numériques d'optimisation. Outre des méthodes type gradient qui sont applicables ici car la fonction $\underline{u} \longrightarrow \Sigma(u' A_i u)^2$ est convexe nous proposons les deux algorithmes suivants pour atteindre l'optimum :

* Algorithme 1 type approximations successives

On diagonalise ΣA_i d'où $\underline{u}_1$

On pose $\alpha_i^{(1)} = \dfrac{\underline{u}_1' A_i \underline{u}_1}{\sqrt{\Sigma (\underline{u}_1' A_j \underline{u}_1)^2}}$

On diagonalise ensuite $\sum \alpha_i A_i$ d'où $\underline{u}_2$ etc.

$$\alpha_i^{(n)} = \frac{\underline{u}_n' A_i \underline{u}_n}{\sqrt{\sum_j (\underline{u}_n' A_j \underline{u}_n)^2}}$$

* Algorithme 2 type alterné

On remarque qu'à l'optimum $\Sigma\alpha_i A_i \underline{u}$ est la combinaison linéaire des vecteurs $A_i \underline{u}$ de norme maximale, c'est-à-dire la première composante principale des $A_i \underline{u}$. $\underline{\alpha}$ est donc vecteur propre de la matrice de covariance des $A_i \underline{u}$ (dont le terme général est $\underline{u}' A_i A_j \underline{u}$) associé à la valeur propre λ^2 car $\left\| \Sigma\alpha_i A_i \underline{u} \right\|^2 = \left\| \lambda \underline{u} \right\|^2 = \lambda$

Il doit donc être possible de trouver λ et $\underline{u}$ par la procédure suivante :

. soit $\underline{u}^{(1)}$ le premier vecteur propre de ΣA_i

. On forme la matrice de terme général $\underline{u}'(1) A_i A_j \underline{u}^{(1)}$ soit $\underline{\alpha}^{(1)}$ son premier vecteur propre normé.

. On cherche alors $\underline{u}^{(2)}$ premier vecteur propre de la matrice $\Sigma\alpha_i^{(1)} A_i$

. $\underline{\alpha}^{(2)}$ est le premier vecteur propre de la matrice de termes $\underline{u}'^{(2)} A_i A_j \underline{u}^{(2)}$

. etc

b - Maximisation de l'inertie sur k axes

Il s'agit ici de maximiser sur $\underline{u}_j$ et α_i l'expression

$$\sum_{i=1}^{p} \sum_{j=1}^{k} \underline{u}_j' \alpha_i A_i \underline{u}_j \text{ avec } \left\| \underline{u}_j \right\|^2 = 1 \text{ et } \Sigma \alpha_i^2 = 1.$$

L'optimum vaut alors $\lambda_1+\lambda_2 + \quad + \lambda_k$ où les λ sont les k plus grandes valeurs propres de $\Sigma\alpha_i A_i$.

On trouve aisément qu'à l'optimum les α_i sont proportionnels à $\sum_{j=1}^{k} \underline{u}_j' A_i \underline{u}_j$ et sont par conséquent tous positifs.

Pour les obtenir nous proposons d'utiliser un algorithme de type itératif dérivé de l'algorithme 1 présenté au paragraphe précédent :

On fixe les α_i et on cherche les k premiers vecteurs propres de $\alpha_i A_i$. Les $\underline{u}_j$ étant fixés on obtient les α_i par la relation $\alpha_i = \sum_{j=i}^{k} \underline{u}_j' A_i \underline{u}_j$ etc.

On choisit comme précédemment $\alpha_i = 1 \forall_i$ pour démarrer l'algorithme.

Les coefficients α_i dépendent donc de k, dimension de l'espace sur lequel on veut faire la représentation.

Que se passe-t-il lorsque k augmente ?

On constate alors que pour k maximal = dim $(\bigoplus_{i=i}^{k} W_i)$:

On a $\sum_{j} \underline{u}_j' A_i \underline{u}_j$ = Trace $A_i = m_i - 1$

Les α_i sont alors proportionnels aux nombres de modalités de chaque variable diminué d'une unité, ce qui ne présente aucun intérêt.

Ceci est dû au fait que le critère de l'inertie pour une analyse des correspondances sur tableau disjonctif complet est d'un intérêt médiocre voire nul [4]. En effet la somme des valeurs propres de ΣA_i vaut toujours $\Sigma(m_i-1)$ et l'inertie totale de l'AFC du tableau disjonctif $(X_1 \mid X_2 \mid \ldots \mid X_p)$ est $\frac{1}{p}$ $\Sigma m_i - 1$ ce qui est indépendant des liaisons entre variables.

Les pourcentages d'inertie expliquée dans une AFC sur tableau disjonctif sont donc souvent extrêmement faibles car bornés par l'inverse de l'inertie puisque la plus grande valeur propre d'une AFC est inférieure ou égale à 1.

Rappelons ici l'exemple classique de la disjonction d'un tableau de contingence :

Si $X_1' X_2$ est le tableau de contingence et $(X_1 \mid X_2)$ le tableau disjonctif complet associé, on a la relation suivante entre les valeurs propres ρ_i de l'AFC de $(X_1' X_2)$ et les valeurs propres λ_i de l'AFC de $(X_1 \mid X_2)$:

$$\rho_i = (1 - 2\lambda_i)^2$$

En fait un meilleur critère de qualité en AFC multiple est fourni par les carrés des valeurs propres dont la somme s'interprète en termes de $\emptyset^2$.

Si les λ_i sont les valeurs propres de $\sum_{i=i}^{p} A_i$ on a :

$$\sum_{i=i}^{n} \lambda_i^2 = (m-p) + 2 \sum\sum_{i \neq j} \emptyset^2_{ij} \qquad \text{avec } m = \sum_{i=i}^{p} m_i$$

voir [5] et [7].

Ceci revient à considérer l'AFC sur le tableau du BURT et non sur le tableau disjonctif, ce qui est plus conforme à la logique : l'extension naturelle d'un tableau de contingence croisant deux variables qualitatives est en effet le tableau de BURT où l'on croise toutes les variables deux à deux (si on se limite aux dépendances d'ordre 1).

2) Critère de la somme des carrés des valeurs propres

Le cas d'une seule valeur propre ayant déjà été vu au paragraphe précédent voyons ce que donne la maximisation de la somme des carrés des k premières valeurs propres :

$$\sum_{l=1}^{k} \lambda_l^2 = \sum_{l=1}^{k} (\underline{u}_l' \; \Sigma\alpha_i \; A_i \; \underline{u}_l)^2$$

$$= \sum_{l=1}^{k} \left\| \sum_{i=1}^{p} \alpha_i \; A_i \; \underline{u}_l \right\|^2$$

$$\text{Donc} \sum_{l=1}^{k} \lambda_l^2 = \sum_{l=1}^{k} \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_i \; \alpha_j \; \underline{u}'_l \; A_i \; A_j \; \underline{u}_l$$

Pour maximiser la quantité précédente sous la contrainte $\sum_{i=i}^{p} \alpha_i^2 = 1$ la méthode des multiplicateurs de Lagrange aboutit aux p équations :

$$\sum_{l=1}^{k} \sum_{j=1}^{p} \alpha_j \; \underline{u}'_l \; A_i \; A_j \; \underline{u}_l = \mu\alpha_i \qquad i = 1,2, \ldots p$$

$\underline{\alpha}$ est donc vecteur propre de la matrice de terme général :

$$\sum_{l=1}^{k} \underline{u}'_l \; A_i \; A_j \; \underline{u}_l$$

D'ou la possibilité d'un algorithme alterné : recherche des k premiers vecteurs propres de $\sum_{i=1}^{p} \alpha_i^{(n)} \; A_i$ puis des vecteurs propres de la matrice de terme général :

$$\sum_{l=1}^{k} \underline{u}'^{(n)}_l \; A_i \; A_j \; \underline{u}^{(n)}_l$$

On peut remarquer au passage une propriété des coefficients α_i en notant que :

$$\sum_{l=1}^{k} \lambda_l^2 = \sum_{l=1}^{k} \left\| \Sigma \alpha_i \; A_i \, \underline{u}_l \right\|^2 = \left\| \sum_{l=1}^{k} \sum_{i=1}^{p} \alpha_i \; A_i \, \underline{u}_l \right\|^2$$

car les $\underline{u}_l$ étant vecteurs propres de $\sum_{i=1} \alpha_i \; A_i$ forment un système orthogonal.

Les α_i sont donc les coefficients de la combinaison linéaire des $\sum_{l=1}^{k} A_i \, \underline{u}_l$ (c'est-à-dire des $A_i \sum_{l=1}^{k} \underline{u}_l$), de norme maximale ; ils définissent donc le facteur associé à la première composante de ces vecteurs.

$\underline{\alpha}$ est donc à la fois vecteur propre de la matrice de terme général

$$\left(\sum_{l=1}^{k} \underline{u}_l'\right) A_i \, A_j \left(\sum_{l=1}^{k} \underline{u}_l\right)$$ et de celle de terme général $\sum_{l=1}^{k} \underline{u}_l' \, A_i \, A_j \, \underline{u}_l$

Si on prend $k = \dim \left(\bigoplus_{i=1}^{k} W_i\right)$ il vient :

$$\Sigma \; \underline{u}_l' \, A_i \, A_j \, \underline{u}_l = \text{Trace } A_i \, A_j = \langle A_i \; ; \; A_j \rangle = t_{ij}$$

On retrouve alors la solution d'ESCOUFIER-PAGES et al. car $\underline{\alpha}$ est alors le premier vecteur propre de la matrice T du paragraphe I.

En effet si on prend pour norme des opérateurs symétriques (ou D - symétriques) la somme des carrés des valeurs propres ceci revient à rendre maximal $\left\| \sum_{i=1}^{p} \alpha_i \; A_i \right\|^2$ sous la contrainte $\Sigma \alpha_i^{\,2} = 1$; ceci définit la première composante principale des opérateurs A_i.

Le critère $\max \sum_{l=1}^{k} \lambda^2_l$ avec $k < \dim \left(\bigoplus_{i=1}^{p} W_i\right)$ revient donc à maximiser la norme de la meilleure approximation de rang k de $\Sigma \alpha_i \, A_i$

<u>La méthode d'ESCOUFIER-PAGES et al. est alors optimale en ce sens qu'elle aboutit à une AFC dont la somme des carrés de toutes les valeurs propres est maximale</u>.

Elle présente d'autre part l'avantage sur les autres méthodes de pondération envisagées précédemment, de fournir des coefficients α_i indépendants de la dimension de l'espace sur lequel la représentation sera faite, et tenant compte des liaisons globales entre variables mesurées par les ϕ^2.

Si on travaille sur la matrice des coefficients de TSCHUPROW, ceci revient à chercher à maximiser $\left\| \sum_{i=1}^{p} \frac{\alpha_i}{\sqrt{m_i - 1}} \; A_i \right\|^2$ avec $\Sigma \alpha_i^{\,2} = 1$

ce qui équivaut à optimiser l'analyse des correspondances, au sens de la somme des carrés des valeurs propres, de $(\beta_1 \, X_1 \mid \beta_2 \, X_2 \mid \ldots \mid \beta_p \, X_p)$ avec pour contrainte $\sum_{i=1}^{p} (m_i - 1) \, \beta_i^{\,2} = 1$

Si les variables ont des nombres de modalités très différents, cette dernière pratique est préférable bien que la contrainte soit moins naturelle, car elle évite l'effet d'entraînement donné par les variables à fort nombre de modalités. Un argument voisin en faveur de l'utilisation du premier vecteur propre du tableau des coefficients de TSCHUPROW est le suivant :

- si tous les coefficients de liaison (au sens de TSCHUPROW qui est moins dépendant du degré de liberté que le ϕ^2) entre les p variables prises deux à deux sont égaux, il est naturel d'exiger qu'elles soient équipondérées. Ceci sera le cas si on travaille avec la matrice des T^2 et non celle des ϕ^2.

III - PONDERATION OPTIMALE A BUT EXPLICATIF

Nous supposons ici qu'il existe une (p+1)ème variable qualitative, à laquelle sont associés le tableau de données X_o et le projecteur A_o, que l'on cherche à expliquer par les p autres variables qualitatives. Il s'agit donc d'un problème du type analyse discriminante sur variables qualitatives. Ayant déjà proposé une méthode de discrimination sur des données de ce type [9] nous ne reviendrons pas sur la solution calculée par DISQUAL mais nous nous attacherons à définir des variantes possible de l'analyse des correspondances pour l'orienter dans un but explicatif.

On sait que les premiers facteurs de l'analyse des correspondances de $(X_1 | \ldots | X_p)$ n'ont en général aucune raison d'être les meilleurs facteurs discriminantes pour X_o. Par ailleurs certains praticiens effectuent une AFC sur la juxtaposition des tableaux de contingence $(X_o' X_1 | X_o' X_2 | \ldots | X_o' X_p)$, cette dernière pratique n'étant équivalente à l'analyse discriminante sur variables qualitatives que dans le cas où les p variables explicatives sont indépendantes deux à deux.

On peut penser à améliorer ces diverses analyses en introduisant comme précédemment des coefficients $\alpha_1, \alpha_2, \ldots \alpha_p$ pour les variables explicatives en sorte que l'on augmentera le pouvoir discriminant des facteurs obtenus.

1) Pondération de l'AFC des variables explicatives

Diverses approches sont possibles dont aucune n'est pleinement satisfaisante.

a) Optimisation de la canonique sur k facteurs.

Cherchons à obtenir des axes factoriels qui soient le plus discriminants possible pour la variable à expliquer. En d'autres termes il s'agit de trouver des coefficients $\alpha_1, \alpha_2 \ldots, \alpha_p$ tels que les k premièrs vecteurs propres $\underline{u}_j$ de $\Sigma\alpha_i A_i$ aient un pouvoir discriminant maximal : les vecteurs propres étant orthogonaux, ce pouvoir discriminant sera la somme des inerties inter-classe : $\sum_{j=1}^{k} \underline{u}_j' A_o \underline{u}_j$

Ceci revient donc à maximiser la qualité de l'analyse canonique de X_o contre les $\underline{u}_j$.

Nous ne proposerons pas d'algorithme pour obtenir les α_i car on s'aperçoit rapidement que cette problématique est absurde.

En effet si on augmente k, les α_i deviennent indéterminés :

Pour $k = \dim(\oplus_{i=1}^{p} W_i)$ l'analyse canonique de X_o contre les $\underline{u}_j$ est alors la même que celle de X_o contre $(\alpha_1 \mid X_1 \mid \alpha_2 \; X_2 \mid \ldots \mid \alpha_p \; X_p)$ c'est-à-dire celle de X_o contre $(X_1 \mid X_2 \mid .. \mid X_p)$ car l'espace engendré par les $\underline{u}_j$ est alors identique à celui engendré par les colonnes de X.

b) Régression sur opérateurs

La première approche revenait en fait à chercher des α_i tels que l'approximation de rang k de $\Sigma\alpha_i A_i$ soit la plus proche de A_o au sens de la canonique. On peut inverser la problématique et rechercher tout d'abord la meilleure approximation de A_o par $\Sigma\alpha_i A_i$ puis trouver l'approximation de rang k de $\Sigma\alpha_i A_i$.

Les résultats obtenus seront différents puisque dans la première procédure les α_i dépendaient manifestement de k, alors qu'ils en sont indépendants dans la deuxième procédure ce qui peut être considéré comme un avantage.

La recherche des coefficients α_i tel que $\Sigma\alpha_i A_i$ soit aussi proche que possible de A_o n'est qu'un problème de régression. Si on prend comme critère de minimiser $\left\| A_o - \Sigma\alpha_i A_i \right\|^2$ avec la norme $\left\| A \right\|^2 = \text{Trace } A^2$ les équations normales s'écrivent :

$F\underline{\alpha} = \underline{v}$ avec F matrice de terme général $f_{ij} = \text{Trace } A_i A_j = \phi^2_{ij}$ et $\underline{v}$ vecteur de terme général $v_i = \text{Trace } A_o A_i = \phi^2_{oi}$

Si F est régulière on obtient $\underline{\alpha} = F^{-1} \underline{v}$ et $\dfrac{\left\| \Sigma\alpha_i A_i \right\|^2}{A_o^2}$ n'est autre que le carré R^2 du coefficient de TSCHUPROW multiple introduit en [8].

Si on extrait les k premiers vecteurs propres de $\Sigma\alpha_i A_i$ de valeurs propres $\lambda_1, \lambda_2 \ldots, \lambda_k$ alors $(\lambda_1^2 + \lambda_2^2 + \ldots + \lambda_k^2) / \Sigma\lambda_j^2$ mesure à la fois le taux de reconstitution de $\Sigma\alpha_i A_i$ au sens des moindres carrés et le taux de reconstitution du R^2 car

$$\Sigma\lambda_j^2 = \left\| \Sigma\alpha_i P_i \right\|^2 = R^2 \left\| A_o \right\|^2 = R^2 (m_o - 1).$$

Les vecteurs propres de $\Sigma\alpha_i A_i$ permettent donc de faire une description des variables explicatives tenant compte de la variable X_o. Une telle procédure semble posséder par rapport à DISQUAL l'avantage d'une plus grande rapidité : au lieu de diagonaliser ΣA_i, retenir les vecteurs propres les plus discriminants et faire enfin une analyse factorielle discriminante, il suffit ici de faire l'AFC de :

$(\alpha_1 X_1 \mid \alpha_2 X_2 \mid \ldots \mid \alpha_p X_p)$ pour obtenir des facteurs très discriminants par X_o, le calcul des α_i étant fort simple.

Un inconvénient mineur est que $\Sigma\alpha_i A_i$ étant en général de rang supérieur à $m_o - 1$ il y aura plus de facteurs que nécessaire pour une discrimination. Un autre défaut de cette régression sur opérateurs est que rien n'assure que $\Sigma\alpha_i A_i$ sera semi-défini-positif car les α_i ne sont pas nécessairement tous positifs (ce qui peut, mais pas obligatoirement, entraîner la non positivité de $\Sigma\alpha_i A_i$). Des valeurs propres négatives, donc dépourvues de signification statistique, sont donc à craindre. Il faudra comme en analyse factorielle sur tableau de distance se limiter aux valeurs propres positives ce qui revient à chercher l'approximation semi-définie-positive de $\Sigma\alpha_i A_i$.

2) <u>Pondération dans une AFC sur juxtaposition de tableaux de contingence.</u>

Lorsque les p variables explicatives sont deux à deux indépendantes $\sum_{i=1}^{p} A_i$ est alors le projecteur sur $\bigoplus_{i=1}^{p} W_i$.

L'analyse discriminante de X_o contre $(X_1 \mid X_2 \mid \ldots \mid X_p)$ revient alors à l'analyse de l'opérateur $A_o (\Sigma A_i) = \sum_{i=1}^{p} A_o A_i$ ce qui équivaut exactement, en cherchant les facteurs et non plus les variables discriminantes, à une AFC sur la juxtaposition des tableaux de contingence croisant la variable à expliquer et les variables explicatives :

$$(X'_o X'_1 \mid X'_o X_2 \mid \ldots \mid X'_o X_p) = K$$

De nombreux praticiens utilisent ce type d'AFC même si les variables ne sont pas indépendantes, car les calculs sont assez légers, mais il ne s'agit pas de l'analyse discriminante exacte.

Puisque ce type d'analyse n'est pas optimal [2]; [7] on peut songer à l'améliorer afin d'obtenir des facteurs plus discriminants en pondérant chaque variable explicative.

Les facteurs associés aux lignes du tableau K définissent des fonctions numériques (codages) transformant la variable à expliquer en variable numérique. La discrimination sera d'autant meilleure que la variance de la variable transformée est plus grande. Les variances en question étant les valeurs propres de l'AFC de K on voit que notre problème sera de maximiser la somme des premières valeurs propres. De plus ces valeurs propres s'interprètent ici comme des contributions à la somme des ϕ^2_{oi} entre la variable à expliquer et les variables explicatives.

$$\text{Trace} \sum_{i=1}^{p} A_o A_i = \sum_{i=1}^{p} \text{Trace } A_o A_i = \sum_{i=1}^{p} \phi^2_{oi}$$

Il faut alors trouver des coefficients $\alpha_1 \ \alpha_2, \ \ldots \ \alpha_p$ avec $\sum_{i=1}^{p} \alpha_i^2 = 1$ tels que $\sum_{l=1}^{k} \underline{u}_l' \left(\sum_{i=1}^{p} \alpha_i A_o A_i\right) \underline{u}_l$ soit maximal avec $\underline{u}_l$ lème vecteur propre de $\Sigma\alpha_i A_o A_i$

On trouve facilement en dérivant par rapport aux α_i la quantité

$\sum_{l=1}^{k} \alpha_i \underline{u}_l' A_o A_i \underline{u}_l - \sum_{l=1}^{k} \lambda_l \underline{u}_l' \underline{u}_l - \mu \sum_{i=1} \alpha_i^2$ que α_i doit être proportionnel

à $\sum_{l=1}^{k} \underline{u}_l' A_o A_i \underline{u}_l$.

La solution s'obtiendra comme précédemment par approximations successives en partant des vecteurs propres de $\Sigma A_o A_i$.

Mais comme bien souvent le nombre des modalités de la variable à expliquer n'est pas très élevé il est préférable de prendre directement l'ensemble des facteurs et de maximiser la trace de $\Sigma\alpha_i A_o A_i$ on trouve alors immédiatement :

$$\boxed{\alpha_i = \phi^2_{oi}}$$ à un coefficient prés.

Chaque variable est donc pondérée par son degré de dépendance avec la variable à expliquer.

L'optimum est alors égal à : $\sqrt{\sum_{i=1}^{k} \phi^4_{oi}}$ alors qu'avec des coefficients α_i égaux entre eux ($\alpha_i = \frac{1}{\sqrt{p}}$ l'inertie totale était $\frac{1}{p} \Sigma \phi^2_{oi}$)

La solution trouvée ici coïncide avec celle de la régression de l'opérateur A_o sur les A_i lorsque les A_i sont orthogonaux deux à deux et que les variables explicatives ont même nombre de modalités.

En pratique ceci reviendra à effectuer une analyse des correspondances sur le juxtaposition suivante de tableaux de contingence pondérés :

$$(\alpha_1 X_o' X_1 \mid \alpha_2 X'_o X_2 \mid \ldots \mid \alpha_p X'_o X_p)$$

Tout se passe donc comme si on pondérait différemment les individus pour chaque croisement proportionnellement à l'intensité de la liaison avec la variable à expliquer.

CONCLUSION

Au delà d'un petit jeu mathématique destiné à éclaircir (?) certaines pratiques, quel peut être l'intérêt de pondérer des variables qualitatives dans une analyse de données ?

Sur le plan descriptif l'intérêt est double : les coefficients α_i permettent de repérer l'importance des variables dans la constitution d'un fait majoritaire et fournissent l'analyse de données le mettant en évidence. La méthode d'Escoufier-Pagès nous semble alors la plus intéressante.

Sur le plan explicatif, dans la mesure où il s'agit souvent d'affecter ultérieurement un nouvel individu à une des modalités de la variable à expliquer, toutes les tentatives pour améliorer le pouvoir discriminant d'analyses non optimales mais peu couteuses en calculs sont a priori recevables pourvu que les coefficients obtenus puissent s'interpréter aisément.

REFERENCES

[1] CAZES P ; BONNEFOUS S ; BAUMERDER A ; PAGES J.P.
"Description cohérente des variables qualitatives prises globalement et de leurs modalités".
Statistique et analyse des données n° 2 1976 p 48-62

[2] CAZES P. : "Propriétés extremales des facteurs issus d'un sous tableau d'un tableau de BURT" Cahiers de l'analyse des données. Vol 2 n° 2 1976 p 143-160.

[3] ESCOUFIER Y. ; CAILLIEZ F. ; PAGES J.P.
"Géométrie et Techniques particulières en analyse factorielle"
European Meeting of the Psychometric Society Uppsala 1978.

[4] LEBART L. "Validité des résultats en analyse des données" CREDOC 1975

[5] LECLERC A. ; AIACH P. "Mesure de l'importance des valeurs propres en analyses des données ..." Revue de Statistique Appliquée 1978 XXVI n° 1 p 5-21.

[6] PAGES JP. ;ESCOUFIER Y. ; CAZES P.
"Opérateurs et analyse des tableaux à plus de deux dimensions" Cahiers du BURO n° 25 1976 p 61-89.

[7] SAPORTA G. "Liaison entre plusieurs ensembles des variables et codage de données qualitatives". Thèse 3e cycle Paris 1975.

[8] SAPORTA G. " Quelques applications des opérateurs d'Escoufier au traitement des variables qualitatives".
Statistique et analyse des données n° 1 1976 p 38-46.

[9] SAPORTA G. "Une méthode et un programme d'analyse discriminante sur variables qualitatives".
Colloque IRIA Analyse des données et Informatique Versailles 1977.