Une méthodologie d'analyse discriminante sur variables qualitatives

Jean-Michel Gautier, Gilbert Saporta

Jean-Michel GAUTIER
C O R E F
47, boulevard du Lycée
92170 VANVES

Gilbert SAPORTA
I U T Université R. Descartes
143, Avenue de Versailles
75016 PARIS

UNE METHODOLOGIE DE DISCRIMINATION SUR VARIABLES QUALITATIVES

Résumé : On propose des principes et un programme pour la discrimination lorsque les variables explicatives sont non-numériques tenant compte des interactions et des différences de dispersion entre groupes.

MOTS CLES : VARIABLES QUALITATIVES : DISCRIMINATION ;
ANALYSE DES DONNEES

I - LE PROBLEME

n individus étant décrits par p variables qualitatives X_1, X_2,... X_p à m_1, m_2,..., m_p modalités et une variable à expliquer Y à k modalités, on désire affecter de nouveaux individus à une des modalités de Y, connaissant les modalités prises par les X_i.

Un exemple de tels problèmes se rencontre en diagnostic automatique où l'on décrit un patient par une série de symptomes et où on cherche à l'affecter à un type de traitement ou de maladie.

Le fait que les variables descriptives soient qualitatives interdit a priori de recourir aux méthodes usuelles de la discrimination (analyse linéaire discriminante, programme BMDO7M) qui sont bâties pour des variables numériques et utilisent de plus des hypothèses de distribution normales à l'intérieur de chacun des k groupes définis par la variable Y.

Si la possibilité de représenter une variable qualitative à m modalités par l'ensemble des m variables indicatrices de ses modalités (forme disjonctive) laisse croire que l'on pourrait se ramener à une analyse discriminante sur $m_1 + m_2 + ..+ m_p$ variables numériques, il faut prendre garde d'une part au fait qu'il y a p relations linéaires liant ces variables et surtout, d'autre part, que l'on ne peut isoler une modalité d'une variable car on est toujours contraint de traiter simultanément toutes les indicatrices d'une même variable qualitative. La représentation des variables par leurs indicatrices permet de définir des scores, c'est-à-dire des codages des modalités, que l'on pourra utiliser directement dans des fonctions de classement.

Le nombre de variables explicatives étant souvent pléthorique faut d'avoir une idée a priori sur les "bonnes" variables, on est généralement conduit à en faire une sélection dans le double but de :

- ne garder que les variables les plus pertinentes
- limiter le volume des calculs et le nombre des scores.

II - LES METHODES EXISTANTES

Divers auteurs ont élaboré depuis quelques années des méthodes et des programmes de discrimination sur variables qualitatives.

Aux Etats-Unis, ANDREWS et MESSENGER avec M.N.A (Multivariate nominal scale analysis) proposent un programme de régression simultanée des k indicatrices de Y sur l'ensemble des indicatrices des X_i. La dépendance linéaire entre les indicatrices de chaque variable est compensée par un système de

contraintes sur les coefficients de régression qui en facilite l'interprétation : les valeurs estimées $\hat{y}_j$ (e_i) des indicatrices des modalités de Y s'interprètent comme les probabilités d'appartenance de l'individu e_i aux différentes classes. Dans la relation $\hat{y}_j (e_i) = \hat{\alpha}_j + \sum_l \sum_h \hat{a}_{jlh} x_{ilh} \hat{\beta}_j$ est la probabilité a priori d'appartenance à la classe j et $\hat{a}_{ilh}$ le correctif à apporter si l'individu i prend la modalité h de X_l.

En France, mis à part l'utilisation de l'analyse des correspondances à des fins de discrimination visuelle ou numérique, effectuée sur la juxtaposition des tableaux de contingence croisant Y et les X_i (dimension k x Σm_i) on relève deux méthodes assez largement utilisées :

1) Celle de M. MASSON (CANOSTEP) qui effectue simultanément le calcul des scores et la sélection des variables, dans le cas de deux groupes seulement, par le procédé suivant : au premier pas on sélectionne X_l optimisant l'analyse canonique de Y contre X_i (plus fort χ^2), on transforme X_l en une variable numérique $\underline{x}_l$ en la codant selon le premier facteur canonique. On détermine ensuite X_2 donnant la meilleure corrélation canonique de Y contre $\underline{x}_l$ et X_i ; on code alors X_2 en $\underline{x}_2$ de façon à écrire la variable canonique $\underline{\xi}$ associée à $(\underline{x}_l\ X_2)$ sous la forme $\underline{\xi} = \underline{x}_l + \underline{x}_2$ on continue en remplaçant $\underline{x}_l$ par $\underline{\xi}$ etc.. Cette méthode est donc semblable à une régression ascendante mais où les coefficients de régression des premières variables introduites ne sont pas modifiés par l'introduction des variables suivantes ce qui n'est optimum que pour des prédicteurs indépendants.

2) Celle de G. SAPORTA (DISQUAL) qui effectue dans un premier temps une sélection progressive des variables en utilisant les coefficients de TSCHUPROW $T_{Y,Xi} = \frac{\chi^2_{Y,Xi}}{\sqrt{(k-1)(m_i-1)}}$ comme des coefficients de corrélation entre variables qualitatives.

Au pas i on introduit X_i rendant maximum le coefficient de corrélation partielle entre Y et X_i à $X_l, X_2, \ldots X_{i-1}$ fixés ; dans un deuxième temps on réalise une analyse discriminante linéaire sur les composantes issues de l'analyse des correspondances multiples du tableau disjonctif des variables sélectionnées.

Aucune de ces méthodes ne traite de façon satisfaisante les interactions d'ordre supérieur à 2 : l'AFC sur tableaux juxtaposés ne tient même pas compte des liaisons simples entre prédicteurs, MNA fournit des résultats absurdes en présence d'intéraction (probabilités négatives ou supérieures à 1), quant aux coefficients de Tschuprow ils ne mesurent qu'imparfaitement les liaisons partielles.

Certes l'introduction de variables croisées fait disparaitre les effets non additifs des intéractions, encore faut-il les détecter : c'est l'ambition du modèle log-linéaire, modèle probabiliste fondé sur la décomposition du tableau de contingence à p+1 dimensions croisant toutes les variables et où l'on prédit par le maximum de vraisemblance le logarithme de la probabilité des cases. En travaillant sur les probabilités conditionnelles à Y fixé on peut en faire un outil de discrimination (DAUDIN) mais l'obstacle tient au nombre des observations souvent insuffisant et à la taille prohibitive du tableau croisé total.

A la suite de nombreuses utilisations de DISQUAL, il nous est apparu nécessaire de proposer une méthodologie de discrimination et quelques perfectionnements au programme actuel concernant les procédures d'affectation tout en conservant la possibilité d'un scoring qui satisfait une majorité d'utilisateurs. Ce programme remanié sera opérationnel à la date du congrès.

III - SELECTION DES VARIABLES ET PRISE EN COMPTE DES INTERACTIONS

A la place de la sélection automatique par les T de TSCHUPROW nous proposons une procédure en deux phases ;

1) A la suite des tris croisant d'une part Y avec chacun des X_i on élimine les X_i présentant une trop faible liaison avec Y en se basant sur la probabilité de dépassement du chi-deux observé.

2) Pour tenir compte des intéractions on utilise ensuite un modèle log-linéaire (par exemple le programme BMDP3F) afin, d'une part, d'obtenir pour des sous-ensembles de variables de taille raisonnable (vu les capacités de calcul) le tableau croisé multi-dimensionnel conditionné par Y et, d'autre part, d'utiliser les tests statistiques permettant d'estimer le niveau des différentes interactions.

Ceci aboutira à créer q nouvelles variables par des croisements des anciennes variables. Dans chacune de ces deux phases en procédera à des regroupements éventuels de modalités afin d'alléger les calculs ultérieurs en se basant sur trois exigences :

- avoir des effectifs suffisants dans les différentes cases.
- ne regrouper que les modalités ayant des profils sur Y peu différents.
- éviter des regroupements illogiques (utilisation du savoir préalable de l'utilisateur).

Tout cela rend impossible une automatisation complète de la sélection qui se déroulera de manière intéractive.

IV - PROCEDURES DE DISCRIMINATION

Désormais, la discrimination s'effectue sur l'ensemble des q variables créées précédemment dont les effets non linéaires ont été réduits. Chaque variable étant éclatée selon les indicatrices de ses modalités, le tableau de données se présente sous la forme :

$$\begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} k \quad m_1 \quad m_2 \quad m_q \\ \left(\begin{array}{c|c|c|c|c|c} Y & X_1 & X_2 & \cdots & X_q \end{array} \right) \end{array} = \left(\begin{array}{c|c} Y & X \end{array} \right)$$

L'espace engendré par les Σm_i colonnes de X est de dimension $\Sigma m_i - q$ si on se limite aux variables de moyenne nulle, car pour chaque X_i la somme des indicatrices vaut $\underline{1}$.

Dans un premier temps, nous remplaçons l'ensemble des indicatrices par un sous-ensemble de variables orthogonales formant une base de l'espace engendré: nous choisissons ici les composantes de l'analyse des correspondances de X qui fournissent la meilleure synthèse de l'information contenue dans X et ont la propriété d'être de variance un et non corrélées deux à deux. Comme le nombre total de composantes risque d'être élevé, on procède ensuite à une réduction fondée sur un double critère : élimination des composantes de faible inertie afin d'éviter de discriminer sur du bruit ; et surtout élimination des composantes de faible pouvoir discriminant, celui-ci étant défini par la variance inter-groupe. Ces pouvoirs discriminants sont additifs en vertu de l'orthogonalité des composantes de l'AFC de X. Le nombre de composantes sélectionnées est fixé soit arbitrairement soit selon un pourcentage du pouvoir discriminant global choisi par l'utilisateur.

Si les composantes retenues sont z_1, $\underline{z}_2$, .., $\underline{z}_s$ on est alors ramené à une discrimination sur s variables numériques explicatives réduites et non corrélées.

La discrimination comprend alors deux volets : d'une part, une visualisation du type analyse de données, d'autre part une procédure de classement.

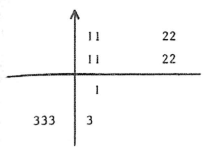
1) Visualisation par l'analyse factorielle discriminante

Cette partie descriptive n'est possible que pour k>3.

L'analyse factorielle discriminante ou analyse canonique de Y contre $Z = (\underline{z}_1 \ \underline{z}_2 \ \cdots \ \underline{z}_s)$ est particulièrement facile à réaliser puisque par construction la matrice de variance de Z est égale à I.

Les facteurs discriminants, au nombre de k-1, sont alors les vecteurs propres de $Z'D_p Y(Y'D_pY)^{-1} Y'D_p Z$ (où D_p est la matrice diagonale des poids

des individus), ils donnent les combinaisons linéaires des $\underline{z}_j$ séparant le mieux possible les k groupes. Les deux premiers caractères discriminants permettent alors de représenter sur un graphique plan les différents groupes :



on représentera les individus repérés par leur numéro de groupe et les modalités des variables explicatives par le barycentre des individus les possédant.

2) Calcul des scores et procédure de classement

a) Rappel de la procédure de DISQUAL

Dans DISQUAL, on utilise pour définir les formules de classement une méthode dérivée de l'analyse linéaire discriminante où on affecte un nouvel individu à la classe dont il est le plus proche au sens de la distance au centre de gravité, la distance étant calculée au moyen de la métrique définie par l'inverse de la matrice de variance totale V des k groupes. Cette métrique avait été choisie pour sa simplicité puisque les $\underline{z}_j$ étant orthonormés, V = I.

Algébriquement, si $\underline{e}$ est le vecteur de description logique de l'individu e :

$$\underline{e}' = \left(\begin{array}{c|c|c|c} m_1 & m_2 & & m_q \\ 0100 & 100 & \cdots & 0010 \end{array} \right)$$

et si U est la matrice des s facteurs sélectionnés de l'AFC de X :

$\underline{z}_j = X \underline{u}_j$. Avec $U = \left(\underline{u}_1 \ \underline{u}_2 \ \cdots \ \underline{u}_s \right)$ de dimension $(\Sigma m_i \ ; s)$

les coordonnées de $\underline{e}$ sur les $\underline{z}_j$ sont données par $U'\underline{e}$; le carré de la distance au centre de gravité g_i du i^e groupe est :

$$d^2 (e \ ; g_i) = \underline{e}'U \, U' \underline{e} + g_i' g_i - 2 \, \underline{g}_i' \, U'\underline{e}$$

Ceci revient à définir un ensemble de k fonctions discriminantes, une par groupe, c'est-à-dire k vecteurs de codages des modalités (les scores), chaque vecteur étant égal, à la constante $\underline{g}_i' \underline{g}_i$ près, à $-2 \, U\underline{g}_i$; on classe alors un individu dans le groupe pour lequel la somme des scores correspondant aux modalités prises est minimale.

Si W désigne la matrice de variance intragroupe (moyenne des matrices de variance W_i de chaque groupe), on sait qu'il est équivalent d'utiliser V^{-1} ou W^{-1} : cette procédure n'est justifiée sur le plan décisionnel que si les matrices de variances observées W_i sont peu différentes, c'est-à-dire si les matrices théoriques sont égales. W est alors l'estimation de Σ matrice de variance commune.

b) Nouvelles possibilités

L'hypothèse d'équicovariance n'est en réalité que rarement vérifiée et la procédure précédente n'est plus optimale.

Dans ce cas nous utiliserons une méthode bayesienne tenant compte à la fois des probabilités a priori p_j d'appartenance à un groupe et de la distribution de chaque groupe.

La méthode bayesienne revient à affecter $\underline{e}$ au groupe j pour lequel la probabilité a posteriori $\frac{p_j \ f_j \ (\underline{e})}{\Sigma p_k \ f_k \ (\underline{e})}$ est maximale, ce qui revient à maximiser $p_j \ f_j \ (e)$ où f_j est la densité de probabilité des variables explicatives sur le groupe j.

Deux options sont alors prévues pour l'estimation des densités : on prend pour f_j une loi normale s-dimensionnelle de moyenne estimée $\underline{g}_j$ et de matrice variance estimée W_j. L'utilisation de lois normales, qui semble surprenante, conduit en fait à d'excellents résultats en pratique (NAKACHE).

On aboutit alors à une méthode de classement non linéaire qui consiste à affecter $\underline{e}$ au groupe j pour lequel :
$(U'\underline{e} - g_j)' \ W_j^{-1} \ (U'\underline{e} - g_j) + \text{Log det } W_j - 2 \text{ Log } p_j$ est minimal. Ceci revient à classer un individu selon la somme des scores des paires de modalités prises.

On procède à une estimation non paramétrique de la densité par la méthode des noyaux de PARZEN : on n'a plus alors de formule explicite de classement. L'algorithme est semblable à celui d'ALLOC (HABBEMA, HERMANS) .

La méthode bayesienne peut cependant conduire à des calculs lourds. De plus l'utilisation de formules non linéaires ne donne de bons résultats qu'avec un échantillon de taille élevée car le nombre de paramètres à estimer devient très important. Dans le cas contraire, ou si l'on veut une formule utilisant des scores additifs, on gardera la discrimination linéaire de DISQUAL. Dans le cas de deux groupes, la méthode d'ANDERSON-BAHADUR lui sera préférée car elle donne la meilleure formule linéaire de classement tenant compte du fait que $W_1 \neq W_2$.

Toutes ces options ainsi que la prise en compte de coûts de classements différents selon les groupes, sont offertes à l'utilisateur de notre programme.

REFERENCES

ANDERSON T.W ; BADAHUR R.R. (1962) "Classification into two multivariate normal distributions with different covariance matrices" Ann. Math-Stat 33 p.420.

ANDREWS F.M. ; MESSENGER R.C. (1973) Multivariate Nominal Scale Analysis University of Michigan.

CAPPE DE BAILLON C ; SAPORTA G. (1976) DISQUAL Manuel d'utilisation Note COREF N°14

DAUDIN J.J (1978) Etude de la liaison entre variables aléatoires. Régression sur variables qualitatives. Thèse 3e cycle Université d'Orsay.

HABBEMA J.D.F. ; HERMANS J. (1976) "The Alloc package Multigroup Discriminant analysis Programs based on direct density estimation". COMPSTAT (Vienne).

KSHIRSAGAR A.M. (1972) Multivariate analysis Marcel Dekker New York.

MASSON M. (1974) Processus linéaires et analyse de données non linéaires. Thèse de doctorat es Sciences Université de Paris VI.

NAKACHE J.P. (1978) Méthodes multidimensionnelles de classement Polycopié ISUP Université Paris VI.

NERLOVE (1973) Univariate and multivariate log-linear and logistic models. Report R 1306 EDA/NIH Rand corporation Santa Monica.

SAPORTA G. (1976) Discriminant analysis when all the variables are nominal. Spring meeting og the Psychometric Society. Murray Hill N.J. U.S.A.

SAPORTA G. (1977) Une méthode et un programme d'analyse discriminante pas à pas sur variables qualitatives. Journées internationales Analyse des données et Informatique Versailles Colloques IRIA p 201-210.