Une méthode et un programme d'analyse discriminante sur variables qualitatives

Gilbert Saporta

UNE MÉTHODE ET UN PROGRAMME D'ANALYSE DISCRIMINANTE PAS A PAS SUR VARIABLES QUALITATIVES

Gilbert SAPORTA

Institut Universitaire de Technologie
Université René Descartes

n individus sont répartis en k groupes définis par les modalités d'une variable Y . On veut "expliquer" cette partition et classer des individus ultérieurs grâce à p variables qualitatives $X_1, X_2, \ldots, X_p$ à $m_1, m_2 \ldots, m_p$ modalités.

Si on remplace chaque X_j par l'ensemble des m_j variables indicatrices de ses modalités on est ramené à un problème d'analyse discriminante usuel avec $m_1 + m_2 + \ldots + m_p$ variables numériques explicatives. L'obtention de fonctions discriminantes s'identifie alors à un problème de codage optimal en ce sens que tout revient à transformer les p variables qualitatives en p variables numériques de sorte que le pouvoir discriminant de la fonction discriminante cherchée soit maximal.

Une première difficulté provient du fait que l'ensemble des variables indicatrices n'est pas de plein rang : il existe donc une infinité de codages conduisant au même optimum. Une seconde difficulté surgit si on veut transposer les méthodes de sélection progressive couramment utilisées en discrimination ordinaire : on ne peut en effet introduire une indicatrice seule sans celles qui correspondent à la même variable qualitative. La sélection ne peut se faire que parmi les p variables initiales donc en traitant les indicatrices par bloc.

Une méthode et un programme nommé DISQUAL (DIScrimination sur variables QUALitatives) ont été développés dans le cadre du contrat COREF-DGRST 75-7-0230 afin de réaliser :

a) Une sélection progressive des variables explicatives en utilisant les propriétés démontrées en [10] du coefficient de TSCHUPROW.

b) L'analyse discriminante sur les variables retenues par la méthode des "facteurs z " [9] pour aboutir d'une part à des formules de classement, d'autre part pour plus de deux groupes à une analyse factorielle discriminante.

I - SELECTION PROGRESSIVE D'UN SOUS-ENSEMBLE DE q VARIABLES EXPLICATIVES

1. Indicatrices associées à une variable X à m modalités

A X associons l'ensemble des m indicatrices $x_1, x_2, \ldots, x_m$ qui lui est équivalent :

$x_j = 1$ pour un individu prenant la modalité j

$x_j = 0$ sinon

Pour les n individus étudiés la variable X est donc équivalente au tableau X :

$$X = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccc} x_1, & x_2, & \ldots, & x_m \end{array} \\ \begin{pmatrix} 0 & 1 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ & & & \\ 0 & 0 & \ldots & 1 \end{pmatrix} \end{array}$$

Le sous-espace vectoriel W de $\mathbb{R}^n$ engendré par les indicatrices est l'ensemble des variables discrètes à m valeurs réalisant une quantification ou codage de X :

$$W = \{\xi \mid \xi = X\,a \quad a \in \mathbb{R}^m\}$$

a est le vecteur des codages, ses coordonnées $a_1, a_2, \ldots a_m$ sont les codes des différentes modalités de X.

Si les n individus sont pondérés par des poids $p_1, p_2, \ldots p_n$ on munira $\mathbb{R}^n$ de la métrique : $D_p = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix}$. Les variables centrées ξ sont alors les variables D_p-orthogonales au vecteur 1.

2. Le coefficient de TSCHUPROW comme cosinus d'angle

A deux variables qualitatives X_1 et X_2 associons les tableaux X_1, X_2 et les espaces W_1 et W_2.

W_1 et W_2 sont caractérisés par les projecteurs D_p-orthogonaux A_1 et A_2

$$\left(A_1 = X_1 \, (X_1' \, D_p \, X_1)^{-1} \, X_1' \, D_p \right).$$

Suivant Y. ESCOUFIER [7], munissons l'ensemble des matrices n,n D_p-symétriques du produit scalaire de la trace :

$$\langle P \,;\, Q \rangle = \text{Trace } (P\,Q)$$

Alors $\langle A_1 \,;\, A_2 \rangle = \text{Trace } A_1\,A_2 = \sum \lambda_i$ où les λ_i sont les carrés des coefficients de corrélation canoniques de X_1 et X_2. Comme X_1 et X_2 sont des tableaux d'indicatrices on sait que : $\text{Trace } A_1\,A_2 = 1 + \phi^2$ où ϕ^2 est le coefficient de K. PEARSON $\phi^2 = \chi^2/n$.

Comme les sous-espaces W_1 et W_2 ont en commun le vecteur 1 il convient de considérer en fait les sous-espaces W_{10} et W_{20} de dimensions m_1-1 et m_2-1 des variables centrées codant X_1 et X_2 et les projecteurs associés A_{10} A_{20}.

On a alors :

$$\text{Trace } (A_{10}\,A_{20}) = \phi^2$$ et le cosinus associé à ce produit scalaire est alors le coefficient de TSCHUPROW entre X_1 et X_2 :

$$\cos\,(A_{10}\,A_{20}) = \frac{\phi^2}{\sqrt{\text{Trace } A_{10}^2 \;\; \text{Trace } A_{20}^2}} = \frac{\phi^2}{\sqrt{(m_1-1)\;(m_2-1)}}$$

Ce coefficient dont la nullité est équivalente à l'indépendance et l'égalité à 1 à la relation fonctionnelle, a ainsi les propriétés d'un cosinus, c'est donc l'analogue d'un coefficient de corrélation entre variables qualitatives si on identifie X_1 au projecteur A_{10}.

3. Méthode de sélection

Nous pouvons donc adopter une méthode semblable à la régression progressive.

La première variable sélectionnée X_1 est celle qui a le plus fort coefficient de Tschuprow avec Y. On calcule alors des coefficients de Tschuprow partiels d'ordre 1 par la formule :

$$T_{YX_\perp/X_1} = \frac{T_{YX_\perp} - T_{YX_1}\,T_{X_\perp X_1}}{\sqrt{(1 - T_{YX_1}^2)\;(1 - T_{X_\perp X_1}^2)}}$$

La deuxième variable sélectionnée X_2 est celle pour laquelle $T_{YX_\perp/X_1}$ est maximal : elle a pour propriété d'être peu liée à X_1 tout en étant la plus liée possible avec Y. On calcule alors des coefficients partiels du second ordre $T_{YX_\perp/X_1\,X_2}$ en itérant la formule précédente et ainsi de suite.

Le processus s'arrête lorsqu'on a sélectionné un nombre de variables fixé à l'avance ou lorsque le coefficient de TSCHUPROW multiple que nous définissons par :

$$1 - T_{Y\,;\,X_1,X_2 \ldots X_q}^2 = (1 - T_{YX_1}^2)\;(1 - T_{YX_2}^2) \ldots (1 - T_{YX_q/X_1,X_2 \ldots X_{q-1}}^2)$$

n'augmente plus suffisamment.

II - ANALYSE DISCRIMINANTE SUR LES q VARIABLES RETENUES

1. Analyse factorielle discriminante

L'analyse factorielle discriminante est l'analyse canonique des tableaux V et X où V est le tableau des indicatrices de Y et X la matrice dont les blocs sont $X_1, X_2, \ldots, X_q$

$$X = (X_1 | X_2 | \ldots | X_q)$$

Dans le cas usuel on sait que les facteurs discriminants sont les vecteurs propres de :

$$(X'\,D_p\,X)^{-1}\,X'\,D_p\,V\,(V'\,D_p\,V)^{-1}\,V'\,D_p\,X$$

mais ici $X' D_p X$ qui est le tableau de BURT des q variables explicatives n'est pas inversible car le vecteur $\mathbb{1}$ est commun à $W_1, W_2, \ldots, W_q$ (et aussi à W_Y). L'espace W engendré par les colonnes de X est donc de dimension $m_1 + m_2 + \ldots + m_q - q$ si on ôte les variables constantes.

Pour résoudre notre problème il faut donc remplacer les colonnes de X par une base de W ce qui revient à choisir une pseudo-inverse de $X' D_p X$. En termes de codages ceci est équivalent à imposer une contrainte au codage de chaque variable X_j. Nous avons montré en [9] que prendre la pseudo-inverse de MOORE-PENROSE était équivalent au choix des contraintes de moyenne nulle sur les X_j.

Une méthode pour obtenir cette pseudo-inverse consiste alors à remplacer X par la matrice Z des facteurs de l'analyse des correspondances de X, ce que les praticiens appellent l'analyse du tableau disjonctif complet des X_j [6]. Les colonnes de Z : $z_1, z_2 \ldots$ sont des variables centrées D_p-orthogonales entre elles (i.e. non corrélées) que l'on choisit de variance unité. Les "facteurs z" sont identiques aux variables auxiliaires définies par J.D. CARROLL pour l'analyse canonique généralisée de $X_1, X_2, \ldots, X_q$ qui ne sont autres que les "composantes principales d'échelle" de GUTTMAN, voir [2],[7],[9].

Rappelons que si D est la diagonale de $X' D_p X$ les z_j sont tels que :

$$\begin{cases} z_j = X u_j \\ D^{-1} X' D_p X u_j = \lambda_j u_j \end{cases}$$

L'analyse factorielle discriminante de Y et Z est alors particulièrement simple car $Z' D_p Z = I$ puisque les z_j sont centrés réduits et non corrélés.

En pratique il est inutile de garder les $m_1 + m_2 + \ldots + m_q - q$ facteurs z et on pourra effectuer la discrimination sur une partie seulement d'entre eux [5]. Ainsi dans DISQUAL on élimine d'abord les z_j de trop faible inertie (λ_j) afin d'éviter de discriminer sur du bruit, puis dans un deuxième temps après avoir reclassé les z_j restant par pouvoir discriminant décroissant (cet ordre est en général différent de celui des λ_j) on ne garde que ceux dont le pouvoir discriminant cumulé est suffisant..

Le pouvoir discriminant de z_j qui est sa variance inter-groupe vaut :

$$z_j' D_p Y (Y' D_p Y)^{-1} Y' D_p z_j$$

et ces pouvoirs sont additifs car les z_j sont non corrélés.

Il existe $k-1$ combinaisons linéaires discriminantes des z_j de la forme $\Sigma c_j z_j$ fournies par l'équation :

$$Z' D_p Y (Y' D_p Y)^{-1} Y' D_p Z c = \mu c$$

donc si $k > 2$ on peut représenter les n individus dans le plan défini par les deux premières variables discriminantes ainsi que les modalités des variables explicatives considérées comme centre de gravité des individus les possèdant ; ce centre de gravité est alors projeté en élément supplémentaire. Dans le cas de 2 groupes DISQUAL sort l'histogramme de l'unique variable discriminante.

Dans tous les cas l'utilisation des facteurs z permet une représentation des individus, qui ne tient compte que des variables explicatives et non de Y mais dont l'intérêt statistique est évident puisqu'il s'agit de l'analyse des correspondances des X_j.

2. Procédure de classement

Les individus étant désormais décrits par les facteurs z retenus à l'étape précédente, l'affectation d'un individu e à l'un des k groupes se fait selon une procédure usuelle de distance au centre de gravité.

Dans sa version actuelle DISQUAL utilise la distance de MAHALANOBIS ce qui revient à la procédure suivante puisque la matrice de variance-covariance des z_j n'est autre que la matrice unité.

Soit U la matrice à $m_1 + m_2 + \ldots + m_q$ lignes dont les colonnes sont les vecteurs u_j tels que $z_j = X u_j$; si e est le vecteur de description logique de l'individu e :

$$e' = (0100|100| \ \ldots \ |0010)$$

les coordonnées de e sur les z_j sont données par $U'e$; le carré de distance de e au centre de gravité g_i du $i^{ème}$ groupe vaut alors :

$$d^2(e, g_i) = e' U U' e + g_i' g_i - 2 g_i' U' e .$$

L'individu e sera affecté au groupe i_o si $g_i' g_i - 2 g_i' U' e$ est minimal pour $i = i_o$ ceci revient en pratique à avoir pour chaque groupe une fonction discriminante du type suivant :

On donne une valeur numérique à chaque modalité des variables $X_1, X_2, \ldots, X_q$; on ajoute les valeurs correspondant aux modalités prises par e et on affecte e là où la somme est minimale.

Si le nombre des données le permet il y a bien sûr intérêt à valider les fonctions discriminantes obtenues sur un échantillon-test.

III - UN EXEMPLE D'APPLICATION : LE CHOIX D'UN MODE DE TRANSPORT

Les résultats suivants sont reproduits avec l'aimable autorisation de A. LEROUX (SNCF) [1]. Afin de respecter le caractère confidentiel de l'étude quelques valeurs numériques ont été modifiées et les noms de certaines variables sont volontairement imprécis.

Il s'agissait d'expliquer le choix d'un des 3 modes de transport suivant : voiture, train, avion, à l'aide de variables socio-économiques (CSP ; composition du ménage par trancne d'âge; possession ou non d'une voiture ; nombre de voyages effectués dans l'année ...) et de variables caractérisant le déplacement (région d'arrivée, de départ, distance parcourue, date ...). Les variables quantitatives ont été discrétisées par découpage en classes. Au total il y avait 32 variables explicatives totalisant 265 modalités. L'échantillon, constitué de 5000 déplacements privés de plus de 100 kms effectués en 1972-73 par des ménages de la moitié nord de la France, a été partagé en un échantillon-test de 2000 déplacements et un échantillon de base de $n = 3000$ déplacements qui par suite de pondération correspondait à un effectif redressé de 4333 individus.

Les trois premières variables discriminantes sont les suivantes :

1 VOIT possession ou non d'un véhicule particulier
2 DIST distance parcourue
3 TGRV taille du groupe de voyage.

La variation du coefficient de Tschuprow multiple (voir figure 1) montre qu'il est inutile d'aller au-delà de q = 8 variables explicatives.

L'analyse discriminante se fera donc à l'aide de ces 8 variables qui totalisent 85 modalités.

L'analyse du tableau disjonctif des 8 variables sélectionnées aboutit donc à 77 facteurs z. Parmi eux seuls les 10 plus discriminants ont été retenus qui sont par ordre décroissant les facteurs numéros :

1, 5, 4, 20, 3, 42, 15, 21, 33, 19

représentant 69 % du pouvoir discriminant total et 22 % de l'inertie.

L'analyse factorielle discriminante sur les 10 facteurs retenus conduit à la représentation de la figure 2 où ont été projetés les centres de gravité des individus possèdant les modalités des 3 variables DIST, TGRV, VOIT, ainsi que les centres des 3 groupes (TRAIN, AVION, VOITURE). Rappelons que puisque k = 3 il n'y a que deux axes discriminants. On voit sur ce graphique que les tranches de distance les plus favorables au train sont DIST 7, DIST 8, DIST 6, DIST 5 et DIST 4.

Le programme calcule ensuite les trois fonctions discriminantes relatives à chaque mode de transport. En additionnant les valeurs obtenues par les modalités prises par un individu on obtient une grandeur homogène à une distance à chaque groupe à une constante près d'où des termes négatifs.

	voiture	train	avion
DIST 1	- .004	.009	.029
DIST 2	- .001	.005	.020
DIST 3	.003	- .003	.017
DIST 4	.005	- .007	.011
DIST 5	.006	- .009	.007
DIST 6	.007	- .011	.009
DIST 7	.014	- .026	.016
DIST 8	.011	- .020	.021
DIST 9	.001	- .001	.008
DIST 10	.003	.003	- .059

Le tableau ci-contre fournit les fonctions discriminantes pour la variable distance. Changées de signe les fonctions peuvent s'interpréter comme des utilités au niveau d'une variable isolée toutes choses égales par ailleurs. La figure 3 représente les courbes d'utilité des trois modes de transport en fonction de la distance.

Les fonctions discriminantes sont ensuite utilisées pour réaffecter les individus de l'échantillon de base dans chacun des groupes ce qui aboutit au tableau de classement :

Groupe attribué / Groupe d'origine	Voiture	Train	Avion
Voiture	1875	677	237
Train	339	852	160
Avion	6	45	142

(effectifs redressés)

Le pourcentage total de bien-classés est de 66 %. Les calculs effectués sur l'échantillon-test conduisant à des résultats comparables ne sont pas reproduits ici.

24 % des déplacements en automobile sont considérés comme effectués par le train ; inversement, 25 % des déplacements en train sont considérés par le modèle comme des déplacements en voiture. Plutôt que d'incriminer le modèle on peut interpréter ces chiffres comme traduisant l'importance de la zone concurrentielle entre ces deux modes de transport. Le programme permet alors de lister les mal-classés afin d'établir plus précisément leurs caractéristiques.

L'ensemble des calculs effectués par DISQUAL sur cette application a demandé 2'50" sur IBM 370-168.

BIBLIOGRAPHIE

[1] Bouroche, J.M., Leroux, A. (1977). *Analyse des données qualitatives à but décisionnel : méthode et application.* Congrès ESOMAR, Oslo.

[2] Bouroche J.M., Saporta G., Tenenhaus M. (1975). *Generalized canonical analysis of qualitative data.* U.S. Japan Seminar on multidimensional scaling and related methods San Diego.

[3] Cappe de Baillon C., Saporta G. (1976). *DISQUAL - Manuel d'utilisation.* Note de travail COREF n° 14.

[4] Carroll J.D. (1968). *A generalization of canonical analysis to three or more sets of variables.* Proceedings of 76th Convention of the A.P.A. 227-228.

[5] Dehedin J. (1975). *Discrimination sur variables qualitatives.* These 3e cycle, Université de Paris VI.

[6] Lebart L. (1975). *L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples.* Consommation n° 2.

[7] Masson M. (1974). *Processus linéaires et analyse de données non linéaire.* These de doctorat es Sciences. Université de Paris VI.

[8] Pages P.J., Escoufier Y., Cazes P. (1976). *Opérateurs et analyse de tableaux à plus de deux dimensions.* Cahiers du BURO, n° 25.

[9] Saporta G. (1975). *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives.* Thèse 3e cycle. Université Paris VI.

[10] Saporta G. (1976). *Discriminant analysis when all the variable are nominal.* Spring meeting of the Psychometric Society. Murray Hill N.J..

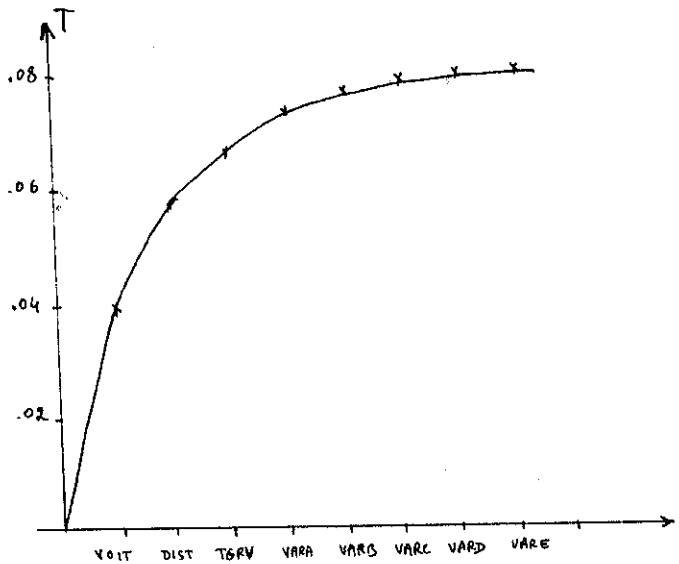

Figure 1 - Evolution du coefficient de TSCHUPROW multiple en fonction des variables introduites.

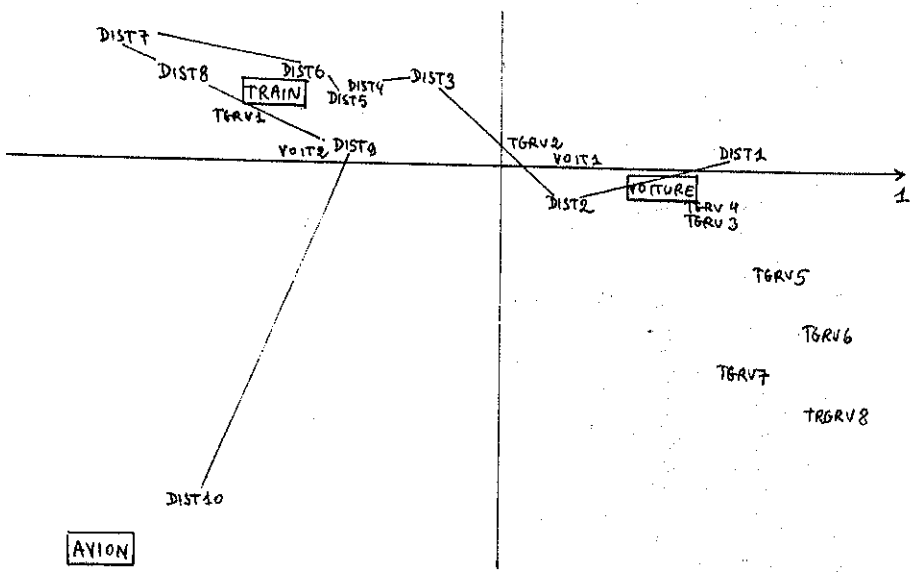

Figure 2 - Représentation des modalités des variables dans le plan discriminant.

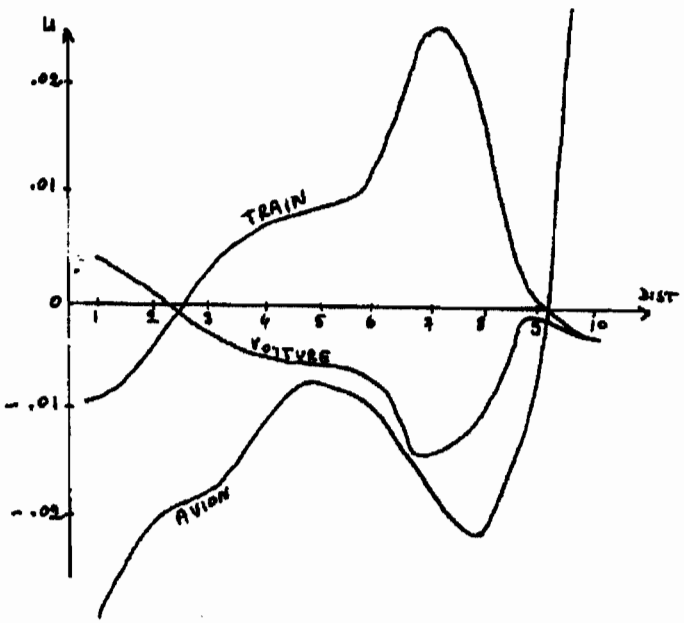

Figure 9 - Courbes d'utilité des 3 modes de transport selon la variable distance toutes choses égales par ailleurs.