Une méthode de Discrimination non paramétrique

S. Mahjoub, Gilbert Saporta

REVUE DE STATISTIQUE APPLIQUÉE

S. MAHJOUB
G. SAPORTA

Une méthode de Discrimination non paramétrique

Revue de statistique appliquée, tome 42, n° 2 (1994), p. 99-113
<http://www.numdam.org/item?id=RSA_1994__42_2_99_0>

Rev. Statistique Appliquée, 1994, **XLII** (2), 99-113

UNE MÉTHODE DE DISCRIMINATION NON PARAMÉTRIQUE

S. Mahjoub (1), G. Saporta (2)

(1) Faculté des Sciences Economiques et de Gestion,
Département des Méthodes Quantitatives
Rte de l'Aérodrome Km 4,5 - BP 69 - 3028 Sfax, Tunisie
(2) Conservatoire National des Arts et Métiers
Département de Mathématiques
292 rue St Martin 75141 Paris cedex 03

RÉSUMÉ

Nous proposons une nouvelle méthode de discrimination non paramétrique s'appliquant aussi bien à des données quantitatives que qualitatives. Nous l'avons comparée aux méthodes classiques par le taux d'erreur, estimé par validation croisée.

Mots-clés : *Discrimination, Analyse factorielle, Validation croisée, Estimateur non paramétrique, Distance en variation de Kolmogoroff.*

SUMMARY

A new non parametric discriminant analysis is presented. The predictors may be quantitative or qualitative. The method is compared to classical linear discrimination and nearest neighbor method by using error rate estimated by cross-validation.

Key-words : *Discrimination, Principal Component Analysis, Cross-Validation, Non Parametric Estimation, Kolmogoroff Distance.*

Introduction

La méthode de discrimination exposée dans cet article repose sur trois idées :

a) l'utilisation des analyses factorielles locales ce qui permet d'orthogonaliser les variables et d'améliorer les estimateurs de densité.

b) l'emploi de noyaux gaussiens pour l'estimateur de densité.

c) une technique non paramétrique de sélection de variables basée sur la distance en variation de Kolmogoroff.

Après avoir remplacé les prédicteurs initiaux par les composantes principales, nous estimons les densités par un estimateur non paramétrique dépendant d'un paramètre α appartenant à l'intervalle $\left]0, \frac{1}{2}\right[$ et ceci pour chaque groupe.

La méthode mise en œuvre pour déterminer les paramètres α optimaux nous fournira en même temps une estimation, par validation croisée, du taux d'erreur.

Nous sélectionnons, enfin, les variables qui maximisent la distance en variation de Kolmogoroff entre les groupes.

1. Analyse factorielle par groupe et les métriques utilisées

Des analyses factorielles effectuées séparément pour chaque groupe fournissent des systèmes de variables quantitatives non correlées bien adaptées à chaque groupe qui seront utilisées pour l'estimation de la fonction de densité du groupe par la méthode des noyaux produits.

Dans le cas où tous les prédicteurs sont quantitatifs, on effectue une Analyse en Composantes Principales (A.C.P) normée sur chaque groupe, tandis que s'ils sont tous qualitatifs, on effectue une Analyse des Correspondances Multiples. Lorsqu'on a un mélange de prédicteurs quantitatifs et qualitatifs, on utilise la méthode suivante :

Soit X le tableau des prédicteurs quantitatifs centré pour le groupe considéré et $A = (A_1 A_2 \cdots A_q)$ le tableau disjonctif des indicatrices des q prédicteurs qualitatifs, on effectuera l'ACP du tableau juxtaposé (XA) avec la métrique M suivante :

$$M = \begin{pmatrix} \mathrm{Diag} X'X & 0 \\ 0 & \mathrm{Diag} A'A \end{pmatrix}^{-1}$$

ce qui revient à réduire les variables quantitatives et à utiliser la métrique du χ^2 pour les indicatrices associées aux variables qualitatives.

2. L'estimateur des densités

La constante de lissage de l'estimateur de Parzen ne tient pas compte des corrélations entre les variables et des différences entre les dispersions des différents prédicteurs.

L'estimateur proposé n'a pas ces défauts [11]. Il s'écrit

$$\tilde{f}_n(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{(2\pi)^{p/2} \left| \frac{\Sigma}{n^{2\alpha}} \right|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - X_j)^T \left(\frac{\Sigma}{n^{2\alpha}} \right)^{-1} (\mathbf{x} - X_j) \right\}$$

α une constante appartenant à l'intervalle $\left]0, \frac{1}{2}\right[$.

$X_1, X_2, \cdots, X_n$ n observations indépendantes de la variable X de fonction de distribution $F(\mathbf{x})$ sur $\mathbb{R}^p$ et de densité $f(\mathbf{x})$ (par rapport à la mesure de Lebesgue), $\sum$ étant la matrice des covariances de X .

Les deux premiers moments de cet estimateur sont proportionnels aux vrais moments [11].

Les propriétés asymptotiques de cet estimateur sont établies [11] sous certaines conditions de régularité.

$\tilde{f}_n(\mathbf{x})$ est un estimateur consistant pour $f(\mathbf{x})$ en tout point de continuité de f (voir [11]).

Si l'ensemble des n observations est réparti en J groupes, ce que nous supposerons par la suite, on estimera J densités (une par groupe) et donc J paramètres $\alpha_1, \cdots, \alpha_j, \cdots, \alpha_J$.

Pour déterminer les paramètres α_j nous proposons la méthode suivante.

2.1. Détermination des α_j

On note par $\alpha_{1,\cdots,J}$ tout J uple $(\alpha_1, ..., \alpha_J)$. On déterminera le $\alpha_{1,\cdots,J}$ qui minimise le taux estimé de mauvais classement. La règle d'affectation est celle qui attribuera une observation $\mathbf{x}$ à la classe ayant la plus grande probabilité *a posteriori*. Cette probabilité *a posteriori* se calcule par la formule de Bayes :

$$\frac{p_l f_l(\mathbf{x})}{\sum p_l f_l(\mathbf{x})}$$

où $f_l(\mathbf{x})$ densité du groupe l est estimée par :

$$\tilde{f}_{n_l}(\mathbf{x}) = \frac{1}{n_l} \sum_{j=1}^{n_l} \frac{1}{(2\pi)^{p/2} \left| \frac{\Sigma}{n_l^{2\alpha_l}} \right|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - X_j)^T \left(\frac{\Sigma}{n_l^{2\alpha_l}} \right)^{-1} (\mathbf{x} - X_j) \right\}$$

$$\text{avec } l = 1, \cdots, J, \quad \alpha_l \in \left] 0, \frac{1}{2} \right[$$

n_l effectif du groupe G_l

En balayant l'intervalle $\left] 0, \frac{1}{2} \right[$, on estime pour chaque $\alpha_1, \cdots, \alpha_J$ le taux d'erreur $TE(\alpha_{1,\ldots,J})$ par la méthode de validation croisée [3].

Le $\alpha^{op}_{1,\ldots,J}$ optimal est celui qui vérifie :

$$TE(\alpha^{op}_{1,\ldots,J}) = \min_{\alpha_1, \ \ldots, J \in]0,1/2[^J} TE(\alpha_{1,\ldots,J})$$

On obtient en même temps une estimation du taux d'erreur apparent par validation croisée.

3. Sélection des Variables

3.1. Cas J = 2

On sait [1] que la probabilité d'erreur de mauvais classement P_e est tel que

$$P_e = \frac{1}{2}\left\{1 - \int_{\mathcal{X}} |P_1(\mathbf{x}) - P_2(\mathbf{x})| d\mu(\mathbf{x})\right\}$$

où :

$\mathcal{X}$ espace d'observations

μ mesure de Lebesgue ou de comptage

$P_i, i = 1, 2$ probabilité *a posteriori* du groupe i

$$P_i(\mathbf{x}) = \frac{\Pi(i) f_i(\mathbf{x})}{f(\mathbf{x})}$$

$f_i(\mathbf{x})$ densité par rapport à μ du groupe i

$\Pi(i)$ probabilité *a priori*

$$f(\mathbf{x}) = \sum_{i=1}^{2} \Pi(i) f_i(\mathbf{x})$$

Notre idée est de sélectionner les variables qui minimisent le taux d'erreur autrement dit qui maximisent $\int_{\mathcal{X}} |P_1(\mathbf{x}) - P_2(\mathbf{x})| d\mu(\mathbf{x})$ qui est la distance en variation de Kolmogoroff entre les deux groupes.

On montre [8] qu'on ne risque pas de biaiser le taux d'erreur P_e du fait qu'on a transformé l'échantillon par une ACP.

3.2. Cas J > 2

Pour généraliser notre idée nous nous basons sur le résultat suivant :

Lemme [6]

$$P_e = \frac{2}{J} \sum_{j=1}^{J-1} \sum_{i=j+1}^{J} P_e(i,j) + \frac{1}{J} \sum_{j=1}^{J-1} \sum_{i=j+1}^{J} \int_{(D_i^* \cup D_j^*)^c} |P_i(\mathbf{x}) - P_j(\mathbf{x})| d\mu(\mathbf{x})$$

et

$$\frac{2}{J} \sum_{j=1}^{J-1} \sum_{i=j+1}^{J} P_e(i,j) \leqslant P_e \leqslant \sum_{j=1}^{J-1} \sum_{i=j+1}^{J} P_e(i,j)$$

avec

$$P_e(i,j) = \frac{1}{2}\left[\Pi(i) + \Pi(j) - \int_{\mathcal{X}} |P_i(\mathbf{x}) - P_j(\mathbf{x})| d\mu(\mathbf{x})\right]$$
$$D_i^* = \{\mathbf{x} \in \mathcal{X} / P_i(\mathbf{x}) > P_j(\mathbf{x}), \forall j = 1, \cdots, J, j \neq i\}$$

Posons,

$$B_{\text{sup}} = \sum_{j=1}^{J-1} \sum_{i=j+1}^{J} P_e(i,j)$$
$$B_{\text{inf}} = \frac{2}{J} B_{\text{sup}}$$

On a donc

$$B_{\text{inf}} \leqslant P_e \leqslant B_{\text{sup}}$$

On voit que si $J = 2$

$$B_{\text{inf}} = P_e = B_{\text{sup}}$$

On retrouve ainsi le cas de deux groupes.

En généralisant le critère de sélection pour le cas $J = 2$, on retiendra les variables qui minimisent B_{sup}.

Autrement dit, on retiendra celles qui maximisent :

$$\sum_{j=1}^{J-1} \sum_{i=j+1}^{J} \int_{\mathcal{X}} |P_i(\mathbf{x}) - P_j(\mathbf{x})| d\mu(\mathbf{x})$$

qui n'est autre que la distance en variation de Kolmogoroff entre les J groupes.

Pour estimer cette distance, on utilise l'estimateur

$$\sum_{j=1}^{J-1} \sum_{i=j+1}^{J} \tilde{j}(i,j)$$

avec

$$\tilde{j}(i,j) = \frac{1}{n} \sum_{k=1}^{n} |\tilde{p}_i(\mathbf{x_k}) - \tilde{p}_j(\mathbf{x_k})|$$

où $\tilde{p}_i$ est un estimateur de $P_i, i = 1, \cdots, J$

où

$$\tilde{p}_i(\mathbf{x}) = \frac{n_i}{n} \frac{\tilde{f}_{n_i}(\mathbf{x})}{\tilde{f}(\mathbf{x})}$$

avec

$$\tilde{f}(\mathbf{x}) = \sum \frac{n_i}{n} \tilde{f}_{n_i}(\mathbf{x})$$

$\tilde{f}_{n_i}$ étant la densité du groupe i calculée à partir de la formule du 2), et n_i est le nombre des observations du groupe i.

$|\tilde{p}_i(\mathbf{x_k}) - \tilde{p}_j(\mathbf{x_k})|$ est calculé à l'aide de l'échantillon total auquel on a enlevé $\mathbf{x_k}$.

On montre [8] que si les $\tilde{p}_i$ sont des estimateurs consistants pout les P_i alors $\tilde{j}(i,j)$ est un estimateur consistant pour $\int_{\mathcal{X}} |P_i(\mathbf{x}) - P_j(\mathbf{x})| d\mu(\mathbf{x})$ pour tout couple (i,j) et $i \neq j$.

3.3. Méthode de sélection des variables [7]

Nous rappelons ici le théorème établi, pour le cas de deux groupes, dans [7] et l'algorithme qu'on a développé. Pour la généralisation à plus de deux groupes et les démonstrations voir [7].

Notons $D_{1,2}(x_1, \cdots, x_p)$ la distance de Kolmogoroff calculée à l'aide des p prédicteurs $x_1, \cdots, x_p$

$R_{i,j} = \{\mathbf{x} \in \mathcal{X} / \Pi_i f_i(\mathbf{x}) > \Pi_j f_j(\mathbf{x})\}, i, j = 1, 2$

$R_{i,j}(x_l, x_{l+1}, \cdots, x_m)$ est la projection de $R_{i,j}$ sur les coordonnées $x_l, x_{l+1}, \cdots, x_m$

Théorème 1.

Soit $x_1, \cdots, x_p$ les p prédicteurs pris dans cet ordre.

on a :

$$D_{1,2}(x_1, \cdots, x_p) = D_{1,2}(x_1) + \sum_{j=2}^{j=p} D_{1,2}(x_j / x_1, \cdots, x_{j-1})$$

avec :

$$D_{1,2}(x_1) = \int_{R_{1,2}(x_1)} \alpha_{1,2}(x_1) \left[\Pi_1^{\frac{1}{p}} f_1(x_1) - \Pi_2^{\frac{1}{p}} f_2(x_1) \right] dx_1$$
$$+ \int_{R_{2,1}(x_1)} \alpha_{2,1}(x_1) \left[\Pi_2^{\frac{1}{p}} f_2(x_1) - \Pi_1^{\frac{1}{p}} f_1(x_1) \right] dx_1$$

$$\alpha_{1,2}(x_1) =$$

$$\int_{R_{1,2}(x_2,\cdots,x_p)} \left[\frac{\Pi_1^{\frac{p-1}{p}} f_1(x_2,\cdots,x_p/x_1) + \Pi_2^{\frac{p-1}{p}} f_2(x_2,\cdots,x_p/x_1)}{2} \right] dx_2 ... dx_p$$

$$D_{1,2}(x_j/x_1,\cdots,x_{j-1}) = \int_{R_{1,2}(x_1,\cdots,x_j)} \alpha_{1,2}(x_j)$$

$$\left[\Pi_1^{\frac{1}{p}} f_1(x_j/x_1,\cdots,x_{j-1}) - \Pi_2^{\frac{1}{p}} f_2(x_j/x_1,\cdots,x_{j-1}) \right] dx_1 ... dx_j$$

$$+$$

$$+ \int_{R_{2,1}(x_1,\cdots,x_j)} \alpha_{2,1}(x_j)$$

$$\left[\Pi_2^{\frac{1}{p}} f_2(x_j/x_1,\cdots,x_{j-1}) - \Pi_1^{\frac{1}{p}} f_1(x_j/x_1,\cdots,x_{j-1}) \right] dx_1 ... dx_j$$

avec :

$$\alpha_{1,2}(x_j) = \left[\frac{\Pi_1^{\frac{1}{p}} f_1(x_1) + \Pi_2^{\frac{1}{p}} f_2(x_1)}{2} \right] \left[\frac{\Pi_1^{\frac{1}{p}} f_1(x_2/x_1) + \Pi_2^{\frac{1}{p}} f_2(x_2/x_1)}{2} \right] \cdots$$

$$\cdot\cdot \left[\frac{\Pi_1^{\frac{1}{p}} f_1(x_{j-1}/x_1,\cdots,x_{j-2}) + \Pi_2^{\frac{1}{p}} f_2(x_{j-1}/x_1,\cdots,x_{j-2})}{2} \right]$$

$$\int_{R_{1,2}(x_{j+1},\cdots,x_p)}$$

$$\left[\frac{\Pi_1^{\frac{p-j}{p}} f_1(x_{j+1},\cdots,x_p/x_1,\cdots,x_j) + \Pi_2^{\frac{p-j}{p}} f_2(x_{j+1},\cdots,x_p/x_1,\cdots,x_j)}{2} \right]$$

$$dx_{j+1} ... dx_p$$

$D(x_1)$ est la contribution de la variable x_1 à la distance $D_{1,2}(x_1,\cdots,x_p)$.

$D_{1,2}(x_j/x_1,\cdots,x_{j-1})$ est la contribution de la variable x_j à la distance $D_{1,2}(x_1,\cdots,x_p)$ conditionnellement à l'information apportée par les variables $x_1,\cdots,x_{j-1}$. Comme on le remarque ce développement dépend de l'ordre des variables. D'un développement à l'autre, la contribution de certaines variables peut varier. En effet, supposons que l'information apportée par la variable x_4 soit la même que celle apportée par les variables x_1, x_2, x_3. Alors dans l'ordre $x_4, x_1, x_2, x_3, x_5, \cdots, x_p$, la contribution de x_4 est grande. Par contre dans l'ordre $x_1, x_2, x_3, x_4, \cdots, x_p$, la contribution de x_4 est négligeable.

3.4. Algorithme de sélection

Il est divisé en deux phases, chacune étant divisée en p étapes s'il y a p prédicteurs. La première phase a pour but d'ordonner les variables par contribution décroissante. La première est celle qui a la plus grande contribution, la deuxième est celle qui a la plus grande contribution connaissant l'information apportée par la première,etc.

La deuxième phase consiste à choisir parmi ces p ensembles de prédicteurs obtenus en respectant l'ordre de la phase 1, celui qui donne le taux d'erreur minimal. Ce groupe n'est pas nécessairement le meilleur r-uplet si on a obtenu r variables, car il s'agit d'un algorithme ascendant sans remise en cause.

PHASE 1

Etape 1 Soit $X = (x_1, \cdots, x_p)$ le vecteur des p prédicteurs. Pour $i = 1, \cdots, p$ calculer $D_{1,2}(x_i)$. Retenir la variable qui a la plus grande contribution.

Soit $x_{(1)}$ cette variable.

Etape 2 Soit $X = (x_1, \cdots, x_{p-1})$ le vecteur des $p-1$ variables qui restent à l'issue de l'étape 1 (sans la variable (1)). Pour $i = 1, \cdots, p-1$ calculer $D_{1,2}(x_i/x_{(1)})$. Retenir la variable qui a la plus grande contribution

Soit $x_{(2)}$ cette variable.

$\vdots$

Etape $i (i < p)$ Soit $X = (x_1, .., x_{p-i+1})$ le vecteur des $p - (i-1)$ variables non encore retenues. Pour $j = 1, \cdots, p - (i-1)$ calculer $D_{1,2}(x_j/x_{(1)}, \cdots, x_{(i-1)})$. Retenir la variable qui a la plus grande contribution.

Soit $x_{(i)}$ cette variable.

Résultat de la phase 1 : on a ordonné les variables par contribution décroissante

$$x_{(1)}, x_{(2)}, \cdots, x_{(p)}$$

PHASE 2 (boucle de p itérations).

Pour $i = 1 \cdots p$
calculer $P_e(x_{(1)}, \cdots, x_{(i)})$
$Fin - pour$

Le groupe des variables sélectionnées est celui qui vérifie l'équation

$$P_e(x_{(1)}, \cdots, x_{(j)}) = \min_{i=1,\ldots,p} P_e(x_{(1)}, \cdots, x_{(i)})$$

S'il y a plus d'un groupe qui vérifie cette équation, on choisit le moins redondant.

Estimateur utilisé

Comme nous utilisons des échantillons de taille finie, nous avons estimé les termes $D_{1,2}(x_k/x_1,\cdots,x_{k-1})$ par les estimateurs obtenus en remplaçant dans les formules du §3.3 définissant $D_{1,2}(x_k/x_1,\cdots,x_{k-1})$ les densités qui interviennent par leurs estimateurs déduits des formules du §2.

C'est ainsi que $f_i(x_k/x_1,\cdots,x_{k-1})$ densité conditionnelle de l'observation $\mathbf{x}$ sur les coordonnées $x_1,\cdots,x_k$ dans le groupe i est estimée par :

$$\tilde{f}_{n_i}(x_k/x_1,\cdots,x_{k-1}) = \begin{cases} \dfrac{\tilde{f}_{n_i}(x_1,\cdots,x_{k-1})}{\tilde{f}_{n_i}(x_1,\cdots,x_k)} & \text{si } \tilde{f}_{n_i}(x_1,\cdots,x_k) > 0, \\ 0 & \text{sinon} \end{cases}$$

$\tilde{f}_{n_i}(x_1,\cdots,x_{k-1})$ et $\tilde{f}_{n_i}(x_1,\cdots,x_k)$ se calculant à partir de la formule donnée au §2.1.

Par ailleurs, les probabilités Π_i ont été estimées par $\dfrac{n_i}{n}$, n_i étant l'effectif du groupe i et n le nombre total d'observations.

On sait [4] que cet estimateur est convergent.

Exemple d'illustration

Nous avons testé notre méthode sur l'échantillon des Infarctus [9]. Il est réparti en deux groupes, l'un est de 51 observations l'autre est de 50. Sept variables quantitatives décrivent cet échantillon Les deux groupes sont assez biens séparés, la distance de Mahalanobis est égale à 4.942

$\alpha_{1,2}^{op} = (0.38, 0.38)$

Sélection des variables

PHASE 1

Etape 1

variable	x_1	x_2	x_3	x_4	x_5	x_6	x_7
$\widetilde{D}(x_i)$	0.66	0.45	0.16	0.23	0.22	0.48	0.50

$$x_{(1)} = x_1$$

Etape 2

variable	x_2	x_3	x_4	x_5	x_6	x_7
$\widetilde{D}(x_i/x_{(1)})$	0.25	0.45	0.55	0.43	0.48	0.45

$$x_{(2)} = x_4$$

Etape 3

variable	x_2	x_3	x_5	x_6	x_7
$\widetilde{D}(x_i/x_{(1)}, x_{(2)})$	0.2	0.48	0.45	0.45	0.43

$$x_{(3)} = x_3$$

Etape 4

variable	x_2	x_5	x_6	x_7
$\widetilde{D}(x_i/x_{(1)}, x_{(2)}, x_{(3)})$	0.2	0.42	0.48	0.49

$$x_{(4)} = x_7$$

Etape 5

variable	x_2	x_5	x_6
$\widetilde{D}(x_i/x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)})$	0.53	0.41	0.46

$$x_{(5)} = x_2$$

Etape 6

variable	x_5	x_6
$\widetilde{D}(x_i/x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)})$	0.4	0.5

$$x_{(6)} = x_6$$

Etape 7

$$x_{(7)} = x_5$$

Résultat de la phase 1 :

$$x_1, x_4, x_3, x_7, x_2, x_6, x_5$$

PHASE 2

On estime le taux d'erreur par la méthode de validation croisée.

$$P_e(x_1) = 0.22$$
$$P_e(x_1, x_4) = 0.22$$
$$P_e(x_1, x_4, x_3) = 0.20$$
$$P_e(x_1, x_4, x_3, x_7) = 0.20$$
$$P_e(x_1, x_4, x_3, x_7, x_2) = 0.20$$
$$P_e(x_1, x_4, x_3, x_7, x_2, x_6) = 0.29$$
$$P_e(x_1, x_4, x_3, x_7, x_2, x_6, x_5) = 0.31$$

Les variables optimales sont x_1, x_3, x_4

$$P_e(x_1, x_3, x_4) = 0.20$$

3.5. Conclusion

Nous résumons les démarches à suivre suivant la nature de l'échantillon dans le tableau ci-dessous.

Méthode non-paramétrique d'analyse discriminante à J groupes

variables explicatives qualitatives	variables explicatives mixtes	variables explicatives quantitatives
↓	↓	↓
métrique $\mathcal{X}^2$	métrique mixte	métrique inverse des variances
Pour chaque groupe Analyse en Composantes Principales Généralisée		
↓		
Pour chaque groupe calcul des densités pour différents $\alpha \in]0, 1/2[$ et choix de celui (α) qui minimise le taux d'erreur (méthode de validation croisée)		
↓		
Pour chaque groupe sélection des variables		

4. Applications

Nous avons testé notre méthode sur trois échantillons, deux sont quantitatifs : les Iris de Fisher et les Conducteurs; le troisième est qualitatif : les nombres digitaux.

L'échantillon des conducteurs est réparti en trois groupes[10]. Il a été obtenu par une enquête auprès des automobilistes. On a recueilli des infomations concernant les trajets parcourus, la consommation et le style de conduite des conducteurs. On cherche à prédire la classe de consommation à l'aide de vingt cinq variables quantitatives.

L'échantillon des nombres digitaux est réparti en cinq classes qui représentent respectivement les chiffres ,un, deux, trois, quatre cinq.Chaque observation est décrite par sept variables dichotomiques [2].

Pour chaque application on divise l'échantillon initial $\mathcal{L}$ en sous-ensembles disjoints $\mathcal{L}_1, \mathcal{L}_2$ tel que $\mathcal{L}_1 \cup \mathcal{L}_2 = \mathcal{L}$.

$\mathcal{L}_1$ servira à déterminer le $\alpha^{op}_{1,\ldots,J}$, à sélectionner les meilleurs prédicteurs et à estimer le taux d'erreur par validation croisée . $\mathcal{L}_2$ servira d'échantillon test.

On note par $TE^{ts}(\alpha^{op}_{1,\ldots,J})$ l'estimation du taux d'erreur obtenue par l'échantillon test $\mathcal{L}_2$ et par $TE^{cv}(\alpha^{op}_{1,\ldots,J})$ l'estimation du taux d'erreur par validation croisée (Mickey-Lachenbruch) obtenue par l'échantillon $\mathcal{L}_1$.

L'échantillon des Iris

L'échantillon est décrit par quatre variables quantitatives.

$n = 150. J = 3$

$\mathcal{L}_1$ est de 120 observations et $\mathcal{L}_2$ 30.

$\alpha^{op}_{1,2,3} = (0.02, 0.02, 0.026)$

$T^{cv}(\alpha^{op}_{1,2,3}) = 0.12;$

$TE^{ts}(\alpha^{op}_{1,2,3}) = 0$

Sélection des variables

Une seule variable est sélectionnée : x_1

$TE^{cv}(\alpha^{op}_{1,2,3}, x_1) = 0.16$

$TE^{ts}(\alpha^{op}_{1,2,3}, x_1) = 0.16$

Échantillon digital

Il est décrit par sept variables qualitatives

$n_1 = 72, n_2 = 15, J = 5$

On a éliminé les variables x_5, x_6, x_7 car elles sont de variance nulle.

$\alpha^{op}_{1,2} = (0.04, 0.04, 0.04, 0.04, 0.004)$

$TE^{cv}(\alpha^{op}_{1,2,3,4,5}) = 0.2$

$TE^{ts}(\alpha^{op}_{1,2,3,4,5}) = 0.2$

Sélection des variables

Deux variables sont sélectionnées : x_1, x_2
$TE^{ts}(\alpha^{op}_{1,2,3,4,5}, x_1, x_2) = 0.2$
$TE^{cv}(\alpha^{op}_{1,2,3,4,5}, x_1, x_2) = 0.2$

Échantillon des conducteurs

Il est décrit par 25 variables quantitatives. Les groupes sont mal séparés.
$n_1 = 150, J = 3$
$\mathcal{L}_1$ est de 132 observations , $n_2 = 18$
$\alpha^{op}_{1,2,3} = (0.2, 0.2, 0.2)$
$TE^{cv}(\alpha^{op}_{1,2,3}) = 0.59$
$TE^{ts}(\alpha^{op}_{1,2,3}) = 0.7$

Sélection des variables

Une seule variable est sélectionnée x_{18}
$TE^{cv}(\alpha^{op}_{1,2,3}, x_{18}) = 0.32$
$TE^{ts}(\alpha^{op}_{1,2,3}, x_{18}) = 0.37$
On voit ici une nette amélioration des taux d'erreur

5. Comparaison avec deux méthodes classiques à l'aide du taux d'erreur estimé par validation croisée

Nous avons comparé notre méthode,à l'aide du taux d'erreur TE estimé par validation croisée, à deux méthodes classiques : L'analyse discriminante classique et la méthode du plus proche voisin (implémentées dans SAS).

Échantillon des Iris

Échantillon	TE
Discrimination linéaire	0.03
Plus proche voisin k=1	0.05
Discrimination non-paramétrique	0.12

Échantillon de l'infarctus

Échantillon	TE
Discrimination linéaire	0.15
Plus proche voisin k=1	0.28
Discrimination non-paramétrique	0.20

Échantillon digital

Échantillon	TE
Discrimination linéaire	0.14
Plus proche voisin k=1	0.16
Discrimination non-paramétrique	0.20

Échantillon des conducteurs

Échantillon	TE
Discrimination linéaire	0.43
Plus proche voisin k=1	0.46
Discrimination non-paramétrique	0.32

6. Conclusion

Notre méthode est facile à mettre en œuvre. Elle a donné des résultats satisfaisants pour ces quatre échantillons.

Elle est applicable à tout problème sans aucune condition préalable.

Du fait qu'elle utilise des composantes principales, au lieu des variables observées l'interprétation peut être délicate.

Pour un échantillon de J groupes, elle nécessite $2 \times J$ analyses factorielles.

BIBLIOGRAPHIE

[1] ANDERSON T.W. (1984). «An Introduction to Multivariate Statistical Analysis». Second Edition, Wiley series in Probability and Matimatical Statistics.

[2] BREIMAN & FRIEDMAN & OLSHEN & STONE (1984), «Classification and Regression Trees». Wadsworth International Group, pp 44,47.

[3] LACHENBRUCH P.A & MICKEY M.R. (1968). «Estimation of Error Rates in Discriminant Analysis». Technometrics 10, pp 1,10.

[4] GAUTIER J.M & SAPORTA G. (1984). «Méthodes non paramétriques en analyse discriminante». Data analysis and informatics, E Diday ed., North Holland, III pp 591, 605.

[5] GLICK N. (1972). «Sample based classification procedures derived from density estimators». Journal of the American statistical association, March pp 116, 121.

[6] LISSACK Y-KING-SUN FU (1976). «Error estimation in pattern recognition via distance between posterior density functions». IEEE transactions of information theory, January pp 34, 44.

[7] MAHJOUB S & SAPORTA G. «A new method of selection of variables in discriminant analysis». Soumis à Applied Stochastic Modells and Data Analysis.

[8] MAHJOUB S. (1992).«Proposition d'une nouvelle méthode de discrimination non paramétrique». Thèse de Doctorat de l'Université de PARIS 9 Dauphine.

[9] NAKACHE J.P., LORENTE P., BENZECRI J.P., CHASTANG C. (1977). «Aspects pronostiques et thérapeutiques de l'Infarctus Myocardique compliqué d'une défaillance sévère de la pompe Cardiaque. Cahiers de l'Analyse des données, Dunod, vol 2, n 4, pp 415, 434.

[10] SAPORTA G. (1990). «Probabilités, Analyse des données et Statistique». Edition Technip.

[11] SHANMUGAM K.S. (1977). «On modified form of Parzen estimator». Pattern recognition vol 9 pp 167, 170.