REVUE DE STATISTIQUE APPLIQUÉE

A. MERBOUHA
A. MKHADRI

# Méthodes de scoring non-paramétriques

*Revue de statistique appliquée*, tome 54, n° 1 (2006), p. 5-26
<http://www.numdam.org/item?id=RSA_2006__54_1_5_0>

*Rev. Statistique Appliquée,* 2006, **LIV** (1), 5-26

# MÉTHODES DE SCORING NON-PARAMÉTRIQUES

A. MERBOUHA[(1)], A. MKHADRI[(2)*]

[(1)] *Département de Mathématiques, FST-Beni-Mellal, Maroc*
[(2)] *Département de Mathématiques, Université Cadi-Ayyad, Marrakech, Maroc*

## RÉSUMÉ

Dans cette note, nous présentons et illustrons les performances de nouvelles méthodes, peu utilisées en scoring ou peu connues, sur un exemple intéressant de données bancaires avec variables mixtes où l'objectif est de prédire le risque de crédit. En plus de la grande dimension de l'espace des observations, les deux groupes (bon et mauvais payeurs) sont très déséquilibrés : les mauvais payeurs représentent moins de 10 %. Dans le cas des coûts de mauvais classement égaux, nous montrons que l'utilisation des méthodes non-paramétriques, fondées sur le modèle de location, peuvent fournir d'excellents résultats dans ce cadre. Elles ont l'avantage de fournir des taux globaux de mauvais classement faibles et des taux conditionnels très équilibrés. Tandis que les autres méthodes fondées sur les $k$ plus proches voisins et la distance de Cuadras ont tendance à fournir des taux conditionnels de mauvais classement très déséquilibrés : elles classent parfaitement les « bons payeurs » et éprouvent d'énormes difficultés à bien classer les « mauvais payeurs ».

***Mots-clés :*** *Discrimination, modèle de location non-paramétrique, $k$-plus proches voisins*

## ABSTRACT

In this paper, we present and illustrate the performance of new nonparametric classification methods, some of them are less known or less used in practice, on an interesting unbalanced bank credit data set. The observations of the latter are described by a relatively large number of mixture of discrete and continuous variables, and where the minority group represents less than $10\%$. It is shown that the nonparametric smoothing approaches of the location model, with adapted weight function, improve the prediction accuracy of the minority group, and have favorable performance compared to the nonparametric methods based on $k$ nearest neighbors model and the distance based discrimination methods.

***Keywords :*** *Credit scoring, nonparametric location model, $k$-nearest neighbors*

* Le second auteur a bénéficié d'une bourse de recherche TWAS 01-159 RG/MATHS/AF/AC.

## 1. Introduction

La motivation principale de cette note est due à la lecture des travaux de scoring de l'équipe du Professeur Härdle (Université de Berlin) sur un exemple intéressant de données bancaires avec variables mixtes. L'objectif étant de prédire le risque de crédit : prédire si un emprunteur sera un bon ou un mauvais payeur et prendre ensuite la décision appropriée. En plus de la grande dimension de l'espace des observations, les deux groupes (bon et mauvais payeurs) sont très déséquilibrés : les mauvais payeurs représentent moins de 10 % qui est considéré comme étant le seuil raisonnable par les spécialistes du crédit.

Les méthodes de scoring (*cf.* Bardos 2001), qui font partie des méthodes d'analyse discriminante (*cf.* McLachlan 1992, Celeux et Nakache 1994, Ripley 1996 et Celeux 2003), ont suscité récemment un intérêt considérable dans le domaine de la finance, l'assurance et le marketing. Ainsi, au moins deux éminents statisticiens (Pr. Härdle de l'université de Berlin et Pr. Hand de London College) ont constitué deux groupes de travail consacrés essentiellement à ce problème de scoring. L'engouement pour ces méthodes, ou ceux de « Data Mining », est dû à la demande urgente des banques, des compagnies d'assurance et des organismes de crédit pour réduire le risque financier, qui n'est pas négligeable. En effet, avec le développement informatique croissant, ces organismes ont constitué de grandes bases ou entrepôts de données qui sont très compliqués à analyser par les moyens classiques. Cela a entraîné l'émergence des méthodes de Data Mining adaptées à ce genre de données et qui connaissent un développement commercial très important (*cf.* Besse *et al.* 2001 pour un très bon exposé à ce sujet). La motivation principale de ces méthodes est la valorisation d'une grande base de données par la recherche d'informations utiles pour l'aide à la décision.

Les méthodes de scoring classiquement utilisées sur données bancaires sont en général de type linéaire (Analyse Discriminante Linéaire (ADL) ou Régression Logistique (RL)) du fait de leur simplicité et leur grande robustesse. Pour tenir compte de la non-linéarité de certaines variables, l'analyse discriminante quadratique (ADQ) ou régression polynomiale est parfois considérée. Toutes ces méthodes sont fondées sur le calcul du score qui est en général une combinaison linéaire des variables explicatives (ou de leurs transformées). Typiquement, le score résume les variables explicatives en une forme prédéfinie (linéaire ou quadratique). Mais ADL et ADQ ne sont pas adaptées aux données avec variables explicatives mixtes où les frontières de séparation des deux groupes ne sont nécessairement ni linéaires ni quadratiques (*cf.* Besse *et al.* 2001 et Müller et Härdle 2002). D'autres méthodes non-linéaires et non-paramétriques, comme les réseaux neuronaux et les arbres de décision, sont préconisées de plus en plus (*cf.* Armingler, Enache et Bonne 1997, Hand et Henley 1997, Henley et Hand 1996 et Hand 2001). En effet, pour l'exemple de données bancaires décrit en Section 2, Komroád (2003) a effectué une étude détaillée de la comparaison de la régression logistique, du perceptron multi-couche et des réseaux à fonctions radiales de base. Les résultats obtenus par les trois méthodes sont similaires : mais le groupe des mauvais payeurs est fortement mal classé par les trois méthodes. Par ailleurs, Müller et Härdle (2002) ont considéré une modification du modèle logistique, qui tient compte de la non-linéarité de certaines variables continues, qui a fourni un

résultat relativement meilleur que la régression logistique linéaire sur une base de données relativement similaire.

Notre objectif dans cette note est de présenter et d'illustrer la performance de nouvelles méthodes, dont certaines sont peu utilisées en scoring ou peu connues, adaptées à la structure de données en question. La première famille est fondée sur le modèle de location (*cf.* Krzanowski 1975) qui suppose que la partie continue du vecteur d'observations conditionnellement à la cellule de la partie discrète suit une loi normale de moyenne dépendante de la cellule et du groupe, et de matrice variance indépendante des deux. Nous montrons que la version non-paramétrique de Asparoukhov et Krzanowski (2000) (avec une modification de la méthode d'estimation de certains paramètres de lissage) fournit un bon résultat. La seconde famille est composée des méthodes non-paramétriques des $k$-plus proches voisins probabilistes et non probabilistes où, à la place de la distance euclidienne classique, nous utilisons une distance adaptée aux données mixtes. La troisième famille est composée de la discrimination barycentrique et de celle due à Cuadras (1989) qui sont simples à mettre en œuvre, mais peu connues et dont la règle d'affectation qu'elles induisent ne nécessite que le calcul de distances entre observations.

Le plan de l'article est le suivant. Nous décrivons en Section 2 les données bancaires utilisées dans toute la suite en montrant leurs caractéristiques de base. En Section 3, nous présentons les trois familles de méthodes de scoring non-paramétriques. Nous commençons par la méthode basée sur le modèle de location adapté aux données avec variables mixtes, et nous détaillons sa version non-paramétrique en précisant notre méthode de choix des paramètres de lissage adaptée à cette structure de données. Ensuite, nous décrivons les méthodes des $k$-plus proches voisins non-probabilistes et probabilistes avec notre choix de distance adaptée aux données traitées. De plus, nous présentons deux méthodes fondées sur la transformation « optimale » de variables qualitatives en variables quantitatives et l'application des $k$ plus proches voisins sur le tableau transformé. Enfin, nous présentons la troisième famille dont la règle d'affectation est simple et fondée juste sur le calcul d'une certaine distance entre observations. Les résultats de ces différentes méthodes sont analysés en Section 4. Finalement, en Section 5 nous résumons les résultats de nos comparaisons.

## 2. Structure des données

Les organismes de crédit utilisent l'analyse discriminante pour prédire si un emprunteur sera un bon ou un mauvais payeur et prendre ensuite la décision adéquate. Pour cela, ils disposent d'une grosse base de données des anciens clients qui ont contracté un certain crédit (immobilier, achat de voiture, ou autre ...) et dont on connaît la qualité payeur résumée par une variable qualitative $Y$ à deux modalités : bon ou mauvais payeur. Les données du dossier de prêt de chaque client sont décrites par les $p$ variables explicatives $(X_1, \ldots, X_p)$ qui sont en général de nature mixte : qualitatives et continues. Dans la suite, nous décrivons brièvement l'ensemble des données de crédit utilisé pour la comparaison de certaines méthodes de scoring.

Les données analysées dans cet article ont été recueillies à partir du site web de Karel Komorád (Humboldt-Universität Berlin). Certains auteurs ont considéré ces

données dans leur illustration des méthodes de scoring (Müller & Rönz 1999, et Müller & Härdle 2002). Ce fichier de données est issu de la Compagnie Bancaire (France), mais la source est confidentielle et les noms de toutes les variables ont été supprimées. Après un premier tri, Komorád (2003) a conservé un échantillon de $6\,178$ clients anonymes décrits par $24$ variables : la variable à classer $Y$ et $23$ variables explicatives $(X_1, \ldots, X_{23})$ dont 8 sont numériques et les autres qualitatives. La variable $Y$ est binaire et représente les « mauvais emprunteur » (codé $1$) et « bon emprunteur » (codé $0$). Le nombre de mauvais clients est relativement faible ($6\,\%$), ce qui est typique pour les données de crédit. Le tableau suivant donne la fréquence des deux modalités de $Y$ sur les deux échantillons d'apprentissage et de test.

TABLEAU 1
*Fréquences des deux modalités de la variable à discriminer Y*

| Y | Apprentissage | Test |
|---|---|---|
| 0 | 3 888 (94 %) | 1 918 (93.9 %) |
| 1 | 247 (6 %) | 125 (6.1 %) |
| Total | 4 135 | 2 043 |

L'étude unidimensionnelle, effectuée par Komorád (2003) et Müller & Härdle (2002), montre que les variables continues $(X_1, \ldots, X_8)$ présentent des distributions très dissymétriques. De plus, les variables $X_5, X_7$ et $X_8$ sont de structure quasi-discrète. Ainsi, Müller & Härdle (2002) se sont concentrés sur les variables $X_1$ à $X_4$ et $X_6$, qui présentent une variation continue, pour leur inclusion de manière non-paramétrique dans le modèle de régression logistique linéaire. De plus, une étude bidimensionnelle sur $X_1$ à $X_3$, effectuée par Müller & Härdle (2002), montre que les hypothèses de ADL ou ADQ (contours circulaires ou elliptiques) sont difficilement justifiables.

Komorád (2003) a divisé, d'une manière aléatoire, l'échantillon global en deux sous ensembles : apprentissage et test. L'apprentissage représente 2/3 de l'échantillon global ($4\,135$ observations), il est utilisé pour construire les règles de décision. Le fichier test contient le 1/3 restant (2043 observations), et il est utilisé pour valider les règles de décision construites sur l'échantillon d'apprentissage. Les deux fichiers précédents, utilisés dans notre étude, sont archivés dans son site web, respectivement sous les noms *data-train.dat* et *data-test.dat*. Le tableau $1$ ci-dessus résume la répartition des observations dans les deux groupes pour les deux échantillons. Le taux de « mauvais payeur » (groupe $2 : Y = 1$) représente $6\,\%$ et est considéré comme normal par les spécialistes de crédit.

## 3. Méthodes de scoring non-paramétriques

Cette section décrit les différentes méthodes, appliquées aux fichiers de données précédents, que nous avons considérées pour comparaison.

### *3.1. Modèle de location non-paramétrique*

Soient $\mathbf{x}$ un vecteur de variables binaires de dimension $r$ et $\mathbf{z}$ un vecteur de variables numériques de dimension $q$. Ici, notre vecteur d'observation est décrit par $p = r + q$ variables mixtes. Les $r$ variables binaires peuvent être exprimées comme un vecteur multinomial $\mathbf{w}^t = (w_1, ..., w_C)$ où $C = 2^r$. Ainsi, chaque vecteur $\mathbf{x}$ définit de manière unique une cellule : $\mathbf{x}^t = (x_1, ..., x_r)$ a pour cellule $m = 1 + \sum_{i=1}^{r} x_i 2^{i-1}$. On suppose qu'on a deux groupes $G_1$ (bon emprunteur : $Y = 0$) et $G_2$ (mauvais emprunteur : $Y = 1$). Le modèle de location (ML) suppose que, conditionnellement au groupe $G_i$ et au fait que le vecteur $\mathbf{x}$ tombe dans la cellule $m$, le vecteur $\mathbf{z}$ suit une loi normale multivariée de moyenne $\mu_i^{(m)}$ et de matrice variance $\Sigma, i = 1, 2; m = 1, ..., C$, i.e.

$$(\mathbf{z} \mid G_i, w_m = 1, w_j = 0, j \neq m) \sim \mathcal{N}_q(\mu_i^{(m)}, \Sigma). \tag{1}$$

Nous désignons par $p_{im}$ la probabilité qu'une observation de la cellule $m$ appartienne au groupe $G_i$ (*cf.* Krzanowski 1975). Si nous supposons que tous les paramètres sont connus, la règle de décision optimale permet d'affecter une observation $(\mathbf{x}^t, \mathbf{z}^t)^t$ au groupe $G_1$ si

$$\xi_m = (\mu_1^{(m)} - \mu_2^{(m)})^t \Sigma^{-1} \{\mathbf{z} - \frac{1}{2}(\mu_1^{(m)} + \mu_2^{(m)})\} - \log(\frac{p_{2m}}{p_{1m}}) \geqslant 0$$

et au groupe $G_2$ sinon, si $\mathbf{x}$ a pour cellule $m$ $(m = 1, \ldots, C)$. Mais, les paramètres sont en général inconnus et doivent être estimés sur l'échantillon d'apprentissage. Ainsi, les estimateurs du maximum de vraisemblance, sans biais, des paramètres sont : $\hat{p}_{im} = n_{im}/n_i, \quad \widehat{\mu}_i^{(m)} = \bar{\mathbf{z}}_i^{(m)}$ où $\bar{\mathbf{z}}_i^{(m)} = (1/n_{im}) \sum_{j=1}^{n_{im}} \mathbf{z}_{ji}^{(m)}$,

$$\widehat{\Sigma} = \mathbf{S} = \frac{1}{\sum_{m=1}^{C} (n_{1m} + n_{2m} - 2)} \sum_{m=1}^{C} \sum_{i=1}^{2} \sum_{j=1}^{n_{im}} (\mathbf{z}_{ji}^{(m)} - \bar{\mathbf{z}}_i^{(m)})(\mathbf{z}_{ji}^{(m)} - \bar{\mathbf{z}}_i^{(m)})^t, \tag{2}$$

où $n_{\mathrm{im}}$ désigne le nombre d'observations du groupe $i$ tombant dans la cellule $m$, $n_i = \#G_i$ et $\mathbf{z}_{ji}^{(m)}$ est le vecteur $\mathbf{z}$ observé pour le $j^{\mathrm{ème}}$ individu du groupe $G_i$ de la cellule $m$.

Deux inconvénients majeurs limitent l'utilisation pratique de cette règle de décision :

- l'estimation du vecteur moyen d'une cellule donnée est fondée seulement sur les variables discrètes et ignore complètement l'information que peuvent apporter les variables continues.
- en présence d'un grand nombre de variables discrètes, l'estimation de probabilités discrètes peuvent mener à des règles de mauvaise performance, car de trop nombreux paramètres sont à estimer.

Une alternative a été proposée par Asparoukhov et Krzanowski (2000) pour éviter ces deux problèmes. Elle consiste d'abord à ajuster la moyenne du vecteur $\mathbf{z}$

dans une cellule $m$ pour le groupe $G_i$ par une moyenne pondérée des $\bar{z}_i^{(\ell)}$ effectuées sur toutes cellules $\ell$ $(1 \leqslant \ell \leqslant C)$. L'estimation de la moyenne de la $j$-ème variable continue de la cellule $m$ du groupe $G_i$ est alors donnée par

$$\tilde{\mu}_{ij}^{(m)} = \{\sum_{\ell=1}^{C} n_{i\ell}\omega_{ij}(m,\ell)\}^{-1} \sum_{\ell=1}^{C} \omega_{ij}(m,\ell) n_{i\ell} \bar{z}_{\mathrm{ij}}^{(\ell)}, \tag{3}$$

où $0 \leqslant \omega_{ij}(m,l) \leqslant 1$ et $\sum_{\ell=1}^{C} n_{i\ell}\omega_{ij}(m,\ell) > 0$ pour $m,l = 1,...,C$; $i = 1,2$; $j = 1,...,q$ et $\bar{z}_{\mathrm{ij}}^{(\ell)}$ désigne la $j^{\mathrm{ème}}$ composante $(1 \leqslant j \leqslant q)$ du vecteur $\bar{\mathbf{z}}_i^{(\ell)}$. Un des choix intéressants de la famille de poids proposée par ces derniers auteurs est de type exponentiel : $\omega_{ij}(m,\ell) = \lambda_{ij}^{d(m,\ell)}, 0 < \lambda_{ij} < 1, \quad i = 1,2; j = 1,\ldots,q; m,\ell = 1,\ldots,C$, où le coefficient de dissimilarité $d(m,\ell) = (\mathbf{x}_m - \mathbf{x}_\ell)^t(\mathbf{x}_m - \mathbf{x}_\ell)$ est une valeur entière, représentant le nombre de désaccords entre les deux cellules $m$ et $\ell$. Ainsi, ce poids est une fonction exponentielle du coefficient de dissimilarité qui décroît dès que le coefficient croît. Afin de réduire le nombre de paramètres de lissage, Asparoukhov et Krzanowski (2000) ont proposé différentes restrictions selon les deux indices $(i,j)$. Leur étude comparative avec d'autres méthodes de discrimination (CART, réseaux de neurones, logistique etc.) conforte ce choix de poids de type exponentiel.

Maintenant, il reste à évaluer l'ensemble $\Lambda$ de tous les paramètres de lissage en maximisant la vraisemblance par validation croisée (ou pseudo-vraisemblace) des observations $(\mathbf{z_1},\ldots,\mathbf{z_n})$ correspondantes aux variables continues (*cf.* Asparoukhov et Krzanowski 2000). Mais le temps de calcul de maximisation de cette pseudo-vraisemblance peut être prohibitif dès que le nombre de variables binaires dépasse $4$ ou $5$, et parfois l'algorithme (fonction *constr* de Matlab) ne converge pas vers une bonne solution (comme dans notre exemple avec $15$ variables qualitatives). Pour réduire le temps de calcul , nous avons fixé une grille de valeurs dans $[0,1]$ pour le paramètre $\lambda$ (i.e $\lambda = 0, 1/10, 2/10, \ldots, 9/10, 1$) et nous avons considéré des poids qui sont indépendants des indices du groupe et des variables continues ($i$ et $j$) : $\omega_{ij}(m,\ell) = \lambda^{d(m,\ell)}, m,\ell = 1,\ldots,C$. Puis nous estimons la valeur de $\lambda$ fournissant le meilleur taux d'erreur sur le fichier apprentissage. Cette démarche fournit un très bon résultat. Néanmoins, la bonne stratégie est de considérer l'estimation du taux d'erreur par validation croisée, mais elle est prohibitive en temps de calcul dans notre exemple.

Par ailleurs, pour éviter le second problème d'instabilité des estimations des probabilités discrètes, Asparoukhov et Krzanowski (2000) ont considéré l'estimateur non-paramétrique de Hall (1981), fondé sur les plus proches voisins pondérés, défini par

$$\tilde{p}_{im} = n_i^{-1} \sum_{j=0}^{r} \omega_{ij} N_{im}^{(j)}, \quad m = 1,...,C; i = 1,2, \tag{4}$$

où $N_{im}^{(j)}$ est le nombre d'individus $\mathbf{x}$ du groupe $G_i$ tels que $d(\mathbf{x}_m,\mathbf{x}) = j$. Le vecteur des poids $\omega_i = (\omega_{i1},\ldots\omega_{ir})^t$ est choisi de sorte à minimiser la moyenne des erreurs standards au carré $\sum_{\ell=1}^{C} \mathbb{E}(\tilde{p}_{i\ell} - p_{i\ell})^2$. Cette minimisation fournit une solution

optimale explicite du vecteur des poids $\mathbf{w}_i, i = 1, 2$ (*cf.* Hall 1981). Cet estimateur optimal est très flexible, mais il a parfois une tendance à un sur-ajustement des données (*cf.* Hall 1981, Mkhadri 1991, Aparoukhov et Danchev 1997). Néanmoins, le résultat performant de ce modèle sur notre exemple de données bancaires confirme l'utilité de cette estimation non-paramétrique.

Une alternative à l'estimation non-paramétrique des probabilités discrètes est de considérer le modèle d'indépendance conditionnelle (MIC). Ce modèle, qui vise à réduire le nombre de paramètres à estimer (juste $r$ paramètres), suppose que les $r$ variables binaires sont indépendantes à l'intérieur de chaque groupe. Les estimations $\widehat{p}_i(\mathbf{x})$ des probabilités discrètes par groupe sont données par, pour $i = 1, 2$

$$\hat{p}_i(\mathbf{x}) = \prod_{j=1}^{r} \frac{\#\{\mathbf{x}_\ell \in G_i | x_\ell^j = x^j\}}{n_i}, \tag{5}$$

où $x^j$ représente la $j^{\text{ème}}$ coordonnée du vecteur $\mathbf{x}$ et $x_\ell^j$ celle de $\mathbf{x}_\ell$. L'implémentation de MIC nécessite une correction du numérateur afin d'éviter les problèmes de la nullité des estimations des probabilités discrètes (*cf.* Celeux et Nakache 1994, p. 27).

Le grand intérêt de MIC est de proposer un nombre réduit de paramètres à estimer pour chaque groupe : $r$ paramètres au lieu de $2^r - 1$ pour le modèle multinomial complet dans le cas de variables binaires. Par ailleurs, par des transformations algébriques simples (*cf.* Celeux et Nakache 1994, p. 27), on peut montrer que la règle de décision de MIC est fonction linéaire des composantes de $\mathbf{x}$. Ainsi, la règle de décision associée au modèle de location est une fonction linéaire de $\mathbf{z}$ et des composantes de $\mathbf{x}$. Ce genre de règle de décision linéaire simple est beaucoup apprécié dans le domaine du scoring pour sa facilité d'interprétation de la contribution de chaque variable à la fonction score.

### *3.2. Méthodes fondées sur les $k$-plus proches voisins*

Nous nous intéressons maintenant aux méthodes non-paramétriques des $k$-plus proches voisins ($k$-ppv). On commence d'abord par la méthode classique non-probabiliste, et nous présentons brièvement la méthode probabiliste proposée récemment par Holmes & Adams (2002). Puis nous décrivons la discrimination barycentrique et l'approche de Buttrey (1998) fondées sur la transformation « optimale » de variables qualitatives en variables quantitatives et l'application respectivement de la méthode $k$-ppv et la distance du khi-deux sur le tableau transformé.

#### *3.2.1. Méthode des $k$-plus proches voisins*

C'est une méthode très ancienne (*cf.* Fix & Hodges 1951) et très répandue dans la communauté de l'intelligence artificielle. En effet, Holmes & Adams (2002) décomptent plus de 900 articles publiés au sujet de cette méthode. Elle est non-paramétrique et ne présuppose aucune forme pour la densité par groupe.

Dans sa version de base, pour chaque vecteur $\mathbf{x}$ à classer, la procédure $k$ppv examine ses $k$ plus proches voisins dans l'échantillon d'apprentissage et l'affecte à

la classe majoritaire. Le terme «proche» est déterminé selon une distance $\rho$ qui est souvent choisie de type euclidienne. Formellement, si l'échantillon d'apprentissage a été obtenu selon un schéma d'échantillonnage rétrospectif, la densité a posteriori du groupe $G_i$ sachant le vecteur d'observation $\mathbf{x}$ peut être approximée par

$$\frac{v_i(\mathbf{x})}{v(\mathbf{x})} \frac{n}{n_i}\pi_i \tag{6}$$

où $v(\mathbf{x})$ (resp. $v_i(\mathbf{x})$) est le nombre de points de l'échantillon d'apprentissage (resp. du groupe $G_i$) tombant dans le petit voisinage de $\mathbf{x}$ et $\pi_i$ est la probabilité *a priori* du $i^{\text{eme}}$ groupe $(i = 1, 2)$ (*cf.* Celeux 2003).

Les choix du nombre $k$ et de la distance $\rho$ sont bien sûr primordiaux. La sélection de $k$ par la méthode de minimisation du taux d'erreur, estimé par validation croisée, est la plus populaire (Ripley 1996). Par contre le choix souvent préconisé pour la métrique $\rho$, pour décider de la distance entre les points, est la métrique euclidienne usuelle. Mais, si les données sont composées de variables mixtes (quantitatives et discrètes), il est préférable dans ce cas de considérer une métrique adaptée à ce genre de données. Comme notre exemple est composé de données mixtes, au lieu d'utiliser la métrique euclidienne usuelle, nous préconisons l'utilisation de la distance de dissimilarité suivante (*cf.* Friedman & Meulman 2002)

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} \delta_k(\mathbf{x}_i, \mathbf{x}_j)/s_k \tag{7}$$

où $\delta_k(\mathbf{x}_i, \mathbf{x}_j) = (x_i^k - x_j^k)^2$ (resp. $\delta_k(\mathbf{x}_i, \mathbf{x}_j) = I(x_i^k \neq x_j^k)$) si la $k^{\text{ieme}}$ variable est continue (resp. si la $k^{\text{ieme}}$ variable est qualitative) et $s_k = \sum_{i=1}^{n}\sum_{j=1}^{n} \delta_k(\mathbf{x}_i, \mathbf{x}_j)/n^2$, $n$ étant le nombre d'individus.

### *3.2.2. Méthode probabiliste des $k$-plus proches voisins*

Néanmoins, la méthode $k$-ppv présente deux inconvénients. D'abord, le choix du nombre de voisins $k$ dans la méthode précédente est soit fixé d'avance, soit sélectionné selon l'approche de minimisation du taux d'erreur par validation croisée. De plus, les prédictions fournies par l'algorithme $k$-ppv n'ont aucune interprétation probabiliste et l'approche standard du comptage des fréquences de chaque groupe entraîne une discrétisation des prédictions qui dépendent de $k$.

Pour contourner ces difficultés, Holmes & Adams (2002) ont proposé un cadre probabiliste pour $k$-ppv qui accommode l'incertain en $k$ ainsi que la force d'interaction entre voisins. Ils ont formulé leur méthode en un algorithme séquentiel en blocs, où il est supposé que les blocs de données arrivent en fonction du temps : i.e. les données observées $\mathcal{D} = \{(Y_1, \mathbf{X}_1), \ldots, (Y_m, \mathbf{X}_m)\}$ sont composées de $m$ blocs, où $Y_s = \{y_1^{(s)}, \ldots, y_{n_s}^{(s)}\}$ est l'ensemble des affectations des $n_s$ observations du $s^{\text{ieme}}$ bloc et $\mathbf{X}_s = \{\mathbf{x}_1^{(s)}, \ldots, \mathbf{x}_{n_s}^{(s)}\}$ est l'ensemble des prédicteurs associés.

Maintenant, soit $Y = (Y_1, \ldots, Y_m) = (y_1, \ldots, y_n)$ l'ensemble des affectations combinées et soit $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_m) = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ l'ensemble des prédicteurs

combinés, où $n = \sum_{s=1}^{m} n_s$. Holmes & Adams (2002) proposent que la distribution *a priori* jointe de $Y$ s'écrive sous la forme

$$p(Y|\mathbf{X}, \beta, k) = \prod_{i=1}^{n} \frac{\exp(\beta \frac{1}{k} \sum_{j \sim i(k)} \delta_{y_i y_j})}{\sum_{q=1}^{Q} \exp(\beta \frac{1}{k} \sum_{j \sim i(k)} \delta_{\mathrm{q}y_j})}, \tag{8}$$

où $\delta_{\mathrm{ab}}$ est la fonction de Dirac, $\delta_{\mathrm{ab}} = 1$ si $a = b, 0$ sinon, $\beta$ est le paramètre d'interaction qui gouverne la force d'association entre les voisins des $y_i (i = 1, \ldots, n)$, et $\sum_{j \sim i(k)}$ signifie que la somme est faite sur les $k$-plus proches voisins de $\mathbf{x}_i$, selon la métrique $\rho(.)$, appartenant à $\{\mathbf{X}_1, \mathbf{X}_2 \ldots, \mathbf{X}_{t_i}\}_{/\mathbf{x}_i}$, où $t_i$ est l'indice du bloc contenant $\mathbf{x}_i$, $A_{/\mathbf{x}_i}$ signifie que $\mathbf{x}_i$ est retiré de l'ensemble $A$ et $Q$ est le nombre de groupes ($Q = 2$ dans cet article). Le terme $\frac{1}{k} \sum_{j \sim i(k)} \delta_{\mathrm{q}y_j}$ décrit la proportion des $k$ plus proches voisins de $\mathbf{x}_i$ du groupe $q$.

La distribution prédictive de la nouvelle observation s'écrit

$$p(y_{n+1}|\mathbf{x}_{n+1}, Y, \mathbf{X}, \beta, k) = \frac{\exp(\beta \frac{1}{k} \sum_{j \sim n+1(k)} \delta_{y_{n+1} y_j})}{\sum_{q=1}^{Q} \exp(\beta \frac{1}{k} \sum_{j \sim n+1(k)} \delta_{\mathrm{q}y_j})}, \tag{9}$$

ainsi le groupe le plus probable pour $y_{n+1}$ est donné par le groupe le plus commun à ses $k$ plus proches voisins. Le paramètre d'interaction $\beta$ joue le rôle d'un coefficient de régression et les deux équations $(8)$ et $(9)$ ont la forme d'une régression logistique locale sur les $k$ plus proches voisins.

Le traitement de $\beta$ et $k$ comme *a priori* connus et fixés n'est pas réaliste, et il ne tient pas compte de l'incertitude d'une composante essentielle dans le modèle. Pour accommoder cette incertitude, Holmes & Adams (2002) affectent des distributions *a priori* à $\beta$ et $k$, qui entraînent que la distribution prédictive marginale s'écrit

$$p(y_{n+1}|\mathbf{x}_{n+1}, Y, \mathbf{X}) = \sum_{k} \int p(y_{n+1}|\mathbf{x}_{n+1}, Y, \mathbf{X}, \beta, k) p(\beta, k|Y, \mathbf{X}) d\beta, \tag{10}$$

où $p(\beta, k|Y, \mathbf{X}) \propto p(Y|\mathbf{X}, \beta, k) p(\beta, k)$. Un algorithme de simulation de chaîne de Markov MCMC est proposé par les auteurs pour approximer l'équation $(10)$.

L'attrait principal de la méthode est qu'aucune hypothèse sur la distribution des prédicteurs n'est supposée. Par ailleurs, la méthode est complètement automatique, avec juste deux paramètres inconnus auxquels ont été affectés des lois *a priori* de type uniformes (Holmes & Adams 2002). Mais, un inconvénient majeur est que les calculs par simulation peuvent être coûteux.

Pour notre exemple de données mixtes, nous avons adapté leur programme à notre cadre en remplaçant la métrique euclidienne par la métrique (7) qui est plus adaptée aux variables mixtes.

#### *3.2.3. Méthode des $k$-plus proches voisins pour variables catégorielles*

Buttrey (1998) a proposé une technique pour adapter les $k$-plus proches voisins aux variables qualitatives. L'idée est de remplacer d'une manière « optimale » chaque modalité d'une variable par un nombre réel. Du coup, on peut appliquer sur le nouveau tableau de données, avec variables quantitatives, la méthode $k$-ppv avec la métrique euclidienne.

Considérons par exemple le cas d'une seule variable $\mathbf{x} = (x_1, \ldots x_n)^t$, où chaque $x_i$ prend une valeur entière parmi $m_1$ valeurs, disons $1, 2, \ldots, m_1$ pour fixer les idées, tandis que $y_i$ prend une valeur de groupe parmi $Q$ valeurs. Soit $\phi(.)$ la transformation qui convertit les $n$ valeurs entières de $\mathbf{x}$ en $n$ nombres réels non nécessairement uniques : $\phi(\mathbf{x}) = (\phi_1, \ldots, \phi_n)^t$. La maximisation du rapport de la variance totale sur la variance intra-groupes permet de sélectionner les $\phi_j = \phi(j)(1 \leqslant j \leqslant m_1)$ «optimaux». Cette approche s'étend de la même manière au cadre de $p$ variables qualitatives (*cf.* Buttrey 1998 pour plus de détails). La solution est en fait donnée par le premier facteur de l'analyse des correspondances du tableau croisant $Y$ et $\mathbf{x}$.

De même, la méthode permet de traiter le problème de discrimination avec variables mixtes. Ce traitement est fondé sur la transformation des variables quantitatives en variables qualitatives, via une représentation similaire à une fonction spline linéaire (*cf.* Buttrey 1998 pour plus de détails). L'approche a été implémentée sous le langage $R$, et elle est disponible dans les «paquetages»sous le nom «knncat».

L'avantage de «knncat» est qu'elle tient compte de la variable à discriminer $Y$, tandis que la méthode Disqual (Saporta 1977), fondée sur le codage des variables qualitatives en utilisant l'analyse des correspondances multiples, n'en tient pas compte (*cf.* Carlier 1994 pour plus de détails). Par ailleurs, les résultats encourageants obtenus par knncat sur plusieurs exemples classiques de dimension importante, disponibles sur le «web», nous ont poussés à la tester sur notre exemple de données bancaires.

### *3.3. Discrimination barycentrique*

La méthode de la discrimination barycentrique (DB) semble, en fait, plus intéressante que Disqal, car elle prend en compte les différences de répartition des prédicteurs entre les groupes (*cf.* Nakache *et al.* 1977). Récemment, Carlier (1994) a présenté une revue intéressante des méthodes exploratoires pour l'analyse discriminante sur variables qualitatives. Contrairement à knncat, DB est facile à mettre en œuvre avec un programme d'analyse factorielle des correspondances (AFC).

En effet, DB se définit ainsi : à partir du tableau des données $\mathbf{X}$ de $p$ variables qualitatives, on construit le tableau $\mathbf{C}$ à $Q$ lignes ($Q$ étant le nombre de groupes *à priori*) et $m = \sum_{j=1}^{p} m_j$ colonnes ($m_j$ est le nombre de modalités de la $j$ème

variable) défini par

$$C_\ell^j = \sum_{\mathbf{x}_i \in G_\ell} x_i^j \quad \text{pour } \ell = 1, \ldots, Q \text{ et } j = 1, \ldots, m$$

où $G_\ell$ est l'échantillon d'apprentissage du groupe $\ell$. Par conséquent, le vecteur $\mathbf{C}_\ell$ est proportionnel au centre de gravité du groupe $\ell$. Ensuite on effectue l'analyse des correspondances du tableau $\mathbf{C}$. La projection des individus $\mathbf{x}_i$ $(i = 1, \ldots, n)$, considérés comme lignes supplémentaires du tableau $\mathbf{C}$, sur les facteurs de $\mathbb{R}^m$ de cette analyse constitue le codage à partir duquel on peut utiliser toute règle de décision. Pour notre exemple, nous avons considéré sur le tableau transformé la règle de décision fondée sur la distance du khi-deux (notée DB), sur $k$-ppv (notée DB$k$-ppv) et sur MP$k$-ppv (notée DBMP$k$-ppv).

### *3.4. Méthode de discrimination fondée sur une distance*

La règle de discrimination basée sur la distance est la plus simple, la plus ancienne et elle est formellement attribuée à Matusita (1956). Dans le cas de deux groupes à discriminer, elle consiste à affecter une observation $\mathbf{v}$ au groupe le plus «proche» : affecter $\mathbf{v}$ au groupe $G_j$ si

$$\rho(\mathbf{v}, G_j) = \min[\rho(\mathbf{v}, G_1), \rho(\mathbf{v}, G_2)], j = 1, 2,$$

où $\rho(\mathbf{v}, G_j)$ désigne la distance entre l'observation $\mathbf{v}$ et le groupe $G_j$. Krzanowski (1993) présente une excellente revue sur les différentes méthodes de discrimination fondées sur les distances et leurs propriétés asymptotiques.

À notre connaissance, l'approche récente la plus intéressante fondée sur la distance est due à Cuadras (1989) en s'inspirant de l'indice de diversité de Rao (1982). Dans le cas de deux groupes à discriminer, si on désigne par $(\mathbf{x}_1^{(1)}, \ldots \mathbf{x}_{n_1}^{(1)})$ et $(\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{n_2}^{(2)})$ les échantillons respectifs des groupes $G_1$ et $G_2$, Cuadras (1989) définit les deux fonctions discriminantes

$$F_1(\mathbf{v}) = \frac{1}{n_1} \sum_{i=1}^{n_1} \rho^2(\mathbf{v}, \mathbf{x}_i^{(1)}) - \frac{1}{2n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \rho^2(\mathbf{x}_i^{(1)}, \mathbf{x}_j^{(1)});$$

$$F_2(\mathbf{v}) = \frac{1}{n_2} \sum_{i=1}^{n_2} \rho^2(\mathbf{v}, \mathbf{x}_i^{(2)}) - \frac{1}{2n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \rho^2(\mathbf{x}_i^{(2)}, \mathbf{x}_j^{(2)}) \tag{11}$$

et affecte $\mathbf{v}$ au groupe $G_j$ si $F_j(\mathbf{v}) = \min(F_1(\mathbf{v}), F_2(\mathbf{v}))$, $j = 1, 2$, où $\rho$ est un indice de distance.

L'avantage de cette approche, qu'on appellera DistCuad, est qu'elle opère exclusivement sur les distances entre observations au lieu de distances entre groupes généralement proposées dans la littérature sur ce sujet (*cf.* Bar-Hen et Daudin 1995

pour une nouvelle distance entre populations pour variables mixtes). Ainsi, nous pouvons utiliser n'importe quelle mesure standard de classification qui tient compte non seulement de variables hétérogènes, mais aussi de certains obstacles comme les données manquantes. Cuadras et Fortiana (1997) ont prouvé que cette méthode peut fournir un bon résultat sur un exemple de dimension modérée. Ils ont considéré l'indice de coefficient de similarité de Gower (1971), similaire à (7) avec $s_k = 1$ pour tout $k$, pour le choix de $\rho$. Pour notre application réelle, nous avons choisi pour $\rho$ l'indice (7) comme pour les méthodes de $k$ppv.

### *3.5. Règle de décision bayésienne*

Pour mieux comprendre le point de vue bayésien adopté dans nos applications, avec l'introduction des coûts de mauvais classement et des probabilités *a priori*, nous rappelons brièvement la manière dont la règle de décision bayésienne est construite.

On dispose d'un échantillon de $n$ individus décrits par $p$ variables explicatives $X^1, \ldots, X^p$ dont l'appartenance à l'un des deux groupes *a priori* $G_1, G_2$ est connu. Le modèle statistique le plus général pour définir une règle de décision optimale est le modèle bayésien. La règle de décision bayésienne est celle qui minimise l'espérance du coût de mauvaise classification. Cette règle optimale dépend essentiellement des probabilités *a priori* $Pr(G_\ell), \ell = 1, 2$ des groupes ($Pr(G_\ell) \geqslant 0$ pour tout $\ell$ et $Pr(G_1) + Pr(G_2) = 1$), des coûts de mauvaise classification $C(\ell, k)$, qui représentent le coût de mauvais classement d'un individu de $G_k$ dans $G_\ell$ (on a bien sûr $C(\ell, \ell) = 0$), et des densités de probabilité par groupe $f_\ell(\mathbf{x}), \ell = 1, 2, \mathbf{x}$ appartenant à l'ensemble des valeurs possibles des variables explicatives. Ainsi la règle de décision optimale est définie par

$$\begin{aligned} \mathbf{x} \text{ est affecté à } \quad G_1 \quad \text{si} \quad C(1,2)P(G_2|\mathbf{x}) < C(2,1)P(G_1|\mathbf{x}) \\ \mathbf{x} \text{ est affecté à } \quad G_2 \quad \text{si} \quad C(1,2)P(G_2|\mathbf{x}) > C(2,1)P(G_1|\mathbf{x}), \end{aligned}$$

$P(G_\ell|\mathbf{x})$ désignant la probabilité *a postériori* du groupe $G_\ell$. En utilisant la formule de Bayes, la règle de Bayes peut donc s'écrire

$$\begin{aligned} \mathbf{x} \text{ est affecté à } \quad G_1 \quad \text{si} \quad C(1,2)Pr(G_2)f_2(\mathbf{x}) < C(2,1)Pr(G_1)f_1(\mathbf{x}) \\ \mathbf{x} \text{ est affecté à } \quad G_2 \quad \text{si} \quad C(1,2)Pr(G_2)f_2(\mathbf{x}) > C(2,1)Pr(G_1)f_1(\mathbf{x}). \end{aligned}$$

Ainsi, la construction effective d'une règle de décision revient à estimer les probabilités *a priori* des groupes, les coûts de mauvais classement et les densités de probabilité par groupe. Les deux premières quantités sont en général spécifiées, et ainsi l'opération principale de la discrimination à but décisionnel est l'estimation des densités par groupe $f_\ell(\mathbf{x}), \ell = 1, 2$. Les méthodes d'estimation les plus utilisées sont fondées sur la méthode d'estimation par maximum de vraisemblance pour les modèles paramétriques (comme le modèle de location) ou sur l'estimation non paramétrique de la densité (comme les $k$-plus proches voisins). Pour notre application sur données réelles, trois options (définies en Section 4) sont considérées pour le choix des coûts de mauvais classement et des probabilités *a priori*.

Par ailleurs, pour adapter les méthodes de discrimination basées sur les distances (DB et Discuad), définies en Sections 3.3 et 3.4, au cadre de discrimination bayésienne, nous avons utilisé une transformation de type $K\exp(-x/2)$, où $K$ est une constante de normalisation et $x$ le carré d'une distance ou son équivalent ($x = F_\ell(v), \ell = 1, 2$ dans le cas de la méthode DistCuad, *cf.* formule $(11)$). Cette transformation permet de retrouver les règles de décision initiales (définies en sections 3.3 et 3.4) dans le cas d'égalité des coûts et d'égalité des probabilités *a priori*.

## 4. Application aux données réelles

Dans cette section, nous comparons les résultats des trois familles de méthodes de scoring précédentes sur les données bancaires décrites en Section 2 et sur un autre exemple de données australiennes de crédit dont les groupes sont relativement équilibrés.

La première famille est composée de trois méthodes fondées sur le modèle de location : le modèle de location linéaire homoscédastique (noté MLH) basé sur les équations $(1)$ et (2), le modèle de location non-paramétrique (noté MLNP-H) basé sur les équations (3) et (4) et enfin, le modèle de location non-paramétrique (noté MLNP-MIC) basé sur les équations (3) et (5).

La seconde famille est composée de quatre méthodes fondées sur les $k$-plus proches voisins qui utilisent la distance adaptée aux variables mixtes définie par $(7)$ : il s'agit de la méthode classique des $k$-plus proches voisins (notée $k$ppv), de la méthode probabiliste des $k$ppv (notée MP$k$ppv), de la méthode de $k$ppv sur variables catégorielles (notée $k$nncat) et de deux méthodes de discrimination barycentrique : l'une avec les $k$ppv notée DB$k$ppv et l'autre avec MP$k$ppv notée DBMP$k$ppv.

La dernière famille se limite à la méthode fondée sur une distance : la distance de Cuadras définie par l'équation (11) (notée DistCuad) en utilisant pour $\rho$ l'indice $(7)$ et la discrimination barycentrique avec une distance du chi-deux (notée DB).

### *4.1. Données de crédit de la Compagnie Bancaire*

Les résultats des différentes méthodes sont résumés dans le tableau 2 et le tableau 3 ci-dessous. Nous affichons les résultats obtenus sur l'échantillon d'apprentissage et l'échantillon test. Dans chaque cas, nous fournissons le nombre d'individus mal classés (respectivement le pourcentage de mauvais classement) dans les deux groupes et le nombre total d'individus mal classés (respectivement le poucentage global de mauvais classement). Nous fournissons aussi, entre parenthèses, pour certaines méthodes, l'estimation de certains paramètres de contrôle (paramètre de lissage $\lambda$, nombre de voisins $k$). Par ailleurs, trois options pour les choix des coûts de mauvais classement et des probabilités *a priori* ont été considérées afin de rendre sens au risque de Bayes inconditionnel :

$(1)$ : $Pr(G_1)=Pr(G_2) = 1/2, C(2,1)=n_2/(n_1+n_2)$ et $C(1,2) = n_1/(n_1+n_2)$,

$(2)$ : $Pr(G_1) = C(1,2) = n_1/(n_1+n_2)$ et $Pr(G_2) = C(2,1) = n_2/(n_1+n_2)$,

$(3)$ : $Pr(G_1) = Pr(G_2) = 1/2$ et $C(1,2) = C(2,1) = 1$.

Toutes les méthodes, sauf knncat, ont été comparées pour les trois options précédentes. La méthode knncat a été considérée uniquement pour l'option (3) pour laquelle elle a été programmée pour le logiciel R. Nous l'avons considérée, pour cette option, seulement pour comparer sa performance par rapport à la méthode similaire DB$k$ppv qui est plus simple à programmer.

TABLEAU 2

*Effectifs et nombre de mauvais classement sur les échantillons d'apprentissage et de test de la compagnie bancaire*

| | Effectifs totaux | 4135 | 3888 | 247 | 2043 | 1918 | 125 |
|---|---|---|---|---|---|---|---|
| Options | Méthodes | | Apprentissage | | | Test | |
| | | Total | G1 | G2 | Total | G1 | G2 |
| (1) | MLH | 247 | 0 | 247 | 125 | 0 | 125 |
| | MLNP-H ($\lambda_{\text{opt}} = 0.10$) | 118 | 0 | 118 | 124 | 0 | 124 |
| | MLNP-MIC ($\lambda_{\text{opt}} = 0.10$) | 118 | 0 | 118 | 124 | 0 | 124 |
| | $k$ppv ($k_{\text{opt}} = 1$) | 185 | 110 | 75 | 191 | 72 | 119 |
| | MP$k$ppv | 301 | 57 | 244 | 142 | 18 | 124 |
| | DB | 247 | 0 | 247 | 125 | 0 | 125 |
| | DB$k$ppv ($k_{\text{opt}} = 7$) | 255 | 9 | 246 | 136 | 11 | 125 |
| | DBMP$k$ppv | 277 | 31 | 246 | 126 | 3 | 123 |
| | DistCuad | 1104 | 1101 | 3 | 588 | 504 | 84 |
| (2) | MLH | 1596 | 1595 | 1 | 846 | 778 | 68 |
| | MLNP-H ($\lambda_{\text{opt}} = 0.10$) | 51 | 40 | 11 | 109 | 38 | 71 |
| | MLNP-MIC ($\lambda_{\text{opt}} = 0.10$) | 212 | 212 | 0 | 154 | 144 | 10 |
| | $k$ppv ($k_{\text{opt}} = 1$) | 185 | 113 | 72 | 192 | 74 | 118 |
| | MP$k$ppv | 301 | 59 | 242 | 143 | 20 | 123 |
| | DB | 247 | 0 | 247 | 125 | 0 | 125 |
| | DB$k$ppv($k_{\text{opt}} = 7$) | 255 | 10 | 245 | 135 | 11 | 124 |
| | DBMP$k$ppv | 277 | 33 | 244 | 125 | 3 | 122 |
| | DistCuad | 10 | 0 | 10 | 125 | 0 | 125 |
| (3) | MLH | 1596 | 1595 | 1 | 846 | 778 | 68 |
| | MLNP-H ($\lambda_{\text{opt}} = 0.10$) | 51 | 40 | 11 | 109 | 38 | 71 |
| | MLNP-MIC ($\lambda_{\text{opt}} = 0.10$) | 212 | 212 | 0 | 154 | 144 | 10 |
| | $k$ppv ($k_{\text{opt}} = 1$) | 148 | 4 | 144 | 141 | 17 | 124 |
| | MP$k$ppv | 247 | 0 | 247 | 125 | 0 | 125 |
| | knncat | 256 | 11 | 245 | 127 | 2 | 125 |
| | DB | 247 | 0 | 247 | 125 | 0 | 125 |
| | DB$k$ppv ($k_{\text{opt}} = 7$) | 247 | 0 | 247 | 125 | 0 | 125 |
| | DBMP$k$ppv | 247 | 0 | 247 | 125 | 0 | 125 |
| | DistCuad | 0 | 0 | 0 | 127 | 2 | 125 |

TABLEAU 3
*Pourcentages de mauvais classement sur les échantillons d'apprentissage et de test de la compagnie bancaire*

| | Effectifs totaux | 4135 | 3888 | 247 | 2043 | 1918 | 125 |
|---|---|---|---|---|---|---|---|
| Options | Méthodes | Apprentissage | | | Test | | |
| | | Total | G1 | G2 | Total | G1 | G2 |
| (1) | MLH | 5.97 | 0 | 100 | 6.12 | 0 | 100 |
| | MLNP-H ($\lambda_{\text{opt}} = 0.10$) | 2.85 | 0 | 44.77 | 6.07 | 0 | 99.20 |
| | MLNP-MIC ($\lambda_{\text{opt}} = 0.10$) | 2.85 | 0 | 44.77 | 6.07 | 0 | 99.20 |
| | $k$ppv ($k_{\text{opt}} = 1$) | 4.47 | 2.83 | 30.36 | 9.35 | 3.75 | 95.20 |
| | MP$k$ppv | 7.28 | 1.47 | 98.79 | 6.95 | 0.94 | 99.20 |
| | DB | 5.97 | 0 | 100 | 6.12 | 0 | 100 |
| | DB$k$ppv ($k_{\text{opt}} = 7$) | 6.17 | 0.23 | 99.60 | 6.66 | 0.57 | 100 |
| | DBMP$k$ppv | 6.70 | 0.80 | 99.60 | 6.17 | 0.16 | 99.40 |
| | DistCuad | 26.70 | 28.32 | 1.21 | 28.78 | 26.28 | 67.20 |
| (2) | MLH | 38.60 | 41.02 | 0.40 | 41.41 | 40.56 | 54.40 |
| | MLNP-H ($\lambda_{\text{opt}} = 0.10$) | 1.23 | 1.03 | 4.45 | 5.34 | 1.98 | 56.80 |
| | MLNP-MIC ($\lambda_{\text{opt}} = 0.10$) | 5.13 | 5.45 | 0 | 7.54 | 7.50 | 8 |
| | $k$ppv ($k_{\text{opt}} = 1$) | 4.47 | 2.91 | 29.15 | 9.40 | 3.86 | 94.40 |
| | MP$k$ppv | 7.28 | 1.52 | 97.98 | 7 | 1.04 | 98.40 |
| | DB | 5.97 | 0 | 100 | 6.12 | 0 | 100 |
| | DB$k$ppv($k_{\text{opt}} = 7$) | 6.17 | 0.26 | 99.19 | 6.61 | 0.57 | 99.20 |
| | DBMP$k$ppv | 6.70 | 0.85 | 98.79 | 6.12 | 0.16 | 97.60 |
| | DistCuad | 2.4 | 0 | 4.05 | 6.2 | 0 | 100 |
| (3) | MLH | 38.6 | 41.01 | 0.4 | 41.41 | 40.56 | 54.4 |
| | MLNP-H ($\lambda_{\text{opt}} = 0.10$) | 1.23 | 1.03 | 4.45 | 5.34 | 1.98 | 56.8 |
| | MLNP-MIC ($\lambda_{\text{opt}} = 0.10$) | 5.13 | 5.45 | 0 | 7.54 | 7.5 | 8 |
| | $k$ppv ($k_{\text{opt}} = 1$) | 3.6 | 1 | 58.3 | 6.9 | .9 | 99.2 |
| | MP$k$ppv | 6 | 0 | 100 | 6 | 0 | 100 |
| | knncat | 6.2 | .3 | 99.2 | 6.2 | .1 | 100 |
| | DB | 6 | 0 | 100 | 6 | 0 | 100 |
| | DB$k$ppv($k_{\text{opt}} = 7$) | 6 | 0 | 100 | 6 | 0 | 100 |
| | DBMP$k$ppv | 6 | 0 | 100 | 6 | 0 | 100 |
| | DistCuad | 0 | 0 | 0 | 6.2 | .05 | 100 |

Il est bien connu que le taux d'erreur apparent d'une règle de décision, estimé sur l'échantillon d'apprentissage, est très optimiste. Par conséquent, nos commentaires suivants sur la performance de différentes méthodes de scoring sera fondée essentiellement sur les résultats sur l'échantillon test. D'après la lecture des résultats sur l'échantillon test, nous distinguons quatre points saillants suivants.

– Pour l'option (1), où les coûts de mauvais classement sont proportionnels aux effectifs des groupes, les différentes méthodes fournissent un pourcentage de mauvais classement inconditionnel sur l'échantillon test comparable de l'ordre de $6\%$, sauf $k$ppv ($9.35\%$) et DistCuad ($28.78\%$). Les pourcentages de mauvais classement, sur le fichier test, de différentes méthodes sont relativement identiques pour l'option $(2)$ et $(3)$, sauf pour $k$ppv qui a produit une réduction de pourcentage de mauvais classement inconditionnel de l'ordre $(2.5\%)$ pour l'option $(3)$.

– Mais, selon le nombre conditionnel de mauvais classements sur l'échantillon test et dans le cadre de l'option $(1)$, DistCuad fournit le meilleur taux d'erreur conditionnel au groupe $G_2$ de l'ordre de $67.20\%$ et le plus fort taux d'erreur conditionnel au groupe $G_1$ ($26.28\%$). Les autres méthodes classent bien les obervations du groupe « bons payeurs » $G_1$ avec un taux d'erreur qui oscille entre $0$ et $3.7\%$, tandis qu'elles classent mal presque tous les éléments du second groupe « mauvais payeurs » $G_2$ d'effectif très faible.

– D'autre part, pour les options $(2)$ et $(3)$, les méthodes fondées sur le modèle de location ont produit les meilleurs pourcentages de mauvais classement par rapport à $G_2$ sur le fichier test. MLNP-MIC domine largement MLNP-H et MLH sur $G_2$ ($8\%$ contre $54.4\%$ pour MLH et $56.8\%$ pour MLNP-H), alors que MLNP-H domine ces deux dernières sur $G_1$ ($1.98\%$ contre $7.5\%$ pour MLNP-MIC et $40.6\%$ pour MLH). Les autres méthodes affectent presque parfaitement les observations du groupe $G_1$ (taux d'erreur oscillant entre $0$ et $3.86\%$) et classent mal presque toutes les observations du groupe $G_2$ (taux d'erreur oscillant entre $94.4$ et $100\%$ pour l'option $(2)$ et $99.2$ et $100\%$ pour l'option $(3)$). Ce qui signifie que la méthode d'estimation de la probabilité discrète dans le modèle de location joue un rôle important dans la règle de décision. Par ailleurs, il faut signaler que le taux global (ou conditionnel à $G_1$) très fort obtenu par MLH justifie amplement la procédure d'ajustement de la moyenne de chaque variable continue dans une cellule par une moyenne pondérée sur toutes les cellules $(3)$, procédure qui est très bénéfique sur cet exemple.

– Par ailleurs, pour l'option $(3)$ d'égalité des coûts de mauvais classement et des probabilités *a priori*, la méthode DB$k$ppv a un comportement similaire à knncat, mais elle nécessite peu de calculs et est donc plus facile à mettre en œuvre que knncat. De plus il est surprenant que DistCuad affecte correctement toutes les observations de l'échantillon d'apprentissage pour l'option $(3)$.

### *4.2. Données australiennes de crédit*

Les données australiennes de crédit ont été téléchargées du site web « UCI Repository : www.liacc.up.pt/ML/statlog/datasets.html ». Elles ont été énormément utilisées, pour la comparaison de différentes méthodes récentes d'apprentissage supervisé, essentiellement par la communauté de l'intelligence artificielle. Au total, on dispose de $690$ observations décrites par $14$ variables mixtes ($8$ quantitatives et $6$ qualitatives) où certaines valeurs sont manquantes. Contrairement à l'exemple précédent, les deux groupes qui composent cette population sont relativement équilibrés : $383$ observations ($55.5\%$) pour $G_1$ et $307$ observations ($44.5\%$) pour $G_2$. Pour éviter le

problème de données manquantes, nous avons éliminé 3 variables qualitatives, et nous avons subdivisé notre échantillon en deux échantillons d'apprentissage et test, dont les fréquences sont résumées dans le tableau 4. Ainsi, les deux groupes sont relativement équilibrés et de plus le nombre de variables discrètes est inférieur à celui des variables quantitatives. Ce qui *a priori* ne va certainement pas avantager les méthodes fondées sur le modèle de location. Les résultats des différentes méthodes sont résumées dans les tableaux 5 et 6.

TABLEAU 4
*Fréquences des deux modalités de la variable à discriminer Y pour données australiennes de crédit*

| Y | Apprentissage | Test |
|---|---|---|
| 0 | 235 (56%) | 148 (55%) |
| 1 | 185 (44%) | 122 (45%) |
| Total | 420 | 270 |

Nous constatons, que dans le cas de l'option $(1)$, DistCuad réalise le plus faible pourcentage de mauvais classement global sur l' échantillons test égal à $11.48\%$, avec un taux d'erreur nul sur $G_1$ et un score correct sur $G_2$ égal à $25.41\%$. Elle est suivie, sur l'échantillon test, par $k$ppv et MLNP-H qui réalisent respectivement un taux global de l'ordre de $25.19\%$ et $30.37\%$. Mais MLNP-H réalise, sur le fichier test, un taux conditionnel remarquable $(13.11\%)$ sur $G_2$ beaucoup plus faible que celui de DistCuad $(25.41\%)$ et $k$ppv $(59.90\%)$, tandis que DB classe mal toutes les observations de $G_2$ $(100\%)$ et classe parfaitement toutes les observations de $G_1$ $(0\%)$, alors que MP$k$ppv et DBMP$k$ppv produisent un résultat complètement opposé à celui de DB.

Dans le cas de l'option $(2)$ et sur le fichier test, DistCuad domine toujours avec un taux global et conditionnel par rapport à $G_1$ et $G_2$ spectaculairement faible. Elle est suivie par respectivement $k$ppv et MLNP-H, alors que les méthodes MP$k$ppv, DB et DBMP$k$ppv se comportent de la même manière que DB dans $(1)$; elles classent parfaitement les observations de $G_1$ et déclassent toutes les observations de $G_2$.

Par contre dans le cas de l'option $(3)$ d'égalité des coûts et de probabilité *a priori*, DB$k$ppv domine avec un taux global et conditionnel par rapport à $G_2$ spectaculairement faible sur le fichier test. Elle est suivie par respectivement knncat, MP$k$ppv, $k$ppv et MLNP-H. DBMP$k$ppv et DistCuad se comportent relativement de la même manière que DBMP$k$ppv dans le cas de l'option $(2)$.

Par conséquent, cet exemple montre que la discrimination fondée sur une distance DistCuad peut fournir, sur fichier test, un taux d'erreur très faible dans le cas des coûts proportionnels aux effectifs des groupes (i.e. option $(1)$ ou $(2)$), mais elle s'effondre complètement dans le cadre usuel d'égalité des coûts et d'égalité de probabilités *a priori* des groupes (i.e. l'option $(3)$) avec un taux d'erreur global très fort. Les trois méthodes fondées sur DB ont fourni des résultats médiocres dans le cadre des options $(1)$ ou $(2)$, mais elles ont fourni un résultat complètement opposé au

précédent (sauf DBMP$k$ppv) dans le cas de l'option (3), avec un résultat spectaculaire pour DB$k$ppv.

TABLEAU 5
*Effectifs et nombre de mauvais classement sur les échantillons d'apprentissage et de test de données australiennes de crédit*

| | Effectifs totaux | 420 | 235 | 185 | 270 | 148 | 122 |
|---|---|---|---|---|---|---|---|
| Options | Méthodes | Apprentissage | | | Test | | |
| | | Total | G1 | G2 | Total | G1 | G2 |
| (1) | MLH | 77 | 67 | 10 | 116 | 74 | 42 |
| | MLNP-H ($\lambda_{opt} = 0.10$) | 1 | 1 | 0 | 82 | 66 | 16 |
| | MLNP-MIC ($\lambda_{opt} = 0.10$) | 0 | 0 | 0 | 99 | 74 | 25 |
| | $k$ppv ($k_{\text{opt}} = 1$) | 62 | 11 | 51 | 68 | 12 | 56 |
| | MP$k$ppv | 235 | 235 | 0 | 148 | 148 | 0 |
| | DB | 185 | 0 | 185 | 122 | 0 | 122 |
| | DB$k$ppv ($k_{\text{opt}} = 1$) | 163 | 0 | 163 | 108 | 1 | 107 |
| | DBMP$k$ppv | 235 | 235 | 0 | 148 | 148 | 0 |
| | DistCuad | 56 | 0 | 56 | 31 | 0 | 31 |
| (2) | MLH | 47 | 27 | 20 | 120 | 59 | 61 |
| | MLNP-H ($\lambda_{\text{opt}} = 0.10$) | 0 | 0 | 0 | 67 | 22 | 45 |
| | MLNP-MIC ($\lambda_{\text{opt}} = 0.10$) | 0 | 0 | 0 | 113 | 61 | 52 |
| | $k$ppv ($k_{\text{opt}} = 1$) | 59 | 19 | 40 | 59 | 14 | 45 |
| | MP$k$ppv | 79 | 6 | 73 | 122 | 0 | 122 |
| | DB | 193 | 12 | 181 | 122 | 0 | 122 |
| | DB$k$ppv ($k_{\text{opt}} = 1$) | 163 | 0 | 163 | 108 | 1 | 107 |
| | DBMP$k$ppv | 185 | 0 | 185 | 122 | 0 | 122 |
| | DistCuad | 56 | 0 | 56 | 31 | 0 | 31 |
| (3) | MLH | 47 | 27 | 20 | 120 | 59 | 61 |
| | MLNP-H ($\lambda_{\text{opt}} = 0.10$) | 0 | 0 | 0 | 67 | 22 | 45 |
| | MLNP-MIC ($\lambda_{\text{opt}} = 0.10$) | 0 | 0 | 0 | 113 | 61 | 52 |
| | $k$ppv ($k_{\text{opt}} = 1$) | 59 | 27 | 32 | 55 | 16 | 39 |
| | MP$k$ppv | 103 | 15 | 88 | 55 | 18 | 37 |
| | knncat | 54 | 25 | 29 | 45 | 13 | 32 |
| | DB | 0 | 0 | 0 | 59 | 0 | 59 |
| | DB$k$ppv ($k_{\text{opt}} = 1$) | 3 | 3 | 0 | 5 | 5 | 0 |
| | DBMP$k$ppv | 180 | 0 | 180 | 122 | 0 | 122 |
| | DistCuad | 134 | 0 | 134 | 114 | 0 | 114 |

TABLEAU 6
*Pourcentages de mauvais classement sur les échantillons d'apprentissage et de test de données australiennes de crédit*

| | Effectifs totaux | 420 | 235 | 185 | 270 | 148 | 122 |
|---|---|---|---|---|---|---|---|
| Options | Méthodes | Apprentissage | | | Test | | |
| | | Total | G1 | G2 | Total | G1 | G2 |
| | MLH | 18.33 | 28.51 | 5.41 | 42.96 | 50 | 34.43 |
| | MLNP-H ($\lambda_{\text{opt}} = 1$) | 0.02 | 0.04 | 0 | 30.37 | 44.59 | 13.11 |
| | MLNP-MIC ($\lambda_{\text{opt}} = 1$) | 0 | 0 | 0 | 41.8 | 41.2 | 42.6 |
| | $k$ppv ($k_{\text{opt}} = 0.10$) | 14.76 | 4.68 | 27.57 | 25.19 | 8.11 | 59.90 |
| (1) | MP$k$ppv | 55.95 | 100 | 0 | 54.81 | 100 | 0 |
| | DB | 44.05 | 0 | 100 | 45.19 | 0 | 100 |
| | DB$k$ppv ($k_{\text{opt}} = 1$) | 38.81 | 0 | 88.11 | 40 | 0.68 | 87.70 |
| | DBMP$k$ppv | 55.95 | 100 | 0 | 54.81 | 100 | 0 |
| | DistCuad | 13.33 | 0 | 30.27 | 11.48 | 0 | 25.41 |
| | MLH | 11.19 | 11.49 | 10.81 | 44.44 | 39.86 | 50 |
| | MLNP-H ($\lambda_{\text{opt}} = 0.10$) | 0 | 0 | 0 | 24.81 | 14.86 | 36.89 |
| | MLNP-MIC ($\lambda_{\text{opt}} = 0.10$) | 0 | 0 | 0 | 41.85 | 33.78 | 42.62 |
| (2) | $k$ppv ($k_{\text{opt}} = 1$) | 14.05 | 8.05 | 21.62 | 21.85 | 9.46 | 36.89 |
| | MP$k$ppv | 18.81 | 2.55 | 39.46 | 45.19 | 0 | 100 |
| | DB | 45.95 | 5.11 | 97.84 | 45.19 | 0 | 100 |
| | DB$k$ppv ($k_{\text{opt}} = 1$) | 38.81 | 0 | 88.11 | 40 | 0.68 | 87.70 |
| | DBMP$k$ppv | 44.05 | 0 | 100 | 45.19 | 0 | 100 |
| | DistCuad | 13.33 | 0 | 30.27 | 11.48 | 0 | 25.41 |
| | MLH | 11.2 | 11.5 | 10.8 | 44.4 | 39.8 | 50 |
| | MLNP-H ($\lambda_{\text{opt}} = 0.10$) | 0 | 0 | 0 | 24.8 | 14.8 | 36.9 |
| | MLNP-MIC ($\lambda_{\text{opt}} = 0.10$) | 0 | 0 | 0 | 41.8 | 41.2 | 42.6 |
| (3) | $k$ppv ($k_{\text{opt}} = 1$) | 14 | 11.5 | 17.3 | 20.4 | 10.8 | 32 |
| | MP$k$ppv | 24.5 | 6.4 | 47.6 | 20.4 | 12.1 | 30.3 |
| | knncat | 12.8 | 10.6 | 15.7 | 16.7 | 8.9 | 26.2 |
| | DB | 0 | 0 | 0 | 21.8 | 0 | 48.3 |
| | DB$k$ppv ($k_{\text{opt}} = 1$) | .7 | 1.3 | 0 | 1.8 | 3.4 | 0 |
| | DBMP$k$ppv | 42.8 | 0 | 97.3 | 45.2 | 0 | 100 |
| | DistCuad | 31.9 | 0 | 72.4 | 42.2 | 0 | 93.4 |

## 5. Conclusion

Dans cet article, nous avons présenté et illustré la performance de trois familles de méthodes de scoring sur deux bons exemples de données de crédit bancaire

(données de la Compagnie Bancaire et australiennes) dont les effectifs des groupes sont respectivement très déséquilibrés et équilibrés. De plus, les variables explicatives sont de nature hétérogène et leur nombre (23) est relativement important pour le premier exemple et modéré pour le second (11), le nombre de variables qualitatives pour ce dernier étant faible (3) et inférieur à celui des variables quantitatives.

Notre expérience sur le premier exemple, d'effectifs de groupes très déséquilibrés, a montré que l'utilisation des méthodes non-paramétriques fondées sur le modèle de location peut fournir d'excellents résultats dans le cadre du scoring pour les cas des options (2) et (3). Elles ont l'avantage de fournir un taux global de mauvais classement faible (surtout MPNP-H) et de très bons taux conditionnels surtout pour le groupe des "mauvais payeurs" d'effectif très faible. Néanmoins, si l'intérêt est porté sur les règles de décision linéaires qui facilitent l'interprétabilité de la contribution de chaque variable à la fonction score, on peut dans ce cas préconiser la méthode MPNP-MIC (qui, de plus, classe mieux les «mauvais payeurs»). Les autres méthodes ont tendance à fournir des taux de mauvais classement très déséquilibrés : elles classent parfaitement les «bons payeurs» et éprouvent d'énormes difficultés à bien classer les «mauvais payeurs». Par ailleurs, l'utilisation de la procédure d'ajustement de la moyenne d'une cellule donnée par la formule (3), et la méthode d'estimation non paramétrique de la probabilité discrète ont été très bénéfiques sur cet exemple.

Le second exemple, d'effectifs de groupes équilibrés, a mis en évidence le comportement opposé de la discrimination fondée sur une distance DistCuad avec les méthodes fondées sur la discrimination barycentrique (sauf MPDB$k$ppv). DistCuad domine toutes les méthodes dans le cas des options (1) et (2), mais elle produit un très mauvais résultat dans le cas de l'option usuelle (3), alors que les méthodes fondées sur DB (sauf MPDB$k$ppv) fournissent un résultat complètement opposé à ce dernier. Les méthodes fondées sur les $k$ppv et le modèle de location non-paramétrique gardent une position intermédiaire pour les différentes options considérées pour le choix des coûts et de probabilités *a priori* des groupes.

**Remerciements**

Les auteurs remercient les deux rapporteurs anonymes pour leur commentaires pertinents qui nous ont aidés à améliorer la présentation de ce travail et Samuel Buttrey pour avoir accepté de nous fournir les résultats de sa méthode knncat.

## Références

ARMINGER G., ENACHE D. et BONNE T. (1997), Analyzing credit risk data : A comparison of logistic discrimination, classification tree analysis, and feedforward networks, *Computational Statistics, Special issue* : 10 Years AG GLM, **12**, 293-310.

ASPAROUKHOV O. et DANCHEV S. (1997), Discrimination and classification in the presence of binary variables, *Biocybernetics and Biomedical Engineering* **17**, (1-2) : 25-39.

ASPAROUKHOV O. et KRZANOWSKI W. J. (2000), Non-parametric smoothing of the location model in mixed variable discrimination, *Statistics and Computing*, **10**, 283-297.

BARDOS M. (2001), *Analyse discriminante, application au risque et scoring financier*, Dunod.

BAR-HEN A. et DAUDIN D. (1995), Generalization of the Mahalanobis distance in mixed case, *Journal of Multivariate Analysis*, **52**, 332-342.

BESSE PH., LE GALL C., RAIMBAULT N. et SARPY S. (2001), Data mining et statistique (avec discussion), *Journal de la SFdS*, Vol. **142**, n° 1, 5-35.

BUTTREY S. E. (1998), Nearest-neighbor classification with categorical variables, *Comput. Stat. & Data Analysis*, **28**,157-169.

CARLIER A. (1994), Méthodes exploratoires, In «*Analyse discriminante sur variables qualitatives*», Eds. G. Celeux et J.P. Nakache, Polytechnica, Paris.

CELEUX G. et NAKACHE J. P. (1994), *Analyse discriminante sur variables qualitatives*, Polytechnica, Paris.

CELEUX G. (2003), Analyse discriminante. Ch. 7 in «*Analyse des données*», Ed. G. Govaert, Hermès, Paris.

CUADRAS C. M. (1989), Distance analysis in discrimination and classification using both continous and categorical variables, In *Statistical data analysis and inference*, Ed. Y. Dodge, Amsterdam : North Holland, 459-473.

CUADRAS C. M. et FORTIANA J (1997), Probability densities from distances and discrimination, *Statistics & Probabilty Letters*, **33**, Issue 4, 405-411.

FIX E. et HODGES J. (1951), Discriminatory analysis-nonparametric discrimination : consistency properties, Technical Report 21-49004, 4, US Air Force, School of Aviation Medecine, Randolph Field, Texas.

FRIEDMAN J. H. et MEULMAN J. J. (2002), Clustering objects on subsets of attributes, *Préprint.* .

GOWER J. C. (1971), A general coefficient of similarity and some of its properties, *Biometrics*, **7**, $857 - 871$.

HALL P. (1981), Optimal near-neighbour estimator for use in discriminant analysis, *Biometrica*, **68**, 572-575.

HAND D. J. (2001), Modelling consumer credit risk, *IMA Journal of Management Mathematics*, **12**, 139-155.

HAND D. J. et HENLEY W. E. (1997), Statistical classification methods in consumer credit scoring : a review, *J. Roy. Statist. Soc.*, Series A, **160**, 523-541.

HENLEY W. E. et HAND D. J. (1996), A $k-$nearest neighbor classifier for assessing consumer credit risk, *Statistician*, **45**, 77-95.

HOLMES C. C. et ADAMS N. M. (2002), A probabilistic nearest-neighbor method for statistical pattern recognition, *J. Roy. Statist. Soc.*, B **64**, 295-306.

KOMROÁD, K. (2003), *On credit scoring estimation*, Master thesis, Homboldt Universität Berlin.

KRZANOWSKI W.J. (1975), Discrimination and classification using both binary and continuous variables, *Journal of the American Statistical Association*, **70**, 782-790.

KRZANOWSKI W. J. (1993), The location model for mixtures of categorical and continuous variables, *Journal of Classification,* **10**, 25-49.

MATUSITA K. (1956), Decision rule, based on the distance, for the classification problem, *Annals of Mathematical Statistics,* **8**, 67-77.

MCLACHLAN G. J. (1992), *Discriminant analysis and statistical pattern recognition*, New York : Wiley.

MKHADRI (1991), Discrimination binaire non-paramétrique : méthodes d'estimation du paramètre de lissage, *Revue de Statistique Appliquée*, **39**, n° 3, 37-55.

MÜLLER M. et RÖNZ B. (1999), Semiparametric Credit scoring, In «*Measuring risk in complex statistical systems*», J. Franke, W. Härdle, G. Stahl (eds.), Springer Verlag.

MÜLLER M. et HÄRDLE W. (2002), Exploring credit data. In «*Credit risk-measurement, evaluation and managementg*», G. Bol, G. Nakhaeizadeh, S.T. Rachev, T. Ridder, K.-H. Vollmer, (eds.), Proceedings Ökonometrie-Workshop 2002 : Kreditrisiko – Messung, Bewertung und Management, University of Karlsruhe, Physica-Verlag.

NAKACHE J.-P., LORENTE P., BENZECRI J. P. et CHASTANG J. F. (1977), Aspects pronostiques et thérapeutiques de l'infarctus myocardique aigu compliqué d'une défaillance sévère de la pompe cardiaque, Application des méthodes de discrimination, *Cahiers d'Anal. des Données*, II n° 4, 415-434.

RAO C. R. (1982), Diversity and dissimilarity coefficients : a unified approach, *Theoretical Population Biology,* **21**, 24-43.

RIPLEY B. (1996), *Pattern recognition and neural network*, Cambridge University Press, Cambridge.

SAPORTA G. (1977), Une méthode et un programme d'analyse discriminante sur variables qualitatives. In «*Analyse des données et informatique*», Ed. E. Diday, INRIA, pp. 201-210.