REVUE DE STATISTIQUE APPLIQUÉE

A. MKHADRI

Discrimination binaire non paramétrique. Méthodes d'estimation du paramètre de lissage

Revue de statistique appliquée, tome 39, nº 3 (1991), p. 37-55

<http://www.numdam.org/item?id=RSA_1991__39_3_37_0>

© Société française de statistique, 1991, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

Rev. Statistique Appliquée, 1991, **XXXIX** (3), 37-55

DISCRIMINATION BINAIRE NON PARAMÉTRIQUE MÉTHODES D'ESTIMATION DU PARAMÈTRE DE LISSAGE

A. MKHADRI

INRIA, Domaine de Voluceau-Rocquencourt
BP 105, 78153 Le Chesnay Cedex-France

RÉSUMÉ

La méthode des noyaux pour l'estimation non paramétrique des probabilités multinomiales, proposée par Aitchison & Aitken (1976), dépend fortement d'un paramètre de lissage λ. Les techniques d'estimation de la densité fondées sur la pseudo-vraisemblance et les fonctions de perte quadratiques sont présentées. Dans ce cadre, nous montrons comment utiliser les techniques de rééchantillonnage (validation croisée et bootstrap) pour estimer explicitement le paramètre de lissage λ. Si l'intérêt principal n'est pas l'estimation de la densité mais la discrimination, d'autres méthodes de choix de λ peuvent donner de meilleures performances pour la séparation des groupes. Les méthodes de ce type ont été considérées récemment par Tutz (1986, 1989) et Hall & Wand (1988). Dans le même cadre, nous proposons une méthode fondée sur la minimisation du taux d'erreur, qui nous fournit explicitement λ sans avoir recours à un algorithme d'optimisation. De plus, on étend aussi la technique du bootstrap à la méthode de Hall & Wand. Une application pratique est présentée pour illustrer le comportement de ces techniques.

Mots-clés : *Estimation de la densité; discrimination; méthode des noyaux; paramètres de lissage; rééchantillonnage; taux d'erreur.*

ABSTRACT

The kernel method for estimating the cell probabilities of multivariate discrete distribution, due to Aitchison & Aitken (1976), depends crucially on an unknown smoothing parameter λ. Most of the methods for choosing the smoothing parameter are discussed in the context of density estimation . The choice may be based on a pseudo-likelihood or on loss functions for the estimation of density. In this setting, we show how to use resampling methods (cross-validation and bootstrap) to estimating the smoothing parameters. If the main interest is not in density estimation but in discrimination, alternative methods for choosing λ from the discrimination viewpoint may yield better performance for separation of groups. Methods of this type have been proposed in Tutz (1986, 1989) for discrete kernels and more recently in Hall & Wand (1988). In the same setting, we propose a method, estimating λ explicitly, based on minimization of the leaving-one-out estimator of the error rate, without using iterative method. Moreover, we extend the method of bootstrap to Hall & Wand

approach's, in the case of two groups. An exemple is given to illustre the pratical behaviour of all these methods.

Key-words *: Density estimation; discrimination; kernel method; smoothing parameters; resampling; Leaving-one-out method.*

Introduction

Aitchison & Aitken (1976) ont proposé une méthode des noyaux, pour l'estimation non paramétrique des probabilités multinomiales pour des données qualitatives. Leur estimateur, construit sur la base d'un échantillon d'apprentissage $E = \{x_1, ..., x_n\}$, avec $x_i \in \{0,1\}^p$, p étant le nombre de variables et $i = 1, ..., n$, peut s'écrire

$$\widehat{f}(x|\lambda, E) = n^{-1} \sum_{y \in E} K(x|y, \lambda),$$

où $K(.|E, \lambda)$ est une fonction appelée noyau et $\lambda \in [1/2, 1]$ un paramètre de lissage. La fonction noyau peut prendre plusieurs formes (Aitken 1983, Habbema & al. 1978 et Titterington 1980; pour les variables ordonnées voir aussi Titterington & Bowman 1985). Rappelons ici la forme très utilisée de Aitchison & Aitken

$$\widehat{f}(x|\lambda, E) = n^{-1} \sum_{y \in E} \lambda^{p-d(x,y)} (1-\lambda)^{d(x,y)} \tag{1}$$

avec $d(x, y) = \sum_{j=1}^{p} |x^j - y^j| (x, y \in \{0,1\}^p)$.

Tous ces estimateurs sont très sensibles au paramètre de lissage, tandis que le type de la fonction noyau a très peu d'importance. Ainsi, le paramètre de lissage devra être choisi avec prudence en s'appuyant sur des considérations pratiques. Le choix du paramètre de lissage peut être basé sur une pseudo-vraisemblance (*i.e.* vraisemblance par validation croisée, Aitchison & Aitken 1976) ou sur des fonctions de perte pour l'estimation de la densité. Cette approche traditionnelle, dans laquelle les paramètres sont déterminés séparément pour chaque groupe, sans tenir compte de la discrimination, a été beaucoup étudiée par plusieurs auteurs (Bowman 1980, Titterington 1980, Hall 1981a, Brown et Rundell 1985, Hand 1982 et Bowman et al. 1984).

Si l'intérêt principal n'est pas l'estimation de la densité mais la discrimination, des méthodes alternatives du choix de λ à partir du point de vue de la discrimination peuvent donner de meilleures performances pour la séparation des groupes. Les méthodes de ce type ont été utilisées par Van Ness & Simpson (1976) et Van Ness (1979) pour les noyaux de Parzen, et plus récemment par Tutz (1986, 1989) et Hall & Wand (1988) pour les noyaux discrets.

Le but de cet article est de discuter certaines de ces méthodes récentes et de montrer comment utiliser des techniques de rééchantillonnage (validation croisée et bootstrap) pour estimer le (ou les) paramètre (s) de lissage λ (ou $\lambda_1, ..., \lambda_K$). Les

méthodes d'estimation fondées sur ces techniques nous fournissent des paramètres optimaux, en un certain sens, de lissage d'une manière explicite et ils sont simples à calculer. De plus, on montrera comment éviter les problèmes numériques des algorithmes d'optimisation, utilisés pour optimiser le taux d'erreur, pour obtenir explicitement λ.

Nous concentrons notre attention sur les données binaires multivariées. Après une revue, au paragraphe 1, des techniques d'estimation de la densité fondées sur le maximum de vraisemblance, les fonctions de perte quadratique et d'information de Kullback-Leibler, nous proposons, dans le même cadre, des variantes fondées sur la validation croisée et le bootstrap. On présente au paragraphe 2 les résultats des méthodes d'estimation liées directement à la discrimination (plus précisément celles de Tutz 1986 et Hall & Wand 1988). De même, on propose une nouvelle méthode, concurrente de celle de Tutz, qui nous fournit explicitement le paramètre optimal de lissage sans avoir recours à un algorithme d'optimisation. De plus, on étend aussi la méthode de bootstrap du paragraphe 1 à la méthode de Hall & Wand. L'application de ces différentes méthodes à un exemple est présentée au paragraphe 3.

1. Procédures de lissage pour l'estimation de la densité

Dans cette section $\widehat{f}(.,\lambda)$ (ou $\widehat{f}(.|E,\lambda)$) représente l'estimateur de la densité par la méthode des noyaux défini sur l'échantillon E représentant un seul groupe a priori.

1.1 Propriétés de la méthode du maximum de vraisemblance

La stratégie de maximisation de la vraisemblance

$$\begin{aligned} V(\lambda|E) &= \prod_{i=1}^{n} \widehat{f}(x_i|E,\lambda) \\ &= \prod_{x\in B^p} \Big\{ \sum_{z\in B^p} (n_z/n) K(x|z,\lambda) \Big\}^{n_x}, \end{aligned} \tag{2}$$

avec $K(x|z,\lambda) = \lambda^{p-d(x,z)}(1-\lambda)^{d(x,z)}$ où $B^p = \{0,1\}^p$ et n_x est la fréquence des observations x dans l'échantillon E, conduit à λ égal à 1 (cf. Hand 1982).

Maintenant la vraisemblance par validation croisée, ou pseudo-vraisemblance, s'écrit

$$\begin{aligned} W(\lambda,E) &= \prod_{i=1}^{n} \widehat{f}(x_i|E-\{x_i\},\lambda) \\ &= \prod_{x\in B^p} \Big\{ \sum_{z\in B^p} [n_z/(n-1)] K(x|z,\lambda) - (\lambda^p/(n-1)) \Big\}^{n_x} \end{aligned} \tag{3}$$

D'après Bowman (1980), choisir λ qui maximise W est équivalent à choisir λ qui maximise

$$n^{-1}\sum_{i=1}^{n} \text{Log}\widehat{f}(x_i, E-\{x_i\}, \lambda)$$

et par conséquent, c'est aussi équivalent à la minimisation de

$$n^{-1}\sum_{i=1}^{n} L\{\delta_{x_i}(x), \widehat{f}(x_i, E-\{x_i\}, \lambda)\},$$

avec $\delta_{x_i}(x) = 1$ si $x = x_i$ et 0 sinon, L étant la fonction de perte Kullback-Leibler définie par $L\{p, q\} = \Sigma_x p(x)\text{Log}\{p(x)/q(x)\}$.

Proposition 1 (Bowman 1980) : *Soit λ_n la valeur maximisant W, alors pour n assez grand, $L(f(.), \widehat{f}(.|\lambda_n))$ converge en probabilité vers $0, f(.)$ étant la densité optimale inconnue.*

1.2 Méthode de Hall et critique de la validation croisée

Hall (1981a) montre que la maximisation de W peut amener à un maximum local en $\lambda = 1$. Pour éviter ce problème, il propose de maximiser l'un des critères

$$J_1(\lambda) = E\Sigma_x\{f(x) - \widehat{f}(x)\}^2$$

ou (4)

$$J_2(\lambda) = E\Sigma_x w(x)\{f(x) - \widehat{f}(x)\}^2,$$

où w est une fonction de poids. Il montre, en utilisant les développements de Taylor de $\widehat{f}$ autour de $(1-\lambda)$, que

$$E(\widehat{f}(x|\lambda)) = f(x) + (1-\lambda)(f_1(x) - pf(x)) + O\{(1-\lambda)^2\} \tag{5}$$

et

$$n\text{var}(\widehat{f}(x|\lambda)) = f(x)(1-f(x)) - 2(1-\lambda)f(x)\{p(1-f(x)) + f_1(x)\} + O\{(1-\lambda)^2\}, \tag{6}$$

où $f_1(x) = \sum_{y:d(x,y)=1} f(y)$. En utilisant ces formules dans la relation

$$E\{f(x) - \widehat{f}(x)\}^2 = \text{var}(\widehat{f}(x|\lambda)) + \{E(f(x|\lambda)) - f(x)\}^2,$$

il en déduit, par approximations jusqu'à l'ordre un, que $\widehat{\lambda}_{1H}$ et respectivement $\widehat{\lambda}_{2H}$ qui minimisent J_1 et respectivement J_2 (en prenant comme fonction de poids

$w(x) = f(x)$ dans J_2) sont définis par

$$\widehat{\lambda}_{1H} = 1 - \frac{p - \Sigma_x f(x)\{f_1(x) - pf(x)\}}{n\Sigma_x\{f_1(x) - pf(x)\}^2} \tag{7a}$$

$$\widehat{\lambda}_{2H} = 1 - \frac{p - \Sigma_x (f(x))^2\{f_1(x) + p(1 - f(x))\}}{n\Sigma_x f(x)\{f_1(x) - pf(x)\}^2}. \tag{7b}$$

Ces formules peuvent être résolues, soit directement en remplaçant $f(x)$ par $N_0(x)/n$ et $f_1(x)$ par $N_1(x)/n$ (cas pratique ; où $N_v(x) = \text{Card}\{b \in E | d(x,b) = v\}$, $v = 0, 1$), soit par une procédure itérative (en regardant cette formule comme une fonction implicite)

$$\widehat{\lambda}_{iH} = g\{\widehat{f}(x|\widehat{\lambda}_{i-1})\}, \ i = 1, 2.$$

$\widehat{\lambda}_0 = 1$ peut être utilisé comme valeur initiale. Hall a considéré le critère J_1 en poussant les développements jusqu'à l'ordre 2 dans les expressions (5) et (6). Ainsi, il obtient l'approximation

$$\widehat{\lambda}'_{1H} = 1 - \frac{p + \Sigma_x f(x)\{f_1(x) - pf(x)\}}{n\Sigma_x f(x)\{2f_2(x) + p(p+1)f(x) - 2pf_1(x)\}}. \tag{8}$$

Il note aussi que la convergence de $\widehat{\lambda}_{iH} (i = 1, 2)$ vers 1 avec une vitesse d'ordre n^{-1} est une condition nécessaire et suffisante pour que l'estimateur associé $\widehat{f}$ soit convergent.

Dans un autre article, Hall (1981b) a considéré une autre forme, plus générale, pour l'estimateur (1) qu'on peut écrire sous forme d'une combinaison linéaire de $N_v(x) (v = 1, ..., r)$, où, pour tout $x \in B^p$, on a

$$\widehat{f}(x, \omega) = n^{-1} \sum_{v=0}^{r} \omega_v N_v(x) \tag{9}$$

avec $\omega = (\omega_0, ..., \omega_r)^t$ un vecteur poids à choisir et $1 \leq r \leq p$.

Remarque 1 : Le cas, $\omega_v = 1$ pour $v \leq r$ et $\omega_j = 0$ pour tout $j > r$, correspond à l'estimateur des r plus proches voisins de Hills (1967). De même, l'estimateur (1) revient à prendre dans (9) $\omega_v = \lambda^p\{(1 - \lambda)/\lambda\}^v$ pour tout v. L'estimateur du modèle multinomial complet correspond au cas où $\omega_0 = 1$ et $\omega_v = 0$ pour tout $v = 1, ..., r$.

On montre facilement (cf. Hall 1981b) que $\Sigma_x N_0(x) = n$ et $\Sigma_x N_1(x) = np$, avec $x \in B^p$.

Théorème 1 (Hall 1981b) : *Le vecteur poids optimal $\omega_{opt} = (\omega_0, ..., \omega_r)^t$ qui minimise J_1 est défini par :*

$$\omega_{opt} = \{Q + n^{-1}(D - Q)\}^{-1} g_0 \tag{10}$$

où Q et D sont des matrices carrées de taille $r+1$ et g_0 un vecteur à $(r+1)$ composantes, tels que

$$g_0 = \Sigma_x f(x)s(x)$$

$$s(x) = (f_0(x), f_1(x), ..., f_r(x))^t, \text{ avec } f_0(x) = f(x) \text{ et } f_v(x) = \sum_{y:d(x,y)=v} f(y)$$

$$Q = \Sigma_x s(x)s^t(x) \text{ et } D = diag\{\tbinom{r}{0}, \tbinom{r}{1}, ..., \tbinom{r}{r}\}.$$

(10) peut s'écrire aussi

$$\omega_{opt} = \{(1-n^{-1})I + n^{-1}Q^{-1}D\}^{-1}\underline{i} = (1+n^{-1})\underline{i} - n^{-1}Q^{-1}D\underline{i} + O(n^{-2})$$

avec $\underline{i} = (1, 0, ..., 0)^t$. Les $f_v(x)$ seront remplacés par leurs estimateurs du maximum de vraisemblance $N_v(x)/n$ qui mènent à $\widehat{Q}$ et ainsi à l'estimateur

$$\widehat{\omega}_{op} = \{(1-n^{-1})I + n^{-1}\widehat{Q}^{-1}D\}^{-1}\underline{i} \approx (1+n^{-1})\underline{i} - n^{-1}\widehat{Q}^{-1}D\underline{i} \qquad (11)$$

où I est la matrice identité d'ordre $r+1$.

On remarque que la somme de ces estimations ne peut pas être égale à 1. Ainsi, Hall modifie ces poids, de telle sorte que leur somme soit égale à 1, en minimisant J_1 sous la contrainte $\omega^t\underline{h} = 1$, où $\underline{h} = (\binom{r}{0}, \binom{r}{1}, ..., \binom{r}{r})^t$. Il obtient dans le cas où $r = p$, en utilisant les multiplicateurs de Lagrange, le résultat suivant.

Corollaire 1 (Hall 1981b, 1983) : *Le vecteur optimal $\widehat{\omega}_1$ minimisant J_1 sous la contrainte $\omega^t\underline{h} = 1$, où $h = (\binom{p}{0}, \binom{p}{1}, ..., \binom{p}{p})^t$, est*

$$\widehat{\omega}_1 = \widehat{\omega}_{op} + \widehat{A}^{-1}\,\underline{u}(1 - \underline{h}^t\widehat{A}^{-1}\underline{i})/(\underline{h}^t\widehat{A}^{-1}\underline{u}) \qquad (12)$$

où $\widehat{A} = (1-n^{-1})I + n^{-1}\widehat{Q}^{-1}D$ et $\underline{u}$ est le vecteur unité à $(p+1)$ composantes.

Les estimations peuvent être négatives. Ce n'est pas forcément un désavantage dans le cadre de la discrimination vu que les comparaisons des scores et donc les classifications peuvent être réalisées.

On note que les résultats précédents nécessitent clairement que la matrice Q soit inversible. Hall a commenté brièvement le cas où Q était singulière. Il note cependant que c'est une occurrence très improbable.

1.3 Nouvelles procédures

Dans cette section, nous proposons différentes procédures d'estimation explicite du paramètre de lissage basées sur les techniques de rééchantillonnage (validation croisée et bootstrap).

1.3.1 Une procédure de validation croisée

En général, le paramètre de lissage λ qui minimise J_1 ou J_2 dépend de la densité inconnue f. En pratique, les données doivent être utilisées pour "estimer" λ. La validation croisée est un moyen de construction de l'estimateur basé sur les données. Pour cela, observons que

$$\Sigma_x E\{f(x) - \widehat{f}(x)\}^2 - \Sigma_x f^2(x) = \Sigma_x E\{\widehat{f}^2(x)\} - 2\Sigma_x E\{\widehat{f}(x)\}f(x),$$

donc choisir λ pour minimiser le membre de droite est équivalent à choisir λ pour minimiser J_1. Ainsi, l'estimateur

$$\begin{aligned} Cval(\lambda) &= \Sigma_x \widehat{f}^2(x, E, \lambda) - \frac{2}{n}\sum_{j=1}^{n} \widehat{f}(x_j|E - \{x_j\}, \lambda) \\ &= \Sigma_x \widehat{f}^2(x, E, \lambda) - \frac{2}{n}\Sigma_x N_0(x)\widehat{f}(x|E - \{x\}, \lambda) \end{aligned} \tag{13}$$

est un estimateur sans biais du membre de droite de l'expression précédente (cf. Rudemo 1982, Bowman 1984 et Hall & Titterington 1987). On a le résultat suivant.

Proposition 2 : *La valeur optimale de λ minimisant (13) est approximée par*

$$\lambda_{val} = 1 - \frac{\Sigma_x N_0(x)N_1(x)}{(n-1)\Sigma_x N_1^2(x)}. \tag{14}$$

Preuve : en effet, en utilisant un développement de Taylor d'ordre deux autour de $(1-\lambda)$ dans (1), on a

$$\begin{aligned} &\widehat{f}(z|E, \lambda) = n^{-1}\{N_0(z) + (1-\lambda)N_1(z) + O(1-\lambda)^2\}, \\ &\widehat{f}(z|E - \{z\}, \lambda) = (n-1)^{-1}\{N_0^*(z) + (1-\lambda)N_1(z) + O(1-\lambda)^2\}, \\ &\widehat{f}^2(z|E, \lambda) = n^{-2}\{N_0^2(z) + 2N_0(z)N_1(z)(1-\lambda) + N_1^2(z)(1-\lambda)^2 + O(1-\lambda)^2\}, \end{aligned}$$

où $N_0^*(z) = N_0(z) - 1$.

Donc $Cval(\lambda)$ s'écrit, en négligeant les termes d'ordre $(1-\lambda)^2/(n-1)$ dans $\widehat{f}(z|E - \{z\}, \lambda)$, approximativement

$$\begin{aligned} Cval(\lambda) \approx\ & \Sigma_z n^{-2} N_0^2(z) - \frac{2}{n(n-1)} N_0^*(z)N_0(z) \\ & + 2(1-\lambda)[n^{-2}N_0(z)N_1(z) - \{n(n-1)\}^{-1}N_0(z)N_1(z)] \\ & + (1-\lambda)^2 n^{-2}\Sigma_z N_1^2(z) \end{aligned}$$

D'où, en posant $h = 1 - \lambda$, et en dérivant $Cval(h)$ par rapport à h, on obtient

$$\begin{aligned} h &= \frac{-\Sigma_z[n^{-2}N_0(z)N_1(z) - \{n(n-1)\}^{-1}N_0(z)N_1(z)]}{\Sigma_z n^{-2}N_1^2(z)} \\ &= \frac{\Sigma_z N_0(z)N_1(z)}{(n-1)\Sigma_z N_1^2(z)}. \end{aligned}$$

Ainsi, quand n devient grand, λ_{val} converge vers 1 (voir Mkhadri 1990b) avec une vitesse d'ordre n^{-1} qui est une condition nécessaire et suffisante pour la convergence de la méthode d'estimation (cf. Hall 1981a).

Remarque 2 : La même procédure peut être appliquée à l'estimateur des plus proches voisins d'ordre r défini par l'expression (9). On en déduit (cf. Mkhadri 1990a) que le poids optimal ω_{val} minimisant $Cval(\omega)$ est défini par :

$$\omega_{val} = n(n-1)^{-1}\mathbf{N}^{-1}P_0,$$

où $\mathbf{N} = \Sigma_x N(x)N(x)^t$ est une matrice carrée de taille $r+1$, et $P_0 = \Sigma_x N_0(x)N^*(x)$ est un vecteur à $r+1$ composantes, où $N(x) = (N_0(x), ..., N_r(x))^t$ et $N^*(x) = (N_0(x) - 1, .N_1(x), ..., N_r(x))^t$. Cette dernière expression de ω_{val} est valable uniquement lorsque $\mathbf{N}$ est inversible. L'avantage de cette expression par rapport à celle obtenue en (10) est qu'elle ne dépend pas de la densité optimale inconnue f.

1.3.2 Procédure basée sur le Bootstrap

C'est une procédure dont l'objectif est similaire à celle basée sur la validation croisée. Elle a été considérée par Taylor (1989), dans le cas continu, en utilisant les noyaux gaussiens. Il étudie le comportement du critère de la moyenne des écarts quadratiques (noté MEQ) basé sur l'échantillon bootstrap lissé $E^* = \{x_1^*, ..., x_m^*\}$ tiré aléatoirement de la loi de $\widehat{f}$. On définit $\widehat{f}^*$ sur E^* de la même manière que $\widehat{f}$ sur E. Ainsi, il arrive à exprimer le critère uniquement en fonction des données initiales sans rééchantillonnage.

On adapte ici cette procédure aux données binaires dans le but d'obtenir le paramètre de lissage optimal dépendant uniquement des données. Nous montrons que cette méthode fournit explicitement un paramètre de lissage optimal.

On suppose que E^* est tiré suivant $\widehat{f}(.|n, \lambda)$, et a la même taille n que E. Nous obtenons la propriété suivante :

Proposition 3 : *La valeur de $h = (1-\lambda)$ minimisant $\widehat{SMEQ}(h) = \Sigma_x \widehat{MEQ}(x, h)$ est approximativement (où $\widehat{MEQ}(x, h) = E\{\widehat{f}^*(x|n, \lambda) - \widehat{f}(x|n, \lambda)|E\}^2$)*

$$h_B \approx \frac{n^{-3}N(n)}{2n^{-3}D_1(n) - 2n^{-2}D_2(n)}, \tag{15}$$

avec

$$N(n) = \Sigma_x\{2pN_0^2(x) - 2N_0(x)N_{01}(x) - 2N_0(x)N_1(x)\} - pn^2,$$
$$D_1(n) = \Sigma_x 2[N_1(x)N_{01}(x)+N_0(x)N_{11}(x)-2pN_0(x)N_1(x)]+\Sigma_x N_1^2(x)+2(pn)^2,$$
$$D_2(n) = \Sigma_x(N_{01}(x) - pN_0(x))^2,$$
$$N_{v1}(x) = \sum_{y:d(x,y)=1} N_v(y), v = 0, 1.$$

Preuve : suivant le développement de Taylor d'ordre deux, on a

$$\widehat{f}^*(x|n,h) = n^{-1}\{M_0(x) + hM_1(x) + O(h^2)\}.$$
$$\widehat{f}(x|n,h) = n^{-1}\{N_0(x) + hN_1(x) + O(h^2)\}$$
$$\widehat{f}_1(x|n,h) = \sum_{y:d(x,y)=1} \widehat{f}(y|n,h) = n^{-1}\{N_{01}(x) + hN_{11}(x) + O(h^2)\}.$$

où $N_v(x) = \text{Card } \{b \in E|d(x,b) = v\}$, $M_v(x) = \text{ Card } \{b \in E^*|d(x,b) = v\}$, et $N_{v1}(x) = \sum_{y:d(x,y)=1} N_v(y), v = 0, 1.$

D'après la formule (5), l'espérance de $\widehat{f}^*(x|n,h)(h = 1 - \lambda)$ s'écrit

$$E\{\widehat{f}^*(x|n,h)\} = \widehat{f}(x|n,h) + h(\widehat{f}_1(x|n,h) - p\widehat{f}(x|n,h)) + O(h^2)$$

et sa variance se déduit de la même manière de (6) par

$$nvar(\widehat{f}^*(x|n,h)) = \widehat{f}(x|n,h)\{1 - \widehat{f}(x|n,h)\} - 2h\widehat{f}(x|n,h)$$
$$\{p(1 - \widehat{f}(x|n,h)) + \widehat{f}_1(x|n,h)\} + O(h^2).$$

Or, on sait que $E\{f(x) - \widehat{f}(x)\}^2 = \text{ var}(\widehat{f}^*(x|n,h)) + \{E(\widehat{f}^*(x|n,\lambda)) - \widehat{f}(x|n,\lambda)\}^2$.

Ainsi, l'estimation bootstrap de $\widehat{SMEQ}(h)$ s'écrit approximativement

$$\begin{aligned}\widehat{SMEQ}(h) &= \Sigma_x E\{\widehat{f}^*(x|n,\lambda) - \widehat{f}(x|n,\lambda)|E\}^2 \\ &\approx \Sigma_x n^{-3}\{N_0(x)(n - N_0(x)) + h[nN_1(x) - 2N_0(x)N_1(x)] \\ &\qquad - h[2pN_0(x)\{n - N_0(x)\} + 2N_0(x)N_{01}(x)] \\ &\qquad - h^2 2[N_1(x)N_{01}(x) + N_0(x)N_{11}(x)] - N_1^2(x)h^2 \\ &\qquad - h^2 2p[nN_1(x) - 2N_0(x)N_1(x)]\} \\ &\quad + \Sigma_x n^{-2}h^2(N_{01}(x) - pN_0(x))^2.\end{aligned}$$

D'où, en dérivant cette expression par rapport à h et en égalant à 0, on en déduit

$$h \approx \frac{n^{-3}\mathbf{N}(n)}{2n^{-3}D_1(n) - 2n^{-2}D_2(n)}.$$

Remarque 3 : Dans un cadre plus général, et indépendamment de Taylor, Hall (1990) propose une méthode d'utilisation du bootstrap classique. On montre (cf. Mkhadri 1990a) que cette procédure nous fournit un paramètre de lissage qui peut être supérieur à 1. De même (cf. §3, tableau 1), la procédure précédente peut amener à un paramètre de lissage supérieur à 1 ; auquel cas on impose un paramètre de lissage égal à 1.

2. Méthodes d'estimation liées a la discrimination

Jusqu'à présent, on s'est intéressé aux méthodes liées directement à l'estimation de la densité, en utilisant des fonctions de perte appropriées. Or, si l'intérêt principal n'est pas l'estimation de la densité mais la discrimination, il peut être très utile de prendre celle-ci en considération. On présente deux méthodes qui sont liées à la discrimination ; on montre comment on peut trouver explicitement le paramètre de lissage pour la première méthode, fondée sur le taux d'erreur, et on propose une nouvelle technique pour estimer explicitement le paramètre de lissage pour la deuxième méthode, basée sur le bootstrap comme dans le paragraphe précédent.

2.1 Méthode basée sur le taux d'erreur

Supposons qu'on ait plusieurs classes $E_1, ..., E_K$, associées à un mélange dans une population de grande taille ; les probabilités a priori des classes sont notées $\delta_1, ..., \delta_K$. Soit $p(x|k)$ la probabilité d'avoir le vecteur $x^t = (x_1, ..., x_p)$ dans la classe k ; $E_k = \{x_1^{(k)}, ..., x_{n_k}^{(k)}\}$ est un échantillon aléatoire de n_k observations issu de la loi $p(x|k)$. Soit $\widehat{D}_. = (\widehat{D}_1, ..., \widehat{D}_K)$ la règle de classification générée par l'utilisation des estimateurs de densité basés sur les noyaux définis en (1), où $x \in \widehat{D}_k$ signifie que

$$\delta_k \widehat{f}(x|E_k, \lambda_k) = \max_i \delta_i \widehat{f}(x|E_i, \lambda_i), i = 1, ..., K,$$

avec $\widehat{f}(x|E_k, \lambda_k) = n_k^{-1} \sum_{y \in E_k} \lambda_k^{p-d(x,y)}(1 - \lambda_k)^{d(x,y)}$ et $\lambda_k \in [1/2, 1]$ $(k = 1, ..., K)$.

Posons $\lambda^t = (\lambda_1, ..., \lambda_K) \in [1/2, 1]^K$; alors le vecteur ligne des scores discriminants au point x est

$$\widehat{f}(x|E, \lambda) = \{\delta_1 \widehat{f}(x|E_1, \lambda_1), ..., \delta_K \widehat{f}(x|E_K, \lambda_K)\}.$$

Soit $p_0(x) = \{\delta_1 p(x|1), ..., \delta_K p(x|K)\}$, le vecteur des scores discriminants optimaux et soit $\widehat{p}(x|k)$ la valeur de la fréquence relative de x dans l'échantillon E_k. Alors l'estimateur t_L du taux de bon classement par validation croisée (leaving-one-out) peut s'écrire (Tutz 1986)

$$t_L(\lambda) = \sum_{k=1}^{K} \delta_k \Sigma_x \widehat{p}(x|k) I_k \tag{16}$$

où $I_k = I_k\{\delta_1 \widehat{f}(x|E_1, \lambda_1), ..., \delta_k \widehat{f}(x|E_k - \{x\}, \lambda_k), ..., \delta_K \widehat{f}(x|E_K, \lambda_K)\}$ (si $x \in E_k$) est défini par :

$$I_k(y_1, ..., y_K) = \begin{cases} 1 & \text{si } y_k > y_i \text{ pour tout } i \neq k \\ 1/r & \text{si } y_k = y_i, i \in \{i_1, ..., i_r\}, y_k > y_i, i \notin \{i_1, ..., i_r\} \\ 0 & \text{sinon} \end{cases}$$

Comme Stone (1977) le constatait, la technique de validation croisée peut ne pas être convergente. Mais Tutz montre que la maximisation de $t_L(\lambda)$ fournit une procédure d'estimation qui est fortement convergente.

Théorème 2 (Tutz 1986) : *Soit* $\lambda^* = (\lambda_1^*, ..., \lambda_K^*) \in [1/2, 1]^K$ *choisi par*

$$t_L(\lambda^*) = \max_{\lambda} t_L(\lambda)$$

Alors

i) $t_L(\lambda^*)$ *converge presque sûrement vers le taux de bon classement optimal* $r(D^*)$.

ii) La règle de classification qui en résulte est fortement convergente au sens de Bayes (i.e. le taux de bon classement de la règle de classification converge presque sûrement vers le taux de bon classement optimal.)

Remarques 4 : La convergence résultant du théorème s'applique aussi dans le cas où, à la place des K échantillons séparés, on utilise un seul échantillon issu d'un mélange ; mais alors les estimateurs des probabilités a priori seront estimés par leurs fréquences relatives (i.e. $\widehat{\delta}_k = n_k/n, k = 1, ..., K$). De même, on obtient le vecteur de paramètres de lissage trivial $1_K = (1, ..., 1)^t$ si, dans la procédure de maximisation, on a utilisé, à la place de la validation croisée, la méthode de resubstitution.

Les problèmes pratiques se posent quand $t_L(\lambda)$ doit être maximisé numériquement : $t_L(\lambda)$ est une fonction discontinue. Pour éviter ces problèmes, une version lissée de $t_L(\lambda)$ peut être utilisée. Elle a été proposée par Glick (1978) dans le cas de deux classes, puis généralisée par Tutz (1985) au cas de K classes. Tutz (1988) montre que le théorème 2 reste vrai lorsque l'on utilise cette version lissée de $t_L(\lambda)$.

Nous proposons une nouvelle approche, plus simple, qui consiste à imposer un seul paramètre de lissage pour toutes les classes ($\lambda = \lambda_1 = ... = \lambda_K$). Des comparaisons pratiques ont en effet montré que d'imposer l'égalité des paramètres de lissage ne détériorait pas les taux de reconnaissance, bien au contaire (cf. Hand 1982 p. 162-164, Tutz 1986). D'après (1), l'estimateur des noyaux pour la classe E_k peut s'écrire

$$\begin{aligned}\widehat{f}(x|E_k,\lambda) &= (\lambda^p/n_k)\sum_{j=1}^{p} N_{jk}(x)\{(1-\lambda)/\lambda\}^j \\ &= (1/n_k(1+\gamma)^p)\sum_{j=1}^{p} N_{jk}(x)\gamma^j\end{aligned}$$

où l'on a posé $\gamma = (1-\lambda)/\lambda$, $\gamma \in [0,1]$ et $N_{jk}(x) = \text{Card}\{b \in E_k | d(x,b) = j\} j = 1,...,p$ et $k = 1,...,K$. Ainsi, l'estimateur par validation croisée pour la classe E_k peut s'écrire

$$\widehat{f}(x|E_k - \{x_i\},\gamma) = \frac{1}{n_k^{(i)}}[N_{0k}^{(i)}(x) + \gamma\{N_{1k}(x) - pN_{0k}^{(i)}(x)\} + O(\gamma^2)];$$

cette expression se réduit à

$$\widehat{f}(x|E_k - \{x_i\},\gamma) \approx (1-\gamma)M_k^{(i)}(x) + \gamma V_k^{(i)}(x), \tag{17a}$$

où

$$V_k^{(i)}(x) = \frac{N_{1k}(x) - (p-1)N_{0k}^{(i)}(x)}{n_k^{(i)}} \quad \text{et} \quad M_k^{(i)}(x) = \frac{N_{0k}^{(i)}(x)}{n_k^{(i)}},$$

avec

$$N_{0k}^{(i)}(x) = \begin{cases} N_{0k}(x) - 1 & \text{si} \quad x \in E_k \\ N_{0k}(x) & \text{sinon} \end{cases} \quad \text{et } n_k^{(i)} = \begin{cases} n_k - 1 & \text{si} \quad x \in E_k \\ n_k & \text{sinon} \end{cases},$$

Nous allons chercher le paramètre de lissage γ qui minimise l'estimateur $T^*(\gamma)$ du taux d'erreur par validation croisée, et qui s'écrit

$$T^*(\gamma) = \sum_{k=1}^{K} \delta_k \sum_{x \in E_k} \frac{1}{n_k}[1 - I_k\{\delta_1 \widehat{f}(x|E_1,\gamma),...,\delta_k \widehat{f}(x|E_k - \{x\},\gamma),...,\delta_K \widehat{f}(x|E_K,\gamma)\}].$$

Pour simplifier la présentation, nous allons détailler les calculs dans le cas de deux groupes E_1 et E_2. On a la propriété suivante :

Proposition 4 : *Le paramètre optimal de lissage est soit $\gamma = 0$, soit $\gamma = 1$, soit γ est de la forme*

$$\frac{\delta_1 M_1^{(i)}(x) - \delta_2 M_2^{(i)}(x)}{\delta_1\{M_1^{(i)}(x) - V_1^{(i)}(x)\} - \delta_2\{M_2^{(i)}(x) - V_2^{(i)}(x)\}} \tag{17b}$$

avec $x \in E$.

Preuve : à γ fixé, il est immédiat de montrer, d'après (17a), que la règle de décision par validation croisée s'écrit, pour tout $x_i (1 \leq i \leq n)$: x_i est affecté à E_1 si et seulement si $C(x_i, \gamma) \geq 0$, où l'on a posé

$$C(x_i, \gamma) = (1-\gamma)[\delta_1 M_1^{(i)}(x_i) - \delta_2 M_2^{(i)}(x_i)] + \gamma[\delta_1 V_1^{(i)}(x_i) - \delta_2 V_2^{(i)}(x_i)].$$

Ainsi, la classe à laquelle x_i est affecté change en fonction de γ, si et seulement si, il existe un γ_0 tel que $C(x_i, \gamma_0) = 0$ et $\gamma_0 \in [0, 1]$. Cela donne bien un γ_0 de la forme (17b). Si $C(x_i, \gamma)$ garde un signe constant sur [0,1], l'affectation de x_i sera indépendante de γ et sera, soit celle fournie par le modèle multinomial complet (si $\gamma = 0$), soit sera de type plus proches voisins (si $\gamma = 1$).

Remarque 5 : La proposition 3 montre que le γ optimal correspond à la valeur à laquelle la décision change d'affectation et, elle est à rechercher parmi 0,1 et les racines des équations $C(x_i, \gamma) = 0$ $(1 \leq i \leq n)$. En pratique, le nombre de ces racines est petit : il représente le nombre d'états où le modèle multinomial complet et le modèle de type plus proches voisins (définis par les $V_k^{(i)}(x)$) diffèrent. Ainsi, contrairement à Tutz, nous sommes en mesure de calculer explicitement le paramètre optimal de lissage par validation croisée. Le cas où le nombre de groupes K est supérieur à deux est analogue (cf. Mkhadri 1990a ch. 5, Celeux & Mkhadri 1991).

2.2 Méthode de Hall & Wand

Dans le cas de deux classes E_1 et E_2, la règle de classification se réduit à : une observation $z (z \in \{0,1\}^p)$ est affectée à la classe E_1 si : $\delta_1 f(x|E_1) \geq \delta_2 f(x|E_2)$, ce qui est équivalent à

$$g(z) = \delta_1 f(x|E_1) - \delta_2 f(x|E_2) \geq 0.$$

$f(x|E_1)$ et $f(x|E_2)$ étant inconnus, on les estime respectivement par $\widehat{f}(x|E_1, \lambda_1)$ et $\widehat{f}(x|E_2, \lambda_2)$ définis par (1). D'où la règle de classification : z est affecté à E_1 si

$$\widehat{g}(z) = \delta_1 \widehat{f}(x|E_1, \lambda_1) - \delta_2 \widehat{f}(x|E_2, \lambda_2) \geq 0.$$

Hall et Wand (1988) proposent de minimiser le critère $J_1(h_1, h_2)$ entre $\widehat{g}(z)$ et $g(z)$, avec $h_i = 1 - \lambda_i (i = 1, 2)$, $J_1(h_1, h_2) = E\Sigma_z\{g(z) - \widehat{g}(z)\}^2$. Ils montrent

que l'un des paramètres optimums peut être négatif. Ainsi, pour éviter ce problème, ils proposent une autre procédure fondée sur la validation croisée. On remarque que minimiser $J_1(h_1, h_2)$ est équivalent à minimiser

$$\begin{aligned}\Delta(h_1,h_2) &= \Sigma_z E\{\delta_1 \widehat{f}(z|E_1,h_1) - \delta_2 \widehat{f}(z|E_2,h_2)\}^2 \\ &- 2[\delta_1^2 f(z|E_1,h_1) E\widehat{f}(z|E_1,h_1) + \delta_2^2 f(z|E_2,h_2) E\widehat{f}(z|E_2,h_2) \\ &- \delta_1\delta_2\{f(z|E_1)E\widehat{f}(z|E_2,h_2) + f(z|E_2)E\widehat{f}(z|E_1,h_1)\}]\end{aligned}$$

Un estimateur sans biais de Δ est (cf. Hall & Wand 1988), si m désigne la taille de E_1, et n celle de E_2

$$\begin{aligned}\widehat{\Delta}(h_1,h_2) &= \Sigma_z\{\delta_1 \widehat{f}(z|E_1,h_1) - \delta_2 \widehat{f}(z|E_2,h_2)\}^2 \\ &- 2[\delta_1^2 m^{-1} \sum_{i=1}^{m} \widehat{f}(x_i|E_1 - \{x_i\}, h_1) + \delta_2^2 n^{-1} \sum_{i=1}^{n} \widehat{f}(y_i|E_2 - \{x_i\}, h_2) \\ &- \delta_1\delta_2\{m^{-1}\sum_{i=1}^{m} \widehat{f}(x_i|E_2,h_2) + n^{-1}\sum_{i=1}^{n} \widehat{f}(y_i|E_1,h_1)\}].\end{aligned}$$

En fait, Hall et Wand n'ont pas précisé les paramètres h_1 et h_2 qui minimisent $\widehat{\Delta}(h_1, h_2)$, nous les donnons ci-dessous.

Proposition 5 : *Lorsque la taille m de l'échantillon E_1 et la taille n de E_2 deviennent grandes, alors les paramètres h_1 et h_2 minimisant $\widehat{\Delta}$ sont définis approximativement par*

$$\begin{aligned}h_1 &\approx (T_{11}T_{21} - S_{12}^2)^{-1}(T_{21}(m-1)^{-1}S_{10} + \rho S_{12}(n-1)^{-1}S_{20}) \\ h_2 &\approx (T_{11}T_{21} - S_{12}^2)^{-1}((n-1)^{-1}T_{11}S_{20} + \rho^{-1}S_{12}(m-1)^{-1}S_{10}),\end{aligned} \tag{18}$$

pour $T_{11}T_{21} - S_{12}^2 \neq 0$, $\rho = \delta_2/\delta_1$ avec

$$\begin{gathered}T_{11} = m^{-2}\Sigma_z\{N_{1,1}(z)\}^2,\ T_{21} = n^{-2}\Sigma_z\{N_{1,2}(z)\}^2, \\ S_{12} = m^{-1}n^{-1}\Sigma_z N_{1,1}(z)N_{1,2}(z), \\ S_{10} = m^{-2}\Sigma_z N_{0,1}(z)N_{1,1}(z),\ S_{20} = n^{-2}\Sigma_z N_{0,2}(z)N_{1,2}(z). \\ N_{v,t}(z) = Card\{x \in E_t/d(x,z) = v\},\ v = 0,1 \text{ et } t = 1,2.\end{gathered}$$

Preuve : Il suffit d'écrire les développements de Taylor d'ordre deux, comme dans la preuve de la proposition 2 (voir Mkhadri 1990a).

Une version du bootstrap lissé

Maintenant on se situe dans les conditions du bootstrap lissé (§1.3.3). On note par E_1^* et E_2^* les échantillons tirés respectivement suivant $\widehat{f}(.|E_1, h_1)$ et

$\widehat{f}(.|E_2, h_2)$. Par analogie avec Hall & Wand, on cherche le couple (h_1, h_2) qui minimise le critère $\Delta^*(h_1, h_2)$, la version de $\Delta(h_1, h_2)$ basée sur les échantillons bootstrap E_1^* et E_2^*,

$$\Delta^*(h_1, h_2) = \Sigma_z E\{\widehat{g}^*(z|h_1, h_2) - \widehat{g}(z|h_1, h_2)\}^2,$$

avec

$$\widehat{g}^*(z|h_1, h_2) = \delta_1 \widehat{f}(z|E_1^*, h_1) - \delta_2 \widehat{f}(z|E_2^*, h_2)$$
$$\text{et } \widehat{g}(z|h_1, h_2) = \delta_1 \widehat{f}(z|E_1, h_1) - \delta_2 \widehat{f}(z|E_2, h_2).$$

De la même manière que précédemment , on a

Théorème 3 : *Lorsque la taille m de l'échantillon E_1 et la taille n de E_2 deviennent grandes, alors les paramètres h_{1B} et h_{2B} minimisant $\widehat{\Delta}^*$ sont définis approximativement par*

$$h_{1B} = \frac{m^{-1}F_2\mathbf{N}_1(m) + \rho m n^{-2}\mathbf{N}_2(n)G_{12}}{2\{F_1F_2 - G_{12}^2\}}$$
$$h_{2B} = \frac{n^{-1}F_1\mathbf{N}_2(n) + \rho^{-1} n m^{-2}\mathbf{N}_1(m)G_{12}}{2\{F_1F_2 - G_{12}^2\}}. \tag{19}$$

où

$$F_1 = m^{-1}D_{11}(m) - D_{21}(m) \text{ et } F_2 = n^{-1}D_{12}(n) - D_{22}(n)$$
$$G_{12} = \Sigma_x\{N_{1,1}(x) - pN_{0,1}(x)\}\{N_{1,2}(x) - pN_{0,2}(x)\} \text{ et } \rho = \delta_2/\delta_1$$
$$\mathbf{N}_t(n_t) = \Sigma_x\{2pN_{0,t}^2(x) - 2N_{01,t}(x)N_{1,t}(x) - 2N_{0,t}(x)N_{1,t}(x)\} - pn_t^2$$
$$D_{1t}(n_t) = \Sigma_x 2[N_{01,t}(x)N_{1,t}(x) + N_{0,t}(x)N_{11,t}(x) - 2pN_{0,t}(x)N_{1,t}(x)] + \Sigma_x N_{1,t}^2(x) + 2(pn_t)^2,$$
$$D_{2t}(n_t) = \Sigma_x(N_{01,t}(x) - pN_{0,t}(x))^2 \text{ et } N_{v1,t}(x) = \sum_{y \in E_t : d(x,y)=1} N_{v,t}(y), v = 0, 1,$$

avec $t = 1, 2$; $n_t = m$ si $t = 1$ et $n_t = n$ si $t = 2$.

Preuve : La démonstration est similaire à celle de la proposition 3 (voir Mkhadri 1990b).

3. Application pratique

Le but de l'exemple est d'étudier l'effet des stratégies présentées pour le choix des paramètres de lissage et l'efficacité de la règle de décision qui en découle. On compare six méthodes d'estimation du paramètre de lissage λ (la méthode de Hall :

disnop, la méthode basée sur la validation croisée §1.3.1 : disval, la méthode basée sur le bootstrap lissé : disbot, la méthode de Hall et Wand par validation croisée : dishw ; la méthode de Hall et Wand boostrapée : dishwb et la méthode du taux d'erreur : Taux. Les trois premières méthodes sont liées à la méthode d'estimation basée sur l'estimation de la densité, tandis que les autres sont liées directement à la discrimination.

Les données sont tirées de Goldstein & Dillon (1978, p. 15-16). Il s'agit d'un questionnaire auquel ont répondu 412 clients dont 154 étaient identifiés comme étant les clients d'un grand magasin (E_1) et 258 étaient identifiés comme étant les clients d'un magasin spécialisé (E_2). Les quatre variables binaires des questionnaires concernent, la volonté de l'information, la connaissance de l'information, l'expérience d'achat direct et l'expérience d'achat par catalogue. A partir de l'échantillon initial de 412 individus, on a tiré aléatoirement un sous-échantillon de 103 individus, qui constitue l'échantillon d'apprentissage et le reste constitue l'échantillon test.

Le tableau 1 résume les résultats des différentes méthodes sur ces données. Pour chaque méthode, nous fournissons le taux de bon classement obtenu sur l'échantillon d'apprentissage, par validation croisée et sur l'échantillon test. Les probabilités a priori ont été prises égales dans les deux groupes.

Sur le fichier d'apprentissage, toutes les méthodes fournissent des résultats analogues (sauf Taux) et peu variables suivant le taux de bon classement conditionnel. Le taux de bon classement par validation croisée est inférieur à celui obtenu sur le fichier d'apprentissage pour toutes les méthodes (sauf pour dishwb qui fournit un résultat certainement biaisé). La différence est de l'ordre de 1 à 2%, ce qui apparaît raisonnable du fait que le taux d'erreur par resubstitution est très optimiste.

Les méthodes dishw et dishwb fournissent un taux de bon classement, par validation croisée, légèrement meilleur que les autres méthodes. Par contre, les méthodes disval, disbot et Taux fournissent un taux de classement global identique, mais le taux de classement conditionnel peut être différent.

Pour l'échantillon test, les méthodes disval, dishwb et Taux fournissent un résultat légèrement meilleur que les autres méthodes.

On remarque aussi que les méthodes Taux et disval lissent plus fortement que les autres méthodes. Par contre, la méthode disbot apparaît peu intéressente, puisqu'elle peut fournir un paramètre de lissage égal à 1.

En conclusion, on constate que les méthodes Taux et disval apparaîssent plus avantageuses que les autres méthodes, du fait qu'elles lissent plus les données, ce qui, nous le pensons, doit être avantageux dans bien des cas.

4. Conclusion

Cette étude montre que l'emploi d'un paramètre de lissage minimisant le taux d'erreur estimé par validation croisée et l'estimateur, par validation croisée, de la somme des écarts moyens au carré fournissent des résultats satisfaisants. D'autre part, les résultats montrent qu'il y a peu de différences entre toutes ces méthodes.

TABLEAU 1
Résultats de la discrimination non paramétrique sur le fichier d'apprentissage, test et par validation croisée avec, pour chaque méthode, le pourcentage de bon classement global (après l'accolade, le taux de bon classement conditionnel) sur chaque fichier et différents paramètres de lissage

Méthode	apprentissage		validation croisée		test	paramètres de lissage λ_1, λ_2	
disnop	76.70	82.50 73.02	74.76	82.50 69.84	67.961	0.977	0.997
disval	76.70	77.50 76.19	75.73	77.50 74.60	69.579	0.881	0.856
disbot	76.70	77.50 76.19	75.73	82.50 71.43	67.961	1.000	1.000
dishw	76.70	82.50 73.02	76.70	75.00 77.78	67.961	0.989	0.991
dishwb	76.70	77.50 76.19	77.67	75.69 79.37	69.579	0.938	0.939
Taux	75.73	82.50 71.43	75.73	77.50 74.60	68.932	0.827	0.827

N. B. : *disnop : méthode de Hall*
disval : méthode fondée sur la validation croisée
disbot : méthode fondée sur le bootstrap lissé
dishw : méthode de Hall - Wand
dishwb : méthode de Hall - Wand Bootstrapée
Taux : méthode basée sur le taux d'erreur par validation croisée

Mais, on peut préférer les techniques de lissage qui visent à minimiser un critère directement lié au problème de discrimination.

Le grand intérêt pratique de ces méthodes est qu'elles sont faciles à programmer du fait qu'on a une estimation explicite du paramètre de lissage et qu'on évite ainsi le recours à un algorithme d'optimisation. Enfin, signalons que le lissage est bénéfique surtout lorsque les données sont clairsemées (grand nombre de variables).

Par ailleurs, la méthode des noyaux s'étend facilement aux variables qualitatives ayant plus de deux modalités (cf. Aitchison & Aitken 1976 pour les variable non ordonnées et Titterington & Bowman 1985 pour les variables ordonnées).

Remerciements

L'auteur remercie G. Celeux et l'éditeur pour leurs suggestions.

Bibliographie

AITCHISON J. & AITKEN C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* ***63***, 413-20.

AITKEN C. G. G. (1983). Kernel methods for the estimation of discrete distributions. *J. Statist. Comput. Simul.* ***16***, 189-200.

BOWMAN A. W. (1980). A note on consistency of kernel method for the analyse of categorical data. *Biometrika* ***67***, 682-4.

BOWMAN A. W. (1984). An alternative method of cross-validation for smoothing of density estimates. *Biometrika* ***71***, 353-60.

BOWMAN A. W., HALL P. & TITTERINGTON D. M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika* ***71***, 341-51.

BROWN P. J. & RUNDELL P. W. K. (1985). Kernel estimates for categorical data. *Technometrics* ***27***, 293-9.

CELEUX G. & MKHADRI A. (1991). Discrete Regularized Discriminant Analysis. *Rapports de recherche INRIA*, N° 1481.

GLICK N. (1978). Additive estimator for probabilities of correct classification. *Pattern Rocognition* ***10***, 211-222.

GOLDSTEIN M. & DILLON W. R. (1978). *Discrete discriminant analysis*. J. Wiley & Sons, New York.

HABBEMA J. D. F., HERMANS J. & REMME J. (1978). Variable kernel density estimation in discriminant analysis. *Compstat. 1978, Proceedings in Computational Statistics*, Physica Verlag.

HALL P. (1981a). On nonparametric multivariate binary discrimination. *Biometrika* ***68***, 287-94.

HALL P. (1981b). Optimal near neighbour estimator for use in discriminant analysis. *Biometrika* ***68***, 572-5.

HALL P. (1983). Orthogonal series methods for qualitative and quantitative data. *Ann. Statist.* ***11***, 1004-7.

HALL P. & TITTERINGTON D. M. (1987). On smoothing sparse multinomial data. *Australian J. Statist.* ***29***, 19-37.

HALL P. & WAND P. (1988). Nonparametric discrimination using density differences. *Biometrika* ***75***, 541-7.

HALL P. (1990). Using bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multiv. Anal.* ***32***, 177-203.

HAND D. J. (1982). *Kernel discriminant analysis*. Chichester : Research Studies Press. Wiley.

HAND D. J. (1983). A comparative of two methods of discriminant anlysis applied to binary data. *Biometrics* ***39***, 683-94.

HILLS M. (1967). Discrimination and allocation with discete data. *J. Roy. Stat. Soc.* ***C16***, 237-250.

MKHADRI A. (1990a). Classification et discrimination des données qualitatives : Discrimination Multinomiale Régularisée. *Thèse de Doctorat de Paris 6.*

MKHADRI A. (1990b). Discrimination binaire nonparamétrique : méthodes d'estimation du paramètre de lissage. *Rapports de recherche INRIA*, N° 1335.

RUDEMO M. (1982). Empirical choice of histograms and kernel density estimation. *Scand. J. Statist.* ***9***, 65-78.

STONE M. (1977). Asymptotics for and against cross-validation. *Biometrika* ***64***, 29-35.

TAYLOR C. C. (1989). Bootstrap choice of smoothing parameter in kernel density estimation. *Biometrika* ***76***, 705-12.

TITTERINGTON D. M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics* ***22***, 259-68.

TITTERINGTON D. M & BOWMAN A. W. (1985). A comparative study of smoothing procedures for ordered categorical data. *J. Statist. Comput. Simul.* ***21***, 291-312.

TUTZ G. (1985). Smoothed additive estimators for nonerror rates in multiples discriminant analysis. *Pattern Rocognition* ***18***, 151-159.

TUTZ G. (1986). An alternative choice of smoothing for kernel-based density estimates in discrete discriminant analysis. *Biometrika* ***73***, 405-11.

TUTZ G. (1988). Smoothing for discrete kernels in discrimination. *Biometrical Journal* ***30***, 729-40.

TUTZ G. (1989). On cross-validation for discrete kernel estimation in discrimination. *Commun. Statist.-Theory Meth.* ***18*** **(11)**, 4145-4162.

VAN NESS J. (1979). On the effects of dimension in discrimant analysis for unequal covariance populations. *Technometrics* ***21***, 119-27.

VAN NESS J. & SIMPSON C. (1976). On the effects of dimension in discrimant analysis. *Technometrics* ***18***, 175-87.