Discrimination binaire non parametrique. Methodes d'estimation du parametre de lissage

A. Mkhadri

INRIA

UNITE DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP105
78153 Le Chesnay Cedex
France
Tel (1) 39 63 55 11

Rapports de Recherche

N° 1335

Programme 5
Automatique, Productique,
Traitement du Signal et des Données

DISCRIMINATION BINAIRE NON PARAMETRIQUE METHODES D'ESTIMATION DU PARAMETRE DE LISSAGE

Abdallah MKHADRI

Novembre 1990

RR_1335

DISCRIMINATION BINAIRE NON PARAMETRIQUE METHODES D'ESTIMATION DU PARAMETRE DE LISSAGE

BINARY NONPARAMETRIC DISCRIMINATION METHODS OF ESTIMATION OF SMOOTHING PARAMETER

Abdallah Mkhadri
INRIA Rocquencourt, Domaine de Voluceau
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 des algorithmes 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 ; rééchantionnage ; 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 f(.| k), like the MISE. 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, fondée sur un estimateur des noyaux. Cet estimateur, construit sur la base de l'échantillon d'apprentissage $\{x_1,..., x_n\}$, peut s'écrire

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

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

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

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

Tous ces noyaux 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 (Aitchison & Aitken (1976)) ou sur des fonctions de perte pour l'estimation de $f(.|E_k)$. Cette approche traditionnelle, dans laquelle les paramètres sont déterminés séparément pour chaque groupe, a été beaucoup étudiée par plusieurs auteurs (Bowman (1980), Titterington (1980), Hall (1981), 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 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 (ou critère de MISE) et d'information de Kullback-Leibler, nous proposons, dans le même cadre, des variantes fondées sur la validation croisée et la technique du bootstrap. On présentera 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 autre méthode, à celle de Tutz, qui nous fournit explicitement le paramètre optimal de lissage sans avoir recours aux algorithmes 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 dans le paragraphe 3.

1. PROCEDURES DE LISSAGE DANS LE CADRE DE L'ESTIMATION DE LA DENSITE

Dans cette section $\hat{f}(.,\lambda)$ (ou $\hat{f}(.|E,\lambda)$) 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

$$V(\lambda|\, E) = \prod_{i=1}^{n} \hat{f}(x_i|E,\lambda) \qquad (2)$$

$$= \prod_{x\in B^p} \{ \sum_{z\in B^p} (n_z/n) K(x|z,\lambda)\}^{n_x},$$

avec $K(x|z,\lambda) = \lambda^{p - d(x,z)}(1 - \lambda)^{d(x,z)}$ où , rappelons-le, $B^p = \{0,1\}^p$ et n_x est la fréquence des observations x, conduit à λ égal à 1 (cf. Hand (1982) ou Mkhadri (1990)).

Maintenant la vraisemblance par validation croisée s'écrit

$$W(\lambda,E) = \prod_{i=1}^{n} \hat{f}(x_i|E-\{x_i\},\lambda)$$

$$= \prod_{x\in B^p} \{ \sum_{z\in B^p} [n_z/(n-1)]\, K(x|z,\lambda) - (\lambda^p/n-1)\}^{n_x} \qquad (3)$$

(car pour x = z, n_z est devenu n_z - 1, il faut donc retrancher $K(x|x,\lambda) = \lambda^p$ (au facteur 1/(n-1) près)).

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

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

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

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

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

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

Il en déduit que la méthode de maximisation de la pseudo-vraisemblance de Aitchison & Aitken est consistante.

1.2 Méthode de Hall et critique de la validation croisée

Hall (1981a) note que si une cellule est vide et si les cellules non vides contiennent plus d'un élément, alors $\frac{\partial \mathrm{Log}\, W}{\partial \lambda}$ peut être positif en $\lambda = 1$, qui peut correspondre ainsi à un maximum local de W. Pour éviter ce problème, il propose de maximiser l'un des critères

$$J_1(\lambda) = E\,\Sigma_X \{f(x) - \hat{f}(x)\}^2$$

ou (4)

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

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

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

et

$$n\,\mathrm{var}(\hat{f}(x|\lambda)) = f(x)(1- f(x)) - 2(1 - \lambda) f(x)\{p(1-f(x)) + f_1(x)\} + O\{(1-\lambda)^2\}, \quad (6)$$

où $f_1(x) = \sum_{y:d(x,y)=1} f(y)$. En remplaçant dans :

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

il en déduit, par approximations jusqu'à l'ordre un, que $\hat{\lambda}_{1H}$ et respectivement $\hat{\lambda}_{2H}$ qui minimisent J_1 et respectivement J_2 (en prenant comme fonction de poids $w(x) = f(x)$ dans J_2) sont définis par

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

$$\hat{\lambda}_{2H} = 1 - \frac{p - \Sigma_x (f(x))2\{f_1(x) + p(1 - f(x))\}}{n\Sigma_x f(x)\{f_1(x) - pf(x)\}^2}. \quad (8)$$

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 $N_v(b) = \text{Card}\{x \in \{0,1\}^p \mid d(x,b) = v\}$, v = 0,1), soit par une procédure itérative (en regardant cette formule comme une fonction implicite)

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

$\hat{\lambda}_0 = 1$ peut être utilisée 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

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

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

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 $N_v(b)$ (v=1,..., r), où, pour tout $b \in B^p$, on a

$$\hat{f}(b,\omega) = n^{-1} \sum_{v=0}^{r} \omega_v N_v(b) \qquad (9)$$

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

Remarque 1 : Le cas, $\omega_v = 1$ pour $v \le r$ et $\omega_j = 0$ pour tout $j > v$, 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.

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

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

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

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

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

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

(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 $\hat{Q}$ et ainsi à l'estimateur

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

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

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

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

où $\hat{A} = (1- n^{-1})I + n^{-1} \hat{Q}^{-1}D$ 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 Autres 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

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

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

$$\text{Cval}(\lambda) = \Sigma_x \hat{f}^2(x,E,\lambda) - \frac{2}{n}\sum_{j=1}^{n} \hat{f}(x_j|E-\{x_j\},\lambda) \qquad (13)$$

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

La valeur optimale de λ minimisant Cval(λ) est approximée par

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

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

$$\hat{f}(z|E,\lambda) = n^{-1}\{ N_0(z) + (1 - \lambda) N_1(z) + O(1 - \lambda)^2 \},$$

$$\hat{f}(z|E-\{x_i\},\lambda) = (n-1)^{-1}\{ N_0^*(z) + (1 - \lambda) N_1(z) + O(1 - \lambda)^2\},$$

$$\hat{f}^2(x|E,\lambda) = n^{-2}\{ N_0^2(x) + 2 N_0(x) N_1(x)(1 - \lambda) + N_1^2(x)(1 - \lambda)^2 + O(1 - \lambda)^3\},$$

où $N_0^*(x_i) = N_0(x_i) - 1$.

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

$$\text{Cval}(\lambda) \approx \Sigma_x\, n^{-2} N_0^2(x) - \frac{2}{n(n-1)} N_0^*(x) N_0(x) + 2(1 - \lambda)[\, n^{-2} N_0(x) N_1(x)$$

$$- \{n(n-1)\}^{-1} N_0(x) N_1(x)] + (1 - \lambda)^2 n^{-2} \Sigma_x N_1^2(x)$$

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

$$h = \frac{-\Sigma_x[\, n^{-2} N_0(x) N_1(x) - \{n(n-1)\}^{-1} N_0(x) N_1(x)]}{\Sigma_x\, n^{-2} N_1^2(x)}$$

$$= \frac{\Sigma_x N_0(x) N_1(x)}{(n-1)\Sigma_x N_1^2(x)}$$

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

Remarque 2 : On remarque qu'on pourra utiliser une autre approximation de (1) sans négliger le terme en facteur λ^p. En effet, posons $\gamma = (1-\lambda)/\lambda$; l'estimateur des noyaux s'écrit alors

$$\hat{f}(x,\gamma) = \{1/(1+\gamma)\}^p \sum_{j=0}^{p} N_j(b)\gamma^j,$$

avec $\gamma \in [0,1]$. Par approximation de Taylor d'ordre 2 en γ, on obtient (cf. Mkhadri (1990))

$$\gamma_{val} \approx \frac{\Sigma_X N_0(x)N_1(x)}{(n-1)\Sigma_X \{N_1(x) - pN_0(x)\}^2}.$$

Remarque 3 : 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 (1990)) que le poids optimal ω_{val} minimisant Cval (ω) est défini par :

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

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 de r+1 composantes. Cette dernière expression de ω_{val} est valable uniquement lorsque **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 $E^* = \{x_1^*,\ldots,x_m^*\}$ tiré aléatoirement de la loi de $\hat{f}$. On définit $\hat{f}^*$ sur E^* de la même manière que $\hat{f}$ sur E. Ainsi, il arrive à exprimer le critère uniquement en fonction des données initiales sans rééchantionnage.

On adaptera ici cette procédure aux données binaires dans le but d'obtenir le paramètre de lissage optimal dépendant uniquement des données. Nous montrerons que cette méthode est consistante et nous fournit un paramètre de lissage optimal explicite.

On suppose que E^* est tiré suivant $\hat{f}(.|n,\lambda)$, et a la même taille n que E. Pour cela, nous utilisons la propriété suivante :

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

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

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

$$\hat{f}^*(x|n,h) = n^{-1}\{M_0(x) + hM_1(x) + O(h^2)\}.$$

$$\hat{f}(x|n,h) = n^{-1}\{N_0(x) + hN_1(x) + O(h^2)\}$$

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

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

D'après la formule (6), l'espérance de $\hat{f}^*(x|n,h)$ ($h = 1 - \lambda$) s'écrit

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

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

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

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

$$\begin{aligned}\hat{SMEQ}(h) \approx \Sigma_x\, n^{-3}\{&N_0(x)(n - N_0(x)) + h[nN_1(x) - 2N_0(x)N_1(x)] \\ &-h[2pN_0(x)\{n - N_0(x)\} + 2N_0(x)N_{01}(x)] \\ &- h^2 2[N_1(x)N_{01}(x) + N_0(x)N_{11}(x)] - N_1^2(x) \\ &- h^2 2p[nN_1(x) - 2N_0(x)N_1(x)]\} \\ &+ \Sigma_x\, n^{-2}h^2(N_{01}(x) - pN_0(x)^2.\end{aligned}$$

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

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

Remarque 4 : Dans un cadre plus général, et indépendamment de Taylor, Hall (1990) propose une méthode d'utilisation du bootstrap classique, pour minimiser le critère global des écarts moyens quadratiques (en englais MISE) (ou pour minimiser un critère analogue utilisant la norme L_q, avec q entier), dans le cadre de l'estimation non paramétrique de la densité. On montre (cf. Mkhadri (1990)) que cette procédure nous fournit un paramètre de lissage qui peut être supérieur à 1.

2. METHODES D'ESTIMATION LIEES A LA DISCRIMINATION

Jusqu'à présent, on s'est intéressé aux méthodes d'estimation 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 de Tutz

Dans le but de lier le choix du paramètre de lissage λ à la discrimination, l'estimation des distributions des groupes séparés doit être considérée simultanément.

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 $\hat{D} = (\hat{D}_1,..., \hat{D}_K)$ la règle de classification générée par l'utilisation des estimateurs de densité basés sur les noyaux définis en (1), où $x \in \hat{D}_k$ signifie que

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

avec $\hat{f}(x|E_k,\lambda_k) = n_k^{-1} \sum_{y \in E_k} \lambda_k^{p - d(x,y)}(1 - \lambda_k)^{d(x,y)}$ 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

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

Soit $p_0(x) = \{\delta_1 p(x|1),...,\delta_K p(x|K)\}$, le vecteur des scores discriminants optimaux et soit $\hat{p}(x|k)$ 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))

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

où $I_k = I_k \{\delta_1 \hat{f}(x|E_1,\lambda_1),...,\delta_k \hat{f}(x|E_k-\{x\},\lambda_k),..., \delta_K \hat{f}(x|E_K,\lambda_K)\}$ (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 consistante. Mais Tutz montre explicitement que la maximisation de $t_L(\lambda)$ fournit une procédure d'estimation qui est fortement consistante.

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 régle de classification qui en résulte est fortement consistante au sens de Bayes.

Remarques 5 : La consistance résultant du théorème s'applique aussi dans le cas où, à la place des échantillons séparés, on utilise un échantillon issu d'une population mixte ; mais alors les estimateurs des probabilités a priori seront estimés par leurs fréquences relatives (i.e. $\hat{\delta}_k = n_k/n$, 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 $t_L(\lambda)$.

Une autre approche consiste à imposer un seul paramètre de lissage pour toutes les classes ($\lambda = \lambda_1 = ... = \lambda_K$). Des comparaisons pratiques ont 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). D'après (§ 1.3.1), l'estimateur des noyaux par validation croisée pour la classe E_k peut s'écrire

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

où l'on a posé $\gamma = (1 - \lambda) / \lambda$, $\gamma \in [0,1]$; qui se réduit à

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

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

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

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

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

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

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

avec $x \in E$.

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

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

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 (17). 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, soit sera de type plus proches voisins.

La proposition 3 montre que le γ optimal est à rechercher parmi 0,1 et les racines des équations $C(x_i,\gamma) = 0$ $(1 \le i \le n)$. 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 (1990) ch. 5, Celeux & Mkhadri (1990)).

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 : $\delta_1 f(x|E_1) \ge \delta_2 f(x|E_2)$, ce qui est équivalent à

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

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

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

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

$$\Delta(h_X, h_Y) = \Sigma_z E\{\delta_1 \hat{f}(z|X,h_X) - \delta_2 \hat{f}(z|Y,h_Y)\}^2$$
$$- 2[\delta_1^2 f(z|X,h_X) E \hat{f}(z|X,h_X) + \delta_2^2 f(z|Y,h_Y) E \hat{f}(z|Y,h_Y)$$
$$- \delta_1 \delta_2 \{f(z|X) E \hat{f}(z|Y,h_Y) + f(z|Y) E \hat{f}(z|X,h_X)]$$

Un estimateur sans biais de Δ est

$$\hat{\Delta}(h_X, h_Y) = \Sigma_z \{\delta_1 \hat{f}(z|X,h_X) - \delta_2 \hat{f}(z|Y,h_Y)\}^2$$
$$- 2[\delta_1^2 m^{-1} \sum_{i=1}^{m} \hat{f}(x_i|X-\{x_i\},h_X) + \delta_2^2 n^{-1} \sum_{i=1}^{n} \hat{f}(y_i|Y-\{y_i\},h_Y)$$
$$- \delta_1 \delta_2 \{ m^{-1} \sum_{i=1}^{m} \hat{f}(x_i|Y,h_Y) + n^{-1} \sum_{i=1}^{n} \hat{f}(y_i|X,h_X)\}].$$

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

Proposition 4 : *Lorsque la taille m de l'échantillon X et la taille n de Y deviennent grandes, alors les paramètres h_X et h_Y minimisant $\hat{\Delta}$ sont définis approximativement par*

$$h_X \approx (T_{X1} T_{Y1} - S_{XY}^2)^{-1} (T_{Y1} (m-1)^{-1} S_{X0} + \rho S_{XY} (n-1)^{-1} S_{Y0}) \qquad (18)$$
$$h_Y \approx (T_{X1} T_{Y1} - S_{XY}^2)^{-1} ((n-1)^{-1} T_{X1} S_{Y0} + \rho^{-1} S_{XY} (m-1)^{-1} S_{X0}),$$

pour $T_{X1} T_{Y1} - S_{XY}^2 \neq 0$, $\rho = \delta_2/\delta_1$ *avec*

$$T_{X1} = m^{-2} \Sigma_z \{N_{1,X}(z)\}^2 , \quad T_{Y1} = n^{-2} \Sigma_z \{N_{1,Y}(z)\}^2,$$
$$S_{XY} = m^{-1} n^{-1} \Sigma_z N_{1,X}(z) N_{1,Y}(z),$$
$$S_{X0} = m^{-2} \Sigma_z N_{0,X}(z) N_{1,X}(z), \; S_{Y0} = n^{-2} \Sigma_z N_{0,Y}(z) N_{1,Y}(z).$$
$$N_{v,T}(z) = Card\{x \in T / d(x,z) = v\}, \; v = 0, 1 \text{ et } T = X, Y.$$

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

Une version du bootstrap lissé

Maintenant on se situe dans les conditions du bootstrap lissé (§ 1.3.3). On note par X^* et Y^* les échantillons tirés respectivement suivant $\hat{f}(.|X,h_X)$ et $\hat{f}(.|X,h_Y)$. Par analogie

avec Hall & Wand, on cherche le couple (h_X,h_Y) qui minimise le critère $\Delta^*(h_X,h_Y)$, la version de $\Delta(h_X,h_Y)$ basée sur les échantillons bootstrap X^* et Y^*,

$$\Delta^*(h_X,h_Y) = \Sigma_z E\{\hat{g}^*(z|h_X,h_Y) - \hat{g}(z|h_X,h_Y)\}^2,$$

avec

$\hat{g}^*(z|h_X,h_Y) = \delta_1 \hat{f}(z|X^*,h_X) - \delta_2 \hat{f}(z|Y^*,h_Y)$ et $\hat{g}(z|h_X,h_Y) = \delta_1 \hat{f}(z|X,h_X) - \delta_2 \hat{f}(z|Y,h_Y)$.

De la même manière que précédemment, on a

Théorème 3 : *Lorsque la taille m de l'échantillon X et la taille n de Y deviennent grandes, alors les paramètres h_{XB} et h_{YB} minimisant $\hat{\Delta}^*$ sont définis approximativement par*

$$h_{XB} = \frac{m^{-1}F_Y N_X(m) + \rho m n^{-2} N_Y(n) G_{XY}}{2\{F_X F_X - G_{XY}^2\}}$$

$$h_{YB} = \frac{n^{-1}F_X N_Y(m) + \rho^{-1} n m^{-2} N_X(n) G_{XY}}{2\{F_X F_X - G_{XY}^2\}}. \qquad (19)$$

où

$$F_X = m^{-1}D_{1X}(m) - D_{2X}(m) \quad \text{et} \quad F_Y = n^{-1}D_{1Y}(n) - D_{2Y}(n)$$

$$G_{XY} = \Sigma_x \{N_{1X}(x) - pN_{0X}(x)\}\{N_{1Y}(x) - pN_{0Y}(x)\} \quad \text{et} \quad \rho = \delta_2/\delta_1$$

$$N_T(n_T) = \Sigma_x\{2pN_{0T}^2(x) - 2N_{01}^T(x)N_{1T}(x) - 2N_{0T}(x)N_{1T}(x)\} - pn_T^2$$

$$D_{1T}(n_T) = \Sigma_x 2[N_{01}^T(x)N_{1T}(x) + N_{0T}(x)N_{11}^T(x) - 2pN_{0T}(x)N_1(x)] + \Sigma_x N_{1T}^2(x) + 2(pn_T)^2,$$

$$D_{2T}(n_T) = \Sigma_x(N_{01}^T(x) - pN_{0T}(x))^2,$$

avec $T = X, Y$; $n_T = m$ si $T = X$ et $n_T = n$ si $T = Y$.

Preuve : La démonstration est similaire à celle de la propriété 3 (voir annexe 2).

3. APPLICATION PRATIQUE

Données Store

Ces données sont tirées de Goldstein & Dillon (1978). Il s'agit de 412 individus décrits par 4 variables binaires. Cet échantillon est divisé en deux groupes E_1 et E_2 de tailles respectives 154 et 258.

Le but de l'exemple est d'étudier l'effet de différencier les fonctions de perte sur le choix des paramètres de lissage et sur l'efficacité de la règle de décision. On compare six méthodes d'estimation du paramètre de lissage (la méthode de Hall : λ_H, la validation

croisée : λ_{val}, le bootstrap lissé : λ_{boot}, la méthode de H-W par validation croisée : $\lambda_{H\text{-}W}$; la méthode de H-W boostrapée : λ_{HWBoot} et la méthode du taux d'erreur : λ_{taux}). Les trois premiers paramètres λ_H, λ_{val}, λ_{boot} sont liés à la méthode d'estimation basée sur l'estimation de la densité, tandis que les autres sont liés directement à la discrimination. Deux cas sont distingués dans cette étude, à savoir : le cas équiprobable et le cas où les probabilités a priori sont proportionnelles aux groupes.

Dans le cas d'équipropabilité, d'après les résultats de la Table 1, les paramètres obtenus par les méthodes d'estimation liées à l'estimation de la densité fournissent des résultats satisfaisants et meilleurs que ceux des méthodes liées à la discrimination. La méthode de bootstrap lissé fournit un résultat similaire à la méthode de Hall et les valeurs des paramètres sont peu différentes. D'un autre côté, la méthode de Hall et Wand bootstrapée fournit un résultat proche de la méthode de Hall et meilleur que celui de Hall-Wand. Les résultats de la Table 1 montrent que les méthodes de Hall et celles basées sur le bootstrap fournissent toujours des estimations des paramètres de lissage plus grandes que les méthodes fondées sur la validation croisée (voir Hand (1983)). Par ailleurs, la méthode fondée sur la minimisation du taux d'erreur fournit un résultat meilleur que les méthodes de Hall, Hall-Wand et celles basées sur le bootstrap, mais un peu moins bon que la méthode basée sur la validation croisée. Néanmoins, ce résultat est intéressant, du fait que la méthode d'estimation et celle de validation sont fondées sur la même fonction de perte et confirme l'idée de l'efficacité d'utilisation du critère dont le but est lié à la séparation des groupes (Tutz (1989)).

Pour le deuxième cas, celui où les probabilités a priori sont proportionnelles à la taille des deux classes, trois points sont à signaler.
Premièrement, on peut faire la même remarque pour les estimations des paramètres de lissage que dans le premier cas ; de même, les méthodes fondées sur la validation croisée ont des tendances opposées, sur chaque groupe a priori, par rapport aux autres méthodes, à savoir que ces dernières méthodes (sauf la méthode H-W) avantagent beaucoup plus le premier groupe (effectif très petit) que le deuxième groupe, contrairement aux autres méthodes. Cet effet s'inversait dans le premier cas (voir Table1).

Le deuxième point est que la méthode H-W fournit toujours un taux d'erreur par resubstitution très petit, mais un taux d'erreur par validation croisée important. Ce qui est surprenant est que la tendance sur chaque groupe par resubstitution est tout à fait opposée

Table 1 : *Résultats de la discrimination non paramétrique sur le fichier Store avec différents paramètres de lissage et probabilités des classes a priori égales*

Paramètre de lissage	apprentissage	validation croisée	λ_1	λ_2
λ_H	70.39 { 80.52 64.34	70.39 { 80.52 64.34	0.996	0.999
λ_{val}	72.09 { 61.69 78.29	72.09 { 61.69 78.29	0.892	0.868
λ_{Boot}	70.39 { 80.52 64.34	70.39 { 80.52 64.34	0.990	0.999
$\lambda_{H\text{-}W}$	66.50 { 84.42 55.81	66.50 { 84.42 55.81	0.858	0.887
$\lambda_{H\text{-}WB}$	69.66 { 82.47 62.02	69.66 { 82.47 62.02	0.961	0.992
λ_{taux}		70.39 { 80.52 64.34	$\lambda = \lambda_1 = \lambda_2 = 0.994$	

N. B. : λ_H *: paramètre obtenu par la méthode de Hall*

λ_{val} *: paramètre obtenu par validation croisée*

λ_{boot} *: paramètre obtenu par la méthode de bootstrap lissé*

$\lambda_{H\text{-}W}$ *: paramètre obtenu par la méthode de Hall - Wand*

$\lambda_{H\text{-}WB}$ *: paramètre obtenu par la méthode de Hall - Wand Bootstrapée*

λ_{taux} *: paramètre obtenu par minimisation du taux d'erreur par validation croisée*

à celle de l'échantillon basée sur la validation croisée. On n'a pas d'explication à ce comportement.

Le troisième point intéressant est que la méthode fondée sur la minimisation du taux d'erreur fournit un taux d'erreur meilleur avec la même tendance sur chaque groupe que les autres méthodes (sauf H-W). Donc cet exemple confirme l'idée de l'efficacité de la méthode d'estimation fondée sur le critère du taux d'erreur par validation croisée qui est liée directement à la discrimination. De même, la méthode basée sur les moindres carrés bootstrapés est plus intéressante que celle de Hall-Wand.

4. CONCLUSION

La méthode des noyaux peut prendre deux formes d'extension d'importance pratique. Premièrement, pour les variables qualitatives avec plus deux modalités, le facteur contribuant au noyau dépendra de la nature d'ordre introduit par les modalités. Aitchison & Aitken (1976) ont considéré le noyau suivant : pour chaque variable j avec m_j modalités non ordonnées (j = 1,..., p), ils posent

$$K_j(y_j \mid x_j, \lambda) = \begin{cases} \lambda & (y_j = x_j) \\ (1 - \lambda) / (m_j - 1) & (y_j \neq x_j) \end{cases}$$

Pour la variable j avec m_j modalités ordonnées, ils ont considéré, dans un exemple avec $m_j = 3$, un noyau avec un poids qui tient compte du voisinage de y_j de la catégorie observée x_j. Rappelons qu'une revue intéressante des méthodes des noyaux utilisant la structure d'ordre sur les modalités des variables a été considérée par Titterington & Bowman (1985). Ils ont comparé plusieurs méthodes des noyaux, sur données simulées, utilisées dans ce cadre.

La deuxième extension possible des méthodes des noyaux a été considérée par Titterington (1980) pour les données manquantes.
Dans ces deux cas d'extension, la méthode du pseudo-vraisemblance est utilisée pour estimer le paramètre de lissage par l'utilisation des algorithmes d'optimisation.

Par ailleurs, Kokalakis & Johnson (1989) ont récemment proposé une méthode bayesienne pour estimer le paramètre de lissage λ de l'estimateur (1), et ont montré que la méthode de Hall était un cas particulier de leur méthode.

Enfin, la théorie asymptotique de tables de contingence très clairsemées a été étudiée, récemment, en détail par Hall & Titterington (1987) et indépendamment d'eux par

Table 2 : *Résultats de la discrimination non paramétrique sur le fichier Store avec différents paramètres de lissage et probabilités des classes a priori proportionnelles aux tailles des classes*

Paramètre de lissage	apprentissage	validation croisée	λ_1	λ_2
λ_H	73.79 { 50.00 87.98	73.30 { 42.86 91.47	0.996	0.999
λ_{val}	72.82 { 39.61 92.64	71.12 { 31.82 94.57	0.892	0.868
λ_{Boot}	73.79 { 50.00 87.98	73.30 { 42.86 91.47	0.996	1.000
$\lambda_{H\text{-}W}$	66.50 { 84.42 55.81	72.82 { 51.27 95.85	0.843	0.922
$\lambda_{H\text{-}WB}$	72.57 { 68.83 74.81	73.30 { 42.86 91.47	0.961	0.992
λ_{Taux}		73.06 { 44.81 89.92	$\lambda = \lambda_1 = \lambda_2 = 0.950$	

probabilité a priori de la classe 1 : 0.374 probabilité a priori de la classe 2 : 0.626

N. B. : λ_H *: paramètre obtenu par la méthode de Hall,*

λ_{val} *: paramètre obtenu par validation croisée*

λ_{boot} *: paramètre obtenu par la méthode de bootstrap lissé*

$\lambda_{H\text{-}W}$ *: paramètre obtenu par la méthode de Hall - Wand*

$\lambda_{H\text{-}WB}$ *: paramètre obtenu par la méthode de Hall - Wand Bootstrapée*

λ_{taut} *: paramètre obtenu par minimisation du taux d'erreur par validation croisée*

Burman (1987). Leurs études ont été considérées dans un cadre asymptotique non standard où le nombre des cellules (ou vecteur état) converge vers l'infini quand $n \to \infty$, mais le rapport m/n reste constant (m étant le nombre des cellules). Ils montrent, en particulier, que leurs estimateurs sont meilleurs que l'estimateur des fréquences observées au sens de la minimisation de la somme des écarts moyens au carré (MISE, en anglais).

REFERENCES

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

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.

BURMAN P. (1987). Smoothing sparse contingency tables. *Sunkhyā : The Indian J. Stat.*, **49**, 24-36.

CELEUX G. & MKHADRI A. (1990). Regularized Multinomial Discriminant. *Rapports de recherche INRIA.* (à paraître)

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

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. & 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.*

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.

KOKOLAKIS G. E. & JOHNSON W. O. (1989). Bayesian estimation of multinomial probabilities and smoothing parameters in the binary classification problem. *Preprint.*

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

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

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. (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.

ANNEXES

Annexe 1 : On montre que λ_{val} converge vers 1 (resp. 0) quand n tend vers l'infini.

En effet , on a

$$h = 1 - \lambda = \frac{\Sigma_x N_0(x) N_1(x)}{(n-1)\Sigma_x N_1^2(x)}.$$

Pour cela, nous montrons que $\Sigma_x N_0(x) N_1(x) \leq \Sigma_x N_1^2(x)$.

Par définition de $N_0(x)$ et $N_1(x)$ et du fait que $\Sigma_x N_0(x) = n$ et $\Sigma_x N_1(x) = np$, on peut toujours ordonner les suites $\{N_0(x)\}$ et $\{N_1(x)\}$, $x \in \{0,1\}^p$, tels que :

$N_0(\ell_1) \leq N_0(\ell_2) \leq ... \leq N_0(\ell_{2^p})$ et $N_1(\tau_1) \leq N_1(\tau_2) \leq ... \leq N_1(\tau_{2^p})$.

Ainsi, il est facile de voir que $N_0(\ell_i) \leq N_1(\tau_i)$ pour tout $i = 1,..., 2^p$, d'où on en déduit que $N_0^2(\ell_i) \leq N_1^2(\tau_i)$. Par conséquent, on a

$$\Sigma_x N_0^2(x)(x) \leq \Sigma_x N_1^2(x)$$

et de plus

$$\Sigma_x N_0(x) N_1(x) \leq \frac{1}{2}\Sigma_x \{N_0^2(x) + N_1^2(x)\} \leq \Sigma_x N_1^2(x).$$

Si n devient grand, h est inférieur à 1/(n-1) et donc il converge vers 0.

Annexe 2 : On montre le résultat du théorème 3 On a

$$E\hat{g}^*(z|h_X,h_Y) = \delta_1 E\hat{f}(z|X^*,h_X) - \delta_2 E\hat{f}(z|Y^*,h_Y)$$
$$Var\hat{g}^*(z|h_X,h_Y) = \delta_1^2 Var\hat{f}(z|X^*,h_X) + \delta_1^2 Var\hat{f}(z|Y^*,h_Y).$$

D'après l'expression (4.6) du § 1.2, on a

$mvar(\hat{f}^*(z|m,h_X)) \approx \hat{f}_X(z|m,h_X)\{1 - \hat{f}_X(z|m,h_X)\} - 2h_X\hat{f}_X(z|m,h_X)\{p(1 - \hat{f}_X(z|m,h_X)) + \hat{f}_{1X}(z|m,h_X)\}$

$nvar(\hat{f}^*(x|n,h_Y)) \approx \hat{f}_X(z|n,h_Y)\{1- \hat{f}_X(z|n,h_Y)\} - 2h_Y\hat{f}_Y(z|n,h_Y)\{p - p\hat{f}_Y(z|n,h_Y) + \hat{f}_{1Y}(z|n,h_Y)\}$

$[E\hat{g}^*(z|h_X,h_Y) - \hat{g}(z|h_X,h_Y)]^2 \approx \delta_1^2 h_X^2 [\hat{f}_{1X}(z|m,h_X) - p\hat{f}_X(z|m,h_X)]^2$

$$+ \delta_2^2 h_Y^2 | \hat{f}_{1Y}(z|n,h_Y) - p\hat{f}_Y(z|n,h_Y)|^2$$

$$+ 2\delta_1\delta_2 h_X h_Y |\hat{f}_{1X}(z|m,h_X) - p\hat{f}_X(z|m,h_X)||\hat{f}_{1Y}(z|n,h_Y) - p\hat{f}_Y(z|n,h_Y)|.$$

Suivant la preuve de la propriété 3, on a approximativement

$\text{var}(\hat{f}^*(z|m,h_X)) \approx m^{-3}\{\Sigma_x N_{0Y}(x)(m - N_{0X}(x)) + h_X N_X(m) - h_X^2 D_{1X}(m)\}$

de même

$\text{var}(\hat{f}^*(z|n,h_Y)) \approx n^{-3}\{\Sigma_x N_{0Y}(x)(n - N_{0Y}(x)) + h_Y N_Y(n) - h_Y^2 D_{1Y}(n)\}$,

où

$$N_T(n_T) = \Sigma_x\{2pN_{0T}^2(x) - 2N_{01}^T(x)N_{1T}(x) - 2N_{0T}(x)N_{1T}(x)\} - pn_T^2$$

$$D_{1T}(n_T) = \Sigma_x 2[N_{01}^T(x)N_{1T}(x) + N_{0T}(x)N_{11}^T(x) - 2pN_{0T}(x)N_1(x)] + \Sigma_x N_{1T}^2(x) + 2(pn_T)^2,$$

$$D_{2T}(n_T) = \Sigma_x(N_{01}^T(x) - pN_{0T}(x))^2,$$

avec $T = X, Y$; $n_T = m$ si $T = X$ et $n_T = n$ si $T = Y$.

De plus, on a

$$[E\hat{g}^*(z|h_X,h_Y) - \hat{g}(z|h_X,h_Y)]^2 \approx \delta_1^2 h_X^2 \Sigma_x m^{-2}(N_{1X}(x) - pN_{0X}(x)^2 + \delta_2^2 h_Y^2 \Sigma_x n^{-2}(N_{1Y}(x)$$

$$- pN_{0Y}(x)^2 + 2\delta_1\delta_2 h_X h_Y n^{-1}m^{-1}\Sigma_x (N_{1X}(x)$$

$$- pN_{0X}(x)(N_{1Y}(x) - pN_{0Y}(x).$$

D'où en remplaçant ces expressions dans $\Delta^*(h_X,h_Y)$ et en dérivant par rapport à h_X et h_Y.

on obtient

$$\frac{\partial\Delta^*(h_X,h_Y)}{\partial h_X} = \delta_1^2 m^{-3}N_X(m) + 2h_X\delta_1^2\{m^{-2}D_{2X}(m) - m^{-3}D_{1X}(m)\}$$

$$+ 2\delta_1\delta_2 h_Y n^{-1}m^{-1}\Sigma_x\{N_{1X}(x) - pN_{0X}(x)\}\{N_{1Y}(x) - pN_{0Y}(x)\}.$$

$$\frac{\partial\Delta^*(h_X,h_Y)}{\partial h_Y} = \delta_1^2 n^{-3}N_Y(n) + 2h_Y\delta_1^2\{n^{-2}D_{2Y}(n) - n^{-3}D_{1Y}(n)\}$$

$$+ 2\delta_1\delta_2 h_X n^{-1}m^{-1}\Sigma_x (N_{1X}(x) - pN_{0X}(x))(N_{1Y}(x) - pN_{0Y}(x)).$$

Ainsi, on a le système d'équations linéaires suivant

$$h_X 2\delta_1^2 m^{-2}F_X - 2\delta_1\delta_2 n^{-1}m^{-1}G_{XY}h_Y = \delta_1^2 m^{-3}N_X(m)$$

$$- h_X 2\delta_1\delta_2 n^{-1}m^{-1}G_{XY} + 2\delta_2^2 n^{-2}F_Y h_Y = \delta_2^2 n^{-3}N_Y(n),$$

où

$$F_X = m^{-1}D_{1X}(m) - D_{2X}(m) \quad \text{et} \quad F_Y = n^{-1}D_{1Y}(n) - D_{2Y}(n)$$

$$G_{XY} = \Sigma_x \{N_{1X}(x) - pN_{0X}(x)\}\{N_{1Y}(x) - pN_{0Y}(x)\}.$$

D'où, on obtient que

$$h_X = \frac{m^{-1}F_Y N_X(m) + \rho mn^{-2}N_Y(n)G_{XY}}{2\{F_X F_X - G_{XY}^2\}}$$

$$h_Y = \frac{n^{-1}F_X N_Y(m) + \rho^{-1}nm^{-2}N_X(n)G_{XY}}{2\{F_X F_X - G_{XY}^2\}}.$$

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249-6399