CLASSIFICATION OF COMPONENTS OF A MIXTURE

UDC 519.21

O. V. SUGAKOVA

ABSTRACT. We consider the problem of classification of individuals sampled from a mixture of several components with different probability distributions. To construct a classifier we use kernel estimators of the density of components in the mixture for a one-dimensional random variable $S_j^N(b) = \sum_{i=1}^d b_i \xi_j^{N,i}$ that is the projection of the vector of observations $\xi_j^N = (\xi_j^{N,1}, \xi_j^{N,2}, \dots, \xi_j^{N,d})$ to a nonrandom direction $b = (b_1, b_2, \dots, b_d)$. We obtain an estimator $\hat{b}$ for the best possible direction b . It is proved that the probability of error for the classifier based on $S(\hat{b})$ converges to the minimal probability of error among all possible classifiers.

1. INTRODUCTION

The classification problem is one of the main problems of statistical analysis. When solving the problem of classification one determines the class containing an individual ξ by using observations of its characteristics [1]. Consider the following setting of the problem.

Let a sample of N observations be given. Each observation of the sample may belong to one of M classes. Denote by $\text{ind}(j)$ the number of the class containing the observation j . The number $\text{ind}(j)$ is unknown, but the probabilities

$$w_{j,N}^k = P\{\text{ind}(j) = k\}$$

are known (we treat these probabilities as the concentrations of the component k for the observation j). In fact, the vector of characteristics $\xi_j^N = (\xi_j^{N,1}, \xi_j^{N,2}, \dots, \xi_j^{N,d}) \in \mathbf{R}^d$ of the individual j is observed. Denote by $h_k(x_1, x_2, \dots, x_d)$ the conditional probability density of ξ_j^N given $\text{ind}(j) = k$. Thus the density of ξ_j^N is

$$f_{\xi_j^N}(x_1, \dots, x_d) = \sum_{k=1}^M w_{j,N}^k h_k(x_1, \dots, x_d).$$

We assume that the weights $w_{j,N}^k$ are known, while the $h_k(x_1, x_2, \dots, x_d)$ are unknown. We also assume that the characteristics ξ_j^N are independent for fixed N .

Given a sample $X = (\xi_j^N, j = 1, \dots, N)$ we want to construct a classifier

$$g: \mathbf{R}^d \rightarrow \{1, \dots, M\}$$

that can give us the number $\text{ind}(O)$ of the class containing an individual O by using the characteristics ξ .

2000 *Mathematics Subject Classification.* Primary 62H30; Secondary 62G07.

Key words and phrases. Kernel estimators of the density, Bayes empirical classifier, estimates of components of a mixture.

According to the classical Bayes approach [2] the classifier (Bayes classifier) with the minimum probability of error is considered to be the best. To construct the Bayes classifier, an observer should know the distributions of the characteristics of the individuals belonging to different classes. Usually the densities are unknown and only a learning sample is available. Also, the number of the class is known for every member of this sample. Using the learning sample, one can construct estimators of unknown densities. Substituting these estimators to the Bayes classifier one obtains the so-called Bayes empirical classifier [2]. This procedure leads to the well-known discriminant analysis [3] in the case of Gaussian observations: the classification in this case is to compare several “discriminating” functions consisting of some linear combinations of observed characteristics.

The probability of error for Bayes empirical classifiers approaches the probability of error for the Bayes classifier in the case of nonparametric estimation of densities [2]. However, the expressions for the Bayes empirical estimators are complicated in the case of multidimensional observations and it is not easy to analyze them. Thus, for multidimensional observations another approach is used. Namely, the characteristics of objects are projected to some direction that reflects the properties of the individual in some sense (in other words, one chooses a linear combination of all observed variables). Such combinations are called aggregated indices in econometrics and social statistics or scales in psychometrics. The classification of observations in this case uses an appropriate aggregated index.

In this paper we consider the problem of constructing a “linear” index giving the minimal error of classification in the asymptotic sense. Our method is to project the observations to different directions and to construct Bayes empirical classifiers for every direction. We estimate the probability of error for every classifier; an index is called the best if the minimum of the probabilities of error is attained at this index. We show that the error of the Bayes empirical classification evaluated by using the indices of this type converges to the minimal probability of error among all linear indices. As a learning sample, one chooses in this case a sample from the mixture with varying concentrations [4, 5]. The classification for mixtures with varying concentrations is considered in [6].

2. SETTING OF THE PROBLEM

In what follows we consider “linear indices” of the form

$$(1) \quad S_j^N(b) = \sum_{i=1}^d b_i \xi_j^{N,i},$$

where $b = (b_1, b_2, \dots, b_d) \in \mathbf{R}^d$; the length of the nonrandom vector b is normalized to be that of a unit vector, that is,

$$(2) \quad b_1^2 + b_2^2 + \dots + b_d^2 = 1.$$

In fact, the index $S(b)$ is the projection of ξ to the direction b . We deal with the classifiers of the form $\tilde{g}_b(\xi) = g(S(b))$, where $g: \mathbf{R} \rightarrow \{1, \dots, M\}$ is an arbitrary measurable function. The probability of error of the classifier $\tilde{g}_b$ is

$$L_{\tilde{g}_b} = P\{\tilde{g}_b(\xi) \neq \text{ind}(O)\}.$$

It is known [2] that the minimum of $L_{\tilde{g}_b}$ among all possible g and for a fixed b is attained at the Bayes classifier

$$g^*(x) = \arg \max_i (p_i u_i(x)),$$

where $p_i = P\{\text{ind}(O) = i\}$ are concentrations of components in the mixture for the observation O and $u_i(x)$ are densities of random variables

$$\eta_i(b) = \sum_{j=1}^d \eta_i^j b_j,$$

where $\eta_i = (\eta_i^1, \eta_i^2, \dots, \eta_i^d)$ has the density h_i . The probability of error for this classifier is given by

$$L_g^* = L^*(b) = 1 - \int_{-\infty}^{\infty} \max_{1 \leq k \leq M} (p_k u_k(x)) dx.$$

The best index to construct the classifier is the projection of ξ to the direction

$$b^* = \arg \min_b L^*(b) = \arg \max_b \int_{-\infty}^{\infty} \max_{1 \leq k \leq M} (p_k u_k(x)) dx.$$

Since the true values of u_k are unknown, we substitute for them the kernel estimators constructed from the sample $(S_j(b))_{j=1}^N$. These estimators are given by

$$(3) \quad \hat{u}_i^N(b, x) = \frac{1}{N\sigma_N} \sum_{j=1}^N a_{j,N}^i K\left(\frac{x - S_j(b)}{\sigma_N}\right)$$

(see [4]), where

$$a_{j,N}^i = \frac{1}{\det \Gamma_N} \sum_{k=1}^M (-1)^{i+k} \gamma_{ik}^N w_{j,N}^k, \quad i, j = 1, \dots, M,$$

$\Gamma_N = (\langle w^k, w^l \rangle)_{k,l=1}^M$ is the Gram matrix of the system of functions $w_{j,N}^k$ for the scalar product

$$\langle w^k, w^l \rangle_N = \frac{1}{N} \sum_{j=1}^N w_{j,N}^k w_{j,N}^l,$$

and γ_{jk}^N is the (j, k) -minor of the matrix Γ_N . Note that $a_{j,N}^i$ are such that

$$(4) \quad \langle a^i, w^k \rangle_N = \frac{1}{N} \sum_{j=1}^N a_{j,N}^i w_{j,N}^k = \chi\{k = i\}.$$

Here $\chi\{A\}$ is the indicator of the event A . We assume that σ_N is a smooth parameter; $\sigma_N \rightarrow 0$ as $N \rightarrow \infty$, but $N\sigma_N \rightarrow \infty$ as $N \rightarrow \infty$; the function K , called the kernel of the estimator, is a density of some random variable η . It is proved in [7] that estimator (3) is asymptotically unbiased and consistent.

The probability of the error classification $L^*(b)$ is estimated by

$$\hat{L}^N(b) = 1 - \int_{-\infty}^{\infty} \max_{1 \leq k \leq M} (p_k \hat{u}_k^N(b, x)) dx,$$

while b^* is estimated by

$$\hat{b} = \arg \min \hat{L}^N(b).$$

The “quasi-Bayes classifier” is constructed from observations ξ as follows:

$$\hat{g}(\hat{b}, x) = \arg \max_{1 \leq i \leq M} (p_i \hat{u}_i^N(\hat{b}, x)).$$

We use the conditional probability of error given a learning sample

$$P\{\hat{g}(\hat{b}, x) \neq \text{ind}(O) \mid X\} = L_{\hat{g}}(\hat{b}) = 1 - \int_{-\infty}^{\infty} p_{\hat{g}(\hat{b}, x)} u_{\hat{g}(\hat{b}, x)} dx$$

to measure the quality of the classifier (see [2]).

3. MAIN RESULTS

Assume that

$$(i) \quad K(x) \leq a, \quad \int_{-\infty}^{\infty} |x|K(x) dx < \infty$$

for some constant a . The collection of sets $S = \{A\}$, where

$$(ii) \quad A = \left\{ y: K \left(\frac{x - \sum_{i=1}^d b_i y_i}{\sigma_N} \right) \geq c \right\}$$

(c is a certain constant), is a Vapnik–Chervonenkis class [1, Chapter X], [5, p. 55].

The partial derivatives of the densities

$$h_i(x_1, \dots, x_d)$$

exist and are bounded for all $i = 1, \dots, M$:

$$(iii) \quad \left| \frac{\partial h_i}{\partial x_s}(x_1, \dots, x_d) \right| \leq c_{is}, \quad s = 1, \dots, d.$$

The weight coefficients $a_{j,N}^i$ are such that

$$(iv) \quad |a_{j,N}^i| \leq \tilde{a},$$

where $\tilde{a}$ is a constant. Furthermore,

$$(v) \quad \frac{1}{\sigma_N} = o(N^{1/4}), \quad N \rightarrow \infty.$$

Theorem 3.1. *If conditions (i)–(v) hold, then*

$$\mathbb{E} L_{\hat{g}}(\hat{b}) \rightarrow L_g^*(b^*) \quad \text{as } N \rightarrow \infty.$$

Corollary 3.1. *If all the assumptions of Theorem 3.1 hold, then*

$$L_{\hat{g}}(\hat{b}) \rightarrow L_g^*(b^*) \quad \text{as } N \rightarrow \infty$$

in probability.

First we prove the following auxiliary results.

Lemma 3.1. *If conditions (i)–(v) hold, then*

$$\mathbb{E} \sup_{b,x} |\hat{u}_k^N(b, x) - u_k(b, x)| \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

for all $k = 1, \dots, M$.

Proof of Lemma 3.1. We have

$$(5) \quad \begin{aligned} & \sup_{b,x} |\hat{u}_k^N(b, x) - u_k(b, x)| \\ & \leq \sup_{b,x} \left| \frac{1}{N\sigma_N} \sum_{j=1}^N a_{j,N}^k K \left(\frac{x - S_j^N(b)}{\sigma_N} \right) - \frac{1}{N\sigma_N} \mathbb{E} \sum_{j=1}^N a_{j,N}^k K \left(\frac{x - S_j^N(b)}{\sigma_N} \right) \right| \\ & \quad + \sup_{b,x} \left| \frac{1}{N\sigma_N} \mathbb{E} \sum_{j=1}^N a_{j,N}^k K \left(\frac{x - S_j^N(b)}{\sigma_N} \right) - u_k(b, x) \right| \\ & = S_1^N + S_2^N. \end{aligned}$$

Consider the term S_1^N . Define the weighted empirical measure by

$$(6) \quad \hat{\mu}_N(A) = \frac{1}{N} \sum_{j=1}^N a_{j,N}^k \chi\{\xi_j^N \in A\}, \quad \bar{\mu}_N(A) = \mathbf{E} \hat{\mu}_N(A).$$

Let $\mathbf{P}(dy)$ be a probability measure. Using the idea of the proof in [1, p. 287] we get

$$(7) \quad \begin{aligned} S_1^N &= \sup_{b,x} \left| \frac{1}{\sigma_N} \int_{\mathbf{R}^d} K \left(\frac{x - \sum_{i=1}^d b_i y_i}{\sigma_N} \right) \mathbf{P}(dy) \right. \\ &\quad \left. - \frac{1}{\sigma_N} \int_{\mathbf{R}^d} K \left(\frac{x - \sum_{i=1}^d b_i y_i}{\sigma_N} \right) \hat{\mu}_N(dy) \right| \\ &\leq \frac{a}{\sigma_N} \sup_{A \in S} |\mathbf{P}(A) - \hat{\mu}_N(A)| \end{aligned}$$

in view of condition (i), where S is the collection of sets mentioned in condition (ii). Furthermore, we note that

$$\sup_{A \in S} |\mathbf{P}(A) - \hat{\mu}_N(A)| \leq \sup_{1 \leq j \leq N} |a_{j,N}^i| + 1 \leq \tilde{a} + 1.$$

Let

$$\tilde{C}_N^k = \sup_{1 \leq j \leq N} |a_{j,N}^i| + \sup_{1 \leq j \leq N} a_{j,N}^i - \inf_{1 \leq j \leq N} a_{j,N}^i.$$

We need the following result.

Theorem 3.2 ([5, p. 60]). *Let $a_{j,N}$ be arbitrary weight coefficients, let S be a VC-class, and denote by $g^G(N)$ its growth function. Let $\hat{\mu}_N(A)$ and $\bar{\mu}_N(A)$ be defined by (6). Then*

$$\mathbf{P} \left\{ \sup \frac{|\hat{\mu}_N(A) - \bar{\mu}_N(A)|}{\tilde{C}_N^k} \geq \lambda \right\} \leq M \left(6Ng^G(2N) \exp \left(-\frac{\lambda^2 N}{32M^2} \right) + 2 \exp \left(-\frac{\lambda^2 N}{8M^2} \right) \right)$$

for all $\lambda > 2M/N$.

Now we come back to the proof of Lemma 3.1:

$$\mathbf{E} \sup_{A \in S} \left| \frac{\mathbf{P}(A) - \hat{\mu}_N(A)}{\tilde{C}_N^k} \right| \leq \int_0^{\frac{\tilde{a}+1}{\tilde{C}_N^k}} \mathbf{P} \left\{ \sup_{A \in S} \left| \frac{\mathbf{P}(A) - \hat{\mu}_N(A)}{\tilde{C}_N^k} \right| \geq \lambda \right\} d\lambda.$$

Since $\tilde{C}_N^k$ is bounded from below, we assume without loss of generality that

$$\frac{2M}{\sqrt[4]{N}} < \frac{\tilde{a}+1}{\tilde{C}_N^k}.$$

We split the integral into two parts and use Theorem 2 to estimate the second part:

$$(8) \quad \begin{aligned} &\mathbf{E} \sup_{A \in S} \left| \frac{\mathbf{P}(A) - \hat{\mu}_N(A)}{\tilde{C}_N^k} \right| \\ &\leq \int_0^{\frac{2M}{\sqrt[4]{N}}} d\lambda + \int_{\frac{2M}{\sqrt[4]{N}}}^{\frac{\tilde{a}+1}{\tilde{C}_N^k}} \left(6MN g^S(2N) \exp \left(-\frac{\lambda^2 N}{32M^2} \right) + 2 \exp \left(-\frac{\lambda^2 N}{8M^2} \right) \right) d\lambda \\ &\leq 2MN^{-1/4} \\ &\quad + \left(6MN g^S(2N) \exp \left(-\frac{1}{8} \sqrt{N} \right) + 2 \exp \left(-\frac{1}{2} \sqrt{N} \right) \right) \left(\frac{\tilde{a}+1}{\tilde{C}_N^k} - \frac{2M}{\sqrt[4]{N}} \right). \end{aligned}$$

Estimates (7) and (8) imply that

$$(9) \quad \begin{aligned} \mathbb{E} S_1^N &\leq O\left(\sigma_N^{-1} N^{-1/4}\right) + O\left(\sigma_N^{-1} N g^S(2N) \exp\left(-\frac{1}{8}\sqrt{N}\right)\right) \\ &\quad + O\left(\sigma_N^{-1} \exp\left(-\frac{1}{2}\sqrt{N}\right)\right). \end{aligned}$$

The right-hand side of (9) approaches zero as $N \rightarrow \infty$. Indeed, the first term approaches zero by condition (v), while the second and third terms tend to zero, since $g^S(N)$, as the growth function of the Vapnik–Chervonenkis class S , grows as a power function and σ_N^{-1} and the exponents approach zero quicker than power functions tend to infinity. Thus

$$(10) \quad \mathbb{E} S_1^N = o(1) \quad \text{as } N \rightarrow \infty.$$

Now we estimate the term S_2^N :

$$\begin{aligned} \frac{1}{N\sigma_N} \mathbb{E} \sum_{j=1}^N a_{j,N}^k K\left(\frac{x - S_j^N(b)}{\sigma_N}\right) &= \frac{1}{N\sigma_N} \sum_{j=1}^N a_{j,N}^k \mathbb{E} K\left(\frac{x - S_j^N(b)}{\sigma_N}\right) \\ &= \frac{1}{N\sigma_N} \sum_{j=1}^N a_{j,N}^k \sum_{k=1}^M w_{j,N}^k \int_{-\infty}^{\infty} K\left(\frac{x-y}{\sigma_N}\right) u_k(b, y) dy \\ &= \int_{-\infty}^{\infty} K(z) u_k(b, x - \sigma_N z) dz. \end{aligned}$$

Here we applied equality (4). Thus S_2^N can be represented in the following form:

$$S_2^N = |\mathbb{E} u_k(b, x - \sigma_N \eta) - u_k(b, X)|,$$

where η is a random variable with the density $K(x)$. Due to condition (2), at least one of the components of the vector $b = (b_1, b_2, \dots, b_d)$ is nonzero; let $b_d \neq 0$. Then the density of the random variable μ_k can be represented as follows:

$$u_k(b, x) = \frac{1}{b_d} \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} h_k\left(x_1, \dots, x_{d-1}, \frac{1}{b_d} \left(x - \sum_{i=1}^{d-1} b_i x_i\right)\right) dx_{d-1}.$$

For brevity we write $h_k(x_d)$ instead of $h_k(x_1, \dots, x_{d-1}, x_d)$. Choose a subsequence $t = t_N \rightarrow \infty$, $N \rightarrow \infty$. We have

$$\begin{aligned} S_2^N &= \frac{1}{|b_d|} \mathbb{E} \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} \left| h_k\left(\frac{1}{b_d} \left(x - \sum_{i=1}^{d-1} b_i x_i - \sigma_N \eta\right)\right) \right. \\ &\quad \left. - h_k\left(\frac{1}{b_d} \left(x - \sum_{i=1}^{d-1} b_i x_i\right)\right) \right| dx_{d-1} \\ &\leq \frac{1}{|b_d|} \mathbb{E} \int_{x_1^2 + \dots + x_{d-1}^2 \leq t^2} dx_1 \cdots dx_{d-1} \left| h_k\left(\frac{1}{b_d} \left(x - \sum_{i=1}^{d-1} b_i x_i - \sigma_N \eta\right)\right) \right. \\ &\quad \left. - h_k\left(\frac{1}{b_d} \left(x - \sum_{i=1}^{d-1} b_i x_i\right)\right) \right| \\ &\quad + \frac{1}{|b_d|} \mathbb{E} \int_{x_1^2 + \dots + x_{d-1}^2 > t^2} dx_1 \cdots dx_{d-1} h_k\left(\frac{1}{b_d} \left(x - \sum_{i=1}^{d-1} b_i x_i - \sigma_N \eta\right)\right) \\ &\quad + \frac{1}{|b_d|} \int_{x_1^2 + \dots + x_{d-1}^2 > t^2} dx_1 \cdots dx_{d-1} h_k\left(\frac{1}{b_d} \left(x - \sum_{i=1}^{d-1} b_i x_i\right)\right) \\ &= A + B + C. \end{aligned}$$

To estimate the term A we apply the mean value theorem:

$$(11) \quad A \leq \mathbf{E} \frac{\sigma_N |\eta|}{b_d^2} \int_{x_1^2 + \dots + x_{d-1}^2 \leq t^2} dx_1 \cdots dx_{d-1} \left| \frac{\partial h_k}{\partial x_d} \left(x_1, \dots, x_{d-1}, \frac{s - \sum_{i=1}^{d-1} b_i x_i}{b_d} \right) \right|.$$

Here s is a random variable whose lower and upper bounds are

$$\frac{1}{b_d} \left(x - \sum_{i=1}^{d-1} b_i x_i - \sigma_N \eta \right) \quad \text{and} \quad \frac{1}{b_d} \left(x - \sum_{i=1}^{d-1} b_i x_i \right),$$

respectively. Changing the variable $x_1 = v b_d^2$ in the integral on the right-hand side of (12) and using condition (iii) on the boundedness of partial derivatives we obtain

$$(12) \quad \begin{aligned} A &\leq \sigma_N \mathbf{E} |\eta| \int_{v^2/b_d^2 + \dots + x_{d-1}^2 \leq t^2} dv \cdots dx_{d-1} \\ &\quad \times \left| \frac{\partial h_k}{\partial x_d} \left(v b_d^2, \dots, x_{d-1}, \frac{1}{b_d} \left(s - \sum_{i=1}^{d-1} b_i x_i \right) \right) \right| \\ &\leq \sigma_N c_{kd} \mathbf{E} |\eta| \text{mes} \left\{ (v, x_2, \dots, x_{d-1}) \in \mathbf{R}^{d-1} : \frac{v^2}{b_d^2} + x_2^2 + \dots + x_{d-1}^2 \leq t^2 \right\} \\ &\leq \sigma_N c_{kd} \mathbf{E} |\eta| \text{mes} \left\{ (v, x_2, \dots, x_{d-1}) \in \mathbf{R}^{d-1} : v^2 + x_2^2 + \dots + x_{d-1}^2 \leq t^2 \right\} \\ &\leq \sigma_N c_{kd} \mathbf{E} |\eta| (2t)^{d-1}. \end{aligned}$$

Now we estimate B :

$$B = \frac{1}{|b_d|} \mathbf{E} \int_{x_1^2 + \dots + x_{d-1}^2 > t^2; x_d = x - \sum_{i=1}^{d-1} b_i x_i - \sigma_N \eta} dx_1 \cdots dx_{d-1} h_k \left(x_1, \dots, x_{d-1}, \frac{x_d}{b_d} \right).$$

Since the integrand is nonnegative, extending the domain of integration we increase the value of the integral, that is,

$$B \leq \frac{1}{|b_d|} \int_{x_1^2 + \dots + x_{d-1}^2 > t^2; -\infty < x_d < \infty} dx_1 \cdots dx_d h_k \left(x_1, \dots, x_{d-1}, \frac{x_d}{b_d} \right).$$

Now we change the variable $v_d = x_d/b_d$ and get

$$(13) \quad B \leq \int_{x_1^2 + \dots + x_{d-1}^2 > t^2; -\infty < v_d < \infty} dx_1 \cdots dv_d h_k(x_1, \dots, v_d).$$

The term C is estimated analogously to B :

$$(14) \quad C \leq \int_{x_1^2 + \dots + x_{d-1}^2 > t^2; -\infty < v_d < \infty} dx_1 \cdots dv_d h_k(x_1, \dots, v_d).$$

Thus (11)–(15) imply that

$$(15) \quad S_2^N \leq c_{kd} \mathbf{E} |\eta| \sigma_N (2t)^{d-1} + 2 \int_{x_1^2 + \dots + x_{d-1}^2 > t^2; -\infty < v_d < \infty} dx_1 \cdots dv_d h_k(x_1, \dots, v_d).$$

One can choose $t = t_N$ such that

$$\sigma_N (2t)^{d-1} \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

The right-hand side of (16) tends to 0 as $N \rightarrow \infty$ for such a choice of $t = t_N$, whence

$$(16) \quad S_2^N \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

uniformly with respect to x and b . Combining (5), (10), and (17) we complete the proof of Lemma 3.1. $\square$

Remark 3.1. If $K(x)$ is a piecewise monotone function with a finite number of change points, then the family of events defined in (ii) is a Vapnik–Chervonenkis class.

Indeed, there is a finite number G of intervals $[d_l, d_{l+1}]$ such that

$$\left\{ y: K\left(\frac{x - \sum_{i=1}^d b_i y_i}{\sigma_N}\right) \geq c \right\} = \bigcup_{l=1}^G \left\{ x - \sigma_N d_{l+1} \leq \sum_{i=1}^d b_i y_i \leq x - \sigma_N d_l \right\}.$$

Every set on the right-hand side of the latter equality is a layer between two hyperplanes; the union of a finite number of such sets is a Vapnik–Chervonenkis class [1, pp. 232, 233].

Lemma 3.2. *If conditions (i)–(iv) hold, then*

$$\mathbf{E} \sup_b \left| \hat{L}_g^N(b) - L_g^*(b) \right| \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

Proof of Lemma 3.2. We need the following Györfi inequality proved in [2].

Theorem 3.3 ([2]).

$$(17) \quad \hat{L}_g^N - L_g^* \leq 2 \sum_{k=1}^M \int_{-\infty}^{\infty} |u_k(x) - \hat{u}_k^N(x)| dx.$$

Let D_N be a sequence such that $D_N \rightarrow \infty$ as $N \rightarrow \infty$. Then

$$(18) \quad \begin{aligned} \mathbf{E} \int_{-\infty}^{\infty} |\hat{u}_k^N(b, x) - u_k(b, x)| dx &\leq \mathbf{E} \int_{-D_N}^{D_N} |\hat{u}_k^N(b, x) - u_k(b, x)| dx \\ &+ \mathbf{E} \left(\int_{D_N}^{\infty} \hat{u}_k^N(b, x) dx + \int_{-\infty}^{-D_N} \hat{u}_k^N(b, x) dx \right) \\ &+ \left(\int_{D_N}^{\infty} u_k(b, x) dx + \int_{-\infty}^{-D_N} u_k(b, x) dx \right) \\ &= A + B + C. \end{aligned}$$

We estimate each term on the right-hand side of (19) separately. We have

$$(19) \quad A \leq 2D_N \mathbf{E} \sup_{b, x} |\hat{u}_k^N(b, x) - u_k(b, x)|.$$

According to Lemma 3.1,

$$\mathbf{E} \sup_{b, x} |\hat{u}_k^N(b, x) - u_k(b, x)| \rightarrow 0, \quad N \rightarrow \infty,$$

whence it follows that there is a sequence $D_N \rightarrow \infty$ such that the right-hand side of (20) tends to zero.

Now we turn to the term C . By the Cauchy–Bunyakovskiĭ inequality and (2),

$$(20) \quad -\sqrt{\sum_{i=1}^d (\eta_k^i)^2} \leq \sum_{i=1}^d b_i \eta_k^i \leq \sqrt{\sum_{i=1}^d (\eta_k^i)^2}.$$

Denote by $F(x)$ the distribution function of the random variable μ_k . Using inequality (21) we get

$$\begin{aligned} \int_{D_N}^{\infty} u_k(b, x) dx &= 1 - F(D_N) \leq \mathbf{P} \left\{ \sqrt{\sum_{i=1}^d (\eta_k^i)^2} \geq D_N \right\}; \\ \int_{-\infty}^{-D_N} u_k(b, x) dx &= F(-D_N) \leq \mathbf{P} \left\{ -\sqrt{\sum_{i=1}^d (\eta_k^i)^2} \leq -D_N \right\} \\ &= \mathbf{P} \left\{ \sqrt{\sum_{i=1}^d (\eta_k^i)^2} > D_N \right\}. \end{aligned}$$

These bounds imply that

$$(21) \quad C \leq 2 \mathbf{P} \left\{ \sum_{i=1}^d (\eta_k^i)^2 \geq D_N^2 \right\} \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

The term B is estimated similarly. Without loss of generality, we assume that $\sigma_N \leq 1$. By the Cauchy–Bunyakovskiĭ inequality

$$\left(\sum_{i=1}^d b_i \eta_k^i + \sigma_N \eta \right)^2 \leq (1 + \sigma_N^2) \left(\sum_{i=1}^d (\eta_k^i)^2 + \eta^2 \right).$$

Thus

$$\begin{aligned} & -\sqrt{\sum_{i=1}^d (\eta_k^i)^2 + \eta^2} \leq \sum_{i=1}^d b_i \eta_k^i + \sigma_N \eta \leq \sqrt{\sum_{i=1}^d (\eta_k^i)^2 + \eta^2}, \\ \mathbf{E} \int_{D_N}^{\infty} \hat{u}_k^N(b, x) dx &= \frac{1}{N \sigma_N} \mathbf{E} \int_{D_N}^{\infty} \sum_{j=1}^N a_{j,N}^k K \left(\frac{x - S_j^N(b)}{\sigma_N} \right) dx \\ &= \int_{D_N}^{\infty} \mathbf{E} u_k(b, x - \sigma_N \eta) dx = \mathbf{E}(1 - F(D_N - \sigma_N \eta)) \\ &= \mathbf{P} \left\{ \sum_{i=1}^d b_i \eta_k^i + \sigma_N \eta \geq D_N \right\} \\ (22) \quad & \leq \mathbf{P} \left\{ 2 \sqrt{\sum_{i=1}^d (\eta_k^i)^2 + \eta^2} \geq D_N \right\}, \\ \mathbf{E} \int_{-\infty}^{-D_N} \hat{u}_k^N(b, x) dx &= \mathbf{E} F(-D_N - \sigma_N \eta) \leq \mathbf{P} \left\{ -2 \sqrt{\sum_{i=1}^d (\eta_k^i)^2 + \eta^2} < -D_N \right\} \\ &= \mathbf{P} \left\{ 2 \sqrt{\sum_{i=1}^d (\eta_k^i)^2 + \eta^2} > D_N \right\}. \end{aligned}$$

Estimates (23) imply that

$$(23) \quad B \leq 2 \mathbf{P} \left\{ \sum_{i=1}^d (\eta_k^i)^2 + \eta^2 \geq \frac{D_N^2}{4} \right\} \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

Combining (20), (22), (24), and (19) we complete the proof of Lemma 3.2 by the Györfi inequality. $\square$

Proof of Theorem 3.1. We have

$$(24) \quad L_{\hat{g}}(\hat{b}) - L_g^*(b^*) = (L_{\hat{g}}(\hat{b}) - \hat{L}_g^N(\hat{b})) + (\hat{L}_g^N(\hat{b}) - \hat{L}_g^N(b^*)) + \hat{L}_g^N(b^*) - L_g^*(b^*).$$

Since

$$\hat{b} = \arg \min_b \hat{L}_g^N(b)$$

and the classifier is such that

$$L_{\hat{g}}(\hat{b}) \leq \hat{L}_g^N(\hat{b}),$$

the first two terms in (25) are negative.

Now (25) implies that

$$(25) \quad 0 \leq L_{\hat{g}}(\hat{b}) - L_g^*(b^*) \leq \left| \hat{L}_g^N(b^*) - L_g^*(b^*) \right| \leq \sup_b \left| \hat{L}_g^N(b) - L_g^*(b) \right|.$$

By Lemma 3.2 the expectation of the right-hand side of (26) tends to 0 as $N \rightarrow \infty$. Theorem 3.1 is proved. $\square$

The corollary follows from Theorem 1 and Chebyshev's inequality.

4. CONCLUDING REMARKS

In this paper, we constructed a quasi-Bayes classifier by using the linear indices $S(b)$ and found a consistent estimator $\hat{b}$ for a direction b ; the probability of error classification for $S(\hat{b})$ converges to the minimal error. The rate of convergence of these estimators will be considered elsewhere.

BIBLIOGRAPHY

1. V. N. Vapnik and A. Ya. Chervonenkis, *Theory of Pattern Recognition*, "Nauka", Moscow, 1974. (Russian) MR0474638 (57:14274)
2. L. Devroye and L. Györfi, *Nonparametric Density Estimation*, Wiley, New York, 1985. MR0780746 (86i:62065)
3. A. Afifi and S. Azen, *Statistical Analysis: A Computer Oriented Approach*, Academic Press, New York, 1972.
4. R. E. Maïboroda, *Nonparametric Statistics of Nonhomogeneous Observations*, Doctoral dissertation, Kyiv, 1994. (Ukrainian)
5. ———, *Statistical Analysis of Mixtures. A Course of Lectures*, Kyiv University, Kyiv, 2004. (Ukrainian)
6. Yu. O. Ivan'ko and R. E. Maïboroda, *Exponential estimates for the empirical Bayes risk in the classification of a mixture with varying concentrations*, Ukr. Mat. Zh. **54** (2002), no. 10, 1421–1428; English transl. in Ukrainian Math. J. **54** (2002), no. 10, 1722–1731. MR2015493 (2004g:62022)
7. O. V. Sugakova, *Asymptotics of a kernel estimate for the density of a distribution constructed from observations of a mixture with varying concentrations*, Teor. Imovirnost. Matem. Stat. **59** (1998), 156–166; English transl. in Theory Probab. Math. Statist. **59** (1999), 161–171. MR1793776

DEPARTMENT OF MATHEMATICS AND THEORETICAL RADIOPHYSICS, FACULTY FOR RADIOPHYSICS, NATIONAL TARAS SHEVCHENKO UNIVERSITY, ACADEMICIAN GLUSHKOV AVENUE 6, KYIV 03127, UKRAINE
E-mail address: `sugak@univ.kiev.ua`

Received 2/APR/2004

Translated by S. KVASKO