Теорія Ймовір. та Матем. Статист.
Вип. 74, 2006

Theor. Probability and Math. Statist.
No. 74, 2007, Pages 37–47
S 0094-9000(07)00696-5
Article electronically published on June 25, 2007

THE ASYMPTOTIC BEHAVIOR OF THRESHOLD-BASED CLASSIFICATION RULES CONSTRUCTED FROM A SAMPLE FROM A MIXTURE WITH VARYING CONCENTRATIONS

UDC 519.21

YU. IVAN'KO AND R. MAĬBORODA

Abstract. We consider a problem on finding the best threshold-based classification rule constructed from a sample from a mixture with varying concentrations. We show that the rate of convergence of the minimal empirical risk estimators to the optimal threshold is of order $N^{-1/3}$ for smooth distributions, while the rate of convergence of the Bayes empirical estimators is of order $N^{-2/5}$ where N is the size of a sample.

1. Introduction

We consider the problem of the classification of an object O from observations after its numerical characteristic $\xi = \xi(O)$. We assume that the object may belong to one of the two classes and restrict the consideration to the case of threshold-based classification rules of the form

$$g_t(\xi) = \begin{cases} 1 & \text{if } \xi \le t, \\ 2 & \text{if } \xi > t. \end{cases} \tag{1}$$

According to this rule, an object is classified to belong to the first class if its characteristic does not exceed a threshold t; otherwise, an object is classified to belong to the second class. A simple example where one can apply this classification rule is to determine whether a person (object) is ill (belongs to the second class) if the temperature of its body (of the characteristic ξ) exceeds 37°C (the threshold t).

It is common to define the best (Bayes) threshold as $t = t^B$ for which g_t has the minimal probability of error. When determining the best threshold, one faces the problem of estimating the threshold by using a learning sample. Widely used methods to estimate t^B, the Bayes empirical classification [2, 11] and minimization of the empirical risk, are based on the assumption that there is a sample, called the learning sample, whose members are classified correctly. The minimal empirical risk method is generalized in this paper to the case where the learning sample is obtained from a sample with varying concentrations [4]. We also study the asymptotic behavior of both methods mentioned above.

The minimal empirical risk method is a comparatively simple technique based on the empirical distribution functions as estimators of the corresponding true distributions. These estimators have a "nice" rate of convergence of order $N^{-1/2}$ where N is the size of a sample [1]. The empirical Bayes classification rule uses the estimators of the distribution

2000 *Mathematics Subject Classification.* Primary 62H30; Secondary 62G07.
Key words and phrases. Minimization of the empirical risk, kernel estimators of densities, Bayes empirical classification rule, estimates of components of a mixture, mixtures with varying concentrations.

densities that have a worse rate of convergence. In the smooth case (considered in this paper), the rate of convergence is of order $N^{-2/5}$ [7]. The first impression is that the minimal empirical risk estimators of the threshold t^B are asymptotically better than the empirical Bayes estimators. However we will show that the rate of convergence of minimal empirical risk estimators is of order $N^{-1/3}$, while that of empirical Bayes classification estimators is of order $N^{-2/5}$.

If the members of a learning sample are classified correctly, the rate of convergence of order $N^{-1/3}$ is obtained in [9] for minimal empirical risk estimators. A rate of order $N^{-1/3}$ is usual for problems of maximizations of nonsmooth functionals of empirical distribution functions [8, 10]. The empirical Bayes classification allows one to account for the smoothness of the distribution functions and to smooth the estimators of densities. This explains why this method has a better rate of convergence.

2. The setting of the problem

A numerical characteristic $\xi = \xi(O)$ is observed for an object O. The object may belong to one of the two prescribed classes. An unknown number of a class containing O is denoted by $\operatorname{ind}(O)$. The a priori probabilities

$$p_i = \mathsf{P}(\operatorname{ind}(O) = i), \qquad i = 1, 2,$$

are assumed to be known. The characteristic ξ is assumed to be random, and its distribution depends on $\operatorname{ind}(O)$:

$$\mathsf{P}(\xi(O) < x \mid \operatorname{ind}(O) = i) = H_i(x).$$

The distributions H_i are unknown; however they have densities h_i with respect to the Lebesgue measure and they are continuous functions.

A classification rule is a function $g\colon \mathbb{R} \to \{1, 2\}$ that assigns a value to $\operatorname{ind}(O)$ by using the characteristic ξ. In general, a classification rule is defined as a general measurable function, but we restrict the consideration in this paper to the so-called threshold-based classification rules of the form (1). The family of threshold-based classification rules is denoted by $\mathcal{G} = \{g_t\colon t \in \mathbb{R}\}$. The probability of error of such a classification rule is given by

$$\begin{aligned} L(g_t) = L(t) &= \mathsf{P}\{g_t(\xi(O)) \neq \operatorname{ind}(O)\} \\ &= \sum_{i=1}^{2} \mathsf{P}\{\operatorname{ind}(O) = i\}\, \mathsf{P}\{g_t(\xi(O)) = 3 - i \mid \operatorname{ind}(O) = i\} = p_1(1 - H_1(t)) + p_2 H_2(t). \end{aligned}$$

A classification rule $g^B \in \mathcal{G}$ is called a Bayes classification rule in the class $\mathcal{G}$ if $L(g)$ attains its minimum at g^B,

$$g^B = \operatorname*{argmin}_{g \in \mathcal{G}} L(g).$$

The threshold t^B for a Bayes classification rule is called the Bayes threshold, $g^B = g_{t^B}$,

$$t^B = \operatorname*{argmin}_{t \in \mathbb{R}} L(t). \tag{2}$$

In what follows, $\operatorname{argmin}_{x \in U} f(x)$ denotes any $x_* \in U$ such that $f(x_*) = \inf_{x \in U} f(x)$. If a function f is random, we additionally assume that x_* is a random variable (in other words, x_* is a measurable function on the main probability space). Since $H_i'(t) = h_i(t)$, t^B is a solution of the equation

$$L'(t) = -p_1 h_1(t) + p_2 h_2(t) = 0. \tag{3}$$

The functions H_i (and, of course, h_i) are assumed to be unknown. One can estimate these functions from the data $\Xi_N = \{\xi_{j:N}\}_{j=1}^N$ being a sample from a mixture with varying concentrations where $\xi_{j:N}$ are independent if N is fixed and

$$\mathsf{P}\{\xi_{j:N} < x\} = w_{j:N} H_1(x) + (1 - w_{j:N}) H_2(x).$$

Here $w_{j:N}$ is a known concentration in the mixture of objects of the first class at the moment when an observation j is made [4].

To estimate the distribution functions H_i, we use weighted empirical distribution functions

$$H_i^N(x) = \frac{1}{N} \sum_{j=1}^N a_{j:N}^i \mathbb{1}\{\xi_j < x\},$$

where $\mathbb{1}\{A\}$ is the indicator of an event A and $a_{j:N}^i$ are known weight coefficients:

$$a_{j:N}^1 = \frac{1}{\Delta_N} \left((1 - s_N^1) w_{j:N} + (s_N^2 - s_N^1) \right), \qquad a_{j:N}^2 = \frac{1}{\Delta_N} (s_N^2 - s_N^1 w_{j:N}),$$

$$s_N^k = \frac{1}{N} \sum_{j=1}^N (w_{j:N})^k, \quad k = 1, 2, \ldots, \qquad \Delta_N = s_N^2 - (s_N^1)^2$$

(see Example 1 of Section 2.2 in [4]). One can apply kernel estimators to estimate the densities of distributions h_i:

$$h_i^N(x) = \frac{1}{N \kappa_N} \sum_{j=1}^N a_{j:N}^i K\left(\frac{x - \xi_{j:N}}{\kappa_N} \right),$$

where K is a kernel (the density of some probability distribution) and $\kappa_N > 0$ is a smoothing parameter [6, 7].

Starting from (2) and (3), one can apply two approaches to estimate t^B. The minimal empirical risk estimator is defined as

$$\hat{t}^{MER} = \operatorname*{argmin}_{t \in \mathbb{R}} L_N(t), \tag{4}$$

where $L_N(t) = p_1(1 - \hat{H}_1^N(t)) + p_2 H_2^N(t)$ is the empirical risk of the classification rule g_t.

The empirical Bayes estimator is constructed as follows. First, one determines the set T_N of all solutions of the equation

$$-p_1 \hat{h}_1^N(t) + p_2 \hat{h}_2^N(t) = 0.$$

Second, one chooses

$$\hat{t}_N^{EBC} = \operatorname*{argmin}_{t \in T_N} L_N(t) \tag{5}$$

as an estimator for t^B.

Remark 2.1. Estimators defined by (4) and (5) do not necessarily exist or may not be unique. The assumptions imposed on the model below imply that the probability of the event that the estimators do not exist tends to 0 as $N \to \infty$. Thus one is free to choose the values of estimators at points where the minimums do not exist, since this does not change the asymptotic behavior of estimators in the sense of the weak convergence.

Analogously, the assumptions of Theorem 3.3 below imply that the probability of the event that the estimator t_N^{EBC} is not uniquely defined tends to 0. The minimum in (4) always is attained at an infinite set of points (on an interval, as a rule). Nevertheless all points of minimum have the same asymptotic behavior under the assumptions of Theorem 3.3. Therefore one can choose any of them in the definition of the empirical Bayes classification estimator.

3. Main results

In what follows we assume that

(A) the threshold t^B defined by (2) exists, is a unique point of the global minimum of $L(t)$, and
$$L(t^B) < \min(p_1, p_2).$$

This inequality excludes the Bayes classification rules that assign the same class to any observation independently of ξ.

(B_k) The limits $s^i = \lim_{N\to\infty} s^i_N$, $i = 1, 2, \dots, k$, exist and $\Delta = s^2 - (s^i)^2 > 0$.

Theorem 3.1. *Let conditions* (A) *and* (B_2) *hold and let H_i be continuous functions on $\mathbb{R}$. Then $\hat{t}^{MER}_N \to t^B$ in probability as $N \to \infty$.*

Theorem 3.2. *Let conditions* (A) *and* (B_2) *hold. Assume that the densities h_i exist and are continuous, $\kappa_N \to 0$ as $N\kappa_N \to \infty$, K is a continuous function, and*
$$d^2 \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} K^2(t)\,dt < \infty.$$
Then $\hat{t}^{EBC} \to t^B$ in probability.

Let the densities h_i exist and be s times continuously differentiable in some neighborhood of the point t^B. Put
$$f_s(t) = (-1)^s \left(p_1 \frac{d^s h_1(t)}{dt^s} - p_2 \frac{d^s h_2(t)}{dt^s} \right).$$
In particular, $f_1(t) = f(x) = p_2 h_2'(x) - p_1 h_1'(x)$,
$$r_N = \left[\frac{1}{N} \sum_{j=1}^{N} \left(p_2 a^2_{j:N} + p_1 a^1_{j:N} \right)^2 \left(w_{j:N} h_1(t^B) + (1 - w_{j:N}) h_2(t^B) \right) \right]^{1/2},$$
$$r = \lim_{N\to\infty} r_N.$$

Remark 3.1. Condition (B_3) implies that the latter limit exists.

In what follows, the symbol $\Rightarrow$ stands for weak convergence.

Theorem 3.3. *Assume that conditions* (A) *and* (B_3) *hold. If the densities h_i exist and are continuously differentiable in a neighborhood of t^B and $f(t^B) \neq 0$, then*
$$N^{1/3} \left(\hat{t}^{MER}_N - t^B \right) \Rightarrow \left(\frac{2r}{f(t^B)} \right)^{2/3} Z,$$
where $Z = \operatorname{argmin}_{t\in\mathbb{R}} (W(t) + t^2)$ and $W(t)$ is a two sided standard Wiener process.

Remark 3.2. It follows from Lemma 2.6 of [10] that $\min_{t\in\mathbb{R}}(W(t) + t^2)$ is almost surely attained at a unique (random) point $Z \in \mathbb{R}$.

Theorem 3.4. *Let conditions* (A) *and* (B_3) *hold. Assume that*

(i) *the derivatives $h_i''(t) = d^2 h_i(t)/dt^2$ exist and are bounded in a neighborhood of t^B and $f(t^B) \neq 0$;*

(ii) *$\int_{-\infty}^{\infty} zK(z)\,dz = 0$, $D^2 \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} z^2 K(z)\,dz < \infty$, and $d^2 < \infty$;*

(iii) *$\kappa_n = c/N^{1/5}$ for some nonrandom $c > 0$.*

Then
$$N^{2/5} \left(\hat{t}^{EBC} - t^B \right) \Rightarrow A + B\eta,$$
where $A = D^2 c^{2/5} f_2(t^B)/(2f(t^B))$, $B = dr/(c^{1/10} f(t^B))$, and η is a standard Gaussian random variable.

4. Proof of theorems

Proof of Theorem 3.1. Note that condition (B_2) implies that the weight coefficients $a^i_{j:N}$ are bounded uniformly with respect to N and j. Theorem 2.4.2 of [4] yields

$$\sup_x \left|H_i^N(x) - H_i(x)\right| \to 0$$

in probability as $N \to \infty$. Thus $\sup_x |L^N(x) - L(x)| \to 0$ in probability. Fix an arbitrary $\lambda > 0$ and $\varepsilon > 0$. Let

$$A_N = \left\{\sup_x \left|L^N(x) - L(x)\right| < \varepsilon/2\right\}.$$

If N is sufficiently large, then

$$\mathsf{P}(A_N) > 1 - \lambda.$$

Since L is a continuous function on $\mathbb{R}$, $L(-\infty) = p_1$, $L(+\infty) = p_2$, and condition (A) holds. Moreover, for all $\delta > 0$ there exists ε such that $L(t) > L(t^B) + \varepsilon$ for all t for which $|t - t^B| > \delta$. Let the event A_N occur. Then

$$L\left(\hat{t}_N^{MER}\right) - \frac{\varepsilon}{2} \le L_N\left(\hat{t}_N^{MER}\right) \le L_N\left(t^B\right) \le L\left(t^B\right) + \frac{\varepsilon}{2},$$

whence $L(\hat{t}_N^{MER}) \le L(t^B) + \varepsilon$ and $|\hat{t}_N^{MER} - t^B| \le \delta$. This completes the proof of the theorem, since δ and λ are arbitrary. □

Proof of Theorem 3.2. According to Theorem 1 of [7], the assumptions of the theorem imply that $\hat{h}_i^N(x) \to h_i(x)$ in probability at every point $x \in \mathbb{R}$. Therefore

$$u_N(x) \stackrel{\text{def}}{=} p_2\hat{h}_2^N(x) - p_1\hat{h}_2^N(x) \to u(x) \stackrel{\text{def}}{=} p_2h_2(x) - p_1h_1(x)$$

in probability. For $\delta > 0$ we let $A_N(\delta) = \{\text{there exists } t\colon |t - t^B| \le \delta, u_N(t) = 0\}$. Now we show that

$$\mathsf{P}(A_N(\delta)) \to 1 \tag{6}$$

as $N \to \infty$.

Since t^B is a point of minimum of $L(t)$ and $L'(t) = u(t)$ is a continuous function, u changes its sign in a neighborhood of the point t^B. This means that there are t^- and t^+ such that

$$t^B - \delta < t^- < t^B < t^+ < t^B + \delta$$

and $u(t^-)u(t^+) < 0$. Thus $\mathsf{P}\{u_N(t^-)u_N(t^+) < 0\} \to 1$. Since u_N is a continuous function, $\{u_N(t^-)u_N(t^+) < 0\} \subseteq A_N(\delta)$. Therefore relation (6) is proved.

Fix δ. In the same way as in the proof of Theorem 3.1 we show that there is $\varepsilon > 0$ such that $L(t) > L(t^B) + \varepsilon$ for all $t \notin [t^B - \delta, t^B + \delta]$. Let $0 < \delta' < \delta$ be such that $L(t) < L(t^B) + \varepsilon/4$ for all $t \in [t^B - \delta', t^B + \delta']$. Put

$$B_N = \left\{\inf_{t \notin [t^B - \delta, t^B + \delta]} L_N(t) > L(t^B) + \varepsilon/2 > \sup_{t \in [t^B - \delta', t^B + \delta']} L_N(t)\right\}.$$

Fix an arbitrary $\lambda > 0$. Using the uniform convergence of L_N to L, we obtain for sufficiently large N that

$$\mathsf{P}(B_N) > 1 - \frac{\lambda}{2}.$$

According to (6),

$$\mathsf{P}(A_N(\delta')) > 1 - \frac{\lambda}{2}$$

for sufficiently large N. If the event $A_N(\delta')$ occurs, then there exists

$$t^* \in T_N \cap \left[t^B - \delta', t^B + \delta'\right]$$

such that $L_N(t^*) < L_N(t)$ for all $t \notin [t^B - \delta, t^B + \delta]$ given the event B_N occurs. Therefore

$$\hat{t}^{EBC} \in [t^B - \delta, t^B + \delta].$$

Thus

$$\mathsf{P}\left\{\left|\hat{t}^{EBC} - t^B\right| < \delta\right\} \geq \mathsf{P}(A_N(\delta') \cap B_N) \geq 1 - \lambda$$

for sufficiently large N.

This completes the proof of the theorem, since λ is arbitrary. □

Proof of Theorem 3.3. Put

$$W_N(\tau) = N^{2/3}\left(L_N\left(t^B + N^{-1/3}\tau\right) - L_N\left(t^B\right) - L\left(t^B + N^{-1/3}\tau\right) + L\left(t^B\right)\right).$$

For the proof of Theorem 3.3 we need some auxiliary results on the asymptotic behavior of the process W_N. We state these results in the following two lemmas.

Lemma 4.1. *On an arbitrary finite interval $U = [\tau_-, \tau_+]$, the stochastic processes W_N weakly converge as $N \to \infty$ to the process rW in the space* $\mathrm{Var}(U)$ *of functions without discontinuities of the second kind equipped with the uniform metric.*

Proof. Since trajectories of W are continuous, one needs to prove that (i) the finite dimensional distributions of W_N are asymptotically Gaussian, (ii) the second moments of increments converge, and (iii) the distributions of W_N are tight in $D(U)$ (see [1]).

First we compute $\mathsf{E}(W_N(\tau_2) - W_N(\tau_1))^2$. Let $\tau_1 < \tau_2$ and put

$$b_{j:N} = p_2 a^2_{j:N} - p_1 a^1_{j:N}.$$

Then

$$W_N(\tau_2) - W_N(\tau_1) = N^{-1/3}\sum_{j=1}^{N} b_{j:N}\left(\mathbb{1}\{\xi_{j:N} \in A_N\} - \mathsf{P}\{\xi_{j:N} \in A_N\}\right), \tag{7}$$

where $A_N = A_N(\tau_1, \tau_2) = [N^{-1/3}\tau_1, N^{-1/3}\tau_2)$. Therefore

$$\mathsf{E}(W_N(\tau_2) - W_N(\tau_1))^2 = N^{-2/3}\sum_{j=1}^{N}(b_{j:N})^2\Big[(w_{j:N}H_1(A_N) + (1 - w_{j:N})H_2(A_N)) - (w_{j:N}H_1(A_N) + (1 - w_{j:N})H_2(A_N))^2\Big].$$

Taking into account that $H_i(A_N) \sim h_i(t^B)N^{-1/3}(\tau_2 - \tau_1)$, we obtain

$$\mathsf{E}(W_N(\tau_2) - W_N(\tau_1))^2 \to r^2(\tau_2 - \tau_1) = \mathsf{E}(rW(\tau_2) - rW(\tau_1))^2$$

as $N \to \infty$. The finite dimensional distributions of W_N are asymptotically Gaussian in view of the central limit theorem under the Lindeberg condition, since all terms in the sum (7) uniformly converge.

It remains to check that the family of distributions of W_N is tight. To apply a known criterion (Theorem 15.6 of [1]), we show that

$$J \stackrel{\text{def}}{=} \mathsf{E}(W_N(\tau) - W_N(\tau_1))^2(W_N(\tau_2) - W_N(\tau))^2 \leq C_1(\tau_2 - \tau_1)^2 \tag{8}$$

for all $\tau_1 < \tau < \tau_2$ where C_1 does not depend on N, τ_1, τ, and τ_2.

Put

$$\eta_j(\tau_1, \tau_2) = b_{j:N}\big(\mathbb{1}\{\xi_{j:N} \in A_N(\tau_1, \tau_2)\} - \mathsf{P}\{\xi_{j:N} \in A_N(\tau_1, \tau_2)\}\big).$$

Then

$$\begin{aligned} J &\le \frac{C_2}{N^{4/3}}\Bigg(\sum_{j\ne k}\big\{\mathsf{E}(\eta_j(\tau,\tau_2))^2(\eta_k(\tau_1,\tau))^2 + |\mathsf{E}\,\eta_j(\tau,\tau_2)\eta_j(\tau_1,\tau)\eta_k(\tau,\tau_2)\eta_k(\tau_1,\tau)|\big\} \\ &\qquad\qquad + \sum_{j=1}^{N}\mathsf{E}(\eta_j(\tau,\tau_2))^2(\eta_j(\tau_1,\tau))^2\Bigg) \\ &\le \frac{C_2}{N^{4/3}}\left(N^2C_3(h^*)^2N^{-1/3}(\tau-\tau_1)N^{-1/3}(\tau_2-\tau) + NC_4(h^*)^4N^{-1/3}(\tau-\tau_1)(\tau_2-\tau)\right) \\ &\le C_1(\tau_2-\tau_1)^2. \end{aligned}$$

Here C_2 is an absolute constant, C_2, C_3, and C_4 depend only on $\sup_{j,N}|b_{j:N}|<\infty$, and $h^*=\sup_{t\in U}(h_1(t)+h_2(t))$.

Now the lemma follows from Theorems 15.4 and 15.6 of [1]. □

Lemma 4.2. *For every $\varepsilon>0$ there exist $D>0$ such that*

$$\mathsf{P}\big\{\textit{there exists } \tau\in\mathbb{R}\colon |W_N(\tau)|>D(1+|\tau|)\big\}<\varepsilon$$

for all N.

Proof. Using inequality (15.30) of [1], we derive from (8) that

$$\begin{aligned} J_1 &\stackrel{\text{def}}{=} \mathsf{P}\left\{\sup_{\tau\in[\tau_-,\tau_+]}\min(|W_N(\tau)-W_N(\tau_1)|,|W_N(\tau_2)-W_N(\tau)|)>\varepsilon/2\right\} \\ &\le \frac{16C_1}{\varepsilon^4}(\tau_+-\tau_-)^2. \end{aligned}$$

Similarly to the proof of (8), we obtain

$$\mathsf{E}\,W_N^4(\tau)\le C_5\tau^2,$$

whence

$$J_2(\tau)\stackrel{\text{def}}{=}\mathsf{P}\{|W_N(\tau)|>\varepsilon/2\}\le\frac{16C_5\tau^2}{\varepsilon^4}$$

by the Chebyshev inequality. Thus

$$\mathsf{P}\left\{\sup_{\tau\in[\tau_-,\tau_+]}|W_N(\tau)|>\varepsilon\right\}\le J_1+J_2(\tau_-)+J_2(\tau_+)\le\frac{C_6}{\varepsilon^4}\left((\tau_+-\tau_-)^2+\tau_-^2+\tau_+^2\right).$$

Therefore

$$\begin{aligned} &\mathsf{P}\big\{\text{there exists } \tau\in\mathbb{R}\colon |W_N(\tau)|>D(1+|\tau|)\big\} \\ &\quad\le\sum_{j=0}^{\infty}\Big[\mathsf{P}\big\{\text{there exists } \tau\in[i,i+1]\colon |W_N(\tau)|>D(1+i)\big\} \\ &\qquad\qquad + \mathsf{P}\big\{\text{there exists } \tau\in[-i-1,-i]\colon |W_N(\tau)|>D(1+i)\big\}\Big] \\ &\quad\le C_7\sum_{i=0}^{\infty}\left(\frac{(i+1-i)^2}{D^4(i+1)^4}+\frac{i^2}{D^4(1+i)^4}+\frac{(i+1)^2}{D^4(1+i)^4}\right)\le\frac{C_8}{D^4}. \end{aligned}$$

This completes the proof of the lemma. □

Now we prove that the weak convergence of stochastic processes implies the convergence of their points of minimum. We need the following notation.

Fix an arbitrary $S > 0$. Let f be a function on $U = [-S, S]$. Put

$$\mathrm{Am}^-(f,\varepsilon) = \inf\left\{x \in U\colon f(x) \le \inf_{y\in U} f(y) + \varepsilon\right\},$$

$$\mathrm{Am}^+(f,\varepsilon) = \sup\left\{x \in U\colon f(x) \le \inf_{y\in U} f(y) + \varepsilon\right\},$$

$$\mathrm{Am}^-(f) = \lim_{\varepsilon\downarrow 0} \mathrm{Am}^-(f,\varepsilon), \qquad \mathrm{Am}^+(f) = \lim_{\varepsilon\downarrow 0} \mathrm{Am}^+(f,\varepsilon).$$

Since $\mathrm{Am}^-(f,\varepsilon)$ and $\mathrm{Am}^+(f,\varepsilon)$ are nonincreasing with respect to ε, the latter limits always exist.

It is easy to see that

$$\mathrm{Am}^-(f) \le x_* \le \mathrm{Am}^+(f)$$

if a point of minimum $x_* = \operatorname{argmin}_{x\in U} f(x)$ exists. Let $|f|_\infty = \sup_{x\in U} |f(x)|$ be the uniform metric. The variable x in the following lemma runs over the interval U, so all the infimums are considered with respect to $x \in U$, that is, for example, $\inf g = \inf_{x\in U} g(x)$.

Lemma 4.3. *The functionals* Am^- *and* Am^+ *are continuous in the space* $\mathrm{Var}(U)$ *equipped with the metric* $|\cdot|_\infty$.

Proof. We show that $\mathrm{Am}^-(f_n) \to \mathrm{Am}^-(g)$ for functions $f_n, g \in \mathrm{Var}(U)$ such that

$$|f_n - g|_\infty \to 0 \quad \text{as } n \to \infty.$$

(The proof for Am^+ is similar.)

By the definition of $\mathrm{Am}^-(g,\varepsilon)$, given arbitrary $\lambda > 0$, there exists x_g such that

$$g(x_g) \le \inf g + \varepsilon$$

and $x_g \le \mathrm{Am}^-(g,\varepsilon) + \lambda$. Let f, $g \in \mathrm{Var}(U)$ be such that $|f - g|_\infty \le \delta$. Then

$$\inf g - \delta \le \inf f \le \inf g + \delta.$$

Therefore

$$f(x_g) \le g(x_g) + \delta \le \inf g + \delta + \varepsilon \le \inf f + \varepsilon + 2\delta.$$

Thus

$$\mathrm{Am}^-(f, \varepsilon + 2\delta) \le \mathrm{Am}^-(g,\varepsilon). \tag{9}$$

Analogously

$$\mathrm{Am}^-(g, \varepsilon + 2\delta) \le \mathrm{Am}^-(f,\varepsilon). \tag{10}$$

Let $|f_n - g|_\infty \to 0$. Given arbitrary $\gamma > 0$, there exists $\varepsilon_0 > 0$ such that

$$\mathrm{Am}^-(g,\varepsilon) - \gamma \le \mathrm{Am}^-(g) \le \mathrm{Am}^-(g,\varepsilon) + \gamma$$

for all $0 \le \varepsilon \le \varepsilon_0$. Choose an arbitrary $0 < \varepsilon' < \varepsilon_0$ and put $\delta = \varepsilon = \varepsilon'/3$. If n is sufficiently large, then

$$|f_n - g|_\infty \le \delta$$

and

$$\mathrm{Am}^-(f_n, \varepsilon') = \mathrm{Am}^-(f_n, \varepsilon + 2\delta) \le \mathrm{Am}^-(g,\varepsilon) \le \mathrm{Am}^-(g) + \gamma$$

by inequality (9). Thus $\mathrm{Am}^-(f_n) \le \mathrm{Am}^-(g) + \gamma$ for sufficiently large n. Using estimate (10), we obtain

$$\mathrm{Am}^-(g) - \gamma \le \mathrm{Am}^-(g, \varepsilon + 2\delta) \le \mathrm{Am}^-(f,\varepsilon)$$

and

$$\mathrm{Am}^-(g) - \gamma \le \mathrm{Am}^-(f_n) + \gamma$$

for sufficiently large n. Since γ is arbitrary, $\mathrm{Am}^-(f_n) \to \mathrm{Am}^-(g)$.

The lemma is proved. □

Continuation of the proof of Theorem 3.3. Let U be a neighborhood of the point t^B where the densities h_i are continuously differentiable. According to Theorem 3.1,

$$\hat{t}^{MER} \to t^B$$

in probability. Thus $\mathsf{P}(A_N) > 1 - \varepsilon/2$ given an arbitrary $\varepsilon > 0$ and for sufficiently large N where $A_N = \{\hat{t}^{MER} \in U\}$. Put $\tau_N = N^{1/3}(\hat{t}_N^{MER} - t^B)$. Now we show that there exists $S = S_\varepsilon$ such that

$$\mathsf{P}\{|\tau_N| < S\} > 1 - \varepsilon \tag{11}$$

for sufficiently large N. Since

$$L_N\left(t^B + N^{-1/3}\tau\right) - L_N\left(t^B\right) = N^{-2/3}W_N(\tau) + L\left(t^B + N^{-1/3}\tau\right) - L\left(t^B\right),$$

we have

$$\tau_N = \operatorname*{argmin}_{\tau}\ v(\tau),$$

where

$$v(\tau) = W_N(\tau) + N^{2/3}\left(L\left(t^B + N^{-1/3}\tau\right) - L\left(t^B\right)\right).$$

In what follows, we assume that the event $\{\hat{t}^{MER} \in U\}$ occurs. Since H_i is a twice differentiable function on U and t^B is a point of minimum of L, we get

$$L\left(t^B + N^{-1/3}\tau\right) - L(t^B) = \frac{1}{2}L''(\zeta)N^{-2/3}\tau^2,$$

where $\zeta \in U$ is a point such that $L''(\zeta) = f(\zeta)$. Since $f(t^B) > 0$, one can find a neighborhood U for which $f(\zeta) > c > 0$ throughout in U. Then $v(\tau) \geq W_N(\tau) + c\tau^2/2$ for $N^{-1/3}\tau + t^B \in U$.

Put $B_N = \{|W_N(\tau) < D(1+|\tau|) \text{ for all } \tau \in \mathbb{R}\}$. Lemma 4.2 implies that for sufficiently large D,

$$\mathsf{P}(B_N) > 1 - \varepsilon/2.$$

If the event B_N occurs, then $v(\tau) > 0 = v(0)$ for $c\tau^2/2 > D(1 + |\tau|)$. Hence $v(\tau)$ does not attain the minimum at those points. Therefore if the event $A_N \cap B_N$ occurs, then

$$c\tau_N^2 < D(1 + |\tau_N|),$$

that is, $|\tau_N| \leq \max(1, 2D/c) = S$. Since $\mathsf{P}(A_N \cap B_N) > 1 - \varepsilon$, inequality (11) is proved.

Analogously we prove that

$$\mathsf{P}\left\{\left|\operatorname*{argmin}_{\tau\in\mathbb{R}} \tilde{W}(\tau)\right| < S\right\} > 1 - \varepsilon \tag{12}$$

for sufficiently large S where $\tilde{W}(\tau) = rW(\tau) + f(0)\tau^2/2$. Another way to achieve the same conclusion is to apply the law of the iterated logarithm for W.

According to Lemma 4.2, the process W_N weakly converges to rW on the interval $[-S, S]$. Since the densities h_i are continuously differentiable,

$$N^{2/3}\left(L(t^B + N^{-1/3}\tau) - L(t^B)\right)$$

converges uniformly to $f(0)\tau^2/2$. Thus v weakly converges to $\tilde{W}$ in the space $\mathrm{Var}[-S, S]$ equipped with the uniform metric.

Since the function $v(\tau)$ is constant between its consecutive jumps, $\min v(\tau)$ is always attained, that is, $\operatorname{argmin}_\tau v(\tau)$ exists. However the point of minimum is not unique. Moreover

$$\mathrm{Am}^-(v) \leq \operatorname*{argmin}_{\tau\in[-S,S]} v(\tau) \leq \mathrm{Am}^+(v).$$

Applying Lemma 4.3, we obtain $\mathrm{Am}^-(v) \Rightarrow \mathrm{Am}^-(\tilde{W})$ and

$$\mathrm{Am}^+(v) - \mathrm{Am}^-(v) \Rightarrow \mathrm{Am}^+(\tilde{W}) - \mathrm{Am}^-(\tilde{W}).$$

Lemma 2.6 of [10] yields that the minimum of $\tilde{W}$ is almost surely attained at a unique point, whence

$$\mathrm{Am}^-(\tilde{W}) = \operatorname*{argmin}_{\tau\in[-S,S]} \tilde{W}(\tau)$$

and $\mathrm{Am}^+(\tilde{W}) - \mathrm{Am}^-(\tilde{W}) = 0$.

This implies that

$$\mathsf{P}\left\{\operatorname*{argmin}_{\tau\in[-S,S]} v(\tau) < x\right\} \to \mathsf{P}\left\{\operatorname*{argmin}_{\tau\in[-S,S]} \tilde{W}(\tau) < x\right\}$$

for all $x \in \mathbb{R}$.

This together with (11)–(12) proves that τ_N converges to

$$\operatorname*{argmin}_{\tau\in\mathbb{R}} \tilde{W}(\tau) = \operatorname*{argmin}_{\tau\in\mathbb{R}} \left(W(\tau) + \alpha\tau^2\right),$$

where $\alpha = f(t^B)/(2r)$. Since $W(\alpha\tau) \doteq \sqrt{\alpha}W(\tau)$ ($\doteq$ denotes the equality of distributions), we obtain

$$\operatorname*{argmin}_{\tau\in\mathbb{R}} \left(W(\tau) + \alpha\tau^2\right) \doteq \operatorname*{argmin}_{\tau\in\mathbb{R}} \left(\alpha^{-1/3}W(\alpha^{2/3}\tau) + \alpha^{-1/3}(\alpha^{2/3}\tau)^2\right) = \alpha^{-2/3}Z.$$

The theorem is proved. □

Proof of Theorem 3.4. As before let $u_N(t) = p_2\hat{h}_2^N(t) - p_1\hat{h}_1^N(t)$. By the definition of t_N^{EBC} we have $u_N(t_N^{EBC}) = 0$. Put $\delta_N = t_N^{EBC} - t^B$. Theorem 3.2 implies that $\delta_N \to 0$ in probability. Thus

$$0 = u_N\left(t^B + \delta_N\right) \approx u_N\left(t^B\right) + \delta_N u_N'\left(t^B\right)$$

and

$$\begin{aligned}\delta_N \approx -\frac{u_N\left(t^B\right)}{u_N'\left(t^B\right)} &= \frac{p_2\hat{h}_2^N\left(t^B\right) - p_1\hat{h}_1^N\left(t^B\right)}{\left(p_2\hat{h}_2^N\left(t^B\right) - p_1\hat{h}_1^N\left(t^B\right)\right)'} \\ &\approx \frac{p_2(\hat{h}_2^N(t^B) - h_2(t^B)) - p_1(\hat{h}_1^N(t^B) - h_1(t^B))}{f_1(t^B)},\end{aligned}$$

since $p_1h_1(t^B) - p_2h_2(t^B) = 0$.

Similarly to the proof of Lemma 2 of [6], we obtain

$$\begin{aligned}&N^{2/5}\left(p_1\left(\hat{h}_1^N\left(t^B\right) - h_1\left(t^B\right)\right) - p_2\left(\hat{h}_2^N\left(t^B\right) - h_2\left(t^B\right)\right)\right) \\ &\quad\Rightarrow D^2c^{2/5}f_2\left(t^B\right)/2 + dr/c^{1/10}\eta\end{aligned}$$

for $\kappa_N = c/N^{1/5}$, where η is a standard Gaussian random variable.

This completes the proof of the theorem. □

5. Concluding remarks

The results obtained in this paper allow one to compare classification rules for the minimal empirical risk method and empirical Bayes classification if the size of a sample is sufficiently large. The rate of convergence of the empirical Bayes classification rule is better than that of the minimal empirical risk method. Nevertheless, if the size of a sample is small and the smoothing parameter is hard to choose in order to estimate densities for the empirical Bayes classification, the minimal empirical risk method may appear to be a more reliable approach.

Bibliography

1. P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, Inc., New York, 1968. MR0233396 (38:1718)
2. V. N. Vapnik, *Inductive principles for the search for empirical laws*, Pattern Recognition. Classification. Prediction, vol. 1, "Nauka", Moscow, 1989, pp. 17–81. (Russian) MR1074215 (92e:68170)
3. L. Devroye and L. Gyorfi, *Nonparametric Density Estimation. The L_1 View*, John Wiley & Sons, Inc., New York, 1985. MR780746 (86i:62065)
4. R. E. Maĭboroda, *Statistical Analysis of Mixtures. A course of lectures*, "Kyiv University", Kyiv, 2003. (Ukrainian)
5. Yu. O. Ivan'ko and R. E. Maĭboroda, *Exponential estimates for the empirical Bayes risk in the classification of a mixture with varying concentrations*, Ukrain. Mat. Zh. **54** (2002), no. 10, 1421–1428; English transl. in Ukrainian Math. J. **54** (2002), no. 10, 1722–1731. MR2015493 (2004g:62022)
6. Yu. O. Ivan'ko, *The asymptotic behavior of kernel estimators and their derivatives constructed from observations from a mixture with varying concentrations*, Visnyk KNU, Ser. Matematika. Mekhanika (2003), no. 9–10, 29–35.
7. O. V. Sugakova, *Asymptotics of a kernel estimate for distribution density constructed from observations of a mixture with varying concentrations*, Teor. Imovirnost. Matem. Statist. **59** (1998), 156–166; English transl. in Theor. Probability Math. Statist. **59** (1999), 161–171. MR1793776
8. H. Chernoff, *Estimation of the mode*, Ann. Inst. Statist. Math. **16** (1964), 31–41. MR0172382 (30:2601)
9. L. Mohammadi and S. van de Geer, *On threshold-based classification rules*, Lecture Notes Monograph Series, Mathematical Statistics and Applications: Festschrift for Constance van Eeden, vol. 42, Institute of Mathematical Statistics, 2003, pp. 261–280. MR2138297 (2006b:62106)
10. J. Kim and D. Pollard, *Cube root asymptotics*, Ann. Stat. **18** (1990), no. 1, 191–219. MR1041391 (91f:62059)
11. V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1996. MR1719582 (2001c:68110)

SK Lemma-Vite, Brats'ka Street, Kyiv, 6, 04070 Ukraine
E-mail address: ivanko@lemma-insur.com.ua

Department of Probability Theory and Mathematical Statistics, Faculty for Mathematics and Mechanics, National Taras Shevchenko University, Glushkov Avenue, 6, Kyiv, 03127, Ukraine
E-mail address: mre@univ.kiev.ua

Received 20/DEC/2004

Translated by OLEG KLESOV