HISTOGRAM ESTIMATORS OF THE SHAPE OF THE CONCENTRATION FUNCTION IN A TWO-COMPONENT MIXTURE

UDC 519.21

D. I. POHYL'KO

ABSTRACT. We construct the projection estimators of the shape of the concentration function and obtain their expansions in the basis of indicator functions (the histogram basis) in the case where the data is a sample from a mixture of two components with unknown distributions whose concentrations are varying with observations. We prove that the estimators are consistent and find the rate of the almost sure convergence.

1. INTRODUCTION

Mixtures of several probability distributions are often used in statistics as a mathematical model for the distribution of the data. The classical model of a mixture can be written as follows:

$$P\{\xi \in A\} = w^1 H_1(A) + w^2 H_2(A) + \cdots + w^M H_M(A),$$

where M is the number of components in a mixture, H_m are some probability distributions, and w^m is the concentration of the component m in the mixture. The above models are used for the cases where observations may belong to one of M known classes (populations, components, etc.). The distribution of characteristics of observations (ξ) is H_m for the component m , while the probability that an observation belongs to the component m is w^m . The classical model of a mixture requires that the class containing an observation is known (see [5] and further references therein).

The problem of estimation of the distribution is studied in [1, 4] for known concentrations. The estimations of the distribution and classification for mixtures with varying concentrations are considered in [10, 11].

We consider the problem of estimation of the concentration function for the case where the concentrations of the components are varying with observations; the mixture contains observations that may belong to one of two components whose distributions are unknown. Thus we deal with the case where the observations ξ_j are jointly independent and

$$P\{\xi_j \in A\} = w(t_j)H_1(A) + (1 - w(t_j))H_2(A),$$

where $w(t_j)$ is the concentration of the first component at the moment t_j when the observation j appears. This problem can be solved by using methods of nonparametric regression (see [8]). Indeed, let η_1 and η_2 be random variables with finite expectations. We treat the pair (η_1, η_2) as a description of characteristics of components. Then ξ_j

2000 *Mathematics Subject Classification.* Primary 62G20; Secondary 62G05.

Key words and phrases. Projection estimates, estimates of the concentration, the shape of the concentration function.

admits the representation $\xi_j = w(t_j)(\mathbf{E}\eta_1 - \mathbf{E}\eta_2) + \mathbf{E}\eta_2 + \varepsilon_j$, where ε_j is the noise (independent centered random variables). Note however that this method leads to losses of information about $w(t_j)$ contained in the noise ε_j . For example, if $\mathbf{E}\eta_1 = \mathbf{E}\eta_2$, then $w(t_j)$ cannot be estimated by using the regression methods. The estimator considered in this paper is consistent. An example of the estimator is given in Section 3 for the case of $\mathbf{E}\eta_1 = \mathbf{E}\eta_2$.

The projection to the histogram basis is a particular case of the method of projection to the wavelet basis (see [7, 6]). An application of wavelet analysis in statistics is considered in [9, 12]. Another application of wavelets to the problems of nonparametric regression is given in [9].

The estimators considered in this paper are analogous to projection estimators studied in [2, 3], where the expansion in a trigonometric basis is used. Estimators of distribution functions of components of a mixture are constructed in [3]; these estimators require the knowledge of the distribution of the concentration function.

2. THE SETTING OF THE PROBLEM AND CONSTRUCTION OF THE ESTIMATOR

Let $\{\xi_j^N, j = 1, \dots, N\}_{N \geq 1}$ be a sequence of series of random vectors in $\mathbb{R}^d$ such that the vectors $\{\xi_j^N, j = 1, \dots, N\}$ are jointly independent whatever N is and

$$(1) \quad \mathbf{P}\{\xi_j^N < x\} = F(t_j, x) = w(t_j)G(x) + (1 - w(t_j))H(x),$$

where G and H are distribution functions on $\mathbb{R}^d$ and $w: [0, 1] \rightarrow [0, 1]$ is the concentration of the component G at the moment $t_j = j/N$.

By Lip_β we denote the Lipschitz class of order β , that is,

$$\begin{aligned} \text{Lip}_\beta[0, 1] = \left\{ f: [0, 1] \rightarrow \mathbb{R} \mid \text{there exists a constant } L' \text{ such that} \right. \\ \left. \sup_{t, s \in [0, 1]} |f(t) - f(s)| \leq L'|t - s|^\beta \right\}, \\ \Sigma(\beta, L') = \left\{ f: [0, 1] \rightarrow \mathbb{R} \mid \sup_{t, s \in [0, 1]} |f(t) - f(s)| \leq L'|t - s|^\beta \right\}. \end{aligned}$$

Below we consider the problem of estimation of w from observations ξ_j^N under the assumption that $w \in \text{Lip}_\beta[0, 1]$, where $\beta \in [0, 1]$ is a fixed (known) number and G and H are unknown.

The function $F(t, x)$ can be represented as follows:

$$(2) \quad F(t, x) = G_1(x) + u(t)G_2(x),$$

where

$$\begin{aligned} G_1(x) = H(x) + abG_2(x), \quad G_2(x) = \frac{G(x) - H(x)}{a}, \\ u(t) = a(w(t) - b), \quad b = \int_0^1 w(t) dt, \quad a = \left(\int_0^1 (w(t) - b)^2 dt \right)^{-1/2}. \end{aligned}$$

It is clear that

$$(3) \quad \int_0^1 (u(t))^2 dt = 1, \quad \int_0^1 u(t) dt = 0.$$

If $w(t)$ is the concentration function, then $u(t)$ is called its shape. It is shown in [2] that the representation (2) is unique up to the sign of $u(t)$ under the normalization conditions (3). If conditions (3) do not hold, then the representation is not unique. In what follows we estimate the function $u(t)$ defined by (2)–(3). Therefore we estimate the

shape of the concentration function; that is, we estimate the concentration function to within the shift and scale parameters.

We find a function $u_N^*(t)$ approximating u for a uniform time grid. As a quality criterion for the approximation we take

$$\sqrt{\frac{1}{N} \sum_{k=1}^N \left(u\left(\frac{k}{N}\right) - u_N^*\left(\frac{k}{N}\right) \right)^2}.$$

We follow the idea of [2] when constructing $u_N^*(t)$. The idea is to project $F(t, x)$ to the class of functions

$$F_v(t, x) = G_1(x) + v(t)G_2(x),$$

where $v \in V$ and V is a finite-dimensional class of functions in $L_2([0, 1], m)$ (here and in what follows m denotes the Lebesgue measure). We also assume that v satisfies the normalization conditions (3). Let $\pi(\cdot)$ be some probability measure on $\mathbb{R}^d$ and let Δ be the support of the measure π . We are looking for a v satisfying (3) and minimizing the expression

$$(4) \quad \sqrt{\frac{1}{N} \int_{\Delta} \sum_{k=1}^N \left(F\left(\frac{k}{N}, x\right) - F_v\left(\frac{k}{N}, x\right) \right)^2 \pi(dx)}.$$

By λ_N we denote the probability counting measure concentrated at the points $t_j^N = j/N$, $j = 1, \dots, N$, that is, $\lambda_N(\{t_j^N\}) = 1/N$. Here and in what follows $\|\cdot\|_N$ and $\langle \cdot, \cdot \rangle_N$ stand for the norm and scalar product in the space $L_2([0, 1], \lambda_N)$, respectively. The symbol $\|\cdot\|_2$ denotes the norm in $L_2([0, 1], m)$. Therefore,

$$\sqrt{\frac{1}{N} \sum_{k=1}^N (u(t_k^N) - u_N^*(t_k^N))^2} = \|u - u_N^*\|_N.$$

If $\mathbb{H} = L_2([0, 1] \times \Delta, \lambda_N \times \pi)$, then the problem of minimization of (4) (in other words, the problem of minimization of $\|F_v - F\|_{\mathbb{H}}$ on V) is equivalent to the maximization of

$$(5) \quad \begin{aligned} D(v) &= \int_{\Delta} \left(\int_{[0,1]} v(t) F(t, x) \lambda_N(dt) \right)^2 \pi(dx) \\ &= \frac{1}{N^2} \int_{\Delta} \int_{\Delta} \left(\sum_{k=1}^N v\left(\frac{k}{N}\right) F\left(\frac{k}{N}, x\right) \right)^2 \pi(dx). \end{aligned}$$

Since the distribution function $F(t, x)$ is unknown, we find an estimator of v by substituting an estimator of F to (5). It is clear that the estimator depends on V . For example, if $\{f_i, i = 1, \dots, l\}$ is an orthonormal basis c.l.s.(V) in $L_2([0, 1], \lambda_N)$, then the estimator u_N^* can be constructed as follows:

$$(6) \quad u_N^* = \sum_{i=1}^l c_N^i f_i,$$

where $c_N = (c_N^i, i = 1, \dots, l)$ is the eigenvector of the matrix $(d_{ij}^N)_{i,j=1}^l$ corresponding to its maximum eigennumber. The entries of the matrix are such that

$$(7) \quad d_{ij}^N = \frac{1}{N^2} \sum_{l,k=1}^N f_i(t_l^N) f_j(t_k^N) \int_{\Delta} \mathbb{1}\{\xi_l^N < x, \xi_k^N < x\} \pi(dx)$$

(see [2, 3]).

We use the following notation:

$$g_k^l(t) := \sqrt{l+1} \mathbb{1} \left\{ t \in \left(\frac{k}{l+1}, \frac{k+1}{l+1} \right] \right\}, \quad V_l = \text{c.l.s.} \{g_k^l, k = 0, \dots, l\},$$

where the closure is considered in the space $L_2((0, 1], \lambda_N)$; $U_l = \text{c.l.s.} \{g_k^l, k = 0, \dots, l\}$ (the closure is considered in the space $L_2((0, 1], m)$) and $\{g_k^l, k = 0, \dots, l\}$ is an orthonormal basis in U_l . Let

$$\{f_j^{Nl}(t), j = 0, \dots, l\}$$

be an orthonormal basis in V_l with respect to the scalar product $\langle \cdot, \cdot \rangle_N$, and let $f_0^{Nl} \equiv 1$, $W_l = \text{c.l.s.} \{f_k^{Nl}, k = 1, \dots, l\}$ (the closure is considered in the space $L_2((0, 1], \lambda_N)$).

Theorem 2.1. *Let (2) and (3) hold,*

$$\int_{\Delta} (G_2(x))^2 \pi(dx) > 0, \quad u \in \Sigma(\beta, L'), \quad \beta \in (0, 1],$$

and let the estimator u_N^* be defined by (6), (7), where W_l is used instead of V . If $l = N^\alpha$ and $\alpha \in (0, \frac{1}{2})$, then

- (1) $\min_{\sigma=\pm 1} \|u - \sigma u_N^*\|_N = O(r_N(\alpha, \beta))$ almost surely as $N \rightarrow \infty$, where
$$r_N(\alpha, \beta) = (\ln N)^{1/4} N^{\alpha/2-1/4} + N^{-\alpha\beta};$$
- (2) $\min_{\sigma=\pm 1} \|u - \sigma u_N^*\|_2 = O(r_N(\alpha, \beta))$ almost surely as $N \rightarrow \infty$.

In what follows we need the following auxiliary results.

Lemma 2.1. *Let $z = (z^1, \dots, z^l) \in \mathbb{R}^l$, $\|z\| \in [\frac{1}{2}, 2]$, $u_z = \sum_{i=1}^l z^i f_i^{Nl}$, and*

$$F(t, x) = G_1(x) + u_z(t)G_2(x) + \Delta_N(t, x),$$

where $\Delta_N(t, x)$ is an arbitrary function. Given $\zeta \in \mathbb{R}^l$ such that $\|\zeta\| = 1$ (the norm is considered in the space $\mathbb{R}^l$) we put

$$\begin{aligned} \Delta_\zeta(x) &= \int_{[0,1]} u_\zeta(t) \Delta_N(t, x) \lambda_N(dt), \quad C = \int_{\Delta} (G_2(x))^2 \pi(dx), \\ D(\zeta) &:= D(u_\zeta) = \int_{\Delta} \left(\int_{[0,1]} u_\zeta(t) (G_1(x) + u_z(t)G_2(x) + \Delta_N(t, x)) \lambda_N(dt) \right)^2 \pi(dx), \\ R(\zeta) &= \int_{\Delta} G_2(x) \Delta_\zeta(x) \pi(dx), \quad Q(\zeta) = \int_{\Delta} (\Delta_\zeta(x))^2 \pi(dx), \\ \Delta_\zeta &= 4|R(\zeta)| + Q(\zeta), \quad \delta = \delta_N = \sup_{\|\zeta\|=1} \Delta_\zeta, \quad v = \operatorname{argmax}_{\|\zeta\|=1} D(\zeta). \end{aligned}$$

Then

$$\min_{\sigma=\pm 1} \|u_z - \sigma u_v\|_N \leq |\|z\| - 1| + 4\sqrt{\delta/C}.$$

Note that $\|z\| = \|u_z\|_N$. Lemma 2.1 is a modification of Lemma 1 in [3].

Proof. One can show that

$$D(\zeta) = C \langle u_\zeta, u_z \rangle_N^2 + 2 \langle u_\zeta, u_z \rangle_N R(\zeta) + Q(\zeta).$$

Put $y = z / \|z\|$. Then $\|y\| = 1$ and $z = y \|z\|$, whence we get the representation

$$D(\zeta) = C \|z\|^2 \langle u_\zeta, u_y \rangle_N^2 + 2R(\zeta) \|z\| \langle u_\zeta, u_y \rangle_N + Q(\zeta).$$

Proceeding in the same way as in [3] we obtain

$$\min_{\sigma=\pm 1} \|u_y - \sigma u_v\|_N \leq 4\sqrt{\frac{\delta}{C}},$$

whence

$$\min_{\sigma=\pm 1} \|u_z - \sigma u_v\|_N \leq \|u_z - u_y\|_N + \min_{\sigma=\pm 1} \|u_y - \sigma u_v\|_N \leq \|z\| - 1 + 4\sqrt{\frac{\delta}{C}}$$

by the triangle inequality. $\square$

The following result is proved in [2, Lemma 3.6.3].

Lemma 2.2. *Let the assumptions of Theorem 2.1 hold,*

$$u(t) = \sum_{i=1}^l c_i^l f_i^{Nl}(t) + u^0(t),$$

where $\|c\| = 1$ (the norm is considered in the space $\mathbb{R}^l$), $\langle u^0, f_i^{Nl} \rangle_N = 0$, $i = 1, \dots, l$,

$$(8) \quad \Delta_N(t, x) = \mathbb{1}\{\xi_{k(t)}^N < x\} - G_1(x) - \sum_{i=1}^N c^i f_i^{Nl}(t) G_2(x),$$

and $k(t) = \min\{y \in \mathbb{Z} \mid y/N \geq t\}$. Then

$$\delta_N = O((\ln N)^{1/2}) N^{\alpha-0.5} \quad a.s.$$

(we use the notation of Lemma 2.1).

If $\Delta_N(t, x)$ is chosen as in (8), then the function u_v in Lemma 2.1 becomes the estimator u_N^* defined by (6).

Applying the mean value theorem one can obtain the following result.

Lemma 2.3. *Let $u: [0, 1] \rightarrow \mathbb{R}$ and $u \in \Sigma(\beta, L')$. Then $\sup_{t \in (0, 1]} |u(t) - v(t)| \leq L'l^{-\beta}$, where $v(t) = \text{Pr}_{U_l} u(t)$ and the projection is constructed in the space $L_2((0, 1], m)$.*

Lemma 2.4. *Let $N \geq 4$, $l = N^\alpha$, $\alpha \in (0, \frac{1}{2})$; $u, v: [0, 1] \rightarrow \mathbb{R}$; $u \in \Sigma(\beta, L')$, and*

$$v(t) = \sum_{k=0}^l c_k \mathbb{1}\left\{t \in \left(\frac{k}{l+1}, \frac{k+1}{l+1}\right]\right\}.$$

Then there exist constants $K_i = K_i(L') \in \mathbb{R}$, $i = 1, 2, 3$, such that

$$\|u - v\|_2 \leq K_1 N^{-\beta} + K_2 \|u - v\|_N + K_3 N^{-\beta/2} \|u - v\|_N^{1/2}.$$

Proof. We have

$$\begin{aligned} \|u - v\|_2^2 &= \int_0^1 (u(t) - v(t))^2 dt = \sum_{k=0}^l \int_{\frac{k}{l+1}}^{\frac{k+1}{l+1}} (u(t) - v(t))^2 dt \\ &\leq \sum_{k=0}^l \sum_{s \in \mathbb{Z}: (\frac{s}{N}, \frac{s+1}{N}] \subset (\frac{k}{l+1}, \frac{k+1}{l+1}]} \int_{\frac{s}{N}}^{\frac{s+1}{N}} (u(t) - v(t))^2 dt \\ &\quad + \sum_{k=0}^l \int_{\frac{s_k}{N}}^{\frac{s_k+1}{N}} (u(t) - v(t))^2 dt \\ &= S_1 + S_2, \end{aligned}$$

where the number $s_k \in \mathbb{Z}$ is such that

$$\frac{Nk-1}{l+1} < s_k \leq \frac{kN}{l+1}.$$

Considering S_1 note that

$$v(t) = v\left(\frac{s}{N}\right)$$

for all $t \in (s/N, (s+1)/N]$. The arguments t at which $v(t)$ may change its value in the interval of integration are included in the sum S_2 . We choose $N \geq 4$ to be sure that the function $v(t)$ does not change its value more than once in an interval of length $1/N$.

Now we estimate S_1 . Since

$$v(t) = v\left(\frac{s+1}{N}\right)$$

for all $t \in (s/N, (s+1)/N]$ and for all terms in the sum S_1 , one can estimate them by

$$\frac{1}{N} \left(u\left(\frac{s+1}{N}\right) - v\left(\frac{s+1}{N}\right) \right)^2 + \frac{2L'N^{-\beta}}{N} \left| u\left(\frac{s+1}{N}\right) - v\left(\frac{s+1}{N}\right) \right| + \frac{L'^2N^{-2\beta}}{N}.$$

Then, for S_2 ,

$$\begin{aligned} & \int_{\frac{s_k}{N}}^{\frac{s_k+1}{N}} (u(t) - v(t))^2 dt \\ & \leq \int_{\frac{s_k}{N}}^{\frac{s_k+1}{N}} \left(u(t) - v\left(\frac{s_k}{N}\right) \right)^2 dt + \int_{\frac{s_k}{N}}^{\frac{s_k+1}{N}} \left(u(t) - v\left(\frac{s_k+1}{N}\right) \right)^2 dt = I_1 + I_2. \end{aligned}$$

Since I_2 is already estimated, we turn to I_1 . Using the mean value theorem for u , one can show that

$$I_1 \leq \frac{L'^2N^{-2\beta}}{N} + \frac{2L'N^{-\beta}}{N} \int_{\frac{s_k-1}{N}}^{\frac{s_k}{N}} \left| u(t) - v\left(\frac{s_k}{N}\right) \right| dt + \int_{\frac{s_k-1}{N}}^{\frac{s_k}{N}} \left(u(t) - v\left(\frac{s_k}{N}\right) \right)^2 dt.$$

Thus there exist constants $C_1, C_2, C_3 \in \mathbb{R}$ such that

$$\begin{aligned} S_1 + S_2 & \leq C_1N^{-2\beta} + \frac{C_2}{N} \sum_{k=1}^N \left(u\left(\frac{k}{N}\right) - v\left(\frac{k}{N}\right) \right)^2 + \frac{C_3N^{-\beta}}{N} \sum_{k=1}^N \left| u\left(\frac{k}{N}\right) - v\left(\frac{k}{N}\right) \right| \\ & \leq C_1N^{-2\beta} + C_2 \|u - v\|_N^2 + C_3N^{-\beta} \|u - v\|_N, \end{aligned}$$

whence Lemma 2.4 follows. $\square$

Proof of Theorem 2.1. Applying the triangle inequality we get

$$\min_{\sigma=\pm 1} \|u - \sigma u_l^*\|_N \leq \|u - \bar{u}_l\|_N + \|\bar{u}_l - \tilde{u}_l\|_N + \min_{\sigma=\pm 1} \|\tilde{u}_l - \sigma u_N^*\|_N,$$

where

$$\bar{u}_l = \text{Pr}_{V_l} u, \quad \tilde{u}_l = \text{Pr}_{W_l} u,$$

and the projection is considered in the space $L_2((0, 1], \lambda_N)$. We estimate the first term. Using the notation of Lemma 2.3 we obtain

$$\|u - \bar{u}_l\|_N = \min_{y \in V_l} \|u - y\|_N \leq \|u - v_l\|_N \leq \sup_{t \in (0, 1]} |u(t) - v_l(t)| \leq L'N^{-\alpha\beta},$$

since $v_l \in V_l$ for $v_l = \text{Pr}_{U_l} \bar{u}_l$. Now we estimate the second term. Since $W_l \subset V_l$ and $\int_{[0, 1]} u(t) dt = 0$, we have

$$\begin{aligned} \|\bar{u}_l - \tilde{u}_l\|_N & = \|\text{Pr}_{V_l \ominus W_l} u\|_N = |\langle u, f_0^{Nl} \rangle_N| = \left| \int_{[0, 1]} u(t) \lambda_N(dt) \right| \\ & = \left| \int_{[0, 1]} u(t) dt - \int_{[0, 1]} u(t) \lambda_N(dt) \right| \leq \sum_{k=1}^N \int_{\frac{k-1}{N}}^{\frac{k}{N}} \left| u(t) - u\left(\frac{k}{N}\right) \right| dt \leq L'N^{-\beta}. \end{aligned}$$

To estimate $|1 - \|\tilde{u}_l\|_N|$ we write

$$\begin{aligned} |1 - \|\tilde{u}_l\|_N| &\leq |1 - \|u\|_N| + \left| \|u\|_N - \|\tilde{u}_l\|_N \right| + \left| \|\tilde{u}_l\|_N - \|\tilde{u}_l\|_N \right| \\ &\leq |1 - \|u\|_N| + \|u - \tilde{u}_l\|_N + \|\tilde{u}_l - \tilde{u}_l\|_N \leq |1 - \|u\|_N| + L'N^{-\alpha\beta} + L'N^{-\beta}, \\ |1 - \|u\|_N| &= \frac{|1 - \|u\|_N^2|}{1 + \|u\|_N} \leq \frac{|1 - \|u\|_N^2|}{1 + \|u\|_N} = \left| \sum_{k=1}^N \int_{\frac{k-1}{N}}^{\frac{k}{N}} \left(u^2(t) - u^2\left(\frac{k}{N}\right) \right) dt \right| \\ &\leq 2 \sup_{t \in (0,1]} |u(t)| \frac{1}{N} \sum_{k=1}^N \sup_{t \in \left(\frac{k-1}{N}, \frac{k}{N}\right]} \left| u(t) - u\left(\frac{k}{N}\right) \right| \leq 2L'^2 N^{-\beta}. \end{aligned}$$

Here we used the inequality $||a| - |b|| \leq \|a - b\|$ for normed spaces and the estimate $\sup_{t \in [0,1]} |u(t)| \leq L'$ that follows from the inclusion $w \in \Sigma(\beta, L')$ and the equality

$$\int_{[0,1]} u(t) dt = 0.$$

Therefore

$$|1 - \|\tilde{u}\|_N| \leq C'_1 N^{-\beta} + L' N^{-\alpha\beta}.$$

Thus the assumptions of Lemma 2.1 hold for sufficiently large N , whence

$$\begin{aligned} \min_{\sigma=\pm 1} \|\tilde{u}_l - \sigma u_N^*\|_N &\leq C'_1 N^{-\beta} + L' N^{-\alpha\beta} + 4\sqrt{\frac{\delta_N}{C}}, \\ \min_{\sigma=\pm 1} \|u - \sigma u_N^*\|_N &\leq C_1 N^{-\beta} + C_2 N^{-\alpha\beta} + C_3 \sqrt{\delta_N}, \end{aligned}$$

where δ_N is estimated in Lemma 2.2. This completes the proof of the first assertion of the theorem. The second assertion follows from the first one and Lemma 2.4. $\square$

The rate of convergence given in Theorem 2.1 is the same as the rate of convergence for estimators constructed by using the decomposition in the trigonometric basis (see [3]).

3. THE BEHAVIOR OF THE ESTIMATOR CONSTRUCTED FROM SIMULATED DATA

Below we consider two examples. Put $w(t) = 4(t - \frac{1}{2})^2 \mathbb{1}\{t \in [0, 1]\}$ for both examples. Let $\pi(A) = m(A \cap [0, 1])$ for all $A \in \mathcal{B}(\mathbb{R})$ in the case of Example A and

$$\pi(A) = \frac{1}{2} m(A \cap [0, 2])$$

for all $A \in \mathcal{B}(\mathbb{R})$ in the case of Example B. The distribution functions are chosen so that

$$\begin{aligned} \text{A)} \quad G(x) &= \begin{cases} 0, & x < 0, \\ x, & 0 \leq x \leq 1, \\ 1, & 1 \leq x, \end{cases} \quad H(x) = \begin{cases} 0, & x < 0.25, \\ 2(x - 0.25), & 0.25 \leq x \leq 0.75, \\ 1, & 0.75 \leq x, \end{cases} \\ \text{B)} \quad G(x) &= \begin{cases} 0, & x < 0, \\ x, & 0 \leq x \leq 1, \\ 1, & 1 \leq x, \end{cases} \quad H(x) = \begin{cases} 0, & x < 1, \\ x - 1, & 1 \leq x \leq 2, \\ 1, & 2 \leq x. \end{cases} \end{aligned}$$

A mixture of components with the above distributions and the concentration function $w(t)$ is simulated with the help of Mathematica 5. Based on the mixture simulated by Mathematica 5, the estimator $u_N^*(t)$ is constructed and $\min_{\sigma=\pm 1} \|u - \sigma u_N^*\|_2$ is evaluated. The averaged results of 30–50 experiments are given in the table below for all sets of parameters. The first column (denoted N) contains the size of the sample; the columns denoted $l = 5, 10, 15$, and 20 correspond to the number of indicator functions in the basis used to construct V_l (that is, $V_l = \text{c.l.s.}\{g_k, k = 1, \dots, l\}$, where

FIGURE 1. Averaged distance $\min_{\sigma=\pm 1} \|u - \sigma u_N^*\|_2$ for Example A FIGURE 2. Averaged distance $\min_{\sigma=\pm 1} \|u - \sigma u_N^*\|_2$ for Example B

$g_k(x) = \sqrt{l} \mathbb{1}\{x \in (\frac{k-1}{l}, \frac{k}{l}]\}$; the columns denoted A and B contain the data for the corresponding examples. The entries of the second row of the table contain the distances $\|u - \Pr_{V_l} u\|_2$ corresponding to $N = 5, 10, 15$, or 20 .

N	5		10		15		20	
	0.440		0.223		0.149		0.112	
	A	B	A	B	A	B	A	B
900	0.497	0.463	0.397	0.259	0.424	0.233	0.459	0.232
1200	0.477	0.459	0.345	0.256	0.367	0.212	0.371	0.206
1500	0.472	0.458	0.302	0.248	0.317	0.200	0.315	0.200
1800	0.464	0.456	0.302	0.241	0.301	0.194	0.294	0.174
2100	0.466	0.454	0.282	0.240	0.255	0.184	0.284	0.175

The results for both examples are depicted in the pictures for three values of l : the lines corresponding to the cases of $l = 10$, $l = 15$, and $l = 20$ are marked with diamonds, stars, and squares, respectively. The axes Y and N contain distances and sizes of samples.

4. CONCLUDING REMARKS

We considered the problem of construction of projection estimators for the shape of the concentration function with the help of the expansion in the histogram basis for the data being a sample from a mixture of two components with unknown distributions whose concentrations vary with observations. We proved that the estimators are consistent and obtained the almost sure rate of convergence. It is shown that the rate of convergence of the estimates constructed in the histogram basis is of the same rate as the rate of convergence for the estimators constructed in the trigonometric basis.

The results are checked for the simulated data. In the case of two distributions that easily can be separated (Example B) the error is small even for small samples; the error decreases as the number of functions in the basis increases.

If the distributions are not too different (Example A), the error is quite large for small sizes (it is almost a half of the norm of the function to be estimated). The error decreases as the number of the basis functions increases; but the rate of decrease is lower than that in Example B and becomes lower as the size of the sample increases. The errors evaluated for the simulated data are not small. This is not surprising, since the assumptions on the model are minimal; we only assume that the distribution functions can be distinguished in some given measure and that the concentration function is Lipschitz.

BIBLIOGRAPHY

1. R. E. Maiboroda, *An asymptotically effective probability estimator constructed from observations of a mixture*, Teor. ĭmovir. Mat. Stat. **59** (1998), 108–115; English transl. in Theory Probab. Math. Statist. **59** (1999), 121–128. MR1793771
2. ———, *Nonparametric Statistics of Nonhomogeneous Observations*, Doctoral dissertation, Kyiv, 1994. (Ukrainian)
3. ———, *Projection estimates for changing concentrations of mixtures*, Teor. ĭmovir. Mat. Stat. **46** (1992), 70–76; English transl. in Theory Probab. Math. Statist. **46** (1993), 71–73. MR1196209 (93j:62102)
4. ———, *An estimate of the distributions of the mixture components with varying concentrations*, Ukrain. Mat. Zh. **48** (1996), no. 4, 562–566; English transl. in Ukrainian Math. J. **48** (1997), 618–622. MR1417019 (97j:62055)
5. ———, *Statistical Analysis of Mixtures*, Kyiv University, Kyiv, 2003. (Ukrainian)
6. C. K. Chui, *Wavelets: A Mathematical Tool for Signal Processing*, SIAM, Philadelphia, PA, 1997. MR1443204 (99b:42012)
7. I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1996. MR1162107 (93e:42045)
8. W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, 1990. MR1161622 (93i:62030)
9. W. Härdle, G. Kerkycharian, D. Picard, and A. Tsybakov, *Wavelets, Approximation, and Statistical Applications*, Springer-Verlag, New York, 1998. MR1618204 (99f:42065)
10. R. E. Maiboroda, *Estimation and classification by mixtures with time-dependent concentrations*, VI International Vilnius Conference on Probability Theory and Math. Statistics. Abstracts of Communications, vol. 2, 1993, p. 48.
11. O. V. Sugakova, *Asymptotics of a kernel estimate for distribution density constructed from observations of a mixture with varying concentration*, Teor. ĭmovir. Mat. Stat. **59** (1999), 156–166; English transl. in Theory Probab. Math. Statist. **59** (2000), 161–171. MR1793776
12. B. Vidakovic, *Statistical Modeling by Wavelets*, John Wiley & Sons, New York, 1999. MR1681904 (2000f:42023)

DEPARTMENT OF PROBABILITY THEORY AND MATHEMATICAL STATISTICS, FACULTY FOR MATHEMATICS AND MECHANICS, NATIONAL TARAS SHEVCHENKO UNIVERSITY, ACADEMICIAN GLUSHKOV AVENUE 6, KYIV 03127, UKRAINE

E-mail address: pokhid@ukr.net

Received 26/MAY/2004

Translated by S. KVASKO