Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Non-parametric adaptive estimation of a multivariate density [⋆]

Vyacheslav A. Vasiliev [*]

[*] *Department of Applied Mathematics and Cybernetics, Tomsk State University, Lenina 36, 634050 Tomsk, Russia (e-mail: vas@mail.tsu.ru).*

Abstract: The properties of adaptive non-parametric kernel estimators for the multivariate probability density $f(x)$ (and its derivatives) of identically distributed random vectors $\varepsilon_n,\ n \geq 1$ at a given point are studied. It is supposed that the vectors $\varepsilon_n,\ n \geq 1$ form a martingale-difference process $(\varepsilon_n)_{n\geq 1}$ and the function to be estimated belongs to a class of densities slightly narrower than the class of densities with the following condition on the highest derivatives of the order ν :

$$|f^{(\nu)}(y) - f^{(\nu)}(x)| \leq \Delta(\|x-y\|), \qquad x, y \in \mathcal{R}^m,$$

where $\Delta(t),\ t \geq 0$, is some positive, bounded from above, monotonously increasing for t, small enough unknown function.

An asymptotic mean square criterion is proposed. The optimality, in asymptotically minimax sense of adaptive estimators of density derivatives, is proved for a class of the Bartlett kernel estimators with a random data-driven bandwidth.

It's well-known that the optimization of the asymptotic value of the mean squared error for the Bartlett kernel density estimators leads to the optimal bandwidth depending on unknown functions. Therefore it is not quite simple to apply these estimators to practice.

The paper proposes an adaptive approach to this problem, which is based on the idea of changing the unknown functions in optimal bandwidth by a sequence of estimators converging to the unknown values of these functions. It is shown, that the constructed adaptive kernel estimators keep all the asymptotic properties of the sharp-optimal non-adaptive Bartlett estimators.

An example of the adaptive estimator, optimal in the sense of the introduced criterion is considered. This estimator has simple structure and may be easily used in real statistical problems. The proposed estimators possess the property of uniform asymptotic normality and almost sure convergence.

1. INTRODUCTION

An important problem in applied and theoretical research is studying the properties of non-parametric estimators of multivariate probability density functions (p.d.f.'s).

Along with the estimation of the p.d.f., the estimation of partial derivatives of a multivariate p.d.f. is of interest. These derivatives are needed in many statistical problems, for example, in estimation of the Fisher information matrix, optimal Bayes estimation of the vector parameter of an exponential distribution, when the prior distribution is unknown, see, for example, Singh (1976). Generally, this problem is important for the construction of stochastic models, including modelling problems of control systems.

Let us consider in more detail the p.d.f. estimation problems. There exist many results on this subject, concerning consistency in different senses for the proposed estimators (see, for example, Delecroix (1996), Devroye and Györfi (1985), Koshkin and Vasiliev (1998), Politis (2003), Pracasa Rao (1983), Pracasa Rao (1996), Singh (1981)). The notion of asymptotic optimality is usually associated with the optimal convergence rate of the minimax risk (see, for example, Ibragimov and Khasminskii (1981) and Stone (1982)). An important question in the development of the non-parametric estimation is to study the exact asymptotic behaviour of the minimax risk and to find an efficient estimator, i.e. an estimator which achieves this asymptote. By applying the minimax approach, it is supposed that the function to be estimated belongs to some class of functions, for example, Hölder, Sobolev, Besov, and so on (see, for example, Devroye and Györfi (1985), Ibragimov and Khasminskii (1981), Stone (1982) among others). If the parameters in the definitions of these classes are unknown, the estimation problem can be treated as a problem of adaptation.

The most general approach to the adaptive estimation problem of a scalar function at a given point in the minimax sense has been developed by Lepski (1990)–Lepski (1992) (see Lepski and Spokoiny (1997) as well).

In papers by Lepski and Spokoiny (1997) the problem of an adaptive bandwidth selection in kernel estimation with a given type of kernel was considered. In particular, the case when a function (of a scalar argument) to be estimated belongs to a given Hölder class $\Sigma(\beta)$ with the unknown smoothness parameter $\beta \leq 2$ was investigated. The p.d.f. from the class $\Sigma(\beta)$ at a given point can be

[⋆] Research was partly supported by RFBR - DFG 05-01-04004 Grant

estimated with the accuracy $n^{-2\beta/2\beta+1}$. At the same time this accuracy is impossible to attain if the parameter β is unknown. The optimal adaptive convergence rate of estimators was calculated in Lepski (1990). It occured to be $(n^{-1}\sqrt{\ln n})^{2\beta/2\beta+1}$ that differs from the non-adaptive one (when the parameter β is known) by the extra log-factor, see also Brown and Low (1996). It was proved in Lepski and Spokoiny (1997) that this estimation procedure is sharp optimal in the adaptive sense over the class of all feasible estimators not only of kernel type. It should be pointed out that for non-adaptive pointwise estimation linear methods are not sharp optimal (see Lepski and Spokoiny (1997), Sacks and Strawderman (1982)).

This talk deals with non-parametric estimation of the multivariate p.d.f. and its derivatives in the case when a function to be estimated belongs to a given class $\Sigma(\nu,\Delta)$ with the following condition on its highest derivative of the ν-th order

$$|f^{(\nu)}(y) - f^{(\nu)}(x)| \le \Delta(\|x-y\|), \qquad x, y \in \mathcal{R}^m, \quad (1)$$

where $\Delta(t),\ t \ge 0$ is some positive, bounded above, monotonously increasing for t small enough function (see Definition 2.1 below). It should be noted, that the condition (1) can be weaker than, for example, the corresponding Lipshitz condition in the definition of the Hölder class $\Sigma(\beta),\ \beta > \nu$.

The order $\nu \ge 1$ of the highest derivative in the definition of $\Sigma(\nu,\Delta)$ is supposed to be known and the function $\Delta(\cdot)$ is assumed to be unknown. The case of unknown order ν could be considered in the future.

We shall investigate an adaptive estimation problem of the partial derivatives of the order α of m-dimensional p.d.f. in the following sense.

First, we consider so-called estimators with reduced bias, see Bartlett (1963), or, by the terminology in Devroye and Györfi (1985), the Bartlett estimators only. It is well-known (see Bartlett (1963), Devroye and Györfi (1985), Epanechnikov (1969) among others), that, by making use of a special class of kernels, we can get the Bartlett estimators of $f \in \Sigma(\nu,\Delta)$ with the principal term of their mean square error (MSE), which do not depends on the unknown function $\Delta(\cdot)$. As follows, such estimators have the rate of convergence equal to $n^{-\frac{2\nu}{m+2\nu}}$, which may differ from the optimal one on the class $\Sigma(\nu,\Delta)$. At the same time this convergence rate can be arbitrary close to the optimal one by appropriate chosen function Δ in the definition of $\Sigma(\nu,\Delta)$. The optimization of the principal term of the MSE of Bartlett's estimators leads to the dependence of their bandwidth from the function to be estimated and its partial derivatives of the order ν. There are different ways to solve the problem of adaptation to this lack of knowledge (see, for example, Berlinet and Devroye (1994), Deheuvels and Hominal (1980), Devroye and Györfi (1985), Donoho (1994), Politis (2003)).

We consider an adaptive approach to this problem, which assumes the usage of non-parametric estimators for these unknown functions in the construction of the optimal bandwidth.

In this talk we propose an asymptotically minimax criterion for the adaptive Bartlett kernel-type estimators of the derivative $f^{(\alpha)}$ of the density $f \in \Sigma(\alpha+\nu,\Delta)$ with random data-driven bandwidth (see formula (10) below), which gives the exact lower bound for the MSE over the class of densities, somewhat narrower than the class $\Sigma(\alpha+\nu,\Delta)$. It is shown that the adaptive rate of convergence is equal to the optimal non-adaptive one, $n^{-\frac{2\nu}{m+2(\alpha+\nu)}}$, of the Bartlett estimators (when the bandwidth of the Bartlett estimators is non-random and unknown).

An example of the optimal estimator in the sense of introduced criterion is considered. The properties of uniform asymptotic normality and almost sure convergence of all presented estimators are investigated.

2. PROBLEM SETTING

Let $\{\mathcal{F}_n\}_{n\ge 0}$ be a filtration in a probability space $(\Omega,\mathcal{F},\mathsf{P})$ and let $\varepsilon = (\varepsilon_n)_{n\ge 1}$ is a martingale-difference process with identically distributed random vectors $\varepsilon_n = (\varepsilon_{n1},\ldots,\varepsilon_{nm})'$ having an unknown p.d.f. $f(\cdot)$, adapted to $\{\mathcal{F}_n\}$, be given (a prime denotes the transposition).

For a fixed vector of nonnegative integers $a = (\alpha_1,\ldots,\alpha_m)$, we consider the estimation problem of a partial derivative

$$f_a^{(\alpha)}(x) = \frac{\partial^\alpha f(x)}{\partial x_1^{\alpha_1}\ldots\partial x_m^{\alpha_m}}, \qquad f_0^{(0)}(x) = f(x)$$

of a p.d.f. $f(x)$ from observations ε at a given point $x \in \mathcal{R}^m$, where $\alpha_1+\alpha_2+\ldots+\alpha_m = \alpha$.

Now we give some needed notation and definitions.

Denote by $\beta(k)$ the set of all vectors $b = (\beta_1,\ldots,\beta_m)$ with nonnegative integer-valued components $\beta_1,\ldots,\beta_m$ such that $\beta_1+\ldots+\beta_m = k$. Omitting the subscript $b = (\beta_1,\ldots,\beta_m)$ of partial derivatives $f_b^{(k)}(x)$ will mean that the set of indices $\beta_1,\ldots,\beta_m$ is not specified.

In the sequel, we denote $c, c_i, C, C_i,\ i = 1,2,\ldots$ as nonnegative constants, possibly different even within the same index.

Definition 2.1. Let a density $f(x),\ x \in \mathcal{R}^m$ be r times differentiable in $\mathcal{R}^m$. We say that a p.d.f. $f(x)$ belongs to the class $\Sigma(r,\Delta)$ if all its partial derivatives of order $r \ge 0$ satisfy the following condition:

$$|f^{(r)}(x) - f^{(r)}(y)| \le \Delta(\|x-y\|), \qquad x, y \in \mathcal{R}^m,$$

where $\Delta(t),\ t \ge 0$ is some positive, possibly unknown, bounded from above, monotonously increasing of t small enough function, i.e. exists some $t_0 > 0$, such that $\Delta(t_1) \le \Delta(t_2)$ for all $t_1 \le t_2 \le t_0$ and $\Delta(0) = 0$, $\|x\|^2 = \sum_{j=1}^m x_j^2$.

We shall denote in the sequel by ν a positive integer, which means the degree of differentiability of the function $f_a^{(\alpha)}(\cdot)$.

As an example of the function $\Delta(\cdot)$ in the definition of $\Sigma(r,\Delta)$, we can take

$$\Delta_\gamma(t) = \frac{\mathcal{L}}{(1+|\ln t|)^\gamma},$$

where $\mathcal{L}$ and γ are some unknown positive constants. Examples of more slowly decreasing functions $\Delta(\cdot)$ can be considered as well.

For positive integer ν we define the following quantities:

$$T_k = \int\limits_{\mathcal{R}^m} u_1^{\alpha_1}\dots u_m^{\alpha_m} K(u)\,du, \qquad \sum_{j=1}^{m}\alpha_j = k,$$

$$T^\nu = (T_1^\nu,\dots,T_s^\nu), \qquad T_j^\nu = \int\limits_{\mathcal{R}^m} u_j^\nu K(u)\,du,$$

$$\omega_f(x) = \frac{(-1)^\nu}{\nu!}\sum_{i=1}^{m} T_i^\nu f^{(\alpha+\nu)}_{a+b_i(\nu)}(x),\ L = \int\limits_{\mathcal{R}^m}\left(K_a^{(\alpha)}(u)\right)^2 du,$$

where $b_i(\nu) = \nu(\delta_{i1},\dots,\delta_{im})$,‘ $i = 1,\dots,m$; δ_{ij} is the Kronecker delta.

Definition 2.2. We say that kernel function $K(\cdot)$ belongs to class $\mathcal{B}^0$ if it is finitely supported, continuously differentiable up to the order α (inclusive), and

$$\int\limits_{\mathcal{R}^m} K(z)\,dz = 1, \quad T_j = 0 \ \text{ for } \ j = 1,\dots,\nu.$$

We say that function $K(\cdot)$ belongs to class $\mathcal{B}$ if in the definition of $\mathcal{B}^0$ we put $T^\nu \neq 0$ and $T_j = 0$ for $j = 1,\dots,\nu$, $\alpha_i < \nu$, $i = 1,\dots,m$.

Definition 2.3. We say that kernel function $K(\cdot)$ belongs to class $\mathcal{B}^*$ if $K(\cdot) \in \mathcal{B}$ and

$$\int\limits_{\mathcal{R}^m} |K(z)|\,dz \le C_1, \qquad \sup_{\mathcal{R}^m}|K_a^{(\alpha)}(z)| \le C_2$$

for some constants C_1 and C_2.

We say that class $\mathcal{B}_\nu^*$ consists of kernel functions $K(\cdot) \in \mathcal{B}^*$, such that T^ν is a fixed known vector.

Define for some (possibly unknown) positive constants c and C the following sets of functions:

$$\widetilde{\Sigma}(\alpha+\nu,\Delta) = \{f \in \Sigma(\alpha+\nu,\Delta):\ \max_{\underline{k=0,\alpha+\nu}} |f^{(k)}(x)| \le C\},$$

$$\Sigma^*(\alpha+\nu,\Delta) = \{f \in \widetilde{\Sigma}(\alpha+\nu,\Delta):\ f(x) \ge c,\ \omega_f^2(x) \ge c\}.$$

Denote by $\mathcal{H}$ the set of monotonously decreasing sequences $h = (h_n)_{n\ge1}$ of real numbers $h_n > 0$ satisfying the condition

$$\lim_{n\to\infty}\left(h_n + (nh_n^{m+2\alpha})^{-1}\right) = 0.$$

Define the set $\Theta_n(\mathcal{H},\mathcal{B})$ of kernel estimators with a bandwidth $h:$

$$\Theta_n(\mathcal{H},\mathcal{B}) = \{f_{a,n}^{(\alpha)}(x) := \frac{1}{nh_n^{m+\alpha}}\sum_{i=1}^{n} K_a^{(\alpha)}\left(\frac{x-\varepsilon_i}{h_n}\right),$$

$$h = (h_n)_{n\ge1} \in \mathcal{H},\ K \in \mathcal{B}\}.$$

Consider the set of estimators $\Theta_n^0 = \Theta_n(\mathcal{H},\mathcal{B}^0)$ and the set $\Theta_n^* = \Theta_n(\mathcal{H},\mathcal{B}_\nu^*)$ of the Bartlett kernel estimators, see Bartlett (1963), Devroye and Györfi (1985). Denote $A = \sup\limits_{z\in\mathcal{R}^m}\{\|z\|:\ K(z) \neq 0\}$ (we shall suppose in the sequel, for simplification, that $A > 1$).

As for the functions from the Hölder class $\Sigma(\beta)$, $\beta > \alpha+\nu$, we can find the principal term of the MSE $u_f^2(f_{a,n}^{(\alpha)}(x)) = \mathsf{E}_f(f_{a,n}^{(\alpha)}(x) - f_a^{(\alpha)}(x))^2$ for the kernel-type estimators $f_{a,n}^{(\alpha)}(x)$ from the class Θ_n^0 for the function $f \in \Sigma(\alpha+\nu,\Delta)$, as $n \to \infty:$

$$u_f^2(f_{a,n}^{(\alpha)}(x)) \sim \frac{1}{nh_n^{m+2\alpha}} + h_n^{2\nu}\cdot\Delta^2(Ah_n), \tag{2}$$

where the first summand on the right-hand side in (2) is proportional to the second moment of the stochastic term in the decomposition of the deviation of the estimator $f_{a,n}^{(\alpha)}(x)$ and the second one - to the upper bound of the squared bias of $f_{a,n}^{(\alpha)}(x)$.

Consider our example $f \in \Sigma(\alpha+\nu,\Delta_\gamma)$ (with the function $\Delta_\gamma(\cdot)$ of known structure). The optimization of the right-hand side in (2) on the bandwidth $h \in \mathcal{H}$ gives the optimal rate of convergence for the MSE:

$$u_f^2(f_{a,n}^{(\alpha)}(x)) \sim (n^\nu \ln^{\gamma(m+2\alpha)} n)^{-\frac{2}{m+2(\alpha+\nu)}} \tag{3}$$

and corresponding bandwidth is proportional to

$$h_n \sim (n^{-1}\ln^{2\gamma} n)^{\frac{1}{m+2(\alpha+\nu)}}.$$

The structure of the function $\Delta(\cdot)$ is unknown, in general, and it is not enough of a'priori information for its estimation. Thus, the optimal bandwidth can not be found from (2).

The standard bandwidth choice for kernels of the type $\mathcal{B}_\nu^*$ and for the mean square loss function is motivated by the balance relation between the principal term of the bias and of the second moment of stochastic term in the decomposition of the MSE for kernel estimators. The principal term of the bias depends on h_n^ν and highest derivatives $f^{(\alpha+\nu)}(x)$ of the function $f(x)$ and the second moment of stochastic term is proportional to $\dfrac{f(x)}{nh_n^{m+2\alpha}}$. Then the minimization on the bandwidth of the principal term of the MSE expansion for the Bartlett kernel estimator $f_{a,n}^{(\alpha)}(x) \in \Theta_n^*$ of the derivative $f_a^{(\alpha)}(x)$, $f(x) \in \Sigma(\alpha+\nu,\Delta)$ gives the following expression (see Bartlett (1963), Devroye and Györfi (1985), Epanechnikov (1969) and Politis (2003) as well), as $n \to \infty$, for the MSE:

$$u_f^2(f_{a,n}^{(\alpha)}(x)) = n^{-\frac{2\nu}{m+2(\alpha+\nu)}}(c_f^{opt} + o(1)). \tag{4}$$

At that $\qquad c_f^{opt} = (m+2(\alpha+\nu))\cdot$

$$\cdot\left(\frac{L^o f(x)}{2\nu}\right)^{\frac{2\nu}{m+2(\alpha+\nu)}}\cdot\left(\frac{\omega_f^2(x)}{m+2\alpha}\right)^{\frac{m+2\alpha}{m+2(\alpha+\nu)}}, \tag{5}$$

where $L^o = \inf\limits_{K\in\mathcal{B}_\nu^*} L$ (here the infimum L^o is assumed to be attained).

The estimator $f_{a,n}^{(\alpha),o}(x)$, satisfying (4) is given by the formula

$$f_{a,n}^{(\alpha),o}(x) = \frac{1}{n(h_n^o)^{m+\alpha}}\sum_{i=1}^{n}(K^o)_a^{(\alpha)}\left(\frac{x-\varepsilon_i}{h_n^o}\right) \tag{6}$$

and has the bandwidth $h^o = (h_n^o)_{n\ge1}$, defined as

$$h_n^o = n^{-\frac{1}{m+2(\alpha+\nu)}} s_{opt}(x), \tag{7}$$

where

$$s_{opt}(x) = \left(\frac{L^o(m+2\alpha)f(x)}{2\nu\omega_f^2(x)}\right)^{\frac{1}{m+2(\alpha+\nu)}}. \tag{8}$$

Condition $T^\nu \neq 0$ in the definition of the Bartlett type kernel class $\mathcal{B}_\nu^*$ excludes the dependence of the bandwidth

h^0 from the unknown function $\Delta(\cdot)$. On the other hand, estimator (6) from Bartlett class Θ_n^* for function $f \in \Sigma(\alpha + \nu, \Delta)$ to be estimated has non-optimal rate of convergence, while estimators from class Θ_n^0 have optimal rate. At the same time, in our example $\Delta(\cdot) = \Delta_\gamma(\cdot)$ it follows from the comparison of the formulae (3) and (4), that the optimal convergence rate $n^{-\frac{2\nu}{m+2(\alpha+\nu)}}$ of Bartlett's estimators differs from the rate $(n^\nu \ln^{\gamma(m+2\alpha)} n)^{-\frac{2}{m+2(\alpha+\nu)}}$ of (non-adaptive) estimators (6) with $K \in \mathcal{B}^0$ on the extra log-factor $(\ln n)^{\frac{2\gamma(m+2\alpha)}{m+2(\alpha+\nu)}}$ only.

The functions $s_{opt}(x)$ and h^o depend on the unknown functions $f(x)$ and $f^{(\alpha+\nu)}(x)$. As follows, the function $f_{a,n}^{(\alpha),o}(x)$ is unobservable and can not be used as an estimator.

Then for solving the problem of adaptation to unknown parameter h^o in the formula (6), it's enough to estimate, similarly to Devroye and Györfi (1985) among others, the unknown function f and its derivatives $f^{(\alpha+\nu)}$ in the definition (8) of the function s_{opt} by the non-parametric or by the cross-validation method, see Berlinet and Devroye (1994).

To describe the class of adaptive estimators of density derivatives we need the following notation. Define sets $\mathcal{H}$ and S of deterministic and respectively random sequences. Specify the set of random bandwidths in the form:

$$\widetilde{\mathcal{H}} = \widetilde{\mathcal{H}}(\mathcal{H}, S) = \{\widetilde{h} = (h_{i-1,n})_{i=\overline{1,n}, n\geq 1} : \ h_{i-1,n} = v_n s_{i-1},$$
$$i = \overline{1,n}, n \geq 1, \ \widetilde{s} = (s_i)_{i\geq 0} \in S(\widetilde{v}), \ \widetilde{v} = (v_n)_{n\geq 1} \in \mathcal{H}\}.$$

For investigating of the optimality properties of adaptive estimators, we consider, instead of the class of estimators $\Theta_n(\mathcal{H}, \mathcal{B})$, the class $\widetilde{\Theta}_n(\widetilde{\mathcal{H}}, \mathcal{B})$ of kernel estimators with a random data-driven bandwidth $\widetilde{h} \in \widetilde{\mathcal{H}}$:

$$\widetilde{\Theta}_n(\widetilde{\mathcal{H}}, \mathcal{B}) = \{f_{a,n}^{(\alpha)}(x) :$$
$$f_{a,n}^{(\alpha)}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_{i-1,n}^{m+\alpha}} K_a^{(\alpha)}\left(\frac{x - \varepsilon_i}{h_{i-1,n}}\right), \widetilde{h} \in \widetilde{\mathcal{H}}, K \in \mathcal{B}\}.$$

From the definition of the kernel estimators set $\widetilde{\Theta}_n$ follows, that their bandwidth $\widetilde{h}$ has the structure (7) with a random sequence $\widetilde{s}$ (for example, with a sequence of estimators for s_{opt}) instead of the unknown parameter s_{opt}. Section 4 gives an example of such estimators for s_{opt}.

We shall define the set S in a different way according to the aims of investigation (see Definition 2.4 below).

Let $\widetilde{\delta} = (\delta_n)_{n\geq 1}$ and $\tilde{r} = (r_n)_{n\geq 1}$ be a sequences of positive numbers, decreasing to zero and monotonously increasing to infinity respectively.

For $n \geq 1$ we put $d_n = n^{\frac{1}{m+2(\alpha+\nu)}} v_n$ and $\widetilde{\Delta} = (\Delta_n)_{n\geq 1}$: $\Delta_n = \Delta(A r_n^{-1})$, $n \geq 1$. Note, that in our example $\Delta = \Delta_\gamma : \Delta_n \sim (\ln r_n)^{-\gamma}$.

Let us suppose, that sequences $\widetilde{\delta}$ and $\widetilde{r}$ satisfy, as $n \to \infty$, the following conditions:

$$\delta_n \cdot n^{\frac{1}{m+2(\alpha+\nu)}} \to \infty, \quad \delta_n / \Delta_n^2 \to \infty, \quad r_n \cdot n^{\frac{1}{m+2(\alpha+\nu)}} \to 0.$$

Denote $a \wedge b = \min(a,b)$, $a \vee b = \max(a,b)$.

Definition 2.4. We say that a sequence $\widetilde{s} = (s_i)_{i\geq 0}$ of positive $\{\mathcal{F}_i\}$-adapted random variables $s_i, i \geq 0$,

i) belongs to set $S(\Sigma, \widetilde{v})$ if for $f \in \Sigma(\alpha + \nu, \Delta)$ it has some non-random limit $s \in (0, \infty)$ in the following sense:

$$\frac{1}{n}\sum_{i=1}^{n} |\mathsf{E}_f[s_{i-1}^{-(m+2\alpha)} - s^{-(m+2\alpha)}]| = o(1),$$
$$\frac{1}{n}\sum_{i=1}^{n} \mathsf{E}_f(s_{i-1}^{\nu} - s^{\nu})^2 = o(1) \ \text{ as } n \to \infty$$

and, besides,

$$\frac{1}{n}\sum_{i=1}^{n} \mathsf{E}_f[s_{i-1}^{2\nu} + s_{i-1}^{1-(m+2\alpha)}] = o(v_n^{-1}), \quad n \to \infty,$$
$$\sup_{i<n} s_i \leq (r_n v_n)^{-1}, \quad n \geq 1; \tag{9}$$

ii) belongs to set $S^*(\widetilde{\Sigma}, \widetilde{v})$ if for some $s \in (0, \infty)$ for all $n \geq 1$

$$\frac{1}{n}\sum_{i=1}^{n} \sup_{f\in\widetilde{\Sigma}(\alpha+\nu,\Delta)} |\mathsf{E}_f[s_{i-1}^{-(m+2\alpha)} - s^{-(m+2\alpha)}]| \leq C\delta_n d_n^{m+2\alpha},$$
$$\frac{1}{n}\sum_{i=1}^{n} \sup_{f\in\widetilde{\Sigma}(\alpha+\nu,\Delta)} \mathsf{E}_f(s_{i-1}^{\nu} - s^{\nu})^2 \leq C\delta_n d_n^{-2\nu}$$

and, besides, the relations (9) hold true and for some $\beta \in (0, 1]$

$$\frac{1}{n}\sum_{i=1}^{n} \sup_{f\in\widetilde{\Sigma}(\alpha+\nu,\Delta)} \mathsf{E}_f[s_{i-1}^{2\nu} + s_{i-1}^{-(m+2\alpha)}]$$
$$\leq C\delta_n[(d_n^{m+2\alpha} v_n^{-1}) \wedge (d_n^{-2\nu}\Delta_n^{-2})] \wedge (\delta_n^{\beta-1} d_n^{m+2\alpha}).$$

This definition allows the dependence of the limits $s \in (0, \infty)$ on the unknown function f.

We introduce the sets $\widetilde{\Theta}(n) = \widetilde{\Theta}_n(\widetilde{\mathcal{H}}(\Sigma), \mathcal{B})$, $\widetilde{\Theta}_\nu^*(n) = \widetilde{\Theta}_n(\widetilde{\mathcal{H}}^*(\Sigma^*), \mathcal{B}_\nu^*)$, $\widetilde{\Theta}^*(n) = \widetilde{\Theta}_n(\widetilde{\mathcal{H}}^*(\widetilde{\Sigma}), \mathcal{B}^*)$, where $\widetilde{\mathcal{H}}(\Sigma) = \widetilde{\mathcal{H}}(\mathcal{H}, S(\Sigma))$, $\widetilde{\mathcal{H}}^*(\Sigma) = \widetilde{\mathcal{H}}(\mathcal{H}, S^*(\Sigma))$.

Define for all kernels $K(\cdot) \in \mathcal{B}_\nu^*$ the following set

$$\Sigma^{**}(\alpha+\nu, \Delta) = \{f \in \Sigma(\alpha+\nu, \Delta) : \ f(x) \neq 0, \ \widetilde{\omega}_f^2(x) \neq 0\}.$$

It will be observed that the optimal bandwidth h^o belongs to all the sets $\widetilde{\mathcal{H}}(\Sigma^*)$, $\widetilde{\mathcal{H}}(\Sigma^{**})$, $\widetilde{\mathcal{H}}^*(\Sigma^*)$, $\widetilde{\mathcal{H}}^*(\Sigma^{**})$.

We prove that the proposed estimation procedure is sharp optimal in the adaptive sense over the class $\widetilde{\Theta}_\nu^*(n)$ of the Bartlett kernel type estimators with random data-driven bandwidth. Namely, in Section 4 (Theorem 4.1) the following assertion for the normalized MSE $u_{f,n}^2(f_{a,n}^{(\alpha)}) = n^{\frac{2\nu}{m+2(\alpha+\nu)}} u_f^2(f_{a,n}^{(\alpha)}(x))$ will be established:

$$\lim_{n\to\infty} \inf_{f_{a,n}^{(\alpha)} \in \widetilde{\Theta}_\nu^*(n)} \sup_{f\in\Sigma^*(\alpha+\nu,\Delta)} (c_f^{opt})^{-1} u_{f,n}^2(f_{a,n}^{(\alpha)}) = 1, \tag{10}$$

where c_f^{opt} is the constant, defined in (5).

3. CONVERGENCE IN THE MEAN SQUARE

In this section we find the principal term of the MSE $u_f^2(f_{a,n}^{(\alpha)})$ and of the bias $b_f(f_{a,n}^{(\alpha)}) = \mathsf{E}_f f_{a,n}^{(\alpha)} - f_a^{(\alpha)}$ of estimators $f_{a,n}^{(\alpha)} \in \widetilde{\Theta}(n)$ and the sharp lower bound for the MSE of estimators $f_{a,n}^{(\alpha)} \in \widetilde{\Theta}_{\nu}^{*}(n)$ in the sense of criterion (10). The quantity c_f^{opt} in (10) is defined in (5).

Put $b_{f,n}(f_{a,n}^{(\alpha)}) = n^{\frac{\nu}{m+2(\alpha+\nu)}} b_f(f_{a,n}^{(\alpha)})$, $L_1 = f(x)L$, $\widetilde{L}_1 = s^{-(m+2\alpha)} L_1$, $\widetilde{L}_2 = s^{\nu}\omega_f(x)$.

Theorem 3.1. Assume that $f \in \Sigma(\alpha+\nu, \Delta)$ for some $\nu \geq 1$ and $f_{a,n}^{(\alpha)} \in \widetilde{\Theta}(n)$, $n \geq 1$. Then we have, as $n \to \infty$:

1°. for the MSE $u_f^2(f_{a,n}^{(\alpha)})$ of estimator $f_{a,n}^{(\alpha)}$ the relations

$$u_f^2(f_{a,n}^{(\alpha)}) = \frac{\widetilde{L}_1}{n v_n^{m+2\alpha}} + v_n^{2\nu}\widetilde{L}_2^2 + o\left(v_n^{2\nu} + \frac{1}{n v_n^{m+2\alpha}}\right)$$

and

$$\sup_{f\in\widetilde{\Sigma}(\alpha+\nu,\Delta)} \sup_{f_{a,n}^{(\alpha)}\in\widetilde{\Theta}^*(n)} |u_{f,n}^2(f_{a,n}^{(\alpha)}) - d_n^{-(m+2\alpha)}\tilde{L}_1$$
$$-d_n^{2\nu}\widetilde{L}_2^2| = o(1)$$

hold true;

2°. for the bias $b_f(f_{a,n}^{(\alpha)})$ the following relations

$$b_f(f_{a,n}^{(\alpha)}) = v_n^{\nu}\widetilde{L}_2 + o(v_n^{\nu})$$

and

$$\sup_{f\in\widetilde{\Sigma}(\alpha+\nu,\Delta)} \sup_{f_{a,n}^{(\alpha)}\in\widetilde{\Theta}^*(n)} |b_{f,n}(f_{a,n}^{(\alpha)}) - d_n^{\nu}\widetilde{L}_2| = o(1)$$

hold;

3°. the equality

$$\sup_{f\in\Sigma_1^*(\alpha+\nu,\Delta)} | \inf_{f_{a,n}^{(\alpha)}\in\widetilde{\Theta}_{\nu}^*(n)} u_{f,n}^2(f_{a,n}^{(\alpha)}) - c_f^{opt}| = o(1) \qquad (11)$$

is fulfilled;

4°. the following inequality

$$\varliminf_{n\to\infty} \inf_{f_{a,n}^{(\alpha)}\in\widetilde{\Theta}_{\nu}^*(n)} \sup_{f\in\Sigma^*(\alpha+\nu,\Delta)} (c_f^{opt})^{-1} n^{\frac{2\nu}{m+2(\alpha+\nu)}} u_f^2(f_{a,n}^{(\alpha)}) \geq 1$$

is valid.

It follows from Theorem 3.1 that the bandwidth structure $\widetilde{h} \in \widetilde{\mathcal{H}}(\Sigma)$ yields the best possible rate of convergence of the MSE for the Bartlett kernel estimators for $f \in \Sigma(\alpha+\nu, \Delta)$ (see Bartlett (1963), Devroye and Györfi (1985), Epanechnikov (1969) and Politis (2003)). The proper choice of the bandwidth $\widetilde{h} \in \widetilde{\mathcal{H}}^*(\Sigma^*)$ gives the minimal asymptotic value c_f^{opt}, defined by (5) for the MSE in the sense of the equality (11), which can be considered as an criterion as well. Assertion 4° of Theorem 3.1 gives the lower bound for the risk function in criterion (10).

4. CONSTRUCTION OF OPTIMAL KERNEL ESTIMATORS

In this section we construct an adaptive variant of the optimal bandwidth (7) and the corresponding kernel estimator of unknown derivative $f_a^{(\alpha)}(x)$. We show, that the proposed estimator is optimal in the sense of criteria (10) and (11) (see (14) and (15) below).

In order to obtain these optimal kernel estimators, first we construct an estimator for the function $s = s_{opt}(x)$, defined in (8). To this end, we will use non-parametric kernel estimators for the functions $f(x)$ and $f^{(\alpha+\nu)}(x)$.

4.1 Pilot estimators of $f(x)$ and $f^{(\alpha+\nu)}(x)$

Define $\widehat{f}_i(x)$ and $\widehat{f}_i^{(\alpha+\nu)}(x)$ the estimators of the type (6) with a kernel $K(\cdot) \in \mathcal{B}^0$ and a known non-random bandwidth $h_n = (n^{-1}\Delta_*^{-2}(n))^{\frac{1}{m+2(\alpha+\nu)}}$, where $\Delta_*(n)$ is a known positive function, satisfying the following conditions:

$$\lim_{n\to\infty} \Delta_*(n) = 0, \quad \varliminf_{n\to\infty} \Delta_*^2(n)\cdot n^{\frac{1}{m+2(\alpha+\nu)}} > 0.$$

Then for $f \in \Sigma(\alpha+\nu, \Delta)$

$$\mathsf{E}_f|\widehat{f}_i(x) - f(x)|^2 = o\left(\tilde{\Delta}_i^2\right)$$

as $i \to \infty$, where $\tilde{\Delta}_i^2 = \Delta_*^2(i) + \Delta^2(Ah_i)$ and for all $i \geq 1$

$$\sup_{f\in\widetilde{\Sigma}(\alpha+\nu,\Delta)} \mathsf{E}_f|\widehat{f}_i(x) - f(x)|^2 \leq C\tilde{\Delta}_i^2;$$

$$\mathsf{E}_f|\widehat{f}_{i,a+b_k(\nu)}^{(\alpha+\nu)}(x) - f_{a+b_k(\nu)}^{(\alpha+\nu)}(x)|^2 = O\left(\tilde{\Delta}_i^2\right), \ k = \overline{1,m}$$

as $i \to \infty$, and for all $i \geq 1$, $k = \overline{1,m}$

$$\sup_{f\in\widetilde{\Sigma}(\alpha+\nu,\Delta)} \mathsf{E}_f|\widehat{f}_{i,a+b_k(\nu)}^{(\alpha+\nu)}(x) - f_{a+b_k(\nu)}^{(\alpha+\nu)}(x)|^2 \leq C\tilde{\Delta}_i^2.$$

The estimators $\widehat{f}_i(x)$ are assumed to be constructed by nonnegative kernels and are positive for all $i \geq 1$.

Define projections $\widetilde{f}_i(x)$ and $\widetilde{\omega}_{f,i}(x)$ of $\widehat{f}_i(x)$ and $\widehat{\omega}_{f,i}(x)$, where

$$\widehat{\omega}_{f,i}(x) = \frac{(-1)^{\nu}}{\nu!}\sum_{k=1}^{m} T_k^{\nu}\widehat{f}_{i,a+b_k(\nu)}^{(\alpha+\nu)}(x),$$

on the interval $[\gamma_i, \Gamma_i]$ by the formulae

$$\widetilde{f}_i(x) = (\widehat{f}_i(x) \wedge \Gamma_i) \vee \gamma_i,$$
$$\widetilde{\omega}_{f,i}(x) = \mathrm{sign}(\widehat{\omega}_{f,i}(x)) \cdot [(|\widehat{\omega}_{f,i}(x)| \wedge \Gamma_i) \vee \gamma_i].$$

Here $(\gamma_i)_{i\geq 1}$ and $(\Gamma_i)_{i\geq 1}$ are known monotonic sequences of positive numbers, decreasing to zero and unboundedly increasing respectively, which satisfy, as $i \to \infty$, the following conditions:

$$\Gamma_i = O(\ln \widetilde{\Delta}_i^{-1}), \quad \gamma_i^{-1} = O(\ln \widetilde{\Delta}_i^{-1}).$$

For the function $\Delta = \Delta_\gamma$ and some known $\gamma_* > 0$ we can put

$$\Delta_*(i) = \frac{1}{\ln^{\gamma_*} i}, \quad \tilde{\Delta}_i^2 = \frac{1}{\ln^{2\gamma_*} i} + \frac{1}{\ln^{2\gamma} i}$$

and $\Gamma_i = \gamma_i^{-1} = \ln\ln(i+1)$, $i \geq 1$.

4.2 Estimators of $s_{opt}(x)$

We specify the estimators s_i^* for the function $s = s_{opt}(x)$ in the definition (7) of the optimal bandwidth h^o in the form:

$$s_i^* = c^*\left(\frac{\tilde{f}_i(x)}{\tilde{\omega}_{f,i}^2(x)}\right)^{\frac{1}{m+2(\alpha+\nu)}},$$

where

$$c^* = \left(\frac{L^o(m+2\alpha)}{2\nu}\right)^{\frac{1}{m+2(\alpha+\nu)}}.$$

The following lemma concerns the basic properties of the estimator s_i^*.

Denote $\overline{\Gamma}_i = (\gamma_i^{-1}\Gamma_i)^3$, $\overline{\Delta}_i = \widetilde{\Delta}_i \cdot \overline{\Gamma}_i$, $i \geq 1$. It should be noted, that for $\Delta = \Delta_\gamma$ we can put $\overline{\Gamma}_i = \ln^6 \ln(i+1)$, $i \geq 1$.

Lemma 4.1. The estimator s_i^* of the function $s_{opt}(x)$ has the properties:

1°. for all $i \geq 1$

$$c^*\gamma_i^* \leq s_i^* \leq c^*\Gamma_i^*,$$
$$\gamma_i^* = (\gamma_i\Gamma_i^{-2})^{\frac{1}{m+2(\alpha+\nu)}}, \quad \Gamma_i^* = (\gamma_i^{-2}\Gamma_i)^{\frac{1}{m+2(\alpha+\nu)}};$$

2°. if $f \in \Sigma^{**}(\alpha+\nu, \Delta)$ for some $\nu \geq 1$, then

– for all integer $k = \overline{-(m+2(\alpha+\nu)), m+2(\alpha+\nu)}$

$$\mathsf{E}_f[(s_i^*)^k - s_{opt}^k]^2 = O\left(\overline{\Delta}_i^2\right) \quad \text{as} \quad i \to \infty,$$

– for all $i \geq 1$

$$\sup_{f\in\Sigma^*(\alpha+\nu,\Delta)} \mathsf{E}_f[(s_i^*)^k - s_{opt}^k]^2 \leq C\overline{\Delta}_i^2.$$

4.3 Optimal adaptive bandwidth and density estimators

Now we define the sequence $h^* = (h_{i-1,n}^*)_{i,n\geq 1}$ as follows

$$h_{i-1,n}^* = n^{-\frac{1}{m+2(\alpha+\nu)}} s_{i-1}^*, \; s_0^* = 1 \tag{12}$$

and the corresponding kernel estimator $f_{a,n}^{(\alpha),*}(x)$ of $f_a^{(\alpha)}(x)$ as

$$f_{a,n}^{(\alpha),*}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(h_{i-1,n}^*)^{m+\alpha}} (K^o)_a^{(\alpha)} \left(\frac{x-\varepsilon_i}{h_{i-1,n}^*}\right). \tag{13}$$

Theorem 4.1 claim the asymptotic optimality of the bandwidth selector (12) and the estimator (13) in the sense of criteria (10) and (11).

Theorem 4.1.

1°. The estimator $f_{a,n}^{(\alpha),*}$ is optimal:

— in the sense of criterion (11)

$$\lim_{n\to\infty} \sup_{f\in\Sigma^*(\alpha+\nu,\Delta)} |u_{f,n}^2(f_{a,n}^{(\alpha)}) - c_f^{opt}| = 0, \tag{14}$$

where the quantity c_f^{opt} is defined by formula (5) and for the bias $b_f(f_{a,n}^{(\alpha),*})$ we have

$$\lim_{n\to\infty} \sup_{f\in\Sigma^*(\alpha+\nu,\Delta)} |b_{f,n}(f_{a,n}^{(\alpha),*}) - s_{opt}^{\nu}\omega_f(x)| = 0;$$

— in the sense of criterion (10)

$$\lim_{n\to\infty} \sup_{f\in\Sigma^*(\alpha+\nu,\Delta)} (c_f^{opt})^{-1} u_{f,n}^2(f_{a,n}^{(\alpha)}) = 1; \tag{15}$$

2°. the equality (10) holds true:

$$\lim_{n\to\infty} \inf_{f_{a,n}^{(\alpha)}\in\widetilde{\Theta}_\nu^*(n)} \sup_{f\in\Sigma^*(\alpha+\nu,\Delta)} (c_f^{opt})^{-1} u_{f,n}^2(f_{a,n}^{(\alpha)}) = 1.$$

Remark 4.1. We have proved also the properties of uniform asymptotic normality and almost sure convergency for the optimal estimators (13) and for the adaptive estimators $f_{a,n}^{(\alpha)}(x)$ from some classes of estimators, defined similarly to introduced above.

REFERENCES

M. S. Bartlett. Statistical estimation of density function. *J. Statist.*, volume A25, pages 245–254. 1963.

A. Berlinet and L. Devroye. A comparison of kernel density estimates. In Publications de I'I.S.U.P. XXXVIII. fasc., volume 3, pages 3–60. 1994.

L. D. Brown and M. G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, volume 24, pages 2384-2398. 1996.

P. Deheuvels and P. Hominal. Estimation automatique de la densite. *Revue de Statistique Appliguee*, volume 28, pages 25–55. 1980.

M. Delecroix. $\mathcal{L}^2$-consistency of functional parameters estimated under ergodicity assumptions. *Publ. Inst. Stat. Univ. de Paris.*, volume 40, pages 33–56. 1996.

L. Devroye and L. Györfi. *Non-parametric density estimation. The L_1 view.* Wiley, New York, 1985.

D. L. Donoho. Statistical estimation and optimal recovery. *Ann. Statist.*, volume 22, pages 238–270. 1994.

V. A. Epanechnikov. Non-parametric estimation of a multivariate density. *Theory Probab. Appl.*, volume 14, (1), pages 156–162. 1969.

I. A. Ibragimov and R. Z. Khasminskii. *Statistical Estimation: Asymptotic Theory.* Springer, Berlin–New York, 1981.

G. M. Koshkin and V. A. Vasiliev. Non-parametric Estimation of Derivatives of a Multivariate Density from Dependent Observations. *Mathematical Methods of Statistics*, NY, volume 7, (4), pages 361–400. 1998.

O. V. Lepski. One problem of Adaptive Estimation in Gaussian White Noise. *Theor. Probab. Appl.*, volume 35, pages 459–470. 1990.

O. V. Lepski. Asymptotic minimax adaptive estimation. 1. Upper bounds. *Theor. Probab. Appl.*, volume 36, pages 645–659. 1991.

O. V. Lepski. Asymptotic minimax adaptive estimation. 2. Statistical model without optimal adaptation. Adaptive estimators. *Theor. Probab. Appl.*, volume 37, pages 468–481. 1992.

O. V. Lepski and V. G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.*, volume 25, (6), pages 2512–2546. 1997.

D. N. Politis. Adaptive bandwidth choice. *J. Nonparam. Statist.*, volume 15, (4-5), pages 517-533. 2003.

B.L.S. Pracasa Rao. *Nonparametric functional estimation.* Academic Press, Orlando, 1983.

B.L.S. Pracasa Rao. Nonparametric estimation of the derivatives of a density by the method of wavelets. *Bull. Inform. Cyb.*, volume 28, (1), pages 91–100. 1996.

J. Sacks and W. Strawderman. Improvements of linear minimax estimates. In S.S.Gupta and J.O.Berger, eds., *Statistical Decision Theory and Related Topics 3.*, volume 2, Academic Press, New York, pages 287–304. 1982.

R. S. Singh. Non-parametric estimation of mixed partial derivatives of a multivariate density. *Multivariate Analysis*, volume 4, pages 111–122. 1976.

R. S. Singh. Speed of convergence in nonparametric estimation of a multivariate density and its mixed partial derivatives. *J. Stat. Plan. Inf.*, volume 5, pages 287–298. 1981.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, volume 10, pages 1040–1053. 1982.