Tatra Mt. Math. Publ. **17** (1999), 197–207

TATRA MOUNTAINS
Mathematical Publications

NONPARAMETRIC FUNCTION RECOVERY FROM QUALITATIVE DATA BY KERNEL SMOOTHING

Ewaryst Rafajłowicz

ABSTRACT. The problem of function recovery from qualitative observations is stated and its solution by the kernel method is considered. Qualitative observations are collected as yes/no answers to the question whether a function at x is below or above a random threshold. Motivations leading to this problem statement are briefly discussed. Consistency of the kernel smoother was proved based on LIL with random number of summands. A heuristic procedure is proposed for choosing the bandwidth. Simulations illustrate the behaviour of the estimate for moderate sample size.

1. Introduction

Our aim in this paper is to propose a new statement of the problem of function recovery and its solution in the spirit of nonparametric techniques. The novelty is in the form of available observations, which are qualitative in nature. More precisely, let $G\colon D \to \mathbb{R}$ be the unknown, but nonnegative and bounded function on a finite interval $D \subset \mathbb{R}$. Let $X_1, X_2, \dots X_n$ denote i.i.d. random variables with values in D and p.d.f. $f_X(x)$. Further assumptions concerning G will be imposed later. The only data concerning values of G come from answers to the questions: "is it true that $G(X_i) \leq W_i$?", where testing levels W_i are random variables generated artificially by the experimenter. W_i are uniformly distributed in the range of G and independent of X_i. We shall consider the simplest setting of the problem, assuming that X_i are distributed uniformly in D, since generalization to arbitrary density $f_X(x)$ is only technically involved.

The problem is to estimate $G(x)$, $x \in D$, from the above data in the spirit of the nonparametric regression estimation theory, i.e., imposing only smoothness type assumptions on G.

2000 Mathematics Subject Classification: 62G05.
Keywords: nonparametric function estimation, qualitative observations, kernel smoother.

This work was supported by the grant of Council for Scientific Research of Poland.

The proposed method is based on the classical idea of kernel smoother, but with a random number of smoothed data. This fact motivates reconsidering its properties (see Remark 1 below for more detailed explanations).

We remark that the method is applicable in the case when errors are present in values of W_i, but answers to the questions: "is it true that $G(X_i) \leq W_i$?" are assumed to be correct. This estimator does not utilizes values of W_i.

Below we summarize more precisely assumptions concerning G and sketch situations when they can be fulfilled.

Assumptions.

G1) For every $x \in D$, there holds $0 \leq G(x) \leq M$, where $M > 0$ is a priori known constant.

G2) $C_G = \int_D G(x)\, dx > 0$ exists and is finite, but unknown.

G3) Let (X_i, W_i) be a pair of random variables, which are uniformly distributed over the rectangle $B = D \times [0, M]$. For every pair (X_i, W_i) the experimenter is able to decide whether $G(X_i) \leq W_i$ or not. We put

$$Z_i = \begin{cases} 1 & \text{if } G(X_i) \geq W_i\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{1}$$

and consider triplets (X_i, W_i, Z_i), $i = 1, 2, \ldots, n$, as available data.

G4) Random vectors (X_i, W_i), $i = 1, 2, \ldots, n$, are mutually independent.

Assumption G1) is important for the proposed approach, but it can be relaxed in two directions. If G can take negative values, but its lower and upper bounds are a priori known, then one can translate it to fit a rectangle in the upper half plane. It is not necessary to have precise bound for G in G1). M can be a rough upper bound, but its closeness to $\sup_{x \in D} G(x)$ is desirable. In this context it is worth to notice independence of X_i and W_i. This fact has consequences for possible applications, since the experimenter need not to generate X_i artificially. In other words, X_i can be generated by the "nature" and the task of the experimenter is to generate W_i and correctly answer the question whether unknown value $G(X_i)$ is smaller than W_i. This resembles the so called random design case in the nonparametric regression estimation. However, the above explanations do not exclude the case when X_i are also artificially generated random variables.

Motivations. Few words are in order concerning motivation of the problem statement. Consider the following example from the area of destructive testing. Let X denote shape parameter(s) of a steel construction element, which determine its ability to bear a certain load without destruction. It is of interest to determine function $G(X)$, defined as the largest nondestructive loading, when an element with parameter(s) X is used. Data for recovery of G can be obtained as follows. A set of testing elements with randomly chosen parameters X_i, $i = 1, 2, \ldots, n$, is prepared. Then, the sequence of uniformly distributed loads W_i is generated and applied to elements with shape parameters X_i, respectively, $i = 1, 2, \ldots, n$, providing yes/no answers on destruction. Note that for generating W_i uniformly distributed in $[0, M]$ it suffices to know a rough upper bound M for G. However, it is desirable to have a good bound, since it influences the quality of reconstructing G.

It seems that possible applications of the above problem statement are wide, including remote sensing, testing high voltage and high temperature materials, as well as influence of medicines on human body.

2. Kernel smoother and its justification

Let $K(\cdot)$ be a kernel from a large family of kernels used in density estimation (see, e.g., [6], [9], [11]), i.e., a function such that

$$K(x) \geq 0\,, \qquad \int_{-\infty}^{\infty} K(x) = 1\,, \qquad \int_{-\infty}^{\infty} xK(x) = 0\,, \qquad \int_{-\infty}^{\infty} x^2 K(x) = \sigma_K^2 < \infty\,. \tag{2}$$

Let N be the random variable defined as the number of those pairs for which $W_i \leq G(X_i)$, $i = 1, 2, \ldots, n$, i.e.,

$$N = \sum_{i=1}^{n} Z_i\,. \tag{3}$$

Estimation method. The proposed estimate $\hat{G}(x)$ of $G(x)$ is of the form:

$$\hat{G}(x) = \frac{M}{nh} \sum_{i=1}^{n} Z_i K\left(\frac{x - X_i}{h}\right), \tag{4}$$

where $h > 0$ is the bandwidth. Sometimes we write $\hat{G}_N$ to stress the dependence of $\hat{G}$ on N.

In the nonparametric regression and density estimation it is usually assumed that $h = h(k)$ is a sequence of numbers such that

$$h(k) > 0\,, \qquad h(k) \xrightarrow{k\to\infty} 0\,, \qquad k \cdot h(k) \xrightarrow{k\to\infty} \infty\,. \tag{5}$$

In our case, one can consider two approaches of choosing the bandwidth sequence. The first one is to take $h = h(n)$, where n is the number of all observations. In the second one, $h = h(N)$ is a random sequence, which depends on the number of pairs (X_i, W_i), $i = 1, 2, \ldots, n$, falling below the curve $\{(x, G(x)) \colon x \in D\}$. We return to this topic later, imposing slightly more restrictive conditions on $h(k)$.

Background of the estimation method. One can justify (4) by invoking the following well known result (see Thm. 3.1 page 40 in [3]).

Theorem 1. (von Neumann) *Let X be a random vector with density g on $\mathbb{R}^d$ and let U be an independent uniform $[0, M]$, $M > 0$ random variable. Define $A = \{(x, u) \colon x \in \mathbb{R}^d,\ 0 \le u \le cg(x)\}$, where $c > 0$ is an arbitrary constant. If (X, U) is a random vector in $\mathbb{R}^{d+1}$ uniformly distributed on A, then X has density g on $\mathbb{R}^d$.*

In order to apply the above theorem to our problem it suffices to note that

$$g(x) = G(x) \Big/ \int_D G(x)\,dx \tag{6}$$

possesses all properties of a probability density function (see G1) and G2)) and to take the normalization constant as follows

$$C_G = \int_D G(x)\,dx\,. \tag{7}$$

Define

$$I_N = \{i \in \{1, 2, \ldots, n\} \colon W_i \le G(X_i)\}\,. \tag{8}$$

Let us note that pairs (X_i, W_i), $i \in I_N$, are uniformly distributed on A defined in Thm. 1. Thus, random variables X_i, $i \in I_N$, are i.i.d. random variables having densities g. This density equals to our unknown function up to constant factor C_G.

In other words, the problem of estimating G from the qualitative data is reduced to the problem of nonparametric estimation of p.d.f. g from observations X_i, $i \in I_N$, and to the problem of estimating constant c.

Density g can be estimated by the kernel type estimator as follows

$$\hat{g}_N(x) = \frac{1}{Nh} \sum_{i \in I_N} K\left(\frac{x - X_i}{h}\right), \tag{9}$$

where K and h are as defined above. One should notice, however, that $\hat{g}_N$ is not the standard estimator, since N is a random variable. Furthermore, N is not the classical stopping time, since its value is determined by all the sample (X_i, W_i), $i = 1, 2, \ldots, n$. Estimation of C_G, defined by (7), is relatively easy. Indeed,

$$\hat{C}_G = N \cdot M/n \tag{10}$$

can be taken, based on the classical Monte-Carlo arguments of estimating integrals.

Summarizing, estimator $\hat{G}$ can be treated as resulting from equation $G(x) = C_G \cdot g(x)$, in which C_G and g are replaced by (9) and (10), respectively.

One can also notice similarity between (4) and the nonparametric kernel type regression estimate when the binary regression is considered.

Remark 1. Indeed, setting $y_i = M \cdot Z_i$ and noticing that $E(y_i|X_i) = G(X_i)$ one can interpret (4) as the Nadaraya-Watson nonparametric estimator of G, where the variance of y_i, conditioned on X_i, is $\frac{G(X_i)}{M}\Big(1-\frac{G(X_i)}{M}\Big)$. Note, however, that number N of nonzero terms in (4) is random. This justifies a study of (4), which takes explicitly randomness of N into account.

Clearly, instead of the kernel type density estimate (9) one can consider also appropriately modified estimates based on orthogonal series expansions (see [8]).

3. Strong pointwise consistency and the convergence rate

Our aim in this section is to prove a.s. convergence of $\hat{G}(x)$ to $G(x)$ as $n \to \infty$. This is done in two steps. Namely, a.s. convergence of $\hat{g}(x)$ to $g(x)$ will be shown and then, $\hat{C}_G \to C_G$ as $n \to \infty$ is stated.

LIL for random subsequences. For a sequence of numbers τ_k, $k = 1, 2, \ldots$, we denote by $\mathcal{C}(\{\tau_k\})$ the set of all cluster points of this sequence.

We need the law of iterated logarithm (LIL) of the Anscombe type for random subsequences in the version, which can be found in [12] (see also [5], [1], [2]).

Theorem 2. (Torrang 1987) *Let $\{Y_k\}_{k=1}^{\infty}$ be i.i.d. random variables with $EY_k = 0$ and $EY_k^2 = \delta^2 < \infty$ and set $S_n = \sum_{k=1}^{n} Y_k$; $n = 1, 2, \ldots$ Let $\{b_k\}_{k=1}^{\infty}$ be a strictly increasing sequence of positive reals such that*

$$b_k \xrightarrow{k\to\infty} \infty, \quad b_{k+1}/b_k \xrightarrow{k\to\infty} B, \qquad 1 \le B < \infty. \tag{11}$$

Further, let $\{\nu_k\}_{k=1}^{\infty}$ be a strictly increasing sequence of positive, integer valued random variables with $\nu_1 \ge 3$ and such that $\nu_k/b_k \xrightarrow{k\to\infty} \theta$ a.s. $0 < \theta < \infty$. Then

$$\mathcal{C}\Big(\big\{S_{\nu_k}/\sqrt{\nu_k \log\log \nu_k}\,\big\}\Big) = \Big[-\delta\sqrt{2}, \delta\sqrt{2}\Big] \quad \text{a.s.} \tag{12}$$

Consistency. In order to apply the above result, let us note that setting $b_k = k$ and $\nu_k = N(k)$, we have $B = 1$ and

$$\frac{N(n)}{n} = \frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{n\to\infty} \frac{C}{M} < \infty \quad \text{a.s.} \tag{13}$$

by the law of large numbers.

Setting, for fixed $x \in D$,

$$Y_k = \frac{1}{h} K\left(\frac{x - X_k}{h}\right) - \frac{1}{h} EK\left(\frac{x - X_k}{h}\right), \qquad k \in I_N\,, \tag{14}$$

we have (see, e.g., [11], [9])

$$\mathrm{var}(Y_k) \le r(x, K)/h\,, \qquad k \in I_N\,, \tag{15}$$

where $r(x, K) = g(x) \cdot R(K)$, while $R(K)$ is a constant, which does not depend on h.

Thus, from (12), (15) it follows that

$$\mathcal{C}\Big(\{\gamma_N \sqrt{h}\tilde{g}_N(x)\}\Big) = \Big[-\sqrt{2r(x, K)}, \sqrt{2r(x, K)}\,\Big] \quad \text{a.s.} \tag{16}$$

where, by definition,

$$\gamma_N = \sqrt{N}/\log\log N\,, \tag{17}$$

while

$$\tilde{g}_N(x) = \hat{g}_N(x) - \frac{1}{h} EK\left(\frac{x - X}{h}\right), \tag{18}$$

where X is a random variable with p.d.f. $g(x)$. The form of $\tilde{g}_N(x)$ follows from the definition of $\hat{g}_N$, taking into account (14) and the fact that for $k \in I_N$ all X_k's have the same p.d.f. $g(x)$. Furthermore, (16) implies that at every point x in which $g(x) > 0$

$$|\tilde{g}_N(x)| = \mathrm{O}\Big(\log\log N \cdot (N \cdot h)^{-1/2}\Big) \quad \text{a.s.} \tag{19}$$

Now, one can consider two ways of choosing h as a function of n. The first one is to take directly $h = h(n) \xrightarrow{n\to\infty} 0$ in such a way that

$$\sqrt{N \cdot h(n)}/\log\log N \xrightarrow{n\to\infty} \infty \quad \text{a.s.} \tag{20}$$

The second approach is to take a random sequence $h = h(N) \xrightarrow{n\to\infty} 0$ a.s. for which

$$\sqrt{Nh(N)}/\log\log N \xrightarrow{n\to\infty} \infty \quad \text{a.s.} \tag{21}$$

Due to (16), both choices imply

$$\tilde{g}_N(x) \xrightarrow{n\to\infty} 0 \quad \text{a.s. for} \quad x \in \mathcal{B}_g\,, \tag{22}$$

where $\mathcal{B}_g$ is the set of all the points where $g(x) > 0$ and it is finite. Furthermore, (13) implies that both (20) and (21) ensure the same convergence rate. From the practical point of view, however, it seems desirable to use $h = h(N)$, since the difference between N and n can be large.

Summarizing, we have

$$\left|\hat{g}_N(x) - g(x)\right| \le \left|\tilde{g}_N(x)\right| + \left|\frac{1}{h} EK\left(\frac{x-X}{h}\right) - g(x)\right|, \tag{23}$$

and it remains to consider the second term in the r.h.s. of (23). To this end it suffices to invoke the result of Wheeden and Zygmund (see [13], Thm. 9.13 page 152), which implies its convergence to zero for almost all $x \in D$, provided that $h \to 0$ and K fulfils the conditions stated below (see also [4] for weaker conditions on K).

The above considerations imply the following theorem.

Theorem 3. *Let $h_N > 0$ be a sequence of numbers such that*

$$h_N \to 0, \quad \sqrt{N \cdot h_N}\big/ \log\log N \to \infty \quad \text{as} \quad N \to \infty. \tag{24}$$

Let kernel K be bounded and conditions (2) hold and $K(x) = O\big(|x|^{-d-\lambda}\big)$ as $|x| \to \infty$, where $d = \dim(x)$, while $\lambda > 0$. If $g \in L_1(\mathbb{R}^d)$, then for $n \to \infty$ $\hat{g}_N(x) \to g(x)$ a.s. at almost all (with respect to the Lebesgue measure) points $x \in \mathcal{B}_g$.

Convergence rate. It is well known (see,e.g., [11] page 39) that if g is twice continuously differentiable, then

$$\left|\frac{1}{h} EK\left(\frac{x-X}{h}\right) - g(x)\right| = h^2 \sigma_K^2 g''(x)/2 + o(h^2). \tag{25}$$

Now, (23), (19) and (25) allow to optimize the choice of $h(N)$, what leads to the following result.

Theorem 4. *If $g \in C^2(D)$, then the choice of $h(N)$ such that*

$$N^{-1/5}(\log\log N)^{2/5} c_1 \le h(N) \le N^{-1/5}(\log\log N)^{2/5} c_2, \quad 0 < c_1 < c_2 < \infty, \tag{26}$$

implies:

a) $\left|g(x) - \hat{g}_N(x)\right| = O(N^{-2/5})(\log\log N)^{4/5}$ *a.s.*

b) $\left|G(x) - \hat{G}_N(x)\right| = O(N^{-2/5})(\log\log N)^{4/5}$ *a.s.*

at almost all points $x \in \mathcal{B}_g \cap \{x\colon g''(x) \ne 0\}$.

Case b) follows directly from a), since $\hat{C}_G \to C_G$ at the rate of the order $1/\gamma_N$ a.s. (see, e.g., [10]).

A faster convergence rates can be obtained, if for more smooth functions of the higher order kernels are chosen [7].

A heuristic method of choosing bandwidth. Conditions (5) and (26) for the bandwidth sequence are asymptotic in nature. In recent years the problem of choosing h for a finite sample size was studied intensively. In particular, one can apply the cross-validation approach or the plug-in techniques to the estimation problem considered here. However, in our problem we have additional information, which can be used for tuning the bandwidth. Namely, we know how many W_i's are below the underlying function and what are the corresponding X_i's.

This motivates the following method. Choose $\hat{h}$, which maximizes the following function with respect to $h > 0$

$$\text{cardinality}\,\Big(\big\{i \in I_N\colon \hat{G}(X_i, h) \geq W_i\big\}\Big)\,, \tag{27}$$

where $\hat{G}(x, h)$ is the kernel estimator $\hat{G}(x)$, but its dependence on h is displayed.

We have no proof that the above $\hat{G}(x, \hat{h})$ is consistent and therefore (27) is called the heuristic method. Simulations performed by the author (excerpt of them is presented below) seem to indicate the heuristic procedure works quite well, at least for moderate sample size.

Kernel estimator with improved efficiency. Estimator $\hat{G}(x)$ uses only part of available qualitative data, namely, those which correspond to $Z_i = 1$, i.e., those pairs (X_i, W_i) below the graph of $G(\cdot)$. For this reason, it will be denoted in this paragraph as $\hat{G}_{\text{lower}}$.

According to the fundamental principle of statistics, an estimator should utilize all the available data. In order to use information contained in pairs (X_i, W_i) above the graph of $G(\cdot)$ (upper data), one can notice that they are uniformly distributed in set $A_{\text{upper}} = \big\{(x, w)\colon G(x) \leq w\big\}$. Furthermore, function $M - G(x)$ is nonnegative and one can again use Thm. 1 to conclude that X_i, $i \in I_N^{\text{upper}} \stackrel{def}{=} \{1, 2, \ldots n\} - I_N$ have p.d.f. proportional to $M - G(x)$. Thus, it is quite natural to define the following estimator

$$\hat{G}_{\text{upper}}(x) = M - \frac{M}{nh} \sum_{i \in I_N^{\text{upper}}} K\left(\frac{x - X_i}{h}\right), \tag{28}$$

where all the symbols are as defined above. Thus, it is natural to define the basic form of estimator for $G(x)$, which uses both the upper and lower data, as follows

$$G_n^*(x) = \big(\hat{G}_{lr}(x) + \hat{G}_{\text{upper}}(x)\big)/2\,. \tag{29}$$

It is clear that for $\hat{G}_{\text{upper}}(x)$ and for $G_n^*(x)$ Thm. 3 and Thm. 4 also hold.

Concerning the heuristic method of choosing h from the upper data, the analog of (27) is the following. Choose $\hat{h}_{\text{upper}}$ as the maximizer of the function

$$\text{cardinality}\ :\ \Big(\big\{i \in I_N^{\text{upper}}\colon \hat{G}_{\text{upper}}(X_i, h) \leq W_i\big\}\Big)\,, \tag{30}$$

with respect to $h > 0$, while $\hat{G}_{\text{upper}}(x, h)$ is defined by (28) with the dependence on h displayed.

When the above method of choosing h is applied to (29), we admit different values of h in $\hat{G}_{\text{lower}}$ and $\hat{G}_{\text{upper}}$.

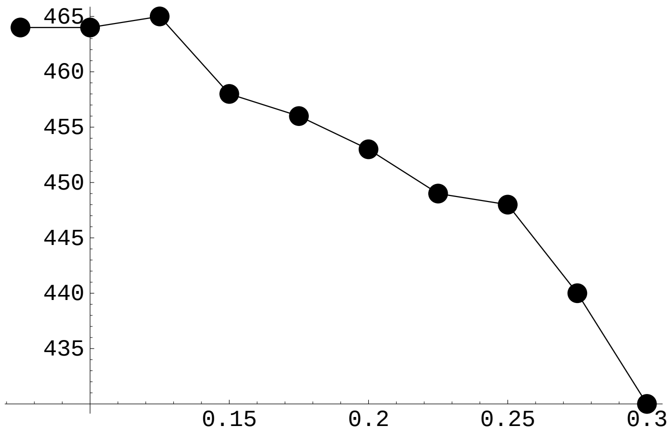

FIGURE 1. Number of observations below the curve $\hat{G}_{\text{lower}}(x)$ at points X_i, $i \in I_N$ as a function of bandwidth h. See Example 1 for details.

4. Kernel smoother estimate in action — simulation example

EXAMPLE 1. To get an insight into the estimator performance, $n = 1000$ qualitative observations from function $G(x) = \sin(2\pi x)/4 + 1/2$, $x \in D = [0, 1]$ were generated. Observations of independent variable were simulated as r.v. uniformly distributed in $[0, 1]$.

The biweight kernel $K(x) = (15/16) * (1 - x^2)^2$ if $|x| \leq 1$ and $K(x) = 0$ otherwise was used. In Fig. 1 the plot of the function (27) is shown for data below the underlying curve. Maximum corresponding to $\hat{h} = 0.125$ is clearly visible. This value was then used to calculate $\hat{G}_{\text{lower}}(x)$, which is shown in Fig. 2 (upper right plot).

Results of estimation are presented in Fig. 2, in which a complementary role of estimates from the both kinds of data is clearly visible. In particular, one can observe compensation of the boundary effects.

Acknowledgements. The author wishes to express his sincere thanks to an anonymous referee for helpful comments. In particular, the explanations given in Remark 1 were induced by these comments.

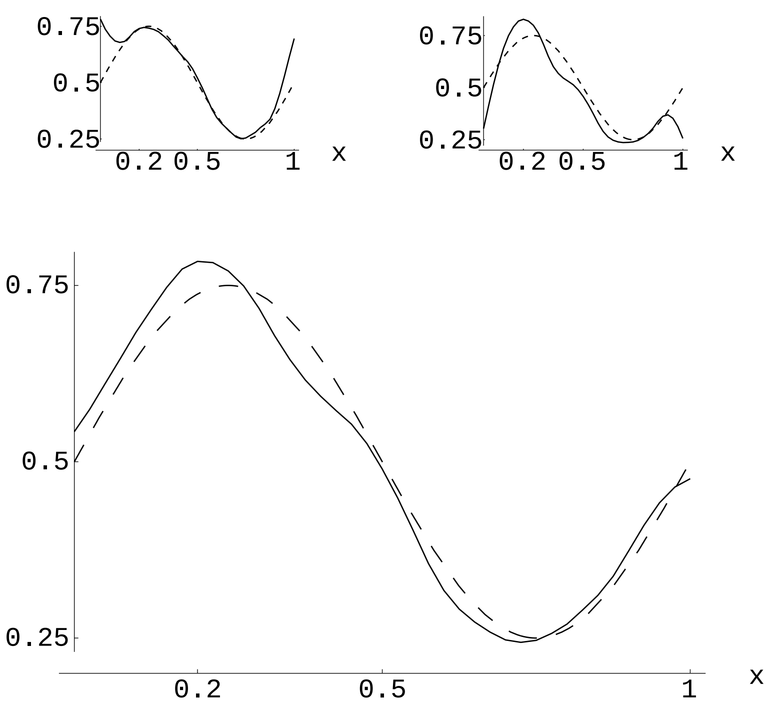

FIGURE 2. Estimates of $G(x) = \sin(2\pi x)/4 + 1/2$ (dashed line) obtained: a) from data above $G(\cdot)$ (upper left figure), b) from data below $G(\cdot)$ (upper right figure) c) as the pointwise mean of solid lines from a) and b) (bottom central figure). See Example 1 for details.

REFERENCES

[1] CHANG, I. S.—HSIUNG, C. A.: *Strassen's invariance principle for random subsequences*, Z. Wahrscheinlichkeitstheor Verw. Geb. **64** (1983), 401–409.

[2] CHOW, Y. S.—TEICHER, H.—WEI, C. Z.—YU, K. F.: *Iterated logarithm laws with random subsequences*, Z. Wahrscheinlichkeitstheor. Verw. Geb. **57** (1981), 235–251.

[3] DEVROYE, L.: *Non-Uniform Random Variate Generation*, Springer-Verlag, Berlin, 1986.

[4] GREBLICKI, W.—KRZYZAK, A.—PAWLAK, M.: *Distribution-free pointwise consistency of kernel regression estimate*, Ann. Statist. **12** (1984), 1570–1575.

[5] HUGGINS, R. M.: *Law of the iterated logarithm for time changed brownian motion and martingales with application to branching processes*, Ann. Probab. **13** (1985), 1148–1156.

[6] GYORFI, L.—DEVROYE, L.: *Nonparametric Density Estimation: The L1 View*, John Wiley, New York, 1985.

[7] MŰLLER, H. G.—GASSER, A. TH.: *Optimal convergence properties of kernel estimates of derivatives of a density function*, in: Proc. Workshop, Heidelberg, Smoothing Techniques for Curve Estimation (Th. Gasser and M. Rosenblatt, ed.), Springer-Verlag, Berlin, 1979, pp. 144–154.

[8] SCHWARTZ, S. C.: *Estimation of probability density by an orthogonal series*, Ann. Math. Stat. **38** (1967), 1261–1265.

[9] SCOTT, D. W.: *Multivariate Density Estimation*, John Wiley, New York, 1992.

[10] SERFLING, R. J.: *Approximation Theorems of Mathematical Statistics*, John Wiley, New York, 1980.

[11] SILVERMAN, B. W.: *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

[12] TORRANG, I.: *Law of the iterated logarithm — cluster points of deterministic and random subsequences*, Probab. Math. Statist. **8** (1987), 133–144.

[13] WHEEDEN, R. L.—ZYGMUND, A.: *Measure and Integral*, Marcel Dekker, New York, Basel, 1977.

Received March 17, 1998

Institute of Engineering Cybernetics
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27
PL–50 370 Wrocław
POLAND

E-mail: raf@ldhpux.immt.pwr.wroc.pl