Estimation of a Density Using Real and Artificial Data

Luc Devroye, Tina Felber, and Michael Kohler

***Abstract*—Let $X, X_1, X_2, \ldots$ be independent and identically distributed $\mathbb{R}^d$-valued random variables and let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function such that a density f of $Y = m(X)$ exists. Given a sample of the distribution of (X, Y) and additional independent observations of X, we are interested in estimating f. We apply a regression estimate to the sample of (X, Y) and use this estimate to generate additional artificial observations of Y. Using these artificial observations together with the real observations of Y, we construct a density estimate of f by using a convex combination of two kernel density estimates. It is shown that if the bandwidths satisfy the usual conditions and if in addition the supremum norm error of the regression estimate converges almost surely faster toward zero than the bandwidth of the kernel density estimate applied to the artificial data, then the convex combination of the two density estimates is L_1-consistent. The performance of the estimate for finite sample size is illustrated by simulated data, and the usefulness of the procedure is demonstrated by applying it to a density estimation problem in a simulation model.**

***Index Terms*—Consistency, density estimation, L_1-error, nonparametric regression.**

I. Introduction

LET $X, X_1, X_2, \ldots$ be independent and identically distributed $\mathbb{R}^d$-valued random variables and let $m : \mathbb{R}^d \to \mathbb{R}$ be an unknown measurable function such that a density f of $Y = m(X)$ exists. The distribution of X is unknown—its measure will be denoted by μ. The density f of $Y = m(X)$ must be estimated, and estimates will be compared on the basis of total variation distance.

This problem is substantially different from that of the estimation of the regression function m, as will be apparent from the discussion below. Note also that X does not have to have a density. In $\mathbb{R}^2$, consider $X = (U, U)$, where U is uniform on $[0, 1]$, and set $Y = m(X) = U$. Then, Y is uniformly distributed, yet X does not have a density. In $\mathbb{R}^1$, a more intricate example involving the Cantor set shows that X has in general not a density. Let the ternary expansion of $X \in (0, 1)$ be $0.b_1b_2\ldots$, where $b_1, b_2, \ldots$ are i.i.d. and uniformly drawn from $\{0, 2\}$. Then, X does not possess a density. Define the mapping m by the binary expansion of $m(X)$, given by $0.(b_1/2)(b_2/2)\ldots$. Since the bits in this expansion are i.i.d. and uniform on $\{0, 1\}$, $Y = m(X)$ is uniformly distributed on $[0, 1]$. However, if Y has a density, then X is nonatomic, i.e., continuous: its distribution function is continuous. We do not wish to assume anything about the underlying distribution of X.

We distinguish between three data models.

(i) In the *classical model*, we have one data size constant, n, and we observe the i.i.d. sequence

$$X_1, \ldots, X_n$$

(drawn from the distribution of X), and

$$Y_i = m(X_i), \quad 1 \le i \le n.$$

(ii) In the finite information model, we have two data size constants, n and N, and we observe the i.i.d. sequence

$$X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+N}$$

(drawn from the distribution of X), and

$$Y_i = m(X_i), \quad 1 \le i \le n.$$

This model is of interest in many applications, where the source of the X_i's is cheap and readily available, but the measurements Y_i are expensive or rare. Especially, internet data fit this setup.

(iii) In the full information model, which corresponds to $N = \infty$, we assume that μ, the distribution of X, is known, and that we have access to

$$X_1, \ldots, X_n$$

(drawn from the distribution of X), and

$$Y_i = m(X_i), \quad 1 \le i \le n.$$

This model is of theoretical interest, as it delineates how far one can push the boundary in the finite information model.

This paper takes a first look at the problem at hand, namely the estimation of the density f of Y for the finite information model (ii). It is of particular interest to learn how the presence of additional X-data (case (ii) with $N > 0$) can aid with the estimation. We present a new estimator and are broadly concerned with its consistency under the widest possible conditions, never assuming anything about the underlying distribution of X. We also point out avenues of future research on this set of problems.

Manuscript received January 05, 2012; revised August 28, 2012; accepted September 22, 2012. Date of publication November 27, 2012; date of current version February 12, 2013. This work was supported by the German Research Foundation (DFG) under the Collaborative Research Center 805.

L. Devroye is with the School of Computer Science, McGill University, Montreal, QC H3A 2K6, Canada (e-mail: lucdevroye@gmail.com).

T. Felber and M. Kohler are with the Department of Mathematics, Technical University Darmstadt, 64289 Darmstadt, Germany (e-mail: tfelber@mathematik.tu-darmstadt.de; kohler@mathematik.tu-darmstadt.de).

Communicated by N. Cesa-Bianchi, Associate Editor for Pattern Recognition, Statistical Learning, and Inference.

Digital Object Identifier 10.1109/TIT.2012.2230053

The safest way to approach this matter is by ignoring the X_i's altogether. In this case, f can be estimated by applying, e.g., a standard kernel density estimate (see [28] and [29]) defined by

$$f_n(y) = \frac{1}{nh_n} \cdot \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_n}\right)$$

with some kernel function $K : \mathbb{R} \to \mathbb{R}$ which is a density (e.g., the naive kernel $K(u) = 1/2 \cdot 1_{[-1,1]}$) and some bandwidth $h_n > 0$, which is a parameter of the estimate. For this estimate, it is known that

$$h_n \to 0 \quad (n \to \infty) \quad \text{and} \quad n \cdot h_n \to \infty \quad (n \to \infty)$$

imply that the estimate is L_1-consistent for all densities (cf., [7] and [24]):

$$\int |f_n(x) - f(x)|\, dx \to 0$$

almost surely as $n \to \infty$. By Scheffé's Lemma (see, e.g., [10]) this implies that the estimated distribution converges to the true distribution in total variation distance and hence the above L_1-consistent density estimate allows simultaneous estimation of all probabilities. For general results in density estimation, we refer to the books of [8], [10], and [12].

Improvements in the performance can be achieved in model (i) if additional information about m is available. This will not be our focus. In model (ii), without assuming anything about m or X, there is indeed help in the form of additional X_i's. We achieve this by estimating m by m_n based on the data

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

which allows us to generate artificial (approximate) observations of $Y = m(X)$ via $\widehat{Y}_i = m_n(X_{n+i})$ $(i = 1, \ldots, N)$. In a second step, we apply separately kernel density estimates to the datasets $Y_1, \ldots, Y_n$ and $\widehat{Y}_1, \ldots, \widehat{Y}_N$ and use a convex combination of the resulting two density estimates as an estimate of f. The L_1-error of this estimate depends in particular on the interplay between the error of the estimate m_n and the bandwidth of the second kernel density estimate. In our main result, we give sufficient conditions for the L_1-consistency of our estimate, and under suitable smoothness conditions on m, we are able to show that our density estimate is indeed L_1-consistent. Furthermore, we indicate that under which conditions our estimate should achieve a better rate of convergence than the simple estimate mentioned above, and these observations are confirmed in our simulation part. As an application, we consider a density estimation problem in a simulation model.

Estimation of m from the data $\mathcal{D}_n$ can be done via regression estimation, a field studied already over many years. The most popular estimates include kernel regression estimate (cf., e.g., [13], [14], [26], [27], [30], [31], or [33]), partitioning regression estimate (cf., e.g., [2] or [18]) nearest neighbor regression estimate (cf., e.g., [6] or [11]), least squares estimates (cf., e.g., [21] or [23]), or smoothing spline estimates (cf., e.g., [22] or [32]). For a detailed introduction to nonparametric regression, we refer to [19].

Our analysis depends critically on the connection between the error of the regression estimate and the error of the density estimate. A similar phenomenon occurs in density estimation of the density of residuals of a regression model (cf., e.g., [1], [5], [9], [16], or [17]), and in our proof, we apply techniques related to the ones in [9]. Our dataset can be considered as one dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$ with labels and one unlabeled dataset $X_{n+1}, X_{n+2}, \ldots$ and our procedure can be considered as semisupervised learning for density estimation, which is usually studied in the context of pattern recognition (cf., e.g., [3], [4], and the wide-ranging literature cited therein.)

The classical model forces one to work with $Y_1, \ldots, Y_n$, i.e., the input data are 1-D, and the nature of the mapping m is irrelevant, as is the underlying distribution of X. It is surprising that with additional information on the distribution of X, one can improve the estimate of the density of $m(X)$. Situations in which this is possible are plentiful. Consider the X_i's as internet data on users or web sites. Let $Y_i = m(X_i)$ be the outcome of an action of a user, such as a purchase or investment. While the number of X_i's is virtually unlimited, the number of users i for which $Y_i = m(X_i)$ is known is limited, because not everyone performs the action. The finite information model (ii) introduced above corresponds to this. In a second example, X_i's represent medical data on patients. For a smallish subset of the patients, the outcome of a treatment is known: $Y_i = m(X_i)$. What we point out is that all data, even those of patients who have not undergone the treatment, can help in the estimation of the density of $m(X)$.

The outline of this paper is as follows: We start in Section II with a discrete analog of the problem, where we illustrate the potential usefulness of the artificial data. Then, we present in Section III a general consistency result for a newly proposed estimate in the general finite information model. We indicate in Section IV how in a special situation, the used regression estimate might be improved drastically. We investigate in Section V the performance of the estimate from Section III for finite sample size by simulated data and illustrate the usefulness of the procedure by applying it to a density estimation problem in a simulation model. We give an outlook in Section VI. The proof of our main result is given in the Appendix.

II. Discrete Analog

In the discrete version of this problem, X is a random variable on the positive integers with

$$p_i = \mathbf{P}\{X = i\}.$$

There is an unknown function m on the positive integers. The objective is to estimate the distribution of the atomic random variable $m(X)$. Since m itself takes only countably many values, the *canonical version* of the problem is such that m itself takes values in the positive integers. We define

$$q_j = \mathbf{P}\{m(X) = j\}$$

and are interested in estimating q_j from data such that the total variation error is small.

We would like to investigate how the additional data $X_{n+1}, \ldots, X_{n+N}$ can help in the estimation. Since models

(i) and (iii) described in Section I correspond to $N = 0$ and $N = \infty$, respectively, it should be clear that we should first try to compare (i) and (iii). In case (i), it is difficult to improve on the empirical estimate

$$q_{n,j} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[m(X_i)=j]}, \quad j \geq 1.$$

Note that $nq_{n,j}$ is binomial (n, q_j), and hence, $q_{n,j}$ is an unbiased estimate of q_j. The expected total variation error is easily bounded:

$$\begin{aligned}\sum_j \mathbf{E}\{|q_j - q_{n,j}|\} &= 2\sum_j \mathbf{E}\{(q_j - q_{n,j})_+\} \\ &\leq 2\sum_j \min\left(q_j, \sqrt{\mathrm{Var}\{q_{n,j}\}}\right) \\ &= 2\sum_j \min\left(q_j, \sqrt{q_j(1-q_j)/n}\right).\end{aligned}$$

If $\sum_j \sqrt{q_j} < \infty$, then the upper bound is $O(1/\sqrt{n})$. In all but the trivial case that $q_j = 1$ for some j, the expected total variation error

$$\sum_j \mathbf{E}\{|q_j - q_{n,j}|\}$$

tends to 0 at the rate $1/\sqrt{n}$ or slower, because if $q_k \in (0,1)$ for some $k \in \mathbb{N}$, then

$$\frac{\mathbf{E}\{|q_k - q_{n,k}|\}}{1/\sqrt{n}} \to 0 \quad (n \to \infty)$$

implies that

$$\frac{\sqrt{n}\cdot(q_{n,k} - q_k)}{\sqrt{q_k \cdot (1-q_k)}} \to 0 \quad \text{in } L_1$$

which is a contradiction to the central limit theorem.

In case (iii), the situation is remarkably different. Let $A = \{X_1, \ldots, X_n\}$ with duplicates removed. If $i \in A$, $m(i)$ is known. If $i \notin A$, m is unknown and cannot possibly be guessed. Since

$$q_j = \mathbf{P}\{m(X) = j\} = \sum_i p_i \mathbb{I}_{[m(i)=j]}$$

we set

$$q_{n,j} = \sum_{i \in A} p_i \mathbb{I}_{[m(i)=j]}.$$

Clearly, $0 \leq q_{n,j} \leq q_j$, and we do not have the unbiasedness we enjoyed in case (i). However, the expected total variation error has a simple and universal expression that is the same (!!!) for all choices of m. First note that the total variation error is

$$\begin{aligned}\sum_j (q_j - q_{n,j}) &= \sum_j \left(\sum_i p_i \mathbb{I}_{[m(i)=j]} - \sum_{i\in A} p_i \mathbb{I}_{[m(i)=j]}\right) \\ &= \sum_i p_i - \sum_{i \in A} p_i \\ &= \sum_{i \notin A} p_i.\end{aligned}$$

Thus, the expected total variation error is

$$\sum_i p_i \mathbf{P}\{i \notin A\} = \sum_i p_i (1-p_i)^n.$$

This error tends to zero with n in all cases, and the rate of decrease depends upon the tail of $\{p_i\}$. However, it is much better than for (i). To wit, consider X with compact support. Then, the expected total variation error tends to 0 at an exponential rate in n.

It is of interest to see how the finite information model interpolates between these two behaviors.

III. General Finite Information Model

In the sequel, we consider the general finite information model, where we introduce a new density estimate and study its consistency.

In the definition of our generic estimate, we first apply a suitably defined regression estimate m_n to the data

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

in order to estimate $m : \mathbb{R}^d \to \mathbb{R}$. Then, we use the estimate m_n of m in order to define an artificial sample of Y. To do this, we choose the size N of this sample and define artificial data via

$$\widehat{Y}_1 = m_n(X_{n+1}), \ldots, \widehat{Y}_N = m_n(X_{n+N}).$$

Next we apply standard kernel density estimates separately to the data $Y_1, \ldots, Y_n$ and $\widehat{Y}_1, \ldots, \widehat{Y}_N$. Let K be the so-called naive kernel defined by

$$K(u) = \frac{1}{2} \cdot \mathbb{I}_{[-1,1]}(u) \quad (u \in \mathbb{R})$$

let $h_n > 0$ and $\widehat{h}_N > 0$ and define

$$f_n(y) = \frac{1}{n \cdot h_n} \cdot \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_n}\right)$$

and

$$\widehat{f}_N(y) = \frac{1}{N \cdot \widehat{h}_N} \cdot \sum_{i=1}^{N} K\left(\frac{y - \widehat{Y}_i}{\widehat{h}_N}\right).$$

Finally, we use a convex combination of these two estimates as estimate of f, i.e., we choose a weight

$$w_n = w_n(\mathcal{D}_n, X_{n+1}, \ldots, X_{n+N}) \in [0,1]$$

and estimate f by

$$g_n = w_n \cdot f_n + (1 - w_n) \cdot \widehat{f}_N.$$

The precise choice of weight function is left open, as is the choice of regression function estimate m_n. The first business at hand is to determine the consistency of the generic estimate. Since we do not impose restrictions on w_n, the choices $w_n = 0$ and $w_n = 1$ imply that both f_n and $\widehat{f}_N$ must be consistent. The latter can only happen if $N \to \infty$. Also, the performance of m_n is critical. For example, if one lets m_n be the nearest neighbor regression function estimate ($m_n(x) = m(X_i)$ if X_i is the nearest neighbor of x), then is the atomic nature of such

m_n a problem for the consistency of $\widehat{f}_N$? The consistency theorem we present in the next section takes a higher view and gives a natural technical condition that links m_n to $\widehat{h}_N$.

Our main result is the following theorem.

Theorem 1: Let $X, X_1, X_2, \ldots$ be independent and identically distributed $\mathbb{R}^d$-valued random variables and let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function such that a density f of $Y = m(X)$ exists. Set $Y_i = m(X_i)$ $(i \in \mathbb{N})$. Let K be the naive kernel, and for $n \in \mathbb{N}$ let $N \in \mathbb{N}$, $h_n > 0$ and $\widehat{h}_N > 0$. Given $(X_1, Y_1), \ldots, (X_n, Y_n), X_{n+1}, X_{n+2}, \ldots$, and a regression estimate m_n, let g_n be the estimate of f as defined in Section II. Assume that $N = N(n) \to \infty$ as $n \to \infty$, and that

$$\begin{aligned} \max(h_n, \widehat{h}_n) &\to 0 \quad (n \to \infty) \\ n \cdot \min(h_n, \widehat{h}_n) &\to \infty \quad (n \to \infty). \end{aligned} \tag{1}$$

Assume furthermore the following on m_n: for every $\epsilon > 0$, $n \in \mathbb{N}$, there exists a (random) set $A_{n,\epsilon} = A_{n,\epsilon}(\mathcal{D}_n) \subseteq \mathbb{R}^d$ such that

$$\lim_{n\to\infty} \mathbf{P}\{\mu\{A_{n,\epsilon}^c\} > \epsilon\} = 0 \tag{2}$$

and

$$\frac{\|m_n - m\|_{\infty, A_{n,\epsilon}}}{\widehat{h}_N} = \frac{\sup_{x \in A_{n,\epsilon}} |m_n(x) - m(x)|}{\widehat{h}_N} = o(1) \tag{3}$$

in probability. Then, regardless of how w_n is chosen

$$\int_{\mathbb{R}} |g_n(y) - f(y)|\, dy \to 0$$

in probability. (In particular, $\int_{\mathbb{R}} \left|\widehat{f}_N(y) - f(y)\right| dy = o(1)$ in probability as well.) If, in addition

$$\limsup_{n\to\infty} \mu\{A_{n,\epsilon}^c\} \le \epsilon \tag{4}$$

almost surely and

$$\frac{\|m_n - m\|_{\infty, A_{n,\epsilon}}}{\widehat{h}_N} = \frac{\sup_{x \in A_{n,\epsilon}} |m_n(x) - m(x)|}{\widehat{h}_N} = o(1) \tag{5}$$

almost surely then

$$\int_{\mathbb{R}} |g_n(y) - f(y)|\, dy \to 0$$

almost surely.

Proof: The proof is given in the Appendix.

Conditions (3) and (5) can be derived from rate of convergence results for nonparametric regression estimates. A less cumbersome, but weaker, consistency result is the following.

Corollary 1: Let $X, X_1, X_2, \ldots$ be independent and identically distributed $\mathbb{R}^d$-valued random variables and let $m : \mathbb{R}^d \to \mathbb{R}$ be a measurable function such that a density f of $Y = m(X)$ exists. Let K be the naive kernel, and for $n \in \mathbb{N}$ let $N \in \mathbb{N}$, $h_n > 0$ and $\widehat{h}_N > 0$. Set $Y_i = m(X_i)$ $(i \in \mathbb{N})$. Given $(X_1, Y_1), \ldots, (X_n, Y_n), X_{n+1}, X_{n+2}, \ldots$, and a regression estimate m_n, let g_n be the estimate of f as defined in Section II. Assume that (1) holds.

(i) If, in addition

$$\begin{aligned} \mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) &= \mathbf{E}\left(|m_n(X) - m(X)|^2\right) \\ &= o\left(\widehat{h}_N^2\right) \end{aligned} \tag{6}$$

then regardless of the choice of w_n

$$\int_{\mathbb{R}} |g_n(y) - f(y)|\, dy \to 0$$

in probability.

(ii) If, in addition

$$\frac{\int |m_n(x) - m(x)|^2 \mu(dx)}{\widehat{h}_N^2} \to 0 \tag{7}$$

almost surely, then regardless of the choice of w_n

$$\int_{\mathbb{R}} |g_n(y) - f(y)|\, dy \to 0$$

almost surely.

Proof: In order to prove (i), choose $a_n \in \mathbb{R}_+$ such that

$$\frac{\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx)}{a_n^2} = o(1) \tag{8}$$

and

$$\frac{a_n}{\widehat{h}_N} = o(1). \tag{9}$$

In order to show (ii), choose $a_n = a_n(\mathcal{D}_n) \in \mathbb{R}_+$ such that

$$\frac{\int |m_n(x) - m(x)|^2 \mu(dx)}{a_n^2} = o(1) \tag{10}$$

almost surely and

$$\frac{a_n}{\widehat{h}_N} = o(1) \tag{11}$$

almost surely. In both cases, set

$$A_{n,\epsilon} = \left\{x \in \mathbb{R}^d : |m_n(x) - m(x)| \le a_n\right\}.$$

Then

$$\frac{\|m_n - m\|_{\infty, A_{n,\epsilon}}}{\widehat{h}_N} \le \frac{a_n}{\widehat{h}_N} \to 0$$

almost surely by (9) or (11), respectively. Furthermore, by Markov's inequality, we have

$$\mu(A_{n,\epsilon}^c) \le \frac{\int |m_n(x) - m(x)|^2 \mu(dx)}{a_n^2}$$

so (2) and (4) are implied by (8) and (10), respectively. Application of Theorem 1 yields the assertion. □

Remark 1: It is an open problem how to choose the bandwidths of the estimates above in a data-dependent way such that the consistency theorem still holds and the estimate behaves well in practice. In view of the minimization of the L_1 error, it would be in particular interesting to try to use the combinatorial method of [12] in this context.

Remark 2: In this paper, Y is defined by applying a deterministic function m to X. In case that the relation is given by a regression model

$$Y = m(X) + \epsilon$$

where ϵ is the random error satisfying $\mathbf{E}(\epsilon \mid X) = 0$ a.s. and the aim is to estimate a density of $m(X)$, it follows from the proofs that Theorem 1 and Corollary 1 remain valid if we set $w_n = 0$.

Remark 3: The techniques introduced in the proofs of Theorem 1 and Corollary 1 can be applied in the context of estimation of the density of the residuals in a nonparametric regression model. Here, we assume that $Y - m(X)$ has a density f, and we try to estimate f using the data $(X_1, Y_1), \ldots, (X_{2n}, Y_{2n})$. To do this, we compute first a regression estimate $m_n(x) = m_n(x, \mathcal{D}_n)$ and use then

$$\widehat{f}_n(x) = \frac{1}{n \cdot h_n} \cdot \sum_{i=1}^{n} K\left(\frac{x - (Y_{n+i} - m_n(X_{n+i}))}{h_n}\right)$$

as estimate of f. Assume $h_n \to 0$ $(n \to \infty)$, $n \cdot h_n \to \infty$ $(n \to \infty)$ and

$$\frac{\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx)}{\widehat{h}_n^2} \to 0 \quad (n \to \infty). \tag{12}$$

We argue as in the proof of Theorem 1, where we use

$$\mathbf{E}\{\widehat{f}_n(x) | \mathcal{D}_n\} = \int \frac{1}{h_n} \cdot K\left(\frac{x - (y - m_n(z))}{h_n}\right) \nu(dz, dy)$$

(where ν is the joint probability measure of (X, Y)) and replace $m_n(X_{n+i})$ by $Y_{n+i} - m_n(X_{n+i})$). Then, we obtain weak consistency:

$$\int_{\mathbb{R}} |\widehat{f}_n(y) - f(y)| \, dy \to 0$$

in probability. It follows from [30] that there exist weakly universally consistent regression estimates, so for each distribution of (X, Y), we can find a sequence of bandwidths h_n (depending on this distribution) such that (12) holds. Observe that this result does not require that X and $Y - m(X)$ are independent.

It is an open problem, whether the same result also holds for a fixed sequence of bandwidths $\{h_n\}_n$ and a fixed sequence of regression estimates.

IV. Linear Interpoland

In this section, we look at a very specific choice of m_n, the linear interpoland, when $d = 1$, and in addition, $m : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable and the distribution function F of X satisfies for some $a < b$: $F(a) = 0$, $F(b) = 1$, and on (a, b), F is twice continuously differentiable with first derivative bounded away from zero and with second derivative uniformly bounded on (a, b). (Equivalently, μ has a continuously differentiable density on $[a, b]$ that is bounded away from 0 on that interval.) These assumptions imply that F^{-1} exists, has on $(0, 1)$ a uniformly bounded second derivative, and that $m \circ F^{-1}$ is twice continuously differentiable with uniformly bounded second derivative on $(0, 1)$.

Two observations are crucial: first of all, the total variation error is invariant under componentwise monotone transformations of X, if we wish to estimate the density of X. Second, if μ is known, as in model (iii), then the probability integral transform can be used and X can be replaced by $F(X)$, where F is the (known) distribution function of X. In particular, for $d = 1$, we may in all cases assume, without loss of generality, that X is uniformly distributed on $[0, 1]$. This is called the *canonical version* of the problem.

The canonical version above corresponds to the case $N = \infty$, where we assume that F (i.e., μ) is explicitly known. In the sequel, we take a slightly more realistic stance and assume instead that $N > n^4$. Let F_N be the empirical distribution function for $X_{n+1}, \ldots, X_{n+N}$, and let $Q_n m$ be the linear interpoland of $(F_N(X_i), m(X_i))$ $(i = 1, \ldots, n)$ (where the first and the last linear part are extended to $-\infty$ and ∞, respectively). Observe that $(Q_n m)(y)$ is well defined for $y \notin \{F_N(X_1), \ldots, F_N(X_n)\}$ even if $F_N(X_i) = F_N(X_j)$ for some $i \neq j$. For $y \in \{F_N(X_1), \ldots, F_N(X_n)\}$, we require that $(Q_n m)(y)$ is equal to some $m(X_k)$ such that $y = F_N(X_k)$. Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics of $X_1, \ldots, X_n$, and set

$$m_n(x) = \begin{cases} (Q_n m)(F_N(x)), & \text{if } X_{(1)} \le x \le X_{(n)} \\ m(X_{(1)}), & \text{if } x < X_{(1)} \\ m(X_{(n)}), & \text{if } x > X_{(n)}. \end{cases}$$

Then

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \, \mu(dx) = O\left(\frac{1}{n^3}\right). \tag{13}$$

Consequently, if we demand that $\lim_{n\to\infty} n^{3/2} \cdot \widehat{h}_N = \infty$ (so that assumption (6) is satisfied), the estimate $\widehat{f}_N$ is weakly L_1-consistent.

Proof of (13): Let $\bar{Q}_n m$ be the linear interpoland of $(F(X_i), m(X_i))$ $(i = 1, \ldots, n)$. For $x \in (a, b)$, set

$$\eta_n(x) = \begin{cases} F(X_{(1)}), & \text{if } F(x) \le F(X_{(1)}) \\ F(X_{(k)}) - F(X_{(k-1)}), & \text{if } F(X_{(k-1)}) < F(x) \\ & \text{and } F(x) \le F(X_{(k)}) \\ 1 - F(X_{(n)}), & \text{if } F(x) \ge F(X_{(n)}). \end{cases}$$

In case $x \in [X_{(1)}, X_{(n)}]$, $\eta(x)$ is an upper bound on the distances of $F(x)$ to the two closests values among $\{F(X_i) : i = 1, \ldots, n\}$. Let $x \in (a, b)$ be arbitrary. First, we assume $x \in [X_{(1)}, X_{(n)}]$. In this case, we have

$$\begin{aligned} &|m_n(x) - m(x)| \\ &\le \left|(Q_n m)(F_N(x)) - \left(\bar{Q}_n m\right)(F_N(x))\right| \\ &\quad + \left|\left(\bar{Q}_n m\right)(F_N(x)) - \left(\bar{Q}_n m\right)(F(x))\right| \\ &\quad + \left|\left(\bar{Q}_n m\right)(F(x)) - \left(m \circ F^{-1}\right)(F(x))\right| \\ &=: T_{1,n} + T_{2,n} + T_{3,n}. \end{aligned}$$

Since

$$\begin{aligned} &\frac{m(X_{(k)}) - m(X_{(k-1)})}{F(X_{(k)}) - F(X_{(k-1)})} \\ &= \frac{m \circ F^{-1}(F(X_{(k)})) - m \circ F^{-1}(F(X_{(k-1)}))}{F(X_{(k)}) - F(X_{(k-1)})} \\ &= (m \circ F^{-1})'(\xi) \end{aligned}$$

for some $\xi \in (F(X_{(k-1)}), F(X_{(k)}))$, $\bar{Q}_n m$ is Lipschitz continuous with Lipschitz constant bounded by $\sup_{z\in(0,1)} \left|\left(m \circ F^{-1}\right)'(z)\right|$, from which we conclude

$$T_{2,n} \le \sup_{z\in(0,1)} \left|\left(m \circ F^{-1}\right)'(z)\right| \cdot \sup_{u\in\mathbb{R}} |F_N(u) - F(u)|.$$

By construction of $Q_n m$ and of $\bar{Q}_n m$, for any $y \in [F_N(X_{(1)}), F_N(X_{(n)})]$, there exists $\bar{y} \in \mathbb{R}$ satisfying

$$(Q_n m)(y) = (\bar{Q}_n m)(\bar{y}) \quad \text{and} \quad |y - \bar{y}| \leq \sup_{u \in \mathbb{R}} |F_N(u) - F(u)|.$$

(In case $y = F_N(X_{(k)}) + \alpha \cdot (F_N(X_{(k+1)}) - F_N(X_{(k)}))$ for some $k \in \{1, \ldots, n-1\}$ and some $\alpha \in (0,1)$, we set $\bar{y} = F(X_{(k)}) + \alpha \cdot (F(X_{(k+1)}) - F(X_{(k)}))$.) Consequently, setting $y = F_N(x)$, we get by using again the Lipschitz property of $\bar{Q}_n m$:

$$\begin{aligned} T_{1,n} &= |(\bar{Q}_n m)(\bar{y}) - (\bar{Q}_n m)(y)| \\ &\leq \sup_{z \in (0,1)} \left| \left(m \circ F^{-1}\right)'(z) \right| \cdot \sup_{u \in \mathbb{R}} |F_N(u) - F(u)|. \end{aligned}$$

Finally, to bound $T_{3,n}$, we observe that $\left(m \circ F^{-1}\right)(u)$ is equal to the value of the first term of the Taylor series of $m \circ F^{-1}$ around u and evaluated at u, and that this Taylor series polynomial p satisfies

$$\bar{Q}_n(p \circ F) = p.$$

Hence, setting $u = F(x)$, it suffices to bound

$$\left| \left(\bar{Q}_n m\right)(u) - \left(\bar{Q}_n (p \circ F)\right)(u) \right|.$$

Since both expressions are linear interpolands of functions at points with x-values $F(X_i)$ $(i = 1, \ldots, n)$ and y-values $(m \circ F^{-1})(F(X_i))$ and $p(F(X_i))$, respectively, their maximum distance for $u \in \left[F(X_{(1)}), F(X_{(n)})\right]$ is bounded by the distance between $m \circ F^{-1}$ and its Taylor series approximant evaluated at the two x-points closest to u, which have distance less than $\eta(x)$ from X.

Summarizing the above results, we get for $x \in [X_{(1)}, X_{(n)}]$

$$\begin{aligned} &|m_n(x) - m(x)| \\ &\leq 2 \cdot \sup_{z \in (0,1)} \left| \left(m \circ F^{-1}\right)'(z) \right| \cdot \sup_{u \in \mathbb{R}} |F_N(u) - F(u)| \\ &\quad + \frac{1}{2} \sup_{z \in (0,1)} \left| \left(m \circ F^{-1}\right)''(z) \right| \cdot \eta_n(x)^2. \end{aligned}$$

For $x < X_{(1)}$, we get (by using the Lipschitz property of $m \circ F^{(-1)}$)

$$\begin{aligned} &|m_n(x) - m(x)| \\ &= |(m \circ F^{-1})(F(X_{(1)})) - (m \circ F^{-1})(F(x))| \\ &\leq \sup_{z \in (0,1)} \left| \left(m \circ F^{-1}\right)'(z) \right| \cdot F(X_{(1)}) \end{aligned}$$

and for $x > X_{(n)}$, we have

$$|m_n(x) - m(x)| \leq \sup_{z \in (0,1)} \left| \left(m \circ F^{-1}\right)'(z) \right| \cdot (1 - F(X_{(n)})).$$

This implies

$$\begin{aligned} &\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \\ &\leq \text{const} \cdot \mathbf{E}\left(\sup_{u \in \mathbb{R}} |F_N(u) - F(u)|^2 \right) + \text{const} \cdot \mathbf{E}\left(\eta_n(X)^4\right) \\ &\quad + \text{const} \cdot \mathbf{E}\left(F(X_{(1)})^3\right) + \text{const} \cdot \mathbf{E}\left((1 - F(X_{(n)}))^3\right). \end{aligned}$$

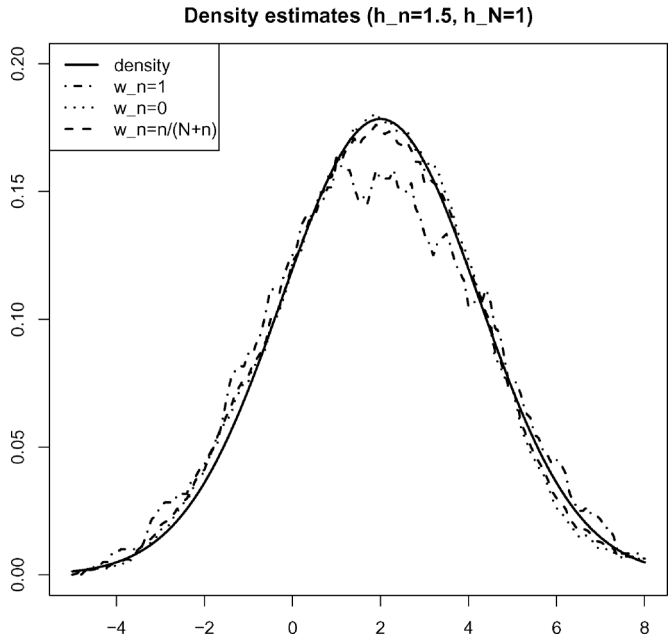

Fig. 1. Density estimates in the first model.

Using the Dvoretzky–Kiefer–Wolfowitz inequality (see [15]), we can conclude

$$\begin{aligned} &\mathbf{E}\left(\sup_{u \in \mathbb{R}} |F_N(u) - F(u)|^2 \right) \\ &\leq \frac{1}{n^3} + \int_{1/n^3}^{\infty} \mathbf{P}\left\{ \sup_{u \in \mathbb{R}} |F_N(u) - F(u)| > \sqrt{t} \right\} dt \\ &\leq \frac{1}{n^3} + \int_{1/n^3}^{\infty} 2 \cdot \exp(-2Nt) dt \\ &= O\left(\frac{1}{n^3}\right). \end{aligned}$$

Furthermore, using the fact that $F(X_1)$ is uniformly distributed on $[0,1]$, we get

$$\begin{aligned} &\mathbf{E}\left(\eta_n(X)^4\right) \\ &= \int_0^{\infty} 4 \cdot t^3 \cdot \mathbf{P}\left[\eta_n(X) > t\right] dt \\ &\leq 4 \cdot \int_0^1 t^3 \cdot \mathbf{P}\left[\forall i \in \{1, \ldots, n\} : \right. \\ &\qquad \left. F(X_i) \notin \left[F(X) - \frac{t}{2},\ F(X)\right],\ F(X) > \frac{t}{2} \right] dt \\ &\quad + 4 \cdot \int_0^1 t^3 \cdot \mathbf{P}\left[\forall i \in \{1, \ldots, n\} : \right. \\ &\qquad \left. F(X_i) \notin \left[F(X),\ F(X) + \frac{t}{2}\right],\ F(X) < 1 - \frac{t}{2} \right] dt \\ &= 4 \cdot \int_0^1 t^3 \cdot 2 \cdot \left(1 - \frac{t}{2}\right)^n dt \\ &\leq 8 \cdot \int_0^1 t^3 \cdot \exp\left(-\frac{n\,t}{2}\right) dt \\ &\leq \frac{384}{n^3} \int_0^1 \exp\left(-\frac{n\,t}{2}\right) dt = O\left(\frac{1}{n^4}\right) \end{aligned}$$

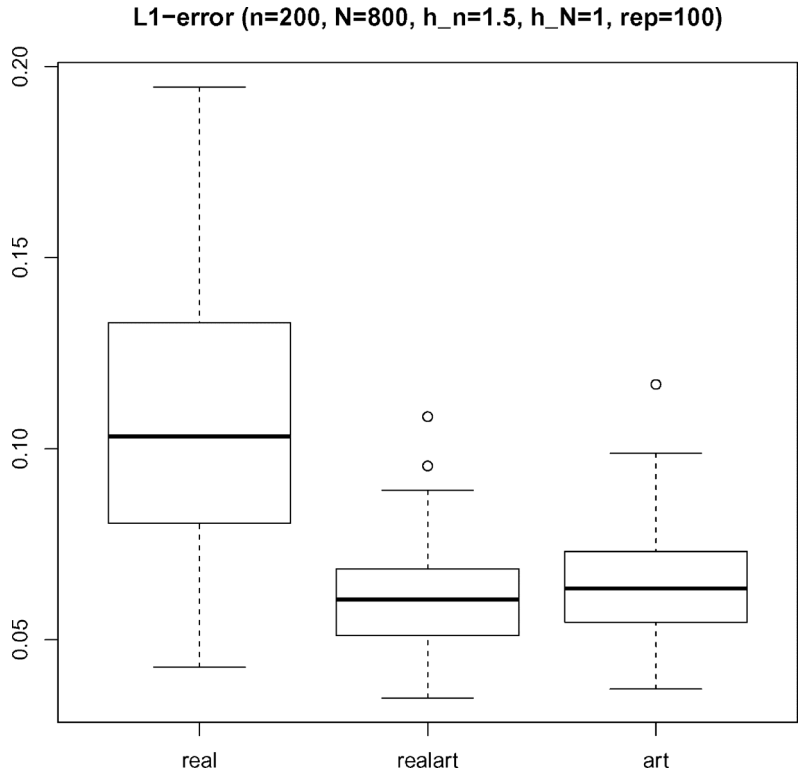

Fig. 2. Boxplots in the first model.

where the last inequality follows by partial integration. Finally, we observe

$$\begin{aligned}
&\mathbf{E}\left((1-F(X_{(n)}))^3\right)\\
&=\mathbf{E}\left(F(X_{(1)})^3\right)\\
&=\int_0^1 3\cdot t^2\cdot \mathbf{P}\left[\forall i\in\{1,\ldots,n\}\,:\,F(X_i)>t\right]\,dt\\
&\le\int_0^1 3\cdot t^2\cdot e^{-n\cdot t}\,dt\\
&<\frac{6}{n^3}=O\left(\frac{1}{n^3}\right).
\end{aligned}$$

□

Remark 6: For $d>1$, we may assume that X has a copula distribution (i.e., a distribution with uniform marginals). In this case, it is not necessary to assume that X has a density g.

V. Application to Simulated Data

In this section, we illustrate the finite sample size performance of our estimates by applying them to simulated data.

In our first example, we set $X=(X^{(1)},X^{(2)})$ for independent standard normally distributed random variables $X^{(1)}$ and $X^{(2)}$ and choose $Y=m(X)$ for

$$m(x_1,x_2)=2\cdot x_1+x_2+2.$$

In this case, Y is normally distributed with expectation 2 and variance $2^2+1^2=5$. We estimate the density of Y by the estimate introduced in Section II, where we use a fully data-driven smoothing spline estimate to estimate the linear function m. For this purpose, we use the routine $Tps()$ from the library $fields$ in the statistics package R. For the weights, we use three different values: $w_n=1$ (in which case, we use only the real data), $w_n=0$ (in which case the estimate is based only on the artificial data) and $w_n=n/(n+N)$ (in which case, we use real and artificial data and all data points have the same weight). We set $n=200$ and $N=800$ and choose the bandwidths by minimizing the L_1-errors of the estimate via comparing the estimated density with the true density (so we assume that we

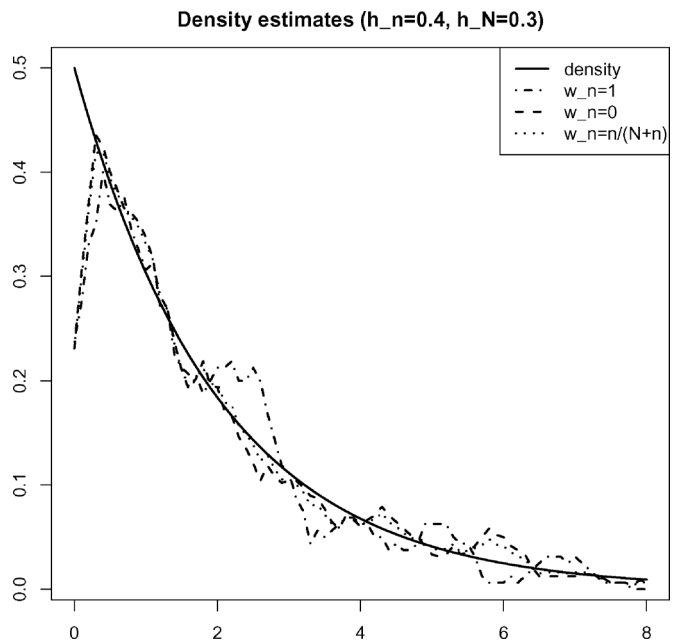

Fig. 3. Density estimates in the second model.

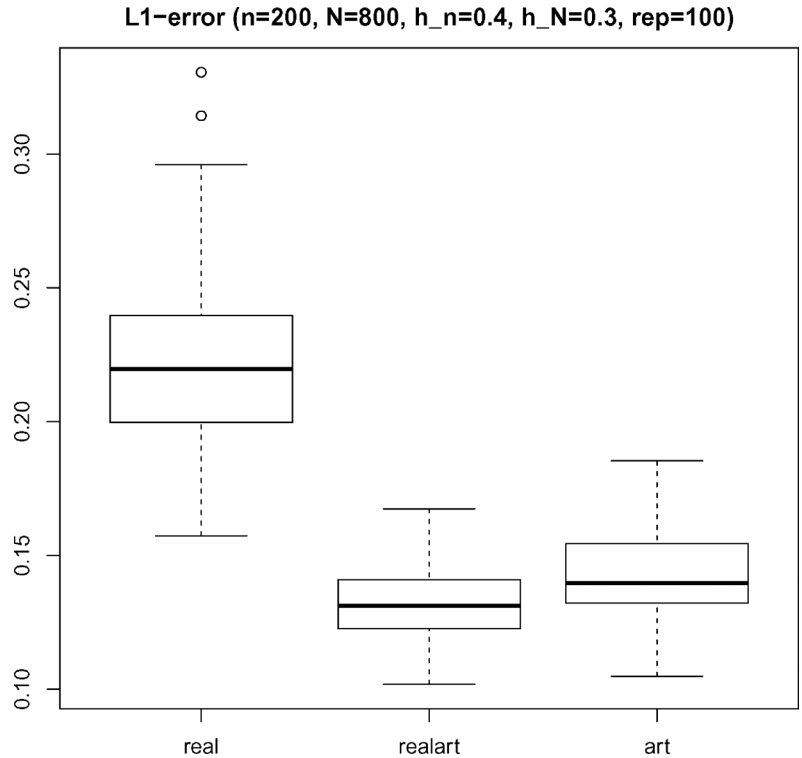

Fig. 4. Boxplots in the second model.

have available an oracle which chooses the optimal bandwidth, so that we can ignore effects occurring because of improper choice of the bandwidths). Fig. 1 shows the three estimates and the true density in a typical simulation. Since the result of our simulation depends on the randomly occurring data points, we repeat the simulation 100 times with independent realizations of the occurring random variables and report in Fig. 2 boxplots of the occurring L_1-errors (where we approximate the integrals by Riemann sums in order to compute the L_1-errors approximately). Comparing the boxplots in Fig. 2, we see that the median of the L_1-errors in case of the estimate which uses only real data (0.1097) is nearly twice as big as the median of the L_1-errors of the estimate which uses only artificial data (0.0648). If we assign the same weight to every data point, it is even smaller (0.0612).

In our second example, we set $X = (X^{(1)}, X^{(2)})$ for independent standard normally distributed random variables $X^{(1)}$ and $X^{(2)}$ and choose $Y = m(X)$ for

$$m(x_1, x_2) = x_1^2 + x_2^2.$$

Then, Y is chi-squared distributed with two degrees of freedom. We define the estimate as in the first example. Again Fig. 3 shows the three estimates and the true density in a typical simulation, and in Fig. 4, we compare boxplots of the L_1-errors of the estimate. From Fig. 4, we see that the median L_1-error of the estimate with $w_n = 0$ (0.1426) is well below the first estimate with $w_n = 1$ (0.2208). The median of the estimate which uses real and artificial data is the smallest (0.1321).

In Figs. 5 and 6, we repeat the same simulation choosing X as a standard-normally distributed random variable and $m(x) = \exp(x)$. In this case, $Y = m(X)$ is log-normally distributed. Fig. 6 shows the same results as before. The estimate which uses only real data is again the worst (0.2221). If every data point has the same weight, the mean L_1-error (0.1341) is smaller than if we use only artificial data (0.1402).

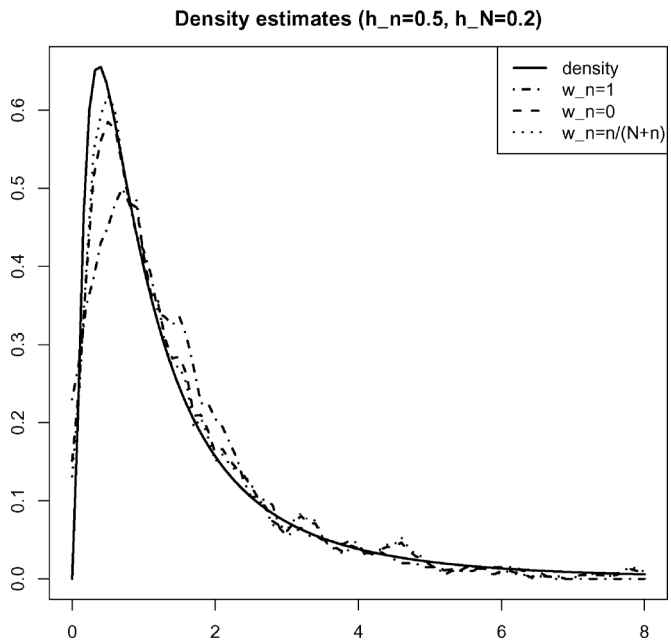

Fig. 5. Density estimates in the third model.

To examine the influence of the parameter N, we set in the third model $n = 200$ and $N = 0, 400, 800, 1200, 1600$. After 100 repetitions, we calculate the empirical mean and standard

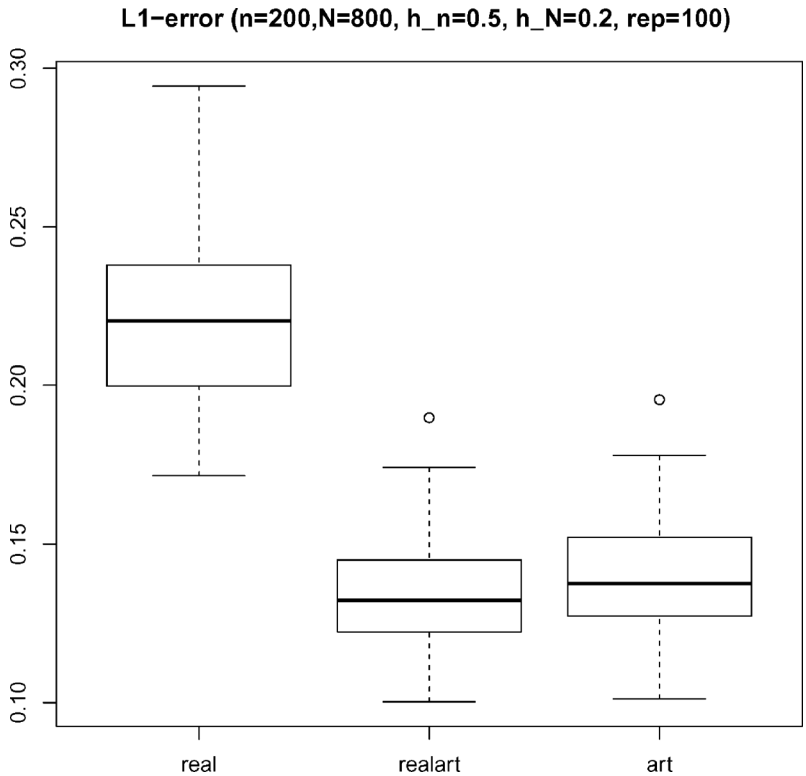

Fig. 6. Boxplots in the third model.

deviation of the approximated L_1-error of the estimate where $w_n = \frac{n}{n+N}$. Both decrease when we raise the parameter N:

N	0	400	800	1200	1600
mean	0.2297	0.1496	0.1270	0.1134	0.1045
sd	0.0274	0.0171	0.0149	0.0134	0.0127

In this example, the speed of decrease of the error does not seem to get smaller for N up to 1600, which might be due to the fact that our regression estimate of $m(X) = \exp(X)$ is rather good.

Finally, we illustrate the usefulness of our estimation procedure by applying it to a density estimation problem in a simulation model. Here, we consider the load distribution in the three legs of a simple tripod. More precisely, a static force is applied on the symmetric tripod to induce mechanical loading equivalent to the weight of 4,5 kg in its three legs. On the bottom side of the legs, force sensors are mounted to measure the leg's axial force. For a safe and stable standing of the tripod, the legs are angled with $\alpha = 5°$ from the middle axis of the connecting devise. Engineers expect that if the holes where the legs are plugged in have a diameter of 15 mm, a third of the general load should be measured in each leg. Unfortunately, a gouging of exactly 15 mm is not possible in the manufacturing process. In the simulation, we assume that the diameters behave like a standard normally distributed random variable with expectation 15 and standard deviation 0.5. Based on the physical model of the tripod, we are able to calculate the resulting load distribution in dependence of the three values of the diameter. Since in this case the real density is unknown, we repeat the simulation 10 000 times to generate a high sample of relative loads. For simplicity, we consider only one leg of the tripod. Application of the routine *density* in the statistics package R to these 10 000 observed values leads to the solid line in Fig. 7. We calculate our estimates as described before using 200 real and 800 artificial data. Again the estimates that use artificial data achieve better results than the estimate with $w_n = 1$. The difference between the estimate with $w_n = 0$ and the estimate with $w_n = n/(N+n)$ is not visible.

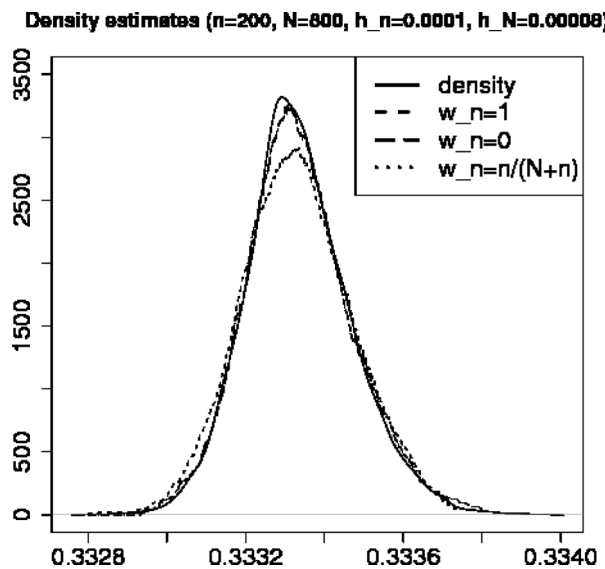

Fig. 7. Density estimates in the fourth model.

VI. Outlook

The consistency result gives us only general guidelines for the joint choice of $\widehat{h}_N$ and m_n. The rate of convergence of $\int |g_n(y) - f(y)|\, dy$ needs to be studied in detail. This can only be attempted if we specify the data-dependent weight function w_n. It is clear that the individual rates of

$$\int |f_n(y) - f(y)|\, dy \quad \text{and} \quad \int |\widehat{f}_N(y) - f(y)|\, dy$$

will form the basis of such a study.

Help can come from $\widehat{f}_N$ only if it performs well. Our examples outlined a few possible situations. The discrete analog of the previous section shows even more dramatically the improvements one can expect if N is very large.

There is also a need to develop a suitable minimax theory for our estimation problem for appropriate subclasses of distributions of X and regression functions m.

Even further afield, one might consider vector-valued regression functions m.

Finally, our generic estimate itself was kernel-based. It is of interest to explore a nearest neighbor or space partitioning version for both m_n and the definition of $\widehat{f}_N$.

In this respect, Györfi and Walk [20] have embarked on a study of the performance of density estimates based on those functions. They obtained explicit rates of convergence under certain smoothness assumptions.

Appendix
Proof of Theorem 1

We prove only the almost sure version of the Theorem 1. The other part can be derived in the same way. Since

$$\begin{aligned}
&\int_{\mathbb{R}} |g_n(y) - f(y)|\, dy \\
&\le w_n \cdot \int_{\mathbb{R}} |f_n(y) - f(y)|\, dy \\
&\quad + (1 - w_n) \cdot \int_{\mathbb{R}} |\widehat{f}_N(y) - f(y)|\, dy \\
&\le \int_{\mathbb{R}} |f_n(y) - f(y)|\, dy + \int_{\mathbb{R}} |\widehat{f}_N(y) - f(y)|\, dy
\end{aligned}$$

it suffices to show

$$\int_{\mathbb{R}} |f_n(y) - f(y)|\, dy \to 0 \tag{14}$$

almost surely and

$$\int_{\mathbb{R}} |\widehat{f}_N(y) - f(y)|\, dy \to 0 \tag{15}$$

almost surely. Equation (14) follows from [7], so it suffices to show (15).

As in the proof of Theorem 1 in [9], we conclude from McDiarmid's inequality (cf., [25]) that (15) is implied by

$$\mathbf{E}\left\{ \int_{\mathbb{R}} |\widehat{f}_N(y) - f(y)|\, dy \middle| \mathcal{D}_n \right\} \to 0 \tag{16}$$

almost surely, which we show in the sequel.

Set $(a)_+ = \max\{a, 0\}$ for $a \in \mathbb{R}$. By Scheffé's Lemma, we know

$$\begin{aligned}
&\int_{\mathbb{R}} |\widehat{f}_N(y) - f(y)|\, dy \\
&= 2 \cdot \int_{\mathbb{R}} (f(y) - \widehat{f}_N(y))_+\, dy \\
&\le 2 \cdot \int_B (f(y) - \widehat{f}_N(y))_+\, dy + 2 \cdot \int_{B^c} f(y)\, dy
\end{aligned}$$

for any $B \subseteq \mathbb{R}$; hence, it suffices to show that we have for any compact set $B \subseteq \mathbb{R}$

$$\mathbf{E}\left\{ \int_B (f(y) - \widehat{f}_N(y))_+\, dy \middle| \mathcal{D}_n \right\} \to 0 \tag{17}$$

almost surely. Since

$$(a)_+ \le |b| + (a - b)_+, \quad \text{for } a, b \in \mathbb{R} \tag{18}$$

this in turn is implied by

$$\mathbf{E}\left\{ \int_B \left| \widehat{f}_N(y) - \mathbf{E}\left\{ \widehat{f}_N(y) | \mathcal{D}_n \right\} \right|\, dy \middle| \mathcal{D}_n \right\} \to 0 \tag{19}$$

almost surely and

$$\int_B \left(f(y) - \mathbf{E}\left\{ \widehat{f}_N(y) | \mathcal{D}_n \right\} \right)_+ dy \to 0 \tag{20}$$

almost surely.

In the first step of the proof, we show (19). By Cauchy–Schwarz and the inequality of Jensen, we have

$$\begin{aligned}
&\mathbf{E}\left\{ \int_B \left| \widehat{f}_N(y) - \mathbf{E}\left\{ \widehat{f}_N(y) | \mathcal{D}_n \right\} \right|\, dy \middle| \mathcal{D}_n \right\} \\
&\le \sqrt{\int_B 1\, dy} \\
&\qquad \cdot \mathbf{E}\left\{ \sqrt{\int_B \left| \widehat{f}_N(y) - \mathbf{E}\left\{ \widehat{f}_N(y) | \mathcal{D}_n \right\} \right|^2 dy} \middle| \mathcal{D}_n \right\} \\
&\le \sqrt{\int_B 1\, dy} \\
&\qquad \cdot \sqrt{\mathbf{E}\left\{ \int_B \left| \widehat{f}_N(y) - \mathbf{E}\left\{ \widehat{f}_N(y) | \mathcal{D}_n \right\} \right|^2 dy \middle| \mathcal{D}_n \right\}}.
\end{aligned}$$

Using the theorem of Fubini and the conditional independence of $\widehat{Y}_1, \ldots, \widehat{Y}_N$, we get

$$\begin{aligned}
&\mathbf{E}\left\{\int_B \left|\widehat{f}_N(y) - \mathbf{E}\left\{\widehat{f}_N(y)\middle|\mathcal{D}_n\right\}\right|^2 dy\middle|\mathcal{D}_n\right\} \\
&= \int_B \mathbf{E}\left\{\left|\widehat{f}_N(y) - \mathbf{E}\left\{\widehat{f}_N(y)\middle|\mathcal{D}_n\right\}\right|^2 \middle|\mathcal{D}_n\right\} dy \\
&\le \int_B \frac{1}{N^2 \cdot \widehat{h}_N^2} \cdot \sum_{i=1}^{N} \mathbf{E}\left\{K^2\left(\frac{y - m_n(X_{n+i})}{\widehat{h}_N}\right)\middle|\mathcal{D}_n\right\} dy \\
&= \frac{1}{N \cdot \widehat{h}_N^2} \cdot \int_B \int K^2\left(\frac{y - m_n(z)}{\widehat{h}_N}\right) \mu(dz)\, dy \\
&= \frac{1}{N \cdot \widehat{h}_N^2} \cdot \int \int_B K^2\left(\frac{y - m_n(z)}{\widehat{h}_N}\right) dy\, \mu(dz) \\
&\le \frac{1}{N \cdot \widehat{h}_N} \cdot \int \int_{\mathbb{R}} K^2(y)\, dy\, \mu(dz) \\
&= \frac{1}{N \cdot \widehat{h}_N} \cdot \int_{\mathbb{R}} K^2(y)\, dy \to 0 \quad (n \to \infty)
\end{aligned}$$

from which we conclude (19) via (1).

In the second step of the proof, we show (20). Let $B \subseteq \mathbb{R}$ be an arbitrary compact set, let $\epsilon > 0$ be arbitrary, and let $A_{n,\epsilon}$ be defined as in the theorem. Then

$$\begin{aligned}
&\int_B \left(f(y) - \mathbf{E}\left\{\widehat{f}_N(y)|\mathcal{D}_n\right\}\right)_+ dy \\
&= \int_B \left(f(y) - \int \frac{1}{\widehat{h}_N} K\left(\frac{y - m_n(x)}{\widehat{h}_N}\right) \mu(dx)\right)_+ dy \\
&\le \int_B \left(f(y) \right. \\
&\qquad \left. - \int \frac{1}{\widehat{h}_N} K\left(\frac{y - m_n(x)}{\widehat{h}_N}\right) \cdot 1_{A_{n,\epsilon}}(x)\, \mu(dx)\right)_+ dy.
\end{aligned}$$

Since K is the naive kernel, we have for any $x \in A_{n,\epsilon}$ in case that $\widehat{h}_N > \|m_n - m\|_{\infty, A_{n,\epsilon}}$

$$\begin{aligned}
&K\left(\frac{y - m(x)}{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}\right) = \frac{1}{2} \\
&\Leftrightarrow m(x) - \widehat{h}_N + \|m_n - m\|_{\infty, A_{n,\epsilon}} \\
&\quad \le \quad y \quad \le m(x) + \widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}} \\
&\Rightarrow m(x) - \widehat{h}_N + (m_n(x) - m(x)) \\
&\quad \le \quad y \quad \le m(x) + \widehat{h}_N - (m(x) - m_n(x)) \\
&\Leftrightarrow m_n(x) - \widehat{h}_N \le y \le m_n(x) + \widehat{h}_N \\
&\Leftrightarrow K\left(\frac{y - m_n(x)}{\widehat{h}_N}\right) = \frac{1}{2}
\end{aligned}$$

which implies

$$K\left(\frac{y - m_n(x)}{\widehat{h}_N}\right) \ge K\left(\frac{y - m(x)}{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}\right).$$

Using this and (18), we conclude that we have in case $\widehat{h}_N > \|m_n - m\|_{\infty, A_{n,\epsilon}}$ [which happens for n sufficiently large with probability one by assumption (5)]

$$\begin{aligned}
&\int_B \left(f(y) - \int \frac{1}{\widehat{h}_N} K\left(\frac{y - m_n(x)}{\widehat{h}_N}\right) \right. \\
&\qquad \left. \cdot \mathbb{I}_{A_{n,\epsilon}}(x)\, \mu(dx)\right)_+ dy \\
&\le \int_B \left(f(y) - \int \frac{1}{\widehat{h}_N} K\left(\frac{y - m(x)}{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}\right) \right. \\
&\qquad \left. \cdot \mathbb{I}_{A_{n,\epsilon}}(x)\, \mu(dx)\right)_+ dy \\
&\le \int_B \left(f(y) - \int \frac{1}{\widehat{h}_N} \right. \\
&\qquad \left. \cdot K\left(\frac{y - m(x)}{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}\right) \mu(dx)\right)_+ dy \\
&\quad + \int_B \int \frac{1}{\widehat{h}_N} K\left(\frac{y - m(x)}{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}\right) \\
&\qquad \cdot \mathbb{I}_{A_{n,\epsilon}^c}(x)\, \mu(dx)\, dy \\
&\le \int_B \left(f(y) - \int_{\mathbb{R}} \frac{1}{\widehat{h}_N} \right. \\
&\qquad \left. \cdot K\left(\frac{y - z}{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}\right) \cdot f(z)\, dz\right)_+ dy \\
&\quad + \frac{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}{\widehat{h}_N} \cdot \int_{\mathbb{R}} K(y)\, dy \int 1_{A_{n,\epsilon}^c}(x)\, \mu(dx).
\end{aligned}$$

By Lebesgue's density theorem (cf., e.g., Theorem 2 or Theorem 3 in [10]) and (1) and (5), we know that

$$\begin{aligned}
&\int_{\mathbb{R}} \frac{1}{\widehat{h}_N} K\left(\frac{y - z}{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}\right) \cdot f(z)\, dz \\
&= \frac{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}{\widehat{h}_N} \cdot \int_{\mathbb{R}} \frac{1}{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}} \\
&\qquad \cdot K\left(\frac{y - z}{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}\right) \cdot f(z)\, dz \\
&\to 1 \cdot f(y) = f(y) \quad (n \to \infty)
\end{aligned}$$

for almost all $y \in \mathbb{R}$ with probability one, which implies (via the dominated convergence theorem)

$$\begin{aligned}
&\int_B \left(f(y) - \int_{\mathbb{R}} \frac{1}{\widehat{h}_N} \right. \\
&\qquad \left. \cdot K\left(\frac{y - z}{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}\right) \cdot f(z)\, dz\right)_+ dy \\
&\to 0
\end{aligned}$$

almost surely. Furthermore

$$\frac{\widehat{h}_N - \|m_n - m\|_{\infty, A_{n,\epsilon}}}{\widehat{h}_N} \cdot \int_{\mathbb{R}} K(y)\, dy \int 1_{A^c_{n,\epsilon}}(x)\, \mu(dx) \leq 1 \cdot 1 \cdot \mu(A^c_{n,\epsilon}).$$

Summarizing the above result, we get

$$\limsup_{n\to\infty} \int_B \left(f(y) - \mathbf{E}\left\{ \widehat{f}_N(y) | \mathcal{D}_n \right\} \right)_+ dy \leq \epsilon$$

almost surely, and with $\epsilon \to 0$ this implies (20). The proof is complete. □

ACKNOWLEDGMENT

The authors would like to thank two anonymous referees for various very helpful comments. Discussions with Adam Krzyżak are gratefully acknowledged.

REFERENCES

[1] I. A. Ahmad, "Residuals density estimation in nonparametric regression," *Statist. Probab. Lett.*, vol. 14, pp. 133–139, 1992.
[2] J. Beirlant and L. Györfi, "On the asymptotic L_2-error in partitioning regression estimation," *J. Statist. Plann. Infer.*, vol. 71, pp. 93–107, 1998.
[3] V. Castelli and T. Cover, "The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2101–2117, Nov. 1996.
[4] O. Chapelle and B. Z. SchölkopfA, *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.
[5] F. Cheng, "Weak and strong uniform consistency of a kernel error density estimator in nonparametric regression," *J. Statist. Plann. Infer.*, vol. 119, pp. 95–107, 2004.
[6] L. Devroye, "Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 61, pp. 467–481, 1982.
[7] L. Devroye, "The equivalence in L1 of weak, strong and complete convergence of kernel density estimates," *Ann. Statist.*, vol. 11, pp. 896–904, 1983.
[8] L. Devroye, *A Course in Density Estimation*. Basel, Switzerland: Birkhäuser, 1987.
[9] L. Devroye, T. Felber, M. Kohler, and A. Krzyżak, "L_1-consistent estimation of the density of residuals in random design regression models," *Statist. Probab. Lett.*, vol. 82, pp. 173–179, 2012.
[10] L. Devroye and L. Györfi, *Nonparametric Density Estimation. The L1 view*, ser. Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. New York: Wiley, 1985.
[11] L. Devroye, L. Györfi, A. Krzyżak, and G. Lugosi, "On the strong universal consistency of nearest neighbor regression function estimates," *Ann. Statist.*, vol. 22, pp. 1371–1385, 1994.
[12] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. New York: Springer-Verlag, 2000.
[13] L. Devroye and A. Krzyżak, "An equivalence theorem for L_1 convergence of the kernel regression estimate," *J. Statist. Plann. Infer.*, vol. 23, pp. 71–82, 1989.
[14] L. Devroye and T. J. Wagner, "Distribution-free consistency results in nonparametric discrimination and regression function estimation," *Ann. Statist.*, vol. 8, pp. 231–239, 1980.
[15] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Ann. Math. Statist.*, vol. 27, pp. 642–669, 1956.
[16] S. Efromovich, "Estimation of the density of regression errors," *Ann. Statist.*, vol. 33, pp. 2194–2227, 2005.
[17] S. Efromovich, "Optimal nonparametric estimation of the density of regression errors with finite support," *Ann. Inst. Statist. Math.*, vol. 59, pp. 617–654, 2006.
[18] L. Györfi, "Recent results on nonparametric regression estimate and multiple classification," *Probl. Control Inf. Theory*, vol. 10, pp. 43–52, 1981.
[19] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer-Verlag, 2002.
[20] L. Györfi and H. Walk, "Rate of convergence of the density estimation of regression residual," *Statist. Risk Model.*, vol. 29, pp. 1–17, 2012.
[21] M. Kohler, "Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression," *J. Statist. Plann. Infer.*, vol. 89, pp. 1–23, 2000.
[22] M. Kohler and A. Krzyżak, "Nonparametric regression estimation using penalized least squares," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 3054–3058, Nov. 2001.
[23] G. Lugosi and K. Zeger, "Nonparametric estimation via empirical risk minimization," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 677–687, May 1995.
[24] R. M. Mnatsakanov and E. V. Khmaladze, "On L_1-convergence of statistical kernel estimators of distribution densities," *Soviet Math. Doklady*, vol. 23, pp. 633–636, 1981.
[25] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics*, ser. London Mathematical Society Lecture Notes Series. Cambridge, U.K.: Cambridge Univ. Press, 1989, vol. 141, pp. 148–188.
[26] E. A. Nadaraya, "On estimating regression," *Theory Probab. Appl.*, vol. 9, pp. 141–142, 1964.
[27] E. A. Nadaraya, "Remarks on nonparametric estimates for density functions and regression curves," *Theory Probab. Appl.*, vol. 15, pp. 134–137, 1970.
[28] E. Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.
[29] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist.*, vol. 27, pp. 832–837, 1956.
[30] C. J. Stone, "Consistent nonparametric regression," *Ann. Statistist.*, vol. 5, pp. 595–645, 1977.
[31] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *Ann. Statist.*, vol. 10, pp. 1040–1053, 1982.
[32] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA: SIAM, 1990.
[33] G. S. Watson, "Smooth regression analysis," *Sankhya Ser. A*, vol. 26, pp. 359–372, 1964.

Luc Devroye was born in Tienen, Belgium, on August 6, 1958. He has a Bachelors and a Ph.D. in Electrical Engineering from KU Leuven (Belgium) and the University of Texas at Austin, respectively. His thesis dealt with nonparametric density estimation.

Immediately after graduating, he joined the School of Computer Science at McGill University in Montreal, Canada, where he still teaches today. He has contributed to the areas of mathematical statistics, probability theory, random number generation, random graphs, and probabilistic analysis of algorithms.

Tina Felber was born in Frankfurt, Germany, on June 4, 1986. She received a degree in mathematics from the Technical University of Darmstadt, Germany, in 2010. Her Ph.D. thesis deals with nonparametric density estimation of the residuals of a stationary and ergodic time series.

Since 2010 she works as a research assistant at the Technical University of Darmstadt, Germany.

Michael Kohler was born in Esslingen, Germany, on July 17, 1969. He received degrees in computer science and mathematics from the University of Stuttgart in 1995, and the Ph.D degree in mathematics from the University of Stuttgart in 1997.

In 1998 he worked as a Visiting Scientist at the Stanford University, Stanford, USA. From 2005 till 2007 he was Professor of Applied Mathematics at the University of Saarbrücken, Germany, since 2007 he is Professor of Mathematical Statistics at the Technische Universität Darmstadt, Germany.

He coauthored with L. Györfi, A. Krzyżak and H. Walk the book *A Distribution-Free Theory of Nonparametric Regression* (New York: Springer, 2002). His main research interests are in the area of nonparametric statistics, especially curve estimation and applications in mathematical finance.