Adaptive density estimation based on real and artificial data *

Tina Felber[1], Michael Kohler[1] and Adam Krzyżak[2,†]

[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: tfelber@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de*

[2] *Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8, email: krzyzak@cs.concordia.ca*

October 2, 2013

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. The problem of estimating f based on a sample of the distribution of (X, Y) and on additional independent observations of X is considered. Two kernel density estimates are compared: the standard kernel density estimate based on the y-values of the sample of (X, Y), and a kernel density estimate based on artificially generated y-values corresponding to the additional observations of X. It is shown that under suitable smoothness assumptions on f and m the rate of convergence of the L_1 error of the latter estimate is better than that of the standard kernel density estimate. Furthermore, a density estimate defined as convex combination of these two estimates is considered and a data-driven choice of its parameters (bandwidths and weight of the convex combination) is proposed and analyzed.

AMS classification: Primary 62G07; secondary 62G20.

Key words and phrases: Density estimation, L_1–error, nonparametric regression, rate of convergence, adaptation.

1 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 a measurable function such that a density f of $Y = m(X)$ exists. In the sequel we study the problem of estimating f from the data

$$(X_1, Y_1), \ldots, (X_n, Y_n), X_{n+1}, \ldots, X_{n+N}$$

*Running title: *Density estimation based on real and artificial data*

†Corresponding author: Tel. +1 514 848 2424, ext. 3007, Fax. +1 514 848 2830

for some $n, N \in \mathbb{N}$.

This problem is motivated by experiments carried out in the Collaborative Research Centre 805 which is interested in the measurement of uncertainty in load-bearing systems like bearing structures of aeroplanes. The simplest example of a load-bearing system in mechanical engineering is a tripod (Figure 1). In the experiments a static force is applied on the tripod. On the bottom side of the legs force sensors are mounted to measure the legs' axial force. If the holes where the legs are plugged in have exactly the same diameter, a third of the general load should be measured in each leg. Unfortunately, such an accurate drilling is not possible in the manufacturing process. Since there is always a small deviation, the force is distributed nonuniformly in the three legs. The random vector $X = (X^{(1)}, X^{(2)}, X^{(3)})$ represents the diameters of the three holes. The function $m : \mathbb{R}^3 \to \mathbb{R}$ describes the physical model of the tripod and $Y = m(X)$ is the resulting load. Here, the measurement of X is very cheap, so there are many observations of the diameters available. Based on the physical model of the tripod we are able to compute the reliability $Y_i = m(X_i)$ for the observed diameters X_i, but due to the fact that this is an expensive and time consuming process we do this only n times and observe additional N values of the random diameter X.

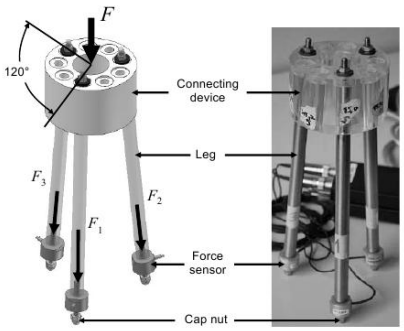

Figure 1: Tripod

The task then is to improve the estimate of the distribution of the reliability by using the additional observations of X. The distribution of the reliability is described by its density, which we assume to exist, and by controlling the L_1 error of an estimate of this density we can bound via the Lemma of Scheffé (cf., e.g., Devroye and Györfi (1985)) the total variation error of the corresponding estimate of the distribution. So we are facing the problem of estimating the density of $Y = m(X)$ given a sample of Y and additional independent observations of X.

The easiest method to do this is to ignore the additional observations of X completely and to simply estimate f by the well-known kernel density estimate f_n (cf., Parzen (1962) and Rosenblatt (1956)) applied to the sample of Y defined by

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

Here $h_n > 0$ is the so-called bandwidth and the kernel $K : \mathbb{R} \to \mathbb{R}$, e. g., naive kernel $K(u) = 1/2 \cdot 1_{[-1,1]}(u)$, is a density. This estimate is L_1 consistent for all densities under the following conditions on the bandwidth

$$h_n \to 0, \quad n \cdot h_n \to \infty \quad (n \to \infty),$$

see Devroye (1983). Further results on density estimation can be found in several books. Devroye and Györfi (1985) present L_1 theory, Devroye (1987) gives a course on density estimation discussing among others the rates of convergence and superkernels, Devroye and Lugosi (2001) introduce combinatorial tools for density estimation, Eggermont and LaRiccia (2001) discuss maximum likelihood approach, a general approach to density estimation is presented in Scott (1992) and Wand and Jones (1995) and L_2 theory is presented in Tsybakov (2008).

In Devroye, Felber and Kohler (2013) it was proposed to consider artificially generated data

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

where $m_n(\cdot) = m_n(\cdot, (X_1, Y_1), \ldots, (X_n, Y_n))$ is a suitable regression estimate of m. For instance we can use kernel regression estimate (cf., e.g., Nadaraya (1964, 1970), Watson (1964), Devroye and Wagner (1980), Stone (1977, 1982) or Devroye and Krzyżak (1989)), partitioning regression estimate (cf., e.g., Györfi (1981) or Beirlant and Györfi (1998)), nearest neighbor regression estimate (cf., e.g., Devroye (1982) or Devroye, Györfi, Krzyżak and Lugosi (1994)), least squares estimates (cf., e.g., Lugosi and Zeger (1995) or Kohler (2000)) or smoothing spline estimates (cf., e.g., Wahba (1990) or Kohler and Krzyżak (2001)). They defined a kernel density estimate based on these artificial data by

$$g_N(y) = \frac{1}{N \cdot h_N} \cdot \sum_{i=1}^{N} K\left(\frac{y - m_n(X_{n+i})}{h_N}\right)$$

(with some (possible different) bandwidth $h_N > 0$) and used a convex combination

$$\begin{aligned}
&\hat{f}_n(y) \\
&= w \cdot f_n(y) + (1 - w) \cdot g_N(y) \\
&= w \cdot \frac{1}{n \cdot h_n} \cdot \sum_{i=1}^{n} K\left(\frac{y - m(X_i)}{h_n}\right) + (1 - w) \cdot \frac{1}{N \cdot h_N} \cdot \sum_{i=1}^{N} K\left(\frac{y - m_n(X_{n+i})}{h_N}\right).
\end{aligned}$$

Devroye, Felber and Kohler (2013) showed that this estimate is under suitable conditions on m, m_n, h_n and h_N universally L_1 consistent, i.e., the L_1 error of the estimate converges to zero (almost surely and in L_1) for all densities f. Furthermore it was shown by using simulated data that the use of the artificial data indeed improves the estimate of f.

In this paper we analyze the rate of convergence of the estimate g_N and identify situations where it is better than the rate of convergence of the standard kernel density estimate f_n. Furthermore we propose and analyze a data-driven method for choosing the parameters of $\hat{f}_n$ (i.e., the weight of the convex combination and the two different bandwidths).

The outline of the paper is as follows: In Section 2 we present our results concerning the rate of convergence of g_N, in Section 3 we introduce a data-driven method for choosing the parameters and present a theoretical adaptation result, the finite sample size performance is illustrated in Section 4 and Section 5 contains the proofs.

2 The density estimate based on artificial data

In this section the estimate

$$g_N(y) = \frac{1}{N \cdot h_N} \cdot \sum_{i=1}^{N} K\left(\frac{y - m_n(X_{n+i})}{h_N}\right)$$

is analyzed, where $m_n(\cdot) = m_n(\cdot; (X_1, Y_1), \ldots, (X_n, Y_n)) : \mathbb{R}^d \to \mathbb{R}$ is a suitable estimate of m based on the sample of (X, Y). In the sequel we assume that K is the naive kernel, i.e., $K : \mathbb{R} \to \mathbb{R}$ is defined by

$$K(u) = \frac{1}{2} \cdot 1_{[-1,1]}(u).$$

In the next theorem we analyze the rate of convergence of the L_1 error of g_N for Hölder continuous f.

Theorem 1 *Assume that*

(i) f has compact support, i.e., there exists a compact set $B \subseteq \mathbb{R}$ such that $f(x) = 0$ for all $x \notin B$,

(ii) f is Hölder continuous with exponent $r \in (0, 1]$ and Hölder constant C, i.e.,

$$|f(x) - f(z)| \leq C \cdot |x - z|^r \quad \text{for all } x, z \in \mathbb{R},$$

and let the estimate g_N of f be defined as above. Then there exist constants $c_1, c_2 > 0$ such that

$$\mathbf{E} \int |g_N(y) - f(y)| \, dy \leq \frac{c_1}{\sqrt{N \cdot h_N}} + c_2 \cdot h_N^r + 2 \cdot \frac{\mathbf{E}\left\{\min\{|m_n(X) - m(X)|, 2 \cdot h_N\}\right\}}{h_N}.$$

The right hand-side above can be bounded from above by

$$\frac{c_1}{\sqrt{N \cdot h_N}} + c_2 \cdot h_N^r + 2 \cdot \frac{\mathbf{E}\left\{|m_n(X) - m(X)|\right\}}{h_N}.$$

Minimizing $c_2 \cdot h_N^r + \frac{\mathbf{E}\{|m_n(X) - m(X)|\}}{h_N}$ with respect to h_N yields

$$h_N \approx \left(\mathbf{E}\left\{|m_n(X) - m(X)|\right\}\right)^{1/(r+1))}$$

so for N sufficiently large (such that the first term on the right-hand side of the above bound is negligible) and we get

$$\mathbf{E} \int |g_N(y) - f(y)| \, dy = O\left(\left(\mathbf{E}\left\{|m_n(X) - m(X)|\right\}\right)^{\frac{r}{r+1}}\right).$$

If we compare this with the optimal minimax rate of convergence

$$n^{-r/(2r+1)}$$

(cf., e.g., Devroye and Györfi (1985) and Devroye (1987)) for the L_1 error of an estimate of a (r, C)-Hölder smooth density we see that the rate of convergence of g_N is better if

$$\mathbf{E}\left\{|m_n(X) - m(X)|\right\} \leq n^{-\frac{r+1}{2r+1}}.$$

Since $(r+1)/(2r+1) \geq 1/2$ this requires a convergence rate better than the usual rate of convergence $n^{-p/(2p+d)}$ which we get for the expected L_1 error of nonparametric regression estimates (cf., e.g., Stone (1982)). However, in our setting we observe Y_i without error, and it is possible to get better rates of convergence in this case.

In the sequel we estimate functions which are (p, C)-smooth in the following sense:

Definition 1 *Let $p = k + \beta$ for some $k \in \mathbb{N}_0$ and $0 < \beta \leq 1$, and let $C > 0$. A function $m : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-smooth, if for every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\sum_{j=1}^d \alpha_j = k$ the partial derivative $\frac{\partial^k m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ exists and satisfies*

$$\left| \frac{\partial^k m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(x) - \frac{\partial^k m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot \|x - z\|^\beta$$

for all $x, z \in \mathbb{R}^d$, where $\mathbb{N}_0$ is the set of non-negative integers.

We estimate m by a piecewise constant nearest neighbor regression estimate defined as follows (see Kohler and Krzyżak (2013)). For $x \in \mathbb{R}^d$ let $X_{(1,n)}(x)$, ..., $X_{(n,n)}(x)$ be a permutation of $X_1, \ldots, X_n$ such that

$$\|x - X_{(1,n)}(x)\| \leq \cdots \leq \|x - X_{(n,n)}(x)\|.$$

In case of ties, i.e., in case $X_i = X_j$ for some $1 \leq i < j \leq n$, we use tie breaking by indices, i.e., we choose the data point with the smaller index first. Then we define the 1-nearest neighbor estimate by

$$m_n(x) = m_n(x, (X_1, m(X_1)), \ldots, (X_n, m(X_n))) = m(X_{(1,n)}). \tag{1}$$

For this estimate the following error bound was shown in Kohler and Krzyżak (2013):

Theorem 2 *Assume $supp(X) \subseteq [0,1]^d$, $0 < p \leq 1$, $C > 0$ and let $m : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary (p, C)-smooth function. Let m_n be defined by (1). Then*

$$\mathbf{E} \int |m_n(x) - m(x)| \, \mathbf{P}_X(dx) \leq \begin{cases} c_1 \cdot n^{-p/d} & \text{if } p < d, \\ c_2 \cdot \frac{\log n}{n} & \text{if } p = d = 1 \end{cases} \tag{2}$$

for some constants c_1, $c_2 \in \mathbb{R}_+$.

Proof. See Theorem 1 in Kohler and Krzyżak (2013) □

Using results of Liitiäinen, Corona and Lendasse, A. (2010) (see Theorem 3.2 therein) one can show that $\log n$ factor can be dropped in the second line of (2) if the ties occur only with probability zero. Furthermore it was shown in Theorem 3 in Kohler and Krzyżak (2013) that the above rates of convergence are optimal in a minimax setting.

From Theorems 1 and 2 we get

Corollary 1 *Let $d = 1$, assume that m is Lipschitz continuous and that the assumptions of Theorem 1 hold. Set $h_N = (\log(n)/n)^{1/(r+1)}$ and let the estimates m_n and g_N be defined as above. Then $N \geq (n/\log(n))^{(2r+1)/(r+1)}$ implies*

$$\mathbf{E} \int |g_N(y) - f(y)| \, dy \leq c_3 \cdot \left(\frac{\log(n)}{n} \right)^{\frac{r}{r+1}}$$

for some $c_3 > 0$. Hence under the above assumptions the rate of of convergence of the expected L_1 error of g_N converges faster to zero than the minimax rate of convergence $n^{-r/(2r+1)}$ for estimation of (r, C)-Hölder smooth densities.

Proof. The assertion follows directly from Theorems 1 and 2. □

The rates for the 1-nearest neighbor estimate are always less than $1/n$. Following the ideas presented in Kohler and Krzyżak (2013) we next show that in case $d = 1$ we can define a nearest neighbor polynomial interpolation estimate which achieves under regularity assumption on the design distribution for smooth functions the rates that are even better than $1/n$. This way we can under appropriate smoothness conditions on m achieve even better rates than in Theorem 2. The corresponding result is an improvement of the results in Section 4 in Devroye et al. (2012), where a slightly weaker rate of convergence was proven for a more complicated estimate. The estimate will depend on a parameter $k \in \mathbb{N}$. Given x and the data $(X_1, Y_1) \ldots, (X_n, Y_n)$, we first choose the k nearest neighbors $X_{(1,n)}(x), \ldots, X_{(k,n)}(x)$ of x among $X_1, \ldots, X_n$, then we choose a polynomial $\hat{p}_x$ of degree $k-1$ interpolating $(X_{(1,n)}(x), m(X_{(1,n)}(x))), \ldots, (X_{(k,n)}(x), m(X_{(k,n)}(x)))$ (such a polynomial always exists and is unique when $X_{(1,n)}(x), \ldots, X_{(k,n)}(x)$ are pairwise disjoint), and define our k-nearest neighbor polynomial interpolating estimate by

$$m_{n,k}(x) = \hat{p}_x(x). \tag{3}$$

For this estimate the following error bound holds:

Theorem 3 *Let $p \in \mathbb{N}$ and $C > 0$, $d = 1$ and assume that $m : \mathbb{R} \to \mathbb{R}$ is (p, C)-smooth and that the distribution of X satisfies supp$(\mathbf{P}_X) \subseteq [0, 1]$,*

$$\mathbf{P}\{X = x\} = 0 \quad \textit{for all } x \in [0, 1] \tag{4}$$

and

$$\mathbf{P}_X(S_r(x)) \geq c_4 \cdot r \tag{5}$$

for all $x \in [0,1]$ and all $r \in (0,1]$ for some constant $c_4 > 0$, where $S_r(x)$ is the closed ball with radius r centered at x. Then for the p-nearest neighbor polynomial interpolating estimate $m_{n,p}$ defined by (3) the following bound holds:

$$\mathbf{E} \int |m_{n,p}(x) - m(x)| \, \mathbf{P}_X(dx) \leq c_5 \cdot n^{-p}$$

for some $c_5 \in \mathbb{R}_+$, where $\mathbb{R}_+$ is the set of positive real numbers.

Proof. See Theorem 2 in Kohler and Krzyżak (2013) □

From Theorems 1 and 3 we can obtain for our density estimate even better rate than in Theorem 2 in case when m is (p, C)-smooth for $p \in \mathbb{N}$, $p \geq 2$:

Corollary 2 *Let $d = 1$ and assume that the assumptions of Theorems 1 and 3 hold. Set $h_N = (n)^{-p/(r+1)}$ and let the estimates m_n and g_N be defined as above. Then $N \geq n^{p \cdot (2r+1)/(r+1)}$ implies*

$$\mathbf{E} \int |g_N(y) - f(y)| \, dy \leq c_6 n^{-p \cdot r/(r+1)}$$

for some $c_6 > 0$.

Proofs. The assertion follows directly from Theorems 1 and 3. □

3 Data dependent choice of the parameters

In this section we present a data-driven choice of the parameter

$$\theta = (h_1, h_2, w) \in \Theta := \{(h_1, h_2, w) \quad : \quad h_1, h_2 > 0, w \in [0,1]\}$$

of the convex combination

$$\begin{aligned} &\hat{f}_{n,\theta}(y) \\ &= w \cdot \frac{1}{n \cdot h_1} \cdot \sum_{i=1}^{n} K\left(\frac{y - m(X_i)}{h_1}\right) + (1-w) \cdot \frac{1}{N \cdot h_2} \cdot \sum_{i=1}^{N} K\left(\frac{y - m_n(X_{n+i})}{h_2}\right). \end{aligned}$$

Here the aim is to minimize the L_1 error of the estimate. We achieve this goal by adapting the so-called combinatorial method for choosing the bandwidth of the kernel density estimate from Devroye and Lugosi (2001) to our situation.

First we choose $l_n \in \{1, \ldots, n-1\}$, e.g., $l_n = \lfloor n/2 \rfloor$. The empirical measure based on $Y_1, \ldots Y_{l_n}$ is denoted by $\hat{\mu}_{l_n}$, i.e.,

$$\hat{\mu}_{l_n}(A) = \frac{1}{l_n} \sum_{i=1}^{l_n} 1_A(Y_i) \quad (A \subseteq \mathbb{R}).$$

Then we compute our density estimate without using these data points via

$$\hat{f}_{n-l_n,\theta}(y)$$

$$= w \cdot \frac{1}{n \cdot h_1} \cdot \sum_{i=l_n+1}^{n} K\left(\frac{y - m(X_i)}{h_1}\right) + (1-w) \cdot \frac{1}{N \cdot h_2} \cdot \sum_{i=1}^{N} K\left(\frac{y - m_{n-l_n}(X_{n+i})}{h_2}\right),$$

where the estimate m_{n-l_n} of m is based only on the data points (X_{l_n+1}, Y_{l_n+1}), ..., (X_n, Y_n), by minimizing

$$\Delta_\theta = \sup_{A \in \mathcal{A}} \left| \int_A \hat{f}_{n-l_n,\theta}(y)\, dy - \hat{\mu}_{l_n}(A) \right|$$

and where

$$\mathcal{A} = \left\{ \left\{ y \in \mathbb{R} : \hat{f}_{n-l_n,\theta_1}(y) > \hat{f}_{n-l_n,\theta_2}(y) \right\} \quad : \quad \theta_1, \theta_2 \in \Theta \right\}.$$

More precisely, we set

$$\hat{f}_n = \hat{f}_{n-l_n,\hat{\theta}_n}(y)$$

where $\hat{\theta}_n \in \Theta$ satisfies

$$\Delta_{\hat{\theta}_n} < \inf_{\theta \in \Theta} \Delta_\theta + \frac{1}{n}.$$

The following result holds:

Theorem 4 *Set $l_n = \lfloor n/2 \rfloor$ and let $\hat{f}_n$ be defined as above. Then*

$$\mathbf{E} \int |\hat{f}_n(y) - f(y)|\, dy$$
$$\leq 3 \cdot \min_{\theta \in \Theta} \mathbf{E} \int |\hat{f}_{n-l_n,\theta}(y) - f(y)|\, dy + 8 \cdot \sqrt{\frac{48 \cdot \log(n) + 16 \cdot \log(N)}{n-2}} + \frac{3}{n}.$$

In case $N \leq n^k$ for some $k \in \mathbb{N}$ our data-driven method chooses the best estimate up to an additional error of order $(\log(n)/n)^{0.5}$, so in case of the 1-nearest neighbor estimate of Theorem 2 we are able to choose the optimal parameter for all Hölder smooth densities. And in case of the nearest neighbor polynomial interpolation estimate our data-driven estimate of the parameters yields a density estimate which achieves the best possible rate for (r, C)-Hölder smooth densities for all $r \leq 1/(2p-1)$.

Remark 1. It is an open problem whether it is possible to construct a data-depend choice of the parameters for which the additional error behaves (especially in the case of a large N) better than $1/\sqrt{n}$.

Remark 2. As pointed out by a referee the results in our paper are related to imputation techniques for missing data, where the missing value of a covariate is replaced by a prediction based on the other covariates (cf., e.g., Section 9.6 in Hastie, Tibshirani and Friedman (2001)). However, in contrast to the methods applied in this context our estimate is applied separately to the predicted missing data and the not missing data and a convex combination of the resulting two estimates is used.

4 Application to simulated and real data

In this section we apply our density estimate to simulated and real data using the data-driven choice of the parameters described in Section 3.

In the next three examples we set the sample size of the observed data $n = 200$, the sample size of the artificial data $N = 800$ and $l_n = \frac{n}{2} = 100$. We use the naive kernel as kernel function K and a fully data-driven smoothing spline estimate to estimate the function m. For this purpose we use the routine *Tps()* from the library *fields* in the statistics package *R*. In our applications below this regression estimate is applied to data where the y-values are observed without error. In this case it is able to compute smooth functions interpolating the data without the need to specify the degree p of the smoothness as in the case of the estimate in Theorem 3.

The parameter $\theta = (h_n, h_N, w)$ is chosen by minimizing Δ_θ over a grid of parameters. The predetermined grid for h_n and h_N is different for the three examples. For the weight we assume $w \in \left\{0, \frac{l_n}{l_n+N}, 1\right\}$. This means, we use only the observed data or only the additional data or every data point is assigned the same weight. The main trick which allows us to compute our estimate in an efficient way is as follows: We approximate the integral in the definition of Δ_θ by a Riemann sum and store all values of $\hat{f}_{n-l_n,\theta}$ at all points needed for computation of the Riemann sum for every θ at the beginning of the computation. In this way the run of one repetition requires only between one and two minutes. Then we compare the proposed estimate with the standard kernel density estimate of Rosenblatt and Parzen and with three density estimates which use a fixed weight $w \in \left\{0, \frac{l_n}{l_n+N}, 1\right\}$ and bandwidths chosen by cross-validation.

In our first example we set $X = (X_1, X_2)$ with independent standard normally distributed random variables X_1 and X_2 and choose $m(x_1, x_2) = 2 \cdot x_1 + x_2 + 2$. In this case $Y = m(X)$ is normally distributed with expectation 2 and variance $2^2 + 1^2 = 5$. Here we choose $h_n \in \{1, 1.25, 1.5, 2\}$ and $h_N \in \{0.6, 0.8, 1\}$. Figure 2 shows both estimates and the underlying density in a typical simulation. The dashed line is the newly proposed estimate and the dotted line the one of Rosenblatt and Parzen.
Since the result of our simulation depends on the randomly occuring data points, we repeat the whole procedure 100 times with independent realizations of the occuring random variables and report boxplots of the L_1-errors in Figure 3. The mean of the L_1-errors of the proposed estimate (0.0927) is less than the mean L_1-error of the Rosenblatt-Parzen density estimate (0.1105) and the one which uses only the observed data (0.1471). The L_1–error of the estimates which use the artificial data are smaller (0.0625 and 0.0607).

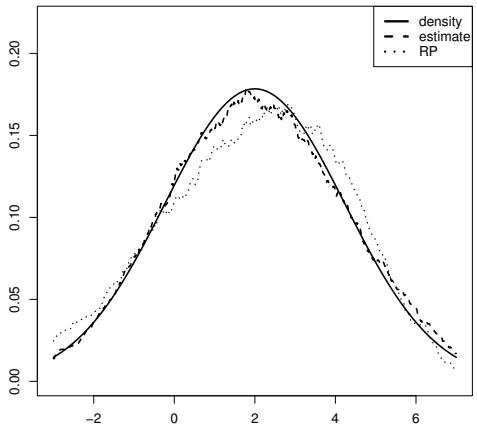

Figure 2: Typical simulation in the first model

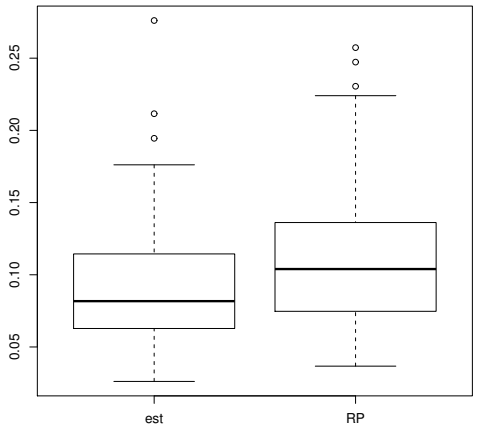

Figure 3: Boxplots of the occuring L_1-errors in the first model

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 $m(x_1, x_2) = x_1^2 + x_2^2$. Then $Y = m(X)$ is chi-squared distributed with two degrees of freedom. Here we choose $h_n \in \{0.25, 0.5, 0.75, 1\}$ and $h_N \in \{0.1, 0.2, 0.3\}$. Figure 4 shows the estimate f_n, the Rosenblatt-Parzen density estimate and the underlying density in a typical simulation.

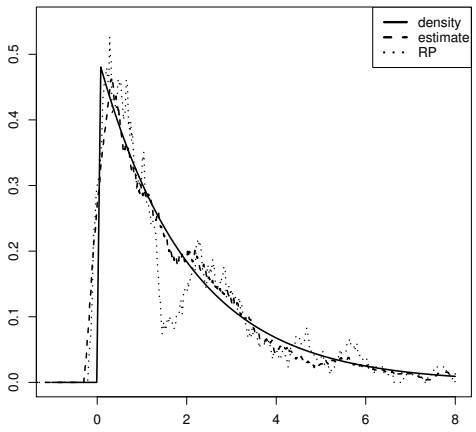

Figure 4: Typical simulation in the second model

In Figure 5 we compare boxplots of the occuring L_1-errors of the estimates. The mean L_1-error of our estimate (0.1701) is much lower than the one of Rosenblatt and Parzen (0.2403) and again much lower than the one which uses only the observed data (0.2969). Again, the estimate which uses only the artifical data (0.1640) and the one where every data point is assigned the same weight (0.1543) are the best.

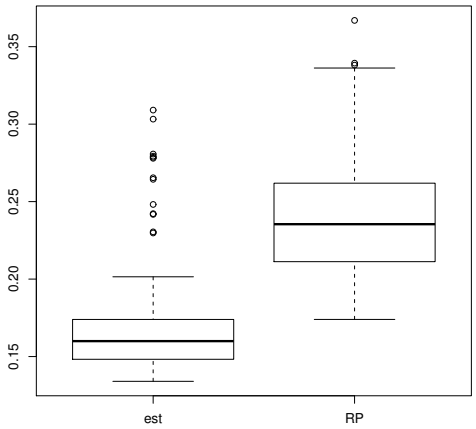

Figure 5: Boxplots of the occuring L_1-errors in the second model

In Figure 6 and Figure 7 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. Here we choose $h_n \in \{0.2, 0.3, 0.4, 0.5\}$ and $h_N \in \{0.1, 0.2, 0.25, 0.3\}$. The mean of the L_1-errors of the estimate f_n is again much lower (0.1531) than the mean error of the Rosenblatt-Parzen estimate (0.2171) and the estimate where $w = 1$ (0.2750). The mean errors of the estimates where $w = 0$ (0.1307) and $w = \frac{l_n}{l_n+N}$ (0.1256) are again the smallest.

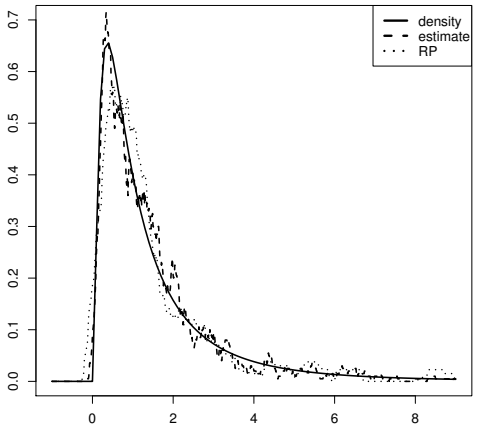

Figure 6: Typical simulation in the third model

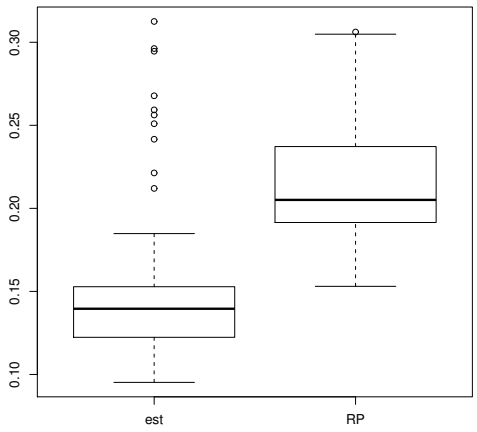

Figure 7: Boxplot of the occuring L_1-errors in the third model

In all three cases our proposed estimate using the data-dependent choice of the parameters works better that the one of Rosenblatt-Parzen and the one which sets $w = 1$ and chooses the bandwidth by cross-validation. However, the L_1–errors of the two estimates which use the additional data but chose the bandwidth by cross-validation are in all three examples a little bit lower, so in principle we could always use e.g. the estimate using only the artificial data. However, in case that the artificial data contains large errors this is certainly not a good estimate. We show next that our data-driven method of the parameters is able to identify such situations. For this purpose, we consider the first model where $X = (X_1, X_2)$ with independent standard normally distributed random variables X_1 and X_2 and $m(x_1, x_2) = 2 \cdot x_1 + x_2 + 2$. Again, we choose $h_n \in \{1, 1.25, 1.5, 2\}$ and $h_N \in \{0.6, 0.8, 1\}$. But instead of the smoothing spline estimate we use $m_n(x_1, x_2) = x_1^2$ as an estimate for the linear function m. Figure 8 shows the boxplots of the occuring L_1–errors.

Here, the estimates which use the additional data are the worst. The data-driven method chooses in every of the 100 simulation the weight $w = 1$ as illustrated in the following tabular:

	$w = 0$	$w = 1$	$w = \frac{l_n}{l_n+N}$
example 1	32	45	23
example 2	42	11	47
example 3	31	14	55
example 4	0	100	0

Finally, we illustrate the usefulness of our estimation procedure by applying it to the density estimation problem of the Collaborative Research Centre 805. Here we consider

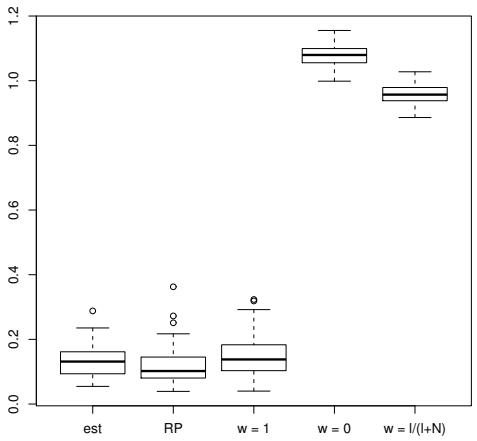

Figure 8: Boxplot of the occuring L_1-errors in the forth model

the load distribution in the three legs of a simple tripod and we assume that the diameters behave like independent 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. For simplicity, we consider only one leg of the tripod. Since in this case the real density is unknown, we repeat the simulation 10.000 times to generate a high sample of relative loads. Application of the routine *density* in the statistics package R to these 10.000 observed values leads to the solid line in Figure 9. We calculate our estimates as described before assuming that $n = 200$ measurements are available. Again, the newly proposed estimate is printed by the dashed line, and the dotted line represents the estimate of Rosenblatt and Parzen. Similarly as before, the run of the curve of our estimate lies much closer to the solid line than the one of Rosenblatt and Parzen.

5 Proofs of the main results

5.1 Proof of Theorem 1

By the Lemma of Scheffe we have

$$\begin{aligned}
&\mathbf{E} \int |g_n(y) - f(y)|\, dy \\
&= 2 \cdot \mathbf{E} \int_B (f(y) - g_N(y))_+\, dy \\
&\leq 2 \cdot \mathbf{E} \int_B |g_N(y) - \mathbf{E}\{g_N(y)|\mathcal{D}_n\}|\, dy + 2 \cdot \mathbf{E} \int_B (f(y) - \mathbf{E}\{g_N(y)|\mathcal{D}_n\})_+\, dy,
\end{aligned}$$

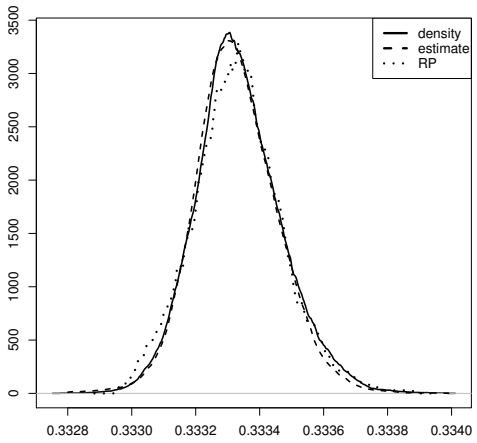

Figure 9: Density estimation in a simulation model

where the last inequality follows from

$$a_+ \leq |b| + (a-b)_+ \quad \text{for } a, b \in \mathbb{R}.$$

As in the proof of Theorem 1 in Devroye, Felber and Kohler (2013) we can bound the first term using the Cauchy-Schwarz inequality and the inequality of Jensen by arguing

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

Next we use the theorem of Fubini and the conditional independence of $m_n(X_{n+1}), \ldots, m_n(X_{n+N})$ and get

$$\begin{aligned}
&\mathbf{E}\left\{\int_B \left|g_N(y) - \mathbf{E}\left\{g_N(y)\middle|\mathcal{D}_n\right\}\right|^2 \, dy \middle| \mathcal{D}_n\right\} \\
&= \int_B \mathbf{E}\left\{\left|g_N(y) - \mathbf{E}\left\{g_N(y)\middle|\mathcal{D}_n\right\}\right|^2 \middle| \mathcal{D}_n\right\} \, dy \\
&\leq \int_B \frac{1}{N^2 \cdot h_N^2} \cdot \sum_{i=1}^{N} \mathbf{E}\left\{K^2\left(\frac{y - m_n(X_{n+i})}{h_N}\right) \middle| \mathcal{D}_n\right\} \, dy \\
&= \frac{1}{N \cdot h_N^2} \cdot \int_B \int K^2\left(\frac{y - m_n(z)}{h_N}\right) \mathbf{P}_X(dz) \, dy
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N \cdot h_N^2} \cdot \int \int_B K^2 \left(\frac{y - m_n(z)}{h_N} \right) \, dy \, \mathbf{P}_X(dz) \\
&\leq \frac{1}{N \cdot h_N} \cdot \int \int_{\mathbb{R}} K^2 \left(y \right) \, dy \, \mathbf{P}_X(dz) \\
&= \frac{1}{N \cdot h_N} \cdot \int_{\mathbb{R}} K^2 \left(y \right) \, dy.
\end{aligned}
$$

From this we conclude

$$
\mathbf{E} \int_B |g_N(y) - \mathbf{E}\{g_N(y)|\mathcal{D}_n\}| \, dy \leq \frac{c_1}{\sqrt{N \cdot h_N}},
$$

hence it suffices to show

$$
\mathbf{E} \int_B (f(y) - \mathbf{E}\{g_N(y)|\mathcal{D}_n\})_+ \, dy \leq c_8 \cdot h_N^r + \frac{\mathbf{E}\left\{\min\{|m_n(X) - m(X)|\}, 2 \cdot h_n\right\}}{h_N}. \quad (6)
$$

By triangle inequality we have

$$
\begin{aligned}
&\mathbf{E} \int_B (f(y) - \mathbf{E}\{g_N(y)|\mathcal{D}_n\})_+ \, dy \\
&\leq \int_B \left| f(y) - \int \frac{1}{h_N} \cdot K \left(\frac{y - m(x)}{h_N} \right) \mathbf{P}_X(dx) \right| \, dy \\
&\quad + \mathbf{E} \int_B \left| \int \frac{1}{h_N} \cdot K \left(\frac{y - m(x)}{h_N} \right) \mathbf{P}_X(dx) - \int \frac{1}{h_N} \cdot K \left(\frac{y - m_n(x)}{h_N} \right) \mathbf{P}_X(dx) \right| \, dy \\
&:= T_{1,n} + T_{2,n}
\end{aligned}
$$

Using that f is the density of $m(X)$, that K is the naive kernel and that f is Hölder continuous we get

$$
\begin{aligned}
T_{1,n} &= \int_B \left| f(y) - \int \frac{1}{h_N} \cdot K \left(\frac{y - x}{h_N} \right) \cdot f(x) \, dx \right| \, dy \\
&\leq \int_B \int \frac{1}{h_N} \cdot K \left(\frac{y - x}{h_N} \right) \cdot |f(y) - f(x)| \, dx dy \\
&\leq \int_B \int \frac{1}{h_N} \cdot K \left(\frac{y - x}{h_N} \right) \cdot C \cdot h_N^r \, dx dy \\
&= C \cdot h_N^r \cdot \int_B 1 \, dy \\
&\leq c_8 \cdot h_N^r.
\end{aligned}
$$

Application of the theorem of Fubini yields

$$
T_{2,n} \leq \frac{1}{h_N} \cdot \mathbf{E} \int \int \left| K \left(\frac{y - m(x)}{h_N} \right) - K \left(\frac{y - m_n(x)}{h_N} \right) \right| \, dy \, \mathbf{P}_X(dx).
$$

An elementary calculation shows that we have for arbitrary $z_1, z_2 \in \mathbb{R}$

$$
\int \left| K \left(\frac{y - z_1}{h_N} \right) - K \left(\frac{y - z_2}{h_N} \right) \right| \, dy
$$

$$= \frac{1}{2} \cdot \int \left| 1_{[z_1-h_N, z_1+h_N]}(y) - 1_{[z_2-h_N, z_2+h_N]}(y) \right| \, dy$$
$$\leq \min\{2 \cdot h_N, |z_1 - z_2|\},$$

which implies

$$\begin{aligned} T_{2,n} &\leq \frac{1}{h_N} \cdot \mathbf{E} \int \min\{2 \cdot h_N, |m_n(x) - m(x)|\} \, \mathbf{P}_X(dx) \\ &= \frac{1}{h_N} \cdot \mathbf{E} \min\{2 \cdot h_N, |m_n(X) - m(X)|\}. \end{aligned}$$

Summarizing the above results we get (6), which in turn implies the assertion. □

5.2 Proof of Theorem 4

Application of Theorem 10.1 in Devroye and Lugosi (2001) yields

$$\int |\hat{f}_n(y) - f(y)| \, dy \leq 3 \cdot \min_{\theta \in \Theta} \int |\hat{f}_{n-l_n,\theta}(y) - f(y)| \, dy + 4 \cdot \Delta + \frac{3}{n}$$

where

$$\Delta = \sup_{A \in \mathcal{A}} \left| \int_A f(y) \, dy - \hat{\mu}_{l_n}(A) \right|.$$

By the well-known results from Vapnik-Chervonenkis theory (cf., e.g., Theorem 3.1 in Devroye and Lugosi (2001)) we get

$$\mathbf{E}\Delta \leq 2 \cdot \sqrt{\frac{\log s_{\mathcal{A}}(l_n)}{l_n}},$$

where $s_{\mathcal{A}}(l)$ is the l-th shatter coefficient of $\mathcal{A}$ defined as the maximal number of subsets of l points which can be picked out by $\mathcal{A}$, i.e.,

$$s_{\mathcal{A}}(l) = \max_{y_1,\ldots,y_l \in \mathbb{R}} |\{\{y_1, \ldots, y_l\} \cap A \quad : \quad A \in \mathcal{A}\}|.$$

In the sequel we will modify the proof of Lemma 11.1 in Devroye and Lugosi (2001) in order to show

$$s_{\mathcal{A}}(l) \leq \left(1 + (l \cdot (n - l_n + 1))^2 \cdot (l \cdot (N+1))^2\right)^4, \tag{7}$$

which implies the assertion.

In order to prove (7) we have to count the number of subsets of $\{y_1, \ldots, y_l\}$ which can be picked out by sets of the form

$$\left\{ y \quad : \quad w \cdot \frac{1}{n \cdot h_1} \cdot \sum_{i=l_n+1}^{n} K\left(\frac{y - m(X_i)}{h_1}\right) \right.$$
$$\left. + (1-w) \cdot \frac{1}{N \cdot h_2} \cdot \sum_{i=1}^{N} K\left(\frac{y - m_{n-l_n}(X_{n+i})}{h_2}\right) \right.$$

$$> \bar{w} \cdot \frac{1}{n \cdot \bar{h}_1} \cdot \sum_{i=l_n+1}^{n} K\left(\frac{y - m(X_i)}{\bar{h}_1}\right)$$
$$+(1-\bar{w}) \cdot \frac{1}{N \cdot \bar{h}_2} \cdot \sum_{i=1}^{N} K\left(\frac{y - m_{n-l_n}(X_{n+i})}{\bar{h}_2}\right) \Bigg\}$$

for arbitrary (h_1, h_2, w), $(\bar{h}_,, \bar{h}_2, \bar{w}) \in \Theta$. This number of sets is upper bounded by the number of subsets of $\{y_1, \ldots, y_l\}$ which can be picked out by sets of the form

$$\Bigg\{ y \quad : \quad c_9 \cdot \sum_{i=l_n+1}^{n} K\left(\frac{y - m(X_i)}{h_1}\right) + c_{10} \cdot \sum_{i=1}^{N} K\left(\frac{y - m_{n-l_n}(X_{n+i})}{h_2}\right)$$
$$> c_{11} \cdot \sum_{i=l_n+1}^{n} K\left(\frac{y - m(X_i)}{\bar{h}_1}\right) + c_{12} \cdot \sum_{i=1}^{N} K\left(\frac{y - m_{n-l_n}(X_{n+i})}{\bar{h}_2}\right) \Bigg\} \quad (8)$$

for arbitrary $c_9, c_{10}, c_{11}, c_{12} \in \mathbb{R}$ and $h_1, h_2, \bar{h}_1, \bar{h}_2 > 0$, which we bound in the sequel.

We start by counting the numbers of vectors of the form

$$\Bigg(\sum_{i=l_n+1}^{n} K\left(\frac{y_j - m(X_i)}{h_1}\right), \sum_{i=1}^{N} K\left(\frac{y_j - m_{n-l_n}(X_{n+i})}{h_2}\right),$$
$$\sum_{i=l_n+1}^{n} K\left(\frac{y_j - m(X_i)}{\bar{h}_1}\right), \sum_{i=1}^{N} K\left(\frac{y_j - m_{n-l_n}(X_{n+i})}{\bar{h}_2}\right) \Bigg)$$

for arbitrary $h_1, h_2, \bar{h}_1, \bar{h}_2 > 0, j = 1, \ldots, l$. Since K is the naive kernel the components of the above vector take on at most $n - l_n + 1$ and $N + 1$ different values, respectively. Consequently the number of different vectors does not exceed $(n - l_n + 1)^2 \cdot (N+1)^2 = L_{n,N}$. Let

$$(z_{1,1}, \ldots, z_{1,4}), \ldots, (z_{L_{n,N},1}, \ldots, z_{L_{n,N},4})$$

be all possible vector values of the above vector. If

$$1_{\{c_9 \cdot \sum_{i=l_n+1}^{n} K\left(\frac{y_j - m(X_i)}{h_1}\right) + c_{10} \cdot \sum_{i=1}^{N} K\left(\frac{y_j - m_{n-l_n}(X_{n+i})}{h_2}\right)}$$
$$> c_{11} \cdot \sum_{i=l_n+1}^{n} K\left(\frac{y_j - m(X_i)}{\bar{h}_1}\right) + c_{12} \cdot \sum_{i=1}^{N} K\left(\frac{y_j - m_{n-l_n}(X_{n+i})}{\bar{h}_2}\right)\}$$
$$\neq 1_{\{\bar{c}_9 \cdot \sum_{i=l_n+1}^{n} K\left(\frac{y_j - m(X_i)}{h_1}\right) + \bar{c}_{10} \cdot \sum_{i=1}^{N} K\left(\frac{y_j - m_{n-l_n}(X_{n+i})}{h_2}\right)}$$
$$> \bar{c}_{11} \cdot \sum_{i=l_n+1}^{n} K\left(\frac{y_j - m(X_i)}{\bar{h}_1}\right) + \bar{c}_{12} \cdot \sum_{i=1}^{N} K\left(\frac{y_j - m_{n-l_n}(X_{n+i})}{\bar{h}_2}\right)\}$$

for some $j = 1 \ldots, l$ then

$$1_{\{c_9 \cdot z_{k,1} + c_{10} \cdot z_{k,2} > c_{11} \cdot z_{k,3} + c_{12} \cdot z_{k,4}\}} \neq 1_{\{\bar{c}_9 \cdot z_{k,1} + \bar{c}_{10} \cdot z_{k,2} > \bar{c}_{11} \cdot z_{k,3} + \bar{c}_{12} \cdot z_{k,4}\}}$$

for some $k = 1 \ldots, L_{n,N}$. Consequently the number of subsets of $\{y_1, \ldots, y_l\}$ which can be picked out by the sets of the form (8) is bounded by $L_{n,N}$-th shatter coefficient of the set

$$\left\{\{(z_1, z_2, z_3, z_4) \in \mathbb{R}^4 \quad : \quad c_9 \cdot z_1 + c_{10} \cdot z_2 > c_{11} \cdot z_3 + c_{12} \cdot z_4 \quad : \quad c_9, c_{10}, c_{11}, c_{12} \in \mathbb{R}\right\}.$$

By Theorem 9.3 and Theorem 9.5 in Györfi et al. (2002) this shatter coefficient is bounded by

$$(1 + L_{n,N})^4,$$

which implies the assertion. □

6 Acknowledgment

The authors would like to thank two anonymous referees and the Associate Editor for valuable comments, which helped to improve the paper. Furthermore the authors would like to thank the German Research Foundation (DFG) for funding this project within the Collaborative Research Centre 805 and acknowledge research support from Natural Sciences and Engineering Research Council of Canada.

References

[1] Beirlant, J. and Györfi, L. (1998). On the asymptotic L_2-error in partitioning regression estimation. *Journal of Statistical Planning and Inference*, **71**, pp. 93–107.

[2] Devroye, L. (1982). Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **61**, pp. 467–481.

[3] Devroye, L. (1983). The equivalence in L1 of weak, strong and complete convergence of kernel density estimates. *Annals of Statistics*, **11**, pp. 896–904.

[4] Devroye, L. (1987). *A Course in Density Estimation.* Birkhäuser, Basel.

[5] Devroye, L., Felber, T., and Kohler, M. (2013). Estimation of a density using real and artificial data. *IEEE Transactions on Information Theory*, **59**, No. 3, pp. 1917-1928.

[6] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation. The L1 view.* Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. John Wiley and Sons, New York.

[7] Devroye, L., Györfi, L., Krzyżak, A., and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, **22**, pp. 1371–1385.

[8] Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation.* Springer-Verlag, New York.

[9] Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, **23**, pp. 71–82.

[10] Devroye, L. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, **8**, pp. 231–239.

[11] Eggermont, P. P. B. and LaRiccia, V. N. (2001). *Maximum Penalized Likelihood Estimation. Volume I: Density Estimation.* Springer-Verlag, New York.

[12] Györfi, L. (1981). Recent results on nonparametric regression estimate and multiple classification. *Problems of Control and Information Theory*, **10**, pp. 43–52.

[13] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics, Springer-Verlag, New York.

[14] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer-Verlag, New York.

[15] Kohler, M. (2000). Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *Journal of Statistical Planning and Inference*, **89**, pp. 1–23.

[16] Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Transactions on Information Theory*, **47**, pp. 3054–3058.

[17] Kohler, M. and Krzyżak, A. (2013). Optimal global rates of convergence for interpolation problems with random design, Statistics and Probability Letters, **83**, pp. 1871-1879.

[18] Liitiäinen, E., Corona, F., Lendasse, A. (2010). Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, **101**, pp. 811-823.

[19] Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, **41**, pp. 677–687.

[20] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, **9**, pp. 141–142.

[21] Nadaraya, E. A. (1970). Remarks on nonparametric estimates for density functions and regression curves. *Theory of Probability and its Applications*, **15**, pp. 134–137.

[22] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, **33**, pp. 1065–1076.

[23] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, pp. 832–837.

[24] Scott, D. W. (1982). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley Series in Probability and Statistics, Wiley, New York.

[25] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statististics*, **5**, pp. 595–645.

[26] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, **10**, pp. 1040–1053.

[27] Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation.* Springer Series in Statistics, Springer-Verlag, New York.

[28] Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

[29] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Chapman and Hall, London.

[30] Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, **26**, pp. 359–372.