Ann Inst Stat Math (2017) 69:189–214
DOI 10.1007/s10463-015-0535-8

Nonparametric estimation of a conditional density

Ann-Kathrin Bott[1] · Michael Kohler[1]

Received: 2 October 2014 / Revised: 9 June 2015 / Published online: 4 August 2015

Abstract In this paper, we estimate a conditional density. In contrast to standard results in the literature in this context we assume that for each observed value of the covariate we observe a sample of the corresponding conditional distribution of size larger than one. A density estimate is defined taking into account the data from all the samples by computing a weighted average using weights depending on the covariates. The error of the density estimate is measured by the L_1-error. Results concerning consistency and rate of convergence of the estimate are presented, and the performance of the estimate for finite sample size is illustrated using simulated data. Furthermore, the estimate is applied to a problem in fatigue analysis.

Keywords Conditional density estimation · L_1-error · Consistency · Rate of convergence

1 Introduction

A well-known problem in the literature is the problem of density estimation. Given an independent sample $Y_1, \ldots, Y_n$ of an $\mathbb{R}^d$-valued random variable Y, the goal is to estimate the density f of the distribution of Y, which is assumed to exist. This can be done, e.g., by the famous Rosenblatt–Parzen kernel density estimate (cf., Rosenblatt 1956; Parzen 1962), defined by

✉ Ann-Kathrin Bott
abott@mathematik.tu-darmstadt.de

Michael Kohler
kohler@mathematik.tu-darmstadt.de

[1] Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7,
64289 Darmstadt, Germany

$$f_n(x) = \frac{1}{n \cdot h_n^d} \cdot \sum_{k=1}^{n} K\left(\frac{x - Y_k}{h_n}\right). \tag{1}$$

Here $h_n > 0$ is the so-called bandwidth and the kernel $K : \mathbb{R}^d \to \mathbb{R}$, e.g., the naive kernel $K(u) = 1/2^d \cdot \mathbb{1}_{[-1,1]^d}(u)$, is a density. This density estimate can be used, e.g., to estimate all probabilities of the underlying distribution, and provided we control the L_1-error of the density estimate 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. It is well known that there exist estimates which are L_1-consistent for all densities, e.g., the above kernel density estimate is L_1-consistent for all densities provided

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

see Devroye (1983). Further results on density estimation can be found in the books (Devroye and Györfi 1985; Devroye 1987; Devroye and Lugosi 2001).

In applications sometimes the sample size is rather small, e.g., in case that a data point corresponds to a rather time-expensive experiment. A concrete application in connection with fatigue analysis, where this effect occurs, is described in Sect. 3 below. This motivates to try to combine several data sets from different (but somehow related) density estimation problems to estimate a general density depending on a covariate.

We do this in the context of conditional density estimation. Usually it is assumed that a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of an $\mathbb{R}^d \times \mathbb{R}$-valued random vector (X, Y) is available. Already Rosenblatt (1969) introduced an estimator of a conditional density. This estimator and many others are motivated by the definition of a conditional density. Let $g_{(X,Y)}(x, y)$ be the joint density of (X, Y) and $g_X(x)$ the marginal density of X. Then the conditional density $g_{Y|X}(y, x)$ of Y given X is given by

$$g_{Y|X}(x, y) = \frac{g_{(X,Y)}(x, y)}{g_X(x)}.$$

Replacing the joint and marginal density by density estimates we obtain an estimator of the conditional density. To estimate the marginal density of X we can directly apply the Rosenblatt–Parzen kernel density estimate (1) with density K and bandwidth $H_n > 0$. Using the product kernel estimator (c.f., e.g., Rosenblatt 1969; Scott 1992; Hyndman et al. 1996) the estimator for the joint density is given by

$$\hat{g}_{(X,Y)}(x, y) = \frac{1}{n \cdot H_n^d \cdot h_n} \sum_{i=1}^{n} K\left(\frac{\|x - X_i\|}{H_n}\right) \cdot K\left(\frac{y - Y_i}{h_n}\right)$$

where $K : \mathbb{R} \to \mathbb{R}_+$ is a density and h_n, $H_n > 0$ are bandwidths. Hence, we can estimate the conditional density by

$$\hat{g}_{Y|X}(y,x) = \frac{\sum_{i=1}^{n} K\left(\frac{\|x-X_i\|}{H_n}\right) \cdot K\left(\frac{y-Y_i}{h_n}\right)}{h_n \sum_{j=1}^{n} K\left(\frac{\|x-X_j\|}{H_n}\right)}. \tag{2}$$

This conditional density estimation problem can also be seen as a nonparametric regression problem. It is well known that

$$\mathbf{E}\left\{\frac{1}{h_n} \cdot K\left(\frac{y-Y}{h_n}\right) \middle| X = x\right\} \to g_{Y|X}(y,x) \quad (n \to \infty)$$

for Lebesgue almost all y and $\mathbf{P}_X$-almost all x (c.f., e.g., Fan et al. 1996). Thus, the estimator (2) can be seen as a kernel regression estimate (cf., e.g., Chapter 5 in Györfi et al. 2002) applied to

$$\left(X_1, \frac{1}{h_n} \cdot K\left(\frac{y-Y_1}{h_n}\right)\right), \ldots, \left(X_n, \frac{1}{h_n} \cdot K\left(\frac{y-Y_n}{h_n}\right)\right),$$

c.f., e.g., Fan and Yim (2004) and Gooijer and Zerom (2003).

Instead of the above kernel density estimate of the conditional density one can also define a partitioning estimate of the conditional density. Results concerning universal consistency and rate of convergence of the L_1-error of such a partitioning estimate have been derived in Györfi and Kohler (2007). Sharp minimax bounds on the L_2-errors of conditional density estimates are presented in Efromovich (2007).

In the sequel we assume that for each covariate X_i $(i \in \{1, \ldots, N_n\})$ we have given not only one observation of the value of Y_i, but instead a whole sample

$$\mathcal{D}_n^{(i)} = \left\{Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_{l_{i,n}}^{(i)}\right\}$$

of size $l_{i,n} \in \mathbb{N}$. Here we assume that for given X_i the data points in $\mathcal{D}_n^{(i)}$ are (conditionally) independent and identically distributed as Y_i, and that all data sets $\mathcal{D}_n^{(i)}$ $(i = 1, \ldots, N_n)$ are independent. For each of these data samples we can estimate the conditional density of Y_i given X_i by

$$\hat{f}_n(y, X_i) = \frac{1}{l_{i,n} \cdot h_n} \cdot \sum_{k=1}^{l_{i,n}} K\left(\frac{y - Y_k^{(i)}}{h_n}\right), \tag{3}$$

where K is a density. Since the amount of data $l_{i,n}$ is decisive for the quality of the above-defined density estimators $\hat{f}_n(\cdot, X_i)$, we will use a local average of kernel density estimates with an additional weighting through the amount of data corresponding to the different densities, and define our estimate via

$$f_n(y,x) = \frac{\sum_{i=1}^{N_n} l_{i,n} \cdot G\left(\frac{\|x-X_i\|}{H_n}\right) \hat{f}_n(y, X_i)}{\sum_{j=1}^{N_n} l_{j,n} \cdot G\left(\frac{\|x-X_j\|}{H_n}\right)}$$
$$= \frac{\sum_{i=1}^{N_n} G\left(\frac{\|x-X_i\|}{H_n}\right) \sum_{k=1}^{l_{i,n}} K\left(\frac{y-Y_k^{(i)}}{h_n}\right)}{h_n \sum_{j=1}^{N_n} l_{j,n} \cdot G\left(\frac{\|x-X_j\|}{H_n}\right)}, \tag{4}$$

where $G = \mathbb{1}_{[0,1]}$.

The estimate (4) is the estimate which we get from (2) if we use there the data

$$\left\{(X_i, Y_k^{(i)}) : i = 1, \dots, N_n, k = 1, \dots, l_{n,i}\right\}. \tag{5}$$

Here the data (5) can in general not be considered as a sample of a random vector (X, Y), since $l_{n,i}$ are deterministic numbers and X might have a density with respect to the Lebesgue–Borel measure.

We measure the quality of our estimate by the average L_1-error

$$\int\int |f_n(y,x) - f(y,x)| \, \mathrm{d}y \, P_X(\mathrm{d}x),$$

which is (via the Lemma of Scheffé) directly linked to the total variation error of the corresponding distribution estimate. We derive sufficient conditions for the L_1-consistency of our estimates and we investigate the rate of convergence of the expected average L_1-error in case of smooth densities. Motivated by an application in fatigue analysis described below we extend all of the above results to the case that the data points $Y_k^{(i)}$ can be observed only with additional measurement errors, which do not need to be independent or have expectation zero, but vanish on average asymptotically. The finite sample size performance of our estimates is illustrated using simulated data.

Throughout the paper the following notation is used: the sets of natural numbers, integers, real numbers and positive real numbers including zero are denoted by $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$ and $\mathbb{R}_+$, resp. $\mathcal{B}$ denotes the set of all Borel sets in $\mathbb{R}$ and $\mathbb{1}_B$ denotes the indicator function of the set B. $\|x\|$ is the Euclidean norm of a vector $x \in \mathbb{R}^d$. The support of a probability measure μ defined on the Borel sets in $\mathbb{R}^d$ is abbreviated by

$$\operatorname{supp}(\mu) = \left\{x \in \mathbb{R}^d \ : \ \mu(S_r(x)) > 0 \quad \text{for all } \ r > 0\right\},$$

where $S_r(x)$ is the ball of radius r around x.

The outline of this paper is as follows: the main results are presented in Sect. 2 and proven in Sect. 4. Section 3 illustrates the finite sample size behavior of our estimate by applying it to simulated data and to a problem in fatigue analysis.

2 Main results

Let (X, Y) be an $\mathbb{R}^d \times \mathbb{R}$-valued random vector such that the conditional distribution of Y given $X = x$ has a density $f(\cdot, x) : \mathbb{R} \to \mathbb{R}_+$ (with respect to the Lebesgue measure). In the sequel we assume that f is a (measurable) real-valued function defined on $\mathbb{R} \times \mathbb{R}^d$. Consequently, since $f(\cdot, x)$ is a density for all $x \in \mathbb{R}^d$, we have

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}} f(y, x)\,\mathrm{d}y\,\mathbf{P}_X(\mathrm{d}x) = 1 < \infty.$$

Let (X, Y), (X_1, Y_1), $\ldots(X_{N_n}, Y_{N_n})$ be independent and identically distributed. We assume that for each $i \in \{1, \ldots, N_n\}$ we observe X_i and a conditionally independent sample

$$\mathcal{D}_n^{(i)} = \left\{Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_{l_{i,n}}^{(i)}\right\}$$

of Y_i. Furthermore we assume that all data sets $\mathcal{D}_n^{(i)}$ are independent. The estimator of $f(y, x)$ is given by

$$f_n(y, x) = \frac{\sum_{i=1}^{N_n} G\left(\frac{\|x-X_i\|}{H_n}\right) \sum_{k=1}^{l_{i,n}} K\left(\frac{y-Y_k^{(i)}}{h_n}\right)}{h_n \sum_{j=1}^{N_n} l_{j,n} \cdot G\left(\frac{\|x-X_j\|}{H_n}\right)}$$

with $G = \mathbb{1}_{[0,1]}$.

Theorem 1 *Assume that* $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}_+$ *is measurable and satisfies*

$$\int_{\mathbb{R}} f(y, x)\mathrm{d}y = 1 \quad (x \in \mathbb{R}^d).$$

Let f_n *be the above-defined density estimate of* f *and let the kernel* $K : \mathbb{R} \to \mathbb{R}_+$ *be a square integrable density. Then*

(A1) $h_n \to 0,\ H_n \to 0, N_n \cdot H_n^d \to \infty \quad (n \to \infty)$,
(A2) $N_n \cdot H_n^d \cdot h_n \cdot \min_{1 \le i \le N_n} l_{i,n} \to \infty \quad (n \to \infty)$
and
(A3) $\limsup_{n\to\infty} \frac{\max_{i=1,\ldots,N_n} l_{i,n}}{\min_{i=1,\ldots,N_n} l_{i,n}} < \infty$

imply

$$\mathbf{E} \int \int |f_n(y, x) - f(y, x)|\,\mathrm{d}y\,P_X(\mathrm{d}x) \to 0 \quad (n \to \infty).$$

Remark 1 In the case $l_{n,i} = 1$ for all i the estimate (4) is equal to the standard kernel density estimate (2) of a conditional density and Theorem 1 implies that this estimate is weakly universally L_1 consistent in case that K is the naive kernel and that the bandwidths satisfy

$$h_n \to 0 \quad (n \to \infty),\ H_n \to 0 \quad (n \to \infty) \quad \text{and} \quad n \cdot H_n^d \cdot h_n \to \infty \quad (n \to \infty).$$

Remark 2 The conditions in $(A1)$ are typical conditions on the bandwidth that are needed to assure consistency of kernel regression and kernel density estimates. The condition $(A2)$ is weaker than the condition

$$h_n \cdot \min_{1 \le i \le N_n} l_{i,n} \to \infty$$

that is needed (besides $h_n \to 0$) to guarantee that all inserted density estimates are consistent.

Remark 3 In Theorem 1 we require that the kernel K is a square integrable density. We conjecture that using arguments from the proof of Theorem 9.2 in Devroye and Lugosi (2001) it should be possible to prove Theorem 1 also for kernels $K : \mathbb{R} \to \mathbb{R}$ which satisfy only

$$\int_{\mathbb{R}} |K(x)|\, \mathrm{d}x < \infty \quad \text{and} \quad \int_{\mathbb{R}} K(x)\, \mathrm{d}x = 1.$$

Corollary 1 *In addition to the assumptions of Theorem 1 we assume that the same amount of data is given for each covariate, i.e., $l_{1,n} = l_{2,n} = \cdots = l_{N_n,n} =: l_n$. Then* (A1) *and*

(A2′) $N_n \cdot H_n^d \cdot l_n \cdot h_n \to \infty \quad (n \to \infty)$

imply

$$\mathbf{E} \int \int |f_n(y,x) - f(y,x)|\, \mathrm{d}y\, P_X(\mathrm{d}x) \to 0 \quad (n \to \infty).$$

Proof Due to the additional assumption we have

$$\min_{1 \le i \le N_n} l_{i,n} = \max_{1 \le i \le N_n} l_{i,n} = l_n$$

and hence condition $(A2)$ of Theorem 1 simplifies to $(A2')$ and condition $(A3)$ trivially holds. □

Next we analyze the rate of convergence of our estimate.

Theorem 2 *Assume that $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}_+$ satisfies*

$$\int_{\mathbb{R}} |f(y,x_1) - f(y,x_2)|\, \mathrm{d}y \le c_1 \cdot \|x_1 - x_2\|^{\alpha} \quad (x_1, x_2 \in \mathbb{R}^d) \tag{6}$$

for some $c_1 > 0$, $\alpha \in (0,1]$. Let f_n be the above-defined density estimate of f with $l_{1,n} = l_{2,n} = \cdots = l_{N_n,n} =: l_n$. Then the conditions

(A4) *The densities* $f(\cdot, x)$ $(x \in \mathbb{R}^d)$ *are Hölder-continuous with exponent* $r \in (0, 1]$, *i.e.,*

$$|f(u, x) - f(v, x)| \leq c_2 \cdot |u - v|^r \quad \textit{for all } u, v \in \mathbb{R},\ x \in \mathbb{R}^d \textit{ and some } c_2 > 0,$$

(A5) *There exists a compact set* $B \in \mathcal{B}$ *such that*

$$f(y, x) = 0 \quad \textit{for all } y \notin B \textit{ and } P_X\textit{-almost all } x \in \mathbb{R}^d,$$

(A6) $C = supp(P_X)$ *is compact*

and

(A7) K *is a density satisfying*

$$\int_{\mathbb{R}} K^2(u)\, \mathrm{d}u < \infty \quad \textit{and} \quad \int_{\mathbb{R}} K(u) \cdot |u|^r\, \mathrm{d}u < \infty$$

imply that there exist constants c_3, c_4, c_5 *and* $c_6 > 0$ *such that for all* $n \in \mathbb{N}$ *we have*

$$\begin{aligned} &\mathbf{E} \int\int |f_n(y, x) - f(y, x)|\, \mathrm{d}y\, P_X(\mathrm{d}x) \\ &\leq \frac{c_3}{\sqrt{N_n \cdot l_n \cdot h_n \cdot H_n^d}} + c_4 \cdot H_n^\alpha + c_5 \cdot h_n^r + \frac{c_6}{N_n \cdot H_n^d}. \end{aligned} \tag{7}$$

Remark 4 Condition (6) and (A4) are, e.g., satisfied, if we set $\alpha = r = 1$ and choose X as uniformly distributed on $[-0.5, 0.5]$ and define Y for given $X = x$ as normally distributed with mean x and variance one. In this case (6) follows from Lemma 1 in Bott et al. (2013).

Remark 5 In Theorem 2 we impose the two smoothness assumptions (6) and (A4) on f. Here (6) is used to derive a rate of convergence result for the kernel regression estimates applied to the different density estimates, and (A4) is needed to bound rate of convergence of the individual density estimates which are averaged in our kernel regression estimate. Both for kernel regression estimation and kernel density estimation similar smoothness assumptions are known to be necessary to derive nontrivial rate of convergence results, however, it is an open problem whether the above smoothness assumptions are really necessary for the above rate of convergence result.

To determine the optimal rate of convergence in Theorem 2 we have to choose the bandwidths h_n and H_n such that the right-hand side of (7) is minimal. We refer to Sect. 4 for details concerning this minimization. This leads to the following result.

Corollary 2 *Assume that the assumptions of Theorem 2 hold.*

(i) *In case that* $N_n^{-\alpha(r+1)} \leq l_n^{-r(\alpha+d)}$ *we set* $h_n = c_7 \cdot (N_n \cdot l_n)^{-\frac{\alpha}{(2\cdot\alpha+d)r+\alpha}}$ *and* $H_n = c_8 \cdot (N_n \cdot l_n)^{-\frac{r}{(2\cdot\alpha+d)r+\alpha}}$. *For suitable chosen* $c_9 > 0$ *it holds*

$$\mathbf{E} \int\int |f_n(y, x) - f(y, x)|\, \mathrm{d}y\, P_X(\mathrm{d}x) \leq c_9 \cdot (N_n \cdot l_n)^{-\frac{\alpha\cdot r}{(2\cdot\alpha+d)r+\alpha}}.$$

(ii) *In case that* $N_n^{-\alpha(r+1)} > l_n^{-r(\alpha+d)}$ *we set* $h_n = c_{10} \cdot N_n^{-\frac{\alpha}{(\alpha+d)(2\cdot r+1)}} \cdot l_n^{-\frac{1}{2\cdot r+1}}$ *and* $H_n = c_{11} \cdot N_n^{-\frac{1}{\alpha+d}}$. *For suitable chosen* $c_{12} > 0$ *it holds*

$$\mathbf{E} \int \int |f_n(y, x) - f(y, x)| \, \mathrm{d}y \, P_X(\mathrm{d}x) \leq c_{12} \cdot N_n^{-\frac{\alpha}{\alpha+d}}.$$

Proof The assertion follows directly from Theorem 2 and the definitions of h_n and H_n. □

Remark 6 Corollary 2 implies the surprising fact that up to some threshold, namely as long as $l_n^{r(\alpha+d)} > N_n^{\alpha(r+1)}$ holds, increasing l_n from one to some greater value does not improve the rate of convergence of our estimate, because our upper bound on the error in (ii) depends only on N_n (i.e., on how many x-values we observe) and not on l_n (i.e., not on the number of y-values we sample for each x-value). However, after this threshold, i.e., as soon as $l_n^{r(\alpha+d)} \leq N_n^{\alpha(r+1)}$, the upper bound becomes a function of $N_n \cdot l_n$ and is hence directly influenced by the value of l_n.

Remark 7 In case of $l_n = 1$, $N_n = n$ and $\alpha = r = 1$ Corollary 2 states that our estimate achieves a rate of convergence of $\mathcal{O}(n^{-\frac{1}{d+3}})$. Györfi and Kohler (2007) obtained the same rate for a partitioning estimate.

We now assume that we observe only data

$$\bar{\mathcal{D}}_n^{(i)} = \left\{ \bar{Y}_1^{(i)}, \bar{Y}_2^{(i)}, \ldots, \bar{Y}_{l_{i,n}}^{(i)} \right\} \quad (i = 1, \ldots, N_n)$$

with additional measurement errors. An application, where this is indeed the case, is described in Sect. 3. The data $\bar{Y}_1^{(i)}, \bar{Y}_2^{(i)}, \ldots, \bar{Y}_{l_{i,n}}^{(i)}$ $(i \in \{1, \ldots, N_n\})$ do not need to be conditionally independent or identically distributed. We assume that the measurement errors are "small" and for this reason we ignore them completely. Consequently we define our density estimate $\bar{f}_n$ of f as

$$\bar{f}_n(y, x) = \frac{\sum_{i=1}^{N_n} G\left(\frac{\|x - X_i\|}{H_n}\right) \sum_{k=1}^{l_{i,n}} K\left(\frac{y - \bar{Y}_k^{(i)}}{h_n}\right)}{h_n \cdot \sum_{j=1}^{N_n} l_{j,n} \cdot G\left(\frac{\|x - X_j\|}{H_n}\right)}.$$

In the following theorem we show that under appropriate assumptions our density estimate remains L_1-consistent.

Theorem 3 *Assume that* $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}_+$ *is measurable and satisfies*

$$\int_{\mathbb{R}} f(y, x) \mathrm{d}y = 1 \quad (x \in \mathbb{R}^d).$$

Let $\bar{f}_n$ *be the above-defined density estimate of* f *with a symmetric density* K, *which is bounded and monotonically decreasing on* $\mathbb{R}_+$. *Then (A1), (A2), (A3) and*

$$\text{(A8)}\ \ \mathbf{E}\left\{\frac{\sum_{i=1}^{N_n} G\left(\frac{\|X-X_i\|}{H_n}\right)\sum_{k=1}^{l_{i,n}}\left|\bar{Y}_k^{(i)}-Y_k^{(i)}\right|}{h_n\cdot\sum_{j=1}^{N_n} l_{j,n}\cdot G\left(\frac{\|X-X_j\|}{H_n}\right)}\right\} \to 0 \quad (n\to\infty)$$

imply

$$\mathbf{E}\int\int |\bar{f}_n(y,x)-f(y,x)|\,\mathrm{d}y\,P_X(\mathrm{d}x)\to 0 \quad (n\to\infty).$$

Remark 8 The conditions $(A1)$, $(A2)$ and $(A3)$ are also needed for data without additional measurement errors. Condition $(A8)$ specifies how the measurement errors need to behave such that our estimator remains consistent.

Theorem 4 *Assume that $f:\mathbb{R}\times\mathbb{R}^d\to\mathbb{R}_+$ satisfies*

$$\int_{\mathbb{R}} |f(y,x_1)-f(y,x_2)|\,\mathrm{d}y \le c_{15}\cdot\|x_1-x_2\|^{\alpha} \quad (x_1,x_2\in\mathbb{R}^d)$$

for some $c_{15}>0$, $\alpha\in(0,1]$ and let f_n be the above-defined density estimate of f with $l_{1,n}=l_{2,n}=\cdots=l_{N_n,n}=: l_n$. Then (A4), (A5), (A6),

(A7′) K is a symmetric density, which is bounded, monotonically decreasing on the positive real axis and which satisfies

$$\int_{\mathbb{R}} K(u)\cdot|u|^r\,\mathrm{d}u<\infty$$

and

(A8′)

$$\frac{1}{l_n}\sum_{k=1}^{l_n}\mathbf{E}\left\{|\bar{Y}_k^{(i)}-Y_k^{(i)}|\,\middle|\,X,X_i\right\}\le c_{16}\cdot\frac{h_n}{\delta_n}$$

almost surely for all $1\le i\le N_n$, some $c_{16}>0$ and some $\delta_n>0$

imply that there exist constants $c_{17}, c_{18}, c_{19}, c_{20}$ and $c_{21}>0$ such that for all $n\in\mathbb{N}$ we have

$$\begin{aligned}\mathbf{E}\int\int |\bar{f}_n(y,x)-f(y,x)|\,\mathrm{d}y\,P_X(\mathrm{d}x) &\le \frac{c_{17}}{\sqrt{N_n\cdot l_n\cdot h_n\cdot H_n^d}}+c_{18}\cdot H_n^{\alpha}+c_{19}\cdot h_n^r\\ &\quad+\frac{c_{20}}{N_n\cdot H_n^d}+c_{21}\cdot\delta_n^{-1}.\end{aligned}$$

Remark 9 We can draw the following conclusions from Corollary 2 and Theorem 4: In case that $N_n^{-\alpha(r+1)}\le l_n^{-r(\alpha+d)}$ we set $h_n=c_{22}\cdot(N_n\cdot l_n)^{-\frac{\alpha}{(2\cdot\alpha+d)r+\alpha}}$ and $H_n=c_{23}\cdot(N_n\cdot l_n)^{-\frac{r}{(2\cdot\alpha+d)r+\alpha}}$. For suitable chosen $c_{24}>0$ it holds

$$\mathbf{E}\int\int |\bar{f}_n(y,x)-f(y,x)|\,\mathrm{d}y\,P_X(\mathrm{d}x)\le c_{24}\cdot\max\left\{(N_n\cdot l_n)^{-\frac{\alpha\cdot r}{(2\cdot\alpha+d)r+\alpha}},\delta_n^{-1}\right\}.$$

And in case that $N_n^{-\alpha(r+1)} > l_n^{-r(\alpha+d)}$ we set $h_n = c_{25} \cdot N_n^{-\frac{\alpha}{(\alpha+d)(2\cdot r+1)}} \cdot l_n^{-\frac{1}{2\cdot r+1}}$ and $H_n = c_{26} \cdot N_n^{-\frac{1}{\alpha+d}}$. For suitable chosen $c_{27} > 0$ it holds

$$\mathbf{E} \int \int |\bar{f}_n(y, x) - f(y, x)| \, \mathrm{d}y \, P_X(\mathrm{d}x) \leq c_{27} \cdot \max \left\{ N_n^{-\frac{\alpha}{\alpha+d}}, \delta_n^{-1} \right\}.$$

3 Application to simulated data

In this section we consider three different examples of simulated data. In all cases the covariate is uniformly distributed, whereas the distribution of the data sets varies. At first we sample $N = 80$ covariates $\{X_1, X_2, \ldots, X_{80}\}$, and afterwards we sample the corresponding data sets $\mathcal{D}_i = \{Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_{25}^{(i)}\}$ where we observe for each value of the covariate 25 points. Overall we sample $n = 2000$ data points beside the covariates. Our estimator uses the corresponding data of those covariates for which the difference of the covariates to the considered covariate is less than the bandwidth H. And for each covariate the density estimate with bandwidth h is considered. We choose both bandwidths $\theta = (h, H)$ with $h, H > 0$ out of a set of parameters Θ by adapting the combinatorial method of Devroye and Lugosi (2001) in our setting. Here we let $\Theta = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}^2$. We define the empirical measure based on the data set $\mathcal{D}_i$ by

$$\hat{\mu}_i(A) = \frac{1}{25} \sum_{k=1}^{25} \mathbb{1}_A(Y_k^{(i)}) \quad (A \subseteq \mathbb{R})$$

and the proposed estimator without the data set $\mathcal{D}_i$ is defined by

$$f_{i,\theta}(y, x) = \frac{\sum_{l=1, l \neq i}^{80} G\left(\frac{\|x - X_l\|}{H}\right) \sum_{k=1}^{25} K\left(\frac{y - Y_k^{(l)}}{h}\right)}{h \sum_{j=1, j \neq i}^{80} 25 \cdot G\left(\frac{\|x - X_j\|}{H}\right)}.$$

We select $\hat{\theta} = (\hat{h}, \hat{H})$ through minimizing

$$\Delta_\theta = \sum_{i=1}^{80} \sup_{A_i \in \mathcal{A}_i} \left| \int_{A_i} f_{i,\theta}(y, X_i) \, \mathrm{d}y - \hat{\mu}_i(A_i) \right|,$$

where

$$\mathcal{A}_i = \left\{ \left\{ y \in \mathbb{R} \, : \, \hat{f}_{i,\theta_1}(y, X_i) > \hat{f}_{i,\theta_2}(y, X_i) \right\} : \theta_1, \theta_2 \in \Theta \right\},$$

i.e., we choose

$$\hat{\theta} = (\hat{h}, \hat{H}) = \arg \min_{\theta \in \Theta} \Delta_\theta.$$

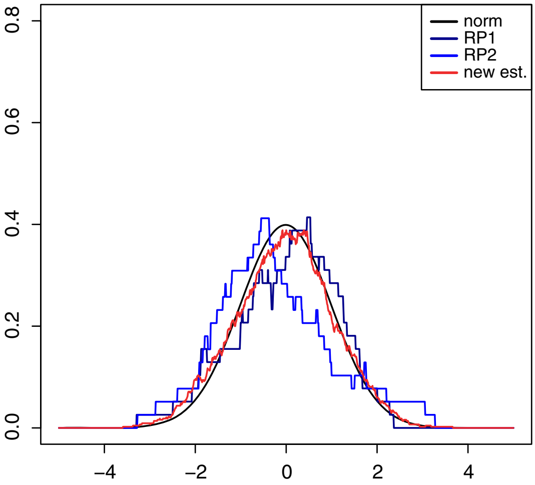


Fig. 1 Typical simulation for $\mu = 0.16$

With this bandwidths $\hat{\theta}$ we define our estimator

$$f_{\hat{\theta}}(y,x) = \frac{\sum_{l=1}^{80} G\left(\frac{\|x-X_l\|}{\hat{H}}\right) \sum_{k=1}^{25} K\left(\frac{y-Y_k^{(l)}}{\hat{h}}\right)}{\hat{h}\sum_{j=1}^{80} 25 \cdot G\left(\frac{\|x-X_j\|}{\hat{H}}\right)}.$$

In the implementation of our estimate we approximate all integrals by Riemann sums. In addition to our proposed estimator we consider two variants of the Rosenblatt–Parzen density estimator. The first one (RP1) is the Rosenblatt–Parzen estimator applied to 25 data points which are specially sampled to the considered covariate. Hence, this estimator uses data that are actually not available and, therefore, it is in practice not applicable to our setting. The second version (RP2) uses those data points (25 points) for which the corresponding covariate of our covariate sample comes closest to the considered covariate. For both the bandwidths are chosen by unbiased cross validation. For all three estimators we use the naive kernel.

In the first simulation model we let the data be independent normally distributed with variance one. In this case the covariate corresponds to the expected value which varies with each data set. We let the covariate be uniformly distributed on $[-0.5, 0.5]$.

Figure 1 shows a typical simulation of the three estimators and the real density for $\mu = 0.16$. While the Rosenblatt–Parzen density estimators used only 25 data points, the proposed estimator used in this case 475 data points. Since the results of our simulation depend on randomly occurring data points, we repeat the whole procedure 100 times and report boxplots in Fig. 2. We compare the estimated average L_1-errors of all these estimators. The mean of the estimated average L_1-errors of the proposed estimate (0.236) is less than the mean of the estimated average L_1-errors of the Rosenblatt–Parzen estimators (0.324, 0.442).

Secondly, we consider exponentially distributed data with parameters λ that are uniformly distributed on $[0.5, 1.5]$. In Fig. 3 an illustrative comparison of the proposed

Fig. 2 *Boxplots* of the estimated average L_1-errors

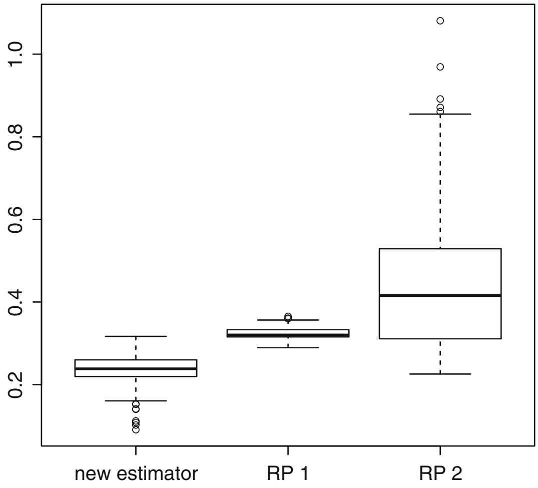
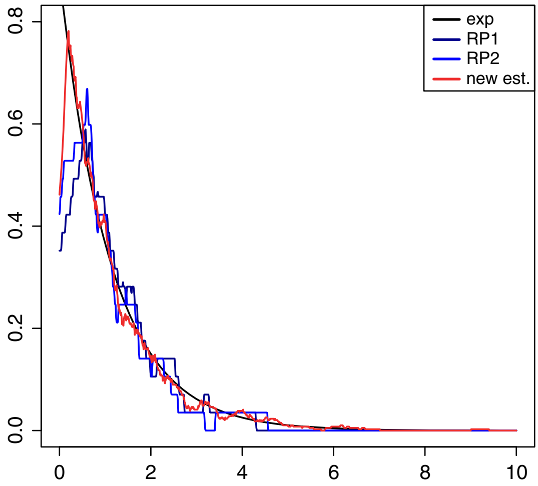


Fig. 3 Typical simulation for $\lambda = 0.89$



estimator (which uses 975 data points) and the Rosenblatt–Parzen density estimators (based on 25 data points) is pictured in case of $\lambda = 0.89$. As before we compare in Fig. 4 the estimated L_1-errors of these estimates. The means of the estimated average L_1-error of the Rosenblatt–Parzen estimators (0.473, 0.550) are nearly twice the mean of the proposed estimate (0.253).

As a third example we consider log-normally distributed data with variance one. As in the first simulation model the covariate corresponds to the expected value and is uniformly distributed on $[-0.5, 0.5]$. Figure 5 shows a simulation example for $\mu = 0$. Here the proposed estimator uses 2000 data points. Comparing the estimated average L_1-errors for 100 repetitions we obtain the boxplot in Fig. 6. Also in this example the proposed estimator outperforms the other estimators. The means of the estimated

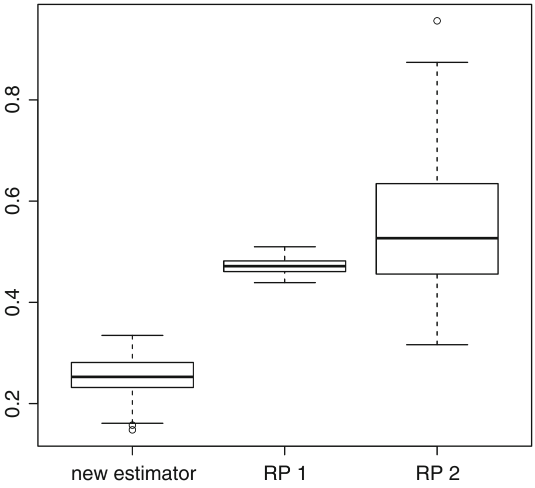


Fig. 4 *Boxplots* of the estimated average L_1-errors

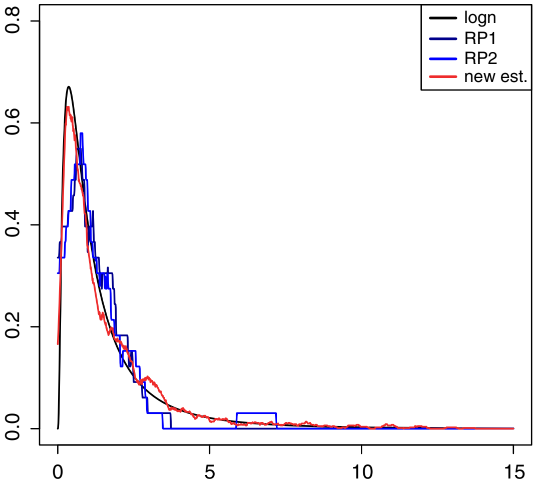


Fig. 5 Typical simulation for $\mu = 0$

average L_1-error of the Rosenblatt–Parzen estimators (0.497, 0.549) are considerably higher than the mean of the proposed estimate (0.243).

The advantages of our estimator become evident in applications where the sample size per covariate is very small or where no sample for the considered covariate is available. If we want to estimate a density in dependence of one particular covariate where a corresponding sample exists, we can also apply the Rosenblatt–Parzen estimator to this sample. This corresponds to the above-introduced RP1. Because of the small sample size this estimator performs worse than our estimator. Clearly, this effect could reverse with a larger data sample per covariate. Simulations showed that in the first example around 50 data points are enough, while in the other two examples the

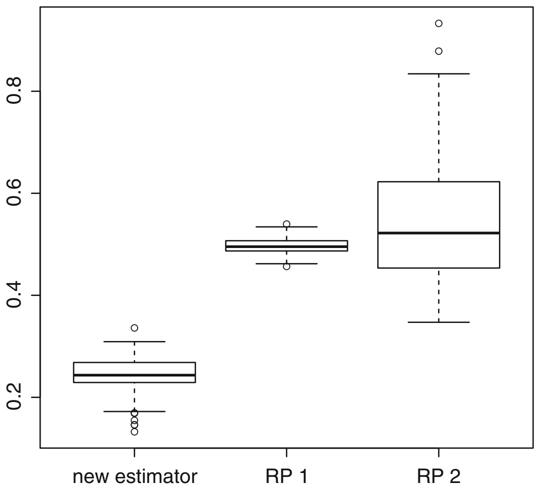


Fig. 6 *Boxplots* of the estimated average L_1-errors

Rosenblatt–Parzen estimator needs around 120 data points to achieve results that are comparable to the ones of our estimator.

Finally we apply our estimator to examine the fatigue behavior of steel under cyclic loading. The data are obtained by relatively time-consuming experiments where for each material m and several adjusted total strain amplitudes ϵ the corresponding numbers of cycles N till failure are determined. We are interested in the random behavior of N. In our model we assume that the behavior of the numbers of cycles N till failure can be described by

$$N(m,\epsilon) = \mu(m,\epsilon) + \sigma(m,\epsilon) \cdot \delta^{(m)}, \tag{8}$$

where $\mu(m,\epsilon)$ is the expected value of $N(m,\epsilon)$ and $\sigma(m,\epsilon)$ is its standard variation. $\delta^{(m)}$ is an error term with expected value zero. Hence, we expect the numbers of cycles till failure to vary around the expected value by a random error term. While the numbers of cycles till failure and accordingly its expected value and variance depend on the material m and the total strain amplitudes ϵ, we assume that the error $\delta^{(m)}$ only depends on the material m and has a density. Our goal is to estimate the density of $\delta^{(m)}$ using given data sets

$$\left\{(\epsilon_1^{(m)}, N_1^{(m)}), \ldots, (\epsilon_{l_m}^{(m)}, N_{l_m}^{(m)})\right\}$$

for each material m. Since these experiments are very time consuming, the number of observations per material is low. Here we consider 26 materials with 305 observations in total. Hence, the sample size per covariate is very low (at most 21 observations, on average 12). We will apply our estimator to estimate the density of $\delta^{(\bar{m})}$, $\bar{m} \in \{1\ldots, 26\}$. But at first we need to construct data of $\delta^{(m)}$ for each material m. Because of (8) we have samples of $\delta^{(m)}$ given by

$$\delta_i^{(m)} = \frac{N_i^{(m)} - \mu(m, \epsilon_i^{(m)})}{\sigma(m, \epsilon_i^{(m)})} \quad (i = 1, \dots, l_m)$$

for each material $m = 1, \dots, 26$. Since μ and σ are unknown, we plug in estimates $\hat{\mu}(m, \epsilon)$ of $\mu(m, \epsilon)$ and $\hat{\sigma}(m, \epsilon)$ of $\sigma(m, \epsilon)$. Due to this estimates we obtain only data with measurement errors. We apply the parametric estimator of Williams et al. (2002) to estimate $\mu(m, \epsilon)$. Therefore, we need to assume that the mean behavior of N is given by the cyclic stress–strain curve (cf., Manson 1965) and consequently we need to estimate only the parameters that determine this curve. The estimation of the variance is more complicated, because we need to apply a nonparametric estimator that usually needs more data. This is the reason, why we generate artificial data points like in Furer and Kohler (2013) and apply the referred smoothing spline estimator to the original data points and artificial data. Thus, for all considered materials we construct a data sample

$$\hat{\delta}_1^{(m)}, \dots, \hat{\delta}_{l_m}^{(m)}$$

via

$$\hat{\delta}_i^{(m)} = \frac{N_i^{(m)} - \hat{\mu}(m, \epsilon_i^{(m)})}{\hat{\sigma}(m, \epsilon_i^{(m)})},$$

where $\hat{\mu}(m, \epsilon)$ and $\hat{\sigma}(m, \epsilon)$ are the above-mentioned estimators. Thereby we can apply our estimator and get a density estimate of the error variable in dependence of the material. We determine the bandwidth of our estimator as in the case of simulated data.

If we consider one particular material, we obtain density estimates of the numbers of cycles till failure in dependence of the total strain amplitudes ϵ, because for each material m and total strain amplitude ϵ, $\mu(m, \epsilon)$ and $\sigma(m, \epsilon)$ and, respectively, $\hat{\mu}(m, \epsilon)$ and $\hat{\sigma}(m, \epsilon)$ are fix. Let $\hat{f}^{(m)}$ be the density estimate of $\delta^{(m)}$, then (8) implies that

$$\hat{g}^{(m)}(\cdot) = \frac{\hat{f}^{(m)}\left(\frac{\cdot\ - \hat{\mu}(m,\epsilon)}{\hat{\sigma}(m,\epsilon)}\right)}{\hat{\sigma}(m, \epsilon)}$$

is a density estimate of $N(m, \epsilon)$. While $\hat{f}^{(m)}$ is fix for all strain amplitudes ϵ, $\hat{\mu}(m, \epsilon)$ and $\hat{\sigma}(m, \epsilon)$ vary with each strain amplitude and thus, $\hat{g}^{(m)}$ alters for each ϵ. In the following Fig. 7 you can see our estimator $\hat{g}^{(m)}$ in dependence of ϵ for one specified material. The maxima to each curve describe the cyclic stress–strain curve with the estimated parameters as before. The figure shows, how the numbers of cycles N till failure vary around its expected value.

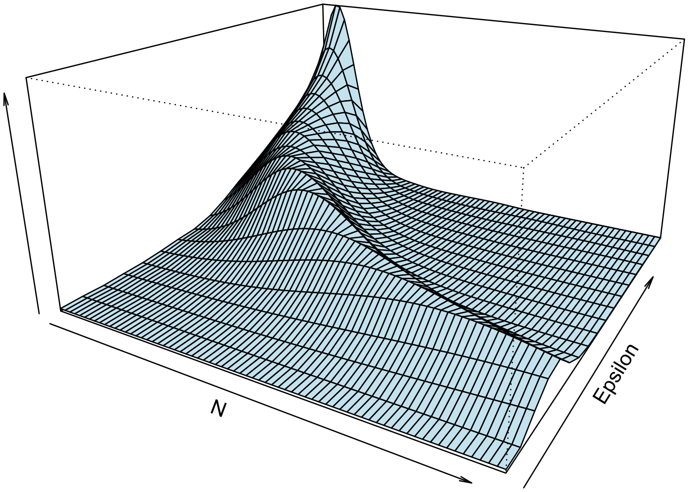


Fig. 7 Estimated densities of N in dependence of ϵ

4 Proofs

4.1 Proof of Theorem 1

Let $B \subset \mathbb{R}$ be compact. According to the Lemma of Scheffé it holds

$$\begin{aligned}
&\int\int |f_n(y,x) - f(y,x)|\, \mathrm{d}y\, P_X(\mathrm{d}x) \\
&\quad = 2 \cdot \int\int (f(y,x) - f_n(y,x))_+\, \mathrm{d}y\, P_X(\mathrm{d}x) \\
&\quad \le 2 \cdot \int\int_B |f_n(y,x) - f(y,x)|\, \mathrm{d}y\, P_X(\mathrm{d}x) + 2 \cdot \int\int_{B^c} f(y,x)\, \mathrm{d}y\, P_X(\mathrm{d}x)
\end{aligned}$$

whenever $f_n(\cdot,x) \neq 0$. Trivially this also holds in case $f_n(\cdot,x) = 0$. For suitable chosen B the second summand is arbitrary small and thus it suffices to show

$$\mathbf{E}\int\int_B |f_n(y,x) - f(y,x)|\, \mathrm{d}y\, P_X(\mathrm{d}x) \to 0 \quad (n \to \infty). \tag{9}$$

With $C \subset \mathbb{R}$ it holds

$$\begin{aligned}
&\mathbf{E}\int\int_B |f_n(y,x) - f(y,x)|\, \mathrm{d}y\, P_X(\mathrm{d}x) \\
&\quad \le \mathbf{E}\int_C\int_B |f_n(y,x) - f(y,x)|\, \mathrm{d}y\, P_X(\mathrm{d}x) + 2 \cdot P_X(C^c).
\end{aligned}$$

We can choose $C \subset \mathbb{R}^d$ compact such that the second summand is arbitrarily small, thus, we consider only the first summand. By Fubini's Theorem and the triangle inequality we get

$$
\begin{aligned}
&\mathbf{E}\int_C\int_B |f_n(y,x)-f(y,x)|\,\mathrm{d}y\,P_X(\mathrm{d}x)\\
&=\int_C\int_B \mathbf{E}\left\{|f_n(y,x)-f(y,x)|\right\}\,\mathrm{d}y\,P_X(\mathrm{d}x)\\
&\le \int_C\int_B \mathbf{E}\left\{|f_n(y,x)-\mathbf{E}\{f_n(y,x)\,|\,X_1,\dots,X_{N_n}\}|\right\}\,\mathrm{d}y\,P_X(\mathrm{d}x)\\
&\quad + C\int_B \mathbf{E}\left\{|\mathbf{E}\{f_n(y,x)\,|\,X_1,\dots,X_{N_n}\}-f(y,x)|\right\}\,\mathrm{d}y\,P_X(\mathrm{d}x)\\
&=: A_{1,n}+A_{2,n}.
\end{aligned}
$$

At first we consider $A_{1,n}$. With the Cauchy–Schwarz inequality, the fact that $B\subset\mathbb{R}$ is compact and Fubini's Theorem we conclude that

$$
\begin{aligned}
&\int_C\int_B \mathbf{E}\left\{|f_n(y,x)-\mathbf{E}\{f_n(y,x)\,|\,X_1,\dots,X_{N_n}\}|\right\}\,\mathrm{d}y\,P_X(\mathrm{d}x)\\
&\le \int_C\int_B \sqrt{\mathbf{E}\left\{|f_n(y,x)-\mathbf{E}\{f_n(y,x)\,|\,X_1,\dots,X_{N_n}\}|^2\right\}}\,\mathrm{d}y\,P_X(\mathrm{d}x)\\
&= \int_C\int_B \sqrt{\mathbf{E}\left\{\mathbf{E}\left\{|f_n(y,x)-\mathbf{E}\{f_n(y,x)\,|\,X_1,\dots,X_{N_n}\}|^2\,\middle|\,X_1,\dots,X_{N_n}\right\}\right\}}\,\mathrm{d}y\,P_X(\mathrm{d}x)\\
&\le c_{28}\cdot\sqrt{\int_C \mathbf{E}\left\{\int_B \mathbf{E}\left\{|f_n(y,x)-\mathbf{E}\{f_n(y,x)\,|\,X_1,\dots,X_{N_n}\}|^2\middle|\,X_1,\dots,X_{N_n}\right\}\,\mathrm{d}y\right\}\,P_X(\mathrm{d}x)}.
\end{aligned}
$$

Using the independence of the data sets $\mathcal{D}_n^{(i)}$ and of the data within each data set we obtain

$$
\begin{aligned}
&\int_B \mathbf{E}\left\{|f_n(y,x)-\mathbf{E}\{f_n(y,x)\,|\,X_1,\dots,X_{N_n}\}|^2\middle|\,X_1,\dots,X_{N_n}\right\}\,\mathrm{d}y\\
&=\int_B \mathbf{E}\left\{\left(\frac{\sum_{i=1}^{N_n} G\left(\frac{\|x-X_i\|}{H_n}\right)\sum_{k=1}^{l_{i,n}}\left(K\left(\frac{y-Y_k^{(i)}}{h_n}\right)-\mathbf{E}\left\{K\left(\frac{y-Y_k^{(i)}}{h_n}\right)\,\middle|\,X_1,\dots,X_n\right\}\right)}{h_n\sum_{j=1}^{N_n} l_{j,n}\cdot G\left(\frac{\|x-X_j\|}{H_n}\right)}\right)^2\right.\\
&\qquad\qquad \left.\middle|\,X_1,\dots,X_{N_n}\right\}\,\mathrm{d}y\\
&=\int_B \frac{\sum_{i=1}^{N_n} G\left(\frac{\|x-X_i\|}{H_n}\right)\sum_{k=1}^{l_{i,n}}\mathbf{E}\left\{\left(K\left(\frac{y-Y_k^{(i)}}{h_n}\right)-\mathbf{E}\left\{K\left(\frac{y-Y_k^{(i)}}{h_n}\right)\,\middle|\,X_1,\dots,X_n\right\}\right)^2\,\middle|\,X_1,\dots,X_{N_n}\right\}}{\left(h_n\sum_{j=1}^{N_n} l_{j,n}\cdot G\left(\frac{\|x-X_j\|}{H_n}\right)\right)^2}\,\mathrm{d}y\\
&=\frac{\sum_{i=1}^{N_n}\frac{l_{i,n}}{h_n^2}G\left(\frac{\|x-X_i\|}{H_n}\right)\int_B \mathbf{E}\left\{\left(K\left(\frac{y-Y_k^{(i)}}{h_n}\right)-\mathbf{E}\left\{K\left(\frac{y-Y_k^{(i)}}{h_n}\right)\,\middle|\,X_i\right\}\right)^2\,\middle|\,X_i\right\}\,\mathrm{d}y}{\left(\sum_{j=1}^{N_n} l_{j,n}\cdot G\left(\frac{\|x-X_j\|}{H_n}\right)\right)^2}.
\end{aligned}
$$

Due to the square-integrability of K we know that there exists a constant $c_{29} > 0$ such that

$$\int_{\mathbb{R}} K^2(y)\,\mathrm{d}y \le c_{29}.$$

With

$$\begin{aligned}
&\int_B \mathbf{E}\left\{\left(K\left(\frac{y-Y_k^{(i)}}{h_n}\right) - \mathbf{E}\left\{K\left(\frac{y-Y_k^{(i)}}{h_n}\right)\Bigg|\, X_i\right\}\right)^2 \Bigg|\, X_i\right\} \mathrm{d}y \\
&\le \int_B \mathbf{E}\left\{K^2\left(\frac{y-Y_k^{(i)}}{h_n}\right)\Bigg|\, X_i\right\} \mathrm{d}y \\
&= \int_B \int K^2\left(\frac{y-u}{h_n}\right) f(u, X_i)\,\mathrm{d}u\,\mathrm{d}y \\
&\le \int_{\mathbb{R}} \int K^2\left(\frac{y-u}{h_n}\right) f(u, X_i)\,\mathrm{d}u\,\mathrm{d}y \\
&= h_n \int \int_{\mathbb{R}} K^2(z)\,dz\, f(u, X_i)\,\mathrm{d}u \\
&\le h_n \cdot c_{29}
\end{aligned}$$

we obtain

$$\begin{aligned}
A_{1,n} &\le c_{28} \cdot \sqrt{\int_C \mathbf{E}\left\{\frac{\frac{c_{29}}{h_n}\sum_{i=1}^{N_n} l_{i,n}\cdot G\left(\frac{\|x-X_i\|}{H_n}\right)}{\left(\sum_{i=1}^{N_n} l_{i,n}\cdot G\left(\frac{\|x-X_i\|}{H_n}\right)\right)^2}\right\} P_X(\mathrm{d}x)} \\
&= c_{28} \cdot \sqrt{\int_C \mathbf{E}\left\{\frac{c_{29}}{h_n\cdot\sum_{i=1}^{N_n} l_{i,n}\cdot G\left(\frac{\|X-X_i\|}{H_n}\right)} \cdot \mathbb{1}_{\left\{\sum_{j=1}^{N_n} l_{i,n}\cdot G\left(\frac{\|x-X_j\|}{H_n}\right)>0\right\}}\right\} P_X(\mathrm{d}x).}
\end{aligned}$$

Applying Lemma 4.1 of Györfi et al. (2002) we get

$$\begin{aligned}
&\int_C \mathbf{E}\left\{\frac{c_{29}}{h_n\cdot\sum_{i=1}^{N_n} l_{i,n}\cdot G\left(\frac{\|x-X_i\|}{H_n}\right)} \cdot \mathbb{1}_{\left\{\sum_{j=1}^{N_n} l_{i,n}\cdot G\left(\frac{\|x-X_j\|}{H_n}\right)>0\right\}}\right\} P_X(\mathrm{d}x) \\
&\le \frac{c_{29}}{\min\limits_{1\le i\le N_n} l_{i,n}\cdot h_n} \cdot \int_C \mathbf{E}\left\{\frac{1}{\sum_{i=1}^{N_n} G\left(\frac{\|x-X_i\|}{H_n}\right)} \cdot \mathbb{1}_{\left\{\sum_{j=1}^{N_n} G\left(\frac{\|x-X_j\|}{H_n}\right)>0\right\}}\right\} P_X(\mathrm{d}x) \\
&\le \frac{c_{29}}{\min\limits_{1\le i\le N_n} l_{i,n}\cdot h_n} \cdot \int_C \frac{2}{(N_n+1)\cdot\mathbf{P}\{\|x-X_1\|\le H_n\}} P_X(\mathrm{d}x).
\end{aligned}$$

Due to compactness of C we can apply Equation (5.1) of the proof of Theorem 5.1 in Györfi et al. (2002) and conclude that

$$\int_C \frac{1}{\mathbf{P}\{\|x - X_1\| \leq H_n\}} P_X(\mathrm{d}x) \leq \frac{c_{30}}{H_n^d}.$$

Hence,

$$A_{1,n} \leq \frac{c_{31}}{\sqrt{N_n \cdot H_n^d \cdot h_n \cdot \min\limits_{1 \leq i \leq N_n} l_{i,n}}}.$$

Due to assumption (A2) $A_{1,n}$ converges to zero. It remains to show that

$$A_{2,n} = \int_C \mathbf{E} \int_B |\mathbf{E}\{f_n(y,x) \,|\, X_1, \ldots, X_{N_n}\} - f(y,x)| \,\mathrm{d}y \, P_X(\mathrm{d}x) \to 0 \quad (n \to \infty).$$

Using $l_{j,n} > 0$ for all $j \in \{1, \ldots, N_n\}$ and $n \in \mathbb{N}$ we obtain

$$\begin{aligned}
&\int_C \mathbf{E} \int_B |\mathbf{E}\{f_n(y,x) \,|\, X_1, \ldots, X_{N_n}\} - f(y,x)| \,\mathrm{d}y \, P_X(\mathrm{d}x) \\
&= \int_C \mathbf{E} \int_B \left| f(y,x) - \frac{\sum_{i=1}^{N_n} l_{i,n} G\left(\frac{\|x-X_i\|}{H_n}\right) \mathbf{E}\left\{K\left(\frac{y-Y_1^{(i)}}{h_n}\right) \,\middle|\, X_1, \ldots, X_{N_n}\right\}}{h_n \sum_{j=1}^{N_n} l_{j,n} \cdot G\left(\frac{\|x-X_j\|}{H_n}\right)} \right| \mathrm{d}y \, P_X(\mathrm{d}x) \\
&= \int_C \mathbf{E} \int_B \left| f(y,x) - \frac{\sum_{i=1}^{N_n} l_{i,n} G\left(\frac{\|x-X_i\|}{H_n}\right) \frac{1}{h_n} \int K\left(\frac{y-u}{h_n}\right) f(u, X_i) \,\mathrm{d}u}{\sum_{j=1}^{N_n} l_{j,n} \cdot G\left(\frac{\|x-X_j\|}{H_n}\right)} \right| \mathrm{d}y \, P_X(\mathrm{d}x) \\
&\leq \int_C \mathbf{E} \int_B f(y,x) \cdot \mathbb{1}_{\left\{\sum_{j=1}^{N_n} G\left(\frac{\|x-X_j\|}{H_n}\right) = 0\right\}} \,\mathrm{d}y \, P_X(\mathrm{d}x) \\
&\quad + \int_C \mathbf{E} \int_B \left| \frac{\sum_{i=1}^{N_n} l_{i,n} G\left(\frac{\|x-X_i\|}{H_n}\right) \cdot \left(f(y,x) - \frac{1}{h_n} \int K\left(\frac{y-u}{h_n}\right) \cdot f(u, X_i) \,\mathrm{d}u\right)}{\sum_{j=1}^{N_n} l_{j,n} \cdot G\left(\frac{\|x-X_j\|}{H_n}\right)} \right| \mathrm{d}y \, P_X(\mathrm{d}x) \\
&\leq \int_C \mathbf{P}\left\{\sum_{j=1}^{N_n} G\left(\frac{\|x - X_j\|}{H_n}\right) = 0\right\} \mathrm{d}y \, P_X(\mathrm{d}x) \\
&\quad + \int_C \mathbf{E} \int_B \frac{\sum_{i=1}^{N_n} l_{i,n} G\left(\frac{\|x-X_i\|}{H_n}\right) \cdot \left| f(y,x) - \int \frac{1}{h_n} K\left(\frac{y-u}{h_n}\right) \cdot f(u, X_i) \,\mathrm{d}u \right|}{\sum_{j=1}^{N_n} l_{j,n} \cdot G\left(\frac{\|x-X_j\|}{H_n}\right)} \,\mathrm{d}y \, P_X(\mathrm{d}x) \\
&=: B_{1,n} + B_{2,n}.
\end{aligned}$$

With $B_{1,n}$ we proceed analogously to the proof of Theorem 5.1, Equation (5.1) in Györfi et al. (2002). We choose $S \subset \mathbb{R}$ compact such that $P_X(S^c)$ is arbitrary small.

Then it holds

$$B_{1,n} = \mathbf{P}\left\{\sum_{j=1}^{N_n} G\left(\frac{\|X - X_j\|}{H_n}\right) = 0\right\} \le \frac{c_{32}}{N_n \cdot H_n^d} + P_X(S^c),$$

which implies $B_{1,n} \to 0$ for $(n \to \infty)$.

In the sequel we bound $B_{2,n}$ from above. Our proof is based on Devroye (2015). We have

$$\begin{aligned}
B_{2,n} \le& \int_C \mathbf{E} \int_B \frac{\sum_{i=1}^{N_n} l_{i,n} G\left(\frac{\|x-X_i\|}{H_n}\right) \cdot \left| f(y,x) - \int \frac{1}{h_n} K\left(\frac{y-u}{h_n}\right) \cdot f(u,x)\, \mathrm{d}u\right|}{\sum_{j=1}^{N_n} l_{j,n} \cdot G\left(\frac{\|x-X_j\|}{H_n}\right)} \mathrm{d}y\, P_X(\mathrm{d}x) \\
&+ \int_C \mathbf{E} \int_B \frac{\sum_{i=1}^{N_n} l_{i,n} G\left(\frac{\|x-X_i\|}{H_n}\right) \cdot \int \frac{1}{h_n} K\left(\frac{y-u}{h_n}\right) \cdot |f(u, X_i) - f(u,x)|\, \mathrm{d}u}{\sum_{j=1}^{N_n} l_{j,n} \cdot G\left(\frac{\|x-X_j\|}{H_n}\right)} \mathrm{d}y\, P_X(\mathrm{d}x) \\
\le& \int_C \int_B \left| f(y,x) - \int \frac{1}{h_n} K\left(\frac{y-u}{h_n}\right) \cdot f(u,x)\, \mathrm{d}u\right| \mathrm{d}y\, P_X(\mathrm{d}x) \\
&+ \int_C \mathbf{E} \int_B \frac{\sum_{i=1}^{N_n} l_{i,n} G\left(\frac{\|x-X_i\|}{H_n}\right) \cdot \int \frac{1}{h_n} K\left(\frac{y-u}{h_n}\right) \cdot |f(u, X_i) - f(u,x)|\, \mathrm{d}u}{\sum_{j=1}^{N_n} l_{j,n} \cdot G\left(\frac{\|x-X_j\|}{H_n}\right)} \mathrm{d}y\, P_X(\mathrm{d}x) \\
=& C_{1,n} + C_{2,n}.
\end{aligned}$$

For $x \in \mathbb{R}^d$ we know that $f(\cdot, x)$ is a density, hence Theorem 1, Chapter 2 in Devroye and Györfi (1985) implies

$$\int_B \left| f(y,x) - \int \frac{1}{h_n} K\left(\frac{y-u}{h_n}\right) \cdot f(u,x)\, \mathrm{d}u\right| \mathrm{d}y \to 0 \quad (n \to \infty).$$

Because of

$$\begin{aligned}
&\int_B \left| f(y,x) - \int \frac{1}{h_n} K\left(\frac{y-u}{h_n}\right) \cdot f(u,x)\, \mathrm{d}u\right| \mathrm{d}y \\
&\quad \le \int_{\mathbb{R}} f(y,x)\, \mathrm{d}y + \int \int \frac{1}{h_n} K\left(\frac{y-u}{h_n}\right) \mathrm{d}y \cdot f(u,x)\, \mathrm{d}u = 2
\end{aligned}$$

this together with the dominated convergence theorem implies

$$C_{n,1} \to 0 \quad (n \to \infty).$$

Furthermore

$$\begin{aligned}
C_{2,n} \le\ & \frac{\max_{i=1,\dots,N_n} l_{i,n}}{\min_{i=1,\dots,N_n} l_{i,n}} \\
& \times \int_C \mathbf{E} \int_B \frac{\sum_{i=1}^{N_n} G\left(\frac{\|x-X_i\|}{H_n}\right) \cdot \int \frac{1}{h_n} K\left(\frac{y-u}{h_n}\right) \cdot |f(u, X_i) - f(u, x)|\, \mathrm{d}u}{\sum_{j=1}^{N_n} G\left(\frac{\|x-X_j\|}{H_n}\right)}\, \mathrm{d}y\, P_X(\mathrm{d}x) \\
\le\ & \frac{\max_{i=1,\dots,N_n} l_{i,n}}{\min_{i=1,\dots,N_n} l_{i,n}} \\
& \times \int_C \mathbf{E} \frac{\sum_{i=1}^{N_n} G\left(\frac{\|x-X_i\|}{H_n}\right) \cdot \int \int_B \frac{1}{h_n} K\left(\frac{y-u}{h_n}\right) \mathrm{d}y \cdot |f(u, X_i) - f(u, x)|\, \mathrm{d}u}{\sum_{j=1}^{N_n} G\left(\frac{\|x-X_j\|}{H_n}\right)}\, P_X(\mathrm{d}x) \\
\le\ & \frac{\max_{i=1,\dots,N_n} l_{i,n}}{\min_{i=1,\dots,N_n} l_{i,n}} \\
& \times \int_C \mathbf{E} \left\{ \frac{\sum_{i=1}^{N_n} G\left(\frac{\|x-X_i\|}{H_n}\right) \cdot \int |f(u, X_i) - f(u, x)|\, \mathrm{d}u}{\sum_{j=1}^{N_n} G\left(\frac{\|x-X_j\|}{H_n}\right)} \right\} P_X(\mathrm{d}x) \\
\le\ & \frac{\max_{i=1,\dots,N_n} l_{i,n}}{\min_{i=1,\dots,N_n} l_{i,n}} \cdot \int \mathbf{E} \left\{ \frac{\sum_{i=1}^{N_n} G\left(\frac{\|X-X_i\|}{H_n}\right) \cdot |f(u, X_i) - f(u, X)|}{\sum_{j=1}^{N_n} G\left(\frac{\|X-X_j\|}{H_n}\right)} \right\} \mathrm{d}u.
\end{aligned}$$

By the proof of Theorem 5.1 in Györfi et al. (2002) (compare there pages 74–75) we get for any $u \in \mathbb{R}$

$$\begin{aligned}
& \mathbf{E} \left\{ \frac{\sum_{i=1}^{N_n} G\left(\frac{\|X-X_i\|}{H_n}\right) \cdot |f(u, X_i) - f(u, X)|}{\sum_{j=1}^{N_n} G\left(\frac{\|X-X_j\|}{H_n}\right)} \right\} \\
& \le \mathbf{E} \left\{ \frac{\sum_{i=1}^{N_n} G\left(\frac{\|X-X_i\|}{H_n}\right) \cdot f(u, X_i)}{\sum_{j=1}^{N_n} G\left(\frac{\|X-X_j\|}{H_n}\right)} \right\} + \int f(u, x) P_X(\mathrm{d}x) \\
& \le \text{const} \cdot \int f(u, x) P_X(\mathrm{d}x) + \int f(u, x) P_X(\mathrm{d}x),
\end{aligned}$$

where

$$\int \int f(u, x) P_X(\mathrm{d}x)\, \mathrm{d}u = 1 < \infty.$$

Hence, by dominated convergence and assumption (A3) it suffices to show

$$\mathbf{E} \left\{ \frac{\sum_{i=1}^{N_n} G\left(\frac{\|X-X_i\|}{H_n}\right) \cdot |f(u, X_i) - f(u, X)|}{\sum_{j=1}^{N_n} G\left(\frac{\|X-X_j\|}{H_n}\right)} \right\} \to 0 \qquad (10)$$

for Lebesgue almost all $u \in \mathbb{R}$.

The weights of the kernel regression estimates satisfies conditions (1)–(3) in Theorem 1 in Stone (1977) (cf., e.g., proof of Theorem 5.1 in Györfi et al. 2002). Hence, if for some $u \in \mathbb{R}$

$$\mathbf{E}|f(u, X)| < \infty$$

holds, then Proposition 1 in Stone (1977) implies that (10) holds. Since

$$\int_{\mathbb{R}} \mathbf{E}|f(u, X)| \, \mathrm{d}u = \int_{\mathbb{R}^d} \int_{\mathbb{R}} f(y, x) \, \mathrm{d}y P_X(\mathrm{d}x) = 1,$$

this implies the assertion. □

4.2 Proof of Theorem 2

Due to (A4) and (A5) we can choose $B \subset \mathbb{R}$ and $C \subset \mathbb{R}^d$ compact such that

$$\mathbf{E} \int \int |f_n(y, x) - f(y, x)| \, \mathrm{d}y \, P_X(\mathrm{d}x) \leq \quad 2 \cdot \mathbf{E} \int_C \int_B |f_n(y, x) - f(y, x)| \, \mathrm{d}y \, P_X(\mathrm{d}x).$$

According to the proof of Theorem 1 and assumptions (6) and (A4) we have

$$\begin{aligned}
&\mathbf{E} \int_C \int_B |f_n(y, x) - f(y, x)| \, \mathrm{d}y \, P_X(\mathrm{d}x) \leq A_{1,n} + B_{1,n} + C_{1,n} + C_{2,n} \\
&\leq \frac{c_{33}}{\sqrt{N_n \cdot H_n^d \cdot h_n \cdot l_n}} + \frac{c_{34}}{N_n \cdot H_n^d} \\
&\quad + \int_C \int_B \int \frac{1}{h_n} \cdot K\left(\frac{y-u}{h_n}\right) \cdot |f(y, x) - f(u, x)| \, \mathrm{d}u \, \mathrm{d}y \, P_X(\mathrm{d}x) \\
&\quad + \mathbf{E} \left\{ \frac{\sum_{i=1}^{N_n} G\left(\frac{\|X - X_i\|}{H_n}\right) \cdot \int |f(u, X_i) - f(u, X)| \, \mathrm{d}u}{\sum_{i=1}^{N_n} G\left(\frac{\|X - X_i\|}{H_n}\right)} \right\} \\
&\leq \frac{c_{33}}{\sqrt{N_n \cdot H_n^d \cdot h_n \cdot l_n}} + \frac{c_{34}}{N_n \cdot H_n^d} \\
&\quad + \int_C \int_B \int \frac{1}{h_n} \cdot K\left(\frac{y-u}{h_n}\right) \cdot c_2 \cdot |y - u|^r \, \mathrm{d}u \, \mathrm{d}y \, P_X(\mathrm{d}x) \\
&\quad + \mathbf{E} \left\{ \frac{\sum_{i=1}^{N_n} G\left(\frac{\|X - X_i\|}{H_n}\right) \cdot c_1 \|X_i - X\|^\alpha}{\sum_{i=1}^{N_n} G\left(\frac{\|X - X_i\|}{H_n}\right)} \right\} \\
&\leq \frac{c_{33}}{\sqrt{N_n \cdot H_n^d \cdot h_n \cdot l_n}} + \frac{c_{34}}{N_n \cdot H_n^d} + c_{35} \cdot h_n^r + c_{36} \cdot H_n^\alpha.
\end{aligned}$$

The proof is complete. □

4.3 Remarks on the minimization of the right-hand side of (7).

In this subsection, we show how we can choose the bandwidths h_n and H_n such that the right-hand side of (7) is minimal.

Minimizing the right-hand side of (7) with respect to h_n leads to

$$h_n = c_{37} \cdot (N_n \cdot l_n)^{-\frac{1}{2r+1}} \cdot H_n^{-\frac{d}{2r+1}},$$

and using this bandwidth we get from (7)

$$\mathbf{E} \int \int |f_n(y,x) - f(y,x)| \, \mathrm{d}y \, P_X(\mathrm{d}x) \leq c_{38} \cdot (N_n \cdot l_n)^{-\frac{r}{2r+1}} \cdot H_n^{-\frac{d \cdot r}{2r+1}} + c_4 \cdot H_n^{\alpha} + \frac{c_6}{N_n \cdot H_n^d}. \tag{11}$$

The optimal bandwidth H_n which minimizes the right-hand side of (11) satisfies

$$c_{39} \cdot (N_n \cdot l_n)^{-\frac{r}{2r+1}} \cdot \frac{d \cdot r}{2r+1} \cdot H_n^{-\frac{d \cdot r}{2r+1}-1} + \frac{c_{40}}{N_n} \cdot d \cdot H_n^{-d-1} = c_{41} \cdot \alpha \cdot H_n^{\alpha-1}.$$

For the optimal bandwidth either the first term on the left-hand side above is greater than or equal to the second term or vice versa. From this we can conclude that the optimal H_n lies either between the two solutions of

$$c_{39} \cdot (N_n \cdot l_n)^{-\frac{r}{2r+1}} \cdot \frac{d \cdot r}{2r+1} \cdot H_n^{-\frac{d \cdot r}{2r+1}-1} = c_{40} \cdot \alpha \cdot H_n^{\alpha-1}$$

and

$$2 \cdot c_{39} \cdot (N_n \cdot l_n)^{-\frac{r}{2r+1}} \cdot \frac{d \cdot r}{2r+1} \cdot H_n^{-\frac{d \cdot r}{2r+1}-1} = c_{40} \cdot \alpha \cdot H_n^{\alpha-1}$$

or between the two solutions of

$$\frac{c_{40}}{N_n} \cdot d \cdot H_n^{-d-1} = c_{41} \cdot \alpha \cdot H_n^{\alpha-1} \quad \text{and} \quad 2 \cdot \frac{c_{40}}{N_n} \cdot d \cdot H_n^{-d-1} = c_{41} \cdot \alpha \cdot H_n^{\alpha-1}.$$

Hence, up to some constant the optimal H_n satisfies in case $(N_n \cdot l_n)^{-\frac{r}{2r+1}} \cdot H_n^{-\frac{d \cdot r}{2r+1}} \geq \frac{1}{N_n} \cdot H_n^{-d}$ the equation

$$(N_n \cdot l_n)^{-\frac{r}{2r+1}} \cdot H_n^{-\frac{d \cdot r}{2r+1}} = H_n^{\alpha}$$

and in case $(N_n \cdot l_n)^{-\frac{r}{2r+1}} \cdot H_n^{-\frac{d \cdot r}{2r+1}} < \frac{1}{N_n} \cdot H_n^{-d}$ the equation

$$\frac{1}{N_n} \cdot H_n^{-d} = H_n^{\alpha}.$$

4.4 Proof of Theorem 3

Let f_n be the estimator of f that uses real data $Y_k^{(i)}$ instead of $\bar{Y}_k^{(i)}$ $(i = 1, \ldots, N_n;\ k = 1 \ldots, l_{i,n})$. It holds

$$\begin{aligned}&\mathbf{E}\int\int |\bar{f}_n(y,x) - f(y,x)|\,\mathrm{d}y\, P_X(\mathrm{d}x)\\&\quad\le \mathbf{E}\int\int |\bar{f}_n(y,x) - f_n(y,x)|\,\mathrm{d}y\, P_X(\mathrm{d}x) + \mathbf{E}\int\int |f_n(y,x) - f(y,x)|\,\mathrm{d}y\, P_X(\mathrm{d}x).\end{aligned}$$

By Theorem 1 we already know, that with the assumptions (A1) and (A2)

$$\mathbf{E}\int\int |f_n(y,x) - f(y,x)|\,\mathrm{d}y\, P_X(\mathrm{d}x) \to 0 \quad (n \to \infty).$$

Thus, it remains to show that

$$\mathbf{E}\int\int |\bar{f}_n(y,x) - f_n(y,x)|\,\mathrm{d}y\, P_X(\mathrm{d}x) \to 0 \quad (n \to \infty).$$

Due to our assumptions on the kernel K we can apply Lemma 1 of Bott et al. (2013) which yields

$$\int \left| K\left(\frac{y-y_1}{h_n}\right) - K\left(\frac{y-y_2}{h_n}\right)\right| \mathrm{d}y \le 2 \cdot K(0) \cdot |y_1 - y_2| \quad \text{for all } y_1, y_2 \in \mathbb{R}.$$

Hence,

$$\begin{aligned}&\mathbf{E}\int\int |\bar{f}_n(y,x) - f_n(y,x)|\,\mathrm{d}y\, P_X(\mathrm{d}x)\\&\quad= \mathbf{E}\int |\bar{f}_n(y,X) - f_n(y,X)|\,\mathrm{d}y\\&\quad\le \mathbf{E}\left\{\frac{\frac{1}{h_n}\sum_{i=1}^{N_n} G\left(\frac{\|X-X_i\|}{H_n}\right)\sum_{k=1}^{l_{i,n}}\int\left|K\left(\frac{y-\bar{Y}_k^{(i)}}{h_n}\right) - K\left(\frac{y-Y_k^{(i)}}{h_n}\right)\right|\mathrm{d}y}{\sum_{j=1}^{N_n} l_{j,n}\cdot G\left(\frac{\|X-X_j\|}{H_n}\right)}\right\}\\&\quad\le 2\cdot K(0)\cdot\mathbf{E}\left\{\frac{\sum_{i=1}^{N_n} G\left(\frac{\|X-X_i\|}{H_n}\right)\sum_{k=1}^{l_{i,n}}\left|\bar{Y}_k^{(i)} - Y_k^{(i)}\right|}{h_n\cdot\sum_{j=1}^{N_n} l_{j,n}\cdot G\left(\frac{\|X-X_j\|}{H_n}\right)}\right\}.\end{aligned}$$

The assertion follows by assumption (A8). □

4.5 Proof of Theorem 4

We know from Theorem 2 and the proof of Theorem 3 that

$$\begin{aligned}
&\mathbf{E}\int\int |\bar{f}_n(y,x) - f(y,x)|\,\mathrm{d}y\, P_X(\mathrm{d}x)\\
&\le \frac{c_{42}}{\sqrt{N_n \cdot l_n \cdot h_n \cdot H_n^d}} + c_{43}\cdot H_n^\alpha + c_{44}\cdot h_n^r + \frac{c_{45}}{N_n \cdot H_n^d}\\
&\quad + c_{46}\cdot \mathbf{E}\left\{\frac{\sum_{i=1}^{N_n} G\left(\frac{\|X-X_i\|}{H_n}\right)\sum_{k=1}^{l_n}\left|\bar{Y}_k^{(i)} - Y_k^{(i)}\right|}{h_n \cdot \sum_{j=1}^{N_n} l_n \cdot G\left(\frac{\|X-X_j\|}{H_n}\right)}\right\}.
\end{aligned}$$

Due to condition $(A8')$ it holds

$$\begin{aligned}
&\mathbf{E}\left\{\frac{\sum_{i=1}^{N_n} G\left(\frac{\|X-X_i\|}{H_n}\right)\sum_{k=1}^{l_n}\left|\bar{Y}_k^{(i)} - Y_k^{(i)}\right|}{h_n \cdot \sum_{j=1}^{N_n} l_n \cdot G\left(\frac{\|X-X_j\|}{H_n}\right)}\right\}\\
&= \mathbf{E}\left\{\frac{\sum_{i=1}^{N_n} G\left(\frac{\|X-X_i\|}{H_n}\right)\sum_{k=1}^{l_n}\mathbf{E}\left\{|\bar{Y}_k^{(i)} - Y_k^{(i)}|\middle|X, X_i\right\}}{h_n \cdot \sum_{j=1}^{N_n} l_n \cdot G\left(\frac{\|X-X_j\|}{H_n}\right)}\right\}\\
&\le \mathbf{E}\left\{\frac{\sum_{i=1}^{N_n} G\left(\frac{\|X-X_i\|}{H_n}\right)\cdot c_{15}\cdot l_n \cdot \frac{h_n}{\delta_n}}{h_n \cdot \sum_{j=1}^{N_n} l_n \cdot G\left(\frac{\|X-X_j\|}{H_n}\right)}\right\}\\
&\le c_{15}\cdot \delta_n^{-1}.
\end{aligned}$$

The proof is complete. □

Acknowledgements The authors would like to thank the German Research Foundation (DFG) for funding this project within the Collaborative Research Center 666, and two anonymous referees and an associate editor for detailed and very helpful comments improving an early version of the manuscript.

References

Bott, A., Felber, T., Kohler, M. (2013). Estimation of a density in a simulation model. *Journal of Nonparametric Statistics, 2015.*

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

Devroye, L. (1987). *A course in density estimation*. Basel: Birkhäuser.

Devroye, L. (2015). Personal communication.

Devroye, L., Györfi, L. (1985). *Nonparametric density estimation. The L1 view. Wiley series in probability and mathematical statistics: Tracts on probability and statistics*. New York: Wiley.

Devroye, L., Lugosi, G. (2001). *Combinatorial methods in density estimation*. New York: Springer.

Efromovich, S. (2007). Conditional density estimation in a regression setting. *Annals of Statistics*, *35*, 2504–2535.

Fan, J., Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika*, *91*, 819–834.

Fan, J., Yao, Q., Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, *83*, 189–206.

Furer, D., Kohler, M. (2013). Smoothing spline regression estimation based on real and artificial data. *Metrika, 2015*.

Gooijer, J. G. D., Zerom, D. (2003). On conditional density estimation. *Statistica Neerlandica*, *57*, 159–176.

Györfi, L., Kohler, M. (2007). Nonparametric estimation of conditional distributions. *IEEE Transactions on Information Theory*, *53*, 1872–1879.

Györfi, L., Kohler, M., Krzyżak, A., Walk, H. (2002). *A distribution-free theory of nonparametric regression*. New York: Springer.

Hyndman, R. J., Bashtannyk, D. M., Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, *5*, 315–336.

Manson, S. S. (1965). Fatigue: A complex subject—some simple approximation. *Experimental Mechanics*, *5*, 193–226.

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

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

Rosenblatt, M. (1969). Conditional probability density and regression estimates. In P. R. Krishnaiah (Ed.), *Multivariate analysis II* (pp. 25–31). New York: Academic Press.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice and visualization*. New York: Wiley.

Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, *5*, 595–645.

Williams, C. R., Lee, Y.-L., Rilly, J. T. (2002). A practical method for statistical analysis of strain-life fatigue data. *International Journal of Fatigue*, *25*, 427–436.