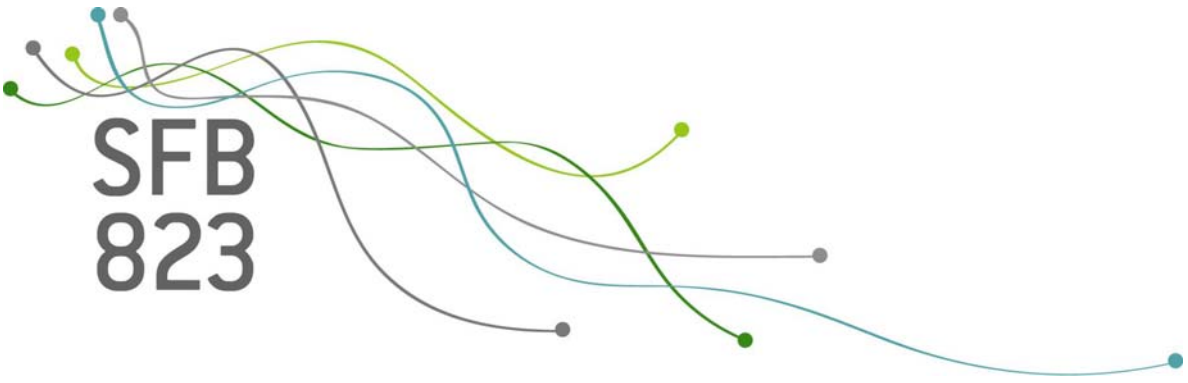
SFB 823

Discussion Paper

Smooth backfitting in additive inverse regression

Nicolai Bissantz, Holger Dette,
Thimo Hildebrandt

Nr. 37/2013



Smooth backfitting in additive inverse regression

Nicolai Bissantz, Holger Dette, Thimo Hildebrandt
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum
Germany
email: nicolai.bissantz@ruhr-uni-bochum.de
holger.dette@ruhr-uni-bochum.de
thimo.hildebrandt@ruhr-uni-bochum.de
FAX: +49 234 32 14559

October 8, 2013

Abstract

We consider the problem of estimating an additive regression function in an inverse regression model with a convolution type operator. A smooth backfitting procedure is developed and asymptotic normality of the resulting estimator is established. Compared to other methods for the estimation in additive models the new approach neither requires observations on a regular grid nor the estimation of the joint density of the predictor. It is also demonstrated by means of a simulation study that the backfitting estimator outperforms the marginal integration method at least by a factor two with respect to the integrated mean squared error criterion.

Keywords: inverse regression; additive models; curse of dimensionality; smooth backfitting
Mathematical subject classification: Primary: 62G20; Secondary 15A29

1 Introduction

In this paper we consider the regression model

$$Y_k = g(\mathbf{X}_k) + \varepsilon_k \qquad k \in \{1, ..., N\}, \tag{1.1}$$

where $\varepsilon_1, ..., \varepsilon_N$ are independent identically distributed random variables and $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are independent identically distributed d-dimensional predictors with components $\mathbf{X}_k = (X_{k,1}, \ldots, X_{k,d})^T$ $(k = 1, \ldots, N)$. We assume that the function g is related to a signal θ by a convolution type operator, that is

$$g(\mathbf{z}) = \int_{\mathbb{R}^d} \psi(\mathbf{z} - \mathbf{t})\theta(\mathbf{t})d\mathbf{t}, \tag{1.2}$$

where $\psi : \mathbb{R}^d \to \mathbb{R}$ is a known function with $\int_{\mathbb{R}^d} \psi(\mathbf{t})d\mathbf{t} = 1$. The interest of the experiment is the nonparametric estimation of the signal θ. Models of the type (1.1) and (1.2) belong to the class of inverse regression models and have important applications in the recovery of images from astronomical telescopes or fluorescence microscopes in biology. Deterministic inverse regression models have been considered for a long time in the literature [Engl et al. (1996), Saitoh (1997)]. However, in the last decade statistical inference in ill-posed problems has become a very active field of research [see Bertero et al. (2009), Kaipio and Somersalo (2010) for a Bayesian approach and Mair and Ruymgaart (1996), Cavalier (2008) and Bissantz et al. (2007) for nonparametric methods].

While most of these methods have been developed for models with a one-dimensional predictor, nonparametric estimation in the multivariate setting is of practical importance because in many applications one has to deal with an at least two-dimensional predictor. A typical example is image reconstruction since a picture is a two-dimensional object. Also in addition to the spatial dimensions, the data might depend on the time thus introducing a third component. For a multivariate predictor the estimation of the signal θ in the inverse regression model (1.1) is a much harder problem due to the curse of dimensionality. In direct regression usually qualitative assumptions regarding the signal such as additivity or multiplicativity are made, which allow the estimation of the regression function at reasonable rates [see Linton and Nielsen (1995), Mammen et al. (1999), Carroll et al. (2002), Hengartner and Sperlich (2005), Nielsen and Sperlich (2005)]. In the present paper we investigate the problem of estimating the signal θ in the inverse regression model with a convolution type operator under the additional assumption of additivity, that is

$$\theta(\mathbf{x}) \;=\; \theta_0 + \theta_1(x_1) + ... + \theta_d(x_d), \tag{1.3}$$

where $\mathbf{x} = (x_1, \ldots, x_d)^T$. In a recent paper Hildebrandt et al. (2013) proposed an estimator of the signal θ if observations are available on a regular grid in $\mathbb{R}^d$. They also considered the case of a random design and investigated the statistical properties of a marginal integration type estimate with known density of the predictor. The asymptotic analysis of both estimates is based on these rather restrictive assumptions regarding the predictor $\mathbf{X}$. A regular grid or explicit knowledge of the density of the predictor $\mathbf{X}$ might not be available in all applications. Moreover, estimation of this density in the marginal integration method cannot be performed at one-dimensional rates [see Hildebrandt et al. (2013)]. In particular it changes the asymptotic properties of additive estimates such that the signal cannot be reconstructed with one-dimensional nonparametric rates.

In the present paper we consider the construction of an estimate in the inverse additive regression model (1.3) with random design, which is applicable under less restrictive assumptions in particular without knowledge of the density of the predictor. For this purpose we combine in Section 2 smooth backfitting [see Mammen et al. (1999)] with Fourier estimation methods in inverse regression models [see Diggle and Hall (1993) or Mair and Ruymgaart (1996)]. Besides several advantages of the smooth backfitting approach observed in the literature in direct regression models [see Nielsen and Sperlich (2005)], the backfitting methodology only requires the estimation of the marginal densities of the predictor. As a consequence, the resulting estimate does not suffer from the curse of dimensionality. Section 3 is devoted to the investigation of the asymptotic properties of the new estimator, while we study the finite sample properties by means of a simulation study in Section 4. In particular we demonstrate that the smooth backfitting approach results in estimates with an at least two times smaller integrated mean squared error than the marginal integration method. Finally, all proofs and technical arguments are presented in Section 5.

2 Smooth backfitting in inverse regression

Note that the linearity of the convolution operator and assumption (1.3) imply that the function g is also additive, and consequently the model (1.1) can be rewritten as

$$Y_k \;=\; g_0 + g_1(X_{k,1}) + ... + g_d(X_{k,d}) + \varepsilon_k, \tag{2.1}$$

where $\mathbf{X}_k = (X_{k,1}, \ldots, X_{k,d})^T$ and the functions $g_0, g_1, \ldots, g_d$ in model (2.1) are related to the components $\theta_0, \theta_1, \ldots, \theta_d$ of the signal θ in model (1.3) by $g_0 = \theta_0$,

$$g_j(z_j) = \int_{\mathbb{R}} \psi_j(z_j - t)\theta_j(t)dt \qquad j = 1, \ldots, d. \tag{2.2}$$

Here ψ_j is the marginal of the convolution function ψ, that is

$$\psi_j(t_j) = \int_{\mathbb{R}^{d-1}} \psi(\mathbf{t})d\mathbf{t}_{-\mathbf{j}} \tag{2.3}$$

and $\mathbf{t} = (t_1, ..., t_d)^T \in \mathbb{R}^d, \mathbf{t}_{-\mathbf{j}} = (t_1, \ldots, t_{j-1}, t_{j+1}, \ldots, t_d)^T \in \mathbb{R}^{d-1}$. The estimation of the additive signal is now performed in several steps and combines Fourier transform estimation methods for inverse regression models [see Diggle and Hall (1993) or Mair and Ruymgaart (1996)] with the smooth backfitting technique developed for direct nonparametric regression models [see Mammen et al. (1999)].

(1) We assume for a moment that the design density is known and denote by f_j and F_j the density and cumulative distribution function of the jth marginal distribution of the random variable $\mathbf{X}$. In a first step all explanatory variables are transformed to the unit cube by

using the probability transformation in each component, that is

$$Z_{k,j} = F_j(X_{k,j}) \qquad j = 1, \ldots, d; \quad k = 1, \ldots, N. \tag{2.4}$$

This transformation is necessary because of two reasons. On the one hand, the asymptotic analysis of methods based on Fourier estimation requires with positive probability observations at points $\mathbf{X}_k$ with a norm $\| \mathbf{X}_k \|$ converging to infinity, because one has to estimate the Fourier transform of the function g_j on the real axis. On the other hand, the asymptotic analysis of the smooth backfitting method requires a distribution of the explanatory variables with a compact support.
In practice the unknown marginal distributions of the predictor are estimated by standard methods and this estimation does not change the asymptotic properties of the statistic. We refer to Remark 2.1 for more details.

(2) The transformation in Step (1) yields the representation

$$Y_k = g_0 + g_1^*(Z_{k,1}) + \cdots + g_d^*(Z_{k,d}) + \varepsilon_k\,; \qquad k = 1, \ldots, N, \tag{2.5}$$

where the functions g_j^* are defined by $g_j^* = g_j \circ F_j^{-1}$ $(j = 1, \ldots, d)$. We now use the smooth backfitting algorithm [see Mammen et al. (1999)] to estimate each function g_j^* in (2.5) from the data $(Z_{1,1}, ..., Z_{1,d}, Y_1), \ldots, (Z_{N,1}, ..., Z_{N,d}, Y_N)$. This algorithm determines estimates of the components $g_0, g_1^*, \ldots, g_d^*$ recursively, where $\hat{g}_0 = \overline{Y}_\cdot = \frac{1}{N}\sum_{k=1}^N Y_k$. For starting values $\hat{g}_1^{*(0)}, \ldots, \hat{g}_d^{*(0)}$ we calculate for $r = 1, 2, \ldots$ the estimators $\hat{g}_1^{*(r)}, \ldots, \hat{g}_d^{*(r)}$ by the recursive relation

$$\begin{aligned} \hat{g}_j^{*(r)}(z_j) \;=\; \hat{g}_j^*(z_j) &- \sum_{k<j} \int \hat{g}_k^{*(r)}(z_k) \Big[\frac{\hat{p}_{jk}(z_j, z_k)}{\hat{p}_j(z_j)} - \hat{p}_{k,[j+]}(z_k) \Big] dz_k \\ &- \sum_{k>j} \int g_k^{*(r-1)}(z_k) \Big[\frac{\hat{p}_{jk}(z_j, z_k)}{\hat{p}_j(z_j)} - \hat{p}_{k,[j+]}(z_k) \Big] dz_k - g_{0,j}^* \;. \end{aligned} \tag{2.6}$$

Here

$$\hat{g}_j^*(z_j) = \frac{\sum_{k=1}^N L\Big(\frac{Z_{k,j} - z_j}{h_B}\Big) Y_k}{\sum_{k=1}^N L\Big(\frac{Z_{k,j} - z_j}{h_B}\Big)} \tag{2.7}$$

denotes the one-dimensional Nadaraya-Watson estimator of the jth component (with kernel L and bandwidth h_B), $\hat{p}_{jk}$ and $\hat{p}_j$ are the (j,k)th and jth marginals of the common kernel

density estimator $\hat{p}$ for the density p of the predictor $(Z_1, ..., Z_d)^T$, and we use the notation

$$\begin{aligned} \hat{p}_{k,[j+]}(z_k) &= \int \hat{p}_{jk}(z_j, z_k)dz_j \Big[\int \hat{p}_j(z_j)dz_j \Big]^{-1} \\ g^*_{0,j} &= \frac{\int \hat{g}^*_j(z_j)\hat{p}_j(z_j)dz_j}{\int \hat{p}_j(z_j)dz_j}. \end{aligned} \tag{2.8}$$

(3) Estimators of the functions g_j in (2.1) are now easily obtained by the transformation

$$\hat{g}_j = \hat{g}_j^{*(r_0)} \circ F_j, \tag{2.9}$$

where $\hat{g}_j^{*(r_0)}$ denotes the estimator obtained after terminating the recursive relation (2.6) at step r_0 ($j = 1, \ldots, d$). In order to recover the signal θ_j from $\hat{g}_j$ we now introduce the random variables

$$U_{k,j} = Y_k - \sum_{\substack{i=1 \\ i \neq j}}^{d} \hat{g}_i(X_{k,i}) - \hat{g}_0 \tag{2.10}$$

and use the data $(X_{1,j}, U_{1,j}), \ldots, (X_{N,j}, U_{N,j})$ to estimate the jth component θ_j of the signal θ by Fourier transform estimation methods [see Diggle and Hall (1993) for example]. For this purpose we note that the relation (2.2) implies for the Fourier transforms Φ_{g_j} and Φ_{θ_j} of the functions g_j and θ_j the relation

$$\Phi_{\theta_j} = \frac{\Phi_{g_j}}{\Phi_{\psi_j}},$$

where

$$\Phi_{\psi_j}(w) = \int_{\mathbb{R}} \psi_j(x) e^{iwx} dx$$

is the Fourier transform of the jth marginal of the convolution function. Now the Fourier transform $\Phi_{g_j}(w)$ of the function g_j is estimated by its empirical counterpart

$$\hat{\Phi}_{g_j}(w) = \frac{1}{N} \sum_{k=1}^{N} e^{iwX_{k,j}} \frac{U_{k,j}}{\max\{f_j(X_{k,j}), f_j(\frac{1}{a_N})\}}, \tag{2.11}$$

where f_j is the density of the jth marginal distribution and a_N is a real valued sequence converging to 0 as $N \to \infty$. The estimator of $\hat{\theta}_j$ is now obtained from a "smoothed" inversion of the Fourier transform, that is

$$\hat{\theta}_j(x_j) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-iwx_j} \Phi_K(hw) \frac{\hat{\Phi}_{g_j}(w)}{\Phi_{\psi_j}(w)} dw, \tag{2.12}$$

where Φ_K is the Fourier transform of a kernel K and h is a bandwidth converging to 0 with increasing sample size.

(4) Finally, the additive estimate of the signal θ is given by

$$\hat{\theta}(\mathbf{x}) = \hat{\theta}_0 + \hat{\theta}_1(x_1) + ... + \hat{\theta}_d(x_d), \tag{2.13}$$

where $\hat{\theta}_0 = \hat{g}_0 = \overline{Y}$. and $\hat{\theta}_j$ is defined in (2.12) for $j = 1, \ldots, d$.

Remark 2.1

(a) Note that we use the term $\max\{f_j(X_{k,j}), f_j(\frac{1}{a_N})\}$ in the denominator of the estimate (2.11) instead of the more intuitive term $f_j(X_{k,j})$. This "truncation" avoids situations where the denominator becomes too small, which would yield unstable estimates with a too large variance.

(b) In practical applications knowledge of the marginal distributions might not be available and in this case the transformation (2.4) can be achieved by

$$\hat{Z}_{k,j} = \hat{\mathbb{F}}_j(X_{k,j}); \qquad j = 1, \ldots, d; \quad k = 1, \ldots, N, \tag{2.14}$$

where for $j = 1, \ldots, d$

$$\hat{\mathbb{F}}_j(x) = \frac{1}{N+1} \sum_{k=1}^{N} \mathbb{I}\{X_{k,j} \leq x\}$$

denotes the empirical distribution function of the jth components $X_{1,j}, ..., X_{N,j}$. Similarly, the density f_j in (2.11) can be estimated by kernel density methods, that is

$$\hat{f}_j(x_j) = \frac{1}{N h_{d,j}} \sum_{k=1}^{N} M\Big(\frac{X_{k,j} - x_j}{h_{d,j}}\Big); \qquad j = 1, \ldots, d, \tag{2.15}$$

where M denotes a kernel and $h_{d,j}$ is a bandwidth proportional to $N^{-1/5}$. We note that the estimators $\hat{\mathbb{F}}_j$ and $\hat{f}_j$ converge uniformly to F_j and f_j at rates $(\frac{\log\log N}{N})^{1/2}$ and $(\frac{\log N}{N h_{d,j}})^{1/2}$, respectively [see van der Vaart (1998), Giné and Guillou (2002)]. The rates of convergence in inverse deconvolution problems are slower and consequently the asymptotic properties of the estimates $\hat{\theta}_j$ do not change if f_j and F_j are replaced by their empirical counterparts $\hat{f}_j$ and $\hat{\mathbb{F}}_j$ defined in (2.14) and (2.15), respectively.

3 Asymptotic properties

In this section we investigate the asymptotic properties of the estimators defined in Section 2. In particular we establish weak convergence. For this purpose we require the following assumptions

(A1) The kernel L in the Nadaraya-Watson estimator $\hat{g}_j^*$ in the backfitting recursion (2.6) is symmetric, Lipschitz continuous and has compact support, say $[-1,1]$. The bandwidth h_B of this estimator is proportional to $N^{-1/5}$.

(A2) $\mathbb{E}[|Y_j|^\alpha] < \infty$ for some $\alpha > \frac{5}{2}$.

(A3) The functions $g_1, \ldots, g_d$ in model (2.1) are bounded and twice differentiable with Lipschitz continuous second order derivatives.

(A4) The Fourier transforms Φ_{ψ_j} of the marginals ψ_j of the convolution function ψ satisfy

$$\int_{\mathbb{R}} \frac{|\Phi_K(w)|}{|\Phi_{\psi_j}(\frac{w}{h})|} dw \leq C_1 h^{-\beta_j}, \qquad \int_{\mathbb{R}} \frac{|\Phi_K(w)|^2}{|\Phi_{\psi_j}(\frac{w}{h})|^2} dw \sim C_2 h^{-2\beta_j},$$
$$\left| \frac{1}{h} \int \int e^{-iw(x-x_j)/h} \frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})} dw \frac{f_j(x)}{\max\{f_j(x), f_j(\frac{1}{a_N}\}} dx \right| = o(h^{-2\beta-1})$$

uniformly with respect to x_j for some constants $\beta_j > 0$ $(j = 1, \ldots, d)$ and constants $C_1, C_2, C_3 > 0$, where the constant C_3 does not depend on x_j.

(A5) The Fourier transform Φ_K of the kernel K is symmetric and supported on the interval $[-1,1]$. Additionally there exists a constant $b \in (0,1]$ such that $\Phi_K(w) = 1$ for all $w \in [-b,b], b > 0$, and $|\Phi_K(w)| \leq 1$ for all $w \in \mathbb{R}$

(A6) The Fourier transforms $\Phi_{\theta_1}, \ldots, \Phi_{\theta_d}$ of the functions $\theta_1, \ldots, \theta_d$ in the additive model (1.3) satisfy

$$\int_{\mathbb{R}} |\Phi_{\theta_j}(w)||w|^{s-1} dw < \infty \quad \text{for some} \quad s > 1 \text{ and } j = 1, ..., d.$$

(A7) The functions $g_1, ..., g_d$ defined in model (2.2) satisfy

$$\int_{\mathbb{R}} |g_j(z)||z|^r dz < \infty \quad \text{for} \quad j = 1, ..., d$$

for some $r > 0$ such that $a_N^{r-1} = o(h^{\beta_j + s})$.

(A8) For each $N \in \mathbb{N}$ let $\mathbf{X}_1, ..., \mathbf{X}_N$ denote independent identically distributed d-dimensional random variables with marginal densities $f_1, ..., f_d$ (which may depend on N) such that $f_j(x) \neq 0$ for all $x \in [-\frac{1}{a_N}, \frac{1}{a_N}]$. We also assume that F_j^{-1} exists, where F_j is the distribution function of $X_{1,j}$. Furthermore we assume, that for sufficiently large $N \in \mathbb{N}$

$$f_j(x) \geq f_j(\frac{1}{a_N}) \quad \text{whenever} \quad x \in [-\frac{1}{a_N}, \frac{1}{a_N}],$$

for all $j = 1, \ldots, d$.

(A9) If $f_{ijk}(t_i, t_j|t_k)$ and $f_{ij}(t_i|t_j)$ denote the densities of the conditional distribution $\mathbb{P}^{X_i,X_j|X_k}$ and $\mathbb{P}^{X_i|X_j}$, respectively, we assume that there exist integrable functions (with respect to the Lebesgue measure), say $U_{ijk} : \mathbb{R}^2 \to \mathbb{R}$, $\eta_{ij} : \mathbb{R} \to \mathbb{R}$, such that the inequalities

$$f_{ijk}(t_i, t_j|t_k) \leq U_{ijk}(t_i, t_j) \ ; \ f_{ij}(t_i|t_j) \leq \eta_{ij}(t_i)$$

are satisfied for all $t_i, t_j, t_k \in \mathbb{R}$.

Remark 3.1 Assumption (A1) - (A3) are required for the asymptotic analysis of the backfitting estimator, while (A4) - (A8) are used to analyze the Fourier estimation methods used in the second step of the procedure. In order to demonstrate that these assumptions are satisfied in several cases of practical importance we consider exemplarily Assumption (A4) and (A6).

(a) To illustrate Assumption (A4) the convolution function ψ and the kernel K are chosen as

$$\psi_j(x) \ = \ \frac{\lambda}{2} e^{-\lambda|x|}; \qquad K(x) = \frac{\sin(x)}{\pi x},$$

respectively. Furthermore we choose f_j as density of a uniform distribution on the interval $[-\frac{1}{a_N}, \frac{1}{a_N}]$ and consider exemplarily the point $x_j = 0$. Note that $\Phi_K(w) = \mathbb{I}_{[-1,1]}(w)$. The integrals in (A4) are obtained by straightforward calculation, that is

$$\int_{\mathbb{R}} \frac{|\Phi_K(w)|}{|\Phi_{\psi_j}(\frac{w}{h})|} dw \ = \ \int_{[-1,1]} \left(1 + \frac{w^2}{h^2}\right) dw = \frac{2}{3h^2} + 2$$

$$\int_{\mathbb{R}} \frac{|\Phi_K(w)|^2}{|\Phi_{\psi_j}(\frac{w}{h})|^2} dw \ = \ \int_{[-1,1]} \left(1 + \frac{w^2}{h^2}\right)^2 dw = \frac{2}{5h^4} + \frac{4}{3h^2} + 2$$

$$\begin{aligned}
& \frac{1}{h} \int_{[-1/a_N, 1/a_N]} \int_{[-1,1]} e^{-iw(x-x_j^*)/h} \frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})} dw \frac{f_j(x)}{\max\{f_j(x), f_j(\frac{1}{a_N}\}} dx \\
= \ & \frac{2}{h} \int_{[-1/a_N, 1/a_N]} \frac{((h^2(x^2-2) + x^2)\sin(\frac{x}{h}) + 2hx\cos(\frac{x}{h}))}{hx^3} dx \\
= \ & \frac{-2a_N \cos(\frac{1}{a_N h}) + 2a_N^2 h \sin(\frac{1}{a_N h}) + 2hSi(\frac{1}{a_N h})}{h}
\end{aligned}$$

and $Si(x)$ denotes the sine-integral $\int_0^x \frac{\sin(y)}{y} dy$. This shows that condition (A4) is satisfied.

(b) In order to illustrate Assumption (A6) let $W^m(\mathbb{R})$ denote the Sobolev space of order $m \in \mathbb{N}$, then the assumption $\theta_j \in W^s(\mathbb{R})$ with $s \in \mathbb{N}\backslash\{1\}$ implies condition (A6). Conversely, if (A6) holds with $s \in \mathbb{N}\backslash\{1\}$, then θ_j is $(s-1)$ times continuously differentiable [see Folland (1984)]. In other words, (A6) is an assumption regarding the smoothness of the components of the signal θ_j $(j = 1, \ldots, d)$.

Our main result, which is proved in the Appendix, establishes the weak convergence of the estimator $\hat{\theta}_j$ for the jth component of the additive signal in model (1.3). Throughout this paper the symbol $\Rightarrow$ denotes weak convergence.

Theorem 3.2 *Consider the additive inverse regression model defined by* (1.1) - (1.3). *If Assumptions (A1) - (A8) are satisfied and additionally the conditions*

$$N^{1/2}h^{\beta_j+1/2}f_j(\frac{1}{a_N})^{1/2} \to \infty \tag{3.1}$$

$$N^{1/2}h^{3/2}f_j(\frac{1}{a_N})^3 \to \infty, \quad N^{1/5}h^{s+\beta_j}f_j(\frac{1}{a_N}) \to \infty \tag{3.2}$$

are fulfilled, then a standardized version of the estimator $\hat{\theta}_j$ defined in (2.12) *converges weakly, that is*

$$V_{N,j}^{-1/2}\left(\hat{\theta}_j(x_j) - \mathbb{E}[\hat{\theta}_j(x_j)]\right) \Rightarrow \mathcal{N}(0,1),$$

where

$$\mathbb{E}[\hat{\theta}_j(x_j)] = \theta_j(x_j) + o(h^{s-1}),$$

and the normalizing sequence is given by

$$V_{N,j} = \frac{1}{Nh^2(2\pi)^2}\int_{\mathbb{R}}\left|\int_{\mathbb{R}} e^{-iw(x_j-y)/h}\frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})}dw\right|^2 \frac{(g_j^2(y)+\sigma^2)f_j(y)}{\max\{f_j(y), f_j(\frac{1}{a_N})\}^2}dy \tag{3.3}$$

and satisfies

$$N^{1/2}h^{\beta_j+1/2}f_j\left(\frac{1}{a_N}\right)^{1/2} \leq V_{N,j}^{-1/2} \leq N^{1/2}h^{\beta_j+1/2}. \tag{3.4}$$

As a consequence of Theorem 3.2 we obtain the weak convergence of the additive estimate $\hat{\theta}$ of the signal θ.

Remark 3.3 If all components except one would be known, it follows from Theorem 3.1 in Hildebrandt et al. (2013) that this component can be estimated at a rate R_N satisfying

$$\frac{c_1}{N^{1/2}h^{1/2+\beta_j}} \leq R_n \leq \frac{c_2}{N^{1/2}h^{1/2+\beta_j}f_j(a_N^{-1})}$$

(with appropriate constants c_1 and c_2). Consequently, it follows from Theorem 3.2 that the smooth backfitting operator $\hat{\theta}_j$ defined in (2.12) has an oracle property and estimates the jth component at the one-dimensional rate.

Corollary 3.4 *Consider the inverse regression model defined by* (1.1) - (1.3) *and assume that the assumptions of Theorem 3.2 are satisfied for all $j = 1, \ldots, d$. Then a standardized version of the*

the additive estimator $\hat{\theta}$ *defined in* (2.13) *converges weakly, that is*

$$V_N^{-1/2}\left(\hat{\theta}(x) - \mathbb{E}[\hat{\theta}(x)]\right) \Rightarrow \mathcal{N}(0,1).$$

Here

$$\mathbb{E}[\hat{\theta}(x)] = \theta(x) + o(h^{s-1}),$$

and the normalizing factor is given by $V_N = \sum_{j=1}^d V_{N,j} + \sum_{1\le k\neq l\le d} V_{N,k,l}$*, where* $V_{N,j}$ *is defined in* (3.3)*,*

$$\begin{aligned} V_{N,k,l} \;=\; & \frac{1}{Nh^2(2\pi)^2}\int_{\mathbb{R}}\int_{\mathbb{R}} e^{-iw(x_k-y)/h}\frac{\Phi_K(w)}{\Phi_{\psi_k}(\frac{w}{h})}dw\overline{\int_{\mathbb{R}} e^{-iw(x_l-z)/h}\frac{\Phi_K(w)}{\Phi_{\psi_l}(\frac{w}{h})}dw} \\ & \times\frac{(\sigma^2 + g_k(y)g_l(z))f_{k,l}(y,z)}{\max\{f_k(y), f_k(\frac{1}{a_N})\}\max\{f_l(y), f_l(\frac{1}{a_N})\}}d(y,z), \end{aligned}$$

and $f_{k,l}$ *denotes the joint density of the pair* $(X_{k,1}, X_{l,1})$*. Moreover* V_N *satisfies*

$$N^{1/2}h^{\beta_{j^*}+1/2}f_{j^*}\left(\frac{1}{a_N}\right)^{1/2} \le V_N^{-1/2} \le N^{1/2}h^{\beta_{j^*}+1/2}.$$

where $j^* = argmin_j h^{\beta_j} f_j(1/a_N)$*.*

4 Finite sample properties

In this section we briefly investigate the finite sample properties of the new backfitting estimators by means of a small simulation study. We also compare the two estimators obtained by the marginal integration method with the backfitting estimator proposed in this paper. All results are based on 500 simulation runs. For the sake of brevity we concentrate on three models with a two-dimensional predictor and two distributions for the predictor. To be precise we consider the models

$$\theta(x_1,x_2) \;=\; \theta_1(x_1) + \theta_2(x_2) = e^{-(x_1-0.4)^2} + e^{-(x_2-0.1)^2}, \tag{4.1}$$

$$\theta(x_1,x_2) \;=\; \theta_1(x_1) + \theta_2(x_2) = x_1 e^{-|x_1|} + (1+x_2^2)^{-1}, \tag{4.2}$$

$$\theta(x_1,x_2) \;=\; \theta_1(x_1) + \theta_2(x_2) = e^{-|x_1|} + (1+x_2^2)^{-1}, \tag{4.3}$$

and assume that the convolution function is given by

$$\psi(x_1,x_2) = \frac{9}{4}e^{-3(|x_1|+|x_2|)}. \tag{4.4}$$

Note that the signals in (4.1) and (4.2) satisfy the assumptions posed in Section 3, while this is not the case for the first component of the signal (4.3). For the distribution of the explanatory

variable we consider an independent and correlated case, that is

(4.5) a uniform distribution on the square $[1/a_N, 1/a_N]^2$

(4.6) a two-dimensional normal distribution with mean $\mathbf{0}$ and variance $\Sigma = \begin{pmatrix} 1 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix}$

The sample size is $N = 701$, the variance is given by $\sigma^2 = 0.25$ and for the sequence a_N we used 0.5. In the simulation the bandwidths are chosen in several (nested) steps. At first the bandwidths $h_{d,j}$ in (2.15) are calculated minimizing the mean integrated squared error of the density estimate. These bandwidths are used in the calculation of the mean integrated squared error of the estimate $\hat{g}_j$ in (2.9), which is then minimized with respect to the choice of h_B. The final step consists of a calculation of the bandwidth h minimizing the mean integrated squared error of the resulting inverse Fourier transform (2.12). In practice this procedure of the mean squared error requires knowledge of the quantities f_j, g_j and for a concrete application we recommend to mimic these calculations by cross validation.
In Figures 1 - 3 we present the estimated mean curves for both components corresponding to model (4.1) - (4.3) respectively. Upper parts of the tables show the results for independent components of the predictor, where the case of correlated explanatory variables is displayed in the lower panels. The figures also contain the (pointwise) estimated 5% and 95%-quantile curves to illustrate the variation of the estimators. We observe that in models (4.1) and (4.2) both components are estimated with reasonable precision [see Figure 1 and 2]. The estimators are slightly more accurate under the assumption of an independent design where the differences are more substantial for the estimators of the second component. The differences between the uncorrelated and correlated case are even more visible for model (4.3), for which the results are displayed in Figure 3. Here we observe that the first component is not estimated accurately in a neighborhood of the origin. This is in accordance with our theoretical analysis, because the first component in model (4.3) does not satisfy the assumptions made in Section 3. Consequently, the resulting estimates of the first component are biased in a neighbourhood of the origin. On the other hand, the second component satisfies these assumptions and the right panels of Figure 3 show that the second component can be estimated with similar precision as in model (4.1) and (4.2).
In order to compare the new method with the marginal integration method proposed in Hildebrandt et al. (2013) we finally display in Table 1 the simulated integrated mean squared error of both estimators for the models (4.1) - (4.3). We observe in the case of independent predictors that the backfitting approach yields an improvement of 50% with respect to the integrated mean squared error criterion. Moreover, in the situation of dependent predictors as considered in (4.6) the improvement is even more substantial and varies between a factor 3 and 4. We expect that the advantages of the backfitting methodology are even larger with an increasing dimension of the predictor $\mathbf{X}$.

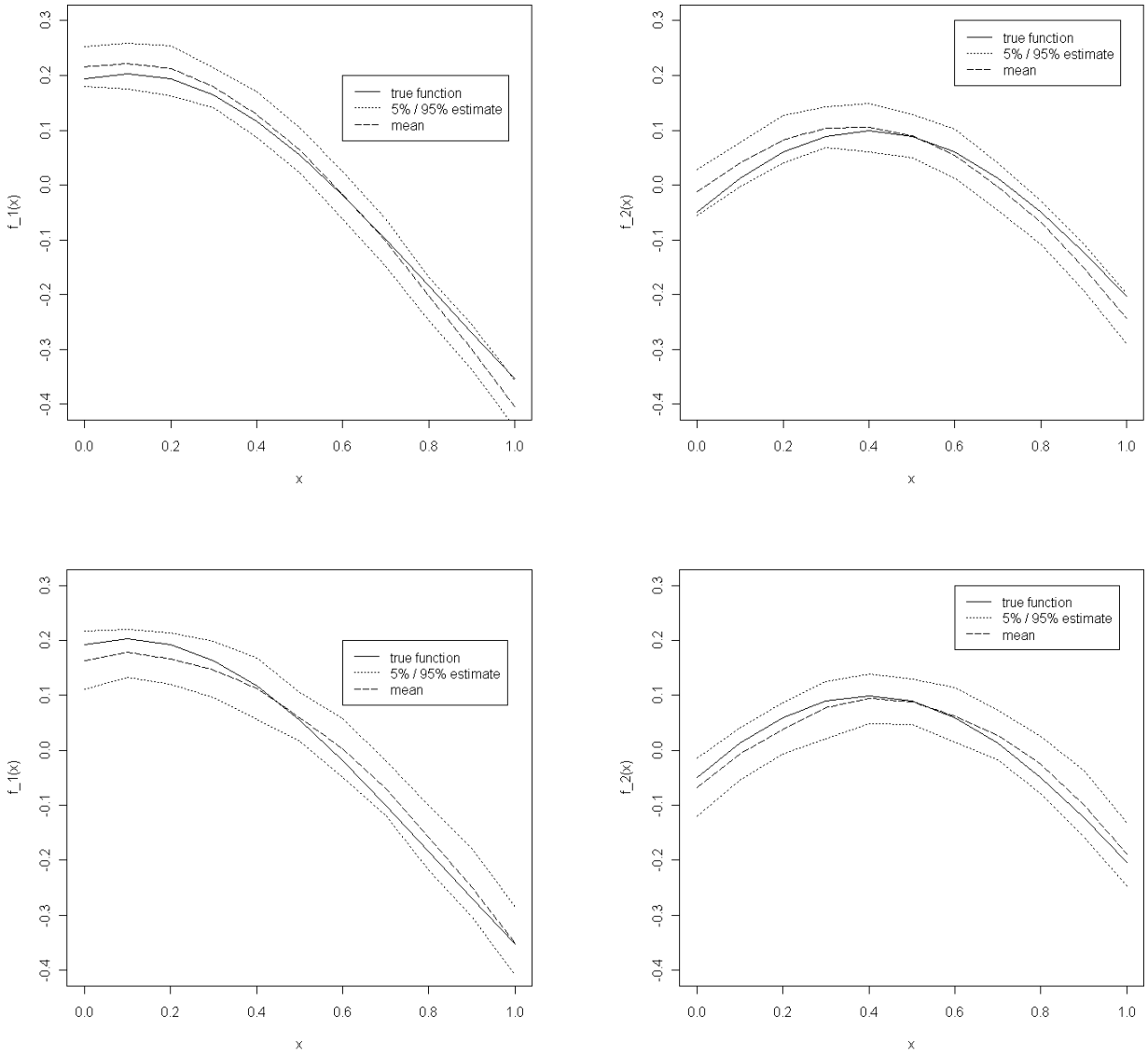


Figure 1: *Simulated mean,* 5%- *and* 95% *quantile of the backfitting estimate on the basis of* 500 *simulation runs, where model is given by* (4.1) *and the design is given by* (4.5) *(upper panel) and* (4.6) *(lower panel). Left part* θ_1*; right part:* θ_2*.*

Acknowledgements. The authors thank Martina Stein and Alina Dette, who typed parts of this manuscript with considerable technical expertise. This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt C1, C4) of the German Research Foundation (DFG).

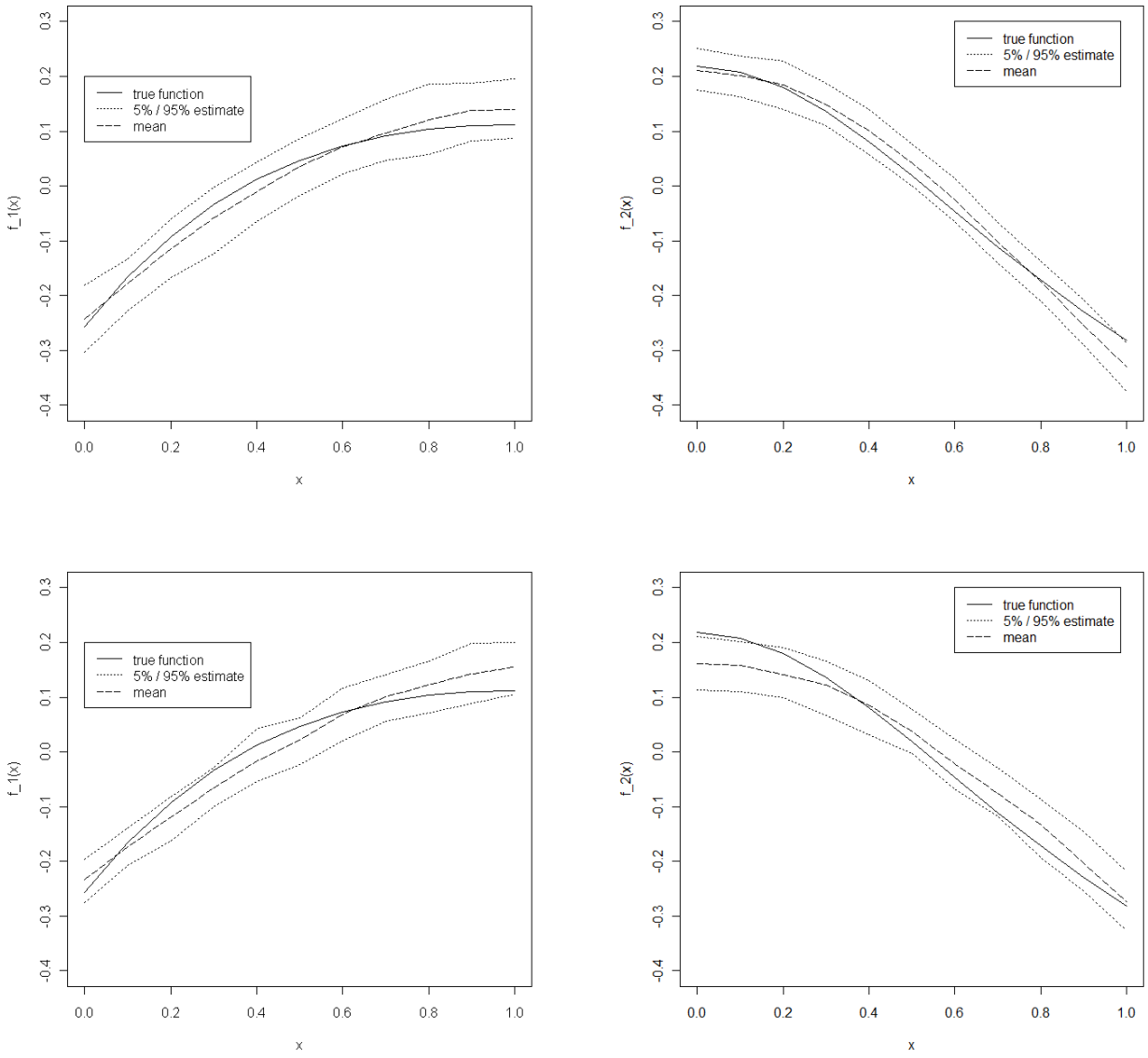


Figure 2: *Simulated mean,* 5%- *and* 95% *quantile of the backfitting estimate on the basis of* 500 *simulation runs, where model is given by* (4.2) *and the design is given by* (4.5) *(upper panel) and* (4.6) *(lower panel). Left part* θ_1*; right part:* θ_2*.*

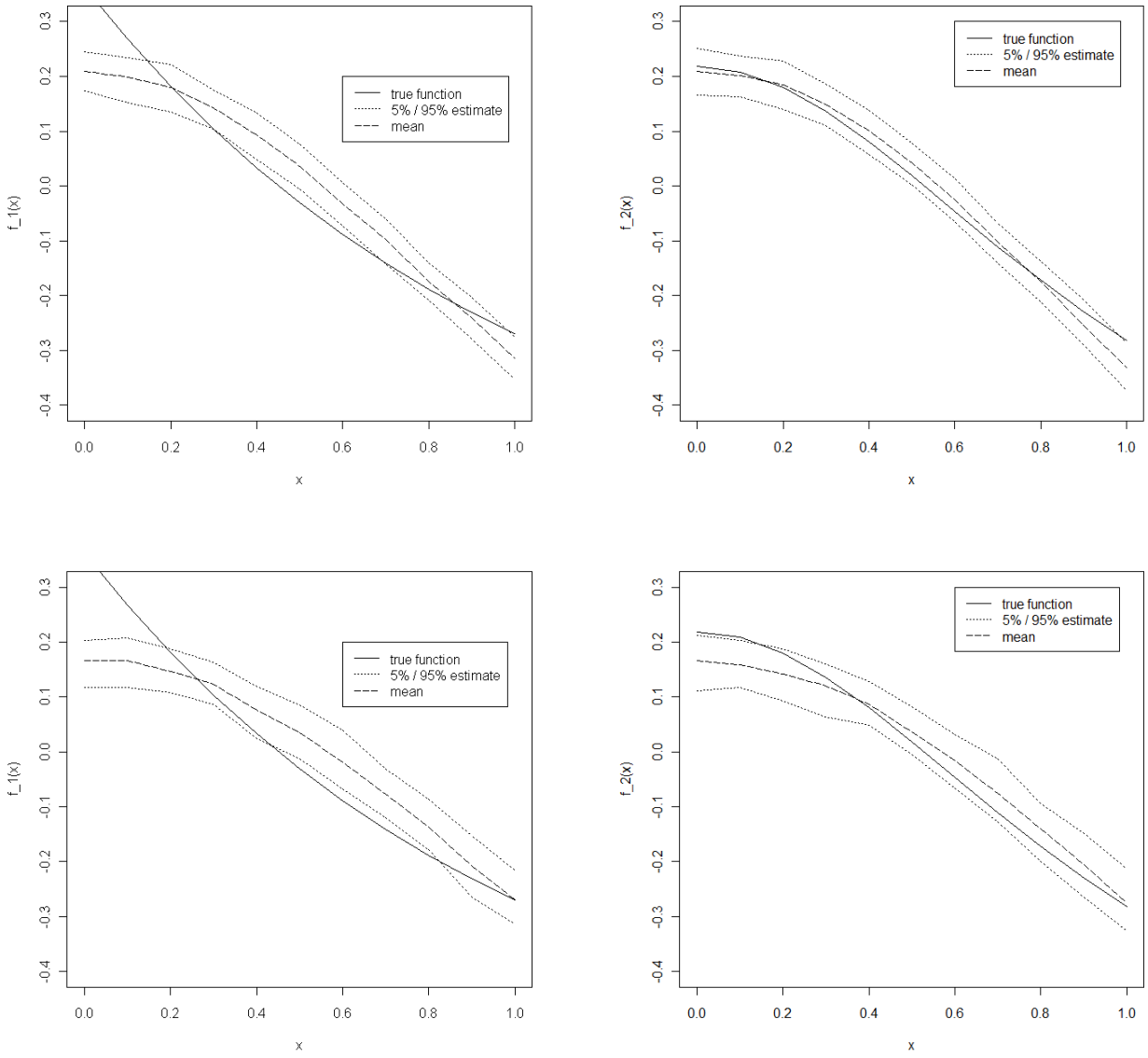


Figure 3: *Simulated mean,* 5%- *and* 95% *quantile of the backfitting estimate on the basis of* 500 *simulation runs, where model is given by* (4.3) *and the design is given by* (4.5) *(upper panel) and* (4.6) *(lower panel). Left part* θ_1*; right part:* θ_2*.*

design	(4.5)		(4.6)	
model	(4.1)	(4.2)	(4.1)	(4.2)
$\hat{\theta}_1$	0.00179	0.00189	0.00500	0.00353
$\hat{\theta}_2$	0.00154	0.00258	0.00488	0.00345
$\hat{\theta}_1^{MI}$	0.00347	0.00365	0.02219	0.00934
$\hat{\theta}_2^{MI}$	0.00311	0.00354	0.01917	0.01092

Table 1: *Simulated mean integrated squared error of the smooth backfitting estimator $\hat{\theta}_j$ ($j = 1, 2$) proposed in this paper and of the marginal estimator $\hat{\theta}_j^{MI}$ proposed by Hildebrandt et al. (2013).*

References

Bertero, M., Boccacci, P., Desiderà, G., and Vicidomini, G. (2009). Image deblurring with Poisson data: From cells to galaxies. *Inverse Problems*, 25(12):123006, 26.

Bissantz, N., Hohage, T., Munk, A., and Ruymgaart, F. (2007). Convergence rates of general regularization methods for statistical inverse problems. *SIAM J. Num. Anal.*, 45:2610–2636.

Brillinger, D. R. (2001). *Time Series Data Analysis and Theory*. SIAM.

Carroll, R. J., Härdle, W., and Mammen, E. (2002). Estimation in an additive model when the parameters are linked parametrically. *Econometric Theory*, 18(4):886–912.

Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004, 19.

Diggle, P. J. and Hall, P. (1993). A Fourier approach to nonparametric deconvolution of a density estimate. *Journal of the Royal Statistical Society, Series B*, 55:523–531.

Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht.

Folland, G. B. (1984). *Real Analysis - Modern Techniques and their Applications*. Wiley, New York.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 38(6):907–921.

Hengartner, N. W. and Sperlich, S. (2005). Rate optimal estimation with the integration method in the presence of many covariates. *Journal of Multivariate Analysis*, 95(2):246–272.

Hildebrandt, T. (2013). *Additive Modelle im inversen Regressionsproblem mit Faltungsoperator*. PhD thesis, Fakultät für Mathematik, Ruhr-Universität Bochum, Germany.

Hildebrandt, T., Bissantz, N., and Dette, H. (2013). Additive inverse regression models with convolution-type operators. Submitted for publication, `http://www.ruhr-uni-bochum.de/mathematik3/research/index.html`.

Kaipio, J. and Somersalo, E. (2010). *Statistical and Computational Inverse Problems*. Springer, Berlin.

Kammler, D. W. (2007). *A first course in Fourier Analysis*. Cambridge University Press.

Linton, O. B. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82(1):93–100.

Mair, B. A. and Ruymgaart, F. H. (1996). Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.*, 56:1424–1444.

Mammen, E., Linton, O. B., and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, 27(5):1443–1490.

Nielsen, J. P. and Sperlich, S. (2005). Smooth backfitting in practice. *Journal of the Royal Statistical Society, Ser. B*, 67(1):43–61.

Saitoh, S. (1997). *Integral Transforms, Reproducing Kernels and their Applications*. Longman, Harlow.

van der Vaart, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge, Cambridge University Press.

5 Appendix: Proof of Theorem 3.2

Let p denote the density of the transformed predictor $(Z_{1,1},\ldots,Z_{1,d})^T$. It is shown in Mammen et al. (1999) that the smooth backfitting algorithm (2.6) produces a sequence of estimates $(\hat{g}_1^{*(r)},\ldots,\hat{g}_d^{*(r)})_{r=0,1,\ldots}$ converging in $L^2(p)$ with geometric rate to a vector $(\overline{g}_1,\ldots,\overline{g}_d)$ which satisfies the system of equations

$$\overline{g}_j(z_j) \;=\; \hat{g}_j^*(z_j) - \sum_{k\neq j}\int \overline{g}_k(z_k)\Big[\frac{\hat{p}_{jk}(z_j,z_k)}{\hat{p}_j(z_j)} - \hat{p}_{k,[j+]}(z_k)\Big]dz_k - g_{0,j}^* \qquad j=1,\ldots,d, \tag{5.1}$$

where $g_{0,j}^*$ is defined in (2.8). Therefore the asymptotic properties of the smooth backfitting operator can be investigated replacing in (2.11) the random variables $U_{k,j}$ defined in (2.10) by their theoretical counterparts

$$\tilde{U}_{k,j} = Y_k - \sum_{\substack{i=1\\ i\neq j}}^{d}\tilde{g}_i(X_{k,i}) - \hat{g}_0,$$

where $\tilde{g}_i(X_{k,i}) = \bar{g}_i(Z_{k,i})$ $(i=1,\ldots d;\ k=1,\ldots,N)$ and $\tilde{g}_i = \bar{g}_i \circ F$ $(i=1,\ldots,d)$. This yields the representation

$$\tilde{U}_{k,j} = g_j(X_{k,j}) + \varepsilon_k + \sum_{\substack{i=1\\ i\neq j}}^{d}(g_i(X_{k,i}) - \tilde{g}_i(X_{k,i})) = g_j(X_{k,j}) + \varepsilon_k + B_{j,k,N}, \tag{5.2}$$

where the last equality defines the random variables $B_{j,k,N}$ in an obvious manner. The results of Mammen et al. (1999) imply

$$B_{j,k,N} = O_p(N^{-1/5}) \tag{5.3}$$

uniformly with respect to $j\in\{1,\ldots,d\}$ and $k\in\{1,\ldots,N\}$.

The assertion of Theorem 3.2 is now proved in four steps establishing the following statements:

$$b_{\hat{\theta}_j}(x_j) \;=\; \mathbb{E}[\hat{\theta}_j(x_j)] - \theta_j(x_j) = o(h^{s-1}) \tag{5.4}$$

$$\mathrm{Var}(\hat{\theta}_j(x_j)) \;=\; V_{N,j}(1+o(1)) \tag{5.5}$$

$$V_{n,j} \qquad \text{satisfies (3.4)} \tag{5.6}$$

$$|\mathrm{cum}_l(V_{N,j}^{-1/2}\hat{\theta}_j(x_j))| \;=\; o(1) \ \text{ for all } \ l\geq 3 \tag{5.7}$$

where $V_{N,j}$ is the normalizing factor defined in (3.3) and cum_l denotes the lth cumulant [see Brillinger (2001)].

Proof of (5.4): We first determine the expectation of the estimator $\hat{\theta}_j$ observing that the

estimator $\hat{\theta}_j$ is linear, i.e.

$$\hat{\theta}_j(x_j) \;=\; \sum_{k=1}^{N} w_{j,N}(x_j, X_{k,j})\tilde{U}_{k,j}, \tag{5.8}$$

where the weights $w_{j,N}(x_j, X_{k,j})$ are defined by

$$w_{j,N}(x_j, X_{k,j}) = \frac{1}{2\pi N h}\int_{\mathbb{R}} e^{-iw(x_j - X_{k,j})/h}\frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})}dw\frac{1}{\max\{f_j(X_{k,j}), f_j(\frac{1}{a_n})\}}, \tag{5.9}$$

and we have replaced the quantities $U_{k,j}$ by $\tilde{U}_{k,j}$ as described at the beginning of the proof. This representation gives

$$\mathbb{E}[\hat{\theta}_j(x_j)] \;=\; E_1 + E_2, \tag{5.10}$$

where the terms E_1 and E_2 are defined by

$$E_1 \;=\; \mathbb{E}\Big[\sum_{k=1}^{N} g_j(X_{k,j})w_{j,N}(x_j, X_{k,j})\Big], \qquad E_2 = \mathbb{E}\Big[\sum_{k=1}^{N} B_{j,k,N}w_{j,N}(x_j, X_{k,j})\Big]. \tag{5.11}$$

Using the definition of $B_{j,k,N}$ and (5.3) the term E_2 can be estimated as follows

$$\begin{aligned} |E_2| \;&\leq\; \mathbb{E}\Big[\sum_{k=1}^{N}\sum_{\substack{i=1\\ i\neq j}}^{d} |g_i(X_{k,i}) - \tilde{g}_i(X_{k,i})| \max_k |w_{j,N}(x_j, X_{k,j})|\Big] \\ &\leq\; \frac{C}{h^{\beta_j+1}f_j(\frac{1}{a_N})}\mathbb{E}\Big[\sum_{\substack{i=1\\ i\neq j}}^{d} |g_i(X_{k,i}) - \tilde{g}_i(X_{k,i})|\Big] \leq \frac{C}{N^{1/5}h^{\beta_j+1}f_j(\frac{1}{a_N})} = o(h^{s-1}), \end{aligned} \tag{5.12}$$

where we used the representation (5.9) and Assumption (A4). The second inequality in (5.12) follows from the fact that

$$\mathbb{E}[|g_i(X_{k,i}) - \tilde{g}_i(X_{k,i})|] = O(N^{-1/5}). \tag{5.13}$$

In order to establish this statement note that $g_i(X_{k,i}) - \tilde{g}_i(X_{k,i}) = O_P(N^{-1/5})$ (uniformly with respect to $k = 1, ..., N$). The proof of the L^1-convergence follows along the lines of the proof of the stochastic convergence in Mammen et al. (1999). Here one additionally shows in each step of the backfitting iteration stochastic convergence and L^1- convergence [see Hildebrandt (2013) for details].

Similarly, we obtain from the definition of the weights $w_{j,N}(x_j, X_{k,j})$ in (5.9) the representation

$$
\begin{aligned}
E_1 &= \frac{1}{2\pi h}\int_{\mathbb{R}} g_j(y)\int_{\mathbb{R}} e^{-iw(x_j-y)/h}\frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})}dw\frac{f_j(y)}{\max\{f_j(y), f_j(\frac{1}{a_N})\}}dy \\
&= \frac{1}{2\pi h}\int_{\mathbb{R}} \Phi_{g_j}\Big(\frac{w}{h}\Big)e^{-iwx_j/h}\frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})}dw \\
&\quad -\frac{1}{2\pi h}\int_{\mathbb{R}} g_j(y)\int_{\mathbb{R}} e^{-iw(x_j-y)/h}\frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})}dw\Big(1-\frac{f_j(y)}{\max\{f_j(y), f_j(\frac{1}{a_N})\}}\Big)dy\ , \\
&= \theta_j(x_j) - F_1 - F_2,
\end{aligned}
\tag{5.14}
$$

where the terms F_1 and F_2 are defined by

$$
\begin{aligned}
F_1 &= \frac{1}{2\pi h}\int_{\mathbb{R}} \Phi_{\theta_j}\left(\frac{w}{h}\right)e^{-iwx_j/h}\left(1-\Phi_K(w)\right)dw, \\
F_2 &= \frac{1}{2\pi h}\int_{\mathbb{R}} g_j(y)\int_{\mathbb{R}} e^{-iw(x_j-y)/h}\frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})}dw\Big(1-\frac{f_j(y)}{\max\{f_j(y), f_j(\frac{1}{a_N})\}}\Big)dy,
\end{aligned}
$$

respectively. The term F_1 can be estimated using Assumption (A6), that is

$$
\begin{aligned}
|F_1| &\le \frac{1}{2\pi h}\int_{\mathbb{R}} |\Phi_{\theta_j}\left(\frac{w}{h}\right)|1-\Phi_K(w)|dw \le \frac{1}{\pi h}\int_{[-b,b]^c} |\Phi_{\theta_j}\left(\frac{w}{h}\right)|dw \\
&\le \frac{1}{\pi}\int_{[-b/h,b/h]^c}\frac{1}{|y|^{s-1}}|y|^{s-1}|\Phi_{\theta_j}(y)|dy \\
&\le \frac{h^{s-1}}{b^{s-1}\pi}\int_{[-b/h,b/h]^c}|y|^{s-1}|\Phi_{\theta_j}(y)|dy = o(h^{s-1}),
\end{aligned}
$$

while the term F_2 is estimated similarly, using Assumption (A4), (A7) and (A8) that is

$$
\begin{aligned}
|F_2| &\le \frac{1}{2\pi h}\int_{\mathbb{R}} |g_j(y)|\int_{\mathbb{R}}\frac{|\Phi_K(w)|}{|\Phi_{\psi_j}(\frac{w}{h})|}dw\Big|1-\frac{f_j(y)}{\max\{f_j(y), f_j(\frac{1}{a_N})\}}\Big|dy \\
&\le \frac{1}{2\pi h}\int_{([-1/a_N,1/a_N])^c}|g_j(y)|dy\int_{\mathbb{R}}\frac{|\Phi_K(w)|}{|\Phi_{\psi_j}(\frac{w}{h})|}dw \;=\; O\left(\frac{a_N^r}{h^{1+\beta_j}}\right) \;=\; o(h^{s-1}).
\end{aligned}
$$

From these estimates and (5.14) we obtain $E_1 = \theta_j(x_j) + o\left(h^{s-1}\right)$, and the assertion (5.4) now follows from the decomposition (5.10) and (5.12).

Proof of (5.5)***:*** Using standard results for cumulants [see Brillinger (2001)] the variance of the estimate $\hat{\theta}_j$ can be calculated as

$$
\mathrm{Var}(\hat{\theta}_j(x_j)) = S_1 + S_2 + S_3 + 2S_4 + 2S_5 + 2S_6,
\tag{5.15}
$$

where

$$
\begin{aligned}
S_1 &= \sum_{k=1}^{N}\sum_{l=1}^{N} \mathrm{cum}\big(\varepsilon_k w_{j,N}(x_j, X_{k,j}), \varepsilon_l \overline{w_{j,N}(x_j, X_{l,j})}\big) \\
S_2 &= \sum_{k=1}^{N}\sum_{l=1}^{N} \mathrm{cum}\big(g_j(X_{k,j}) w_{j,N}(x_j, X_{k,j}), g_j(X_{l,j}) \overline{w_{j,N}(x_j, X_{l,j})}\big) \\
S_3 &= \sum_{k=1}^{N}\sum_{l=1}^{N} \mathrm{cum}\big(B_{j,k,N} w_{j,N}(x_j, X_{k,j}), B_{j,l,N} \overline{w_{j,N}(x_j, X_{l,j})}\big) \\
S_4 &= \sum_{k=1}^{N}\sum_{l=1}^{N} \mathrm{cum}\big(\varepsilon_k w_{j,N}(x_j, X_{k,j}), g_j(X_{l,j}) \overline{w_{j,N}(x_j, X_{l,j})}\big) \\
S_5 &= \sum_{k=1}^{N}\sum_{l=1}^{N} \mathrm{cum}\big(\varepsilon_k w_{j,N}(x_j, X_{k,j}), B_{j,l,N} \overline{w_{j,N}(x_j, X_{l,j})}\big) \\
S_6 &= \sum_{k=1}^{N}\sum_{l=1}^{N} \mathrm{cum}\big(g_j(X_{k,j}) w_{j,N}(x_j, X_{k,j}), B_{j,l,N} \overline{w_{j,N}(x_j, X_{l,j})}\big).
\end{aligned}
$$

It is easy to see that $S_4 = 0$ because of $\mathbb{E}[\varepsilon_k] = 0$ and the independence of ε_k and $\mathbf{X}_k$. We will show that the first two terms S_1 and S_2 determine the variance and that the terms S_3, S_5 and S_6 are of smaller order. For a proof of the latter result we concentrate on the sixth term because the results for the terms S_3 and S_5 can be treated analogously.
As ε_k, ε_l, $X_{k,j}$ and $X_{l,j}$ are independent for $k \neq l$ the term S_1 can be written as

$$
\begin{aligned}
N\mathrm{cum}\big(\varepsilon_k w_{j,N}(x_j, X_{k,j}), \varepsilon_k \overline{w_{j,N}(x_j, X_{k,j})}\big) &= N\mathrm{cum}\big(\varepsilon_k, \varepsilon_k\big)\mathrm{cum}\big(w_{j,N}(x_j, X_{k,j}), \overline{w_{j,N}(x_j, X_{k,j})}\big) \\
&+ N\mathrm{cum}\big(\varepsilon_k, \varepsilon_k\big)\mathrm{cum}\big(w_{j,N}(x_j, X_{k,j})\big)\mathrm{cum}\big(\overline{w_{j,N}(x_j, X_{k,j})}\big),
\end{aligned}
$$

where we used the product theorem for cumulants and $\mathbb{E}[\varepsilon_k] = 0$. Now a straightforward calculation gives

$$
S_1 = \frac{\sigma^2}{Nh^2(2\pi)^2} \int_{\mathbb{R}} \Big| \int_{\mathbb{R}} e^{-iw(x_j - y)/h} \frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})} dw \Big|^2 \frac{f_j(y)}{\max\{f_j(y), f_j(\frac{1}{a_N})\}^2} dy \cdot (1 + o(1)).
$$

The second summand in (5.16) can be calculated in the same way and we obtain

$$
S_2 = \frac{1}{Nh^2(2\pi)^2} \int_{\mathbb{R}} \Big| \int_{\mathbb{R}} e^{-iw(x_j - y)/h} \frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})} dw \Big|^2 \frac{g_j^2(y) f_j(y)}{\max\{f_j(y), f_j(\frac{1}{a_N})\}^2} dy \cdot (1 + o(1)).
$$

In a last step we investigate the sixth summand of (5.16) (the other terms S_3 and S_5 are treated in the same way). By the product theorem and the definition of the cumulants we obtain for this

term

$$S_6 \;=\; -\sum_{k\neq l}\sum_{\substack{i=1\\ i\neq j}}^{d} \mathrm{Cov}\big(g_j(X_{k,j})w_{j,N}(x_j,X_{k,j}),\overline{g}_i(F(X_{l,i}))w_{j,N}(x_j,X_{l,j}\big)\cdot(1+o(1)),$$

where we used the definitions of $B_{j,l,N}=\sum_{i\neq j}(g_i(X_{l,i})-\tilde{g}_i(X_{l,i}))$ and $\overline{g}_i=\tilde{g}_i\circ F_i^{-1}$. We introduce the weights

$$q_{mj}(X_{l,i})=\frac{L\big(\frac{F_j(X_{m,i})-F_j(X_{l,i})}{h_B}\big)}{\sum_{s=1}^{N}L\big(\frac{F_j(X_{s,i})-F_j(X_{l,i})}{h_B}\big)}\quad l,m=1,...,N;\quad i=1,...,d,$$

denote by

$$\hat{g}_i^*(F_i(X_{l,i}))\;=\;\sum_{m=1}^{N}q_{mi}(X_{l,i})Y_m\quad l=1,...,N;\quad i=1,...,d \tag{5.16}$$

the one-dimensional Nadaraya-Watson estimator from the data $F_i(X_{1,i}),\ldots,F_i(X_{N,i})$ evaluated at the point $F_i(X_{l,i})$ and define

$$v_{mi}(X_{l,i},z_m)=\frac{\hat{p}_{im}(F_i(X_{l,i}),z_m)}{\hat{p}_i(F_i(X_{l,i}))}-\hat{p}_{m,[i+]}(z_m)\quad i,m=1,...,d;\quad l=1,...,N$$

as the integrand in equation (5.1). This yields for the term S_6 the decomposition

$$S_6\;=\;(B-A)(1+o(1)), \tag{5.17}$$

where the terms A and B are defined by

$$A\;=\;\sum_{k\neq l}\sum_{\substack{i=1\\ i\neq j}}^{d}\mathrm{Cov}\big(g_j(X_{k,j})w_{j,N}(x_j,X_{k,j}),w_{j,N}(x_j,X_{l,j})\sum_{m=1}^{N}q_{mi}(X_{l,i})Y_m\big)$$

and

$$B\;=\;\sum_{k\neq l}\sum_{\substack{i=1\\ i\neq j}}^{d}\sum_{\substack{m=1\\ m\neq i}}^{d}\mathrm{Cov}\Big(g_j(X_{k,j})w_{j,N}(x_j,X_{k,j}),\big(\int\tilde{g}_m(z_m)v_{mi}(X_{l,i},z_m)dz_m+g_{0,i}^*\big)w_{j,N}(x_j,X_{l,j})\Big),$$

respectively. We start with the estimation of the term A calculating each covariance separately, that is

$$\Big|\mathrm{Cov}\big(g_j(X_{k,j})w_{j,N}(x_j,X_{k,j}),w_{j,N}(x_j,X_{l,j})\sum_{m=1}^{N}q_{mi}(X_{l,i})Y_m\big)\Big|\leq\big(H_1+H_2\big)(1+o(1)), \tag{5.18}$$

where the terms H_1 and H_2 are defined by

$$H_1 = \frac{1}{Nh_B}\Big|\sum_{r=1}^{d}\mathbb{E}\Big[g_j(X_{k,j})w_{j,N}(x_j,X_{k,j})L\Big(\frac{F_i(X_{k,i})-F_i(X_{l,i})}{h_B}\Big)g_r(X_{k,r})\overline{w_{j,N}(x_j,X_{l,j})}\Big]$$

$$H_2 = \frac{1}{Nh_B}\Big|\mathbb{E}\Big[g_j(X_{k,j})w_{j,N}(x_j,X_{k,j})\Big]\mathbb{E}\Big[L\Big(\frac{F_i(X_{k,i})-F_i(X_{l,i})}{h_B}\Big)\sum_{r=1}^{d}g_r(X_{k,r})\overline{w_{j,N}(x_j,X_{l,j})}\Big]$$

and we used the fact that the kernel density estimate

$$\frac{1}{Nh_B}\sum_{m}L\Big(\frac{F_i(X_{m,i})-F_i(X_{l,i})}{h_B}\Big)=\hat{p}_i(F_i(X_{l,i}))$$

in the denominator of the Nadaraya-Watson estimate (5.16) converges uniformly to 1 as $F_i(X_{l,i})$ is uniformly distributed on the interval $[0,1]$ [see Giné and Guillou (2002)]. We first investigate the term H_1 and obtain by a tedious calculation using assumption (A4) and (A9)

$$\begin{aligned}
H_1 &\le \frac{(1+o(1))}{N^3h^2}\Big|\int_{\mathbb{R}^2}\Big(\int_{\mathbb{R}}g_j(t_j)\int_{\mathbb{R}}e^{-iw(x_j-t_j)/h}\frac{\Phi_K(w)}{\Phi_{\psi_j}\left(\frac{w}{h}\right)}dw\frac{f_j(t_j)f_{irj}(t_i,t_r|t_j)}{\max\{f_j(t_j),f_j(1/a_N)\}}dt_j\Big)\\
&\times\int_{\mathbb{R}}\Big(\int_{\mathbb{R}}e^{-iw(x_j-s_j)/h}\frac{\Phi_K(w)}{\Phi(w/h)}dw\frac{f_j(s_j)f_{ij}(t_i|s_j)}{\max\{f_j(t_j),f_j(1/a_N)\}}ds_j\Big)g_j(t_r)dt_idt_r\Big|\\
&\le \frac{C}{N^3h^2}\int_{\mathbb{R}^2}\Big|\int_{\mathbb{R}}e^{-iw(x_j-t_j)/h}\frac{\Phi_K(w)}{\Phi_{\psi_j}(w/h)}dw\frac{f_j(t_j)}{\max\{f_j(t_j),f_j(1/a_N)\}}dt_j\Big|\\
&\times\Big|\int_{\mathbb{R}}e^{-iw(x_j-s_j)/h}\frac{\Phi_K(w)}{\Phi_{\psi_j}(w/h)}dw\frac{f_j(s_j)}{\max\{f_j(t_j),f_j(1/a_N)\}}ds_j\Big|U_{irj}(t_i,t_r)\eta_{ij}(t_i)dt_idt_r\\
&= o\Big(\frac{1}{N^3h^{2\beta+1}}\Big)
\end{aligned}$$

uniformly with respect to k,l. A similar calculation yields

$$H_2\le\frac{1}{Nh_B}\Big|\frac{E_1}{N}\,\mathbb{E}\Big[L\Big(\frac{F_i(X_{k,i})-F_i(X_{l,i})}{h_B}\Big)\sum_{r=1}^{d}g_r(X_{k,r})\overline{w_{j,N}(x_j,X_{l,j})}\Big]\Big|=o\Big(\frac{1}{N^3h^{2\beta+1}}\Big)$$

(uniformly with respect to k,l) where we use the estimate (5.11) in the first step. Consequently the term A in (5.18) can be bounded by $A=o(1/Nh^{2\beta+1})$ A tedious calculation using similar arguments yields for the term $B=O(1/Nh^{2\beta+1})$ and by (5.17) the sum S_6 is of the same order. Moreover, it will be shown in the proof of (5.6) below that this order is smaller than the order of the first two summands S_1 and S_2 in (5.16) which gives

$$S_6=O\Big(\frac{1}{Nh^{2\beta+1}}\Big)=o(S_j)\quad j=1,2.$$

A similar calculation for the terms S_3 and S_5 finally yields

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_j(x_j)) &= \frac{1}{Nh^2(2\pi)^2}\int_{\mathbb{R}}\Big|\int_{\mathbb{R}} e^{-iw(x_j-y)/h}\frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})}dw\Big|^2\frac{(g_j^2(y)+\sigma^2)f_j(y)}{\max\{f_j(y),f_j(\frac{1}{a_N})\}^2}dy\times(1+o(1))\ ,\\ &= V_{N,j}(1+o(1))\ ,\end{aligned}$$

which proves (5.5).

Proof of (5.6). As g_j is bounded for all $j=1,...,d$ and $\max\{f_j(y),f_j(\frac{1}{a_N})\}^2 \geq f_j(y)f_j(\frac{1}{a_N})$ the term $V_{N,j}$ defined in (3.3) can be estimated as follows

$$|V_{N,j}| \leq \frac{C}{Nh(2\pi)^2 f_j(\frac{1}{a_N})}\int_{\mathbb{R}}\Big|\int_{\mathbb{R}} e^{-iw(x_j/h-y)}\frac{\Phi_K(w)}{\Phi_{\psi_j}(\frac{w}{h})}dw\Big|^2dy = \frac{C}{Nh(2\pi)^2 f_j(\frac{1}{a_N})}\int_{\mathbb{R}}\frac{|\Phi_K(w)|^2}{|\Phi_{\psi_j}(\frac{w}{h})|^2}dw\ ,$$

where C is a constant and we used Parseval's equality for the last identity [see Kammler (2007)]. Now assumption (A4) yields the upper bound, that is $|V_{N,j}| \leq C/Nh^{1+2\beta_j}f_j(\frac{1}{a_N})$. From the assumption $f_j(x)^{-1} \geq C$ and again Parsevals equality we also get the lower bound $|V_{N,j}| \geq C/Nh^{1+2\beta_j}$, which completes the proof of the estimate (3.4).

Proof of (5.7)***:*** Observing the representation (5.8) the lth cumulant of the estimate $\hat{\theta}_j$ can be estimated as follows

$$|\mathrm{cum}_l(\hat{\theta}_j(x_j))| = \Big|\sum_{k_1,...,k_l=1}^{N}\mathrm{cum}\Big(\tilde{U}_{k_1,j}w_{j,N}(x_j,X_{k_1,j}),...,\tilde{U}_{k_l,j}w_{j,N}(x_j,X_{k_l,j})\Big)\Big| \leq G_1+G_2,$$

where the terms G_1 and G_2 are defined by

$$\begin{aligned}G_1 &= \Big|\sum_{k_1,...,k_l=1}^{N}\mathrm{cum}\Big(A_{k_1,j}w_{j,N}(x_j,X_{k_1,j}),...,A_{k_l,j}w_{j,N}(x_j,X_{k_l,j})\Big)\Big|\\ G_2 &= \Big|\sum_{k_1,...,k_l=1}^{N}\sum_{s=1}^{l}\binom{l}{s}\mathrm{cum}\Big(B_{j,k_1,N}w_{j,N}(x_j,X_{k_1,j}),\ldots,B_{j,k_s,N}w_{j,N}(x_j,X_{k_s,j}),\\ &\qquad A_{k_{s+1},j}w_{j,N}(x_j,X_{k_{s+1},j}),...,A_{k_l,j}w_{j,N}(x_j,X_{k_l,j})\Big)\Big|\end{aligned}$$

and we introduce the notation $A_{k_i,j}=g_j(X_{k_i,j})+\varepsilon_{k_i}$. Exemplarily we investigate the first term of this decomposition, the term G_2 is treated similarly. As the random variables $A_{k_1,j}w_{j,N}(x_j,X_{k_1,j})$ and $A_{k_2,j}w_{j,N}(x_j,X_{k_2,j})$ are independent for $k_1\neq k_2$ and identically distributed for $k_1=k_2$ it follows that

$$G_1 = N\Big|\mathrm{cum}_l(A_{k,j}w_{j,N}(x_j,X_{k,j}))\Big| \leq N\sum_{s=0}^{l}\binom{l}{s}\sum_{\substack{\mathbf{j}\in\{0,1\}^l\\ j_1+...+j_l=s}}\Big|\sum_{\nu}\prod_{k=1}^{p}\mathrm{cum}(A_{ij},ij\in\nu_k)\Big|,$$

where we used the product theorem for cumulants [see Brillinger (2001)] and the third sum extends over all indecomposable partitions of the table

$$\begin{array}{ccc} A_{i1} & & A_{i2} \\ \vdots & & \vdots \\ A_{i1} & & A_{i2} \\ & A_{ij} & \\ & \vdots & \\ & A_{ij} & \end{array}$$

with $A_{i1} = \varepsilon_1$ $(1 \leq i \leq s)$, $A_{i2} = w_{j,N}(x_j, X_{1,j})$ $(1 \leq i \leq s)$ and $A_{ij} = g_j(X_{1,j})w_{j,N}(x_j, X_{1,j})$ $(s+1 \leq i \leq l)$. In order to illustrate how to estimate this expression we consider exemplarily the case $l = 3$, where G_1 reduces to

$$G_1 \;=\; N \sum_{s=0}^{3} \binom{3}{s} \sum_{\substack{\mathbf{j} \in \{0,1\}^3 \\ j_1 + \ldots + j_3 = s}} \Big| \sum_{\nu} \prod_{k=1}^{p} \operatorname{cum}(A_{ij}, ij \in \nu_k) \Big|.$$

As ε is independent of X_1 and has mean 0 the partitions in G_1 with $s = 1$ vanish. The terms corresponding to $s = 0, 2, 3$ contain only quantities of the form

$$\begin{aligned} &\operatorname{cum}_3(g_j(X_{1,j})w_{j,N}(x_j, X_{1,j})), \\ &\sigma^2 \operatorname{cum}\big(w_{j,N}(x_j, X_{1,j}), w_{j,N}(x_j, X_{1,j}), g_j(X_{1,j})w_{j,N}(x_j, X_{1,j})\big), \\ &\sigma^2 \operatorname{cum}\big(w_{j,N}(x_j, X_{1,j})\big) cum\big(w_{j,N}(x_j, X_{1,j}), g_j(X_{1,j})w_{j,N}(x_j, X_{1,j})\big), \\ &\kappa_3 \operatorname{cum}\big(w_{j,N}(x_j, X_{1,j}), w_{j,N}(x_j, X_{1,j}), w_{j,N}(x_j, X_{1,j})\big), \\ &\kappa_3 \operatorname{cum}\big(w_{j,N}(x_j, X_{1,j}), w_{j,N}(x_j, X_{1,j})\big) cum\big(w_{j,N}(x_j, X_{1,j})\big), \\ &\kappa_3 \operatorname{cum}\big(w_{j,N}(x_j, X_{1,j})\big) cum\big(w_{j,N}(x_j, X_{1,j})\big) cum\big(w_{j,N}(x^*, X_{1,j})\big), \end{aligned}$$

where κ_3 denotes the third cumulant of ε_1. As the inequality

$$\mathbb{E}\Big[|g_j(X_{1,j})w_{j,N}(x_j, X_{1,j}))|^{b_r} |w_{j,N}(x_j, X_{1,j}))|^{a_r - b_r} \Big] \leq \frac{C}{N^{a_r} h^{a_r(\beta_j+1)} f_j(\frac{1}{a_n})^{a_r}}$$

holds for $0 \leq b_r \leq a_r$ all terms can be bounded by $C/(N^3 h^{3(\beta_j+1)} f_j(\frac{1}{a_n})^3)$. This yields

$$N^{3/2} h^{3\beta_j + 3/2} G_1 \leq C N^{3/2+1} h^{3\beta_j+3/2} \frac{1}{N^3 h^{3(\beta_j+1)} f_j(\frac{1}{a_n})^3} = o(1),$$

where we used the conditions on the bandwidth in the last step. Similar calculations for the

general case show

$$N^{l/2}h^{l\beta_j+l/2}G_1 \;=\; O\big((N^{l/2-1}h^{l/2}f_j(\frac{1}{a_N})^l)^{-1}\big) = o(1)$$

whenever $l \geq 3$. The term G_2 can be calculated in the same way, where for example one additionally has to use the estimate $\mathrm{Cov}(B_{j,k,N}, \varepsilon_l) = O(1/N)$ uniformly with respect to all $j = 1, ..., d$, and $k, l = 1, ..., N$, which follows from the definition of the backfitting estimator.