The Annals of Statistics
1999, Vol. 27, No. 5, 1443–1490

THE EXISTENCE AND ASYMPTOTIC PROPERTIES OF A BACKFITTING PROJECTION ALGORITHM UNDER WEAK CONDITIONS[1]

BY E. MAMMEN,[2] O. LINTON[3] AND J. NIELSEN

Reprecht-Karls-Universität Heidelberg, Yale University and Codanhus

We derive the asymptotic distribution of a new backfitting procedure for estimating the closest additive approximation to a nonparametric regression function. The procedure employs a recent projection interpretation of popular kernel estimators provided by Mammen, Marron, Turlach and Wand and the asymptotic theory of our estimators is derived using the theory of additive projections reviewed in Bickel, Klaassen, Ritov and Wellner. Our procedure achieves the same bias and variance as the oracle estimator based on knowing the other components, and in this sense improves on the method analyzed in Opsomer and Ruppert. We provide "high level" conditions independent of the sampling scheme. We then verify that these conditions are satisfied in a regression and a time series autoregression under weak conditions.

1. Introduction. Separable models are important in exploratory analyses of nonparametric regression. The backfitting technique has long been the state of the art method for estimating these models; see Hastie and Tibshirani (1991). While backfitting has proved very useful in application and simulation studies, it has been somewhat difficult to analyze theoretically, which has long been a drawback to its universal acceptance. Recently, a new method, called marginal integration, has been proposed; see Linton and Nielsen (1995), Tjøstheim and Auestad (1994) and Newey (1994) [see also earlier work by Auestad and Tjøstheim (1991)]. This method is perhaps easier to understand for nonstatisticians since it involves averaging rather than iterative solution of nonlinear equations. Its statistical properties are trivial to obtain and have been established in the aforementioned papers. Although tractable, marginal integration is not generally efficient. Linton (1997) and Fan, Mammen and Härdle (1998) showed how to improve on the efficiency of the marginal integration estimator in regression. In the former paper, this was achieved by carrying out one backfitting iteration from this

Received July 1997; revised April 1999.
[1] Research on this paper was started when the authors were visiting the Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse," Humboldt-Universität zu Berlin.
[2] Supported in part by the Deutsche Forschungsgemeinschaft, project MA1026/6-1.
[3] Supported in part by the NSF and NATO.
AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.
Key words and phrases. Additive models, alternating projections, backfitting, kernel smoothing, local polynomials, nonparametric regression.

initial consistent starting point. This modification actually achieves full oracle efficiency, that is, one achieves the same result as if one knew the other components. This suggests that backfitting itself is also efficient in the same sense. Moreover, backfitting, since it relies only on one-dimensional smooths, is free from the curse of dimensionality.

Recent work by Opsomer and Ruppert (1997) and Opsomer (1998) has addressed the algorithmic and statistical properties of backfitting. Specifically, they gave sufficient conditions for the existence and uniqueness of a version of backfitting, or rather an exact solution to the empirical projection equations, suitable for any (recentered) smoother matrix. They also derived an expansion for the conditional mean squared error of their version of backfitting: the asymptotic variance is equal to the oracle bound while the precise form of the bias, as for the integration method, depends on the way recentering is carried out, but in any case the bias is not oracle, except when the covariates are mutually independent. This important work confirms the efficiency, at least with respect to variance, of (their version of) backfitting. Unfortunately, their version of backfitting is not design adaptive, which is somewhat surprising given that they use local polynomial smoothers throughout. Furthermore, their proof technique required one rather strong condition: specifically, the amount of dependence in the covariates was strictly limited.

In this paper, we define a new backfitting-type estimator for additive nonparametric regression. We make use of an interpretation of the Nadaraya–Watson estimator and the local linear estimator as projections in an appropriate Hilbert space, which was first provided by Mammen, Marron, Turlach and Wand (1997). Our additive estimator is defined as the further projection of these multivariate estimators down on the space of additive functions. We examine this estimator and show how, in both the Nadaraya–Watson case and the local linear case, the estimator can be interpreted as a backfitting estimator defined through iterative solution of the empirical equations. We establish the geometric convergence of the backfitting equations to the unique solution using the theory of additive projections; see Bickel, Klaassen, Ritov and Wellner (1993). We use this result to establish the limiting behavior of the estimates: we give both the asymptotic distribution and a uniform convergence result. Our procedure achieves the same bias and variance as the oracle estimator based on knowing the other components, and in this sense improves on the method analyzed in Opsomer and Ruppert (1997). Although the criterion function is defined in terms of the high-dimensional estimates, we show that the estimator is also characterized by equations that only depend on one- and two-dimensional marginals, so that the curse of dimensionality truly does not operate here. Our first results are established using ideas from Hilbert space mathematics and hold under "high level" conditions, which are formulated independently of specific sampling assumptions. We then verify these conditions in an i.i.d. regression model and in a time series autoregression with strong mixing data. Our

conditions are weaker than those of Opsomer and Ruppert (1997) and do not restrict the dependence between the covariates in any way.

The paper is organized as follows. In Section 2 we show how local polynomial estimators can be interpreted as projections. In Section 3 we introduce our additive estimators in the simplest situation, that is, for the Nadaraya–Watson-like pilot estimator, establishing the convergence of the backfitting algorithm and the asymptotic distribution of the estimator under high level conditions that are suitable for a range of sampling schemes. In Section 4 we extend the analysis to local polynomials. In Section 5 we give primitive conditions in a time series autoregression that imply the high level conditions. All proofs are contained in the Appendix.

2. A projection interpretation of the local polynomials. Let Y, X be random variables of dimensions 1 and d, respectively, and let $(Y^1, X^1), \ldots, (Y^n, X^n)$ be a random sample drawn from (Y, X). We first provide a new interpretation of local polynomial estimators of the regression function $m(x_1, \ldots, x_d) = E(Y|X = x)$ evaluated at the vector $x = (x_1, \ldots, x_d)$, based on Mammen, Marron, Turlach and Wand (1997). This new point of view will be useful for interpreting our estimators of the restricted additive function $m(x) = m_0 + m_1(x_1) + \cdots + m_d(x_d)$.

The full-dimensional qth order local polynomial regression smoother which we denote by $\hat{\mathbf{m}}(x) = (\hat{m}^0(x), \ldots, \hat{m}^{s-1}(x))^T$ satisfies

$$\begin{aligned}\hat{\mathbf{m}}(x) = \arg \min_{\theta^0, \ldots, \theta^{s-1}} \sum_{i=1}^{n} \Bigg\{ Y^i - \theta^0 - \left(\frac{X_1^i - x_1}{h}\right)\theta^1 \\ - \cdots - \left(\frac{X_d^i - x_d}{h}\right)^q \theta^{s-1} \Bigg\}^2 \\ \times \prod_{l=1}^{d} K_h\left(X_l^i - x_l\right),\end{aligned} \tag{1}$$

where $K_h(\cdot) = K(\cdot/h)/h$ with $K(\cdot)$ a univariate kernel and $h = h(n)$ a positive bandwidth sequence, while q is the order of the polynomial approximation and $s = \sum_{l=0}^{q} \binom{l+d-1}{d-1}$ is the total number of distinct partial derivatives up to and including the qth order. In fact, for simplicity of notation we will concentrate on the local linear case considered in Ruppert and Wand (1994) for which $q = 1$ and $s = d + 1$. The Nadaraya–Watson case, for which $q = 0$ and $s = 1$, is even simpler; see below. For simplicity of notation, we use product kernels that have the same kernel and the same bandwidth in each component. Our results can be easily extended to the case of different kernels and bandwidths.

For the new interpretation of local linear estimators we shall think of the data $\mathbf{Y} = (Y^1, \ldots, Y^n)^T$ as an element of the space of tuples of $n(d + 1)$

functions

$$\mathscr{F} = \{(f^{i,j}: i = 1, \ldots, n; j = 0, \ldots, d): \text{Here, } f^{i,j} \text{ are functions from } \mathbb{R}^d \text{ to } \mathbb{R}\}.$$

We do this by putting $f^{i,0}(x) \equiv Y^i$ and $f^{i,j}(x) \equiv 0$ for $j \neq 0$. We define the following seminorm on $\mathscr{F}$:

$$(2) \quad \|f\|_*^2 = \int \frac{1}{n} \sum_{i=1}^{n} \left[f^{i,0}(x) + \sum_{j=1}^{d} f^{i,j}(x) \frac{x_j - X_j^i}{h} \right]^2 \prod_{j=1}^{d} K_h(X_j^i - x_j)\, dx.$$

Consider now the following subspaces of $\mathscr{F}$:

$\mathscr{F}_{\text{full}} = \{f \in \mathscr{F}: f^{i,j}$ does not depend on i for $j = 0, \ldots, d\}$,
$\mathscr{F}_{\text{add}} = \{f \in \mathscr{F}_{\text{full}}: f^{i,0}(x) = g_1(x_1) + \cdots + g_d(x_d)$ for some functions $g_j: \mathbb{R} \mapsto \mathbb{R}[j = 1, \ldots, d]$ and $f^{i,j}(x) = g^j(x_j)$ for some functions $g^j: \mathbb{R} \mapsto \mathbb{R}$ for $j = 1, \ldots, d\}$.

The estimate $\hat{\mathbf{m}}(x)$ defines an element of $\mathscr{F}_{\text{full}}$ by putting $f^{i,j}(x) = \hat{m}^j(x)$, $j = 0, 1, \ldots, d$. It is easy to see that $\hat{\mathbf{m}}$ is the orthogonal projection, with respect to $\| \ \|_*$, of $\mathbf{Y}$ onto $\mathscr{F}_{\text{full}}$. Below we introduce our version $\tilde{\mathbf{m}}$ of the backfitting estimator as the orthogonal projection of $\hat{\mathbf{m}}$ onto $\mathscr{F}_{\text{add}}$ (with respect to $\| \ \|_*$). For an understanding of $\tilde{\mathbf{m}}$ it will be essential that it is the orthogonal projection of $\mathbf{Y}$ onto $\mathscr{F}_{\text{add}}$. For the definition of such norms and linear spaces for higher order local polynomials and for other smoothers we refer to Mammen, Marron, Turlach and Wand (1997). Each local polynomial estimator corresponds to a specific choice of inner product in a Hilbert space, and the definition of the corresponding additive estimators is then the projection further down on $\mathscr{F}_{\text{add}}$. In particular, for the local constant estimator (Nadaraya–Watson-like smoothers) one chooses

$$\mathscr{F} = \{(f^i: i = 1, \ldots, n): \text{Here, } f^i \text{ are functions from } \mathbb{R}^d \text{ to } \mathbb{R}\},$$

$$\mathscr{F}_{\text{full}} = \{f \in \mathscr{F}: f^i \text{ does not depend on } i\},$$

$$\mathscr{F}_{\text{add}} = \{f \in \mathscr{F}_{\text{full}}: f^i(x) = g_1(x_1) + \cdots + g_d(x_d) \text{ for some functions } g_j: \mathbb{R} \to \mathbb{R}\},$$

$$\|f\|_*^2 = \int \frac{1}{n} \sum_{i=1}^{n} [f^i(x)]^2 \prod_{j=1}^{d} K_h(X_j^i - x_j)\, dx.$$

Note that for functions $\mathbf{m}$ in $\mathscr{F}_{\text{full}}$ (i.e., $m := m^1 = \cdots = m^n$) we get

$$\|\mathbf{m}\|_*^2 = \int m(x)^2 \hat{p}(x)\, dx,$$

where $\hat{p}(x) = n^{-1} \sum_{i=1}^{n} \{\prod_{j=1}^{d} K_h(X_j^i - x_j)\}$ is the kernel density estimate of the design density. In particular, in this case $\tilde{\mathbf{m}}$ is the projection of the full-dimensional Nadaraya–Watson estimate onto the subspace of additive functions with respect to the norm of the space $\mathbf{L}_2(\hat{p})$. We give a slightly

different motivation for the projection estimate $\tilde{\mathbf{m}}$ in the next section; see (7). There we will discuss the case of local constant smoothing in detail.

3. Estimation with Nadaraya–Watson-like smoothers. In this section we will discuss how our projection idea can be applied to define Nadaraya–Watson backfitting smoothers. The first subsection will give details about the implementation for the Nadaraya–Watson smoother. In the second subsection we will discuss asymptotic properties of our backfitting estimates. This will be done for a more general setup than Nadaraya–Watson smoothing. We will show that the backfitting algorithm converges numerically and we will give simple expansions for the stochastic and deterministic part of the backfitting estimate. The conditions under which these expansions hold will be verified in Section 5 for Nadaraya–Watson smoothers in both an i.i.d. and an autoregression setting. The expansions will imply that the asymptotic variance of our estimate does not depend on the number of additive components (and that in particular, they coincide with the case of only one component). Furthermore, the asymptotic bias is given by a simple geometric operation. It is the projection of the usual asymptotic bias expansion of a full-dimensional estimate onto the space of additive functions.

3.1. *A backfitting Nadaraya–Watson estimator.* In this subsection we will motivate our backfitting estimate for Nadaraya–Watson regression smoothers with product kernels,

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} \prod_{l=1}^{d} K_h(x_l - X_l^i) Y^i}{\sum_{i=1}^{n} \prod_{l=1}^{d} K_h(x_l - X_l^i)}. \tag{3}$$

The specific choice of (3) is not so important. One can show that the discussion of this subsection can be extended to smoothers that have the ratio form

$$\hat{m}(x) = \frac{\hat{r}(x)}{\hat{p}(x)}, \tag{4}$$

where $\hat{p}(x)$ is an estimator of $p(x)$, the marginal density of X, which depends only on $\mathscr{X}^n = \{X^1, \ldots, X^n\}$. The assumption that the pilot estimate $\hat{m}$ exists (i.e., is everywhere and always finite uniformly in n with probability tending to 1) will be dropped in our asymptotic analysis in the next section, which will allow us to include the case of high dimensions d. We assume for the most part that

$$m(x) = m_0 + m_1(x_1) + \cdots - m_d(x_d), \tag{5}$$

for some functions $m_j(\cdot)$, $j = 1, \ldots, d$ and constant m_0, although our definitions make sense more generally, that is, when the regression function is not additive, in which case the asymptotic behavior of our estimate is more difficult to analyze. For identifiability we assume that

$$\int m_j(x_j) p_j(x_j)\, dx_j = 0, \qquad j = 1, \ldots, d, \tag{6}$$

where $p_j(\cdot)$ is the marginal density of X_j. Denote also the marginal density of (X_j, X_k) by $p_{jk}(\cdot,\cdot)$, respectively $(j, k = 1, \ldots, d)$. The vector $(X_k\colon k \neq j)$ is denoted by X_{-j} and its Lebesgue density by p_{-j}.

Recall that backfitting is motivated as solving an empirical version of the set of equations

$$\begin{aligned} m_1(x_1) &= E(Y|X_1 = x_1) - m_0 - E\{m_2(X_2)|X_1 = x_1\} \\ &\quad - \cdots - E\{m_d(X_d)|X_1 = x_1\}, \\ &\vdots = \vdots \\ m_d(x_d) &= E(Y|X_d = x_d) - m_0 - E\{m_1(X_1)|X_d = x_d\} \\ &\quad - \cdots - E\{m_{d-1}(X_{d-1})|X_d = x_d\}. \end{aligned}$$

With only sample information available, one replaces the population quantity $E(Y|X_j = x_j)$ by one-dimensional smoothers $\hat{m}_j(\cdot)$, and iterates from some arbitrary starting values for $m_j(\cdot)$; see Hastie and Tibshirani [1991, page 108]. Let $\hat{p}(x) = n^{-1}\sum_{i=1}^n \prod_{l=1}^d K_h(x_l - X_l^i)$ be the multidimensional kernel density estimate and let $\hat{m}(x)$ be the multidimensional Nadaraya–Watson estimate as defined in (3). We define the "empirical projection" estimates $\{\tilde{m}_j(\cdot),\ j = 0, \ldots, d\}$ as the minimizers of the following criterion:

$$\|\hat{m} - \overline{m}\|_{\hat{p}}^2 = \int \left[\hat{m}(x) - \overline{m}_0 - \overline{m}_1(x_1) - \cdots - \overline{m}_d(x_d)\right]^2 \hat{p}(x)\, dx, \tag{7}$$

where the minimization runs over all functions $\overline{m}(x) = \overline{m}_0 + \sum_j \overline{m}_j(x_j)$, with $\int \overline{m}_j(x_j)\hat{p}_j(x_j)\, dx_j = 0$, where $\hat{p}_j(x_j) = \int \hat{p}(x)\, dx_{-j}$ is the marginal of the density estimate $\hat{p}(x)$. This is the one-dimensional density estimate $\hat{p}_j(x_j) = n^{-1}\sum_{i=1}^n K_h(x_j - X_j^i)$. A minimizer of (7) exists if the density estimate $\hat{p}$ is nonnegative. Equation (7) means that $\tilde{m}(x) = \tilde{m}_0 + \tilde{m}_1(x_1) + \cdots + \tilde{m}_j(x_d)$ is the projection in the space $\mathbf{L}_2(\hat{p})$ of $\hat{m}$ onto the subspace of additive functions $\{m \in \mathbf{L}_2(\hat{p})\colon\ m(x) = m_0 + m_1(x_1) + \cdots + m_d(x_d)\}$. This is a central point of our thesis. For projection operators, backfitting is well understood (as a method of alternating projections; see below). Therefore, this interpretation will enable us to understand convergence of the backfitting algorithm and the asymptotics of $\tilde{m}_j$. We remark that not every backfitting algorithm based on iterative smoothing can be interpreted as an alternating projection method.

The solution to (7) is characterized by the following system of equations $(j = 1, \ldots, d)$:

$$\tilde{m}_j(x_j) = \int \hat{m}(x)\frac{\hat{p}(x)}{\hat{p}_j(x_j)}\, dx_{-j} - \sum_{k \neq j} \int \tilde{m}_k(x_k)\frac{\hat{p}(x)}{\hat{p}_j(x_j)}\, dx_{-j} - \tilde{m}_0, \tag{8}$$

$$0 = \int \tilde{m}_j(x_j)\hat{p}_j(x_j)\, dx_j. \tag{9}$$

Straightforward algebra gives

$$\int \hat{m}(x)\frac{\hat{p}(x)}{\hat{p}_j(x_j)}\,dx_{-j} = \frac{n^{-1}\sum_{i=1}^{n} K_h\big(x_j - X_j^i\big)Y^i}{\hat{p}_j(x_j)} \equiv \hat{m}_j(x_j), \tag{10}$$

because of $\int \prod_{l\neq j} K_h(x_l - X_l^i)\,dx_{-j} = 1$, where $\hat{m}_j(x_j)$ is exactly the corre-3sponding univariate Nadaraya–Watson estimator. Furthermore, $\tilde{m}_0 = \int \hat{m}(x)\hat{p}(x)\,dx$, and because of $\int \prod_{l=1}^{d} K_h(x_l - X_l^i)\,dx_{-j} = 1$, we find, as in Hastie and Tibshirani (1991), that $\tilde{m}_0 = n^{-1}\sum_{i=1}^{n} Y^i$, that is, that $\tilde{m}_0$ is the sample mean. Therefore, $\tilde{m}_0$ is a $\sqrt{n}$-consistent estimate of the population mean and the randomness from this estimation is of smaller order and can be effectively ignored. Note also that

$$\tilde{m}_0 = \int \hat{m}_j(x_j)\,\hat{p}_j(x_j)\,dx_j \quad \text{for } j = 1,\ldots,d. \tag{11}$$

We therefore define a backfitting estimator $\tilde{m}_j(x_j)$, $j = 1,\ldots,d$, as a solution to the system of equations $[j = 1,\ldots,d]$

$$\tilde{m}_j(x_j) = \hat{m}_j(x_j) - \sum_{k\neq j}\int \tilde{m}_k(x_k)\frac{\hat{p}(x)}{\hat{p}_j(x_j)}\,dx_{-j} - \tilde{m}_0,$$

$$0 = \int \tilde{m}_j(x_j)\,\hat{p}_j(x_j)\,dx_j.$$

with $\tilde{m}_0$ defined by (11). Up to now we have assumed that multivariate estimates of the density and of the regression function exist for all x. This assumption is not reasonable for large dimensions d (or at least such estimates can perform very poorly). Furthermore, this assumption is not necessary. Note that (8) can be rewritten as

$$\tilde{m}_j(x_j) = \hat{m}_j(x_j) - \sum_{k\neq j}\int \tilde{m}_k(x_k)\frac{\hat{p}_{j,k}(x_j, x_k)}{\hat{p}_j(x_j)}\,dx_k - \tilde{m}_0, \tag{12}$$

where $\hat{p}_{j,k}(x_j, x_k) = n^{-1}\sum_{i=1}^{n} K_h(x_j - X_j^i)K_h(x_k - X_k^i)$ is the two-dimensional marginal of the full-dimensional kernel density estimate $\hat{p}(x)$. In this equation only one- and two-dimensional marginals of $\hat{p}$ are used.

Up to now we have implicitly assumed that the support of X is unbounded or at least that the density approaches zero at the boundary suitably fast. We now consider a generalization of the method which takes care of the boundary effects that are present when the densities have compact support. We do not require that (11) holds [i.e., $\int \hat{m}_j(x_j)\hat{p}_j(x_j)\,dx_j$ may depend on j], nor that $\hat{p}_j$ be a probability density, and we allow that $\hat{p}_j$ is not the marginal density of $\hat{p}_{j,k}$; that is, it may not hold for all $j \neq k$ that

$$\hat{p}_j(x_j) = \int \hat{p}_{j,k}(x_j, x_k)\,dx_k. \tag{13}$$

For instance, this may be the case for kernel density estimates of a density with compact support. For details see Section 5. For this more general setting

we want to find now an appropriate modification of (12). We rewrite (12) as

$$\tilde{m}_j(x_j) = \hat{m}_j(x_j) - \sum_{k \neq j} \int \tilde{m}_k(x_k) \frac{\tilde{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)}\, dx_k - \tilde{m}_{0,j}, \tag{14}$$

where $\tilde{m}_{0,j}$ is chosen such that $\int \tilde{m}_j(x_j)\hat{p}_j(x_j)\, dx_j = 0$ for all j. Under the assumption of (11), (13) and $\int \hat{p}_j(x_j)\, dx_j = 1$, this gives (12). In general, (14) can be rewritten as

$$\begin{aligned} \tilde{m}_j(x_j) &= \hat{m}_j(x_j) - \tilde{m}_{0,j} \\ &\quad - \sum_{k \neq j} \int \tilde{m}_k(x_k) \left[\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} - \hat{p}_{k,[j+]}(x_k) \right] dx_k, \end{aligned} \tag{15}$$

where for $k \neq j$,

$$\hat{p}_{k,[j+]}(x_k) = \int \hat{p}_{jk}(x_j, x_k)\, dx_j \left[\int \hat{p}_j(x_j)\, dx_j \right]^{-1}, \tag{16}$$

$$\tilde{m}_{0,j} = \frac{\int \hat{m}_j(x_j)\hat{p}_j(x_j)\, dx_j}{\int \hat{p}_j(x_j)\, dx_j}. \tag{17}$$

In the next section we will discuss estimates $\tilde{m}_j$ that are defined by (15) along with their asymptotic properties. In practice, our backfitting algorithm works as follows. One starts with an arbitrary initial guess $\tilde{m}_j^{[0]}$ for $\tilde{m}_j$; for example $\tilde{m}_j^{[0]} = \hat{m}_j$ or $\tilde{m}_j^{[0]}$ is the marginal integration estimator of Linton and Nielsen (1995). In the jth step of the rth iteration cycle one puts

$$\begin{aligned} \tilde{m}_j^{[r]}(x_j) &= \hat{m}_j(x_j) - \sum_{k<j} \int \tilde{m}_k^{[r]}(x_k) \left[\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} - \hat{p}_{k,[j+]}(x_k) \right] dx_k \\ &\quad - \sum_{k>j} \int \tilde{m}_k^{[r-1]}(x_k) \left[\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} - \hat{p}_{k,[j+]}(x_k) \right] dx_k - \tilde{m}_{0,j}, \end{aligned} \tag{18}$$

and the process is iterated until a desired convergence criterion is satisfied. The integrals are computed numerically; see Section 4 below for further comments.

3.2. *Asymptotics for the Nadaraya–Watson-like estimator*. We now consider estimates $\tilde{m}_j$ that are defined by (15), where $\hat{m}_j$, $\hat{p}_{jk}$ and $\hat{p}_j$ are some given estimates. The next theorem gives conditions under which, with probability tending to 1, there exists a solution $\tilde{m}_j$ of (15) that is unique and that can be calculated by backfitting. Furthermore, the backfitting algorithm converges with geometric rate. Our assumptions, given below, are "high-level" and only refer to properties of $\hat{m}_j$, $\hat{p}_{jk}$ and $\hat{p}_j$ [e.g., we do not require that p be the underlying density of X or that $\hat{m}_j$, $\hat{p}_{jk}$, and $\hat{p}_j$ are kernel estimates] —these properties can be verified for a range of smoothers under quite general heterogeneous and dependent sampling schemes, as we show in

Section 5. In the sequel, all integrals are taken over the support of the relevant variables. We use the convention that $0/0 = 0$.

ASSUMPTIONS. We suppose that there exists a density function p on $\mathbb{R}^d$ with marginals

$$p_j(x_j) = \int p(x)\, dx_{-j}$$

and

$$p_{j,k}(x_j, x_k) = \int p(x)\, dx_{-(j,k)} \quad \text{for } j \neq k.$$

(A1) For all $j \neq k$, it holds that

$$\int \frac{p_{j,k}^2(x_j, x_k)}{p_k(x_k)\, p_j(x_j)}\, dx_j\, dx_k < \infty.$$

(A2) For all $j \neq k$, it holds that

$$\int \left[\frac{\hat{p}_j(x_j) - p_j(x_j)}{p_j(x_j)} \right]^2 p_j(x_j)\, dx_j = o_P(1),$$

$$\int \left[\frac{\hat{p}_{j,k}(x_j, x_k)}{p_k(x_k)\, p_j(x_j)} - \frac{p_{j,k}(x_j, x_k)}{p_k(x_k)\, p_j(x_j)} \right]^2 p_k(x_k)\, p_j(x_j)\, dx_j\, dx_k = o_P(1),$$

$$\int \left[\frac{\hat{p}_{j,k}(x_j, x_k)}{p_k(x_k)\, \hat{p}_j(x_j)} - \frac{p_{j,k}(x_j, x_k)}{p_k(x_k)\, p_j(x_j)} \right]^2 p_k(x_k)\, p_j(x_j)\, dx_j\, dx_k = o_P(1).$$

Furthermore, $\hat{p}_j$ vanishes outside the support of p_j, $\hat{p}_{j,k}$ vanishes outside the support of $p_{j,k}$ and $\hat{p}_{j,k}(x_j, x_k) = \hat{p}_{k,j}(x_k, x_j)$.

(A3) There exists a finite constant C such that with probability tending to 1 for all j

$$\int \hat{m}_j^2(x_j)\, p_j(x_j)\, dx_j \leq C.$$

(A4) For some finite intervals $S_j \subset \mathbb{R}$ that are contained in the support of p_j $[1 \leq j \leq d]$ we suppose that there exists a finite constant C such that with probability tending to 1 for all $j \neq k$,

$$\sup_{x_k \in S_k} \int \frac{\hat{p}_{j,k}^2(x_j, x_k)}{\hat{p}_k^2(x_k)\, p_j(x_j)}\, dx_j \leq C.$$

For the statement of our next assumption we suppose that the one-dimensional smoothers $\hat{m}_j$ can be decomposed as

$$\hat{m}_j = \hat{m}_j^A + \hat{m}_j^B.$$

For $s = A$ and $s = B$, we define $\tilde{m}_j^s$ as the solution of the following equation:

$$\begin{aligned}\tilde{m}_j^s(x_j) &= \hat{m}_j^s(x_j) \\ &\quad - \sum_{k \neq j} \int \tilde{m}_k^s(x_k) \left[\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} - \hat{p}_{k,[j+]}(x_k) \right] dx_k \\ &\quad - \tilde{m}_{0,j}^s,\end{aligned} \tag{19}$$

where $\tilde{m}_{0,j}^s = \int \hat{m}_j^s(x_j) \hat{p}_j(x_j)\, dx_j / \int \hat{p}_j(x_j)\, dx_j$. Existence and uniqueness of $\tilde{m}_j^A$ and $\tilde{m}_j^B$ is stated in the next theorem (using the following assumption). Note that $\tilde{m}_j^s$ is defined as $\tilde{m}_j$ in (15) with $\hat{m}_j$ replaced by $\hat{m}_j^s$. We get that $\tilde{m}_j = \tilde{m}_j^A + \tilde{m}_j^B$.

(A5) There exists a finite constant C such that with probability tending to 1 for all j,

$$\int \hat{m}_j^A(x_j)^2 p_j(x_j)\, dx_j \leq C$$

and

$$\int \hat{m}_j^B(x_j)^2 p_j(x_j)\, dx_j \leq C.$$

In the applications of our results we will put $\hat{m}_j^A$ as the stochastic part and $\hat{m}_j^B$ as the expectation part of $\hat{m}_j$ (or in case of a random design, as the conditional expectation of $\hat{m}_j$ given the design.) In particular, in the case of Nadaraya–Watson smoothing of i.i.d. tuples (X^i, Y^i) with $Y^i = m(X^i) + \varepsilon^i$ where ε^i is mean zero, we will put $\hat{m}_j^A(x_j) = n^{-1} \sum_{i=1}^n K_h(x_j - X_j^i) \varepsilon^i / \hat{p}_j(x_j)$ and $\hat{m}_j^B(x_j) = n^{-1} \sum_{i=1}^n K_h(x_j - X_j^i) m(X^i) / \hat{p}_j(x_j)$. Note that (in this case) conditions on $\hat{m}_j^A$ and $\hat{m}_j^B$ are easy to verify (because only one-dimensional smoothing is applied) whereas conditions on $\tilde{m}_j^A$ and $\tilde{m}_j^B$ are harder to treat because these variables are defined only implicitly. The next assumption states a condition on $\hat{m}_j^A$ that can be used to treat the stochastic part $\tilde{m}_j^A$.

(A6) We suppose that for a sequence $\Delta_n \to 0$, the first component $\hat{m}_j^A$ satisfies for $j \neq k$,

$$\sup_{x_k \in S_k} \left| \int \frac{\hat{p}_{j,k}(x_j, x_k)}{\hat{p}_k(x_k)} \hat{m}_j^A(x_j)\, dx_j \right| = o_P(\Delta_n), \tag{20}$$

$$\left\| \int \frac{\hat{p}_{j,k}(x_j, x_k)}{\hat{p}_k(x_k)} \hat{m}_j^A(x_j)\, dx_j \right\|_2 = o_P(\Delta_n). \tag{21}$$

where $\| \cdots \|_2$ denotes the norm in the space $L_2(p_k)$. For simplicity of notation the index k is suppressed in the notation. The sets S_k have been introduced in (A4).

For the expectation term $\tilde{m}_j^B$ we suppose in the following assumption that it stabilizes asymptotically around a nonrandom term. Below we will give

assumptions on $\hat{m}_j^B$ that are easier to check and that will imply the condition on $\tilde{m}_j^B$.

(A7) We suppose that there exist (deterministic) functions $\mu_{n,j}(\cdot)$ such that the term $\tilde{m}_j^B$ satisfies

$$\sup_{x_j \in S_j} \left| \tilde{m}_j^B(x_j) - \mu_{n,j}(x_j) \right| = o_P(\Delta_n),$$

where the sets S_j are introduced in assumption (A4).

These conditions, which we discuss further below, are all straightforward to verify, except (A7). They are weaker than those made by Opsomer and Ruppert (1997); in particular, we do not restrict the dependence between the covariates.

The following result is crucial in establishing the asymptotic properties of the estimates.

THEOREM 1 (Convergence of backfitting). *Suppose that conditions* A1–A3 *hold. Then, with probability tending to* 1, *there exists a solution* $\tilde{m}_j$ *of* (15) *that is unique. Furthermore, there exist constants* $0 < \gamma < 1$ *and* $c > 0$ *such that, with probability tending to* 1, *the following inequality holds*:

$$\begin{aligned} &\int \left[\tilde{m}_j^{[r]}(x_j) - \tilde{m}_j(x_j) \right]^2 p_j(x_j)\, dx_j \\ &\qquad \le c\gamma^{2r} \left(1 + \sum_{j=1}^{d} \int \left\{ \tilde{m}_j^{[0]}(x_j) \right\}^2 p_j(x_j)\, dx_j \right). \end{aligned} \tag{22}$$

Here, the functions $\tilde{m}_1^{[0]}(x_1), \ldots, \tilde{m}_d^{[0]}(x_d)$ *are the starting values of the backfitting algorithm. For* $r > 0$ *the functions* $\tilde{m}_1^{[r]}(x_1), \ldots, \tilde{m}_d^{[r]}(x_d)$ *are defined by* (18).

Furthermore, for $s = A$ *and* $s = B$ *under the additional assumption of* (A5), *with probability tending to* 1 *there exists a solution* $\tilde{m}_j^s$ *of* (19) *that is unique.*

Our next theorem states that the stochastic part of the backfitting estimate is easy to understand. It coincides with the stochastic part of a one-dimensional smooth. Therefore, for an understanding of the asymptotic properties of the backfitting estimate it remains to study its asymptotic bias. This will be done after the theorem under additional assumptions.

THEOREM 2. *Suppose that conditions* (A1)–(A6) *hold for a sequence* Δ_n *and intervals* S_j $(1 \le j \le n)$. *Then it holds that*

$$\sup_{x_j \in S_j} \left| \tilde{m}_j^A(x_j) - \left[\hat{m}_j^A(x_j) - \tilde{m}_{0,j}^A \right] \right| = o_P(\Delta_n).$$

If in addition (A7) *holds, then one gets*

$$\sup_{x_j \in S_j} \left| \tilde{m}_j(x_j) - \left[\hat{m}_j^A(x_j) - \tilde{m}_{0,j}^A + \mu_{n,j}(x_j) \right] \right| = o_P(\Delta_n). \tag{23}$$

Typically the asymptotic stochastic behavior of $\hat{m}_j^A$ is easy to understand because it is a one-dimensional linear smoother. So if Δ_n is small enough, Theorem 2 gives the asymptotics of $\tilde{m}_j^A$. We will discuss this below in detail.

We come now to the study of the expectation term $\tilde{m}_j^B$. The asymptotic expectation $\mu_{n,j}(x_j)$ can be calculated by a projection under the following assumptions:

(A8) Suppose that for $j \neq k$,

$$\sup_{x_j \in S_j} \int \left| \frac{p_{j,k}(x_j, x_k)}{p_j(x_j) p_k(x_k)} - \frac{\hat{p}_{j,k}(x_j, x_k)}{\hat{p}_j(x_j) \hat{p}_k(x_k)} \right| p_k(x_k)\, dx_k = o_P(1). \tag{24}$$

(A9) There exist deterministic functions $\alpha_{n,1}(x_1), \ldots, \alpha_{n,d}(x_d)$, constants $\alpha_{n,0} \gamma_{n,1}, \ldots, \gamma_{n,d}$ and a function $\beta(x)$ (not depending on n), such that

$$\int \alpha_{n,j}(x_j)^2 p_j(x_j)\, dx_j < \infty,$$

$$\int \beta(x)^2 p(x)\, dx < \infty,$$

$$\sup_{x_1 \in S_1, \ldots, x_d \in S_d} |\beta(x)| < \infty,$$

$$\int \alpha_{n,j}(u) \hat{p}_j(u)\, du = \gamma_{n,j} - o_P(\Delta_n),$$

$$\sup_{x_j \in S_j} \left| \hat{m}_j^B(x_j) - \hat{\mu}_{n,0} - \hat{\mu}_{n,j}(x_j) \right| = o_P(\Delta_n), \tag{25}$$

$$\int \left| \hat{m}_j^B(x_j) - \hat{\mu}_{n,0} - \hat{\mu}_{n,j}(x_j) \right|^2 p_j(x_j)\, dx_j = o_P\left(\Delta_n^2\right), \tag{26}$$

for a random variable $\hat{\mu}_{n,0}$ and where

$$\hat{\mu}_{n,j}(x_j) = \alpha_{n,0} + \alpha_{n,j}(x_j) + \sum_{k \neq j} \int \alpha_{n,k}(x_k) \frac{\hat{p}_{j,k}(x_j, x_k)}{\hat{p}_j(x_j)}\, dx_k + \Delta_n \int \beta(x) \frac{p(x)}{p_j(x_j)}\, dx_{-j}.$$

We will discuss these assumptions after the following theorem.

THEOREM 3. *Suppose that conditions* (A1)–(A6), (A8), (A9) *hold. Define a constant* β_0 *and functions* β_j *on* $\mathbb{R}$ [*with* $\int \beta_j(x_j) p_j(x_j)\, dx_j = 0$] *by*

$$(\beta_0, \beta_1, \ldots, \beta_d) = \arg \min_{\beta_0, \ldots, \beta_d} \int \left[\beta(x) - \beta_0 - \beta_1(x_1) - \cdots - \beta_d(x_d) \right]^2 p(x)\, dx. \tag{27}$$

Then

$$\sup_{x_j \in S_j} \left| \tilde{m}_j^B(x_j) - \mu_{n,j}(x_j) \right| = o_P(\Delta_n),$$

where

$$\mu_{n,j}(x_j) = \alpha_{n,j}(x_j) - \gamma_{n,j} + \Delta_n \beta_j(x_j);$$

that is, (A7) *holds with this choice of* $\mu_{n,j}(x_j)$.

Theorems 2 and 3 give the asymptotic behavior of $\tilde{m}_j(x_j)$ in terms of Δ_n, $\hat{m}_j^A(x_j)$, $\alpha_{n,j}$ and $\beta_j(x_j)$, which quantities can be analyzed by standard techniques. In Section 5 we will verify conditions (A1)–(A6), (A8), (A9) for Nadaraya–Watson smoothing. In this case, as discussed in the last subsection, $\hat{m}_j(x_j)$ is defined as

$$\hat{m}_j(x_j) = \sum_{i=1}^{n} K_h\left(X_j^i - x_j\right) Y^i / \hat{p}_j(x_j) \tag{28}$$

and $\hat{p}_j$ and $\hat{p}_{jk}$ are kernel density estimates [of the densities of X_j and (X_j, X_k), respectively]. We will show that conditions (A1)–(A6), (A8), (A9) hold under the assumptions (B1)–(B7), stated there; see Theorem 4. This will be done with h of order $n^{-1/5}$ and kernels K with boundary corrections. It will turn out that the conditions hold with $\Delta_n = h^2$ and where $\alpha_{n,j}(x_j)$ is equal to $m_j(x_j)$ plus a correction term $O_P(h)$ at the boundary and where

$$\beta(x) = \sum_{j=1}^{d} \left[m_j'(x_j) \frac{\partial}{\partial x_j} \log p(x) + \frac{1}{2} m_j''(x_j) \right] \int u^2 K(u)\, du. \tag{29}$$

We remark that under strong conditions (that we do not apply here) $h^2\beta(x)$ is the asymptotic bias of a full-dimensional Nadaraya–Watson estimate. So Theorem 3 shows that the bias terms of the backfitting estimates are given by projections of the "theoretical" bias of a full-dimensional Nadaraya–Watson estimate.

In the discussion of Section 5 we will assume that the additive model (5) holds. The discussion of the expectation part $\tilde{m}_j^B$ becomes very complicated when the regression function is not additive. Then if the full-dimensional kernel density estimate $\hat{p}$ exists, one would expect that in first-order $\tilde{m}_1^B(x_1) + \cdots + \tilde{m}_d^B(x_d)$ is equivalent to the $L_2(\hat{p})$ projection of the regression function onto the space of additive functions. Because of the slow convergence of $\hat{p}$ to p we conjecture that this differ from the $L_2(p)$ projection by terms that are larger than $O_P(n^{-2/5})$.

4. Estimation with local polynomials. For simplicity of notation we consider only local linear smoothing. All arguments and theoretical results given for this special case can be generalized to local polynomials of higher degree.

Define the matrices [of dimension $n \times (d+1)$ and $n \times n$, respectively]

$$\mathbf{X}(x) = \begin{pmatrix} 1 & \dfrac{X_1^1 - x_1}{h} & \cdots & \dfrac{X_d^1 - x_d}{h} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \dfrac{X_1^n - x_1}{h} & \cdots & \dfrac{X_d^n - x_d}{h} \end{pmatrix},$$

$$\mathbf{K}(x) = \frac{1}{n}\operatorname{diag}\left(\prod_{l=1}^{d} K_h(X_l^1 - x_l), \ldots, \prod_{l=1}^{d} K_h(X_l^n - x_l) \right). \tag{30}$$

With these quantities the local linear estimate $\hat{\mathbf{m}}(x)$ is defined as

$$\hat{\mathbf{m}}(x) = \left\{ \mathbf{X}(x)^T \mathbf{K}(x) \mathbf{X}(x) \right\}^{-1} \mathbf{X}(x)^T \mathbf{K}(x) \mathbf{Y} \equiv \hat{\mathbf{V}}^{-1}(x) \hat{\mathbf{R}}(x), \tag{31}$$

where $\mathbf{Y} = (Y^1, \ldots, Y^n)^T$, $\hat{\mathbf{V}}(x) = \mathbf{X}(x)^T \mathbf{K}(x) \mathbf{X}(x)$ and $\hat{\mathbf{R}}(x) = \mathbf{X}(x)^T \mathbf{K}(x) \mathbf{Y}$.

Backfitting estimators based on local polynomials can be written in the form of (7) by choosing $\hat{p}(x) = \hat{V}_{0,0}(x) - \hat{\mathbf{V}}_{0,-0}^T(x) \hat{\mathbf{V}}_{-0,-0}^{-1}(x) \hat{\mathbf{V}}_{0,-0}(x)$, where

$$\hat{\mathbf{V}}(x) = \begin{pmatrix} \hat{V}_{0,0}(x) & \hat{\mathbf{V}}_{0,-0}(x) \\ \hat{\mathbf{V}}_{-0,0}(x) & \hat{\mathbf{V}}_{-0,-0}(x) \end{pmatrix} \equiv \mathbf{X}(x)^T \mathbf{K}(x) \mathbf{X}(x),$$

with the scalar $\hat{V}_{0,0}(x) = n^{-1} \sum_{i=1}^{n} \prod_{l=1}^{d} K_h(X_l^i - x_l)$, and $\hat{\mathbf{V}}_{-0,0}(x)$, $\hat{\mathbf{V}}_{-0,-0}(x)$ defined appropriately. This approach has two disadvantages. First, it may work only in low dimensions, since for the asymptotics, existence of the matrix $\hat{\mathbf{V}}_{-0,-0}^{-1}(x)$ and convergence of $\hat{\mathbf{V}}_{-0,-0}(x)$ is required under our assumptions (and this may hold only for low-dimensional argument x). Second, the corresponding backfitting algorithm does not consist of iterative local polynomial smoothing.

We now discuss another approach based on local polynomials that works in higher dimensions and that is based on iterative local polynomial smoothing. We motivate this approach for the case in which $\hat{\mathbf{V}}(x)$ does exist, but we will see that the definition of the backfitting estimate is based on only one- and two-dimensional "marginals" of $\hat{\mathbf{V}}(x)$. So its asymptotic treatment requires only consistency of these marginals, and the asymptotics work also for higher dimensions. This is similar to the discussion in the last section where consistency has been needed only for one- and two-dimensional marginals of the kernel density estimate $\hat{p}$.

For functions $f = (f^0, \ldots, f^d)$ with components f^j: $\mathbb{R}^d \mapsto \mathbb{R}$ and $d+1$ by $d+1$ positive (semi-)definite matrix function $M(\cdot)$, define the (semi-)norm

$$\|f\|_M = \int f(x)^T M(x) f(x)\, dx.$$

There is a one-to-one correspondence between functions f and functions in $\mathscr{F}_{\text{full}}$. Furthermore, taking $M = \hat{\mathbf{V}}$ we get that $\|\cdot\|_M$ is simply the seminorm induced by $\|\cdot\|_*$. In Section 2 our version $\tilde{\mathbf{m}}(x) = (\tilde{m}^0(x), \ldots, \tilde{m}^d(x))^T$ of the backfitting estimate was defined as the projection of (the function in $\mathscr{F}_{\text{full}}$ corresponding to) $\hat{\mathbf{m}}$ [see (1)] with respect to $\|\cdot\|_*$ onto the space $\mathscr{F}_{\text{add}}$. Therefore, $\tilde{\mathbf{m}}$ coincides with the $L_2(\hat{\mathbf{V}})$ projection, with respect to the (semi-)norm $\|f\|_{\hat{\mathbf{V}}}$, of $\hat{\mathbf{m}}$ onto the subspace $\mathscr{M}_{\text{add}}$, where

$$\begin{aligned}\mathscr{M}_{\text{add}} = \Big\{ &\mathbf{u}(x) = \big(u^0(x), \ldots, u^d(x)\big)^T \in \mathscr{M} \mid u^0(x) \\ &= u_0 - u_1(x_1) + \cdots + u_d(x_d), u^l(x) = w_l(x_l) \\ &\text{for } l = 1, \ldots, d, \text{ where } u_1, \ldots, u_d \text{ are functions } \mathbb{R} \to \mathbb{R} \\ &\text{with } \int \hat{V}^j_{0,0}(x_j) u_j(x_j)\, dx_j = 0 \text{ for } j = 1, \ldots, d, \text{ where } u_0 \\ &\text{is a constant and where } w_l \colon l = 1, \ldots, d \text{ are functions } \mathbb{R} \to \mathbb{R} \Big\},\end{aligned}$$

where for each j the $(d+1) \times (d+1)$ matrix $\hat{\mathbf{V}}^j(x_j) = \int \hat{\mathbf{V}}(x)\, dx_{-j}$ and where $\hat{V}^j_{l,l'}(x_j)$ $[0 \le l,\ l' \le d]$ denote the elements of $\hat{\mathbf{V}}^j(x_j)$. Note that the estimate $\hat{V}^j_{0,0}$ coincides with the marginal kernel density estimate $\hat{p}_j$ and that therefore the norming $\int \hat{V}^j_{0,0}(x_j) u_j(x_j)\, dx_j = 0$ makes sense. This norming makes the definition of the additive components u_j unique. (Clearly, the definition of the set $\mathscr{M}_{\text{add}}$ would not change if we omit this norming.) The class $\mathscr{M}_{\text{add}}$ contains functions that are additive in the first component (for $l = 0$) and where the other components (for $l = 1, \ldots, d$) depend only on a one-dimensional argument. A function f in $\mathscr{M}_{\text{add}}$ is specified by a constant f_0 and $2d$ functions $\mathbb{R} \mapsto \mathbb{R}$. Because f^l, $l = 1, \ldots, d$, depend only on one argument, in abuse of notation we write also $f^l(x_l)$ instead of $f^l(x)$. Note that there is a one-to-one correspondence between elements of $\mathscr{M}_{\text{add}}$ and $\mathscr{F}_{\text{add}}$.

We now discuss how $\tilde{\mathbf{m}}$ is calculated by backfitting. Note that $\tilde{\mathbf{m}}$ is defined as the minimizer of $\|\hat{m} - m\|_{\hat{\mathbf{V}}}$. Recall that this is equivalent to minimizing $\|\mathbf{Y} - \mathbf{m}\|^2_*$ over $\mathscr{F}_{\text{add}}$. We discuss now minimization of this quantity with respect to the jth components $m^j(x_j)$ and $m_0 + m_j(x_j)$. Define for each j,

$$\|f\|^2_j(x_j) = \int \frac{1}{n} \sum_{i=1}^{n} \left[f^{i,0}(x) + \sum_{j=1}^{d} f^{i,j}(x) \frac{x_j - X^i_j}{h} \right]^2 \prod_{j=1}^{d} K_h\big(X^i_j - x_j\big)\, dx_{-j}$$

and note the obvious fact that

$$\|f\|^2_* = \int \|f\|^2_j(x_j)\, dx_j, \qquad j = 1, \ldots, d.$$

Therefore, because such an integral is minimized by minimizing the integrand, our problem is solved by minimizing $\|\mathbf{Y} - \mathbf{m}\|^2_j(x_j)$, for fixed x_j, with respect to $m^j(x_j)$ and $m_0 + m_j(x_j)$, for $j = 1, \ldots, d$. After some standard

calculations, this leads to the following first order conditions:

$$\begin{aligned}
&\tilde{m}_j(x_j)\hat{V}^j_{0,0}(x_j) + \tilde{m}^j(x_j)\hat{V}^j_{j,0}(x_j) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} K_h\big(X_j^i - x_j\big)Y^i - \tilde{m}_0\hat{V}^j_{0,0}(x_j) \\
&\quad\quad - \sum_{l\neq j}\int \tilde{m}_l(x_l)\hat{V}^{l,j}_{0,0}(x_l, x_j)\,dx_l \\
&\quad\quad - \sum_{l\neq j}\int \tilde{m}^l(x_l)\hat{V}^{l,j}_{l,0}(x_l, x_j)\,dx_l,
\end{aligned} \tag{32}$$

$$\begin{aligned}
&\tilde{m}_j(x_j)\hat{V}^j_{j,0}(x_j) + \tilde{m}^j(x_j)\hat{V}^j_{j,j}(x_j) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} \frac{X_j^i - x_j}{h} K_h\big(X_j^i - x_j\big)Y^i - \tilde{m}_0\hat{V}^j_{J,0}(x_j) \\
&\quad\quad - \sum_{l\neq j}\int \tilde{m}_l(x_l)\hat{V}^{l,j}_{0,j}(x_l, x_j)\,dx_l \\
&\quad\quad - \sum_{l\neq j}\int \tilde{m}^l(x_l)\hat{V}^{l,j}_{l,j}(x_l, x_j)\,dx_l.
\end{aligned} \tag{33}$$

Here we have used one- and two-dimensional marginals of the matrix $\hat{\mathbf{V}}$,

$$\hat{\mathbf{V}}^r(x_r) = \int \hat{\mathbf{V}}(x)\,dx_{-r}, \tag{34}$$

$$\hat{\mathbf{V}}^{r,s}(x_r, x_s) = \int \hat{\mathbf{V}}(x)\,dx_{-(r,s)}. \tag{35}$$

The elements of these matrices are denoted by $\hat{V}^r_{p,q}(x_r)$ and $\hat{V}^{r,s}_{p,q}(x_r, x_s)$ with $p, q = 0, \ldots, d$. Together with the norming condition

$$\int \tilde{m}_j(x_j)\hat{V}^j_{0,0}(x_j)\,dx_j = 0, \tag{36}$$

(32) and (33) define $\tilde{m}_0$, $\tilde{m}_j$ and $\tilde{m}^j$ for given $\mathbf{Y}$ and $[\tilde{m}_l, \tilde{m}^l\colon l \neq j]$.

Equations (32), (33) and (36) can be rewritten as

$$\tilde{m}_j(x_j) = \hat{m}_j(x_j) - \check{m}_j(x_j), \tag{37}$$

$$\tilde{m}^j(x_j) = \hat{m}^j(x_j) + \check{m}^j(x_j), \tag{38}$$

where $\hat{m}_j(x_j)$, $\check{m}_j(x_j)$, $\hat{m}^j(x_j)$ and $\check{m}^j(x_j)$ are defined by

$$\hat{m}_j(x_j)\hat{V}^j_{0,0}(x_j) + \hat{m}^j(x_j)\hat{V}^j_{j,0}(x_j) = \frac{1}{n}\sum_{i=1}^{n} K_h\big(X_j^i - x_j\big)Y^i, \tag{39}$$

$$\hat{m}_j(x_j)\hat{V}^j_{j,0}(x_j) + \hat{m}^j(x_j)\hat{V}^j_{j,j}(x_j) = \frac{1}{n}\sum_{i=1}^{n}\frac{X^i_j - x_j}{h}K_h\big(X^i_j - x_j\big)Y^i, \tag{40}$$

$$\begin{aligned}\check{m}_j(x_j)\hat{V}^j_{0,0}(x_j) + \check{m}^j(x_j)\hat{V}^j_{j,0}(x_j)\\ = -\tilde{m}_0\hat{V}^j_{0,0}(x_j) - \sum_{l\neq j}\int \tilde{m}_l(x_l)\hat{V}^{l,j}_{0,0}(x_l, x_j)\,dx_l\\ - \sum_{l\neq j}\int \tilde{m}^l(x_l)\hat{V}^{l,j}_{l,0}(x_l, x_j)\,dx_l,\end{aligned} \tag{41}$$

$$\begin{aligned}\check{m}_j(x_j)\hat{V}^j_{j,0}(x_j) + \check{m}^j(x_j)\hat{V}^j_{j,j}(x_j)\\ = -\tilde{m}_0\hat{V}^j_{j,0}(x_j) - \sum_{l\neq j}\int \tilde{m}_l(x_l)\hat{V}^{l,j}_{0,j}(x_l, x_j)\,dx_l\\ - \sum_{l\neq j}\int \tilde{m}^l(x_l)\hat{V}^{l,j}_{l,j}(x_l, x_j)\,dx_l,\end{aligned} \tag{42}$$

$$\int \check{m}_j(x_j)\hat{V}^j_{0,0}(x_j)\,dx_j = -\int \hat{m}_j(x_j)\hat{V}^j_{0,0}(x_j)\,dx_j. \tag{43}$$

Note that $(\hat{m}_j, \hat{m}^j)$ is the one-dimensional local linear fit of the observations Y^i onto X^i_j.

Again, (37)–(43) define $\tilde{m}_0$, $\tilde{m}_j$ and $\tilde{m}^j$ for given $\mathbf{Y}$ and $[\tilde{m}_l, \tilde{m}^l\colon l \neq j]$. In the jth step of every cycle of the backfitting algorithm an update of $\tilde{m}_0$, $\tilde{m}_j$ and $\tilde{m}^j$ will be calculated by solving (37)–(43). In the next subsection we will discuss asymptotics for the backfitting estimate in a more general setup. In particular, there we will not assume that $(\hat{m}_l, \hat{m}^l)$ is a one-dimensional local linear fit nor that $\hat{\mathbf{V}}^l$ and $\hat{\mathbf{V}}^{l,l'}$ are motivated by local linear smoothing. Furthermore, we will not make any assumptions on the stochastic nature of the sample. For arbitrary choices of $(\hat{m}_l, \hat{m}^l)$, we will define $\tilde{m}_j$ and $\tilde{m}^j$ by

$$\begin{aligned}\hat{\mathbf{M}}_j(x_j)\begin{pmatrix}\{\tilde{m}_j - \hat{m}_j\}(x_j)\\ \{\tilde{m}^j - \hat{m}^j\}(x_j)\end{pmatrix} = -\tilde{m}_{0,j}\begin{pmatrix}\hat{V}^j_{0,0}(x_j)\\ \hat{V}^j_{j,0}(x_j)\end{pmatrix}\\ - \sum_{l\neq j}\int \hat{\mathbf{S}}_{l,j}(x_l, x_j)\begin{pmatrix}\tilde{m}_l(x_l)\\ \tilde{m}^l(x_l)\end{pmatrix}dx_l.\end{aligned} \tag{44}$$

$$\int \tilde{m}_j(x_j)\hat{V}^j_{0,0}(x_j)\,dx_j = 0, \tag{45}$$

where

$$\hat{\mathbf{M}}_j(x_j) = \begin{pmatrix}\hat{V}^j_{0,0}(x_j) & \hat{V}^j_{j,0}(x_j)\\ \hat{V}^j_{j,0}(x_j) & \hat{V}^j_{j,j}(x_j)\end{pmatrix}, \tag{46}$$

$$\hat{\mathbf{S}}_{l,j}(x_l, x_j) = \begin{pmatrix} \hat{V}_{0,0}^{l,j}(x_l, x_j) & \hat{V}_{l,0}^{l,j}(x_l, x_j) \\ \hat{V}_{j,0}^{l,j}(x_l, x_j) & \hat{V}_{l,j}^{l,j}(x_l, x_j) \end{pmatrix}. \tag{47}$$

Note that again as for Nadaraya–Watson smoothing we allow $\tilde{m}_{0,j}$ to depend on n. In particular, this may be the case if it does not hold that

$$\int \hat{V}_{r,s}^{l,j}(x_l, x_j)\, dx_j = \hat{V}_{r,s}^{l}(x_l) \tag{48}$$

for $r \in \{0, l\}$ and $s \in \{0, j\}$.

Not $\tilde{m}^j(x_j)$, but $h\,\tilde{m}^j(x_j)$ is an estimate of the derivative $m_j'(x_j)$ of $m_j(x_j)$. The reason is that in our definition of the seminorm $\| \ \|_*$ we have the linear term $f^{i,j}(x)(x_j - X_j^i)/h$ and not the term $f^{i,j}(x)(x_j - X_j^i)$; see (2) and see also the definition (30) of the matrix $X(x)$. Typically, estimates of derivatives have variance of order $(nh^3)^{-1}$, compared to the order $(nh)^{-1}$ for estimates of the functions itself. For this reason, one can show that, because of our norming by the factor h^{-1}, $\tilde{m}^j(x_j)$ has variance that is of the same asymptotic order as the variance of $\tilde{m}_j(x_j)$. The same holds for $\hat{m}^j(x_j)$. This is the reason why we have introduced the factor h^{-1} in $\| \ \|_*$ and $X(x)$.

Let us finish this section by some computational remarks.

1. The backfitting algorithm runs now with the following iteration step $(a = 0, 1, \ldots)$:

$$\begin{pmatrix} \hat{f}_j(x_j) \\ \tilde{m}^{[a+1],j}(x_j) \end{pmatrix} = \begin{pmatrix} \hat{m}_j(x_j) \\ \hat{m}^j(x_j) \end{pmatrix} - \hat{\mathbf{M}}_j(x_j)^{-1} \times \sum_{l \neq j} \int \hat{\mathbf{S}}_{l,j}(x_l, x_j) \begin{pmatrix} \tilde{m}_l^{[a]}(x_l) \\ \tilde{m}^{[a],l}(x_l) \end{pmatrix} dx_l. \tag{49}$$

$$\tilde{m}_j^{[a+1]}(x_j) = \hat{f}_j(x_j) - \int \hat{f}_j(u_j) \hat{V}_{0,0}^{j}(u_j)\, du_j. \tag{50}$$

2. For the case in which (48) holds, in a faster implementation, the norming of $\tilde{m}_j$ done in (50) could be omitted, that is, one could put $\tilde{m}_j^{[a+1]}(x_j) = f_j(x_j)$. After the final cycle all functions $\tilde{m}_j$ could be replaced by $\tilde{m}_j(x_j) - \int \tilde{m}_j(x_j) \hat{V}_{0,0}^{j}(x_j)\, dx_j$ and $\tilde{m}_0$ defined appropriately. It is easy to see that this algorithm does the same. If one is interested only in the estimation of the sum $m_0 + m_1(x_1) + \cdots + m_d(x_d)$, the final norming could be omitted or replaced by another norming.
3. A possible initialization of backfitting is given by putting $\tilde{m}_0 = 0$, $\tilde{m}_l = \hat{m}_l$ and $\tilde{m}^l = \hat{m}^l$ for $l = 1, \ldots, d$.
4. Note that the estimates $\hat{m}_l$ and $\hat{m}^l$ have to be calculated only at the beginning and do not have to be updated in each backfitting iteration.
5. For an implementation of backfitting, all estimates {i.e., $\hat{m}_l$, $\hat{m}^l$, $\check{m}_l$, $\check{m}^l$, $\tilde{m}_l$, $\tilde{m}^l$, $\hat{\mathbf{V}}^l$ and $\hat{\mathbf{V}}^{l,l'}$) have to be calculated on a grid and the integrals in (41) and (42) have to be replaced by averages. It should be emphasized that the grid need not coincide with the set of design points. In particular,

for large data sets it may not be necessary or desirable that it contain the same number of points.

4.1. *Asymptotics for local polynomials.* We discuss now asymptotics for the backfitting local polynomials estimate. As for Nadaraya–Watson smoothing, this will be done in a general setup. We assume that some estimates $\hat{m}_l$, $\hat{m}^l$, $\hat{\mathbf{V}}^l$ and $\mathbf{V}^{l,l'}$ $[l, l' = 1, \ldots, d]$ are given and that $\tilde{m}_{0,l}$, $\tilde{m}_l$ and $\tilde{m}^l$ $[l = 1, \ldots, d]$ are defined by (44)–(47). In particular, we will not assume that $(\hat{m}_l, \hat{m}^l)$ is a one-dimensional local linear fit and that $\hat{\mathbf{V}}^l$ and $\hat{\mathbf{V}}^{l,l'}$ are motivated by local linear smoothing. Furthermore, we will not make any assumptions on the stochastic nature of the sample.

ASSUMPTIONS. We suppose that there exists a density function p on $\mathbb{R}^d$ with marginals

$$p_j(x_j) = \int p(x)\, dx_{-j}$$

and

$$p_{j,k}(x_j, x_k) = \int p(x)\, dx_{-(j,k)} \quad \text{for } j \neq k$$

and a positive definite $(d+1) \times (d+1)$ (deterministic) matrix $\mathbf{W}$ with elements $W_{r,s}$: $0 \leq r,\ s, \leq d$. We define $\hat{\mathbf{M}}_j(x_j)$ and $\hat{\mathbf{S}}_{l,j}(x_l, x_j)$ as in (46) and (47) and we put

$$\mathbf{M}_j(x_j) = \begin{pmatrix} W_{0,0} & W_{j,0} \\ W_{j,0} & W_{j,j} \end{pmatrix} p_j(x_j),$$

$$\mathbf{S}_{l,j}(x_l, x_j) = \begin{pmatrix} W_{0,0} & W_{l,0} \\ W_{j,0} & W_{l,j} \end{pmatrix} p_{l,j}(x_l, x_j).$$

We suppose that $W_{0,0} = 1$.

(A1′) For all $j \neq k$, it holds that

$$\int \frac{p_{j,k}^2(x_j, x_k)}{p_k(x_k)\, p_j(x_j)}\, dx_j\, dx_k < \infty.$$

(A2′) For all $j \neq k$, it holds that

$$\int \left[\frac{\hat{V}_{0,0}^j(x_j) - p_j(x_j)}{p_j(x_j)} \right]^2 p_j(x_j)\, dx_j = o_P(1),$$

$$\int \left[\frac{\hat{V}_{0,0}^{j,k}(x_j, x_k)}{p_k(x_k)\, p_j(x_j)} - \frac{p_{j,k}(x_j, x_k)}{p_k(x_k)\, p_j(x_j)} \right]^2 p_k(x_k)\, p_j(x_j)\, dx_j\, dx_k = o_P(1),$$

$$\int \left[\hat{\mathbf{M}}_j(x_j)^{-1} \hat{\mathbf{S}}_{k,j}(x_k, x_j) - \mathbf{M}_j(x_j)^{-1} \mathbf{S}_{k,j}(x_k, x_j) \right]_{r,s}^2 p_k(x_k)^{-1} p_j(x_j)\, dx_j\, dx_k = o_P(1)$$

for $r, s = 1, 2$. Here $[\cdots]_{r,s}$ denotes the (r, s) element of a matrix $[\cdots]$. Furthermore, $\hat{\mathbf{M}}_j$ vanishes outside the support of p_j, $\hat{\mathbf{S}}_{j,k}$ vanishes outside the support of $p_{j,k}$ and $\hat{\mathbf{S}}_{j,k}(x_j, x_k)^T = \hat{\mathbf{S}}_{k,j}(x_k, x_j)$.

(A3′) There exists a constant C such that with probability tending to 1 for all j,

$$\int \hat{m}_j(x_j)^2 p_j(x_j)\, dx_j \le C$$

and

$$\int \hat{m}^j(x_j)^2 p_j(x_j)\, dx_j \le C.$$

(A4′) For some finite intervals $S_j \subset \mathbb{R}$ that are contained in the support of p_j $[1 \le j \le d]$ we suppose that there exists a finite constant C such that with probability tending to 1 for all $j \ne k$,

$$\sup_{x_j \in S_j} \int \operatorname{trace}\left[\hat{\mathbf{S}}_{k,j}(x_k, x_j)\hat{\mathbf{M}}_j(x_j)^{-2}\hat{\mathbf{S}}_{k,j}(x_k, x_j)\right] p_k(x_k)^{-1}\, dx_k \le C.$$

We decompose the smoothers $\hat{m}_j$ and $\hat{m}^j$ as $\hat{m}_j = \hat{m}_j^A + \hat{m}_j^B$ and $\hat{m}^j = \hat{m}^{j,A} + \hat{m}^{j,B}$. For $s = A$ and $s = B$ we define $\tilde{m}_{0,j}^s$, $\tilde{m}_j^s$ and $\tilde{m}^{j,s}$ as the solution of the following equations:

$$\begin{aligned}
&\hat{\mathbf{M}}_j(x_j)\begin{pmatrix} \{\tilde{m}_j^s - \hat{m}_j^s\}(x_j) \\ \{\tilde{m}^{j,s} - \hat{m}^{j,s}\}(x_j) \end{pmatrix} \\
&\quad = -\tilde{m}_{0,j}^s \begin{pmatrix} \hat{\mathbf{V}}_{0,0}^j(x_j) \\ \hat{\mathbf{V}}_{j,0}^j(x_j) \end{pmatrix} - \sum_{l \ne j} \int \hat{\mathbf{S}}_{l,j}(x_l, {}_j) \begin{pmatrix} \tilde{m}_l^s(x_l) \\ \tilde{m}^{l,s}(x_l) \end{pmatrix} dx_l,
\end{aligned} \tag{51}$$

$$\int \tilde{m}_j^s(x_j)\hat{V}_{0,0}^j(x_j)\, dx_j = 0. \tag{52}$$

Existence and uniqueness of $\tilde{m}_j^A$, $\tilde{m}_j^B$, $\tilde{m}^{j,A}$ and $\tilde{m}^{j,B}$ is stated in the next theorem. Note that $(\tilde{m}_j^s, \tilde{m}^{j,s})$ is defined as $(\tilde{m}_j, \tilde{m}^j)$ in (44) and (45) with $(\hat{m}_j, \hat{m}^j)$ replaced by $(\hat{m}_j^s, \hat{m}^{j,s})$.

(A5′) There exists a constant C such that with probability tending to 1 for all j,

$$\int \hat{m}_j^s(x_j)^2 p_j(x_j)\, dx_j \le C, \qquad s = A, B$$

and

$$\int \hat{m}^{j,s}(x_j)^2 p_j(x_j)\, dx_j \le C, \qquad s = A, B.$$

In the applications of our results we will put $(\hat{m}_j^A, \hat{m}^{j,A})$ as the stochastic part and $(\hat{m}_j^B, \hat{m}^{j,B})$ as the expectation part of $(\hat{m}_j, \hat{m}^j)$ [or in case of a random design, as the conditional expectation of $(\hat{m}_j, \hat{m}^j)$ given the design].

In particular, in the case of local linear smoothing of i.i.d. tuples (X^i, Y^i) with $Y^i = m(X^i) + \varepsilon^i$ where ε^i is mean zero, $(\hat{m}_j^A, \hat{m}^{j,A})$ is the local linear fit to (X_j^i, ε^i) and $(\hat{m}_j^B, \hat{m}^{j,B})$ is the local linear fit to $(X_j^i, m(X^i))$.

(A6′) We suppose that for a sequence Δ_n we have

$$\sup_{x_k \in S_k} \left\| \int \hat{\mathbf{M}}_k(x_k)^{-1} \hat{\mathbf{S}}_{k,j}(x_k, x_j) \begin{pmatrix} \hat{m}_j^A(x_j) \\ \hat{m}^{j,A}(x_j) \end{pmatrix} dx_j \right\|_2 = o_P(\Delta_n),$$

$$\left\| \int \hat{\mathbf{M}}_k(x_k)^{-1} \hat{\mathbf{S}}_{k,j}(x_k, x_j) \begin{pmatrix} \hat{m}_j^A(x_j) \\ \hat{m}^{j,A}(x_j) \end{pmatrix} dx_j \right\|_{\mathbf{M}_k,2} = o_P(\Delta_n),$$

where $\|\cdots\|_2$ denotes the L_2 norm in $\mathbb{R}^2$ and where for functions g: $\mathbb{R} \mapsto \mathbb{R}^2$ we define $\|g\|_{\mathbf{M}_k,2}^2 = \int g(u)^T \mathbf{M}_k(u) g(u)\, du$. The sets S_k have been introduced in (A4′).

For the expectation term $\tilde{m}_j^B$ we suppose in the following assumption that it stabilizes asymptotically around a nonrandom term. Below we will give assumptions on $(\tilde{m}_j^B, \tilde{m}^{j,B})$ that are easier to check and that will imply the condition on $\tilde{m}_j^B$.

(A7′) We suppose that there exist deterministic functions $\mu_{n,j}(\cdot)$ such that

$$\sup_{x_j \in S_j} \left| \tilde{m}_j^B(x_j) - \mu_{n,j}(x_j) \right|,$$

where the sets S_j have been introduced in assumption (A4′).

We remark again that these conditions are all straightforward to verify, except perhaps (A7′). Note that we shall not require $\hat{\mathbf{V}}(x)$ to converge in probability to $\mathbf{W}p(x)$, because this would be affected by the curse of dimensionality, a necessary condition would be that $nh^d \to \infty$ for kernel smoothing, which rules out the one-dimensional convergence rate when $d > 4$.

We state now results that are similar to the ones for Nadaraya–Watson smoothing in Section 3.

THEOREM 1′ (Convergence of backfitting). *Suppose that conditions* (A1′)–(A3′) *hold. Then, with probability tending to* 1, *there exists a solution* $[\tilde{m}_{0,l}, \tilde{m}_l, \tilde{m}^l\colon\ l = 1, \ldots, d]$ *of* (44)–(47) *that is unique. Furthermore, there exist constants* $0 < \gamma < 1$ *and* $c > 0$ *such that, with probability tending to* 1, *the following inequalities hold*:

$$\int \left[\tilde{m}_j^{[r]}(x_j) - \tilde{m}_j(x_j) \right]^2 p_j(x_j)\, dx_j \le c\gamma^{2r}\Gamma,$$

$$\int \left[\tilde{m}^{j,[r]}(x_j) - \tilde{m}^j(x_j) \right]^2 p_j(x_j)\, dx_j \le c\gamma^{2r}\Gamma,$$

where

$$\Gamma = 1 + \sum_{l=1}^{d} \int \left[\tilde{m}_l^{[0]}(x_l)\right]^2 p_l(x_l)\,dx_l + \int \left[\tilde{m}^{l,[0]}(x_l)\right]^2 p_l(x_l)\,dx_l.$$

Here, for $r = 0$ *the functions* $\tilde{m}_{0,l}^{[0]}$, $\tilde{m}_l^{[0]}$ *and* $\tilde{m}^{l,[0]}$ *are the starting values of the backfitting algorithm. For* $r > 0$ *the functions* $\tilde{m}_l^{[r]}$ *and* $\tilde{m}^{l,[r]}$ *are defined by* (49) *and* (50).

Furthermore, provided (A5′) *holds also, for* $s = A$ *and* $s = B$, *with probability tending to* 1, *there exists a solution* $[\tilde{m}_0^s\ \tilde{m}_j^s$ *and* $\tilde{m}^{j,s}\colon j = 1, \ldots, d]$ *of* (51)–(52) *that is unique.*

Just as Theorem 2 stated for Nadaraya–Watson smoothing, the stochastic part of the backfitting estimate coincides with a one-dimensional local linear fit. This is stated in the following theorem. Under conditions analogous to (59) we get the following result.

THEOREM 2′. *Suppose that conditions* (A1′)–(A6′) *hold for a sequence* Δ_n *and intervals* S_j $(1 \le j \le n)$. *Then it holds that*

$$\sup_{x_j \in S_j} \left| \tilde{m}_j^A(x_j) - \left[\hat{m}_j^A(x_j) - \tilde{m}_{0,j}^A\right] \right| = o_P(\Delta_n).$$

In addition, if (A7′) *holds, one gets*

$$\sup_{x_j \in S_j} \left| \tilde{m}_j(x_j) - \left[\hat{m}_j^A(x_j) - \tilde{m}_{0,j}^A + \mu_{n,j}(x_j)\right] \right| = o_P(\Delta_n). \tag{53}$$

We show now how the asymptotic expectation $\mu_{n,j}(x_j)$ can be calculated. This can be done by a more direct argument as for Nadaraya–Watson smoothing. We use the following assumptions:

(A8′) Suppose that for all $j \ne k$,

$$\sup_{x_j \in S_j} \int \left| \left[\hat{\mathbf{M}}_j(x_j)^{-1} \hat{\mathbf{S}}_{k,j}(x_k, x_j) - \mathbf{M}_j^{-1}(x_j)\mathbf{S}_{k,j}(x_k, x_j)\right]_{r,s} \right| p_k(x_k)\,dx_k = o_P(1).$$

for $r, s = 1, 2$.

(A9′) There exist deterministic functions $\alpha_{n,1}(x_1), \ldots, \alpha_{n,d}(x_d)$, $\alpha_n^1(x_1), \ldots, \alpha_n^d(x_d)$ and constants $\alpha_{n,0}$, $\gamma_{n,1}, \ldots, \gamma_{n,d}$ such that

$$\int \alpha_{n,j}(x_j)^2 p_j(x_j)\,dx_j < \infty,$$

$$\int \alpha_n^j(x_j)^2 p_j(x_j)\,dx_j < \infty,$$

$$\int \alpha_{n,j}(u)\hat{V}^j_{0,0}(u)\,du = \gamma_{n,j} + o_P(\Delta_n),$$

$$\sup_{x_j \in S_j} \left|\hat{m}^B_j(x_j) - \hat{\mu}_{n,0} - \hat{\mu}_{n,j}(x_j)\right| = o_P(\Delta_n),$$

$$\int \left|\hat{m}^B_j(x_j) - \hat{\mu}_{n,0} - \hat{\mu}_{n,j}(x_j)\right|^2 p_j(x_j)\,dx_j = o_P\left(\Delta_n^2\right),$$

$$\sup_{x_j \in S_j} \left|\hat{m}^{j,B}(x_j) - \hat{\mu}^0_n - \hat{\mu}^j_n(x_j)\right| = o_P(\Delta_n),$$

$$\int \left|\hat{m}^{j,B}(x_j) - \hat{\mu}^j_n(x_j)\right|^2 p_j(x_j)\,dx_j = o_P\left(\Delta_n^2\right),$$

for random variables $\hat{\mu}_{n,0}$ and where

$$\begin{pmatrix} \hat{\mu}_{n,j}(x_j) \\ \hat{\mu}^j_n(x_j) \end{pmatrix} = \begin{pmatrix} \alpha_{n,0} + \alpha_{n,j}(x_j) \\ \alpha^j_n(x_j) \end{pmatrix} + \sum_{k \neq j} \int \hat{\mathbf{M}}_j(x_j)^{-1} \hat{\mathbf{S}}_{k,j}(x_k, x_j) \begin{pmatrix} \alpha_{n,k}(x_k) \\ \alpha^k_n(x_k) \end{pmatrix} dx_k.$$

THEOREM 3′. *Suppose that conditions* (A1′)–(A6′) (A8′), (A9′) *hold. Then*

$$\sup_{x_j \in S_j} \left|\tilde{m}^B_j(x_j) - \mu_{n,j}(x_j)\right| = o_P(\Delta_n),$$

$$\sup_{x_j \in S_j} \left|\tilde{m}^{j,B}(x_j) - \mu^j_n(x_j)\right| = o_P(\Delta_n),$$

where $\mu_{n,j}(x_j) = \alpha_{n,j}(x_j) - \gamma_{n,j}$ *and* $\mu^j_n(x_j) = \alpha^j_n(x_j)$. *In particular*, (A7′) *holds with this choice of* $\mu_{n,j}(x_j)$.

From Theorems 2′ and 3′ we get the asymptotic behavior of the backfitting estimates defined in (44)–(47). It turns out that for the local linear estimator itself, the conditions hold with $\Delta_n = h^2$, $\alpha_{n,j}(x_j) = m_j(x_j) + h^2 \frac{1}{2} m''_j(x_j) \int u^2 K(u)\,du$ and $\alpha^j_n(x_j) = h m'_j(x_j)$. We remark that under strong conditions (that we do not apply here) $\sum_{j=1}^d \alpha_{n,j}(x_j) - m(x)$ is the asymptotic bias of a full-dimensional local linear estimate.

5. Verification of conditions. We now provide sufficient conditions for (A1)–(A6), (A8), (A9) to hold in a time series setting for the Nadaraya–Watson smoother. We suppose that $\{Y^i, X^i\}_{i=1}^\infty$ is a jointly stationary process. This includes autoregression, where $X^i = (Y^{i-1}, \ldots, Y^{i-d})'$, and regular cross-sectional regression where X^i is of dimensions d and the joint process is i.i.d., as special cases. Let $\mathcal{F}_a^b$ be the σ-algebra of events generated by the random variables $\{Y^i, X^i;\ a \le j \le b\}$. The stationary processes $\{Y^i, X^i\}$ are

called strongly mixing [Rosenblatt (1956)] if

$$\sup_{A\in\mathcal{F}_{-\infty}^{0},\, B\in\mathcal{F}_{k}^{\infty}} \left|P(A\cap B)-P(A)P(B)\right| \equiv \alpha(k)\to 0 \quad \text{as } k\to\infty.$$

We assume that the additive model holds, that is,

$$E[Y|X=x] = m_0 + m_1(x_1) + \cdots + m_d(x_d) \tag{54}$$

for x in a compact set ($[0,1]^d$, say). For identifiability we suppose that $Em_j(X_j)\mathbf{1}(X_j\in[0,1]) = 0$. Let N be the number of points X^i that lie in $[0,1]^d$. We define

$$\hat{m}_j(x_j) = N^{-1}\sum_{i=1}^{n}\mathbf{1}\left(X^i\in[0,1]^d\right)K_h\left(x_j, X_j^i\right)Y^i/\hat{p}_j(x_j), \tag{55}$$

$$\hat{p}_j(x_j) = N^{-1}\sum_{i=1}^{n}\mathbf{1}\left(X^i\in[0,1]^d\right)K_h\left(x_j, X_j^i\right), \tag{56}$$

$$\hat{p}_{j,k}(x_j,x_k) = N^{-1}\sum_{i=1}^{n}\mathbf{1}\left(X^i\in[0,1]^d\right)K_h\left(x_j, X_j^i\right)K_h\left(x_k, X_k^i\right), \tag{57}$$

where now

$$K_h(u,v) = \mathbf{1}(u,v\in[0,1])\frac{K_h(u-v)}{\int_0^1 K_h(w-v)\,dw} \tag{58}$$

with, again, $K_h(u) = h^{-1}K(h^{-1}u)$. We will suppose that the kernel K has compact support $[-C_1, C_1]$, see (B1). For this reason we get that $K_h(u,v) = K_h(u-v)$ for $v\in[C_1, h, 1-C_1h]$ or for $u\in[2C_1h, 1-2C_1h]$. So $K_h(u,v)$ differs from $K_h(u-v)$ only on the boundary. This boundary modification of the kernel will be needed for the verification of assumption (A9). All other assumptions can be verified for the unmodified kernel $K_h(u-v)$. Assumption (A9) was needed to get an asymptotic expansion for the bias of $\tilde{m}_j$; see Theorem 3. The norming (58) gives that $\int_0^1 K_h(u,v)\,du = 1$. Therefore we have $\int_0^1\hat{p}_{j,k}(x_j,x_k)\,dx_k = \hat{p}_j(x_j)$ and $\int_0^1\hat{p}_j(x_j)\,dx_j = 1$. Because of these properties $\tilde{m}_j$ is defined by (12).

For simplicity of notation, again we assume that the kernels and the bandwidths do not depend on j.

(B1) The kernel K is bounded, has compact support ($[-C_1, C_1]$, say), is symmetric about zero and is Lipschitz continuous; that is, there exists a positive finite constant C_2 such that $|K(u)-K(v)|\le C_2|u-v|$.

(B2) The density q_0 of X^i and the densities $q_{0,l}$ of (X^i, X^{i+l}), $l=1,\ldots,$ are uniformly bounded. Furthermore, q_0 is bounded away from zero on $[0,1]$.

(B3) For some $\theta>2$, $E(|Y|^\theta)<\infty$. Let $\sigma_j^2(x_j) = \operatorname{var}[Y-m(X)|X_j=x_j]$.

(B4) The second partial derivatives of the function m exist and are Lipschitz continuous. The first partial derivatives of q_0 exist and are continuous.

(B5) The conditional densities $f_{X|Y}(x|y)$ of X given Y and $f_{X^i,X^{i+l}|Y^i,Y^{i+l}}(x^0,x^l|y^0,y^l)$ of (X^i,X^{i+l}) given (Y^i,Y^{i+l}), $l=1,\ldots,$ exist and are bounded from above.

(B6) The process $\{Y^i, X^i\}$ is strongly mixing with $\sum_{i=1}^{\infty} i^a\{\alpha(i)\}^{1-2/\nu} < \infty$ for some $2 < \nu \le \theta$ and $a > 1 - 2/\nu$.

(B7) The strong mixing coefficients satisfy $\sum_{j=1}^{\infty}\varphi(j;c) < \infty$ and $\sum_{j=1}^{\infty}\psi(j;c) < \infty$ for $c = 1, 2$, where $\varphi(n;c) = (nL_1(n)/r_1(n))(nT_n^2/h^c \log n)^{1/4}\alpha\{r_1(n)\}$ with $r_1(n) = (nh^c/T_n \log n)^{1/2}$ and $L_1(n) = (nT_n^2/h^{c+2}\log n)^{c/2}$ with $T_n = \{n \log n(\log\log n)^{1+\delta}\}^{1/\theta}$ for some $1 > \delta > 0$, while $\psi(n;c) = (nL_2(n)/r_2(n))(n/h^c \log n)^{1/4}\alpha\{r_2(n)\}$ with $r_2(n) = (nh^c/\log n)^{1/2}$ and $L_2(n) = (n/h^{c+2}\log n)^{c/2}$.

These conditions are slight modifications of assumptions used in Masry (1996a, b). We will use results of these papers to achieve the main results of this section. We conjecture that a direct proof works under weaker conditions.

When (Y^i, X^i) are i.i.d., we can dispense with (B5)–(B7), and replace (B2)–(B4) by:

(B2′) The d-dimensional vector X has compact support $[0, 1]^d$ and its density q_0 is bounded away from zero and infinity on $[0, 1]^d$.

(B3′) For some $\theta > 5/2$, $E(|Y|^{\theta}) < \infty$. Let $\sigma_j^2(x_j) = \operatorname{var}[Y - m(X)|X_j = x_j]$.

(B4′) The second partial derivatives of the function m exist and are continuous. The first partial derivatives of q_0 exist and are continuous.

Condition (B3′) ensures that $\sup_{1 \le i \le n}|Y_i| = o_P(n^{2/5})$. The following theorem could also be stated for the case of a stationary sequence (Y^i, X^i) where X^i has compact support.

THEOREM 4. *Suppose that the model* (54) *applies and that conditions* (B1)–(B7) *hold, or* (B1), (B2′), (B3′) *and* (B4′) *hold in the i.i.d. case, and that Nadaraya–Watson backfitting smoothing is used; that is,* $\hat{m}_j$, $\hat{p}_j$ *and* $\hat{p}_{j,k}$ *are defined according to* (55)–(57) *and* $\tilde{m}_j$ *is defined by* (12). *Suppose additionally that* $n_0^{1/5}h \to c_h$ *for a constant* c_h *with* $n_0 = EN = nP(X \in [0, 1]^d)$. *Then, for closed subsets* $S_1, \ldots, S_d$ *of* $(0, 1)$ *conditions* (A1)–(A6), (A8), (A9) *are satisfied with* $\Delta_n = h^2$, *with* β *as defined by* (29), *with* $\alpha_{n,j}(x_j) = m_j(x_j) + m_j'(x_j)\int K_h(x_j, u)(u - x_j)\,du[\int K_h(x_j, v)\,dh]^{-1}$, $\gamma_{n,j} = 0$, $p(x) = q_0(x)\mathbf{1}(x \in [0, 1]^d)/P(X \in [0, 1]^D)$, *and with* $\hat{m}_j^A(x_j) = N^{-1}\sum_{i=1}^n K_h(x_j, X_j^i)(Y^i - E[Y^i|X^i])/\hat{p}_j(x_j)$. *In particular, the uniform expansion* (23) *holds and the following convergence holds in distribution for any* $x_1, \ldots, x_d \in (0, 1)$,

$$n_0^{2/5}\begin{bmatrix} \tilde{m}_1(x_1) - m_1(x_1) \\ \vdots \\ \tilde{m}_d(x_d) - m_d(x_d) \end{bmatrix} \times N\left(\begin{bmatrix} c_h^2\beta_1(x_1) \\ \vdots \\ c_h^2\beta_d(x_d) \end{bmatrix}, \begin{bmatrix} v_1(x_1) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & v_d(x_d) \end{bmatrix}\right),$$

where β_j *is defined by* (27) *and where* $v_j(x_j) = c_h^{-1} c_K \sigma_j^2(x_j)/p_j(x_j)$, $j = 1, \ldots, d$ *with* $c_K = \int K(u)^2\, du$. *Consequently*,

$$n_0^{2/5}[\tilde{m}(x) - m(x)] \times N\left(c_h^2 \sum_{j=1}^{d} \beta_j(x_j),\ \sum_{j=1}^{d} v_j(x_j)\right).$$

It is illuminating to relate the estimate $\tilde{m}_j$ to the corresponding infeasible estimate $\ddot{m}_j$ that uses the knowledge of the other components m_l with $l \neq j$. Specifically, let $\ddot{m}_j(x_j)$ be the one-dimensional kernel smooth of the unobserved data $Y_*^i = Y^i - m_0 - \sum_{k \neq j} m_k(X_k^i)$ on X_j^i, thus

$$\ddot{m}_j(x_j) = \frac{\sum_{i=1}^n K_h\left(X_j^i, x_j\right) Y_*^i}{\sum_{i=1}^n K_h\left(X_j^i, x_j\right)}, \qquad j = 1, \ldots, d. \tag{59}$$

Under standard regularity conditions [see, e.g., Härdle (1991) for the i.i.d. case],

$$n_0^{2/5}\left\{\ddot{m}_j(x_j) - m_j(x_j)\right\} \times N\left(\ddot{b}_j(x_j), \ddot{v}_j(x_j)\right), \qquad j = 1, \ldots, d, \tag{60}$$

where $\ddot{b}_j(x_j) = c_h^2\{m_j'(x_j)p_j'(x_j)/p_j(x_j) + (1/2)m_j''(x_j)\}\int u^2 K(u)\, du$ and $\ddot{v}_j(x_j) = v_j(x_j)$. Define also the centered version of $\ddot{m}_j(x_j)$,

$$\ddot{m}_j^c(x_j) = \ddot{m}_j(x_j) - \frac{1}{N}\sum_{i=1}^{n} \ddot{m}_j\left(X_j^i\right)\mathbf{1}\left(X^i \in [0,1]^d\right), \tag{61}$$

which has the same asymptotic variance as $\ddot{m}_j(x_j)$ but bias $\ddot{b}_j^c(x) = \ddot{b}_j(x) - \int \ddot{b}_j(x)p_j(x_j)\, dx_j$. Because in the construction of $\ddot{m}_j^c$ knowledge of the other components is used, this estimate gives a target that we may not expect to beat by using $\tilde{m}_j$. We see that $\tilde{m}_j$ and the theoretical target estimate $\ddot{m}_j^c$ have the same asymptotic variance, whereas they differ in their asymptotic bias. We will see below that backfitting estimates based on local linear will have the same asymptotic bias and variance as their target estimate. The basic reason is that the function $\beta(x)$ is not additive whereas the corresponding function in the local linear case is. Recall that $\beta(x)$ corresponds to the asymptotic bias of the full-dimensional estimate $\hat{m}(x)$ and that it is well known that for the Nadaraya–Watson estimate the asymptotic bias depends on the design density p whereas for the local linear estimate it does not.

We next state the theorem for the local linear estimator. We define now the marginal estimates $\hat{m}_j(x_j)$ and $\hat{m}^j(x_j)$ by

$$\begin{aligned}\hat{\mathbf{M}}_j(x_j)\begin{pmatrix}\hat{m}_j(x_j)\\ \hat{m}^j(x_j)\end{pmatrix} &= \frac{1}{N}\sum_{i=1}^{n} \mathbf{1}\left(X^i \in [0,1]^d\right)\\ &\quad \times K_h\left(x_j, X_j^i\right)\begin{pmatrix}1\\ h^{-1}\left[X_j^i - x_j\right]\end{pmatrix} Y^i,\end{aligned} \tag{62}$$

where $K_h(u,v)$ is defined as in (58) and where

$$\hat{\mathbf{M}}_j(x_j) = \begin{pmatrix} \hat{V}^j_{0,0}(x_j) & \hat{V}^j_{j,0}(x_j) \\ \hat{V}^j_{j,0}(x_j) & \hat{V}^j_{j,j}(x_j) \end{pmatrix}$$

$$= \frac{1}{N}\sum_{i=1}^{n} \mathbf{1}\big(X^i \in [0,1]^d\big) \times K_h\big(x_j, X^i_j\big) \begin{pmatrix} 1 & h^{-1}\big[X^i_j - x_j\big] \\ h^{-1}\big[X^i_j - x_j\big] & h^{-2}\big[X^i_j - x_j\big]^2 \end{pmatrix}. \tag{63}$$

Furthermore we put

$$\hat{\mathbf{S}}_{l,j}(x_l, x_j) = \begin{pmatrix} \hat{V}^{l,j}_{0,0}(x_l, x_j) & \hat{V}^{l,j}_{l,0}(x_l, x_j) \\ \hat{V}^{l,j}_{j,0}(x_l, x_j) & \hat{V}^{l,j}_{l,j}(x_l, x_j) \end{pmatrix}$$

$$= \frac{1}{N}\sum_{i=1}^{n} \mathbf{1}\big(X^i \in [0,1]^d\big) K_h\big(x_j, X^i_j\big) K_h\big(x_l, X^i_l\big) \times \begin{pmatrix} 1 & h^{-1}\big[X^i_l - x_l\big] \\ h^{-1}\big[X^i_j - x_j\big] & h^{-2}\big[X^i_j - x_j\big]\big[X^i_l - x_l\big] \end{pmatrix}. \tag{64}$$

We get now our result for this version of the backfitting local linear estimate. Now, the asymptotic bias is explicitly given and its formula does not require a projection step.

THEOREM 4′. *Suppose that the model* (54) *applies and that conditions* (B1)–(B7) *hold, or* (B1), (B2′), (B3′) *and* (B4′) *hold in the i.i.d. case, and that local linear backfitting smoothing is used, that is,* $\hat{m}_j(x_j)$, $\hat{m}^j(x_j)$, $\hat{\mathbf{M}}_j(x_j)$ *and* $\hat{\mathbf{S}}_{l,j}$ *are defined according to* (62)–(64) *and* $\tilde{m}_{0,j}$, $\tilde{m}_j$ *and* $\tilde{m}^j$ *are defined by* (44), (45). *Suppose additionally that* $n_0^{1/5}h \to c_h$ *for a constant* c_h *with* $n_0 = EN = nP(X \in [0,1]^d)$. *Then, for closed subsets* $S_1, \ldots, S_d$ *of* $(0,1)$, *conditions* (A1′)–(A6′), (A8′), (A9′) *are satisfied with*

$$\Delta_n = h^2,$$

$$\mathbf{W} = \begin{pmatrix} 1 & 0 \\ 0 & \int u^2 K(u)\,du \end{pmatrix},$$

$$\begin{pmatrix} \hat{m}^A_j(x_j) \\ \hat{m}^{j,A}(x_j) \end{pmatrix} = \hat{\mathbf{M}}_j(x_j)^{-1} \frac{1}{N}\sum_{i=1}^{n} \mathbf{1}\big(X^i \in [0,1]^d\big) \times K_h\big(x_j, X^i_j\big) \begin{pmatrix} 1 \\ h^{-1}\big[X^i_j - x_j\big] \end{pmatrix} \big(Y^i - E\big[Y^i | X^i\big]\big),$$

$$\alpha_{n,j}(x_j) = m_j(x_j) + \frac{h^2}{2} m_j''(x_j) \int u^2 K(u)\, du,$$

$$\alpha_n^j(x_j) = h m_j'(x_j),$$

$$\gamma_{n,j} = \nu_{n,j} + \frac{h^2 \int u^2 K(u)\, du}{2} \int m_j''(x_j)\, p_j(x_j)\, dx_j,$$

$$\nu_{n,j} = \int m_j(x_j) K_h(x_j, u)\, p_j(u)\, du\, dx_j.$$

In particular, the uniform expansion (53) *holds and the following convergence holds in distribution for any* $x_1, \ldots, x_d \in (0,1)$,

$$n_0^{2/5} \begin{bmatrix} \tilde{m}_1(x_1) - m_1(x_1) + \nu_{n,1} \\ \tilde{m}_d(x_d) - m_d(x_d) + \nu_{n,d} \end{bmatrix} \times N\left(\begin{bmatrix} c_h^2 \delta_1(x_1) \\ \vdots \\ c_H^2 \delta_d(x_d) \end{bmatrix}, \begin{bmatrix} v_1(x_1) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & v_d(x_d) \end{bmatrix} \right),$$

where

$$\delta_j(x_j) = \frac{\int u^2 K(u)\, du}{2} \left\{ m_j''(x_j) - \int m_j''(x_j)\, p_j(x_j)\, dx_j \right\}$$

and where $v_j(x_j) = c_h^{-1} c_K \sigma_j^2(x_j)/p_j(x_j)$, $j = 1, \ldots, d$ *with* $c_K = \int K(u)^2\, du$. *Furthermore*,

$$n_0^{2/5}[\tilde{m}(x) - m(x)] \times N\left(c_h^2 \sum_{j=1}^d \delta_j(x_j), \sum_{j=1}^d v_j(x_j) \right).$$

In this case, the bias functions coincide with the biases $\ddot{b}_j^c(x_j)$ of the centered oracle estimate $\ddot{m}_j^c(x_j)$ for $j = 1, \ldots, d$. So, in this case, the asymptotic bias and the asymptotic variance are identical to the bias and variance of the centered oracle estimator (based also on local linear estimation). That means our estimate achieves the same first-order asymptotics as if the other components were known. In particular, our estimate is design adaptive. This is in contrast to Opsomer and Ruppert (1997) who propose a backfitting estimate, based on the local linear smoother, that has design dependent bias.

Finally, the variance $\sigma_j^2(x_j)$ can be consistently estimated from the residuals $\tilde{\varepsilon}_i = Y^i - \tilde{m}(X^i)$, $i = 1, \ldots, n$, which, along with the usual estimates of $p_j(x_j)$, enables consistent estimation of $v_j(x_j)$ and $\sum_{j=1}^d v_j(x_j)$.

APPENDIX

The proofs will make use of Lemmas 1–4 which we give below. Before we come to this, let us collect some facts about iterative projections. Define the

following spaces of additive functions:

$$\mathscr{H} = \{m \in \mathbf{L}_2(p)\colon m(x) = m_1(x_1) + \cdots + m_d(x_d)\ (p \text{ a.s.}) \text{ for some functions } m_1 \in \mathbf{L}_2(p_1), \ldots, m_d \in \mathbf{L}_2(p_d)\},$$

$$\mathscr{H}^0 = \left\{m \in \mathscr{H}\colon m(x) = m_1(x_1) + \cdots + m_d(x_d)\ (p \text{ a.s.}), \int m(x)p(x)\,dx = 0\right\},$$

$$\mathscr{H}^{0,n} = \left\{m \in \mathscr{H}\colon m(x) = m_1(x_1) + \cdots + m_d(x_d)\ (p \text{ a.s.}), \int m_j(x_j)\hat{p}_j(x_j)\,dx_j = 0 \text{ for } j = 1, \ldots, d\right\},$$

$$\mathscr{H}_j = \{m \in \mathscr{H}^0\colon m(x) = m_j(x_j)\ (p \text{ a.s.}) \text{ for a function } m_j \in \mathbf{L}_2(p_j)\},$$

$$\mathscr{H}_j^n = \{m \in \mathscr{H}^{0,n}\colon m(x) = m_j(x_j)\ (p \text{ a.s.}) \text{ for a function } m_j \in \mathbf{L}_2(p_j)\}.$$

The norm in the space $\mathscr{H}$ is denoted by $\|m\|_2^2 = \int m^2(x)p(x)\,dx$ for $m \in \mathscr{H}$. For $m \in \mathscr{H}_j$ we get with $m_j(x_j) = m(x)$ (p a.s.) that $\|m\|_2^2 = \int m^2(x)p(x)\,dx = \int m_j^2(x_j)p_j(x_j)\,dx_j$. Here and in the following for simplicity of notation we identify functions $m_j \in \mathscr{H}_j$ (or in $\mathscr{H}_j^n$) that map $\mathbb{R}^d$ into $\mathbb{R}$ with functions m_j: $\mathbb{R} \to \mathbb{R}$ by putting $m_j(x_j) = m_j(x)$.

The projection of an element of $\mathscr{H}$ onto $\mathscr{H}_j$ is denoted by $\prod_j$, that is, $\prod_j m(x) = E[m(X)|X_j = x_j] - E[m(X)]$. The operator $\Psi_j = I - \prod_j$ gives the projection onto the linear space

$$\mathscr{H}_j^\perp = \left\{m \in \mathscr{H}\colon \int m(x)\phi(x_j)p(x)\,dx = 0 \text{ for all } \phi \in \mathscr{H}_j\right\} = \left\{m \in \mathscr{H}\colon \int m(x)p(x)\,dx_{-j} = \int m(x)p(x)\,dx\ (p_j \text{ a.s.})\right\}.$$

For $m(x) = m_1(x_1) + \cdots + m_d(x_d) \in \mathscr{H}$ we get

$$\begin{aligned}\Psi_j m(x) &= m(x) - E\big[m(X)|X_j = x_j\big] + E\big[m(X)\big] \\ &= m_1(x_1) + \cdots + m_{j-1}(x_{j-1}) + m_j^*(x_j) + m_{j+1}(x_{j+1}) \\ &\quad + \cdots + m_d(x_d),\end{aligned} \tag{65}$$

where

$$m_j^*(x_j) = -\sum_{k \neq j} \int m_k(x_k)\frac{p_{jk}(x_j, x_k)}{p_j(x_j)}\,dx_k + \sum_k \int m_k(u_k)p_k(u_k)\,du_k. \tag{66}$$

For $m \in \mathscr{H}$ the additive components $m_1, \ldots, m_d$ are only unique up to an additive constant. Note, however, that the value of $\Psi_j m$ does not depend on the special choice of $m_1, \ldots, m_d$.

For functions $m \in \mathscr{H}^{0,n}$ with $m(x) = m_1(x_1 + \cdots + m_d(x_d)$, $m_j \in \mathscr{H}_j^n$ we define the operator $\hat{\Psi}_j$ as Ψ_j but with $m_j^*(x_j)$ on the right-hand side of (65) replaced by

$$\hat{m}_j^*(x_j) = -\sum_{k \neq j} \int m_k(x_k) \left[\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} - \hat{p}_{k,[j+]}(x_k) \right] dx_k, \tag{67}$$

where the function $\hat{p}_{k,[j+]}$ has been defined in (16). Note that for functions $m_j \in \mathscr{H}_j$ we get $\Psi_j m_j(x) = 0$, while

$$\Psi_k m_j(x) = m_j(x_j) - \int m_j(u) \frac{p_{kj}(x_k, u)}{p_k(x_k)} du. \tag{68}$$

Put $T = \Psi_d \cdots \Psi_1$ and $\hat{T} = \hat{\Psi}_d \cdots \hat{\Psi}_1$. We will see below that in our setup the backfitting algorithm is based on iterative applications of $\hat{T}$. A central tool for understanding backfitting will be given by the next lemma, which describes iterative applications of T. For linear operators $S: \mathscr{H} \mapsto \mathscr{H}$ we define

$$\|S\| = \sup\{\|Sf\|_2 : f \in \mathscr{H}, \|f\|_2 \leq 1\},$$
$$\|S\|_0 = \sup\{\|Sf\|_2 : f \in \mathscr{H}^0, \|f\|_2 \leq 1\},$$
$$\|S\|_{0,n} = \sup\{\|Sf\|_2 : f \in \mathscr{H}^{0,n}, \|f\|_2 \leq 1\}.$$

LEMMA 1 [Norm of the operator T]. *Suppose that condition* (A1) *holds. Then* $T: \mathscr{H} \mapsto \mathbf{L}_2(p)$ *is a positive self-adjoint operator with operator norm* $\|T\|_0 < 1$. *Hence, for every* $m \in \mathscr{H}^0$ *we get*

$$\|T^r m\|_2 \leq \|T\|_0^r \|m\|_2. \tag{69}$$

Furthermore, for every $m \in \mathscr{H}^0$ *there exist* $m_j \in \mathscr{H}_j$ $(1 \leq j \leq d)$ *such that* $m(u) = m_1(u_1) + \cdots + m_d(u_d)$ $(p.\ a.s.)$ *and for some constant* $c > 0$,

$$\|m\|_2 \geq c \max\{\|m_1\|_2, \ldots, \|m_d\|_2\}. \tag{70}$$

PROOF. We start by proving (69). It is known that (69) holds with $\|T\|_0^2 \leq 1 - \prod_{j=1}^d \sin^2(\tau_j)$ where $\cos \tau_j = \rho(\mathscr{H}_j, \mathscr{H}_{j+1} + \cdots + \mathscr{H}_d)$ and where for two subspaces L_1 and L_2, the quantity $\rho(L_1, L_2)$ is the cosine of the minimal angle between L_1 and L_2; that is, $\rho(L_1, L_2) = \sup\{\int h_1(x) h_2(x) p(x)\, dx : h_j \in L_j \cap (L_1 \cap L_2)^{\perp}, \|h_j\|_2 \leq 1 (j = 1, 2)\}$. This result was shown in Smith, Solomon and Wagner (1977). For a discussion, see Deutsch (1985) and Bickel, Klaassen, Ritov and Wellner [(1993), Appendix A.4]. We will show now that for $1 \leq j \leq d$ the subspaces $\mathscr{M}_j = \mathscr{H}_1 + \cdots + \mathscr{H}_j$ are closed subsets of $L_2(p)$. This implies that $\rho(\mathscr{H}_{j+1}, \mathscr{M}_j) < 1$ for $j = 1, \ldots, d-1$; see again Deutsch [(1985), Lemma 2.5] and Bickel, Klaassen, Ritov and Wellner [(1993), Appendix A.4, Proposition 2]. To prove that $\mathscr{M}_j$ is closed we will use the following two facts. For two closed subspaces L_1 and L_2 of $L_2(p)$ it holds that $L_1 + L_2$ is closed if and only if there exists a constant $c > 0$ such that for all $m \in L_1 + L_2$ there exist $m_1 \in L_1$ and $m_2 \in L_2$ with $m(u) = m_1(u_1) + m_2(u_2)$ $(p$ a.s.) and

$$\|m\|_2 \geq c \max\big[\|m_1\|_2, \|m_2\|_2\big]. \tag{71}$$

Furthermore, $L_1 + L_2$ is closed if the projection of L_2 onto L_1 is compact. For the proof of these two statements, see Bickel, Klaassen, Ritov and Wellner [(1993), Appendix A.4, Proposition 2]. Suppose now that it has already been proved for $j \le j_o - 1$ that $\mathscr{M}_j$ is closed and that we want to show that $\mathscr{M}_{j_o}$ is closed. As mentioned above, for this claim it suffices to show that $\prod_{j_o} | \mathscr{M}_{j_o - 1}$ is compact. We remark first that (71) implies that for every $m \in \mathscr{M}_{j_o - 1}$ there exist $m_j \in \mathscr{H}_j$ $(j \le j_o - 1)$ such that $m(u) = m_1(u_1) + \cdots + m_{j_o - 1}(u_{j_o - 1})$ (p a.s.) and with a constant $c > 0$,

$$\|m\|_2 \ge c \max\left[\|m_1\|_2, \ldots, \|m_{j_o - 1}\|_2\right]. \tag{72}$$

We will prove that

$$\left\| \prod_{j_o} m \right\|_2^2 \le \text{const.} \left[\sum_{j=1}^{j_o - 1} \int R_{j, j_o}^2(x_j, x_{j_o}) p_j(x_j) p_{j_0}(x_{j_0})\, dx_j\, dx_{j_0} \right] \|m\|_2^2 \tag{73}$$

with

$$R_{j, j_o}(x_j, x_{j_o}) = \frac{p_{j, j_o}(x_j, x_{j_o})}{p_{j_o}(x_{j_o}) p_j(x_j)}.$$

Inequality (73) implies compactness of $\prod_{j_o} | \mathscr{M}_{j_o - 1}$. To see this one uses (A1) and argues as in the standard proofs for compactness of Hilbert–Schmidt operators; see, for example, Example 3.2.4 in Balakrishnan (1981).

It remains to show (73). This follows from (72) with applications of the Cauchy–Schwarz inequality.

Equation (70) follows as (72). □

The next lemma extends this result to the stochastic operator $\hat{T}$.

LEMMA 2 (Norm of the operator $\hat{T}$). *Suppose that conditions* (A1)–(A2) *hold. Then*

$$\|\hat{\Psi}_j - \Psi_j\|_{0, n} = o_P(1), \tag{74}$$

$$\|\hat{T} - T\|_{0, n} = o_P(1). \tag{75}$$

Choose γ *with* $\|T\|_0 < \gamma < 1$. *Then, with probability tending to* 1,

$$\|\hat{T}\|_{0, n} < \gamma. \tag{76}$$

Furthermore, for some constant $c > 0$ *with probability tending to* 1 *it holds that for every* $m \in \mathscr{H}^{0, n}$

$$\|m\|_2 \ge c \max\{\|m_1\|_2, \ldots, \|m_d\|_2\}, \tag{77}$$

where $m_j \in \mathscr{H}_j^n$ $(1 \le j \le d)$ *with* $m(u) = m_1(u_1) + \cdots + m_d(u_d)$ (p *a.s.*).

PROOF. For a function $m \in \mathscr{H}^{0, n}$ we get $m(x) = m_1(x_1) + \cdots + m_d(x_d)$ with functions $m_j \in \mathscr{H}_j^n$. We remark first that the distance between m_j^* and

$\hat{m}_j^*$ [see (66) and (67)] can be bounded with $\int m_k(x_k)\hat{p}_k(x_k)\,dx_k = 0$ and with the help of the Cauchy–Schwarz inequality as follows:

$$\begin{aligned}
\|\hat{m}_j^* - m_j^*\|_2 \le{}& \sum_{k\neq j}\left\|\int m_k(x_k)\left[\frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)} - \frac{p_{jk}(x_j,x_k)}{p_j(x_j)}\right]dx_k\right\|_2 \\
&+ \sum_{k\neq j}\left|\int m_k(x_k)\hat{p}_{k,[j+]}(x_k)\,dx_k\right| \\
&+ \sum_{k}\left|\int m_k(x_k)p_k(x_k)\,dx_k\right| \\
={}& \sum_{k\neq j}\left\|\int m_k(x_k)\left[\frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)p_k(x_k)} - \frac{p_{jk}(x_j,x_k)}{p_j(x_j)p_k(x_k)}\right]p_k(x_k)\,dx_k\right\|_2 \\
&+ \sum_{k\neq j}\left|\int m_k(x_k)\left[\frac{\hat{p}_k(x_k) - \hat{p}_{k,[j+]}(x_k)}{p_k(x_k)}\right]p_k(x_k)\,dx_k\right| \\
&+ \sum_{k}\left|\int m_k(x_k)\left[\frac{\hat{p}_k(x_k) - p_k(x_k)}{p_k(x_k)}\right]p_k(x_k)\,dx_k\right| \\
\le{}& \sum_{k\neq j}\|m_k\|_2(U_{jk} + R_{jk}) + \sum_k \|m_k\|_2 Q_k,
\end{aligned}$$

with

$$\begin{aligned}
U_{jk}^2 &= \int\left[\frac{p_{j,k}(x_j,x_k)}{p_k(x_k)p_j(x_j)} - \frac{\hat{p}_{j,k}(x_j,x_k)}{p_k(x_k)\hat{p}_j(x_j)}\right]^2 p_k(x_k)p_j(x_j)\,dx_j\,dx_k, \\
R_{jk}^2 &= \int\left[\frac{\hat{p}_k(x_k) - \hat{p}_{k,[j+]}(x_k)}{p_k(x_k)}\right]^2 p_k(x_k)\,dx_k, \\
Q_k^2 &= \int\left[\frac{\hat{p}_k(x_k) - p_k(x_k)}{p_k(x_k)}\right]^2 p_k(x_k)\,dx_k.
\end{aligned}$$

With $T_j = \max_{k\neq j}|U_{jk} + R_{jk}| + \max_k|S_k|$, this and (70) imply with a constant C (not depending on m),

$$\|\hat{m}_j^* - m_j^*\|_2 \le C\|m\|_2 T_j.$$

Now because of (A2), $U_{jk} = o_P(1)$ and $Q_k = o_P(1)$. Furthermore,

$$\begin{aligned}
&\int\left[\frac{p_k(x_k) - \hat{p}_{k,[j+]}(x_k)}{p_k(x_k)}\right]^2 p_k(x_k)\,dx_k \\
&\quad= \int\left[\int\left\{\frac{p_{jk}(x_j,x_k)}{p_j(x_j)p_k(x_k)} - \frac{\hat{p}_{jk}(x_j,x_k)}{p_k(x_k)p_j(x_j)}\right\}p_j(x_j)\,dx_j\right]^2 p_k(x_k)\,dx_k
\end{aligned}$$

$$\leq \int \left[\frac{p_{jk}(x_j, x_k)}{p_j(x_j)\, p_k(x_k)} - \frac{\hat{p}_{jk}(x_j, x_k)}{p_k(x_k)\, p_j(x_j)} \right]^2 p_k(x_k)\, p_j(x_j)\, dx_j\, dx_k$$
$$= o_P(1);$$

therefore $R_{jk} = o_P(1)$ and $T_j = o_P(1)$. This shows (74) and (75). Claim (76) follows from (75) and

$$\|T\|_{0,n} = \|T\|_0 + o_P(1). \tag{78}$$

It remains to show (78). This follows immediately from

$$\inf_{f \in \mathscr{H}^{0,n}} \sup_{g \in \mathscr{H}^0, \|g\|_2 = 1} \|f - g\|_2 = o_P(1), \tag{79}$$

$$\inf_{f \in \mathscr{H}^{0}} \sup_{g \in \mathscr{H}^{0,n}, \|g\|_2 = 1} \|f - g\|_2 = o_P(1). \tag{80}$$

For the proof of (79) and (80) note, for example, that for $m_j \in \mathscr{H}_j^n$ one has

$$\left| \int m_j(x_j)\, p(x_j)\, dx_j \right|^2 = \left| \int m_j(x_j) \left[p_j(x_j) - \hat{p}_j(x_j) \right] dx_j \right|^2$$
$$\leq \|m_j\|_2^2 \left\| \left[p_j - \hat{p}_j \right] / p_j \right\|_2^2$$
$$= \|m_j\|_2^2 o_P(1)$$

because of (A2). Similarly, one shows (77); see also (70). □

Our next lemma builds on Lemma 2 to establish a stochastic expansion for $\tilde{m}(x) = \tilde{m}_1(x_1) + \cdots + \tilde{m}_d(x_d)$ in terms of $\hat{m}_j$ $[1 \leq j \leq d]$.

LEMMA 3 [Stochastic expansion of $\tilde{m}$]. *Suppose that conditions* (A1)–(A3) *hold. Then there exist constants* $0 < \gamma < 1$ *and* $0 < C, C' < \infty$ *such that with probability tending to* 1, *the following stochastic expansion holds for all* $s \geq 1$:

$$\tilde{m}(x) = \sum_{r=0}^{s} \hat{T}^r \hat{\tau}(x) + R^{[s]}(x),$$

where

$$\hat{\tau}(x) = \hat{\Psi}_d \cdots \hat{\Psi}_2 \left[\hat{m}_1(x) - \tilde{m}_{0,1} \right] + \cdots + \hat{\Psi}_d \left[\hat{m}_{d-1}(x) - \tilde{m}_{0,d-1} \right] + \hat{m}_d(x_d) - \tilde{m}_{0,d}$$

and where $R^{[s]}(x) = R_1^{[s]}(x_1) + \cdots + R_d^{[s]}(x_d)$ *is a function in* $\mathscr{H}^{0,n}$ *with*

$$\|R_j^{[s]}\|_2 \leq C\gamma^s. \tag{81}$$

Under the additional assumption of (A4) *it holds that*

$$\sup_{x_j \in S_j} \left| R_j^{[s]}(x_j) \right| \leq C'\gamma^s. \tag{82}$$

PROOF. We remark first that (15) can be rewritten as

$$\tilde{m}(x) = \hat{\Psi}_j \tilde{m}(x) - \hat{m}_j(x_j) - \tilde{m}_{0,j}. \tag{83}$$

Iterative applications of this equation for $j = 1, \ldots, d$ gives

$$\tilde{m}(x) = \hat{T}\tilde{m}(x) + \hat{\tau}(x). \tag{84}$$

Iterative applications of (84) gives

$$\tilde{m}(x) = \sum_{r=0}^{\infty} \hat{T}^r \hat{\tau}(x).$$

The operator norm $\|\hat{T}\|_{0,n}$ is smaller than γ, with probability tending to 1, for $\gamma < 1$ large enough. This was shown in the last lemma and it shows that the infinite series expansion in the last equation is well defined. Furthermore, this can be used to prove that for $C_1 > 0$ large enough, with probability tending to 1, $\|R^{[s]}\|_2 \le C_1 \gamma^s$. This implies claim (81) because of (77).

Assume now (A4). For the proof of (82) note that for $C_2 > 0$ large enough with probability tending to 1 for all functions g in $\mathscr{H}_j$ with $\|g\|_2 \le 1$, it holds for $k \neq j$ that

$$\sup_{x_k \in S_k} \left| \int \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_k(x_k)} g(x_j)\, dx_j \right| \le C_2, \tag{85}$$

$$\left\| \int \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_k(x_k)} g(x_j)\, dx_j \right\|_2 \le C_2. \tag{86}$$

Inequality (85) follows from assumption (A4) by application of the Cauchy–Schwarz inequality:

$$\begin{aligned} &\sup_{x_k \in S_k} \left| \int \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_k(x_k)} g(x_j)\, dx_j \right| \\ &\quad = \sup_{x_k \in S_k} \left| \int \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_k(x_k) p_j(x_j)} p_j(x_j) g(x_j)\, dx_j \right| \\ &\quad \le \left| \sup_{x_k \in S_k} \int \frac{\hat{p}_{jk}^2(x_j, x_k)}{\hat{p}_k^2(x_k) p_j(x_j)}\, dx_j \int g^2(x_j) p_j(x_j)\, dx_j \right|^{1/2}. \end{aligned}$$

For the proof of (86) one applies again the Cauchy–Schwarz inequality and

$$\int \frac{\hat{p}_{jk}^2(x_j, x_k)}{\hat{p}_k^2(x_k) p_j^2(x_j)} p_k(x_k) p_j(x_j)\, dx_j\, dx_k \le C_3 \tag{87}$$

for a constant C_3 (with probability tending to 1). Claim (87) follows from assumptions (A1) and (A2).

Equations (85) and (86) imply that for $C_4 > 0$ large enough with probability tending to 1 for all functions h in $\mathscr{H}$ with $\|h\| \le 1$ it holds for $1 \le j \le d$

that

$$\sup_{x\in S}\left|\hat{T}h(x)\right| \le C_4, \tag{88}$$

where $S = \{x\colon x_j \in S_j\}$. Now, because of

$$R^{[s]}(x) = \sum_{r=s+1}^{\infty} \hat{T}^r\hat{\tau}(x) = \hat{T}R^{[s-1]}(x),$$

claim (82) now follows from

$$\begin{aligned}\sup_{x\in S}\left|R^{[s]}(x)\right| &\le C_4\|R^{[s-1]}\| \\ &\le C_4C_1\gamma^{s-1}.\end{aligned}$$

□

LEMMA 4 (Behavior of the stochastic component of $\tilde{m}$). *Suppose that* (A1)–(A6) *hold. Then we have that*

$$\sup_{x_j\in S_j}\left|\tilde{m}_j^A(x_j) - \hat{m}_j^A(x_j) + \tilde{m}_{0,j}^A\right| = o_P(\Delta_n). \tag{89}$$

PROOF. We will show Lemma 4 for $j = 1$. Proceeding as in the last lemma we get that, with probability tending to 1,

$$\tilde{m}^A(x) = \sum_{r=0}^{\infty} \hat{T}^r\hat{\tau}^A(x),$$

where

$$\begin{aligned}\hat{\tau}^A(x) &= \hat{\Psi}_d\cdots\hat{\Psi}_2\left[\hat{m}_1^A - \tilde{m}_{0,1}^A\right](x) + \cdots + \hat{\Psi}_d\left[\hat{m}_{d-1}^A - \tilde{m}_{0,d-1}^A\right](x) \\ &\quad + \hat{m}_d^A(x_d) - \tilde{m}_{0,d}^A, \\ \tilde{m}^A(x) &= \tilde{m}_1^A(x_1) + \cdots + \tilde{m}_d^A(x_d).\end{aligned}$$

We argue now that the statement of the lemma follows from

$$\sup_{x\in S}\left|\sum_{r=1}^{\infty}\hat{T}^r\hat{\tau}^A(x)\right| = o_P(\Delta_n), \tag{90}$$

where as above $S = \{x\colon x_j \in S_j\}$. For seeing this, note that (90) implies that

$$\sup_{x\in S}\left|\tilde{m}^A(x) - \hat{\tau}^A(x)\right| = o_P(\Delta_n). \tag{91}$$

Only the first summand of $\hat{\tau}^A(x)$, that is, $\hat{\Psi}_d\cdots\hat{\Psi}_2\hat{m}_1^A(x)$ depends on x_1. Furthermore, the operators $\hat{\Psi}_2,\ldots,\hat{\Psi}_d$ do not change the additive component of a function that depends on x_1. Therefore $\hat{\tau}^A(x)$ is of the form $\hat{\tau}^A(x) = \hat{m}_1^A(x_1) + \hat{\tau}_{-1}^A(x_2,\ldots,x_d)$ where $\hat{\tau}_{-1}^A$ is a function that does not depend on x_1. For this reason the claim of the lemma follows for $j = 1$. [Note also that $\int\hat{p}_1(x_1)[\hat{m}_1^A(x_1) - \tilde{m}_{0,1}^A]\,dx_1 = \int\hat{p}_1(x_1)\tilde{m}_1^A(x_1)\,dx_1 = 0$.]

For the proof of (90) note first that

$$\|\hat{T}\hat{\tau}^A\|_2 = o_P(\Delta_n). \tag{92}$$

This follows from (21), $\|\hat{T}\|_{0,n} \leq 1$ and $\|\hat{\Psi}_j\|_{0,n} \leq 1$ (with probability tending to 1); see Lemma 2. Because of $\|\hat{T}\|_{0,n} \leq \gamma$ (with probability tending to 1 for a $\gamma < 1$) (92) shows that

$$\left\| \sum_{r=1}^{\infty} \hat{T}^r \hat{\tau}^A \right\|_2 = o_P(\Delta_n). \tag{93}$$

With (88) this shows

$$\sup_{x \in S} \left| \sum_{r=2}^{\infty} \hat{T}^r \hat{\tau}^A(x) \right| = o_P(\Delta_n).$$

So for claim (90) it remains to show

$$\sup_{x \in S} \left| \hat{T}\hat{\tau}^A(x) \right| = o_P(\Delta_n).$$

This can be done using (20), (21), $\|\hat{\Psi}_j\|_{0,n} \leq 1$ (with probability tending to 1), and (88). □

PROOF OF THEOREM 1. For the proof, note first that by definition of our backfitting algorithm [see (18)],

$$\tilde{m}^{[r]}(x) = \hat{T}\tilde{m}^{[r-1]}(x) + \hat{\tau}(x).$$

Iterative application of this equation gives

$$\tilde{m}^{[t]}(x) = \sum_{s=0}^{r-1} \hat{T}^s \hat{\tau}(x) + \hat{T}^r \tilde{m}^{[0]}(x).$$

Because of Lemma 3 this shows

$$\tilde{m}^{[r]}(x) - \tilde{m}(x) = -\sum_{s=r}^{\infty} \hat{T}^s \hat{\tau}(x) + \hat{T}^r \tilde{m}^{[0]}(x).$$

Because of (A3) and $\|\hat{\Psi}_j\| = \|\Psi_j\| + o_P(1) = 1 + o_P(1)$, we have for a constant C' that $\|\hat{\tau}\|_2 \leq C'$ with probability tending to 1. So with Lemma 2 we get that

$$\|\tilde{m}^{[r]} - \tilde{m}\|_2 \leq \left[\frac{C'}{1-\gamma} + \|\tilde{m}^{[0]}\|_2 \right] \gamma^r$$

with probability tending to 1. Claim (22) follows now by application of (70). For the proof of existence and uniqueness of $\tilde{m}_j^A$ and $\tilde{m}_j^B$, one proceeds similarly. □

Theorem 2 follows from Lemma 4.

PROOF OF THEOREM 3. We put for $1 \le j \le d$,

$$\hat{m}_j^{B,1}(x_j) = \overline{\alpha}_{n,j}(x_j) + \sum_{k \neq j} \int \overline{\alpha}_{n,k}(x_k) \left[\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} - \hat{p}_{k,[j+]}(x_k) \right] dx_k,$$

$$\hat{m}_j^{B,2}(x_j) = \Delta_n \int \beta(x) \frac{p(x)}{p_j(x_j)}\, dx_{-j},$$

$$\hat{m}_j^{B,3}(x_j) = \hat{m}_j^{B}(x_j) - \frac{\int \hat{m}_j^{B}(u)\, \hat{p}_j(u)\, du}{\int \hat{p}_j(u)\, du} - \hat{m}_j^{B,1}(x_j) - \hat{m}_j^{B,2}(x_j),$$

where

$$\overline{\alpha}_{n,j}(x_j) = \alpha_{n,j}(x_j) - \int \alpha_{n,j}(u)\, \hat{p}_j(u)\, du.$$

For $r = 1, \ldots, 3$; $j = 1, \ldots, d$ we define now $\tilde{m}_j^{B,r}$ by

$$\tilde{m}_{0,j}^{B,r} = \frac{\int \hat{m}_j^{B,r}(x_j)\, \hat{p}_j(x_j)\, dx_j}{\int \hat{p}_j(x_j)\, dx_j},$$

$$\begin{aligned} \tilde{m}_j^{B,r}(x_j) = {} & \hat{m}_j^{B,r}(x_j) \\ & - \sum_{k \neq j} \int \tilde{m}_k^{B,r}(x_k) \left[\frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} - \hat{p}_{k,[j+]}(x_k) \right] dx_k \\ & - \tilde{m}_{0,j}^{B,r}. \end{aligned} \tag{94}$$

By these equations the quantities $\tilde{m}_j^{B,r}$ are uniquely defined. This has been shown in Theorem 1.

Note that $\tilde{m}_j^{B}(x_j) = \tilde{m}_j^{B,1}(x_j) + \tilde{m}_j^{B,2}(x_j) + \tilde{m}_j^{B,3}(x_j)$. We will show

$$\tilde{m}_j^{B,1}(x_j) = \overline{\alpha}_{n,j}(x_j), \tag{95}$$

$$\sup_{x_j \in S_j} \left| \tilde{m}_j^{B,2}(x_j) - \Delta_n \beta_j(x_j) \right| = o_P(\Delta_n), \tag{96}$$

$$\sup_{x_j \in S_j} \left| \tilde{m}_j^{B,3}(x_j) \right| = o_P(\Delta_n). \tag{97}$$

These claims imply the statement of the theorem. For the proof of (95) note that $\tilde{m}_{0,j}^{B,1} = 0$ and that $\tilde{m}_j^{B,1}(x_j) = \overline{\alpha}_{n,j}(x_j)$ solves (94). This shows (95).

For $r = 2, 3$ we get for $\tilde{m}^{B,r}(x) = \tilde{m}_1^{B,r}(x_1) + \cdots + \tilde{m}_d^{B,r}(x_d)$,

$$\tilde{m}^{B,r}(x) = \sum_{k=0}^{\infty} \hat{T}^k \hat{\tau}^{B,r}(x),$$

where

$$\hat{\tau}^{B,r}(x) = \hat{\Psi}_d \cdots \hat{\Psi}_2\big[\hat{m}_1^{B,r} - \tilde{m}_{0,1}^{B,r}\big](x) + \cdots + \hat{\Psi}_d\big[\hat{m}_{d-1}^{B,r} - \tilde{m}_{0,d-1}^{B,r}\big](x) + \hat{m}_d^{B,r}(x_d) - \tilde{m}_{0,d}^{B,r}.$$

For the proof of (96) we will show that

$$\sup_{x\in S}\left|\tilde{m}^{B,2}(x) - \sum_{k=0}^{\infty} T^k \tau^{B,2}(x)\right| = o_P(\Delta_n), \tag{98}$$

where

$$\tau^{B,2}(x) = \Psi_d \cdots \Psi_2\big[\hat{m}_1^{B,2} - \overline{m}_0^{B,2}\big](x) + \cdots + \Psi_d\big[\hat{m}_{d-1}^{B,2} - \overline{m}_0^{B,2}\big](x) - \hat{m}_d^{B,2}(x_d) - \overline{m}_0^{B,2},$$

$$\overline{m}_0^{B,2} = \Delta_n \int \beta(x)p(x)\,dx = \tilde{m}_{0,j}^{B,2} + o_P(\Delta_n).$$

By the same arguments as in the beginning of the proof of Lemma 3 (with $\hat{T}$ replaced by T) one can see that

$$\Delta_n\{\beta_1(x_1) + \cdots + \beta_d(x_d)\} = \sum_{k=0}^{\infty} T^k \tau^{B,2}(x).$$

Therefore (98) implies (96). For the proof of (98) we write, with $W = \sum_{k=0}^{\infty} T^k$,

$$\begin{aligned}
&\tilde{m}^{B,2}(x) - \sum_{k=0}^{\infty} T^k \tau^{B,2}(x) \\
&\quad = -\sum_{k=1}^{\infty}\big[T^k - \hat{T}^k\big]\hat{\tau}^{B,2}(x) + W\big[\hat{\tau}^{B,2}(x) - \tau^{B,2}(x)\big] \\
&\quad = -\sum_{k=1}^{\infty}\sum_{l=0}^{k-1} T^l\big[T - \hat{T}\big]\hat{T}^{k-1-l}\hat{\tau}^{B,2}(x) + W\big[\hat{\tau}^{B,2}(x) - \tau^{B,2}(x)\big] \\
&\quad = -T\hat{V}\hat{\tau}^{B,2}(x) + \big[\hat{T} - T\big]\hat{U}\hat{\tau}^{B,2}(x) + W\big[\hat{\tau}^{B,2}(x) - \tau^{B,2}(x)\big],
\end{aligned}$$

where

$$\hat{V} = \sum_{k=1}^{\infty}\sum_{l=1}^{k-1} T^{l-1}\big[T - \hat{T}\big]\hat{T}^{k-1-l},$$

$$\hat{U} = \sum_{k=1}^{\infty} \hat{T}^{k-1}.$$

One applies now that $\|\hat{\tau}^{B,2}\|_2 = O_P(\Delta_n)$ and that

$$\begin{aligned}
&\sup_{x\in S}|Tg(x)| = O_P(1), \\
&\sup_{x\in S}\big|\big[\hat{T} - T\big]g(x)\big| = o_P(1)
\end{aligned} \tag{99}$$

for functions g with $\|g\|_2 = O_P(1)$; see the proof of (88) and apply (A8).

Because of $\|\hat{V}\|_{0,n} = o_P(1)$ and $\|\hat{U}\|_{0,n} = O_P(1)$ this shows $\sup_{x \in S}|T\hat{V}\hat{\tau}^{B,2}(x) + [\hat{T} - G]\hat{U}\hat{\tau}^{B,2}(x)| = o_P(\Delta_n)$. For the proof of (98) it remains to show

$$\sup_{x \in S} \left| W\left[\hat{\tau}^{B,2}(x) - \tau^{B,2}(x)\right]\right| = o_P(\Delta_n). \tag{100}$$

Claim (100) follows from (99) and

$$\sup_{x \in S} \left|\hat{\tau}^{B,2}(x) - \tau^{B,2}(x)\right| = o_P(\Delta_n), \tag{101}$$

$$\|\hat{\tau}^{B,2} - \tau^{B,2}\|_2 = o_P(\Delta_n). \tag{102}$$

For the proof of (101) and (102) one proceeds similarly to the proof of (88). For the statement of the theorem it remains to prove (97). For this claim one shows that

$$\sup_{x \in S} \left|\hat{\tau}^{B,3}(x)\right| = o_P(\Delta_n),$$

$$\|\hat{\tau}^{B,3}\|_2 = o_P(\Delta_n).$$

This can be done by showing for $j = 1, \ldots, d$,

$$\sup_{x_j \in S_j} \left|\hat{m}_j^{B,3}(x_j)\right| = o_P(\Delta_n),$$

$$\|\hat{m}_j^{B,3}\|_2 = o_P(\Delta_n). \qquad \square$$

PROOFS OF THEOREMS 1′ AND 2′. The theorems follow as Theorems 1 and 2 by essentially the same arguments. In particular, instead of $L_2(p)$ we consider now $L_2(Wp) = \{f = (f^0, \ldots, f^d)\colon\ f^j\colon\ \mathbb{R}^d \mapsto \mathbb{R}$ with $\int f^T(x)Wf(x)p(x)\,dx < \infty\}$. Furthermore, now the spaces $\mathscr{H}$, $\mathscr{H}^0$, $\mathscr{H}_j$, $\mathscr{H}^{0,n}$ and $\mathscr{H}_j^n$ are defined as

$$\mathscr{H} = \Big\{m = (m^0, \ldots, m^d) \in L_2(Wp)\colon m^0(x) = m_1(x_1) + \cdots + m_d(x_d)$$
$$(p \text{ a.s.}) \text{ for functions } m_1 \in L_2(p_1), \ldots, m_d \in L_2(p_d), \text{ the}$$
$$\text{functions } m^j \text{ depend only on } x_j \text{ for } j = 1, \ldots, d\Big\},$$

$$\mathscr{H}^0 = \left\{m \in \mathscr{H}\colon \int m^0(x)\,p(x)\,dx = 0\right\},$$

$$\mathscr{H}_j = \Big\{m \in \mathscr{H}\colon m^0(x) \text{ depends only on } x_j\ (p \text{ a.s.}) \text{ and for } l \neq j$$
$$\text{it holds that } m^l(x) \equiv 0\ (p \text{ a.s.})\Big\},$$

$$\mathscr{H}^{0,n} = \Big\{m \in \mathscr{H}\colon m^0(x) = m_1(x_1) + \cdots + m_d(x_d)\ (p \text{ a.s.}) \text{ for functions}$$
$$m_1 \in L_2(p_1), \ldots, m_d \in L_2(p_d) \text{ with } \int m_j(u_j)\hat{V}_{0,0}^j(u_j)\,du_j = 0\Big\},$$

$$\mathscr{H}_j^n = \Big\{m \in \mathscr{H}^{0,n}\colon m(x) \text{ depends only on } x_j\ (p \text{ a.s.})\Big\}.$$

For a function $m \in \mathscr{H}$ with $m^0(x) = m_1(x_1) + \cdots + m_d(x_d)$ for some functions m_j we define now $\Psi_j m$,

$$\left[\Psi_j m\right]^0(x) = f_1(x_1) + \cdots + f_d(x_d),$$
$$\left[\Psi_j m\right]^k(x) = f^k(x_k),$$

where for $k \neq j$,

$$f_k(x_k) = m_k(x_k),$$
$$f^k(x_k) = m^k(x_k),$$

and where

$$\begin{pmatrix} f_j(x_j) \\ f^j(x_j) \end{pmatrix} = -\sum_{k \neq j} \int \mathbf{M}_j^{-1}(x_j)\mathbf{S}_{j,k}(x_j, x_k) \times \begin{pmatrix} m_k(x_k) - \int m_k(u_k)\, p_k(u_k)\, du_k \\ m^k(x_k) \end{pmatrix} dx_k + \begin{pmatrix} \int m_j(u_j)\, p_j(u_j)\, du_j \\ 0 \end{pmatrix}.$$

Furthermore, for a function $m \in \mathscr{H}^{0,n}$ with $m^0(x) = m_1(x_1) + \cdots + m_d(x_d)$ for some functions m_j with $\int m_j(u_j)\hat{V}_{0,0}^j(u_j)\, du_j = 0$ we define now $\hat{\Psi}_j m$:

$$\left[\hat{\Psi}_j m\right]^0(x) = f_1(x_1) + \cdots + f_d(x_d),$$
$$\left[\hat{\Psi}_j m\right]^k(x) = f^k(x_k),$$

where for $k \neq j$,

$$f_k(x_k) = m_k(x_k),$$
$$f^k(x_k) = m^k(x_k),$$

and where

$$f_j(x_j) = g_j(x_j) - \int g_j(u_j)\hat{V}_{0,0}^j(u_j)\, du_j,$$
$$\begin{pmatrix} g_j(x_j) \\ f^j(x_j) \end{pmatrix} = -\sum_{k \neq j} \int \hat{\mathbf{M}}_j^{-1}(x_j)\hat{\mathbf{S}}_{j,k}(x_j, x_k) \begin{pmatrix} m_k(x_k) \\ m^k(x_k) \end{pmatrix} dx_k.$$

Proceeding as above, one can show that the norm of the operators $T = \Psi_d \cdots \Psi_1$ and $\hat{T} = \hat{\Psi}_d \cdots \hat{\Psi}_1$ is smaller than $\gamma < 1$ (with probability tending to 1). Theorems 1′ and 2′ follow by stochastic expansions of $\tilde{\mathbf{m}}$. □

The proof of Theorem 3′ is similar to the proof of Theorem 3 and is omitted.

PROOF OF THEOREM 4. We have to verify conditions (A1)–(A6), (A8), (A9). Continuity of q_0 implies that $\inf_{0 \le x_j \le 1} p_j(x_j) > 0$ for all j and $\sup_{0 \le x_j \le 1,\, 0 \le x_k \le 1} p_{j,k}(x_j, x_k) < \infty$. This shows (A1).

In the proof we will make repeated use of

$$\sup_{x_j \in I_h,\, x_k \in I_h} \left| \hat{p}_{j,k}(x_j, x_k) - p_{j,k}(x_j, x_k) \right| = O_P\left([\log n]^{1/2} n^{-3/10} \right), \tag{103}$$

$$\sup_{x_j \in I_h} \left| \hat{p}_j(x_j) - p_j(x_j) \right| = O_P\left([\log n]^{1/2} n^{-2/5} \right), \tag{104}$$

$$\sup_{0 \le x_j,\, x_k \le 1} \left| \hat{p}_{j,k}(x_j, x_k) - \int_0^1 K_h(x_j, u)\, du \int_0^1 K_h(x_k, v)\, dv\, p_{j,k}(x_j, x_k) \right| = O_P(n^{-1/5}), \tag{105}$$

$$\sup_{0 \le x_j \le 1} \left| \hat{p}_j(x_j) - \int_0^1 K_h(x_j, u)\, du\, p_j(x_j) \right| = O_P(n^{-1/5}), \tag{106}$$

where $I_h = [2C_1 h, 1 - 2C_1 h]$, $I_h^c = [0, 2C_1 h) \cup (1 - 2C_1 h, 1]$ and $I_h^{2,c} = (I_h^c \times [0,1]) \cup ([0,1] \times I_h^c)$.

A proof of (103) and (104) can be found in Masry (1996b). Claims (105) and (106) can be shown by a modification of the arguments in Masry (1996b).

Note that (105) and (106) imply that

$$\sup_{(x_j, x_k) \in I_h^{2,c}} \left| \hat{p}_{j,k}(x_j, x_k) \right| = O_P(1), \tag{107}$$

l02

$$\sup_{x_j \in I_h^c} \left| \hat{p}_j(x_j)^{-1} \right| = O_P(1), \tag{108}$$

$$\sup_{x_j \in I_h^c} \left| \hat{p}_j(x_j) \right| = O_P(1). \tag{109}$$

Assumptions (A2), (A4) and (A8) can be easily proved by application of (103)–(109). Assumptions (A3) and (A5) follow from

$$\sup_{x_j \in [0,1]} \left| \hat{m}_j^A(x_j) \right| = O_P\left\{ \left(\frac{\log n}{nh} \right)^{1/2} \right\}, \tag{110}$$

$$\sup_{x_j \in [0,1]} \left| \hat{m}_j^B(x_j) \right| = O_P(1). \tag{111}$$

For a proof of (110) see again Masry (1996b). Claim (111) follows from

$$\sup_{x_j \in I_h} \left| \hat{m}_j^B(x_j) - \hat{\mu}_{n,j}(x_j) \right| = o_P(h^2), \tag{112}$$

$$\sup_{x_j \in I_h^c} \left| \hat{m}_j^B(x_j) - \hat{\mu}_{n,j}(x_j) \right| = o_P(h). \tag{113}$$

Note that because of (112) and (113), for the proof of (A9) it suffices to check that $\gamma_{n,j}$ can be chosen as $\gamma_{n,j} \equiv 0$. This follows from

$$\int \alpha_{n,j}(x_j)\,\hat{p}_j(x_j)\,dx_j = o_P(\Delta_n). \tag{114}$$

So it remains to establish (A6), (114), (112) and (113).

PROOF OF (114). By definition of $\alpha_{n,j}$ we get

$$\begin{aligned}
&\int \alpha_{n,j}(x_j)\,\hat{p}_j(x_j)\,dx_j \\
&\quad = \int m_j(x_j)\,\hat{p}_j(x_j)\,dx_j \\
&\qquad + \int m_j'(x_j)\,K_h(x_j,u)(u-x_j)\left[\int K_h(x_j,v)\,dv\right]^{-1}\hat{p}_j(x_j)\,dx_j\,du.
\end{aligned}$$

By standard kernel arguments one can show that the right-hand side is equal to

$$\begin{aligned}
&\int m_j(x_j)\,K_h(x_j,u)\,p_j(u)\,du\,dx_j \\
&\qquad + \int m_j'(x_j)\,K_h(x_j,u)(u-x_j)\left[\int K_h(x_j,v)\,dv\right]^{-1} \\
&\qquad\quad \times K_h(x_j,w)\,p_j(w)\,du\,dw\,dx_j + o_P(\Delta_n).
\end{aligned}$$

We argue now that the second term is equivalent to

$$\begin{aligned}
&\int m_j'(x_j)\,K_h(x_j,u)(u-x_j)\left[\int K_h(x_j,v)\,dv\right]^{-1} \\
&\qquad \times K_h(x_j,w)\,p_j(x_j)\,du\,dw\,dx_j + o_P(\Delta_n) \\
&\quad = \int m_j'(x_j)\,K_h(x_j,u)(u-x_j)\,p_j(x_j)\,du\,dx_j + o_P(\Delta_n).
\end{aligned}$$

Putting these expansions together we get that

$$\begin{aligned}
&\int \alpha_{n,j}(x_j)\,\hat{p}_j(x_j)\,dx_j \\
&\quad = \int m_j(x_j)\,K_h(x_j,u)\,p_j(u)\,du\,dx_j \\
&\qquad + \int m_j'(x_j)\,K_h(x_j,u)(u-x_j)\,p_j(x_j)\,du\,dx_j + o_P(\Delta_n) \\
&\qquad + \int m_j(u)\,K_h(x_j,u)\,p_j(u)\,du\,dx_j + o_P(\Delta_n) \\
&\quad = \int m_j(u)\,p_j(u)\,du + o_P(\Delta_n) \\
&\quad = o_P(\Delta_n).
\end{aligned}$$

□

PROOF OF A6. We will give only the proof of (20). Claim (21) follows from (107), (108), (110) and (20). By the triangle inequality,

$$\begin{aligned}
&\sup_{x_k\in I_h}\left|\int_0^1 \frac{\hat p_{j,k}(x_j,x_k)}{\hat p_k(x_k)}\hat m_j^A(x_j)\,dx_j\right| \\
&\quad\le \sup_{x_k\in I_h}\left|\int_0^1 \frac{p_{j,k}(x_j,x_k)}{p_j(x_j)p_k(x_k)}\hat v_j(x_j)\,dx_j\right| \\
&\qquad+\sup_{x_k\in I_h}\left|\int_0^1\left[\frac{\hat p_{j,k}(x_j,x_k)}{\hat p_j(x_j)\hat p_k(x_k)}-\frac{p_{j,k}(x_j,x_k)}{p_j(x_j)p_k(x_k)}\right]\hat v_j(x_j)\,dx_j\right| \\
&\quad\le \sup_{x_k\in I_h}\left|\int_0^1 \frac{p_{j,k}(x_j,x_k)}{p_j(x_j)p_k(x_k)}\hat v_j(x_j)\,dx_j\right| + o_P(h^2),
\end{aligned}$$

because of (103)–(108), (110), where

$$\hat v_j(x_j)=\frac{1}{N}\sum_{i\in J_n}K_h\bigl(x_j-X_j^i\bigr)\varepsilon^i,$$

where

$$J_n=\bigl\{i\colon X^i\in[0,1]^d,\,1\le i\le n\bigr\}.$$

Therefore,

$$\int_0^1 \frac{\hat p_{j,k}(x_j,x_k)}{\hat p_k(x_k)}\hat m_j^A(x_j)\,dx_j=\frac{1}{N}\sum_{i\in J_n}\varepsilon^i\xi_{ni}(x_k)+o_P(h^2)$$

uniformly for $x_k\in I_h$ with

$$\xi_{ni}(x_k)=\int K(u)\frac{p_{j,k}\bigl(X_j^i-uh,x_k\bigr)}{p_j\bigl(X_j^i-uh\bigr)p_k(x_k)}\,du$$

by straightforward change of variables. The argument is now quite similar to that given in Masry (1996b). We drop the k subscript for convenience. The interval $[0,1]$ can be covered by a finite number $c(n)$ of cubes $I_{n,r}$ with centers u_r and with side length $l(n)$. We then have

$$\begin{aligned}
\sup_{u\in I_h}\left|\frac{1}{N}\sum_{i\in J_n}\varepsilon^i\xi_{ni}(u)\right| &= \max_{1\le r\le c(n)}\sup_{u\in I_h\cap I_{n,r}}\left|\frac{1}{N}\sum_{i\in J_n}\varepsilon^i\xi_{ni}(u)\right| \\
&\le \max_{1\le r\le c(n)}\sup_{u\in I_h\cap I_{n,r}}\left|\frac{1}{N}\sum_{i\in J_n}\varepsilon^i\xi_{ni}(u)-\frac{1}{N}\sum_{i\in J_n}\varepsilon^i\xi_{ni}(u_r)\right| \\
&\quad+\max_{1\le r\le c(n)}\left|\frac{1}{N}\sum_{i\in J_n}\varepsilon^i\xi_{ni}(u_r)\right| \\
&\equiv Q_1+Q_2,\quad\text{say.}
\end{aligned}$$

It is straightforward to see that $|\xi_{ni}(u) - \xi_{ni}(u_r)| \le al(n)$ for some constant a and that $Q_1 = O(l(n))$ with probability 1. To handle the second term we must use an exponential inequality and a blocking argument as in Masry's proof. In conclusion, by appropriate choice of $c(n)$, we obtain $Q_1 + Q_2 = O(\log n/\sqrt{n})$ with probability 1. □

PROOF OF (112) AND (113). Note that by definition,

$$\hat{m}_j^B(x_j) = N^{-1} \sum_{i \in J_n} K_h\big(x_j, X_j^i\big) m(X^i)/\hat{p}_j(x_j)$$
$$= N^{-1} \sum_{i \in J_n} K_h\big(x_j, X_j^i\big)\big[m_0 + m_1\big(X_1^i\big) + \cdots + m_d\big(X_d^i\big)\big]/\hat{p}_j(x_j)$$

and

$$\hat{\mu}_{n,j}(x_j) = m_j(x_j) + m_j'(x_j) \int K_h(x_j, u)(u - x_j)\, du \left[\int_0^1 K_h(x_j, u)\, du\right]^{-1}$$
$$+ \sum_{k \ne j,\, k,\, j \in J_n} \int m_k(x_k) \frac{\hat{p}_{j,k}(x_j, x_k)}{\hat{p}_j(x_j)}\, dx_k$$
$$+ \sum_{k \ne j,\, k,\, j \in J_n} \int m_k'(x_k) \frac{\hat{p}_{j,k}(x_j, x_k)}{\hat{p}_j(x_j)} K_h(x_k, u)(u - x_k)$$
$$\times \left[\int_0^1 K_h(x_k, v)\, dv\right]^{-1} du\, dx_k$$
$$+ h^2 d_K p_j(x_j)^{-1} \sum_{k=1}^{d} \int \left[\frac{\partial p(x)}{\partial x_k} m_k'(x_k) + \frac{1}{2} p(x) m_k''(x_k)\right] dx_{-j}$$

with $d_k = \int u^2 K(u)\, du$. We argue now that for $j = 1, \ldots, d$,

$$N^{-1} \sum_{i \in J_n} K_h\big(x_j, X_j^i\big) m_j\big(X_j^i\big)/\hat{p}_j(x_j)$$
$$= m_j(x_j) + m_j'(x_j)$$
$$\times \int K_h(x_j, u)(u - x_j)\, du \left[\int_0^1 K_h(x_j, u)\, du\right]^{-1} \tag{115}$$
$$+ h^2 \int u^2 K(u)\, du\, p_j(x_j)^{-1}\left[p_j'(x_j) m_j'(x_j) + \tfrac{1}{2} p_j(x_j) m_j''(x_j)\right]$$
$$+ R_{n,j}(x_j)$$

with $\sup_{x_j \in I_h} |R_{n,j}(x_j)| = o_P(h^2)$ and $\sup_{x_j \in I_h^c} |R_{n,j}(x_j)| = O_P(h^2)$. Furthermore, we argue for $j \ne k$ that

$$
\begin{aligned}
&N^{-1}\sum_{i\in J_n} K_h\bigl(x_j, X_j^i\bigr)m_k\bigl(X_k^i\bigr)/\hat{p}_j(x_j)\\
&\quad=\int m_k(x_k)\frac{\hat{p}_{j,k}(x_j,x_k)}{\hat{p}_j(x_j)}\,dx_k\\
&\qquad+\int m_k'(x_k)\frac{\hat{p}_{j,k}(x_j,x_k)}{\hat{p}_j(x_j)}\\
&\qquad\quad\times K_h(x_k,u)(u-x_k)\biggl[\int_0^1 K_h(x_k,v)\,dv\biggr]^{-1}du\,dx_k\\
&\qquad+h^2 d_K p_j(x_j)^{-1}\\
&\qquad\quad\times\int\biggl[\frac{\partial p_{j,k}(x_j,x_k)}{\partial x_k}m_k'(x_k)+\tfrac{1}{2}p_{j,k}(x_j,x_k)m_k''(x_k)\biggr]dx_k\\
&\qquad+R_{n,j,k}(x_j)
\end{aligned}
\tag{116}
$$

with $\sup_{x_j\in I_h}|R_{n,j,k}(x_j)|=o_P(h^2)$ and $\sup_{x_j\in I_h^c}|R_{n,j,k}(x_j)|=O_P(h^2)$. It can be easily verified that (115) and (116) imply (112) and (113). So it remains to show (115) and (116). The proof of (115) is straightforward and will be omitted. For the proof of (116) note that for $k\neq j$ and uniformly for $x_j\in[0,1]$,

$$
\begin{aligned}
&\frac{1}{N}\sum_{i\in J_n} K_h\bigl(x_j,X_j^i\bigr)m_k\bigl(X_k^i\bigr)\\
&\quad=\frac{1}{N}\sum_{i\in J_n}\int K_h\bigl(x_j,X_j^i\bigr)K_h\bigl(x_k,X_k^i\bigr)m_k\bigl(X_k^i\bigr)\,dx_k\\
&\quad=\frac{1}{n}\sum_{i\in J_n}\int K_h\bigl(x_j,X_j^i\bigr)K_h\bigl(x_k,X_k^i\bigr)\\
&\qquad\qquad\times\Bigl[m_k(x_k)+\bigl(X_k^i-x_k\bigr)m_k'(x_k)+\tfrac{1}{2}\bigl(X_k^i-x_k\bigr)^2 m_k''(x_k)\Bigr]dx_k\\
&\qquad+o_P(h^2)\\
&\quad=\int\hat{p}_{jk}(x_j,x_k)m_k(x_k)\,dx_k+\frac{1}{N}\sum_{i\in J_n}\bigl[U_i(x_j)+V_i(x_j)\bigr]+o_P(h^2),
\end{aligned}
$$

where

$$
\begin{aligned}
U_i(x_j)&=\int K_h\bigl(x_j,X_j^i\bigr)K_h\bigl(x_k,X_k^i\bigr)\bigl(X_k^i-x_k\bigr)m_k'(x_k)\,dx_k,\\
V_i(x_j)&=\int K_h\bigl(x_j,X_j^i\bigr)K_h\bigl(x_k,X_k^i\bigr)\tfrac{1}{2}\bigl(X_k^i-x_k\bigr)^2 m_k''(x_k)\,dx_k.
\end{aligned}
\tag{117}
$$

For $x_j \in I_h$, claim (116) follows now from (104) and

$$\sup_{x_j \in I_h} \Bigg| E\big[U_i(x_j)\big] - \int m_k'(x_k)\, p_{j,k}(x_j, x_k) \times K_h(x_k, u)(u - x_k)\, du\, dx_k - h^2 d_K \int \frac{\partial p_{j,k}(x_j, x_k)}{\partial x_k} m_k'(x_k)\, dx_k \Bigg| = o(h^2), \tag{118}$$

$$\sup_{x_j \in I_h} \Big| E\big[V_i(x_j)\big] - h^2 d_K \int \tfrac{1}{2} p_{j,k}(x_j, x_k) m_k''(x_k)\, dx_k \Big| = o(h^2), \tag{119}$$

$$\sup_{x_j \in I_h} \Bigg| \int m_k'(x_k) \Bigg[\hat{p}_{j,k}(x_j, x_k) \Bigg[\int_0^1 K_h(x_k, w)\, dw \Bigg]^{-1} - p_{j,k}(x_j, x_k) \Bigg] \times K_h(x_k, u)(u - x_k)\, du\, dx_k \Bigg| = o_P(h^2), \tag{120}$$

$$\sup_{x_j \in I_h} \big| U_i(x_j) - E\big[U_i(x_j)\big] \big| = o_P(h^2), \tag{121}$$

$$\sup_{x_j \in I_h} \big| V_i(x_j) - E\big[V_i(x_j)\big] \big| = o_P(h^2), \tag{122}$$

Claims (118) and (119) follow by standard kernel arguments. For the proof of (12) one applies (103) and (105). For the proof of (121) and (122) one proceeds similarly to Masry (1996b); see also the proof of (A6). So it remains to show (116) for $x_j \in I_h^c$. This can be done by similar arguments. □

PROOF OF THEOREM 4′. Theorem 4′ can be shown by arguments similar to the proof of Theorem 4. First one shows uniform convergence of $\hat{\mathbf{M}}_j(x_j)$ to $\mathbf{M}_j(x_j)$ and of $\hat{\mathbf{S}}_{l,j}(x_l, x_j)$ to $\mathbf{S}_{l,j}(x_l, x_j)$. For the proof of (A9′) one needs an expansion of

$$\begin{pmatrix} \hat{m}_j^B(x_j) \\ \hat{m}^{j,B}(x_j) \end{pmatrix} = \hat{\mathbf{M}}_j(x_j)^{-1} \frac{1}{N} \sum_{i=1}^n \mathbf{1}\big(X^i \in [0,1]^d\big) K_h\big(x_j, X_j^i\big) \begin{pmatrix} 1 \\ h^{-1}\big[X_j^i - x_j\big] \end{pmatrix} \times \big[m_0 + m_1\big(X_1^i\big) + \cdots + m_d\big(X_d^i\big)\big].$$

For the treatment of this quantity one has to consider for $k \neq j$ the term

$$\frac{1}{N} \sum_{i=1}^n \mathbf{1}\big(X^i \in [0,1]^d\big) K_h\big(x_j, X_j^i\big) \begin{pmatrix} 1 \\ h^{-1}\big[X_j^i - x_j\big] \end{pmatrix} m_i\big(X_k^i\big).$$

Using $\int K_h(x_k, X_k^i)\,dx_k = 1$ and with $V_i(x_j)$ defined as in (117) one gets that this term is equal to

$$\frac{1}{N}\sum_{i=1}^{n} \mathbf{1}\big(X^i \in [0,1]^d\big)\int K_h\big(x_j, X_j^i\big)K_h\big(x_k, X_k^i\big)\begin{pmatrix}1\\ h^{-1}\big[X_j^i - x_j\big]\end{pmatrix} m_k\big(X_k^i\big)\,dx_k$$

$$= \frac{1}{N}\sum_{i=1}^{n} \mathbf{1}\big(X^i \in [0,1]^d\big)\int K_h\big(x_j, X_j^i\big)K_h\big(x_k, X_k^i\big)\begin{pmatrix}1\\ h^{-1}\big[X_j^i - x_j\big]\end{pmatrix}$$

$$\times\big[m_k(x_k) + m_k'(x_k)\{X_k^i - x_k\}\big]\,dx_k + \frac{1}{N}\sum_{i\in J_n}\begin{pmatrix}V_i(x_j)\\ 0\end{pmatrix} + o_P(h^2)$$

$$= \int \hat{\mathbf{S}}_{j,k}(x_j, x_k)\begin{pmatrix}m_k(x_k)\\ hm_k'(x_k)\end{pmatrix} dx_k + \frac{1}{N}\sum_{i\in J_n}\begin{pmatrix}V_i(x_j)\\ 0\end{pmatrix} + o_P(h^2).$$

For a further treatment of this expansion one uses now (119) and (122) and proceeds similarly to the proof of Theorem 4. □

REFERENCES

AUESTAD, B. and TJØSTHEIM, D. (1991). Functional identification in nonlinear time series. In *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.) 493–507. Kluwer, Amsterdam.

BALAKRISHNAN, A. V. (1981). *Applied Functional Analysis*. Springer, New York.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. John Hopkins Univ. Press.

DEUTSCH, F. (1985). Rate of convergence of the method of alternating projections. In *Parametric Optimization and Approximation* (B. Brosowski and F. Deutsch, ed.) 96–107. Birkhäuser, Basel.

FAN, J., MAMMEN, E. and HÄRDLE, W. (1998). Direct estimation of low-dimensional components in additive models. *Ann. Statist.* **26** 943–971.

HÄRDLE, W. (1991). *Applied Nonparametric Regression*. Cambridge Univ. Press.

HASTIE, T. and TIBSHIRANI, R. (1991). *Generalized Additive Models*. Chapman and Hall, London.

LINTON, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika* **84** 469–474.

LINTON, O. B. and NIELSEN, J. P. (1995). Estimating structured nonparametric regression by the kernel method. *Biometrika* **82** 93–101.

MAMMEN, E., MARRON, J. S., TURLACH, B. and WAND, M. P. (1997). A general framework for smoothing. Preprint.

MASRY, E. (1996a). Multivariate regression estimation: local polynomial fitting for time series. *Stochastic Processes Appl.* **65** 81–101.

MASRY, E. (1996b). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.* **17** 571–599.

NEWEY, W. K. (1994). Kernel estimation of partial means. *Econom. Theory* **10** 233–253.

OPSOMER, J. D. (1998). On the existence and asymptotic properties of backfitting estimators. *Ann. Statist.* To appear.

OPSOMER, J. D. and RUPPERT, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25** 186–211.

ROSENBLATT, M. (1956). A central limit theorem and strong mixing conditions. *Proc. Nat. Acad. Sci. U.S.A.* **4** 43–47.

RUPPERT, D. and WAND, M. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22** 1346–1370.

SMITH, K. T., SOLOMON, D. C. and WAGNER, S. L. (1977). Practical and mathematical aspects of the problem of reconstructing objects from radiographs. *Bull. Amer. Math. Soc.* **83** 1227–1270.

TJØSTHEIM, D. and AUESTAD, B. (1994). Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.* **89** 1398–1409.

E. MAMMEN
INSTITUT FÜR ANGEWANDTE MATHEMATIK
RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
69120 HEIDELBERG
GERMANY
E-MAIL: mammen@statlab.uni-heidelberg.de

O. LINTON
COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
30 HILLHOUSE AVENUE
NEW HAVEN, CONNECTICUT 06520-8281
E-MAIL: linton@econ.yale.edu

J. NIELSEN
CODANHUS
60 GAMMEL KONGEVEJ
DK-1790 COPENHAGEN V
DENMARK
E-MAIL: npj@codan.dk