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

O. Linton[1], E. Mammen[2], and J. Nielsen[3]

May 8, 1998

Abstract

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 et al. (1997), and the asymptotic theory of our estimators is derived using the theory of additive projections reviewed in Bickel et al. (1995). 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). We provide 'high level' conditions independent of the sampling scheme. We then verify that these conditions are satisfied in a time series autoregression under weak conditions.

AMS 1991 subject classifications. primary 62G07 , secondary 62G20

Keywords and phrases. Additive models; Alternating projections; Backfitting; Kernel Smoothing; Local Polynomials; Nonparametric Regression.

Short title. Backfitting 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 proven 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 non-statisticians 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. Fan, Mammen, and Härdle (1996) and Linton (1996) showed how to improve on the efficiency of the marginal integration estimator in regression – in the latter paper, this was achieved by carrying out one backfitting iteration from this initial consistent

[1] Cowles Foundation for Research in Economics, Yale University, 30 Hillhouse Avenue, New Haven, CT 06520-8281, USA. Phone: (203) 432-3699. Fax: (203) 432-6167. http://www.econ.yale.edu/~linton. Supported by the National Science Foundation and the North Atlantic Treaty Organization.

[2] Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany; Supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation ökonomischer Prozesse", Humboldt-Universität zu Berlin.

[3] PFA Pension, Sundkrogsgade 4, DK-2100 Copenhagen, Denmark

starting point. This modification actually achieves full oracle efficiency, i.e., 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 (1997) 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 (recentred) 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 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 rather strong conditions; 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 et al. (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 in 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 et al. (1995). We use this result to establish the limiting behaviour 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 a time series regression with strong mixing data. Our conditions are strictly weaker than those of Opsomer and Ruppert (1997), and do not necessarily restrict the dependence between the covariates in any way.

This 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, i.e., 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 investigate a time series setting and give primitive conditions that imply the high level conditions. In section 6 we illustrate our procedure on financial data. 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 $x = (x_1, \ldots, x_d)^T$, see Mammen, Marron, Turlach and Wand (1997). This new point of view will be useful

for interpreting our estimators of the restricted additive function $m(x) = \mu + m_1(x_1) + \cdots + m_d(x_d)$.

The full dimensional local [q'th order] polynomial regression smoother, which we denote by $\widehat{\mathbf{m}}(x) = (\widehat{m}^0(x), \ldots, \widehat{m}^{qd}(x))^T$, satisfies

$$\widehat{\mathbf{m}}(x) = \underset{\mathbf{m}=(m^0,\ldots,m^{qd})^T}{\arg\min} \sum_{i=1}^{n} \left\{ Y^i - m^0 - \left(\frac{X_1^i - x_1}{h}\right) m^1 - \ldots - \left(\frac{X_d^i - x_d}{h}\right)^q m^{qd} \right\}^2 \prod_{\ell=1}^{d} K_h(X_\ell^i - x_\ell), \tag{1}$$

where q is the order of the polynomial approximation. In fact, for simplicity of notation we will concentrate on the local linear case considered in Ruppert and Wand (1995) for which $q = 1$ – the Nadaraya-Watson case, for which $q = 0$, is even simpler, see below. Define the matrices [of dimension $n \times (d+1)$ and $n \times n$, respectively]

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

$$\mathbf{K}(x) = \operatorname{diag}\left(\textstyle\prod_{\ell=1}^{d} K_h(X_\ell^1 - x_\ell), \quad \cdots \quad , \prod_{\ell=1}^{d} K_h(X_\ell^n - x_\ell) \right),$$

and write

$$\widehat{\mathbf{m}}(x) = \left\{ \mathbf{X}(x)^T \mathbf{K}(x) \mathbf{X}(x) \right\}^{-1} \mathbf{X}(x)^T \, \mathbf{K}(x) \mathbf{Y} \equiv \widehat{\mathbf{V}}^{-1}(x) \widehat{\mathbf{R}}(x), \tag{2}$$

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

For the new interpretation of local linear estimators we think of the data $\mathbf{Y} = (Y^1, \ldots, Y^n)^T$ as an element of the space of tuples of $2n$ functions

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

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 norm on $\mathcal{F}$,

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

where $K_h(\cdot) = K(\cdot/h)/h$ with $K(\cdot)$ a univariate kernel. Consider now the following subspaces of $\mathcal{F}$:

$$\begin{aligned}
\mathcal{F}_{full} &= \{f \in \mathcal{F} : f^{i,j} \text{ does not depend on } i \text{ for } j = 0, \ldots, d\} \\
\mathcal{F}_{add} &= \{f \in \mathcal{F}_{full} : f^{i,0}(x) = g_1(x_1) + \ldots g_d(x_d) \text{ for some functions } g_j : \mathbb{R} \to \mathbb{R} \text{ for } j = 1, \ldots, d \\
&\qquad \text{and } f^{i,j}(x) = g^j(x_j) \text{ for some functions } g^j : \mathbb{R} \to \mathbb{R} \text{ for } j = 1, \ldots, d \text{ if } j \neq 0\}.
\end{aligned}$$

The estimate $\widehat{\mathbf{m}}(x)$ defines an element of $\mathcal{F}$ by putting $f^{i,j}(x) = \widehat{m}^j(x)$, $j = 0, 1, \ldots, d$. This is an element of $\mathcal{F}_{full}$. It is easy to see that, with respect to $\| \quad \|_*$, $\widehat{\mathbf{m}}$ is the orthogonal projection of $\mathbf{Y}$ onto $\mathcal{F}_{full}$. Below we introduce our version $\widetilde{\mathbf{m}}$ of the backfitting estimator as the orthogonal projection of $\widehat{\mathbf{m}}$ onto $\mathcal{F}_{add}$ [with respect to $\| \quad \|_*$]. For an understanding of $\widetilde{\mathbf{m}}$ it will become essential that it be the orthogonal projection of $\mathbf{Y}$ onto $\mathcal{F}_{add}$. 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 $\mathcal{F}_{add}$. In particular, for the local constant estimator (Nadaraya Watson-like smoothers) one chooses:

$$\begin{aligned}
\mathcal{F} &= \left\{(f^i : i = 1, \ldots, n) : \text{ Here, } f^i \text{ are functions from } \mathbb{R}^d \text{ to } \mathbb{R}\right\} \\
\mathcal{F}_{full} &= \left\{f \in \mathcal{F} : f^i \text{ does not depend on } i\right\} \\
\mathcal{F}_{add} &= \left\{f \in \mathcal{F}_{full} : f^i(x) = g_1(x_1) + \ldots g_d(x_d) \text{ for some functions } g_j : \mathbb{R} \to \mathbb{R}\right\} \\
\|f\|_*^2 &= \int \frac{1}{n} \sum_{i=1}^{n} [f^i(x)]^2 \prod_{j=1}^{d} K_h(X_j^i - x_j)\, dx.
\end{aligned}$$

Note that for functions $\mathbf{m}$ in $\mathcal{F}_{full}$ [i.e. $m := m^1 = \ldots = m^n$] we get

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

where $\hat{p}(x) = n^{-1} \sum_{j=1}^{n} K_h(X_j^i - x_j)$ is the kernel density estimate of the design density. In particular, in this case $\widetilde{\mathbf{m}}$ is the projection of the full dimensional Nadaraya-Watson estimate onto the subspace

of additive with respect to the norm of the space $\mathbf{L}_2(\widehat{p})$. We give a slightly different motivation for the projection estimate $\widetilde{\mathbf{m}}$ 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 motivate our backfitting estimate based on regression smoothers like the Nadaraya-Watson

$$\widehat{m}(x) = \frac{n^{-1}\sum_{i=1}^{n}\prod_{\ell=1}^{d}K_h(x_\ell - X_\ell^i)Y^i}{n^{-1}\sum_{i=1}^{n}\prod_{\ell=1}^{d}K_h(x_\ell - X_\ell^i)}. \tag{3}$$

The specific choice of the Nadaraya-Watson estimator is not important, but the smoother is supposed to have the ratio form

$$\widehat{m}(x) = \frac{\widehat{r}(x)}{\widehat{p}(x)} = \sum_{i=1}^{n} w_i(x)Y^i, \tag{4}$$

where $\widehat{p}(x)$, which depends only on $\mathcal{X}^n = \{X^1, \ldots, X^n\}$, is an estimator of $p(x)$, the marginal density of X. Here, the weighting sequence $\{w_i(x)\}_{i=1}^n$ only depends on $\mathcal{X}^n$, as does the weighting sequence $\{w_i^*(x)\}_{i=1}^n$ of the numerator $\widehat{r}(x) = \sum_{i=1}^n w_i^*(x)Y^i$. The assumption that the pilot estimate $\widehat{m}$ exists [i.e., is everywhere and always finite] 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

$$m(x) = \mu + m_1(x_1) + \ldots + m_d(x_d), \tag{5}$$

although our definitions make sense more generally i.e., when the regression function is not additive, in which case the target function is the closest additive approximation to the regression function. For identifiability we assume that

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

where the marginal density of X_j is denoted by $p_j(\cdot)$. Denote also the marginal density of (X_i, X_j) by $p_{ij}(\cdot,\cdot)$ respectively $(i, j = 1, \ldots, d)$. The vector $(X_i : i \neq j)$ is denoted by X_{-j} and its density by p_{-j}.

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

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

In the sample, one replaces $E(Y|X_j = x_j)$ by one-dimensional smoothers $\widehat{m}_j(\cdot)$, and iterates from some arbitrary starting values for $m_j(\cdot)$ see Hastie and Tibshirani (1991, p. 108). Let $\widehat{p}(x)$ and $\widehat{m}(x)$ be multidimensional density and regression smoothers defined above. We define backfitting estimates $\widetilde{m}_j$ as the minimizers of the following norm

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

where the minimization runs over all functions $\overline{m}(x) = \mu + \sum_j \bar{m}_j(x_j)$, with $\int \bar{m}_j(x_j)\widehat{p}_j(x_j)dx_j = 0$, see Nielsen and Linton (1996) [we suppose that the density estimate $\widehat{p}$ is non-negative]. This means that $\widetilde{m}(x) = \widehat{\mu} + \widetilde{m}_1(x_1) + \ldots + \widetilde{m}_j(x_d)$ is the projection in the space $\mathbf{L}_2(\widehat{p})$ of $\widehat{m}$ onto the affine subspace of additive functions $\{m \in \mathbf{L}_2(\widehat{p}) : m(x) = \mu + m_1(x_1) + \ldots + m_d(x_d)\}$. This is a central point of our discussion. For projection operators backfitting is well understood (method of alternating projections, see below). Therefore, this interpretation will enable us to understand convergence of the backfitting algorithm and the asymptotics of $\widetilde{m}_j$. 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 ($j = 1, \ldots, d$):

$$\widetilde{m}_j(x_j) = \int \widehat{m}(x)\frac{\widehat{p}(x)}{\widehat{p}_j(x_j)}dx_{-j} - \sum_{k\neq j}\int \widetilde{m}_k(x_k)\frac{\widehat{p}(x)}{\widehat{p}_j(x_j)}dx_{-j} - \widehat{\mu} \tag{8}$$

$$\widehat{\mu} = \int \widehat{m}(x)\widehat{p}(x)dx, \tag{9}$$

where $\widehat{m}_j(x_j) = n^{-1}\sum_{i=1}^n K_h(x_j - X_j^i)Y^i/\widehat{p}_j(x_j)$ is the univariate Nadaraya-Watson regression smoother, in which $\widehat{p}_j(x_j) = \int \widehat{p}(x)dx_{-j}$ is the marginal of the density estimate $\widehat{p}(x)$. Straightforward algebra gives

$$\begin{aligned}\int \widehat{m}(x)\frac{\widehat{p}(x)}{\widehat{p}_j(x_j)}dx_{-j} &= \widehat{p}_j^{-1}(x_j)n^{-1}\sum_{i=1}^n K_h(x_j - X_j^i)Y^i \int \prod_{\ell\neq j} K_h(x_\ell - X_\ell^i)dx_{-j} \\ &= \widehat{m}_j(x_j).\end{aligned}$$

Furthermore, $\widehat{\mu} = \int \widehat{m}(x)\widehat{p}(x)dx = \int \widehat{r}(x)dx$, and when $\int w_j(x)dx = 1$, we find, as in Hastie and Tibshirani (1991), that $\widehat{\mu} = n^{-1}\sum_{i=1}^n Y^i$, i.e., that $\widehat{\mu}$ is the sample mean. So $\widehat{\mu}$ 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

$$\widehat{\mu} = \int \widehat{m}_j(x_j)\widehat{p}_j(x_j)\ dx_j \quad \text{for } j = 1, \ldots, d. \tag{10}$$

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

$$\widetilde{m}_j(x_j) = \widehat{m}_j(x_j) - \sum_{k\neq j}\int \widetilde{m}_k(x_k)\frac{\widehat{p}(x)}{\widehat{p}_j(x_j)}dx_{-j} - \widehat{\mu}, \quad j = 1, \ldots, d,$$

with $\widehat{\mu}$ defined by (10). Up to now we have assumed that multivariate estimates of the density and of the regression function exist. 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

$$\widetilde{m}_j(x_j) = \widehat{m}_j(x_j) - \sum_{k \neq j} \int \widetilde{m}_k(x_k) \frac{\widehat{p}_{jk}(x_j, x_k)}{\widehat{p}_j(x_j)} dx_k - \widehat{\mu}. \tag{11}$$

In this equation only two dimensional marginals of $\widehat{p}$ are used. Note also that the solutions $\widetilde{m}_j(x_j)$ to (11) inherit the smoothness properties of $\widehat{m}(x)$ and $\widehat{p}(x)$. We can therefore estimate the derivatives of $m_j(x_j)$, for example, by

$$\frac{d^r \widetilde{m}_j(x_j)}{dx_j^r} = \frac{d^r \widehat{m}_j(x_j)}{dx_j^r} - \sum_{k \neq j} \int \widetilde{m}_k(x_k) \frac{d^r}{dx_j^r} \left\{ \frac{\widehat{p}_{j,k}(x_j, x_k)}{\widehat{p}_j(x_j)} \right\} dx_k, \ r = 1, 2, \ldots$$

In the next section we will discuss estimates $\widetilde{m}_j$ that are defined by (11) along with their asymptotic properties. In practice, our backfitting algorithm works as follows. One starts with an arbitrary initial guess $\widetilde{m}_j^{[0]}$ for $\widetilde{m}_j$. In the j-th step of the r-th iteration cycle one puts

$$\widetilde{m}_j^{[r]}(x_j) = \widehat{m}_j(x_j) - \sum_{k<j} \int \widetilde{m}_k^{[r]}(x_k) \frac{\widehat{p}_{jk}(x_j, x_k)}{\widehat{p}_j(x_j)} dx_k - \sum_{k>j} \int \widetilde{m}_k^{[r-1]}(x_k) \frac{\widehat{p}_{jk}(x_j, x_k)}{\widehat{p}_j(x_j)} dx_k - \widehat{\mu},$$

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

3.1 Asymptotics for the Nadaraya-Watson-like Version

We consider now estimates $\widetilde{m}_j$ that are defined by (11) with $\widehat{\mu}$ defined by (10), where $\widehat{m}_j$, $\widehat{p}_{jk}$, and $\widehat{p}_j$ are some given estimates. The next theorem gives conditions under which, with probability tending to one, there exists a solution $\widetilde{m}_j$ of (11) 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 $\widehat{m}_j$, $\widehat{p}_{jk}$, and $\widehat{p}_j$ [for example, we do not require that p is the underlying density of X or that $\widehat{m}_j$, $\widehat{p}_{jk}$, and $\widehat{p}_j$ are kernel estimates] – these properties can be verified for a range of smoothers under quite general heterogeneous and dependent sampling schemes, and we investigate this in section 5 below.

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*

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

Furthermore,

$$\int \widehat{m}_j(x_j)\widehat{p}_j(x_j)\ dx_j \equiv \text{ const.}$$

By definition this constant is equal to $\widehat{\mu}$, see (10).

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

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

(A4) *There exists a constant C such that with probability tending to one for all $j \neq k$,*

$$\sup_{x_k} \int \frac{\widehat{p}_{j,k}^2(x_j, x_k)}{\widehat{p}_k^2(x_k)\widehat{p}_j(x_j)} dx_j \leq C.$$

(A5) *We suppose that for a sequence* $\Delta_n \downarrow 0$ *the one-dimensional smoothers* $\widehat{m}_j$ *can be decomposed as* $\widehat{m}_j = \widehat{m}_j^A + \widehat{m}_j^B$ *with* $\int \widehat{m}_j(x_j)\widehat{p}_j(x_j)\,dx_j$ *not depending on* j *and, where the first component* $\widehat{m}_j^A$ *is mean zero and satisfies*

$$\sup_{x_k} \left| \int \frac{\widehat{p}_{j,k}(x_j, x_k)}{\widehat{p}_k(x_k)} \widehat{m}_j^A(x_j) dx_j \right| = o_P\left(\frac{\Delta_n}{\log n}\right).$$

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

$$\widetilde{m}_j^s(x_j) = \widehat{m}_j^s(x_j) - \sum_{k \neq j} \int \widetilde{m}_k^s(x_k) \frac{\widehat{p}_{jk}(x_j, x_k)}{\widehat{p}_j(x_j)} dx_k - \widehat{\mu}^s, \tag{12}$$

where $\widehat{\mu}^s = \int \widehat{m}^s(x)\widehat{p}(x)dx$. *Existence and uniqueness of* $\widetilde{m}_j^A$ *and* $\widetilde{m}_j^B$ *is stated in the next theorem. Note that* $\widetilde{m}_j^s$ *is defined as* $\widetilde{m}_j$ *in equation (11) with* $\widehat{m}_j$ *replaced by* $\widehat{m}_j^s$. *We suppose that for (deterministic) functions* $\mu_{j,n}(\cdot)$ *the term* $\widetilde{m}_j^B$ *satisfies*

$$\widetilde{m}_j^B(x_j) = \mu_{j,n}(x_j) + o_P(\Delta_n).$$

These conditions, which we discuss further below, are all straightforward to verify, except perhaps A5, and turn out to be weaker than those made by Opsomer and Ruppert (1997).

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

THEOREM 1 [CONVERGENCE OF BACKFITTING]. *Suppose that conditions A1-A2 hold. Then, with probability tending to one, there exists a solution* $\widetilde{m}_j$ *of (11) and (10) that is unique. Furthermore there exist constants* $0 < \gamma < 1$ *and* $c > 0$ *such that, with probability tending to one, the following inequality holds*

$$\int \left[\widetilde{m}_j^{[r]}(x_j) - \widetilde{m}_j(x_j)\right]^2 p_j(x_j)dx_j \leq c\gamma^{2r} \int \left[\widetilde{m}^{[0]}(x)\right]^2 p(x)dx. \tag{13}$$

Here, for $r = 0$ *the function* $\widetilde{m}^{[r]}(x) = \widetilde{m}_1^{[r]}(x_1) + \ldots + \widetilde{m}_d^{[r]}(x_d)$ *is the starting value of the backfitting algorithm.*

Furthermore, for $s = A$ and $s = B$, with probability tending to one there exists a solution $\widetilde{m}_j^s$ of (12) 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 for the special case that an asymptotic theory is available for the pilot estimate $\widehat{m}$.

THEOREM 2. *Suppose that conditions A1 - A5 hold for a sequence Δ_n. Then, it holds that*

$$\sup_{x_j} |\widetilde{m}_j^A(x_j) - \widehat{m}_j^A(x_j)| = o_P(\Delta_n).$$

In particular, one gets

$$\widetilde{m}_j(x_j) = \widehat{m}_j^A(x_j) + \mu_{j,n}(x_j) + o_P(\Delta_n).$$

We now apply Theorem 2 to the case that full dimensional pilot estimates $\widehat{p}(x)$, $\widehat{r}(x)$ and $\widehat{m}(x) = \widehat{r}(x)/\widehat{p}(x) = \sum_{i=1}^n w_i(x)Y^i$ exist and that $\widehat{\mu}, \widetilde{m}_1, \ldots, \widetilde{m}_d$ are defined as minimizers of (7) [i.e., $\widehat{\mu} + \widetilde{m}_1 + \ldots + \widetilde{m}_d$ is the projection of $\widehat{m}$ onto the class of additive functions in $\mathrm{L}_2(\widehat{p})$.] For the one-dimensional smooths, $\widehat{m}_j$, we have, with appropriate weights $w_{ji}(x_j)$, that

$$\widehat{m}_j(x_j) = \int \widehat{m}(x) \frac{\widehat{p}(x)}{\widehat{p}_j(x_j)} \, dx_{-j} = \sum_{i=1}^n w_{ji}(x) Y^i.$$

We compare now the estimate $\widetilde{m}_j$ with the infeasible estimate $\ddot{m}_j$ that uses the knowledge of the other components m_l with $l \neq j$. More precisely, we define the infeasible estimator $\ddot{m}_j(x_j)$ to be the one-dimensional smooth of the unobserved data $Y_*^i = m_j(X_j^i) + \varepsilon^i$ [with $\varepsilon^i = Y^i - \mu - \sum_{k=1}^n m_k(X_k^i)$] on X_j^i, thus

$$\ddot{m}_j(x_j) = \sum_{i=1}^n w_{ji}(x_j) Y_*^i, \quad j = 1, \ldots, d. \tag{14}$$

Then, under appropriate regularity conditions,

$$n^{2/5}\left\{\ddot{m}_j(x_j) - m_j(x_j)\right\} \Longrightarrow N\left\{\ddot{b}_j(x_j), \ddot{v}_j(x_j)\right\}, \quad j = 1, \ldots, d, \tag{15}$$

for certain functions $\ddot{b}_j(\cdot)$ and $\ddot{v}_j(\cdot)$. Moreover, because of $\text{cov}\left\{\ddot{m}_j(x_j), \ddot{m}_k(x_k)\right\} = o(n^{-4/5})$ one has

$$n^{2/5}\left\{\ddot{m}_j(x_j) - m_j(x_j)\right\} \text{ and } n^{2/5}\left\{\ddot{m}_k(x_k) - m_k(x_k)\right\} \text{ are asymptotically independent for } j \neq k. \tag{16}$$

The additional information that $\int m_j(x_j)p_j(x_j)dx_j = 0$ may have some value, and we can define the mean corrected version of $\ddot{m}_j(x_j)$, by $\ddot{m}_j^c(x_j) = \ddot{m}_j(x_j) - n^{-1}\sum_{i=1}^n \ddot{m}_j(X_j^i)$, which has the same asymptotic variance as $\ddot{m}_j(x_j)$ but bias $\ddot{b}_j^c(x) = \ddot{b}_j(x) - \int \ddot{b}_j(x)p_j(x_j)dx_j$.

We suppose now that our conditions hold with $\widehat{m}^A(x) = \sum_{i=1}^n w_i(x)\varepsilon^i$ and $\widehat{m}^B(x) = \sum_{i=1}^n w_i(x)$ $m(X^i)$. One can decompose

$$\begin{aligned}
\widehat{m}_j^A(x_j) &= \int \widehat{m}^A(x)\frac{\widehat{p}(x)}{\widehat{p}_j(x_j)}dx_{-j} = \sum_{i=1}^n w_{ji}(x_j)\varepsilon^i \\
\widehat{m}_j^B(x_j) &= \int \widehat{m}^B(x)\frac{\widehat{p}(x)}{\widehat{p}_j(x_j)}dx_{-j} = \sum_{i=1}^n w_{ji}(x_j)m(X_j^i).
\end{aligned}$$

Suppose now that it can be shown for a function b that

$$\widehat{m}^B(x) = m(x) + n^{-2/5}b(x) + o_P(n^{-2/5}). \tag{17}$$

We have the following

COROLLARY 1. *Suppose that conditions A1-A5 hold with $\Delta_n = n^{-2/5}$, and that (14) - (17) apply. Then*

$$n^{2/5}\begin{bmatrix} \widetilde{m}_1(x_1) - m_1(x_1) \\ \vdots \\ \widetilde{m}_d(x_d) - m_d(x_d) \end{bmatrix} \Longrightarrow N\left(\begin{bmatrix} b_1(x_1) \\ \vdots \\ b_d(x_d) \end{bmatrix}, \begin{bmatrix} v_1(x_1) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & v_d(x_d) \end{bmatrix}\right),$$

where $v_j(x_j) = \ddot{v}_j(x_j)$, $j = 1, \ldots, d$, *are defined above, while* $b_j(x_j)$ *are solutions to the following minimization problem*

$$\min_{\mu, b_1(\cdot), \ldots, b_d(\cdot)} \int \left[b(x) - \mu - b_1(x_1) - \ldots - b_d(x_d) \right]^2 p(x) dx, \quad \text{s.t.} \int b_j(x_j) p_j(x_j) dx_j = 0, \; j = 1, \ldots, d.$$

For the special case that the function b *is already of additive form* $b(x) = b_1^*(x_1) + \ldots + b_d^*(x_d)$, *the bias functions* $b_j(x_j)$ *coincide with the bias* $\ddot{b}_j^c(x_j)$ *of the 'corrected' oracle estimate* $\ddot{m}_j^c(x_j)$. *Also*

$$n^{2/5} \left\{ \widetilde{m}(x) - m(x) \right\} \Longrightarrow N \left[b_+(x), v_+(x) \right],$$

where $b_+(x) = \sum_j b_j(x_j)$ *and* $v_+(x) = \sum_j v_j(x_j)$.

Suppose additionally that for a sequence δ_n *with* $n^{-2/5} = o(\delta_n)$

$$\begin{aligned} \sup_x |\widehat{m}^B(x) - m(x) - n^{-2/5} b(x)| &= O_P(\delta_n) \\ \sup_x |\ddot{m}_j(x_j) - m_j(x_j)| &= O_P(\delta_n) \quad \textit{for } j = 1, \ldots, d. \end{aligned}$$

Then, we have for $j = 1, \ldots, d$,

$$\sup_x |\widetilde{m}_j(x) - m_j(x)| = O_P(\delta_n).$$

4 Estimation with Local polynomials

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

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

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

with the scalar $\widehat{V}_{0,0}(x) = n^{-1} \sum_{i=1}^{n} \prod_{\ell=1}^{d} K_h(X_\ell^i - x_\ell)$, and $\widehat{\mathbf{V}}_{-0,0}(x), \widehat{\mathbf{V}}_{-0,-0}(x)$ defined appropriately. This approach has two disadvantages. First, it may work only in low dimensions – since for the asymptotics, existence of the matrix $\widehat{\mathbf{V}}_{-0,-0}^{-1}(x)$ and convergence of $\widehat{\mathbf{V}}_{-0,-0}(x)$ is required under our assumptions [and this may hold only for low dimensional argument x]. Second, the corresponding backfitting algorithm does not consist in 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 that $\widehat{\mathbf{V}}(x)$ does exist, but we will see that the definition of the backfitting estimate is based on only one- and two-dimensional 'marginals' of $\widehat{\mathbf{V}}(x)$. 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 $\widehat{p}$.

For functions $f = (f^0, \ldots, f^d)$ with components $f^j : \mathbb{R}^d \to \mathbb{R}$ and $d+1$ by $d+1$ positive definite matrix function $M(\cdot)$, define the 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 $\mathcal{F}_{full}$. Furthermore, taking $M = \widehat{\mathbf{V}}$ the norm $\|\cdot\|_M$ is simply the norm induced by the norm $\|\cdot\|_*$. In Section 2 our version $\widetilde{\mathbf{m}}(x) = (\widetilde{m}^0(x), \ldots, \widetilde{m}^d(x))^T$ of the backfitting estimate was defined as the projection of [the function in $\mathcal{F}_{full}$ corresponding to] $\widehat{\mathbf{m}}$ [see (1)] with respect to $\|\cdot\|_*$ onto the space $\mathcal{F}_{add}$. Therefore, $\widetilde{\mathbf{m}}$ coincides with the $\mathrm{L}_2(\widehat{\mathbf{V}})$ projection, with respect to the norm $\|f\|_{\widehat{\mathbf{V}}}$, of $\widehat{\mathbf{m}}$ onto the subspace $\mathcal{M}_{add}$,

where

$$\begin{aligned}\mathcal{M}_{add} &= \{\mathbf{u}(x) = (u^0(x), \ldots, u^d(x))^T \in \mathcal{M} | \\ & u^0(x) = \mu + u_1(x_1) + \ldots + u_d(x_d), u^\ell(x) = w_\ell(x_\ell) \text{ for } \ell = 1, \ldots, d, \\ & \text{where } u_1, \ldots, u_d \text{ are functions } \mathbb{R} \to \mathbb{R} \text{ with } \int \widehat{\mathbf{V}}^j_{j,0}(x_j) u_j(x_j) dx_j = 0 \text{ for } j = 1, \ldots, d \\ & \text{and where } w_\ell : \ell = 1, \ldots, d \text{ are functions: } \mathbb{R} \to \mathbb{R}\},\end{aligned}$$

where for each j the $(d+1) \times (d+1)$ matrix $\widehat{\mathbf{V}}^j(x_j) = \int \widehat{\mathbf{V}}(x) dx_{-j}$. The class $\mathcal{M}_{add}$ contains functions that are additive in the first component [for $\ell = 0$] and where the other components [for $\ell = 1, \ldots, d$] depend only on a one-dimensional argument. A function f in $\mathcal{M}_{add}$ is specified by a constant μ and $2d$ functions $\mathbb{R} \to \mathbb{R}$. Because f^ℓ, $\ell = 1, \ldots, d$, depend only on one argument, in abuse of notation we write also $f^\ell(x_\ell)$ instead of $f^\ell(x)$. Note that there is a one-to-one correspondence between elements of $\mathcal{M}_{add}$ and $\mathcal{F}_{add}$.

We now discuss how $\widetilde{\mathbf{m}}$ is calculated by backfitting. Note that $\widetilde{\mathbf{m}}$ is defined as minimizer of $\|\widehat{m} - m\|_{\widehat{\mathbf{V}}}$. Recall that this is equivalent to minimize $\|\mathbf{Y} - \mathbf{m}\|^2_*$ over $\mathcal{F}_{add}$. We discuss now minimization of this term with respect to the j-th components $m^j(x_j)$ and $\mu + m_j(x_j)$. Define for each j,

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

and note the obvious fact that

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

Therefore, because 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 $\mu + m_j(x_j)$, for $j = 1, \ldots, d$. After some standard calculations, this leads to:

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

$$
\begin{aligned}
& & & - \sum_{\ell \neq j} \int \widetilde{m}_\ell(x_\ell) \widehat{V}^{\ell,j}_{0,0}(x_\ell, x_j)\, dx_\ell \\
& & & - \sum_{\ell \neq j} \int \widetilde{m}^\ell(x_\ell) \widehat{V}^{\ell,j}_{\ell,0}(x_\ell, x_j)\, dx_\ell \qquad (18) \\
\widetilde{m}_j(x_j) \widehat{V}^j_{j,0}(x_j) + \widetilde{m}^j(x_j) \widehat{V}^j_{j,j}(x_j) &=& \frac{1}{n} \sum_{i=1}^n \frac{X^i_j - x_j}{h} K_h(X^i_j - x_j) Y^i - \hat{\mu} \widehat{V}^j_{j,0}(x_j) \\
& & & - \sum_{\ell \neq j} \int \widetilde{m}_\ell(x_\ell) \widehat{V}^{\ell,j}_{0,j}(x_\ell, x_j)\, dx_\ell \\
& & & - \sum_{\ell \neq j} \int \widetilde{m}^\ell(x_\ell) \widehat{V}^{\ell,j}_{\ell,j}(x_\ell, x_j)\, dx_\ell. \qquad (19)
\end{aligned}
$$

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

$$
\begin{aligned}
\widehat{\mathbf{V}}^r(x_r) &=& \int \widehat{\mathbf{V}}(x)\, dx_{-r} \qquad (20) \\
\widehat{\mathbf{V}}^{r,s}(x_r, x_s) &=& \int \widehat{\mathbf{V}}(x)\, dx_{-(r,s)}. \qquad (21)
\end{aligned}
$$

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

$$
\int \widetilde{m}_j(x_j) \widehat{V}^j_{j,0}(x_j)\, dx_j = 0, \qquad (22)
$$

equations (18) and (19) define $\hat{\mu}$, $\widetilde{m}_j$ and $\widetilde{m}^j$ for given $\mathbf{Y}$ and $[\widetilde{m}_\ell, \widetilde{m}^\ell : \ell \neq j]$.

Equations (18) and (19) can be rewritten as

$$
\begin{aligned}
\widetilde{m}_j(x_j) &=& \bar{m}_j(x_j) + \check{m}_j(x_j) \qquad (23) \\
\widetilde{m}^j(x_j) &=& \bar{m}^j(x_j) + \check{m}^j(x_j), \qquad (24)
\end{aligned}
$$

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

$$
\bar{m}_j(x_j) \widehat{V}^j_{0,0}(x_j) + \bar{m}^j(x_j) \widehat{V}^j_{j,0}(x_j) = \frac{1}{n} \sum_{i=1}^n K_h(X^i_j - x_j) Y^i \qquad (25)
$$

$$\bar{m}_j(x_j)\widehat{V}^j_{j,0}(x_j) + \bar{m}^j(x_j)\widehat{V}^j_{j,j}(x_j) = \frac{1}{n}\sum_{i=1}^{n}\frac{X^i_j - x_j}{h}K_h(X^i_j - x_j)Y^i \tag{26}$$

$$\begin{aligned}\check{m}_j(x_j)\widehat{V}^j_{0,0}(x_j) + \check{m}^j(x_j)\widehat{V}^j_{j,0}(x_j) &= -\hat{\mu}\widehat{V}^j_{0,0}(x_j) - \sum_{\ell\neq j}\int \widetilde{m}_\ell(x_\ell)\widehat{V}^{\ell,j}_{0,0}(x_\ell, x_j)\, dx_\ell \\ &\quad - \sum_{\ell\neq j}\int \widetilde{m}^\ell(x_\ell)\widehat{V}^{\ell,j}_{\ell,0}(x_\ell, x_j)\, dx_\ell \end{aligned} \tag{27}$$

$$\begin{aligned}\check{m}_j(x_j)\widehat{V}^j_{j,0}(x_j) + \check{m}^j(x_j)\widehat{V}^j_{j,j}(x_j) &= -\hat{\mu}\widehat{V}^j_{j,0}(x_j) - \sum_{\ell\neq j}\int \widetilde{m}_\ell(x_\ell)\widehat{V}^{\ell,j}_{0,j}(x_\ell, x_j)\, dx_\ell \\ &\quad - \sum_{\ell\neq j}\int \widetilde{m}^\ell(x_\ell)\widehat{V}^{\ell,j}_{\ell,j}(x_\ell, x_j)\, dx_\ell \end{aligned} \tag{28}$$

$$\int \check{m}_j(x_j)\widehat{V}^j_{j,0}(x_j)\, dx_j = 0, \tag{29}$$

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

Again, together with the norming condition (22), equations (23)-(29) define $\hat{\mu}$, $\widetilde{m}_j$ and $\widetilde{m}^j$ for given $\mathbf{Y}$ and $[\widetilde{m}_\ell, \widetilde{m}^\ell : \ell \neq j]$. In the j-th step of every cycle of the backfitting algorithm an update of $\hat{\mu}$, $\widetilde{m}_j$ and $\widetilde{m}^j$ will be calculated by solving equations (23)-(29). In the next subsection we will discuss asymptotics for the backfitting estimate in a more general set up. In particular, there we will not assume that $(\bar{m}_\ell, \bar{m}^\ell)$ is a one-dimensional local linear fit nor that $\widehat{\mathbf{V}}^\ell$ and $\widehat{\mathbf{V}}^{\ell,\ell'}$ are motivated by local linear smoothing. Furthermore, we will not make any assumptions on the stochastic nature of the sample. For arbitrary choices of $(\bar{m}_\ell, \bar{m}^\ell)$ is a one-dimensional local linear fit nor that $\widehat{\mathbf{V}}^\ell$ and $\widehat{\mathbf{V}}^{\ell,\ell'}$ we will assume that $\widetilde{m}_j$ and $\widetilde{m}^j$ are defined by

$$\begin{aligned}\hat{\mathbf{M}}_j(x_j)\begin{pmatrix}\{\widetilde{m}_j - \bar{m}_j\}(x_j)\\ \{\widetilde{m}^j - \bar{m}^j\}(x_j)\end{pmatrix} &= -\hat{\mu}\begin{pmatrix}\widehat{\mathbf{V}}^j_{0,0}(x_j)\\ \widehat{\mathbf{V}}^j_{j,0}(x_j)\end{pmatrix} \\ &\quad - \sum_{\ell\neq j}\int \hat{\mathbf{S}}_{\ell,j}(x_\ell, x_j)\begin{pmatrix}\widetilde{m}_\ell(x_\ell)\\ \widetilde{m}^\ell(x_\ell)\end{pmatrix} dx_\ell.\end{aligned} \tag{30}$$

$$\int \widetilde{m}_j(x_j)\widehat{V}^j_{j,0}(x_j)\, dx_j = 0. \tag{31}$$

where

$$\hat{\mathbf{M}}_j(x_j) = \begin{pmatrix} \widehat{V}^j_{0,0}(x_j) & \widehat{V}^j_{j,0}(x_j) \\ \widehat{V}^j_{j,0}(x_j) & \widehat{V}^j_{j,j}(x_j) \end{pmatrix}, \tag{32}$$

$$\hat{\mathbf{S}}_{\ell,j}(x_\ell, x_j) = \begin{pmatrix} \widehat{V}^{\ell,j}_{0,0}(x_\ell, x_j) & \widehat{V}^{\ell,j}_{\ell,0}(x_\ell, x_j) \\ \widehat{V}^{\ell,j}_{j,0}(x_\ell, x_j) & \widehat{V}^{\ell,j}_{\ell,j} x_\ell, x_j) \end{pmatrix}. \tag{33}$$

Let us finish this section by some computational remarks.

- In a faster implementation, the norming of $\widetilde{m}_j$ done in (22) could be omitted and one could put always $\hat{\mu} = 0$. After the final cycle all functions $\widetilde{m}_j$ could be replaced by $\widetilde{m}_j(x_j) - \int \widetilde{m}_j(x_j) V^j_{j,0}(x_j)\, dx_j$ and $\hat{\mu}$ defined appropriately. It is easy to see that this algorithm does the same. If one is interested only in the estimation of the sum $\mu + m_1(x_1) + \ldots + m_d(x_d)$ the final norming could be omitted or replaced by another norming.

- A possible initialization of backfitting is given by putting $\hat{\mu} = 0$, $\widetilde{m}_\ell = \bar{m}_\ell$ and $\widetilde{m}^\ell = \bar{m}^\ell$ for $\ell = 1, \ldots, d$.

- Note that the estimates $\bar{m}_\ell$ and $\bar{m}^\ell$ have to be calculated only at the beginning and have not to be updated in each backfitting iteration.

- For an implementation of backfitting, all estimates [i.e., $\bar{m}_\ell$, $\bar{m}^\ell$, $\check{m}_\ell$, $\check{m}^\ell$, $\widetilde{m}_\ell$, $\widetilde{m}^\ell$, $\widehat{\mathbf{V}}^\ell$ and $\widehat{\mathbf{V}}^{\ell,\ell'}$] have to be calculated on a grid and the integrals in (27) and (28) 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 contains the same number of points.

4.1 Asymptotics for Local Polynomials

We discuss now asymptotics for the backfitting estimate. This will be done in a general set up. We assume that some estimates $\bar{m}_\ell$, $\bar{m}^\ell$, $\widehat{\mathbf{V}}^\ell$ and $\widehat{\mathbf{V}}^{\ell,\ell'}$ $[\ell,\ell' = 1,\ldots,d]$ are given and that $\hat{\mu}$, $\widetilde{m}_\ell$ and $\widetilde{m}^\ell$ $[\ell = 1,\ldots,d]$ are defined by (30) - (33). In particular, we will not assume that $(\bar{m}_\ell, \bar{m}^\ell)$ is a one dimensional local linear fit and that $\widehat{\mathbf{V}}^\ell$ and $\widehat{\mathbf{V}}^{\ell,\ell'}$ 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 $\mathbf{V}$. We define $\hat{\mathbf{M}}_j(x_j)$ and $\hat{\mathbf{S}}_{\ell,j}(x_\ell, x_j)$as in (32) and (33) and we put

$$\mathbf{M}_j = \begin{pmatrix} V_{0,0} & V_{j,0} \\ V_{j,0} & V_{j,j} \end{pmatrix},$$

$$\mathbf{S}_{\ell,j} = \begin{pmatrix} V_{0,0} & V_{\ell,0} \\ V_{j,0} & V_{\ell,j} \end{pmatrix}.$$

(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*

$$\int \left[\hat{\mathbf{M}}_j(x_j)^{-1}\hat{\mathbf{S}}_{\ell,j}(x_\ell, x_j) - \mathbf{M}_j^{-1}\mathbf{S}_{\ell,j}\frac{p_{j,\ell}(x_j, x_\ell)}{p_j(x_j)}\right]_{r,s}^2 p_\ell(x_\ell)^{-1} p_j(x_j) dx_j dx_\ell = o_P(1).$$

for $r,s = 1,2$. Here $[\ldots]_{r,s}$ denotes the (r,s) element of a matrix $[\ldots]$.

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

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

and

$$\int \bar{m}^j(x_j)^2 p_j(x_j) dx_j \leq C.$$

Furthermore $\int \bar{m}_j(x_j)^2 \widehat{V}^j_{j,0}(x_j)\, dx_j$ does not depend on j and it is equal to $\hat{\mu}$.

(A4′) *There exists a constant C such that with probability tending to one for all $j \neq k$*

$$\sup_{x_j} \int \mathrm{trace}[\hat{\mathbf{M}}_j(x_j)^{-1} \hat{\mathbf{S}}_{\ell,j}(x_\ell, x_j) \hat{\mathbf{M}}_\ell(x_\ell)^{-1} \hat{\mathbf{S}}_{\ell,j}(x_\ell, x_j) \hat{\mathbf{M}}_j(x_j)^{-1}] dx_\ell \leq C.$$

(A5′) *We suppose that for a sequence Δ_n the smoothers $\bar{m}_j$ and $\bar{m}^j$ can be decomposed as $\bar{m}_j = \bar{m}^A_j + \bar{m}^B_j$ and $\bar{m}^j = \bar{m}^{j,A} + \bar{m}^{j,B}$, where the first components $\bar{m}^A_j$ and $\bar{m}^{j,A}$ are mean zero and satisfy*

$$\sup_{x_k} \left| \int \hat{\mathbf{M}}_j(x_j)^{-1} \hat{\mathbf{S}}_{\ell,j}(x_\ell, x_j) \begin{pmatrix} \bar{m}^A_j(x_j) \\ \bar{m}^{j,A}(x_j) \end{pmatrix} dx_j \right| = o_P\left(\frac{\Delta_n}{\log n}\right).$$

For $s = A$ and $s = B$ we define $\hat{\mu}^s$, $\widetilde{m}^s_j$ and $\widetilde{m}^{j,s}$ as the solution of the following equations

$$\hat{\mathbf{M}}_j(x_j) \begin{pmatrix} \{\widetilde{m}^s_j - \bar{m}^s_j\}(x_j) \\ \{\widetilde{m}^{j,s} - \bar{m}^{j,s}\}(x_j) \end{pmatrix} = -\hat{\mu} \begin{pmatrix} \widehat{\mathbf{V}}^j_{0,0}(x_j) \\ \widehat{\mathbf{V}}^j_{j,0}(x_j) \end{pmatrix} - \sum_{\ell \neq j} \int \hat{\mathbf{S}}_{\ell,j}(x_\ell, x_j) \begin{pmatrix} \widetilde{m}^s_\ell(x_\ell) \\ \widetilde{m}^{\ell,s}(x_\ell) \end{pmatrix} dx_\ell. \tag{34}$$

$$\int \widetilde{m}^s_j(x_j) \widehat{V}^j_{j,0}(x_j)\, dx_j = 0. \tag{35}$$

Existence and uniqueness of $\widetilde{m}^A_j$,$\widetilde{m}^B_j$,$\widetilde{m}^{j,A}$ and $\widetilde{m}^{j,B}$ is stated in the next theorem. Note that $(\widetilde{m}^s_j, \widetilde{m}^{j,s})$ is defined as $(\widetilde{m}_j, \widetilde{m}^j)$ in equations (30) and (31) with $(\bar{m}_j, \bar{m}^j)$ replaced by $(\bar{m}^s_j, \bar{m}^{j,s})$. We suppose that for (deterministic) functions $\mu_{j,n}(\cdot), \mu^{j,B}_n(\cdot)$ the term $\widetilde{m}^B_j$ satisfies:

$$\widetilde{m}^B_j(x_j) = \mu_{j,n}(x_j) + o_P(\Delta_n)$$

$$\widetilde{m}^{j,B}(x_j) = \mu_n^j(x_j) + o_P(\Delta_n).$$

We remark again that these conditions are all straightforward to verify, except perhaps A5′. Note that we shall not require $\widehat{\mathbf{V}}(x)$ to converge in probability to $\mathbf{V}(x)$ because this would depend on the curse of dimensionality.

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′-A2′ hold. Then, with probability tending to one, there exists a solution [$\hat{\mu}$, $\widetilde{m}_\ell$, $\widetilde{m}^\ell : \ell = 1, \ldots, d$] of (30) - (33) that is unique. Furthermore, there exist constants $0 < \gamma < 1$ and $c > 0$ such that, with probability tending to one, the following inequality holds*

$$\begin{aligned} \int \left[\widetilde{m}_j^{[r]}(x_j) - \widetilde{m}_j(x_j)\right]^2 p_j(x_j)dx_j &\leq c\gamma^{2r}\Gamma, \\ \int \left[\widetilde{m}^{j,[r]}(x_j) - \widetilde{m}^j(x_j)\right]^2 p_j(x_j)dx_j &\leq c\gamma^{2r}\Gamma, \end{aligned}$$

where

$$\Gamma = [\hat{\mu}^{[0]}]^2 + \sum_{\ell=1}^{d} \int \left[\widetilde{m}_\ell^{[0]}(x_\ell)\right]^2 p_\ell(x_\ell)dx_\ell + \int \left[\widetilde{m}^{\ell,[0]}(x_\ell)\right]^2 p_\ell(x_\ell)dx_\ell.$$

Here, for $r = 0$ the functions $\hat{\mu}^{[0]}$, $\widetilde{m}_\ell^{[0]}$ and $\widetilde{m}^{\ell,[0]}$ are the starting values of the backfitting algorithm.

Furthermore for $s = A$ and $s = B$, with probability tending to one, there exists a solution [$\hat{\mu}^s$, $\widetilde{m}_j^s$ and $\widetilde{m}^{j,s} : j = 1, \ldots, d$] of (34) - (35) that is unique.

Just as Theorem 2 stated for Nadaraya-Watson smoothing, the stochastic part of the backfitting estimate coincides again with a one-dimensional local linear fit. This is stated in the following theorem:

THEOREM 2′. *Suppose that conditions A1′ - A5′ hold for a sequence* Δ_n. *Then it holds that*

$$\sup_{x_j} \left| \widetilde{m}_j^A(x_j) - \widehat{m}_j^A(x_j) \right| = o_P(\Delta_n).$$

In particular, one gets

$$\widetilde{m}_j(x_j) = \widehat{m}_j^A(x_j) + \mu_{j,n}(x_j) + o_P(\Delta_n).$$

We suppose now that a full dimensional estimate $\widehat{\mathbf{m}}$ as described at the beginning of this section exists. Under conditions analogous to (14)-(17) we get the following corollary.

COROLLARY 1′. *Suppose that conditions A1′-A5′ hold with* $\Delta_n = n^{-2/5}$ *and that conditions of type (14)-(17) apply for* $\widehat{\mathbf{m}}$. *Then,*

$$n^{2/5} \begin{bmatrix} \widetilde{m}_1(x_1) - m_1(x_1) \\ \vdots \\ \widetilde{m}_d(x_d) - m_d(x_d) \end{bmatrix} \Longrightarrow N \left(\begin{bmatrix} b_1(x_1) \\ \vdots \\ b_d(x_d) \end{bmatrix}, \begin{bmatrix} v_1(x_1) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & v_d(x_d) \end{bmatrix} \right),$$

where $v_j(x_j) = \ddot{v}_j(x_j)$, $j = 1, \ldots, d$, *are the variances of the infeasible 'oracle estimate'* $\ddot{m}_j(x_j)$ *[defined similarly as for Nadaraya-Watson smoothing in the last section], while* $b_j(x_j)$ *are solutions to the following minimization problem*

$$\min_{\mu, b_1(\cdot), \ldots, b_d(\cdot)} \int \left[b(x) - \mu - b_1(x_1) - \ldots - b_d(x_d) \right]^2 p(x) dx, \quad \text{s.t.} \int b_j(x_j) p_j(x_j) dx_j = 0, \; j = 1, \ldots, d.$$

For the case that the function b *is already of additive form* $b(x) = b_1^*(x_1) + \ldots + b_d^*(x_d)$ *the bias functions* b_j *coincide with the bias* $\ddot{b}_j^c(x_j)$ *of the 'corrected' oracle estimate* $\ddot{m}_j^c(x_j)$. *Also*

$$n^{2/5} \left\{ \widetilde{m}(x) - m(x) \right\} \Longrightarrow N \left[b_+(x), v_+(x) \right],$$

where $b_+(x) = \sum_j b_j(x_j)$ *and* $v_+(x) = \sum_j v_j(x_j)$.

Suppose additionally that for a sequence δ_n *with* $n^{-2/5} = o(\delta_n)$

$$\begin{aligned} \sup_x |\widehat{m}^B(x) - m(x) - n^{-2/5} b(x)| &= O_P(\delta_n) \\ \sup_x |\ddot{m}_j(x_j) - m_j(x_j)| &= O_P(\delta_n) \quad \textit{for } j = 1, \ldots, d. \end{aligned}$$

Then, we have for $j = 1, \ldots, d$,

$$\sup_x |\widetilde{m}_j(x) - m_j(x)| = O_P(\delta_n).$$

REMARK. For example, when the data are independent and identically distributed (i.i.d.) and the smoother is the local linear, then the bias b is of additive form $b(x) = b_1^*(x_1) + \ldots + b_d^*(x_d)$. Then the bias functions b_j coincide with the bias $\ddot{b}_j^c(x_j)$ of the 'corrected' oracle estimate $\ddot{m}_j^c(x_j)$:

$$b_j(x_j) = \lim_{n\to\infty} \left\{h^2 n^{2/5}\right\} \frac{\mu_2(K)}{2} \left\{ m_j''(x_j) - \int m_j''(x_j) p_j(x_j) dx_j \right\}$$

for $j = 1, \ldots, d$, where $\mu_2(K) = \int t^2 K(t) dt$. In this case, the asymptotic bias and the asymptotic variance are identical to bias and variance of the mean corrected 'oracle' estimator [based also on local linear estimation]. That means our estimate achieves the same first order asymptotics as if the other components would be 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.

5 Verification of Conditions

We now provide sufficient conditions for A1-A5 to hold in a time series setting for the Nadaraya-Watson smoother. We suppose that $\{Y^i, Z^i\}_{i=-\infty}^{\infty}$ is a jointly stationary process on the real line

and let $X_i = (Z^i, \ldots, Z^{i-d+1})'$. In this case, $m(\cdot)$ is the d'th order autoregressive mean. Let $\mathcal{F}_a^b$ be the σ-algebra of events generated by the random variables $\{Y^i, Z^i; \quad a \leq j \leq b\}$. The stationary processes $\{Y^i, Z^i\}$ are called strongly mixing [Rosenblatt (1956)] if

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

We assume

(B1) *The kernel K is bounded, has compact support, is symmetric about zero, and is Lipschitz continuous, i.e., there exists a positive finite constant C such that $|K(u) - K(v)| \leq C|u - v|$.*

(B2) *The density q_0 $[= p]$ of X_0 and the densities $q_{0,\ell}$ of (X_0, X_ℓ), $\ell = 1, \ldots,$ are bounded away from zero and infinity on their compact support.*

(B3) *For some $\theta > 2$, $E(|Y|^\theta) < \infty$.*

(B4) *The conditional densities $f_{X_0|Y_0}(x_0|y_0)$ of X_0 given Y_0 and $f_{X_0,X_\ell|Y_0,Y_\ell}(x_0, x_\ell|y_0, y_\ell)$ of (X_0, X_ℓ) given (Y_0, Y_ℓ), $\ell = 1, \ldots,$ exist and are bounded from above.*

(B5) *The processes $\{Y_j, Z_j\}$ are strong mixing with $\sum_{j=1}^\infty j^a \{\alpha(j)\}^{1-2/\nu} < \infty$ for some $\nu > 2$ and $a > 1 - 2/\nu$.*

(B6) *The strong mixing coefficients satisfy $\sum_{j=1}^\infty \varphi(j;\theta) < \infty$ and $\sum_{j=1}^\infty \psi(j;c) < \infty$ for $c = 1, 2$, where: $\varphi(n;\theta) = (nL_1(n)/r_1(n))(nT_n^2/\log n)^{1/4} \alpha\{r_1(n)\}$ with $r_1(n) = (n/T_n \log n)^{1/2}$ and $L_1(n) = (nT_n^2/\log n)^{1/2}$ with $T_n = \left\{n \log n (\log\log n)^{1+\delta}\right\}^{1/\theta}$ for some $\delta > 0$, 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 $L_2(n) = (n/h^{c+2} \log n)^{c/2}$.*

(B7) *The functions m and p are twice continuously differentiable.*

We apply some results of Masry (1996) to establish the conditions A1-A5.

THEOREM 3. *Suppose that conditions B1-B7 hold. Then conditions A1-A5 hold.*

6 Illustration

We applied our backfitting method to the estimation of nonparametric AR-ARCH models on stock return data. Specifically, we fit the following model

$$Y^i = \mu + m_1(Y^{i-1}) + m_2(Y^{i-2}) + \sigma^i \varepsilon^i \quad ; \quad \sigma^{2i} = \lambda + v_1(Y^{i-1}) + v_2(Y^{i-2}),$$

first applying our method to the raw data to obtain estimates $\widehat{\mu}, \widetilde{m}_1(\cdot)$, and $\widetilde{m}_2(\cdot)$, and then applying it to the squared residuals $\{Y^i - \widehat{\mu} - \widetilde{m}_1(Y^{i-1}) - \widetilde{m}_2(Y^{i-2})\}^2$ to obtain estimates $\widehat{\lambda}, \widetilde{v}_1(\cdot)$, and $\widetilde{v}_2(\cdot)$. As in Härdle and Yiang (1996), we expect estimation of the mean not to affect the estimation of the variance and we have computed standard errors accordingly.

Our data is monthly return on the S&P500 index whose stocks were traded on the New York Stock Exchange between 1946-1986. The results are shown below

*** FIGURES HERE***

The confidence intervals reveal that the mean effect is not well determined, as is to be expected, but that the variance effects are highly significant. The asymmetry in $v_1(\cdot)$ has been found before in stock returns, see Bollerslev, Engle, and Nelson (1994, pp. 3028-3029), and is possibly explained by the 'leverage effect' – the limited liability of public companies makes downturns more risky than upturns for investors.

A Appendix: Proofs

Before we come to the proofs of our results let us collect some facts about iterative projections. Let us define the following spaces of additive functions

$$\mathcal{H} = \{m \in \mathbf{L}_2(p) : m(x) = m_1(x_1) + \ldots + m_d(x_d) \ (p \text{ a.s.}), \int m(x)p(x)dx = 0\},$$

$$\mathcal{H}_j = \{m \in \mathcal{H} : m(x) = m_j(x_j) \ (p \text{ a.s.}) \text{ for a function } m_j \in \mathbf{L}_2(p_j)\}.$$

The norm in the space $\mathcal{H}$ is denoted by $\|m\|^2 = \int m^2(x)p(x)dx$ for $m \in \mathcal{H}$. For $m \in \mathcal{H}_j$ we get with $m_j(x_j) = m(x)$ (p a.s.) that $\|m\|^2 = \int m^2(x)p(x)dx = \int m_j^2(x_j)p_j(x_j)dx_j$. The projection of an element of $\mathcal{H}$ onto $\mathcal{H}_j$ is denoted by Π_j. The operator $\Psi_j = I - \Pi_j$ gives the projection onto the linear space

$$\begin{aligned} \mathcal{H}_j^{\perp} &= \{m \in \mathcal{H} : \int m(x)\phi(x_j)p(x)dx = 0 \text{ for all } \phi \in \mathcal{H}_j\} \\ &= \{m \in \mathcal{H} : \int m(x)p(x)dx_{-j} = 0 \ (p_j \text{ a.s.})\}. \end{aligned}$$

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

$$\Psi_j m(x) = m_1(x_1) + \ldots + m_{j-1}(x_{j-1}) + m_j^*(x_j) + m_{j+1}(x_{j+1}) + \ldots + m_d(x_d) \tag{36}$$

with

$$m_j^*(x_j) = -\sum_{k \neq j} \int m_k(x_k) \frac{p_{jk}(x_j, x_k)}{p_j(x_j)} dx_k. \tag{37}$$

We define the operator $\widehat{\Psi}_j$ as Ψ_j but with $m_j^*(x_j)$ on the right hand side of (36) replaced by

$$\widehat{m}_j^*(x_j) = -\sum_{k \neq j} \int m_k(x_k) \frac{\widehat{p}_{jk}(x_j, x_k)}{\widehat{p}_j(x_j)} dx_k. \tag{38}$$

Put $T = \Psi_d \cdots \Psi_1$ and $\widehat{T} = \widehat{\Psi}_d \cdots \widehat{\Psi}_1$. We will see below that in our set up the backfitting algorithm is based on iterative applications of $\widehat{T}$. A central tool for understanding backfitting will be given by the next lemma that describes iterative applications of T.

LEMMA [NORM OF THE OPERATOR T]. *Suppose that condition A1 holds. Then* $T : \boldsymbol{L}_2(p) \to \boldsymbol{L}_2(p)$ *is a positive self adjoint operator with operator norm* $\tau = \sup\{\|Tf\| : \|f\| \leq 1\} < 1$. *Hence, for every* $m \in \mathcal{H}$ *we get*

$$\|T^r m\| \leq \tau^r \|m\|. \tag{39}$$

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

$$\|m\| \geq c \max\{\|m_1\|, \ldots, \|m_d\|\}. \tag{40}$$

PROOF OF LEMMA. We start by proving (39). It is known that (39) holds with $\tau^2 \leq 1 - \prod_{j=1}^{d} \sin^2(\tau_j)$ where $\cos \tau_j = \rho(\mathcal{H}_j, \mathcal{H}_{j+1} + \ldots + \mathcal{H}_r)$ 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, i.e., $\rho(L_1, L_2) = \sup\{\int h_1(x)h_2(x)\, p(x)dx : h_j \in L_j \cap (L_1 \cap L_2)^{\perp}, \|h_j\| \leq 1 (j = 1, 2\}$. 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 $\mathcal{M}_j = \mathcal{H}_1 + \ldots + \mathcal{H}_j$ are closed subsets of $\mathrm{L}_2(p)$. This implies that $\rho(\mathcal{H}_{j+1}, \mathcal{M}_j) < 1$ 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 $\mathcal{M}_j$ is closed we will use the following two facts. For two closed subspaces L_1 and L_2 of $\mathrm{L}_2(p)$ 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

$$\|m\| \geq c \max[\|m_1\|, \|m_2\|]. \tag{41}$$

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 $j \leq j_o - 1$ that $\mathcal{M}_j$ is closed and that we want to show that $\mathcal{M}_{j_o}$ is closed. As mentioned above, for this claim it suffices to show that $\Pi_{j_o}|\mathcal{M}_{j_o-1}$ is compact. We remark first that (41) implies that for every $m \in \mathcal{M}_{j_o-1}$ there exist $m_j \in \mathcal{H}_j$ $(j \leq j_o - 1)$ such that $m(u) = m_1(u_1) + \ldots + m_{j_o-1}(u_{j_o-1})$ (p a.s.) and with a constant $c > 0$

$$\|m\| \geq c \max[\|m_1\|, \ldots, \|m_{j_o-1}\|]. \tag{42}$$

We will prove that

$$\|\Pi_{j_o} m\|^2 \leq const. \left[\sum_{j=1}^{j_o-1} \int R^2_{j,j_o}(x_j, x_{j_o}) p_j(x_j) p_{j_0}(x_{j_0}) dx_j dx_{j_0}\right] \|m\|^2 \tag{43}$$

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 (43) implies compactness of $\Pi_{j_o}|\mathcal{M}_{j_o-1}$. To see this one argues as in the standard proofs for compactness of Hilbert-Schmidt operators, see e.g., Example 3.2.4 in Balakrishnan (1981).

It remains to show (43). This follows from (42) with applications of the Cauchy-Schwarz inequality.

Equation (40) follows as (42). ■

PROOF OF THEOREM 1. The following lemma establishes the result.

LEMMA [NORM OF THE OPERATOR $\widehat{T}$]. *Suppose that conditions A1-A2 hold. Choose γ with $\tau < \gamma < 1$. Then, with probability tending to one, the operator norm* $\sup\left\{\|\widehat{T} f(x)]\| : \|f\| \leq 1\right\}$ *is bounded by γ.*

PROOF OF LEMMA. We remark first that the distance between m_j^* and $\widehat{m}_j^*$, see (37)-(38) can be bounded as follows.

$$\|\widehat{m}_j^* - m_j^*\| \leq \sum_{k \neq j} \|m_k\| S_{jk},$$

with

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

With $S_j = \max_{k \neq j} |S_{jk}|$ this and equation (40) imply

$$\|\widehat{m}_j^* - m_j^*\| \leq \frac{d}{c} \|m\| S_j.$$

Now because of (A2), $S_j = o_P(1)$. This gives $\|\widehat{\Psi}_j - \Psi_j\| = o_P(1)$. Now the statement of the lemma follows from

$$\|\widehat{T} - T\| = o_P(1).$$

■

LEMMA [STOCHASTIC EXPANSION OF $\widetilde{m}$]. *Suppose that conditions A1-A2 hold. Then there exist constants $0 < \gamma < 1$ and $C > 0$ such that with probability tending to one, for $\widetilde{m}$ the following stochastic expansion holds for $s \geq 1$:*

$$\widetilde{m}(x) = \sum_{r=0}^{s} \widehat{T}_1^r \widehat{m}_1(x) + \ldots + \sum_{r=0}^{s} \widehat{T}_d^r \widehat{m}_d(x) + R^{[s]}(x),$$

where $\widehat{T}_j = \widehat{\Psi}_j \widehat{\Psi}_{j-1} \cdots \widehat{\Psi}_1 \widehat{\Psi}_d \widehat{\Psi}_{d-1} \cdots \widehat{\Psi}_{j+1}$ and $R^{[s]}(x) = R_1^{[s]}(x_1) + \ldots + R_d^{[s]}(x_d)$ is a function in $\mathcal{H}$ with

$$\|R_j^{[s]}\| \leq C \gamma^s. \tag{44}$$

Under the additional assumption of (A3) it holds that

$$\sup_{x_j} |R_j^{[s]}(x_j)| \leq C \gamma^s. \tag{45}$$

PROOF OF LEMMA. We remark first that (11) can be rewritten as

$$\widetilde{m}(x) = \widehat{\Psi}_j \widetilde{m}(x) + \widehat{m}_j(x_j).$$

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

$$\widetilde{m}(x) = \widehat{T}\widetilde{m}(x) + \widehat{\delta}(x),$$

where

$$\widehat{\delta}(x) = \widehat{\Psi}_d \cdots \widehat{\Psi}_1 \widehat{m}_1(x) + \ldots + \widehat{\Psi}_d \widehat{m}_{d-1}(x) + \widehat{m}_d(x_d).$$

With the last equality we get the following expansion

$$\widetilde{m}(x) = \sum_{r=0}^{\infty} \widehat{T}^r \widehat{\delta}(x).$$

Plugging the definition of $\widehat{\delta}$ into this equation gives

$$\widetilde{m}(x) = \sum_{r=0}^{\infty} \widehat{T}_1^r \widehat{m}_1(x) + \ldots + \sum_{r=0}^{\infty} \widehat{T}_d^r \widehat{m}_d(x).$$

The operator norms of $\widehat{T}_1, \ldots, \widehat{T}_d$ are smaller than γ, with probability tending to one, for $\gamma < 1$ large enough. This follows from 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 one, $\|R_j^{[s]}\| \leq C_1 \gamma^s$. This implies claim (44) because of (40). Assume now (A4). For the proof of (45) note that for $C_2 > 0$ large enough with probability tending to one for all functions f, g in $\mathcal{H}_j$ with $\sup_{x_j} |f(x_j)| \leq 1$ and $\|g\| \leq 1$ it holds for $k \neq j$ that

$$\left| \int \frac{\widehat{p}_{jk}(x_j, x_k)}{\widehat{p}_k(x_k)} f(x_j) dx_j \right| \leq C_2, \tag{46}$$

$$\left| \int \frac{\widehat{p}_{jk}(x_j, x_k)}{\widehat{p}_k(x_k)} g(x_j) dx_j \right| \leq C_2. \tag{47}$$

Equation (47) follows from assumption (A4) by application of the Cauchy Schwarz inequality. Equations (46) and (47) imply that for $C_3 > 0$ large enough with probability tending to one for all functions h in $\mathcal{H}$ with $\|h\| \leq 1$ it holds that

$$\|\widehat{T}h\| \leq C_3. \tag{48}$$

Claim (45) can be shown by using (44) and (48). ■

PROOF OF THEOREM 2. The following lemma establishes the result.

LEMMA [BEHAVIOUR OF THE STOCHASTIC COMPONENT OF $\widetilde{m}$]. *Suppose (A1) - (A5). Then we have that*

$$\sup_{x_j} |\widetilde{m}_j^A(x_j) - \widehat{m}_j^A(x_j)| = O_P(\log n \ \Delta_n). \tag{49}$$

PROOF OF LEMMA. Proceeding as in the last lemma we get with $s = C^* \log n$ (where C^* is chosen large enough)

$$\widetilde{m}^A(x) - \widehat{m}^A(x) = \sum_{r=1}^{s} \widehat{T}_1^r \widehat{m}_1^A(x) + \ldots + \sum_{r=1}^{s} \widehat{T}_d^r \widehat{m}_d^A(x) + R^{[s]}(x),$$

where $R^{[s]}(x) = R_1^{[s]}(x_1) + \ldots + R_d^{[s]}(x_d)$ is a function in $\mathcal{H}$ with

$$\sup_{x_j} |R_j^{[s]}(x_j| \leq \Delta_n.$$

It remains to show

$$\sup_{x} \|\widehat{T}_1^r \widehat{m}_1^A(x)\| = O_P(\Delta_n).$$

This follows from assumption (A5) by arguments as in the proof of the last lemma. ■

PROOFS OF THEOREMS 1′ AND 2′. The theorems follow as Theorems 1 and 2 by essentially the same arguments. In particular, instead of $\mathrm{L}_2(p)$ we consider now $\mathrm{L}_2(Vp) = \{f = (f^0, \ldots, f^d) : f^j :$

$\mathbb{R}^d \to \mathbb{R}$ with $\int f^T(x) V f(x) p(x)\, dx < \infty\}$. Furthermore, now the spaces $\mathcal{H}$ and $\mathcal{H}_j$ are defined as

$$\begin{aligned}
\mathcal{H} &= \{m \in \mathbf{L}_2(Vp) : m^0(x) = m_1(x_1) + \ldots + m_d(x_d) \ (p \text{ a.s.}) \text{ for functions } m_1 \in \mathbf{L}_2(p_1), \ldots, \\
&\qquad m_d \in \mathbf{L}_2(p_d), \ \int m^0(x) p(x) dx = 0 \text{ and for } j = 1, \ldots, d \text{ the functions } m^j \text{ depend only} \\
&\qquad \text{on } x_j\}, \\
\mathcal{H}_j &= \{m \in \mathcal{H} : m^0(x) = m_j(x_j) \ (p \text{ a.s.}) \text{ for a function } m_j \in \mathbf{L}_2(p_j) \\
&\qquad \text{and for } \ell \not= j \text{ it holds that } m^\ell(x) \equiv 0\}.
\end{aligned}$$

Note that again every function f in $\mathcal{H}$ is a sum of functions in $\mathcal{H}_j$: there exist functions $f_j : \mathbb{R} \to \mathbb{R}^2$ with

$$x \ \to \ \left(\begin{array}{cc} e_0 & e_j \end{array} \right) f_j(x_j) \text{ is a function in } \mathcal{H}_j,$$

$$f(x) = \sum_{j=1}^{d} \left(\begin{array}{cc} e_0 & e_j \end{array} \right) f_j(x_j).$$

Here for $j = 0, \ldots, d$ the vector e_j denotes the $(j+1)$st eigenvector of $\mathbb{R}^{d+1}$. The operators Ψ_j is now defined as in (36) with

$$m_j^*(x_j) = -\sum_{k \not= j} \int \mathbf{M}_j^{-1} \mathbf{S}_{j,k} \frac{p_{jk}(x_j, x_k)}{p_j(x_j)} m_k(x_k) dx_k.$$

Furthermore, we define the operator $\widehat{\Psi}_j$ now as Ψ_j but with $m_j^*(x_j)$ on the right hand side of (36) replaced by

$$m_j^*(x_j) = -\sum_{k \not= j} \int \hat{\mathbf{M}}_j^{-1}(x_j) \hat{\mathbf{S}}_{j,k}(x_j, x_k) m_k(x_k) dx_k.$$

Proceeding as above one can show that the norm of the operators $T = \Psi_d \cdots \Psi_1$ and $\widehat{T} = \widehat{\Psi}_d \cdots \widehat{\Psi}_1$ is smaller than $\gamma < 1$ [with probability tending to one]. Theorems $1'$ and $2'$ follow by stochastic expansions of $\widetilde{\mathbf{m}}$, compare the last two lemmas. ■

PROOF OF THEOREM 3. Let $\|g\|_\infty = \sup_x |g(x)|$. Then, under these conditions,

$$\left\|\widehat{p}_{j,k}-E(\widehat{p}_{j,k})\right\|_\infty = O\left\{\left(\frac{\log n}{nh^2}\right)^{1/2}\right\} \quad a.s. \quad ; \quad \left\|\widehat{p}_j-E(\widehat{p}_j)\right\|_\infty = O\left\{\left(\frac{\log n}{nh}\right)^{1/2}\right\} \quad a.s \tag{50}$$

$$\left\|E(\widehat{p}_{j,k})-p_{j,k}\right\|_\infty = O(h) \quad a.s. \quad ; \quad \left\|E(\widehat{p}_j)-p_j\right\|_\infty = O(h) \quad a.s. \tag{51}$$

$$\left\|\widehat{m}_j^A\right\|_\infty = O\left\{\left(\frac{\log n}{nh}\right)^{1/2}\right\} \quad a.s. \quad ; \quad \left\|\widehat{m}_j^B-m_j\right\|_\infty = O(h^2) \quad a.s., \tag{52}$$

see Masry (1996) for a proof. Since $\inf_{x_j} p_j(x_j) > 0$ for all j and $\sup p_{j,k}(x_j,x_k) < \infty$, A1 is satisfied. Furthermore, since

$$\inf_{x_j} \widehat{p}_j(x_j) \geq \inf_{x_j} p_j(x_j) - \sup_{x_j} |\widehat{p}_j(x_j)-p_j(x_j)| \geq \inf_{x_j} p_j(x_j) - o(1) \quad a.s.,$$

by (50) and (51), assumptions A2 and A4 are also satisfied by straightforward use of the geometric series expansion and the above result. Specifically, we have

$$\frac{1}{\widehat{p}_j(x_j)} = \frac{1}{p_j(x_j)} - \frac{\widehat{p}_j(x_j)-p_j(x_j)}{\widehat{p}_j(x_j)p_j(x_j)}.$$

Likewise, assumption A3 is satisfied by B2, B3, and (52). By the triangle inequality,

$$\begin{aligned}
\sup_{x_k}\left|\int \frac{\widehat{p}_{j,k}(x_j,x_k)}{\widehat{p}_k(x_k)}\widehat{m}_j^A(x_j)dx_j\right| &\leq \sup_{x_k}\left|\int \frac{p_{j,k}(x_j,x_k)}{p_k(x_k)}\widehat{m}_j^A(x_j)dx_j\right| + \\
&\quad \sup_{x_k}\left|\int\left[\frac{\widehat{p}_{j,k}(x_j,x_k)}{\widehat{p}_k(x_k)}-\frac{p_{j,k}(x_j,x_k)}{p_k(x_k)}\right]\widehat{m}_j^A(x_j)dx_j\right| \\
&\leq \sup_{x_k}\left|\int \frac{p_{j,k}(x_j,x_k)}{p_k(x_k)}\widehat{m}_j^A(x_j)dx_j\right| \\
&\quad + C\left\{\left\|\widehat{p}_{j,k}-p_{j,k}\right\|_\infty + \left\|\,\widehat{p}_k-p_k\right\|_\infty\right\}\left\|\widehat{m}_j^A\right\|_\infty
\end{aligned}$$

where the second term on the right hand side is $o(n^{-2/5})$ with probability one when $h = O(n^{-1/5})$. As for the first term, without loss of generality, we can suppose that

$$\widehat{m}_j^A(x_j) = n^{-1}\sum_{i=1}^{n}\frac{K_h(x_j - X_j^i)\eta_j^i}{p_j(x_j)} \equiv \frac{\widehat{v}_j(x_j)}{p_j(x_j)},$$

where $E(\eta_j^i|X_j^i) = 0$. Therefore,

$$\int \frac{p_{j,k}(x_j,x_k)}{p_k(x_k)}\widehat{m}_j^A(x_j)dx_j = \frac{1}{n}\sum_{i=1}^{n}\eta_j^i\xi_{ni}(x_k) \quad ; \quad \xi_{ni}(x_k) = \int K(u)\frac{p_{j,k}(X_j^i - uh, x_k)}{p_j(X_j^i - uh)p_k(x_k)}du$$

by straightforward change of variables. The argument is now quite similar to that given in Masry (1996). We drop the k subscript for convenience. Since the support of X is compact, it can be covered by a finite number $c(n)$ of cubes $I_{n,r}$ with centres x_r with dimension $l(n)$. We then have

$$\begin{aligned}
\sup_{x\in\mathcal{X}}\left|\int \frac{p_{j,k}(x_j,x)}{p_k(x)}\widehat{m}_j^A(x_j)dx_j\right| &= \max_{1\le r\le c(n)}\sup_{x\in\mathcal{X}\cap l_{n,r}}\left|\frac{1}{n}\sum_{i=1}^{n}\eta_j^i\xi_{ni}(x)\right| \\
&\le \max_{1\le r\le c(n)}\sup_{x\in\mathcal{X}\cap l_{n,r}}\left|\frac{1}{n}\sum_{i=1}^{n}\eta_j^i\xi_{ni}(x) - \frac{1}{n}\sum_{i=1}^{n}\eta_j^i\xi_{ni}(x_r)\right| \\
&\quad + \max_{1\le r\le c(n)}\left|\frac{1}{n}\sum_{i=1}^{n}\eta_j^i\xi_{ni}(x_r)\right| \\
&\equiv Q_1 + Q_2, \quad \text{say.}
\end{aligned}$$

It is straightforward to see that $|\xi_{ni}(x) - \xi_{ni}(x_r)| \le al(n)$ for some constant a and that $Q_1 = O_p(l(n))$. 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 $Q_1 + Q_2 = O(\log n/n^{1/2})$ with probability one. ■

References

[1] Auestad, B. and Tjøstheim, D., (1991). Functional identification in nonlinear time series. In *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, Kluwer Academic: Amsterdam. pp 493–507.

[2] Balakrishnan, A. V. (1981). *Applied functional analysis*, Springer, New York, Heidelberg, Berlin.

[3] Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models.* The John Hopkins University Press, Baltimore and London.

[4] Bollerslev, T., Engle, R.F., and D.B. Nelson (1994). ARCH Models. In *The Handbook of Econometrics*, vol. IV, eds. D.F. McFadden and R.F. Engle III. North Holland.

[5] Deutsch, F. (1985). Rate of convergence of the method of alternating projections. In: *Parametric optimization and approximation.* Ed. by B. Brosowski and F. Deutsch 96 - 107 Birkhäuser, Basel, Boston, Stuttgart.

[6] Fan, J, E. Mammen, and W. Härdle (1996). Direct estimation of low dimensional components in additive models. *Preprint.*

[7] Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

[8] Härdle, W. and L. Yiang (1996) 'Nonparametric Autoregression with Multiplicative Volatility and Additive Mean,' Forthcoming in *J. Time Ser. Anal.*

[9] Hastie, T. and R. Tibshirani (1991). *Generalized Additive Models.* Chapman and Hall, London.

[10] Linton, O.B. (1996). Efficient estimation of additive nonparametric regression models. *Biometrika*, To Appear.

[11] Linton, O.B. and W. Härdle. (1996). Estimating additive regression models with known links. *Biometrika* **83**, .

[12] Linton, O.B. and J.P. Nielsen. (1995). Estimating structured nonparametric regression by the kernel method. *Biometrika* **82**, 93-101.

[13] Mammen, E., Marron, J. S., Turlach, B. and Wand, M. P. (1997). A general framework for smoothing. *Preprint.* Forthcoming.

[14] Masry, E. (1996). Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and their Applications.* **65**, 81-101.

[15] Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *J. Time Ser. Anal.* **17**, 571-599.

[16] Newey, W.K. (1994). Kernel estimation of partial means. *Econometric Theory.* **10**, 233-253.

[17] Nielsen, J.P., and O.B. Linton (1997). An optimization interpretation of integration and backfitting estimators for separable nonparametric models. *J. Roy. Statist. Soc., Ser. B,* Forthcoming.

[18] Opsomer, J. D., (1997). On the existence and asymptotic properties of backfitting estimators. *Preprint.*

[19] Opsomer, J. D. and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25**, 186 - 211.

[20] Robinson, P.M. (1983). Nonparametric estimators for time series. *J. Time Ser. Anal.* **4**, 185-197.

[21] Rosenblatt, M. (1956). A central limit theorem and strong mixing conditions, *Proc. Nat. Acad. Sci.* **4**, 43-47.

[22] Ruppert, D., and M. Wand (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346-1370.

[23] Smith, K. T., D. C. Solomon, and S. L. Wagner (1977). Practical and mathematical aspects of the problem of reconstructing objects from radiographs. *Bull. Amer. Math. Soc.* **83**, 1227 -1270.

[24] Tjøstheim, D., and B. Auestad (1994). Nonparametric identification of nonlinear time series: projections. *J. Am. Stat. Assoc.* **89**, 1398-1409.