Available at
WWW.MATHEMATICSWEB.ORG
POWERED BY SCIENCE DIRECT

Journal of Statistical Planning and Inference 115 (2003) 521–542

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Derivative estimation and testing in generalized additive models[☆]

Lijian Yang[a, *], Stefan Sperlich[b], Wolfgang Härdle[c]

[a] *Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA*
[b] *Depto. de Estadistica y Econometria, Universidad Carlos III de Madrid, c/Madrid 126, E-28903, Getafe-Madrid, Spain*
[c] *Center for Applied Statistics and Economics, Institut für Statistik und Ökonometrie, Humboldt Universität zu Berlin, Spandauer Str. 1, D-10178 Berlin, Germany*

Received 22 November 2000; received in revised form 22 February 2002; accepted 1 March 2002

Abstract

Estimation and testing procedures for generalized additive (interaction) models are developed. We present extensions of several existing procedures for additive models when the link is the identity. This set of methods includes estimation of all component functions and their derivatives, testing functional forms and in particular variable selection. Theorems and simulation results are presented for the fundamentally new procedures. These comprise of, in particular, the introduction of local polynomial smoothing for this kind of models and the testing, including variable selection. Our method is straightforward to implement and the simulation studies show good performance in even small data sets. © 2002 Elsevier Science B.V. All rights reserved.

MSC: primary 62G07; 62G10; secondary 62H12; 62H15

Keywords: Component analysis; Derivative estimation; Generalized additive models; Local polynomial estimator; Marginal integration

1. Introduction

Additive models in nonparametric regression analysis are rather popular for mainly two reasons. In economic theory additivity is equivalent to the well-known property

[☆] This research was supported by NSF grant DMS 9971186, the Spanish "Dirección General de Enseñanza Superior" BEC2001-1270, and the Deutsche Forschungsgemeinschaft, SFB 373 "Quantifikation und Simulation ökonomischer Prozesse", Berlin.

[*] Corresponding author.
E-mail address: yang@stt.msu.edu (L. Yang).

PII: S0378-3758(02)00163-5

called strong separability and has many straightforward consequences for analysis, interpretation and decision making. In statistics it is well known due to articles of Stone (1985, 1986) that additive regression models can be estimated at the univariate rate of convergence. Most flexible model estimators suffer from the so-called "curse of dimensionality". This problem disappears if the impact of the regressors $X_1, X_2, \ldots, X_d$ on the response Y is in some sense separable, e.g. when the regression function $E[Y|X] = m(X)$ is additive

$$m(X) = c + \sum_{\beta=1}^{d} f_\beta(X_\beta). \tag{1}$$

Here, c is a constant and $\{f_\beta(\cdot)\}_{\beta=1}^{d}$ is a set of unknown functions. They are assumed to be smooth but otherwise arbitrary up to the identifiability condition $Ef_\beta(X_\beta) = 0$ for every $1 \leqslant \beta \leqslant d$.

Among the existing procedures such as backfitting (Hastie and Tibshirani, 1990; Mammen et al., 1999) and series estimator (Andrews and Whang, 1990), the marginal integration estimator (Tjøstheim and Auestad, 1994, or Linton and Nielsen, 1995) attracted a fair amount of attention thanks to the appealing simplicity in implementation as well as in theory. Further, its interpretation and extension to interactive models is well understood (see Nielsen and Linton, 1997; Sperlich et al., 1999, or Sperlich et al., 2002). Although the backfitting is easy to implement, its iterative structure has made its theoretical properties and correct interpretation difficult, see Mammen et al. (1999). Moreover, there is no theory for extensions as we know them already for the marginal integration estimator, e.g. by Fan et al. (1998), Severance-Lossin and Sperlich (1999), Linton and Härdle (1996) or Sperlich et al. (2002). Nevertheless, due to their very different interpretation when the real model is not additive, backfitting and marginal integration should not be considered as competitors.

Apparently, model (1) excludes a wide variety of situations. The most natural and often used extension is

$$G(m(X)) = c + \sum_{\beta=1}^{d} f_\beta(X_\beta), \tag{2}$$

where $G(\cdot)$ is a monotone link function. This is needed for many situations when model (1) is inappropriate, e.g. for binary and survival data. Widely used link functions include the logit and probit links for binary data, and the logarithm transform for Poisson count data. One can also let G be the logarithm function and so the regression function becomes multiplicative. Without loss of generality but along general practice, we assume the link function $G(\cdot)$ to be known a priori. Testing the specification of this link is beyond the scope of this paper but is discussed, e.g. in Härdle et al. (2001)

For the generalized additive model (2) (GAM) there is still need for investigation. On the one hand, there is little theory for the many existing backfitting procedures. Derivative estimation is a very important matter, especially in economics, but so far not investigated for these kinds of models. Indeed, consistent, direct estimation of derivatives is essential in economic studies, e.g. for estimating elasticities, returns to scale, substitution rates, average derivatives and much more. Often, these indices or functions are much more of interest than the regression function itself. Certainly, derivative estimates can be obtained from kernel estimates of the additive components. But already for the most simple case $G = identity$, Severance-Lossin and Sperlich (1999) showed that direct estimators for $f_\gamma^{(v)}$, $v > 0$, using local polynomials outperform by far taking derivatives of kernel estimates for the f_γ. Further, the need of testing methods for various problems as, e.g. variable selection, functional forms and additivity is obvious. As will be seen in the following, also here derivative estimation can be a very helpful tool. Since our methods do not restrict the form of link function G, they generalize the work of Hjellvik et al. (1998) which deals exclusively with additive models.

We introduce local polynomial estimation scheme for the components f_β in model (2) and their derivatives. For the ease of notation, asymptotic theory is explicitly derived only for the more complicated case of estimating derivatives. Having these estimates at hand they can be used for testing. This can be either testing against parametric, in our case polynomial, specification or it can be used for variable selection procedures. We construct our test statistics in analog to the one of Härdle and Mammen (1993). We performed a simulation study for the two original new contributions, i.e. derivative estimation and variable selection.

The paper is organized as follows. In the next section, we provide the technical setting for the problems and describes the marginal integration estimators of $f_\alpha^{(v)}(\cdot)$. In Section 3 we discuss important extensions. Section 4 presents procedures and theorems for a general testing method. Simulation studies are given in Section 5. All proofs are deferred to the appendix.

2. Estimation of functions and derivatives

As indicated before the main purpose of this paper is to complete the set of tools for the analysis of marginal impact functions in regression models, especially for GAM with known link function. We will present first procedures and theory for local polynomial smoother in models with possibly nonidentical link function. For brevity, we give our results in terms of derivative estimation, of which the estimation of component function is a particular case. Before coming closer to the here applied marginal integration method we need some general considerations about derivative estimation in generalized additive regression models.

To make inference on the derivative $f_\alpha^{(v)}(\cdot)$, we first want to express it in terms of the known G and the unknown m. Denote the variable $X = (X_\alpha, \bar{X})$ to highlight a particular direction α, where $\bar{X} = (X_1, \ldots, X_{\alpha-1}, X_{\alpha+1}, \ldots, X_d)$. The marginal density of X_α, that of $\bar{X}$ and the joint density of $X = (X_\alpha, \bar{X})$, are denoted by $\varphi_\alpha(x_\alpha)$, $\bar{\varphi}(\bar{x})$, and $\varphi(x_\alpha, \bar{x})$, respectively. We define $F_\alpha(x_\alpha) = \int G\{m(x)\}\bar{\varphi}(\bar{x})\,\mathrm{d}\bar{x} = c + f_\alpha(x_\alpha)$ for every $1 \leqslant \alpha \leqslant d$, then

$$G\{m(x)\} = \sum_{\alpha=1}^{d} F_\alpha(x_\alpha) - (d-1)c. \tag{3}$$

Taking derivatives on both sides and working by induction on v gives

Lemma 1. *For $v \geqslant 1$, define $J_v = \{(j_1, j_2, \ldots, j_v) \mid 0 \leqslant j_1, j_2, \ldots, j_v \leqslant v$, and $j_1 + 2j_2 + \cdots + vj_v = v\}$, the vth derivative $f_\alpha^{(v)}(x_\alpha)$ satisfies the following formula*:

$$f_\alpha^{(v)}(x_\alpha) = v! \sum_{(j_1,j_2,\ldots,j_v)\in J_v} G^{(j_1+j_2+\cdots+j_v)}\{m(x_\alpha,\bar{x})\} \prod_{\lambda=1}^{v} \frac{\{\partial_\alpha^{(\lambda)} m(x_\alpha,\bar{x})\}^{j_\lambda}}{(\lambda!)^{j_\lambda} j_\lambda!}, \tag{4}$$

where $\partial_\alpha^\lambda m(x) = \frac{\partial^\lambda m(x)}{\partial x_\alpha^\lambda}$.

Note from this lemma that a function of the vector variable x reduces to a function of a scalar variable x_α. Integrating both sides of (4) yields

Lemma 2. *For* $v \geqslant 1$

$$f_\alpha^{(v)}(x_\alpha)=v! \sum_{(j_1,j_2,\ldots,j_v)\in J_v} \int G^{(j_1+j_2+\cdots+j_v)}\{m(x_\alpha,\bar{x})\}\prod_{\lambda=1}^{v} \frac{\{\partial_\alpha^{(\lambda)} m(x_\alpha,\bar{x})\}^{j_\lambda}}{(\lambda!)^{j_\lambda} j_\lambda!}\, \bar{\varphi}(\bar{x})\,\mathrm{d}\bar{x}. \tag{5}$$

Eq. (5) implies that for an i.i.d. sample X_i, $i = 1,2,\ldots,n$

$$f_\alpha^{(v)}(x_\alpha) = \frac{1}{n}\sum_{i=1}^{n} G^{(j_1+j_2+\cdots+j_v)}\{m(x_\alpha,\bar{X}_i)\} \prod_{\lambda=1}^{v} \frac{\{\partial_\alpha^{(\lambda)} m(x_\alpha,\bar{X}_i)\}^{j_\lambda}}{(\lambda!)^{j_\lambda} j_\lambda!} + \mathrm{O_p}(1/\sqrt{n}). \tag{6}$$

This is used in the next paragraph to obtain estimators of $f_\alpha^{(v)}(x_\alpha)$ with low-dimensional rates typical for the dimension of the considered derivative function. Later we will also introduce a statistic for testing $f_\alpha^{(v)}(\cdot) \equiv 0$ based on its estimates.

For statistical inference, let (X_i, Y_i), $i=1,2,\ldots,n$ be an i.i.d. sample following model (2). The marginal integration estimator for $f_\alpha^{(v)}(x_\alpha)$, respectively $F_\alpha(x_\alpha)$ from (3) is defined by replacing in Eq. (5) the unknown expression $m(\cdot)$ by a local polynomial smoother $\tilde{m}(\cdot)$. The integral over the marginal density $\bar{\varphi}(\bar{x})$ we replace by (marginal) averaging over $\tilde{m}(x_\alpha,\bar{X}_i)$. The multidimensional local polynomial estimator has been introduced in detail by Ruppert and Wand (1994) and by Severance-Lossin and Sperlich (1999) in the context of marginal integration. We need the following notation.

Set for all $l=1,2,\ldots,n$ and $\lambda=0,1,2,\ldots,p$, where p is an integer such that $p-v > 0$ is odd

$$Z_\alpha = \{(X_{i\alpha} - x_\alpha)^\lambda\}_{n\times(p+1)}, \quad W_{l,\alpha} = \mathrm{diag}\left\{\frac{1}{n} K_h(X_{i\alpha} - x_\alpha)L_g(\bar{X}_i - \bar{X}_l)\right\}_{i=1}^{n},$$

$Y = (Y_i)_{n\times 1}$, and E_λ is a $(p+1)$ vector of zeros whose $(\lambda+1)$-element is 1,

where K and L are kernel functions, while for any function K, we denote $K_h(u) = K(u/h)/h$, and here h and g are bandwidths. In the following, $K^{(i)}$ denote the ith convolution of a function K with itself, and $\mu_r(K) = \int u^r K(u)\,\mathrm{d}u$. Note that $E_0'(Z_\alpha' W_{l,\alpha} Z_\alpha)^{-1} Z_\alpha' W_{l,\alpha} Y$ is a special local polynomial smoother to get our $\tilde{m}$.

Now we can give a closed expression for the estimators

$$\hat{F}_\alpha(x_\alpha) = n^{-1}\sum_{l=1}^{n} G\{E_0'(Z_\alpha' W_{l,\alpha} Z_\alpha)^{-1} Z_\alpha' W_{l,\alpha} Y\}$$

for $F_\alpha(x_\alpha)$ and consequently,

$$\hat m(x) = G^{-1}\left\{\sum_{\beta=1}^{d}\hat F_\beta(x_\beta) - (d-1)\frac{1}{nd}\sum_{l=1}^{n}\sum_{\beta=1}^{d}\hat F_\beta(X_{j\beta})\right\}$$

and

$$\widehat{\partial_\alpha^\lambda m}(x_\alpha, \bar X_l) = \lambda! E_\lambda'(Z_\alpha' W_{l,\alpha} Z_\alpha)^{-1} Z_\alpha' W_{l,\alpha} Y, \tag{7}$$

$$\hat f_\alpha^{(v)}(x_\alpha) = \frac{v!}{n}\sum_{l=1}^{n}\sum_{(j_1,j_2,\ldots,j_v)\in J_v} G^{(\sum_{\lambda=1}^{v} j_\lambda)}\{\hat m(x_\alpha,\bar X_l)\}\prod_{\lambda=1}^{v}\frac{\{\widehat{\partial^{(\lambda)} m}(x_\alpha,\bar X_l)\}^{j_\lambda}}{(\lambda!)^{j_\lambda} j_\lambda!} \tag{8}$$

for $m(x)$ and the derivatives $\partial_\alpha^\lambda m(x)$ or $f_\alpha^{(\lambda)}$, respectively. It can be seen easily that in the case of estimating f_α we take

$$\hat f_\alpha(x_\alpha) = \hat F_\alpha(x_\alpha) - \frac{1}{nd}\sum_{j=1}^{n}\sum_{\beta=1}^{d}\hat F_\beta(X_{j\beta}).$$

To establish the asymptotics for these estimators we need the following assumptions:

(A1) The kernel $K(\cdot)$ is a symmetric, compactly supported and Lipschitz continuous probability density; while the kernel $L(\cdot)$ is symmetric, compactly supported and Lipschitz continuous with $\int L(u)\,\mathrm{d}u = 1$ and order q *where* $q > \frac{1}{4}(d-1)$ for estimation and $q > (d-1)((p+1)/(p+3v))$ for testing hypotheses (which in effects, can even be relaxed to $q > (d-1)((p+1-v)/(p+3v))$ as one can see from the proof);

(A2) Bandwidths satisfy $nhg^{d-1}/\ln(n) \to \infty$, $g^{2q}/h^{p+1} \to 0$ and $h = h_e = h_0 n^{-1/(2p+3)}$ for estimation in Section 2 and $h = h_t = h_0 n^{-2/(p+3v+2)}$ for testing hypotheses in Section 4.

(A3) The functions $f_s(\cdot)$'s have bounded Lipschitz continuous $(p+1)$th derivatives.

(A4) The variance function, $\sigma^2(\cdot)$, is bounded and Lipschitz continuous.

(A5) φ and $\bar\varphi$ are uniformly bounded away from zero and infinity and have bounded Lipschitz continuous $(p+1)$th derivatives.

(A6) G is uniformly bounded away from zero and infinity and have bounded Lipschitz continuous $(p+1)$th derivative.

Note that in kernel regression, using marginal integration, these assumptions are standard and we therefore skip their discussion except one remark. In assumption A1 kernel L has to be a higher kernel because bias reduction for the nuisance directions is needed. Certainly, one could also allow K to be of higher order and introduce the corresponding changes in the asymptotic expressions. However, if we do not want to have a first-order impact of the nuisance directions on the bias, one has always to respect a certain trade-off between the order of K and L. The asymptotics of our estimators are given in the next theorem.

Theorem 1. *Under assumptions* A1–A6, *for any* α *and for* $v \geqslant 1$, *the estimated* v*th derivative* $\hat f_\alpha^{(v)}(x_\alpha)$ *satisfies*

$$\sqrt{nh^{2v+1}}\{\hat f_\alpha^{(v)}(x_\alpha) - f_\alpha^{(v)}(x_\alpha) - h^{p+1-v} b_{v\alpha}(x_\alpha)\} \xrightarrow{\mathrm{D}} N\{0, v_{v\alpha}(x_\alpha)\},$$

where

$$b_{v\alpha}(x_\alpha) = \frac{v!\mu_{p+1}(K_v^*)}{(p+1)!}\int \{(G' \circ m)\partial_\alpha^{(p+1)} m\}(x_\alpha, \bar{x})\bar{\varphi}(\bar{x})\,\mathrm{d}\bar{x},$$

$$v_{v\alpha}(x_\alpha) = (v!)^2 \|K_v^*\|_2^2 \int \left\{\frac{(G' \circ m)^2\sigma^2}{\varphi}\right\}(x_\alpha, \bar{x})\bar{\varphi}^2(\bar{x})\,\mathrm{d}\bar{x},$$

where $K_v^*(u) = \sum_{t=0}^p s_{vt}u^t K(u)$ *with* $(s_{st})_{0 \leqslant s,t \leqslant p} = S^{-1} = \{\mu_{s+t}(K)\}_{0 \leqslant s,t \leqslant p}^{-1}$.

Note here by the definition of matrix S that the λth equivalent kernel $K_\lambda^*(u)$ has the following property:

$$\mu_q(K_\lambda^*) = \begin{cases} 0, & q \leqslant p, \quad q \neq \lambda, \\ 1, & q = \lambda, \\ \Lambda_\lambda, & q = p+1, \end{cases} \tag{9}$$

where Λ_λ is some finite constant depending on λ. Note that the notation "equivalent kernel" refers to the equivalence between higher order kernel and local polynomial smoothing, see Lejeune (1985). It should not be confused with the notion of canonical kernels mentioned in Marron and Nolan (1988), which refers to the classes of kernels which are rescales of each other.

Now we write

$$\frac{1}{v!}(\hat{f}_\alpha^v(x_\alpha) - f_\alpha^v(x_\alpha)) = h^{p+1-v}\frac{1}{v!}b_{v\alpha}(x_\alpha) + \sum_{j=1}^n w_{j\alpha}\varepsilon_j + \mathrm{O}_\mathrm{p}\left(\frac{1}{\sqrt{n}} + h^{p+2-v}\right), \tag{10}$$

where we defined $\sigma(X_i)$, ε_i by $Y_i - m(X_i) = \sigma(X_i)\varepsilon_i$, and

$$w_{j\alpha} = \frac{1}{h^v n}K_{vh}^*(x_\alpha - X_{j\alpha})\frac{\bar{\varphi}(\bar{X}_j)\sigma(X_j)(G' \circ m)(x_\alpha, \bar{X}_j)}{\varphi(x_\alpha, \bar{X}_j)}. \tag{11}$$

It is easy to verify that this holds whether $h = h_e$ or $h = h_t$, compare assumption A2, and it will be made use of it in the next sections. We use residuals $\hat{u}_i := Y_i - \hat{m}(X_i)$, to approximate $\sigma(X_i)\varepsilon_i$.

3. Discussion of extensions

We now discuss briefly possible extensions which allow to consider more general models as done in Section 2. This ordering has been chosen as otherwise the notation would have become much too confusing in Section 2 and especially the appendix.

So far we have considered the GAM

$$G\{m(x)\} = c + \sum_{\beta=1}^d f_\beta(x_\beta) \quad \text{with } E[f_\alpha(X_\alpha)] = 0,$$

where all component functions were univariate. Clearly, the marginal integration idea to estimate marginal impact functions and its derivatives works through also for any other dimension of regressor X_α.

Consequently, the regressors $X_1,\dots,X_d$ could be grouped into $q \leqslant d$ (nonoverlapping) groups $Z_1 \in \mathbb{R}^{d_1},\dots,Z_q \in \mathbb{R}^{d_1}$, $\sum_{l=1}^{q} d_l = d$ ending up in model

$$G\{m(x)\} = c + \sum_{l=1}^{q} g_l(z_l) \quad \text{with } E[g_l(Z_l)] = 0. \tag{12}$$

The common problem is now to find the correct groups. But this question is equivalent to finding the significant interactions between the original, univariate regressors $X_1,\dots,X_d$. For this reason we decompose the regression as follows:

$$G\{m(x)\} = c + \sum_{\alpha=1}^{d} f_\alpha(x_\alpha) + \sum_{\alpha<\beta}^{d} f_{\alpha,\beta}(x_\alpha,x_\beta) + \sum_{\alpha<\beta<\gamma}^{d} f_{\alpha,\beta,\gamma}(x_\alpha,x_\beta,x_\gamma) + \cdots .$$

In practice one would stop after the second-order interaction to get an idea about the (correct) grouping in Eq. (12). Therefore, the interesting model is usually

$$G\{m(x)\} = c + \sum_{\alpha=1}^{d} f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha,\beta}(x_\alpha,x_\beta), \tag{13}$$

which can be identified when imposing the centering conditions

$$\int f_\alpha(u)\varphi_\alpha(u)\,\mathrm{d}u = 0$$

and

$$\int f_{\alpha,\beta}(u,v)\varphi_\alpha(u)\,\mathrm{d}u = \int f_{\alpha,\beta}(u,v)\varphi_\beta(v)\,\mathrm{d}v = 0.$$

An intensive discussion of the estimation of additive interaction models when the link $G(\cdot)$ is the identity can be found in Sperlich et al. (2002). Therefore, we only sketch here the procedure for the case when G is not trivial. Our consideration has been motivated by finding the right grouping in (12), so it is enough to estimate consistently the $f_{\alpha,\beta}$ up to a constant. With the methods presented in Section 4 these estimates can be used for testing significance.

Analogous to F_α we can define $F_{\alpha,\beta}(x_\alpha,x_\beta) = \int G\{m(x_\alpha,x_\beta,\check{x})\check{\varphi}(\check{x})\,\mathrm{d}\check{x}\}$ where $\check{x}$ is now the subvector of x containing all elements except x_α, x_β and $\check{\varphi}$ the marginal density of $\check{x}$. Some small calculations show that $(F_{\alpha,\beta} - F_\alpha - F_\beta)(\cdot)$ is equal to $f_{\alpha,\beta}$ up to an additive constant. Now we estimate

$$\hat{F}_{\alpha,\beta}(x_\alpha,x_\beta) = n^{-1}\sum_{l=1}^{n} G\{\tilde{m}(x_\alpha,x_\beta,\check{x})\},$$

where $\tilde{m}$ can be defined similar to above, see Sperlich et al. (2002), and proceed with $(\hat{F}_{\alpha,\beta} - \hat{F}_\alpha - \hat{F}_\beta)(\cdot)$.

One referee wondered whether our work can be extended to other smoothers such as, e.g. splines or series estimators. Certainly, the marginal integration idea can always be applied. However, apart from the fact that there is little distribution theory for splines in multivariate regression, series estimators (Andrews and Whang, 1990) as well as splines (Stone, 1994) in additive models are rather used to project the data directly into the space of additive models and look there for the optimal regression fit. As discussed in Sperlich et al. (2002), mixing such a projection with marginal integration is quite problematic in practice for interpretation reasons. Concerning derivative estimation in multivariate regression, to our knowledge kernel methods are so far the only approach that provides distribution theory, see e.g. Ruppert and Wand (1994), Severance-Lossin and Sperlich (1999) or Mammen et al. (1999). Finally, for the special context of testing problems Dette et al. (2001) discussed advantages and disadvantages of both, marginal integration and kernel-based methods.

4. Hypothesis testing on derivatives

We now turn to componentwise testing. The presented procedures will be useful to check significance or such polynomial structure as linearity for the considered functions. The interest in testing whether a function is significant at all is obvious as it enables us to perform variable selection as well as looking for interaction, see Section 3. Testing polynomial structure is motivated by both economic and statistic arguments. Especially, linearity has many important consequences in economics and thus is an important assumption to check. On the other hand, if a wanted parametric specification cannot be rejected, the empirical researcher will always prefer to use it. This is due to interpretation, facilities in modeling, etc.

As in the preceding sections we will condense the presentation on the case of test statistics with one-dimensional derivative functions. Let us first specify the hypothesis we focus on. We want to test the null hypothesis

$$\mathrm{H}_0: \int f_\alpha^{(v)}(x_\alpha)^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha = 0 \quad \text{vs. local alternatives}$$

$$\mathrm{H}_n: \frac{1}{v!^2} \int f_\alpha^{(v)}(x_\alpha)^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha > C\rho_n,$$

where $\pi(x)$ is a weight function with Lipschitz continuous $(p+1)$th derivative, $\rho_n = n^{-1}h^{-(2v+1/2)}$ and C is

$$(z_{1-\alpha_I} + z_{1-\alpha_{II}})\sigma_T, \tag{14}$$

where

$$\sigma_T^2 = \frac{\|K_v^{*(2)}\|_{L^2}^2}{2} \int \left[\int \left\{ \frac{(G' \circ m)^2 \sigma^2}{\varphi} \right\} (x_\alpha, \bar{x}) \bar{\varphi}^2(\bar{x})\,\mathrm{d}\bar{x} \right]^2 \pi(x_\alpha)^2\,\mathrm{d}x_\alpha. \tag{15}$$

Further, $z_{1-\alpha_I}$ is the upper $(1-\alpha_I)$th point of the standard normal variable, $\alpha_I \in (0,1)$ is the pre-specified significance level, while α_{II} is the pre-specified type II error. We define the test statistic

$$T = \int \frac{1}{(v!)^2} \hat{f}_\alpha^{(v)}(x_\alpha)^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha, \tag{16}$$

which is an estimate for $(1/(v!)^2) \int f_\alpha^{(v)}(x_\alpha)^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha$. The next theorem will show that T is a suitable statistic for testing H_0.

Theorem 2. *For any given* α *and* $h = h_t = h_0 n^{-2/(p+3v+2)}$ *as specified in* A2, *under assumptions* A1–A6, *the limiting distribution of* T *is*

$$h^{2v+1/2} nT - h^{2v+1/2} n \frac{1}{(v!)^2} \int f_\alpha^{(v)}(x_\alpha)^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha$$

$$- h^{-1/2} K_v^{*(2)}(0) \int \left\{ \frac{(G' \circ m)^2 \sigma^2}{\varphi} \right\} (x_\alpha, \bar{x}) \bar{\varphi}(\bar{x})^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha\,\mathrm{d}\bar{x}$$

$$- \frac{nh^{p+v+3/2} \mu_{p+1}(K_v^*)}{v!(p+1)!} \int \{(G' \circ m) \partial_\alpha^{(p+1)} m\}(x_\alpha, \bar{x}) \bar{\varphi}(\bar{x}) f_\alpha^{(v)}(x_\alpha) \pi(x_\alpha)\,\mathrm{d}x_\alpha\,\mathrm{d}\bar{x}$$

$$\xrightarrow{\mathrm{D}} N(0, \sigma_T^2). \tag{17}$$

The test rule is to reject H_0 *if*

$$h^{2v+1/2} nT \geqslant h^{-1/2} K_v^{*(2)}(0) \int \left\{ \frac{(G' \circ m)^2 \sigma^2}{\varphi} \right\} (x_\alpha, \bar{x}) \bar{\varphi}(\bar{x})^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha\,\mathrm{d}\bar{x}$$

$$+ z_{1-\alpha_I} \sigma_T. \tag{18}$$

The probability of type II error is smaller than α_{II} *as* $n \to \infty$: *for any function* $f_\alpha(x)$,

$$P[\mathrm{H}_0 \textit{ is retained} \mid \mathrm{H}_n \textit{ is true}] \leqslant \alpha_{II} + \mathrm{o}(n^{-(p-v+1)/(p+3v+2)}),$$

where the term $\mathrm{o}(n^{-(p-v+1)/(p+3v+2)})$ *implicitly depends on* $f_\alpha(\ )$.

Consider the test problem $\mathrm{H}_0 : f_\alpha$ *is linear*. Note that if looking at $\mathrm{H}_0 : \int f_\alpha^{(2)}(x_\alpha)^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha = 0$, taking $p = 3$, then $n^{-(p-v+1)/(p+3v+2)} = n^{-2/11}$ is the rate. Alternatively, we could look on the first derivative, i.e. $\mathrm{H}_0 : \int f_\alpha^{(2)}(x_\alpha)^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha = \text{const}$. Then, with $v = 1$, $p = 2$ we get even a rate of $n^{-2/7}$ although testing against zero or against a constant is basically the same. Apart from the rate, in small samples it can often be preferable for numerical reasons to look on lower degrees of derivatives if possible.

The remaining question is how to make these tests feasible, i.e. how to get the critical values. There are obviously two standard ways: estimating the asymptotics of the test or applying (wild) bootstrap.

To apply the rejection rule (18) from Theorem 2 we need to estimate σ_T, see (15) and the bias expression

$$h^{-1/2} K_v^{*(2)}(0) \int \left\{ \frac{(G' \circ m)^2 \sigma^2}{\varphi} \right\} (x_\alpha, \bar{x}) \bar{\varphi}(\bar{x})^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha\,\mathrm{d}\bar{x}. \tag{19}$$

For both expressions one would have to estimate function $m(\cdot)$, density $\varphi(\cdot)$ and its marginal $\bar{\varphi}(\cdot)$. This could be done by any nonparametric consistent estimator. Alternatively, for estimating the whole bias expression in once, one could also try one of the various bias-estimators for non or semiparametric models. Apart from the obvious fact that this task imposes one more crucial step, it is well known that the first-order asymptotics derived in Theorem 2 are not very helpful when applying the test in practice. Hjellvik et al. (1998) showed that maybe several hundred thousand observations are necessary to reach some reasonable accuracy by this method. This even gets worse when the asymptotic expressions also have to be estimated, again depending on some smoothing parameters.

Instead, (wild) bootstrap (see e.g. Liu, 1988 or Wu, 1986) is used to better approximate the distribution of the statistic T under hypothesis H_0. A detailed introduction, discussion and theory of this method in the context of testing problems combined with marginal integration can be found in Gozalo and Linton (2001), Härdle et al. (2001) or Sperlich et al. (2002). The idea is that random variables u_i^*, $i=1,\ldots,n$ are drawn from a distribution equal to that of the residuals $\hat{u}$ up to the second (or higher) moment. Then, bootstrap samples $Y_1^*,\ldots,Y_n^*$ can be constructed by $Y_i^* = G^{-1}\{c^0 + \sum_{\beta=1}^d \hat{f}_\beta^0(X_i)\} + u_i^*$, where the notations $f_\beta^0(\cdot)$ and c^0 indicate that the constant and the additive components f_β were estimated under the null hypothesis H_0. Having a bootstrap sample, calculate the (bootstrap) test statistic $\hat{T}^*$ out from sample $(X_i, Y_i^*)_{i=1}^n$. As the $\hat{T}^*$ are distributed as T under H_0, repeating this many times one gets thus a simulated critical value under H_0, respectively, a simulated p-value for T. Note that in case of limited dependent observations with binary response (let us still call it Y) as e.g. in probit or logit models, G^{-1} represents the error distribution and we can draw Y_i^* from $G^{-1}\{c^0 + \sum_{\beta=1}^d \hat{f}_\beta^0(X_i)\}$ directly.

The consistency of the (wild) bootstrap for this test procedure can be either concluded from Härdle et al. (2001) or, more directly, from Gozalo and Linton (2001). The latter considers an additivity test for our model and uses marginal integration as well. Not surprisingly, in their consistency proof Gozalo and Linton used the same decomposition as we [their Eq. (A.25)] but have two terms more [called U_{4i}, U_{5i}] due to the different testing problem. Further, they just handle with a simple kernel $K(\cdot)$ instead of $K_\nu^*(\cdot)$ and with the $f_\beta(\cdot)$ instead of their derivatives. As these technicalities are clearly irrelevant to their consistency proof for the bootstrap [their Proof of Theorem 2], we can conclude directly the consistency just by following their arguments [starting from their Eq. (A.39)]. We therefore skip here a detailed proof of consistency for the wild bootstrap.

5. Some simulation results

We investigated the performance of our procedures in finite samples; first for the derivative and function estimation, then for the variable selection, i.e. component wise testing for significance of the impact functions.

5.1. Function and derivative estimation

Although the introduction of a nontrivial link function $G(\cdot)$ looks straight forward for the marginal integration, in practice it unfortunately can cause strong negative effects on the small sample performance. We will illustrate this in the following by doing the same simulations twice, first for the identical link and then for $G(\cdot) = \ln(\cdot)$.

We drew $n=200$ independent variables $X \sim U[-2,2]^3$ and considered the models

$$G_\gamma\{m(X)\} = c + \sum_{\alpha=1}^{3} f_\alpha(X_\alpha), \quad \gamma = 1,2 \quad \text{with } f_1(X_1) = 1.5\sin(-1.5X_1)$$

$$f_2(X_2) = X_2^2 - \tfrac{4}{3}, \quad f_3(X_3) = X_3 \tag{20}$$

and $c = 3.0$. Further, G_1 is the identity and G_2 the logarithm. Finally, we added a standard normal disturbance ε to Eq. (20).

To get $\tilde{m}$ we used local linear smoother. In the previous section we did not discuss the question of bandwidth choice. Yang and Tschernig (1999) proposed plug-in optimal bandwidth for the simpler problem of multivariate regression, but there does not seem to be a solution for testing. In our current setting, the optimal bandwidth depends on the direction of estimation, the different G_γ and whether the function or its derivatives are being estimated. For function estimation, cross-validation (CV) is a commonly used bandwidth selector. However, the CV aims to minimize the mean squared error of estimating the whole regression, not any particular components. In addition, CV loses its appeal for derivative estimation. Therefore, we believe that plug-in methods might be more appropriate here. Recall that $h = h_0 n^{-1/(2p+3)}$, see assumption (A2). Then, following Theorem 1 the h_0 that minimizes the squared error at x_α is

$$h_0(x_\alpha) = \left[\frac{(2v+1)\|K_v^*\|_2^2 \int\{(G'\circ m)^2\sigma^2/\varphi\}(x_\alpha,\bar{x})\bar{\varphi}^2(\bar{x})\,\mathrm{d}\bar{x}}{2(p+1-v)((\mu_{p+1}(K_v^*)/(p+1)!)\int\{(G'\circ m)\partial_\alpha^{(p+1)}m\}(x_\alpha,\bar{x})\bar{\varphi}(\bar{x})\,\mathrm{d}\bar{x})^2}\right]^{1/(2p+3)}, \tag{21}$$

whereas the globally optimal one is

$$h_0 = \left[\frac{(2v+1)\|K_v^*\|_2^2 \int\{(G'\circ m)^2\sigma^2/\varphi\}(x_\alpha,\bar{x})\bar{\varphi}^2(\bar{x})\varphi_\alpha(x_\alpha)\,\mathrm{d}\bar{x}\,\mathrm{d}x_\alpha}{2(p+1-v)(\mu_{p+1}(K_v^*)/(p+1)!)^2 \int(\int\{(G'\circ m)\partial_\alpha^{(p+1)}m\}(x_\alpha,\bar{x})\bar{\varphi}(\bar{x})\,\mathrm{d}\bar{x})^2\varphi_\alpha(x_\alpha)\,\mathrm{d}x_\alpha}\right]^{1/(2p+3)}$$

minimizing the integrated mean squared error [MISE(f_α)]. Both expressions could be pre-estimated, either nonparametrically, or maybe better, approximated by parametric estimates. More discussion about this in the context of marginal integration estimation can also be found in Sperlich et al. (1999) or, in the context of testing, in Dette et al. (2001).

On the other hand, one has to understand that decreasing the bandwidth is somehow like decreasing the degrees of freedom, or, oversmoothing means to approximate the "true model" by a smooth one. This is indeed a philosophical question as already the existence of a true, estimable model is discussable. Empirical researchers in many practical fields may have a minimum degree of smoothness in mind and thus chooses the bandwidth "by eye" regardless of any asymptotic optimality criteria, see the recent work of Chaudhuri and Marron (1999) for the so-called SiZer approach to smoothing.

We consider only global bandwidths what is reasonable here as we assume no prior knowledge of the functionals f_γ, $\gamma = 1,\ldots,d$ and the distribution of the explanatory variables was uniform. As kernel we chose the quartic one. Then the exact solution for (21) is $n^{1/5}h_0 = (0.65157, 0.70568, \infty)$ (with $p = 1$) if the link is $G_1(\cdot)$. However, for only $n = 200$ this bandwidth led sometimes to numerical problems, and the impractical ∞ we replaced by 2.0 to allow for some nonlinearity in the estimation.

So we chose finally the bandwidth vector (for the estimation when the link is G_1) $h_1 = (1.0, 1.5, 2.0)$ what is a fair compromise between the optimality considerations and numerical necessities for the different curvatures. For simplification we set further $g = h$. When the link function is $G_2(\cdot)$, similar considerations lead to the choice $h_2 = (1.5, 1.75, 2.0)$.

After running 500 repetitions we had to skip about 1% of the results which still suffered from numerical problems when the link was G_2. In Figs. 1 (for G_1) and Fig. 2 (for G_2) are given the data generating functions, respectively their derivatives, as dotted lines together with the 99% confidence bands (solid lines) for the estimator resulting from the 500 repetitions. Note that we did no bias reduction here. For that reason the real data generating functions (dotted lines) do not lie inside the bands. Instead, we see clearly the structural biases.

Though the procedures seem to work reasonably well, we recognize an enormous loss of exactness when the link is not trivial. Not surprisingly, the derivative estimation with only $n = 200$ observations seems to be pretty hard, especially for G_2. We can recognize further the biases and boundary effects. As indicated before, the chosen bandwidths do not seem to be optimal but quite reasonable.

5.2. Testing the component functions

As the testing problem is easier than estimation, we considered here a more complicated model

$$G_T\{m(X)\} = \sum_{\alpha=1}^{3} f_\alpha(X_\alpha) \quad \text{with } f_3(X_3) = a \cdot X_3 \tag{22}$$

with f_1, f_2 as in (20) and $G_T(u) = -\ln(1/u - 1)$. So one observes

$$Y = \begin{cases} 1 & \text{if } \displaystyle\sum_{\alpha=1}^{3} f_\alpha(X_\alpha) > \varepsilon, \\ 0 & \text{else,} \end{cases}$$

where ε was logit distributed. Again we drew $n = 200$ independent variables $X \sim U[-2,2]^3$.

To implement the test, T was computed by

$$\hat{T} = \frac{1}{n}\sum_{j=1}^{n} \frac{\hat{f}_\alpha^{(v)}(X_{j\alpha})^2 \pi(X_{j\alpha})}{(1/n)\sum_{t=1}^{n} K_h(X_{j\alpha} - X_{t\alpha})} \frac{1}{(v!)^2}. \tag{23}$$

For the wild bootstrap, we took observations Y_i^*, $i = 1,\ldots,n$ drawn from the (estimated) data generating process under H_0, given $(X_i)_{i=1}^n$ and calculated the corresponding test

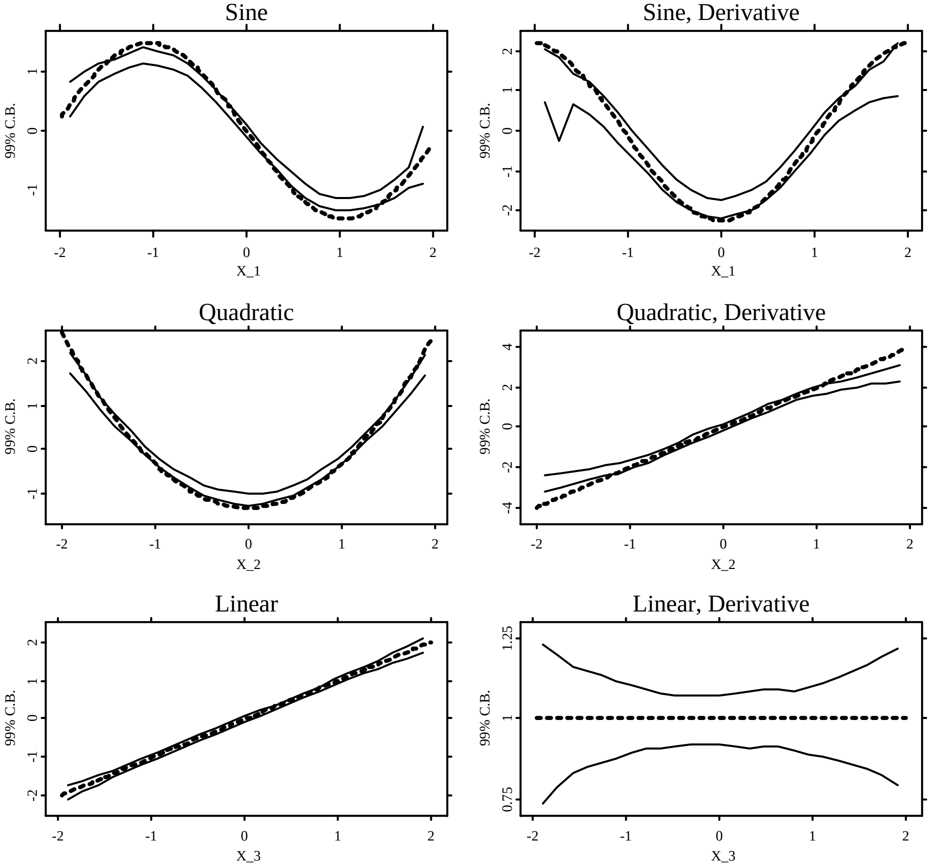

Fig. 1. Model (20) with identity link G_1. Functions on the left, derivatives on the right. Dotted lines are the data generating functions, respectively their derivatives, solid lines are the 99% confidence bands after 500 runs.

statistic $\hat{T}^*$. For this (pre-) estimation undersmoothing is recommended, see Härdle and Marron (1991). In our simulation study we used local linear smoother with $h = g = 1.5$ for all directions to estimate the data generating process under H_0. For the simulation study we drew only 249 bootstrap samples to approximate the distribution of $\hat{T}$. In practice one should certainly draw about 1000.

Our aim was to test $\mathrm{H}_0 : f_3 \equiv 0$ for increasing a, see (22). We first compared the test statistics based on function estimates with the one based on derivative estimates. It is certainly known that the one based on derivatives is especially of interest when the considered function is not smooth, e.g. has a peek or a jump. On the other hand, it is also known that in those cases kernel smoother cannot always be recommended.

In Theorem 2, for a first derivative-based test, the local quadratic smoother has been suggested. For those, larger bandwidths are necessary and we had to set the bandwidth

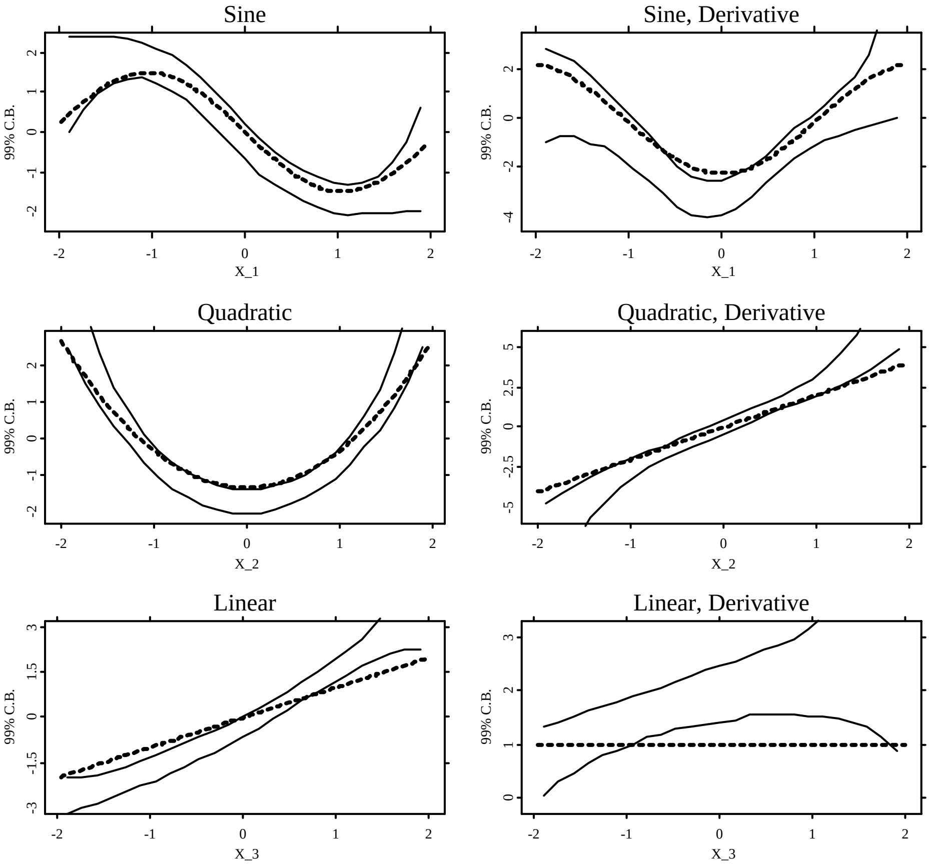

Fig. 2. Model (20) with log link G_2. Functions on the left, derivatives on the right. Dotted lines are the data generating functions, respectively their derivatives, solid lines are the 99% confidence bands after 500 runs.

Table 1
Relative rejection frequencies and p-values for testing $H_0: f_3 \equiv 0$ with tests based on function estimate $(v=0)$ and based on derivative estimates $(v=1)$, using local quadratic smoother with $h=g=3.0$

Significance level		1%	5%	10%	15%	p-value
$f_3(u)=0$	$v=0$	6.0	10.0	14.8	20.8	48.7
	$v=1$	2.0	2.8	6.4	8.4	55.0
$f_3(u)=u$	$v=0$	100	100	100	100	0.0
	$v=1$	24	42	49	58	16.5

to $h=g=3.0$. In Table 1 the relative frequencies of rejections for function-based test $(v=0)$ as well as for derivative-based $(v=1)$ tests are given, all after 500 repetitions. Additionally, in Table 2 we give the corresponding variances over the 500 repetitions.

We tried thereby other bandwidths and different degrees for the local polynomial smoother. We found that for $n=200$ the bandwidth choice can be very crucial when

Table 2
Variances for the relative rejection frequencies and p-values for testing $H_0 : f_3 \equiv 0$ with tests based on function estimate ($v = 0$) and based on derivative estimates ($v = 1$), see Table 1

Significance level		1%	5%	10%	15%	p-value
$f_3(u) = 0$	$v = 0$	5.7	9.0	12.7	16.5	9.6
	$v = 1$	2.0	2.7	6.0	7.7	7.6
$f_3(u) = u$	$v = 0$	0.0	0.0	0.0	0.0	0.0
	$v = 1$	17.9	24.5	25.1	24.5	3.4

Table 3
Relative rejection frequencies and p-value for testing $H_0 : f_3 \equiv 0$ with tests based on function estimate, using local linear smoother with $h = g = 1.75$. Left column refers to the alternative $f_3(u) = a \cdot u$

a	Significance level in (%)				
	1	5	10	15	p-value
0.00	1.2	4.4	8.4	11.8	54.4
0.25	6.4	16.4	28.0	36.8	34.5
0.50	37.2	60.0	71.6	78.4	10.4
0.75	84.0	93.6	97.2	98.8	0.01
1.00	98.4	99.6	100	100	0.00

using local quadratic or higher order polynomials. It can also be seen in Table 1 that though the p-value is fitted well under H_0, the quantiles are not. Thus we conclude that our estimates are just too wiggly or, the data are too sparse causing numerical problems. A very intensive simulation study would be necessary to investigate in detail the performance and usefulness of derivative-based tests in this context if the sample size is "small". We see e.g. in Table 1 that it is pretty conservative for these samples.

There are much more encouraging findings for tests based on the function estimate ($v = 0$). We present for this case also results when using local linear smoother with bandwidth $h = g = 1.75$. Again we did a simulation study of 500 repetitions. In Table 3 it can be seen how fast the power increases with a from (22).

All in all we conclude that even for small data sets but pretty complex model structures our procedures work reasonable well. The function and derivative estimates give clearly the wanted functional forms. The test procedures are more crucial. For such small samples we recommend to use only statistics based on function estimate (if possible) and low polynomial degrees. They perform well in both, fitting the correct quantiles under the hypothesis and showing strong power against the alternative.

Acknowledgements

The authors thank Oleg Lepskii and two anonymous referees for helpful comments and discussion.

Appendix

Our estimation procedure makes uses of the following lemma which is a generalization of the result of Linton and Härdle (1996).

Lemma A.1. *Under assumptions* A1–A6, *for any* α

$$\sqrt{nh}(\hat{F}_\alpha(x_\alpha) - F_\alpha(x_\alpha) - h^{p+1}b_\alpha(x_\alpha)) \xrightarrow{\mathrm{D}} N\{0, v_\alpha(x_\alpha)\},$$

where

$$b_\alpha(x_\alpha) = \frac{\mu_{p+1}(K_0^*)}{(p+1)!} \int \left[\frac{(G' \circ m)\{\partial_\alpha^{(p+1)}(m\varphi) - m\partial_\alpha^{(p+1)}(\varphi)\}}{\varphi} \right] (x_\alpha, \bar{x})\bar{\varphi}(\bar{x})\,\mathrm{d}\bar{x},$$

$$v_\alpha(x_\alpha) = \|K_0^*\|_2^2 \int \left\{ \frac{(G' \circ m)^2\sigma^2}{\varphi} \right\} (x_\alpha, \bar{x})\bar{\varphi}^2(\bar{x})\,\mathrm{d}\bar{x}.$$

Furthermore

$$\sqrt{nh}(\hat{m}(x) - m(x) - h^{p+1}b(x)) \xrightarrow{\mathrm{D}} N\{0, v(x)\}, \tag{A.1}$$

where

$$b(x) = (G^{-1})' \circ G \circ m(x) \sum_{\alpha=1}^{d} b_\alpha(x_\alpha)$$

and

$$v(x) = \{(G^{-1})' \circ G \circ m(x)\}^2 \sum_{\alpha=1}^{d} v_\alpha(x_\alpha).$$

Proof. It follows the same asymptotic reasoning as Linton and Härdle (1996) with minor changes because of the use of equivalent kernel K_0^* instead of K. The bias here is of order h^{p+1} instead of h^2. □

Proof of Theorem 1. The steps are similar to Severance-Lossin and Sperlich (1999). The special features here is the use of formula (4) and its empirical version (8), and the fact that for $\lambda = 0, 1, 2, \ldots, v$, the partial derivative estimates $\widehat{\partial^{(\lambda)}m}(x_\alpha, \bar{X}_l)$ have the bias rates of $h^{p+1-\lambda}$ and variance rates of $1/nh^{2\lambda+1}$, which together with the previous lemma, gives that

$$\sqrt{nh^{2v+1}}\,\frac{1}{v!}\,(\hat{f}_\alpha^{(v)}(x_\alpha) - f_\alpha^{(v)}(x_\alpha))$$

$$= n^{-1} \sum_{l=1}^{n} G'(m(x_\alpha, \bar{X}_l))(\widehat{\partial^{(v)}m}(x_\alpha, \bar{X}_l) - \partial^{(v)}m(x_\alpha, \bar{X}_l))$$

$$+ \mathrm{O}(\sqrt{nh^{2v+1}}h^{p+2-v} + h),$$

where the asymptotics of $n^{-1}\sum_{l=1}^{n}(G' \circ m)(\widehat{\partial^{(v)}m} - \partial^{(v)}m)(x_\alpha, \bar{X}_l)$ is treated as in Severance-Lossin and Sperlich (1999). □

The proof of Theorem 2 is essentially the same with trivial link or more general links. Therefore, to simplify notation, we give the proof in the case of trivial link function. In this case, $G' \circ m \equiv 1$ and $m(X) = c + \sum_{\beta=1}^{d} f_\beta(X_\beta)$. Therefore,

$$b_{v\alpha}(x) = \frac{v!\mu_{p+1}(K_v^*)}{(p+1)!} \int \{\partial_\alpha^{(p+1)} m\}(x_\alpha, \bar{x})\bar{\varphi}(\bar{x})\,\mathrm{d}\bar{x} = \frac{v!\mu_{p+1}(K_v^*)}{(p+1)!} f_\alpha^{(p+1)}(x_\alpha)$$

and the expression for T is

$$\int \left\{ \frac{f_\alpha^{(v)}(x_\alpha)}{v!} + \frac{h^{p+1-v}\mu_{p+1}(K_v^*)}{(p+1)!} f_\alpha^{(p+1)}(x_\alpha) + \sum_{j=1}^{n} w_{j\alpha}\varepsilon_j \right\}^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha$$

$$+ \mathrm{O_p}\left(\frac{1}{n} + h^{2p+4-2v}\right),$$

which can be reduced to

$$Q + \int \frac{f_\alpha^{(v)}(x_\alpha)^2}{(v!)^2} \pi(x_\alpha)\,\mathrm{d}x_\alpha + \frac{h^{p+1-v}\mu_{p+1}(K_v^*)}{(p+1)!v!} \int f_\alpha^{(p+1)}(x_\alpha) f_\alpha^{(v)}(x_\alpha)\pi(x_\alpha)\,\mathrm{d}x_\alpha$$

$$+ \mathrm{O_p}(h^{2p+4-2v}) \tag{A.2}$$

with the quadratic term $Q = \int \{\sum_{j=1}^{n} w_{j\alpha}\varepsilon_j\}^2 \pi(x_\alpha)\,\mathrm{d}x_\alpha$, and $w_{j\alpha}$ from (11).

We leave out here the routine verification that the following cross term is negligible:

$$\int 2\left\{ \frac{f_\alpha^{(v)}(x_\alpha)}{v!} + \frac{h^{p+1-v}\mu_{p+1}(K_v^*)}{(p+1)!} f_\alpha^{(p+1)}(x_\alpha) \right\} \sum_{j=1}^{n} w_{j\alpha}\varepsilon_j \pi(x_\alpha)\,\mathrm{d}x_\alpha.$$

The formula of σ_T^2 in the case of trivial link is simplified to

$$\sigma_T^2 = \frac{\|K_v^{*(2)}\|_{L^2}^2}{2} \int \left\{ \int \frac{\sigma^2(x)\bar{\varphi}^2(\bar{x})}{\varphi(x)}\,\mathrm{d}\bar{x} \right\}^2 \pi^2(x_\alpha)\,\mathrm{d}x_\alpha. \tag{A.3}$$

To derive the asymptotics of Q, write it as $\sum_{j,k=1}^{n} \varepsilon_j\varepsilon_k A(X_j, X_k)\,\sigma(X_j)\sigma(X_k)$ where

$$A(X_j, X_k) = \frac{1}{h^{2v}n^2} \int K_{vh}^*(x_\alpha - X_{j\alpha})K_{vh}^*(x_\alpha - X_{k\alpha})$$

$$\times \frac{\bar{\varphi}(\bar{X}_j)\bar{\varphi}(\bar{X}_k)}{\varphi(x_\alpha, \bar{X}_j)\varphi(x_\alpha, \bar{X}_k)} \pi(x_\alpha)\,\mathrm{d}x_\alpha. \tag{A.4}$$

Separating the diagonal and the cross terms, one gets $Q = Q_1 + Q_2$ with

$$Q_1 = \sum_{j=1}^{n} \varepsilon_j^2 A(X_j, X_j)\sigma(X_j)\sigma(X_j) = \sum_{j=1}^{n} A(X_j, X_j)\{Y_j - m(X_j)\}^2 \tag{A.5}$$

and

$$Q_2 = \sum_{1 \leqslant j < k = n} 2A(X_j, X_k)\{Y_j - m(X_j)\}\{Y_k - m(X_k)\}.$$

We simplify the expressions $A(X_j, X_k)$ and Q_1 in the following lemmata.

Lemma A.2. *$A(X_j, X_k)$ from* (A.4) *can be written as*

$$\frac{1}{h^{2\nu+1}n^2}(K_\nu^* * K_\nu^*)\left(\frac{X_{j\alpha} - X_{k\alpha}}{h}\right)\frac{\bar{\varphi}(\bar{X}_j)\bar{\varphi}(\bar{X}_k)}{\varphi(X_{j\alpha}, \bar{X}_j)\varphi(X_{j\alpha}, \bar{X}_k)}\pi(X_{j\alpha})\{1 + \mathrm{O_p}(h)\}. \quad \text{(A.6)}$$

Proof. By definition

$$\begin{aligned} A(X_j, X_k) &= \frac{1}{h^{2\nu}n^2}\int K_{\nu h}^*(x_\alpha - X_{j\alpha})K_{\nu h}^*(x_\alpha - X_{k\alpha})\frac{\bar{\varphi}(\bar{X}_j)\bar{\varphi}(\bar{X}_k)}{\varphi(x_\alpha, \bar{X}_j)\varphi(x_\alpha, \bar{X}_k)}\pi(x_\alpha)\,\mathrm{d}x_\alpha \\ &= \frac{1}{h^{2\nu+1}n^2}(K_\nu^* * K_\nu^*)\left(\frac{X_{j\alpha} - X_{k\alpha}}{h}\right)\frac{\bar{\varphi}(\bar{X}_j)\bar{\varphi}(\bar{X}_k)}{\varphi(X_{j\alpha}, \bar{X}_j)\varphi(X_{j\alpha}, \bar{X}_k)} \\ &\quad \times \pi(X_{j\alpha})\{1 + \mathrm{O_p}(h)\}. \end{aligned}$$

Lemma A.3. *As $n \to \infty$ it holds in* (A.5) *that*

$$Q_1 = \frac{(K_\nu^* * K_\nu^*)(0)}{h^{2\nu+1}n}\int \frac{\bar{\varphi}(\bar{x})^2\sigma(x)^2}{\varphi(x)}\pi(x_\alpha)\,\mathrm{d}x_\alpha\,\mathrm{d}\bar{x} + \mathrm{O_p}\left(\frac{1}{h^{2\nu}n} + \frac{1}{h^{2\nu+1}n^{3/2}}\right). \quad \text{(A.7)}$$

Proof. We calculate the mean and the variance of Q_1

$$\begin{aligned} EQ_1 &= nE\{A(X_1, X_1)\sigma(X_1)^2\} \\ &= \frac{(K_\nu^* * K_\nu^*)(0)}{h^{2\nu+1}n}\int \frac{\bar{\varphi}(\bar{x})^2\sigma(x)^2}{\varphi(x)}\pi(x_\alpha)\,\mathrm{d}x_\alpha\,\mathrm{d}\bar{x}\,\{1 + \mathrm{O}(h)\} \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(Q_1) &= n\,\mathrm{Var}\{A(X_1, X_1)\sigma(X_1)^2\} \leqslant nE\{A(X_1, X_1)^2\sigma(X_1)^4\} \\ &= \frac{(K_\nu^* * K_\nu^*)(0)^2}{h^{4\nu+2}n^3}E\frac{\bar{\varphi}(\bar{X}_1)^4\sigma(X_1)^4}{\varphi(X_1)^4}\pi^2(X_{1\alpha})\{1 + \mathrm{O}(h)\} \\ &= \mathrm{O_p}\left(\frac{1}{h^{2\nu+1}n^{3/2}}\right). \end{aligned}$$

Therefore,

$$Q_1 = \frac{(K_\nu^* * K_\nu^*)(0)}{h^{2\nu+1}n}\int \frac{\bar{\varphi}(\bar{x})^2\sigma(x)^2}{\varphi(x)}\pi(x_\alpha)\,\mathrm{d}x_\alpha\,\mathrm{d}\bar{x} + \mathrm{O_p}\left(\frac{1}{h^{2\nu}n} + \frac{1}{h^{2\nu+1}n^{3/2}}\right)$$

as is in (A.7). □

Note because $E[\varepsilon_i]=0$, $i=1,2,\dots,n$ and the random vectors (X_i,ε_i), $i=1,2,\dots,n$ are i.i.d., Q_2 is an U-statistic, symmetric and *nondegenerate* because $E_j\varepsilon_j\varepsilon_k A(X_j,X_k)\sigma(X_j)\sigma(X_k)=0$, where $E_j=E_{\varepsilon_j,X_j}$. To apply central limit theorem to this U-statistic, we calculate the following three quantities:

1. The variance of one term: $A_n=E[\varepsilon_1\varepsilon_2 A(X_1,X_2)\sigma(X_1)\sigma(X_2)]^2$
2. The fourth moment of one term: $B_n= E[\varepsilon_1\varepsilon_2 A(X_1,X_2)\sigma(X_1)\sigma(X_2)]^4$
3. The $C_n=E[J_n(\varepsilon_1,X_1,\varepsilon_2,X_2)]^2$, where

$$J_n(\varepsilon,X,\delta,Y)=E_1[\varepsilon_1\varepsilon A(X_1,X)\sigma(X_1)\sigma(X)\varepsilon_1\delta A(X_1,Y)\sigma(X_1)\sigma(Y)]$$

and then verify that

$$\frac{C_n+(1/n)B_n}{A_n^2}\to 0,\quad \text{as } n\to\infty \tag{A.8}$$

see, Hall (1984).

Lemma A.4. *As $n\to\infty$ in* (A.8) *one has*

$$A_n=\frac{2\sigma_T^2}{h^{4v+1}n^4}+\mathrm{O}\left(\frac{1}{h^{4v}n^4}\right). \tag{A.9}$$

Proof. We start with the definition of A_n and Eq. (A.6) in Lemma A.2

$$A_n=E\left[\frac{1}{h^{2v+1}n^2}(K_v^* * K_v^*)\left(\frac{X_{1\alpha}-X_{2\alpha}}{h}\right)\frac{\bar{\varphi}(\bar{X}_1)\bar{\varphi}(\bar{X}_2)\sigma(X_1)\sigma(X_2)}{\varphi(X_{1\alpha},\bar{X}_1)\varphi(X_{1\alpha},\bar{X}_2)}\right.$$
$$\left.\times\,\pi(X_{1\alpha})\{1+\mathrm{O}_{\mathrm{p}}(h)\}\right]^2$$

or

$$\frac{1}{h^{4v+2}n^4}\int\left[(K_v^* * K_v^*)\left(\frac{x_\alpha-y_\alpha}{h}\right)\frac{\bar{\varphi}(\bar{x})\bar{\varphi}(\bar{y})\sigma(x)\sigma(y)}{\varphi(x_\alpha,\bar{x})\varphi(x_\alpha,\bar{y})}\pi(x_\alpha)\{1+\mathrm{O}_{\mathrm{p}}(h)\}\right]^2$$
$$\times\varphi(x_\alpha,\bar{x})\varphi(y_\alpha,\bar{y})\,\mathrm{d}x_\alpha\,\mathrm{d}\bar{x}\,\mathrm{d}y_\alpha\,\mathrm{d}\bar{y}$$

and which equals, by change of variables $y_\alpha=x_\alpha+hu$

$$\frac{1}{h^{4v+1}n^4}\int\left[(K_v^* * K_v^*)(u)\frac{\bar{\varphi}(\bar{x})\bar{\varphi}(\bar{y})\sigma(x)\sigma(x_\alpha+hu,\bar{y})}{\varphi(x_\alpha,\bar{x})\varphi(x_\alpha,\bar{y})}\pi(x_\alpha)\right]^2$$
$$=\frac{\|(K_v^* * K_v^*)\|_{L^2}^2}{h^{4v+1}n^4}\int\frac{\bar{\varphi}(\bar{x})^2\bar{\varphi}^2(\bar{y})\sigma^2(x)\sigma^2(x_\alpha,\bar{y})}{\varphi(x_\alpha,\bar{x})\varphi(x_\alpha,\bar{y})}\pi^2(x_\alpha)\,\mathrm{d}x_\alpha\,\mathrm{d}\bar{x}\,\mathrm{d}\bar{y}\,\{1+\mathrm{O}(h)\}$$
$$=\frac{2}{h^{4v+1}n^4}\frac{\|K_v^{*(2)}\|_{L^2}^2}{2}\int\left\{\int\frac{\sigma^2(x_\alpha,\bar{x})\bar{\varphi}^2(\bar{x})}{\varphi(x)}\,\mathrm{d}\bar{x}\right\}^2\pi(x_\alpha)^2\,\mathrm{d}x_\alpha\,\{1+\mathrm{O}(h)\}$$
$$=\frac{2\sigma_T^2}{h^{4v+1}n^4}+\mathrm{O}\left(\frac{1}{h^{4v}n^4}\right).\qquad\square$$

Lemma A.5. *As $n \to \infty$, in* (A.8) *one has*

$$B_n = \frac{\|K_v^{*(2)}\|_{L^4}^4}{h^{8v+3}n^8} \int \left\{ \int \frac{\sigma^4(x)\bar{\varphi}^4(\bar{x})}{\varphi^3(x)}\,\mathrm{d}\bar{x} \right\}^2 \pi^4(x_\alpha)\,\mathrm{d}x_\alpha + \mathrm{O}\left(\frac{1}{h^{8v+2}n^8}\right). \tag{A.10}$$

Proof. Like for A_n, we start with the definition of B_n and Eq. (A.6) in Lemma A.2

$$B_n = E\left[\frac{1}{h^{2v+1}n^2}(K_v^* * K_v^*)\left(\frac{X_{1\alpha}-X_{2\alpha}}{h}\right) \frac{\bar{\varphi}(\bar{X}_1)\bar{\varphi}(\bar{X}_2)\sigma(X_1)\sigma(X_2)}{\varphi(X_{1\alpha},\bar{X}_1)\varphi(X_{1\alpha},\bar{X}_2)} \right.$$

$$\left. \times\, \pi(X_{1\alpha})\{1+\mathrm{O}_\mathrm{p}(h)\} \right]^4$$

$$= \frac{1}{h^{8v+4}n^8} \int \left[(K_v^* * K_v^*)\left(\frac{x_\alpha - y_\alpha}{h}\right) \frac{\bar{\varphi}(\bar{x})\bar{\varphi}(\bar{y})\sigma(x)\sigma(y)}{\varphi(x_\alpha,\bar{x})\varphi(x_\alpha,\bar{y})}\,\pi(x_\alpha)\{1+\mathrm{O}_\mathrm{p}(h)\} \right]^4$$

$$\times\varphi(x_\alpha,\bar{x})\varphi(y_\alpha,\bar{y})\,\mathrm{d}x_\alpha\,\mathrm{d}\bar{x}\,\mathrm{d}y_\alpha\,\mathrm{d}\bar{y}$$

$$= \frac{\|K_v^{*(2)}\|_{L^4}^4}{h^{8v+3}n^8} \int \left\{ \int \frac{\sigma^4(x_\alpha,\bar{x})\bar{\varphi}^4(\bar{x})}{\varphi^3(x)}\,\mathrm{d}\bar{x} \right\}^2 \pi^4(x_\alpha)\,\mathrm{d}x_\alpha\,\{1+\mathrm{O}(h)\}. \qquad \square$$

Now we want to calculate $C_n = E[J_n(\varepsilon_1,X_1,\varepsilon_2,X_2)]^2$, where

$$J_n(\varepsilon,X,\delta,Y) = E_1[\varepsilon_1\varepsilon A(X_1,X)\sigma(X_1)\sigma(X)\varepsilon_1\delta A(X_1,Y)\sigma(X_1)\sigma(Y)].$$

Lemma A.6. *It holds that*

$$J_n(\varepsilon,X,\delta,Y) = \frac{\varepsilon\delta\bar{\varphi}(\bar{X})\sigma(X)\bar{\varphi}(\bar{Y})\sigma(Y)\pi^2(X_\alpha)}{h^{4v+1}n^4\varphi(X_\alpha,\bar{X})\varphi(X_\alpha,\bar{Y})} K_v^{*(4)}\left(\frac{Y_\alpha - X_\alpha}{h}\right)$$

$$\times \int \frac{\bar{\varphi}^2(\bar{x})\sigma^2(X_\alpha,\bar{x})}{\varphi(X_\alpha,\bar{x})}\,\mathrm{d}\bar{x}\,\{1+\mathrm{O}_\mathrm{p}(h)\}. \tag{A.11}$$

Proof. By definition of J_n and Eq. (A.6) in Lemma A.2

$$J_n(\varepsilon,X,\delta,Y) = \frac{\varepsilon\delta}{h^{4v+2}n^4} \int K_v^{*(2)}\left(\frac{x_\alpha - X_\alpha}{h}\right) \frac{\bar{\varphi}(\bar{x})\bar{\varphi}(\bar{X})\sigma(X)}{\varphi(x_\alpha,\bar{x})\varphi(x_\alpha,\bar{X})}\,\sigma^2(x)$$

$$\times K_v^{*(2)}\left(\frac{x_\alpha - Y_\alpha}{h}\right) \frac{\bar{\varphi}(\bar{x})\bar{\varphi}(\bar{Y})\sigma(Y)}{\varphi(x_\alpha,\bar{x})\varphi(x_\alpha,\bar{Y})}\,\pi^2(x_\alpha)\varphi(x_\alpha,\bar{x})\,\mathrm{d}x_\alpha\,\mathrm{d}\bar{x}$$

$$\times\{1+\mathrm{O}_\mathrm{p}(h)\},$$

which, by a change of variable $x_\alpha = X_\alpha + hu$, becomes

$$J_n(\varepsilon, X, \delta, Y) = \frac{\varepsilon\delta}{h^{4v+1}n^4} \int K_v^{*(2)}(u) \frac{\bar{\varphi}(\bar{x})\bar{\varphi}(\bar{X})\sigma(X)}{\varphi(X_\alpha + hu, \bar{x})\varphi(X_\alpha + hu, \bar{X})}\, \sigma^2(X_\alpha + hu, \bar{x})$$

$$\times K_v^{*(2)}\left(\frac{X_\alpha - Y_\alpha}{h} + u\right) \frac{\bar{\varphi}(\bar{x})\bar{\varphi}(\bar{Y})\sigma(Y)}{\varphi(X_\alpha + hu, \bar{x})\varphi(X_\alpha + hu, \bar{Y})}$$

$$\times\, \pi^2(X_\alpha + hu)\varphi(X_\alpha + hu, \bar{x})\, \mathrm{d}u\, \mathrm{d}\bar{x}\, \{1 + \mathrm{O_p}(h)\}$$

and thus

$$J_n(\varepsilon, X, \delta, Y) = \frac{\varepsilon\delta\bar{\varphi}(\bar{X})\sigma(X)\bar{\varphi}(\bar{Y})\sigma(Y)\pi^2(X_\alpha)}{h^{4v+1}n^4\varphi(X_\alpha, \bar{X})\varphi(X_\alpha, \bar{Y})} (K_v^{*(4)}) \left(\frac{Y_\alpha - X_\alpha}{h}\right)$$

$$\times \int \frac{\bar{\varphi}^2(\bar{x})\sigma^2(X_\alpha, \bar{x})}{\varphi(X_\alpha, \bar{x})}\, \mathrm{d}\bar{x}\, \{1 + \mathrm{O_p}(h)\}. \qquad \square$$

Lemma A.7.

$$C_n = \frac{\|K_v^{*(4)}\|_{L^2}^2}{h^{8v+1}n^8} \int \left\{ \int \frac{\bar{\varphi}(\bar{x})^2\sigma(x_\alpha, \bar{x})^2}{\varphi(x_\alpha, \bar{x})}\, \mathrm{d}\bar{x} \right\}^4 \pi^4(x_\alpha)\, \mathrm{d}x_\alpha + \mathrm{O}\left(\frac{1}{h^{8v}n^8}\right). \tag{A.12}$$

Proof. By definition, applying substitution and some algebra. □

Lemma A.8. *As $n \to \infty$ it holds*

$$\sqrt{h^{4v+1}n^2}Q_2 \xrightarrow{\mathrm{D}} N(0, \sigma_T^2). \tag{A.13}$$

Proof. we have established in (A.9), (A.10), and (A.12) that $A_n \propto 1/h^{4v+1}n^4$, $B_n \propto 1/h^{8v+3}n^8$, and $C_n \propto 1/h^{8v+1}n^8$, and hence

$$\frac{C_n + (1/n)B_n}{A_n^2} = \mathrm{O}\left(h + \frac{1}{nh}\right) \to 0, \quad \text{as } n \to \infty.$$

Therefore, by the central limit theorem for nondegenerate U-statistic as in Hall (1984), $\sqrt{h^{4v+1}n^2}Q_2$ is asymptotically normal with asymptotic variance

$$\frac{n^2}{2} h^{4v+1}n^2 A_n = \frac{n^2}{2} h^{4v+1}n^2 \frac{2\sigma_T^2}{h^{4v+1}n^4} = \sigma_T^2. \qquad \square$$

Proof of Theorem 2. Now combining the results on Q_1 and Q_2, namely (A.7) in Lemma A.3 and (A.13) in Lemma A.8, plus Eq. (A.2), we obtain Eq. (17) in Theorem 2. □

References

Andrews, D.W.K., Whang, Y.J., 1990. Additive interactive regression models: circumvention of the curse of dimensionality. Econom. Theory 6, 466–479.

Chaudhuri, P., Marron, J.S., 1999. SiZer for exploration of structures in curves. J. Amer. Statist. Assoc. 94, 807–823.

Dette, H., von Liers und Wilkau, C., Sperlich, S., 2001. A comparison of different nonparametric methods for inference on additive models. Working Paper 01-28, Carlos III de Madrid, Spain.

Fan, J., Härdle, W., Mammen, E., 1998. Direct estimation of low dimensional components in additive models. Ann. Statist. 26, 943–971.

Gozalo, P.L., Linton, O.B., 2001. Testing additivity in generalized nonparametric regression models. J. Econom. 104, 1–48.

Hall, P., 1984. Central limit theorem for integrated square error of multivariate nonparametric density estimators. J. Multivariate Anal. 14, 1–16.

Härdle, W., Mammen, E., 1993. Comparing nonparametric versus parametric regression fits. Ann. Statist. 21, 1926–1947.

Härdle, W., Marron, J.S., 1991. Bootstrap simultaneous error bars for nonparametric regression. Ann. Statist. 19, 778–796.

Härdle, H., Huet, S., Mammen, E., Sperlich, S., 2001. Bootstrap inference in semiparametric generalized additive models. Working Paper 00-70, Carlos III de Madrid, Spain.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London.

Hjellvik, V., Yao, Q., Tjøstheim, D., 1998. Linearity testing using local polynomial approximation. J. Statist. Plann. Inference 68, 295–321.

Lejeune, M., 1985. Estimation non-parametrique par noyaux: regression polynomiale mobile. Rev. Statist. Appl. XXXIII, 43–67.

Linton, O.B., Härdle, W., 1996. Estimation of additive regression models with known links. Biometrika 83, 529–540.

Linton, O.B., Nielsen, J.P., 1995. A kernel method of estimating structured nonparametric regression based on marginal integration. Biometrika 82, 93–101.

Liu, R., 1988. Bootstrap procedures under some non i.i.d. models. Ann. Statist. 16, 1696–1708.

Mammen, E., Linton, O.B., Nielsen, J.P., 1999. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Ann. Statist. 27, 1443–1490.

Marron, J.S., Nolan, D., 1988. Canonical kernels for density estimation. Statist. Probab. Lett. 7, 195–199.

Nielsen, J.P., Linton, O.B., 1997. An optimization interpretation of integration and backfitting estimators for separable nonparametric models. J. Roy. Statist. Soc. Ser. B 60, 217–222.

Ruppert, D., Wand, M.P., 1994. Multivariate locally weighted least squares regression. Ann. Statist. 22, 1346–1370.

Severance-Lossin, E., Sperlich, S., 1999. Estimation of derivatives for additive separable models. Statistics 33, 241–265.

Sperlich, S., Linton, O.B., Härdle, W., 1999. Integration and backfitting methods in additive models: finite sample properties and comparison. Test 8, 419–458.

Sperlich, S., Tjøstheim, D., Yang, L., 2002. Nonparametric estimation and testing of interaction in additive models. Econom. Theory 18, 197–251.

Stone, C.J., 1985. Additive regression and other nonparametric models. Ann. Statist. 13, 689–705.

Stone, C.J., 1986. The dimensionality reduction principle for generalized additive models. Ann. Statist. 14, 90–606.

Stone, C.J., 1994. The use of polynomial splines and their tensor products in multivariate function estimation. Ann. Statist. 22, 118–184.

Tjøstheim, D., Auestad, B.H., 1994. Nonparametric identification of nonlinear time series: projections. J. Amer. Statist. Assoc. 89, 1398–1409.

Wu, C.F.J., 1986. Jacknife, bootstrap and other resampling methods in regression analysis (with discussion). Ann. Statist. 14, 1261–1350.

Yang, L., Tschernig, R., 1999. Multivariate bandwidth selection for local linear regression. J. Roy. Statist. Soc. Ser. B 61, 793–815.