Bernoulli **7**(4), 2001, 669–697

Testing additivity by kernel-based methods – what is a reasonable test?

HOLGER DETTE* and CARSTEN VON LIERES UND WILKAU**

Ruhr-Universität Bochum, Fakultät für Mathematik, 44780 Bochum, Germany.
*E-mail: *holger.dette@ruhr-uni-bochum.de; **carsten.von.lieres@ruhr-uni-bochum.de*

In the common nonparametric regression model with high-dimensional predictor, several tests for the hypothesis of an additive regression are investigated. The corresponding test statistics are based either on the differences between a fit under the assumption of additivity and a fit in the general model, or on residuals under the assumption of additivity. For all tests asymptotic normality is established under the null hypothesis of additivity and under fixed alternatives with different rates of convergence corresponding to both cases. These results are used for a comparison of the different methods. It is demonstrated that a statistic based on an empirical L^2-distance of the Nadaraya–Watson and the marginal integration estimator yields the (asymptotically) most efficient procedure, if these are compared with respect to the asymptotic behaviour under fixed and local alternatives. The finite-sample properties of the proposed procedures are investigated by means of a simulation study, which qualitatively confirms the asymptotic results.

Keywords: additive models; curse of dimensionality; dimension reduction; marginal integration estimate; test of additivity

1. Introduction

Consider the common nonparametric regression model

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1.1}$$

where $X = (X_1, \ldots, X_d)^{\mathrm{T}}$ is a d-dimensional random variable, Y is the real-valued response, ε denotes the real-valued error (independent of X) with mean 0 and variance 1, and m, σ are unknown (smooth) functions. Much effort has been devoted to the problem of estimating the regression function m. While for a one-dimensional predictor nonparametric methods such as kernel and local polynomial estimators have become increasingly popular, the regression in the case of a high-dimensional predictor cannot be estimated efficiently because of the so-called curse of dimensionality.

For this reason many methods of dimensionality reduction have been proposed in the literature (see, for example, Friedman and Stuetzle 1981; Li 1991). Buja *et al.* (1989) and Hastie and Tibshirani (1990) promoted the additive regression model

$$H_0 : m(x) = C + \sum_{\alpha=1}^{d} k_\alpha(x_\alpha), \tag{1.2}$$

1350–7265

where $k_1, \ldots, k_d$ are unknown smooth functions normalized by $\mathrm{E}[k_\alpha(X_\alpha)] = 0$ and $x = (x_1, \ldots, x_d)^{\mathrm{T}}$. A theoretical motivation for this model is that under the assumption of additivity the regression can be estimated with the same rate of estimation error as in the univariate case (see Stone 1985). Buja *et al.* (1989) proposed backfitting, where the idea is to project the data on the space of additive functions. Basically, this method estimates the orthogonal projection of the regression function $m(\cdot)$ onto the subspace of additive functions in the Hilbert space induced by the density of the predictor X. The asymptotic properties of a related backfitting procedure have recently been analysed by Opsomer and Ruppert (1997) and Mammen *et al.* (1999). Because of the implicit definition of these estimates, several authors have proposed a direct method based on marginal integration (see, for example, Tjøstheim and Auestad 1994; Tjøstheim 1994; Linton and Nielsen 1995). This method does not require the iterative solution of a system of nonlinear equations and yields an alternative projection onto the subspace of additive functions which is not necessarily orthogonal. Several authors have proposed modifications of the marginal integration estimator, (see, for example, Fan *et al*., 1998; Linton 1997; 2000; Hengartner 1996; Severance-Lossin and Sperlich 1999). For a more detailed discussion of the difference between the backfitting and the marginal integration estimator, we refer to the work of Nielsen and Linton (1998) or Sperlich, Linton and Härdle (1999). A rather different approach to estimating an additive regression function can be obtained by Fourier series estimation and is discussed by Andrews and Whang (1990). This method was used by Eubank *et al.* (1995) for the construction of a test of additivity if the data are observable on a grid.

Because the additive structure is important in terms of interpretability and its ability to deliver fast rates of convergence in the problem of estimating the regression, the additive model (1.2) should be accompanied by an adequate model check. Although early work dates back to Tukey (1949), it is only recently that the problem of testing additivity has been of real interest (see for example, Hastie and Tibshirani 1990; Barry 1993; Eubank *et al.* 1995; Sperlich, Tjøstheim and Yang 1999; Gozalo and Linton 2001). Various authors argue that, even if the null hypothesis (1.2) is accepted with a rather large p-value, there need not be any empirical evidence for the additive model (see Berger and Delampady 1987; Staudte and Sheather 1990). These authors point out that it is often preferable to reformulate the hypothesis (1.2) as

$$H_\eta : M^2 > \eta, \qquad H_1 : M^2 \leqslant \eta, \tag{1.3}$$

where M^2 is a measure of additivity and η is a given sufficiently small constant such that the experimenter agrees to analyse the data under the assumption of additivity whenever $M^2 \leqslant \eta$. From a mathematical point of view this approach requires the determination of the distribution of an appropriate estimator for M^2 not only under the classical null hypothesis (1.2) ($M^2 = 0$) but also at any point of the alternative ($M^2 > 0$).

In this paper we investigate several tests for the hypothesis of additivity which are based on kernel methods. For the sake of simplicity we will mainly concentrate on a U statistic formed from the residuals of a marginal integration fit – see also Zheng (1996), who used a similar idea for testing a parametric form of the regression. We prove asymptotic normality of the corresponding test statistic under the null hypothesis of additivity and fixed alternatives with different rates of convergence corresponding to both cases. The results are

then extended to several related concepts of testing model assumptions proposed in the literature (see González-Manteiga and Cao 1993; Dette 1999; Gozalo and Linton 2001). The main difference between our approach and the work of the last-named authors is that we are able to find the asymptotic properties of the tests under any fixed alternative of non-additivity. By way of an application, we identify a most efficient procedure in the class of tests based on the kernel method by looking at the asymptotic distribution under any fixed alternative. In Section 2 we give a motivation of the test statistic, while the main results are given in Section 3, which includes the corresponding results for several related tests. Section 4 contains a comprehensive comparison of the finite-sample performance of the different test statistics by means of a simulation study, which essentially reflects our asymptotic findings for moderate sample sizes. The proofs of our results, which are in the main rather cumbersome, are deferred to the Appendix.

2. Marginal integration revisited

Our main reason for using the marginal integration estimator for the construction of our test procedures is its direct definition, which allows an asymptotic treatment using central limit theorems for degenerate U statistics (see Zheng 1996; Hall 1984). A similar approach based on the backfitting estimator seems to be intractable, because our method would not require the asymptotic properties of the estimators of the additive regression function as recently derived by Opsomer and Ruppert (1997) and Mammen *et al.* (1999) but an explicit representation of the residuals from a fit by the backfitting estimate. On the other hand, Dette and Munk (1998) pointed out several drawbacks in the application of Fourier series estimation for checking model assumptions (see Section 5.2 of that paper) and we did not use series estimation for the construction of the test.

Let f denote the density of the explanatory variable $X = (X_1, \ldots, X_d)^{\mathrm{T}}$ with marginal densities f_α of X_α, $\alpha = 1, \ldots, d$. For a d-dimensional vector $x = (x_1, \ldots, x_d)$, let $x_{\underline{\alpha}}$ be the $(d-1)$-dimensional vector obtained by removing the αth coordinate from x, that is, $x_{\underline{\alpha}} = (x_1, \ldots, x_{\alpha-1}, x_{\alpha+1}, \ldots, x_d)$. If L^2_{add} denotes the subspace of addititive functions in the Hilbert space $L^2(f)$, we consider the projection P_0 from $L^2(f)$ onto L^2_{add} defined by

$$m_0(x) = (P_0 m)(x) = \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) - (d-1)c, \tag{2.1}$$

where

$$m_\alpha(x_\alpha) = \int m(x_\alpha, x_{\underline{\alpha}}) f_{\underline{\alpha}}(x_{\underline{\alpha}}) \mathrm{d}x_{\underline{\alpha}} = \int m(x_1, \ldots, x_{\alpha-1}, x_\alpha, x_{\alpha+1}, \ldots, x_d) f_{\underline{\alpha}}(x_{\underline{\alpha}}) \mathrm{d}x_{\underline{\alpha}}, \tag{2.2}$$

$$c = \int m(t) f(t) \mathrm{d}t. \tag{2.3}$$

Here we have used the notation

$$f_{\underline{\alpha}}(t_{\underline{\alpha}}) = \int f(t_1, \ldots, t_{\alpha-1}, t_\alpha, t_{\alpha+1}, \ldots, t_d)\mathrm{d}t_\alpha$$

and write in (2.2), with some abuse of terminology, $x = (x_\alpha, x_{\underline{\alpha}})$ to highlight the particular coordinate x_α. The representation (2.1) can be rewritten as

$$m_0(x) = C + \sum_{\alpha=1}^{d} k_\alpha(x_\alpha),$$

where

$$C = c + \sum_{\alpha=1}^{d} \left\{ \int m(t_\alpha, t_{\underline{\alpha}}) f_\alpha(t_\alpha) f_{\underline{\alpha}}(t_{\underline{\alpha}}) \mathrm{d}t_\alpha \mathrm{d}t_{\underline{\alpha}} - c \right\}$$

and

$$k_\alpha(x_\alpha) = m_\alpha(x_\alpha) - \int m(t_\alpha, t_{\underline{\alpha}}) f_\alpha(t_\alpha) f_{\underline{\alpha}}(t_{\underline{\alpha}}) \mathrm{d}t_\alpha \mathrm{d}t_{\underline{\alpha}},$$

which corresponds to the normalization given in Section 1. Note that P_0 is not necessarily an orthogonal projection with respect to the Hilbert space $L^2(f)$, where f is the joint density of X. However, one can easily verify that it is an orthogonal projection in the case of independent predictors.

Unless otherwise stated, let $K_i(\cdot)$, $i = 1, 2$, denote one- and $(d-1)$-dimensional Lipschitz continuous kernels of order 2 and $q \geqslant d$ respectively, with compact support, and define, for a bandwidth $h_i > 0$, $t_1 \in \mathbb{R}$, $t_2 \in \mathbb{R}^{d-1}$,

$$K_{1,h_1}(t_1) = \frac{1}{h_1} K_1\left(\frac{t_1}{h_1}\right), \qquad K_{2,h_2}(t_1) = \frac{1}{h_2^{d-1}} K_2\left(\frac{t_2}{h_2}\right). \tag{2.4}$$

For an independent and identically distributed sample $(X_i, Y_i)_{i=1}^n$, $X_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}}$, we consider the empirical counterparts of the components of m_0 in (2.1):

$$\hat{m}_\alpha(x_\alpha) = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{j=1}^{n} \frac{K_{1,h_1}(X_{j\alpha} - x_\alpha) K_{2,h_2}(X_{j\underline{\alpha}} - X_{k\underline{\alpha}})}{\hat{f}^{(\alpha)}(x_\alpha, X_{k\underline{\alpha}})} \cdot Y_j, \tag{2.5}$$

$$\hat{c} = \frac{1}{n} \sum_{j=1}^{n} Y_j, \tag{2.6}$$

where

$$\hat{f}^{(\alpha)}(x_\alpha, x_{\underline{\alpha}}) = \frac{1}{n} \sum_{i=1}^{n} K_{1,h_1}(X_{i\alpha} - x_\alpha) K_{2,h_2}(X_{i\underline{\alpha}} - x_{\underline{\alpha}}) \tag{2.7}$$

is an estimator of the joint density of X. Note that

$$\hat{m}_\alpha(x_\alpha) = \frac{1}{n} \sum_{j=1}^{n} \tilde{m}^{(\alpha)}(x_\alpha, X_{j\underline{\alpha}}),$$

where

$$\tilde{m}^{(\alpha)}(x_\alpha, x_{\underline{\alpha}}) = \frac{\frac{1}{n}\sum_{j=1}^{n} K_{1,h_1}(X_{j\alpha} - x_\alpha) K_{2,h_2}(X_{j\underline{\alpha}} - x_{\underline{\alpha}}) Y_j}{\hat{f}^{(\alpha)}(x_\alpha, x_{\underline{\alpha}})} \tag{2.8}$$

is the Nadaraya–Watson estimator at the point $(x_\alpha, x_{\underline{\alpha}})$; see Nadaraya (1965) or Watson (1964). The marginal integration estimator of $m_0 = P_0 m$ is now defined by

$$\hat{m}_0(x) = \sum_{\alpha=1}^{d} \hat{m}_\alpha(x_\alpha) - (d-1)\hat{c}, \tag{2.9}$$

and the corresponding residuals are denoted by $\hat{e}_j = Y_j - \hat{m}_0(X_j)$, $j = 1, \ldots, n$. As a first test statistic we consider the U statistic

$$T_{0,n} = \frac{1}{n(n-1)} \sum_{i \neq j} L_g(X_i - X_j) \hat{e}_i \hat{e}_j \pi(X_i) \pi(X_j), \tag{2.10}$$

where L is a d-dimensional symmetric kernel of order 2 with compact support, $L_g(\cdot) = (1/g^d) L_g(\cdot/g)$, $g > 0$, an additional bandwidth and π a given continuous weight function. We note that this type of statistic was originally introduced by Zheng (1996) in the problem of testing linearity of the regression, and independently discussed by Gozalo and Linton (2001) in the problem of testing additivity in a more general context. A theoretical justification for the application of this statistic to testing additivity will be given in Section 3. For a heuristic argument at this point we replace the residuals $\hat{e}_i$ by $\Delta(X_i) = m(X_i) - m_0(X_i)$ in $T_{0,n}$ and obtain from results of Hall (1984) or Zheng (1996) that in this case the corresponding statistic

$$V_{6n} = \frac{1}{n(n-1)} \sum_{i \neq j} L_g(X_i - X_j) \Delta(X_i) \Delta(X_j) \pi(X_i) \pi(X_j) \tag{2.11}$$

converges with limit

$$\begin{aligned} \mathrm{E}[V_{6n}] &= \int L_g(x - y) \Delta(x) \Delta(y) f(x) f(y) \pi(x) \pi(y) \mathrm{d}x \, \mathrm{d}y \qquad (2.12) \\ &= \int [m(x) - m_0(x)]^2 f^2(x) \pi^2(x) \mathrm{d}x + o(1). \end{aligned}$$

For this reason a test of the classical hypothesis of additivity can be obtained by rejecting (1.2) for large values of $T_{0,n}$.

There are several alternative ways of defining an appropriate statistic for the problem of testing additivity:

$$T_{1,n} = \frac{1}{n}\sum_{i=1}^{n}[\hat{m}(X_i) - \hat{m}_0(X_i)]^2\pi(X_i),$$

$$T_{2,n} = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i[\hat{m}(X_i) - \hat{m}_0(X_i)]\pi(X_i), \tag{2.13}$$

$$T_{3,n} = \frac{1}{n}\sum_{i=1}^{n}[\hat{e}_i^2 - \hat{d}_i^2]\pi(X_i).$$

Here $\hat{m}$ is the Nadaraya–Watson estimator with kernel L, and $\hat{d}_i = Y_i - \hat{m}(X_i)$ denotes the corresponding residual. The estimate $T_{1,n}$ compares a completely nonparametric fit with the marginal integration estimate and extends concepts of González-Manteiga and Cao (1993) and Härdle and Mammen (1993) to the problem of testing additivity. $T_{3,n}$ is essentially a (weighted) difference of estimators for the integrated variance function in the additive and non-restricted model. This concept was firstly proposed by Dette (1999) in the context of testing parametric structures of the regression function; see also Azzalini and Bowman (1993) for a similar statistic based on residuals. Finally, the statistic $T_{2,n}$ was introduced by Gozalo and Linton (2001), motivated by Lagrange multiplier tests of classical statistics.

In the following section we investigate the asymptotic behaviour of these statistics under the hypothesis (1.2) and fixed alternatives. We note that the asymptotic results under the null hypothesis of additivity have been independently found in a slightly more general context by Gozalo and Linton (2001) using different techniques in the proofs. It is the main purpose of the present paper to show that the asymptotic behaviour of the statistics $T_{j,n}$, $j = 0, \ldots, 3$, under fixed alternatives is rather different and to demonstrate potential applications of such results.

Remark 2.1. Several authors have proposed modifications of the marginal integration estimator; see the discussion of variance minimization in Fan *et al*. (1998), the definition of the efficient estimator in Linton (1997; 2000) or the application of local polynomials to bias reduction in Severance-Lossin and Sperlich (1999). It is worthwhile mentioning that the results in Theorem 3.2 and Theorem 3.5 remain valid with slight modifications of the asymptotic bias and variance terms. This follows by a careful inspection of the proof in the appendix, albeit with a substantial increase in algebraic complexity.

3. Main results and a comparison

We start with a detailed discussion of the asymptotic behaviour of the statistic $T_{0,n}$ and its consequences for the problem of testing additivity. Then the corresponding results for the statistics $T_{1,n}$, $T_{2,n}$, $T_{3,n}$ will be briefly stated and the different methods compared. In order to state and prove our main results we need a few regularity assumptions.

Assumption 1. *The explanatory variable X has a density f supported on $Q = [0, 1]^d$. f is bounded from below by a positive constant $c > 0$ and has continuous partial derivatives of order $q \geqslant d$.*

Assumption 2. *$m \in C_b^q(Q)$, where $C_b^q(Q)$ denotes the class of bounded functions (defined on Q) with continuous partial derivatives of order q.*

Assumption 3. *$\sigma \in C_b(Q)$, where $C_b(Q)$ denotes the class of bounded continuous functions (defined on Q).*

Assumption 4. *The distribution of the error has finite fourth moment,* $\mathrm{E}[\varepsilon^4] < \infty$.

Assumption 5. *As $n \to \infty$, the bandwidths g, h_1, $h_2 > 0$ satisfy*

$$h_1 \sim n^{-1/5}, \qquad h_2^q = o(h_1^2), \qquad \frac{\log n}{n h_1 h_2^{d-1}} = o(h_1^2), \qquad g^d = o(h_1^2), \qquad n g^d \to \infty.$$

Note that the optimal order for a twice continuously differentiable regression function $h_1 \sim n^{-1/5}$ in Assumption 5 requires $q > d - 1$ in order to satisfy

$$h_2^q = o(h_1^2) \quad \text{and} \quad \frac{\log n}{n h_1 h_2^{d-1}} = o(h_1^2)$$

simultaneuously. Our first result specifies the asymptotic distribution of the statistic $T_{0,n}$ under the null hypothesis of additivity.

Theorem 3.1. *If Assumptions* 1–5 *and the hypothesis of additivity are satisfied, then the statistic $T_{0,n}$ defined in* (2.10) *is asymptotically normally distributed, that is,*

$$n g^{d/2} T_{0,n} \xrightarrow{\mathscr{D}} N(0, \lambda_0^2), \tag{3.1}$$

where the asymptotic variance is given by

$$\lambda_0^2 = 2 \int L^2(x) \mathrm{d}x \int \sigma^4(x) \pi^4(x) f^2(x) \mathrm{d}x \tag{3.2}$$

and L is the d-dimensional kernel used in the definition of $T_{0,n}$.

Note that Theorem 3.1 has been found independently by Gozalo and Linton (2001) and provides a test for the hypothesis of additivity by rejecting H_0 for large values of $T_{0,n}$, that is,

$$n g^{d/2} T_{0,n} > u_{1-\alpha} \hat{\lambda}_{0,n}, \tag{3.1}$$

where $u_{1-\alpha}$ denotes the $1 - \alpha$ quantile of the standard normal distribution and $\hat{\lambda}_{0,n}$ is an appropriate estimator of the limiting variance (3.2). A simple estimator could be obtained by similar arguments to those given in Zheng (1996):

$$\hat{\lambda}_{0,n}^2 = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} L_g^2(X_i - X_j) \hat{e}_i^2 \hat{e}_j^2 \pi^2(X_i) \pi^2(X_j).$$

Our next result discusses the asymptotic behaviour of the statistic $T_{0,n}$ under a fixed alternative and proves – as a by-product – consistency of the test (3.3). On the other hand, it also provides the interesting possibility of an alternative formulation of the classical hypothesis of additivity, which will be described at the end of this section.

Theorem 3.2. *If Assumptions* 1–5 *are satisfied and the regression is not additive,* $\Delta = m - P_0 m \neq 0$, *then*

$$\sqrt{n}\{T_{0,n} - \mathrm{E}[T_{0,n}]\} \xrightarrow{\mathscr{D}} N(0, \mu_0^2). \tag{3.4}$$

Here

$$\mathrm{E}[T_{0,n}] = \mathrm{E}(\Delta^2 \pi^2 f(X_1)) - 2\mathrm{E}[\Delta \pi^2 f(X_1) \cdot b(X_1)] \cdot h_1^2 + o(h_1^2) + O(g^2), \tag{3.5}$$

with $b(x) = \sum_{\alpha=1}^{d} b_\alpha(x_\alpha)$, *in which*

$$b_\alpha(x_\alpha) = c_2(K_1) \int \left(\frac{1}{2} \frac{\partial^2 m}{\partial x_\alpha^2} + \frac{1}{f} \frac{\partial f}{\partial x_\alpha} \frac{\partial m}{\partial x_\alpha} \right) (x_\alpha, t_{\underline{\alpha}}) f_{\underline{\alpha}}(t_{\underline{\alpha}}) \mathrm{d}t_{\underline{\alpha}}, \tag{3.6}$$

where $c_2(K_1) = \int t_1^2 K_1(t_1) \mathrm{d}t_1$. *The asymptotic variance is given by*

$$\mu_0^2 = 4\mathrm{E}[\sigma^2(X_1)\{P_1(\Delta \pi^2 f)(X_1)\}^2] \tag{3.7}$$

$$+ 4 \operatorname{var}\left[(\Delta^2 \pi^2 f)(X_1) - \mathrm{E}\left(\Delta \pi^2 f(X_2) \left\{ \sum_{\alpha=1}^{d} m(X_{2\alpha}, X_{1\underline{\alpha}}) - (d-1) m(X_1) \right\} | X_1 \right) \right],$$

where $P_1 m = m - P_0^* m$, *in which the mapping* P_0^* *is defined by*

$$P_0^* g(x) = \sum_{\alpha=1}^{d} \frac{f_{\underline{\alpha}}(x_{\underline{\alpha}})}{f(x)} \int (gf)(x_\alpha, t_{\underline{\alpha}}) \mathrm{d}t_{\underline{\alpha}} - (d-1) \int (gf)(t) \mathrm{d}t. \tag{3.8}$$

Remark 3.3. Note that the mapping P_0^* defined in (3.8) is not a projection on the space of additive functions. In the case of independent predictors one can easily show that $P_0^* = P_0$. Moreover, if additionally the weight function is given by $\pi = 1/\sqrt{f}$, the asymptotic variance in (3.7) simplifies to

$$\mu_0^2 = 4\mathrm{E}[\sigma^2(X_1)\Delta^2(X_1)] + 4 \operatorname{var}[\Delta^2(X_1)],$$

where $\Delta = m - m_0$.

Remark 3.4. A careful analysis of the proof of Theorem 3.2 shows that for a sufficiently smooth regression and kernels L and K_i, $i = 1, 2$, of sufficiently high order we have

$$\mathrm{E}[T_{0,n}] = \mathrm{E}[\Delta^2(X_1)(\pi^2 f)(X_1)] + o\left(\frac{1}{\sqrt{n}} \right),$$

where the term $M^2 := \mathrm{E}[\Delta^2(X_1)(\pi^2 f)(X_1)]$ on the right-hand side serves as a measure of additivity. In this case Theorem 3.2 provides an interesting advantage to many of the commonly applied goodness-of-fit tests which will be explained in the following. It is well known that for model checks the type II error of a test is more important than the type I error, because, in the case of acceptance of the null hypothesis, the subsequent data analysis is adapted to the assumed model. From Theorem 3.2 we obtain as an approximation for the probability of the type II error of the test (3.3),

$$P(\text{rejection}) \approx \Phi\left(\sqrt{n}\frac{M^2}{\mu_0} - \frac{u_{1-\alpha}}{\sqrt{ng^d}}\frac{\lambda_0}{\mu_0}\right),$$

where $u_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution. On the other hand, the result can also be used for testing precise hypotheses (see Berger and Delampady 1987) of the form (1.3). Finally, we note that Theorem 3.2 could also be used for the construction of confidence intervals for the measure of additivity M^2.

Theorem 3.5. *Suppose that Assumptions* 1–5 *are satisfied and that* $T_{1,n}$, $T_{2,n}$, $T_{3,n}$ *are as defined in* (2.13).

(i) Under the hypothesis of additivity we have

$$ng^{d/2}\{T_{j,n} - \mathrm{E}_{H_0}[T_{j,n}]\} \xrightarrow{\mathscr{D}} N(0, \lambda_j^2), \quad j = 1, \ldots, 3,$$

where

$$B_1 = \mathrm{E}_{H_0}[T_{1,n}] = \frac{1}{ng^d}\int L^2(x)\mathrm{d}x \int \sigma^2(x)\pi(x)\mathrm{d}x + o\left(\frac{1}{ng^d}\right),$$

$$B_2 = \mathrm{E}_{H_0}[T_{2,n}] = \frac{1}{ng^d}L(0)\int \sigma^2(x)\pi(x)\mathrm{d}x + o\left(\frac{1}{ng^d}\right),$$

$$B_3 = \mathrm{E}_{H_0}[T_{1,n}] = \frac{1}{ng^d}\left(2L(0) - \int L^2(x)\mathrm{d}x\right)\int \sigma^2(x)\pi(x)\mathrm{d}x + o\left(\frac{1}{ng^d}\right)$$

and

$$\lambda_1^2 = 2\int \sigma^4(x)\pi^2(x)\mathrm{d}x \int (L * L)^2(x)\mathrm{d}x,$$

$$\lambda_2^2 = 2\int \sigma^4(x)\pi^2(x)\mathrm{d}x \int L^2(x)\mathrm{d}x,$$

$$\lambda_3^2 = 2\int \sigma^4(x)\pi^2(x)\mathrm{d}x \int (2L - (L * L))^2(x)\mathrm{d}x,$$

in which $f * g$ *denotes the convolution of the functions* f *and* g.

(ii) If the regression is not additive, $\Delta = m - m_0 \neq 0$, *then*

$$\sqrt{n}\{T_{j,n} - \mathrm{E}_{H_1}[T_{j,n}]\} \xrightarrow{\mathcal{D}} N(0, \mu_j^2), \quad j = 1, \ldots, 3,$$

where

$$\begin{aligned}
\mathrm{E}_{H_1}[T_{1,n}] &= B_1 + \nu_0 - 2\nu_1 + 2\nu_2,\\
\mathrm{E}_{H_1}[T_{2,n}] &= B_2 + \nu_0 - 2\nu_1 + \nu_2,\\
\mathrm{E}_{H_1}[T_{3,n}] &= B_3 + \nu_0 - 2\nu_1,\\
\nu_0 &= \mathrm{E}[(\Delta^2\pi)(X_1)],\\
\nu_1 &= \mathrm{E}[(\Delta\pi)(X_1)b(X_1)] \cdot h_1^2 + o(h_1^2),\\
\nu_2 &= \mathrm{E}[(\Delta\pi)(X_1)b_{\mathrm{NW}}(X_1)] \cdot g^2 + o(g^2),
\end{aligned}$$

b is defined in Theorem 3.2*, b_{NW} is the bias of the Nadaraya–Watson estimate, the asymptotic variances are given by*

$$\begin{aligned}
\mu_j^2 &= 4\mathrm{E}[\sigma^2(X_1)\{P_1(\Delta\pi)(X_1)\}^2]\\
&\quad + \mathrm{var}\left[(\Delta^2\pi)(X_1) - 2\mathrm{E}\left(\Delta\pi(X_2)\left\{\sum_{\alpha=1}^{d} m(X_{2\alpha}, X_{1\underline{\alpha}}) - (d-1)m(X_1)\right\}|X_1\right)\right],
\end{aligned}$$

$j = 1, \ldots, 3$ and the mapping P_1 is defined in Theorem 3.2.

In the remaining part of this section we will use Theorems 3.2 and 3.5 to compare the tests of additivity induced by the statistics $T_{j,n}$, $j = 0, \ldots, 3$. For the sake of a transparent presentation we assume for this comparison a sufficient smoothness for the regression and sufficiently large order for the kernel, such that the asymptotic bias of $T_{j,n}$ under a fixed alternative is given by

$$\mathrm{E}_{H_1}[T_{j,n}] = M_j^2 + B_j + o\left(\frac{1}{\sqrt{n}}\right), \qquad j = 0, \ldots, 3,$$

where $B_0 = 0$, B_1, B_2, B_3 are defined in Theorem 3.5, and

$$\begin{aligned}
M_0^2 &= \mathrm{E}[\Delta^2(X_1)(\pi^2 f)(X_1)],\\
M_j^2 &= \mathrm{E}[\Delta^2(X_1)\pi(X_1)], \qquad j = 1, \ldots, 3.
\end{aligned}$$

In this case the probability of rejection is approximately given by

$$P(\text{rejection}) \approx \Phi\left(\frac{1}{\mu_j}\left\{\sqrt{n}M_j^2 - \frac{u_{1-\alpha}\lambda_j}{\sqrt{ng^d}}\right\}\right), \qquad j = 0, \ldots, 3, \tag{3.9}$$

where μ_j, λ_j are as defined in Theorems 3.1, 3.2 and 3.5. From this representation we see that, in general, there is no clear recommendation for one of the statistics $T_{j,n}$. The appropriate choice of a test depends sensitively on the relation between the variance function

σ, weight function π, regression m and alternative Δ. A fair comparison seems to be possible by adjusting with respect to the measure of additivity. This can be done by replacing the weight function π in $T_{0,n}$ by $\pi/\sqrt{f}$ (in practice, an estimator of f has to be used), which gives

$$M_j^2 = \mathrm{E}[\Delta^2(X_1)\pi(X_1)], \qquad j = 0, \ldots, 3,$$

and (by the definition of μ_j^2 in Theorems 3.2 and 3.5)

$$\mu_0^2 > \mu_j^2, \qquad j = 1, \ldots, 3. \tag{3.10}$$

Looking at the dominating term in (3.9), we thus obtain that (asymptotically) tests based on the statistics $T_{j,n}$, $j = 1, \ldots, 3$, will be more powerful than the test based on the statistic $T_{0,n}$. We note, however, that for realistic sample sizes this improvement will only be substantial if the variance function is 'small' compared to the deviation Δ of the additive approximation from the model. For a comparison of the remaining statistics, observe that for the corresponding tests the terms with factor $\sqrt{n}$ in (3.9) are identical and, consequently, a most efficient procedure is obtained by minimizing the variance λ_j^2 of the asymptotic distribution under the null hypothesis of additivity. This comparison coincides with the concept of considering local alternatives which converge to the null hypothesis at a rate $(ng^{d/2})^{-1/2}$. The following lemma shows that the statistics $T_{1,n}$ and $T_{2,n}$ should be prefered to $T_{3,n}$ with respect to this criterion. This result was also conjectured by Gozalo and Linton (2001) without proof. A rigorous derivation will be given at the end of the Appendix.

Lemma 3.6. *If K is an arbitrary density, we have*

$$\int (K * K)^2(x)\mathrm{d}x \leqslant \int K^2(x)\mathrm{d}x \leqslant \int (2K - K * K)^2(x)\mathrm{d}x \tag{3.11}$$

or, equivalently,

$$\lambda_1^2 \leqslant \lambda_2^2 \leqslant \lambda_3^2.$$

We finally note that the arguments in favour of $T_{1,n}$ and $T_{2,n}$ are only based on the discussion of the asymptotic variances, which is correct from an asymptotic point of view. For realistic sample sizes, however, the bias has to be taken into account. Here we observe exactly the opposite behaviour, namely that the statistic $T_{0,n}$ is preferable because its standardized version has no bias converging to infinity. The simulation results presented in Section 4 indicate that the asymptotic arguments in favour of $T_{1,n}$, $T_{2,n}$ are valid for sample sizes $N \geqslant 100$ and 'small' variances of the error distribution.

Remark 3.7. Note that Gozalo and Linton (2001) study the asymptotic distribution of the statistics $T_{j,n}$, $j = 0, \ldots, 3$, under the null hypothesis of additivity in the context of generalized nonparametric regression models including discrete covariates. The results of the present paper can also be extended to this more general situation at the cost of some additional notation. For the sake of a simple notation we have not formulated the results in

full detail, but indicate the generalization of Theorems 3.1 and 3.2 in the situation of a known link function as considered in Linton and Härdle (1996). In the nonparametric regression model

$$\mathrm{E}[Y|X = x] = m(x),$$

we are interested in testing the hypothesis

$$H_0^G : G(m(x)) = C + \sum_{\alpha=1}^{d} k_\alpha(x_\alpha),$$

where G is a given link function. The definition of the marginal integration estimator of m is straightforward (see, for example, Linton and Härdle 1996). To be precise, let

$$\tilde{m}_\alpha(x_\alpha) = \frac{1}{n}\sum_{i=1}^{n} G(\tilde{m}^{(\alpha)}(x_\alpha, X_{i\underline{\alpha}}))$$

denote the estimator of

$$\int G(m(x_\alpha, x_{\underline{\alpha}}))f_{\underline{\alpha}}(t_{\underline{\alpha}})\mathrm{d}t_{\underline{\alpha}},$$

where $\tilde{m}^{(\alpha)}$ is defined in (2.8). Furthermore, let

$$\hat{c} = \frac{1}{d}\sum_{\alpha=1}^{d}\frac{1}{n}\sum_{i=1}^{m} G(\tilde{m}^{(\alpha)}(X_{i\alpha}, X_{i\underline{\alpha}}))$$

denote an estimator of $\int G(m(x))f(x)\mathrm{d}x$. Defining

$$\hat{m}_0(x) = \sum_{\alpha=1}^{d}\tilde{m}_\alpha(x_\alpha) - (d-1)\hat{c},$$

the marginal integration estimator of the regression function m is obtained as

$$\hat{m}(x) = F(\hat{m}_0(x)), \tag{3.12}$$

where $F = G^{-1}$ is the inverse of the link function. The statistic $T_{0,n}$ is now exactly defined as in (2.10) (with residuals obtained from (3.12)), and under the hypothesis H_0^G and certain regularity assumptions for the link function (see, for example, Linton and Härdle 1996; Gozalo and Linton 2001) Theorem 3.1 remains valid. On the other hand, under a fixed alternative $\sqrt{n}(T_{0,n} - \mathrm{E}[T_{0,n}])$ is asymptotically normal, with asymptotic variance given by

$$\mu_0^2 = 4\mathrm{E}[\sigma^2(X_1)P_1^G(\Delta\pi^2)(X_1)]$$
$$+ 4\,\mathrm{var}\left[(\Delta\pi)^2(X_1)f(X_1) - \mathrm{E}\left((\Delta\pi^2 f)(X_2)\left\{\sum_{\alpha=1}^{d} G(m(X_{2\alpha}, X_{1\underline{\alpha}})) - (d-1)G(m(X_1))\right\}|X_1\right)\right],$$

where $\sigma^2(x) = \mathrm{var}[Y|X = x]$ denotes the conditional variance of the response, $\Delta = m - Fm_0$, $m_0 = P_0 \circ G \circ m$, P_0 is the projection defined in (2.1), $P_1^G = I - P_0^G$ and the mapping P_0^G is defined by

$$(P_0^G g)(x) = G'(m(x))\left\{\sum_{\alpha=1}^{d}\frac{f_{\underline{\alpha}}(x_{\underline{\alpha}})}{f(x)}\int (gf)(x_\alpha, t_{\underline{\alpha}})F'(m_0(x_\alpha, t_{\underline{\alpha}}))\mathrm{d}t_{\underline{\alpha}} - (d-1)\int (gf)(t)F'(m_0(t))\mathrm{d}t\right\}.$$

The proof of this result follows essentially the steps given in the Appendix, observing that for a smooth link function the residuals are given by

$$\begin{aligned} Y_i - \hat{m}(X_i) &= Y_i - m(X_i) + m(X_i) - F(m_0(X_i)) - \{F(\hat{m}_0(X_i)) - F(m_0(X_i))\} \\ &\approx Y_i - m(X_i) + \Delta(X_i) - F'(m_0(X_i))\{\hat{m}_0(X_i) - m_0(X_i)\}. \end{aligned}$$

Therefore in the analysis of the statistic $T_{0,n}$ the terms $V_{1,n}$, V_{4n}, V_{6n} (see the proof in the Appendix) are treated exactly in the same way as for $G(x) = x$. For the remaining terms one uses a careful analysis of the proof in the appendix and a further Taylor expansion of $\hat{m}_0(X_i) - m_0(X_i)$ which yields the additional terms $G'(m(X_1))$ in the asymptotic variance.

4. A finite-sample comparison

In order to investigate qualitatively the finite-sample performance of the different procedures we have conducted a small simulation study. Consider the bivariate regression model

$$Y_i = m(X_{i1}, X_{i2}) + \sigma\varepsilon_i, \qquad i = 1, \ldots, n,$$

where the (X_{i1}, X_{i2}), $i = 1, \ldots, n$, are assumed to be independent and uniformly distributed on the unit square $[0, 1]^2$; the ε_i, $i = 1, \ldots, n$, are independent, standard normally distributed, and independent of the (X_{i1}, X_{i2}), $i = 1, \ldots, n$; and $\sigma = 0.1$. For the kernel in all estimators we use the Epanechnikov kernel

$$K(t) = \tfrac{3}{4}(1 - t^2)\, I_{[-1,1]}(t),$$

and a product of two kernels of this type as a two-dimensional kernel. In similar problems it has been observed by several authors (for example, Azzalini and Bowman 1993; Hjellvik and Tjøstheim 1995; Alcalá *et al.* 1999) that the asymptotic normal distribution under the null hypothesis does not provide a satisfactory approximation for the distribution of the statistics $T_{j,n}$, $j = 0, \ldots, 3$. For these reasons most authors propose the application of the wild bootstrap in this context (see, for example, Härdle and Mammen 1993; Hjellvik and Tjøstheim 1995). It is worthwhile mentioning that the approximation by the limiting distribution under a fixed alternative is comparable with the classical central limit theorem (see the proof of Theorem 3.2 in the Appendix) and is therefore more accurate compared to the approximation under the null hypothesis. Nevertheless, the asymptotic distribution in this case depends on certain features of the data-generating process, which are difficult to estimate except in rare circumstances. For this reason we also recommend the application of the wild bootstrap for testing precise hypotheses of the form (1.3). Note that the calculation of the test statistics $T_{j,n}$, $j = 1, \ldots, 3$, requires the specification of the bandwidths h_1 and h_2 for estimation the regression function under the null hypothesis (the marginal integration estimate $\hat{m}_0$ defined in (2.9)) and the bandwidth h appearing in the Nadaraya–Watson

estimator $\hat{m}$. This choice is based on a cross-validation procedure in a preliminary simulation under the null hypothesis of an addititve model $m(x_1, x_2) = x_1 + x_2$ using the bandwidths

$$h_1 = h_2 = \gamma_1 n^{-1/5}, \qquad h = \gamma n^{-1/6}.$$

This minimization yields $\gamma_1 = 0.4$, $\gamma = 0.44$, which was used throughout this study. The statistic $T_{0,n}$ requires the specification of a further bandwidth g, which was chosen as $0.2\, n^{-2/5}$. The weight function π is used to exclude boundary effects and is given by

$$\pi(x_1, x_2) = (1 - 2\delta)^{-2} I_{[\delta, 1-\delta]^2}(x_1, x_2),$$

where $\delta = 0.05$. For the resampling we used the wild bootstrap (see Wu 1986; Härdle and Mammen 1993), where

$$Y_i^* = \hat{m}_0(X_{i1}, X_{i2}) + \varepsilon_i^*,$$

$$\varepsilon_i^* = (u_i(1 - \sqrt{5})/2 + (1 - u_i)(1 - \sqrt{5})/2)\hat{e}_i,$$

$\hat{e}_i$ is defined in Section 2 and the u_i, $i = 1, \ldots, n$, are independent and identically distributed random variables with *Bernoulli*(p) distribution independent of the original sample with $p = (5 + \sqrt{5})/10$. The hypothesis of additivity is rejected if $T_{k,n} \geqslant t^*_{k,n,1-\alpha}$, where $t^*_{k,n,1-\alpha}$ denotes the critical value obtained from the bootstrap distribution,

$$P^*(T^*_{k,n} \geqslant t^*_{k,n,1-\alpha}) = 1 - \alpha, \qquad k = 0, 1, 2, 3,$$

where P^* denotes the conditional distribution given the sample (Y_i, X_i), $i = 1, \ldots, n$. The number of bootstrap replications for the estimation of $t^*_{k,n,1-\alpha}$ was chosen as $B = 500$. We have simulated the rejection probabilities of these tests for different models on the basis of 500 replications of each experiment. We considered the models

$$m(x_1, x_2) = x_1 + x_2 + a x_1 x_2, \tag{4.1a}$$

$$m(x_1, x_2) = (x_1 + x_2)^b, \tag{4.1b}$$

$$m(x_1, x_2) = \sin(c\pi(x_1 + x_2)), \tag{4.1c}$$

where the parameters a, b and c specify the deviation from the null hypothesis of additivity. The corresponding results are depicted in Tables 4.1–4.3 for sample sizes $n = 100$ and $n = 200$.

We observe a reasonable approximation of the level by all test procedures, with only slight advantages for the statistic $T_{3,n}$. A comparison of the power shows larger differences and a similar picture in all considered cases. The test based on $T_{0,n}$ – Zheng's (1996) approach – yields substantial smaller rejection probabilities in all cases considered in our simulation study, which confirms our asymptotic findings of Section 3 (note that the variance is small and that we used a uniform distribution corresponding to the case considered in the comparison (3.10)). A comparison of the remaining statistics shows that the test based on $T_{3,n}$ has lower power than the procedures based on $T_{1,n}$ and $T_{2,n}$, in accordance with our asymptotic findings in Lemma 3.6. Finally, we note that the power behaviour of the tests based on $T_{1,n}$ and $T_{2,n}$ is very similar, which is also in agreement

Table 4.1. Simulated rejection probabilities of the wild bootstrap tests based on the statistics $T_{i,n}(i = 0, 1, 2, 3)$ in model (4.1a) for the 5% and 10% level and various alternatives. The variance is constant and given by $\sigma^2 = 0.01$, and the choice $a = 0$ corresponds to the additive model

		$T_{0,n}$		$T_{1,n}$		$T_{2,n}$		$T_{3,n}$	
n	a	5%	10%	5%	10%	5%	10%	5%	10%
	0	0.046	0.166	0.064	0.128	0.052	0.110	0.054	0.118
100	1	0.278	0.494	0.866	0.920	0.816	0.904	0.698	0.788
	2	0.494	0.780	0.976	0.984	0.976	0.990	0.948	0.974
	0	0.046	0.122	0.040	0.070	0.042	0.094	0.054	0.104
200	1	0.438	0.658	0.998	0.998	0.996	0.998	0.986	0.990
	2	0.896	0.986	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.2. Simulated rejection probabilities of the wild bootstrap tests based on the statistics $T_{i,n}(i = 0, 1, 2, 3)$ in model (4.1b) for the 5% and 10% level and various alternatives. The variance is constant and given by $\sigma^2 = 0.01$

		$T_{0,n}$		$T_{1,n}$		$T_{2,n}$		$T_{3,n}$	
n	b	5%	10%	5%	10%	5%	10%	5%	10%
	1/2	0.072	0.168	0.238	0.324	0.168	0.266	0.144	0.210
100	1/3	0.074	0.170	0.238	0.334	0.182	0.282	0.146	0.228
	1/4	0.058	0.168	0.196	0.300	0.166	0.250	0.132	0.226
	1/2	0.100	0.180	0.494	0.618	0.374	0.522	0.304	0.394
200	1/3	0.068	0.152	0.440	0.556	0.324	0.448	0.236	0.356
	1/4	0.060	0.160	0.362	0.484	0.250	0.398	0.188	0.292

Table 4.3. Simulated rejection probabilities of the wild bootstrap tests based on the statistics $T_{i,n}(i = 0, 1, 2, 3)$ in the model (4.1c) for the 5% and 10% level and various alternatives. The variance is constant and given by $\sigma^2 = 0.01$

		$T_{0,n}$		$T_{1,n}$		$T_{2,n}$		$T_{3,n}$	
n	c	5%	10%	5%	10%	5%	10%	5%	10%
	1/2	0.680	0.876	1.000	1.000	0.996	1.000	0.990	0.996
100	1	0.920	0.988	0.994	0.998	0.996	0.998	0.996	0.998
	2	0.982	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	1/2	0.972	0.990	1.000	1.000	1.000	1.000	1.000	1.000
200	1	0.994	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

with Lemma 3.6. Observing the more precise approximation of the level by the test based on $T_{2,n}$, we recommend the use of this approach for the problem of testing additivity.

Appendix A. Proofs

For the sake of a transparent notation we consider the case $d = 2$. In addition, we use $\pi(x) \equiv 1$ as our weight function; the general case is treated exactly in the same way. Because all results are essentially proved similarly, we restrict ourselves to a proof of the asymptotic behaviour of the statistic $T_{0,n}$ (that is, Theorem 3.1 and 3.2).

A.1. Proof of Theorem 3.1

Observing that under the hypothesis of additivity $m_0 = P_0 m = m$, we obtain from (1.1) the decomposition $\hat{e}_j = \sigma(X_j)\varepsilon_j - \delta(X_j)$, $\delta(x) = \hat{m}_0(x) - m_0(x)$ and

$$T_{0,n} = V_{1n} - 2V_{2n} + V_{3n}, \tag{A.1}$$

where

$$V_{1n} = \frac{1}{n(n-1)} \sum_{i \neq j} L_g(X_i - X_j)\sigma(X_i)\sigma(X_j)\varepsilon_i\varepsilon_j, \tag{A.2}$$

$$V_{2n} = \frac{1}{n(n-1)} \sum_{i \neq j} L_g(X_i - X_j)\sigma(X_i)\varepsilon_i\delta(X_j), \tag{A.3}$$

$$V_{3n} = \frac{1}{n(n-1)} \sum_{i \neq j} L_g(X_i - X_j)\delta(X_i)\delta(X_j). \tag{A.4}$$

The first term can be treated as in Zheng (1996) using the results of Hall (1984), and we obtain

$$ngV_{1n} \to N(0, \lambda_0^2), \tag{A.5}$$

where the variance λ_0^2 is defined in (3.2). The estimation of the remaining terms is more delicate.

With the notation $\delta(x) = \delta_1(x_1) + \delta_2(x_2) - \delta_0$, where

$$\delta_r(x_r) = \hat{m}_r(x_r) - m_r(x_r), \qquad r = 1, 2,$$

$$\delta_0 = \frac{1}{n} \sum_{k=1}^{n} Y_k - c, \tag{A.6}$$

we derive the decomposition

$$V_{2n} = V_{2n}^{(1)} + V_{2n}^{(2)} - V_{2n}^{(0)},$$

where

$$V_{2n}^{(r)} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_g(X_i - X_j)\sigma(X_i)\varepsilon_i \cdot \delta_r(X_{jr}), \qquad r = 1, 2,$$

and

$$V_{2n}^{(0)} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_g(X_i - X_j)\sigma(X_i)\varepsilon_i \cdot \delta_0.$$

First, we will show that

$$V_{2n}^{(r)} = O_P\left(\frac{1}{nh_1}\right), \qquad r = 1, 2.$$

Obviously it suffices just to treat the case $r = 1$. Recalling definition (2.5), we rewrite $\hat{m}_1(x_1)$ as

$$\hat{m}_1(x_1) = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} w_{kl}^{(1)}(x_1) \cdot Y_l,$$

where

$$w_{kl}^{(1)}(x_1) = \frac{K_{1,h_1}(X_{l1} - x_1)K_{2,h_2}(X_{l2} - X_{k2})}{\hat{f}^{(1)}(x_1, X_{k2})} \tag{A.7}$$

and $\hat{f}^{(1)}$ is defined in (2.7). Observing that

$$m_1(x_1) = \frac{1}{n} \sum_{k=1}^{n} m(x_1, X_{k2}) + O\left(\sqrt{\frac{\log\log n}{n}}\right) P\text{-a.s.}$$

(by the law of the iterated logarithm) we obtain

$$\begin{aligned} \delta_1(x_1) = {} & \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} w_{kl}^{(1)}(x_1) \cdot \sigma(X_l)\varepsilon_l \\ & + \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} w_{kl}^{(1)}(x_1) \cdot (m(X_{l1}, X_{l2}) - m(x_1, X_{k2})) + O\left(\sqrt{\frac{\log\log n}{n}}\right) \end{aligned} \tag{A.8}$$

(noting that $(1/n)\sum_{l=1}^{n} w_{kl}^{(1)}(x_1) = 1$) and

$$V_{2n}^{(1)} = (V_{2n}^{(1.1)} + V_{2n}^{(1.2)})(1 + o_P(1)),$$

where

$$V_{2n}^{(1.1)} = \frac{1}{n^3(n-1)} \sum_{i,k,l=1}^{n} \sum_{j \neq i} L_g(X_i - X_j)\sigma(X_i)\varepsilon_i w_{kl}^{(1)}(X_{j1}) \cdot \sigma(X_l)\varepsilon_l,$$

$$V_{2n}^{(1.2)} = \frac{1}{n^3(n-1)} \sum_{i,k,l=1}^{n} \sum_{j \neq i} L_g(X_i - X_j)\sigma(X_i)\varepsilon_i w_{kl}^{(1)}(X_{j1}) \cdot (m(X_{l1}, X_{l2}) - m(X_{j1}, X_{k2})).$$

Computing the expectation of the first term, we obtain

$$\mathrm{E}(V_{2n}^{(1.1)}) = \frac{1}{n^3(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} \mathrm{E}[L_g(X_i - X_j)\sigma^2(X_i) w_{ki}^{(1)}(X_{j1})].$$

Now, by definition (A.7),

$$\begin{aligned} \mathrm{E}(w_{ki}^{(1)}(X_{j1}) | X_i, X_j) &= K_{1,h_1}(X_{i1} - X_{j1}) \mathrm{E}\left(\frac{K_{2,h_2}(X_{i2} - X_{k2})}{\hat{f}^{(1)}(X_{j1}, X_{k2})} | X_i, X_j \right) \\ &= K_{1,h_1}(X_{i1} - X_{j1}) \mathrm{E}\left(\frac{K_{2,h_2}(X_{i2} - X_{k2})}{f(X_{j1}, X_{k2})} | X_i, X_j \right)(1 + o(1)), \end{aligned}$$

where the second equality is obtained by the strong uniform consistency of the kernel density estimate $\hat{f}^{(1)}$; see, for example, Silverman (1978). For $k \neq i, j$, Taylor expansion gives

$$\mathrm{E}\left(\frac{K_{2,h_2}(X_{i2} - X_{k2})}{f(X_{j1}, X_{k2})} | X_i, X_j \right) = \frac{f_2(X_{i2})}{f(X_{j1}, X_{i2})} + O(h_2^q),$$

and the boundedness of the density and the kernels K_1 and K_2 yields

$$\mathrm{E}(V_{2n}^{(1.1)}) = O\left(\frac{1}{nh_1} \right) + O\left(\frac{1}{n^2 h_1 h_2} \right),$$

where the O terms correspond to the cases $k \neq i, j$ and $k = i$ (or $k = j$), respectively.

Next we compute the variance of $V_{2n}^{(1.1)}$ by discussing the individual terms in the sum

$$\begin{aligned} (V_{2n}^{(1.1)})^2 = \frac{1}{n^6(n-1)^2} &\sum_{i,i'=1}^{n} \sum_{j \neq i, j' \neq i'} \sum_{k,k'=1}^{n} \sum_{l,l'=1}^{n} L_g(X_i - X_j)\sigma(X_i)\varepsilon_i w_{kl}^{(1)}(X_{j1})\sigma(X_l)\varepsilon_l \\ &\times L_g(X_{i'} - X_{j'})\sigma(X_{i'})\varepsilon_{i'} w_{k'l'}^{(1)}(X_{j'1})\sigma(X_{l'})\varepsilon_{l'}. \end{aligned}$$

The terms in this sum have expectation zero exept for the case when $i' = i$ and $l' = l$; $i' = l$ and $i = l'$; $i = l$ and $i' = l'$; or $i' = i = l' = l$.

Consider the first case, $i' = i$ and $l' = l$. Conditioning on X_i, X_l and taking the expectation of the corresponding terms yields

$$\frac{1}{n^6(n-1)^2} \sum_{i,l=1}^{n} \sum_{j \neq i, j' \neq i'} \sum_{k,k'=1}^{n} \mathrm{E}[\mathrm{E}(L_g(X_i - X_j) w_{kl}^{(1)}(X_{j1}) | X_i, X_l)^2 \sigma^2(X_i)\sigma^2(X_l)](1 + o(1)),$$

which is of order $O(1/n^2 h_1^2)$ by the same reasoning as above. The other cases are treated

in the same way, showing that $\mathrm{var}(V_{2n}^{(1.1)}) = O(1/n^2 h_1^2)$. It follows by Chebyshev's inequality that

$$V_{2n}^{(1.1)} = O_P\left(\frac{1}{nh_1}\right). \tag{A.9}$$

For the second term in the decomposition of $V_{2n}^{(1)}$ we obviously have

$$\mathrm{E}(V_{2n}^{(1.2)}) = 0.$$

In order to find the corresponding variance we note that

$$(V_{2n}^{(1.2)})^2 = \frac{1}{n^6(n-1)^2} \sum_{i,i'=1}^{n} \sum_{j\neq i, j'\neq i'} \sum_{k,k'=1}^{n} \sum_{l,l'=1}^{n} L_g(X_i - X_j)\sigma(X_i)\varepsilon_i L_g(X_{i'} - X_{j'})\sigma(X_{i'})\varepsilon_{i'}$$
$$\times w_{kl}^{(1)}(X_{j1})(m(X_{l1}, X_{l2}) - m(X_{j1}, X_{k2}))w_{k'l'}^{(1)}(X_{j'1})(m(X_{l'1}, X_{l'2}) - m(X_{j'1}, X_{k'2})). \tag{A.10}$$

If $i' = i$, and all other indices are pairwise different, we have, for the expectation of the corresponding terms in the sum (A.10),

$$\frac{1}{n}\mathrm{E}[\sigma^2(X_i)\mathrm{E}(L_g(X_i - X_j)\mathrm{E}(w_{kl}^{(1)}(X_{j1})(m(X_{l1}, X_{l2}) - m(X_{j1}, X_{k2}))|X_i, X_j)|X_i)^2]. \tag{A.11}$$

Using the strong uniform consistency of $\hat{f}$ again and the assumption $(\log n)/nh_1h_2 = o(h_1^2)$, we obtain, by a lengthy argument,

$$\mathrm{E}(w_{kl}^{(1)}(X_{j1})(m(X_{l1}, X_{l2}) - m(X_{j1}, X_{k2}))|X_i, X_j)$$
$$= \mathrm{E}\left(\frac{K_{1,h_1}(X_{l1} - X_{j1})K_{2,h_2}(X_{l2} - X_{k2})}{f(X_{j1}, X_{k2})}(m(X_{l1}, X_{l2}) - m(X_{j1}, X_{k2}))|X_j\right)(1 + o(1)),$$

which is asymptotically equal to

$$\left\{\mathrm{E}\left(\frac{K_{1,h_1}(X_{l1} - X_{j1})f_2(X_{l2})}{f(X_{j1}, X_{l2})}(m(X_{l1}, X_{l2}) - m(X_{j1}, X_{l2}))|X_j\right) + O(h_2^q)\right\}(1 + o(1))$$
$$= O(h_1^2) + O(h_2^q),$$

the O terms being independent of X_j. So the term (A.11) is of order

$$O\left(\frac{h_1^4 + h_2^{2q}}{n}\right) = O\left(\frac{1}{n^2 h_1}\right),$$

this equality being a consequence of Assumption 5. The terms in the sum (A.10) with $i' = i$ and $l' = l$ (all other indices pairwise different) have expectation

$$\frac{1}{n^2}\mathrm{E}[\sigma^2(X_i)\mathrm{E}(L_g(X_i - X_j)$$

$$\times \mathrm{E}(w_{kl}^{(1)}(X_{j1})(m(X_{l1}, X_{l2}) - m(X_{j1}, X_{k2}))|X_i, X_j, X_l)|X_i, X_l)^2]$$

$$= \frac{1}{n^2}\mathrm{E}\left[\sigma^2(X_i)\mathrm{E}\left(L_g(X_i - X_j)\right.\right.$$

$$\left.\left.\times K_{1,h_1}(X_{l1} - X_{j1})\left(\frac{f_2(X_{l2})}{f(X_{j1}, X_{l2})}(m(X_{l1}, X_{l2}) - m(X_{j1}, X_{l2})) + o(1)\right)|X_i, X_l\right)^2\right]$$

$$= O\left(\frac{1}{n^2 h_1^2}\right),$$

again by boundedness. By a similar argument for the remaining terms in the sum (A.10) we obtain the result

$$V_{2n}^{(1.2)} = O_P\left(\frac{1}{nh_1}\right) \tag{A.12}$$

Combining (A.9) and (A.12), we obtain

$$V_{2n}^{(1)} = O_P\left(\frac{1}{nh_1}\right)$$

Clearly, the same holds for $V_{2n}^{(2)}$. Finally, it is not hard to show that $V_{2n}^{(0)} = O_P(1/n)$, and a combination of these results gives

$$V_{2n} = O_P\left(\frac{1}{nh_1}\right).$$

It follows from Assumption 5 that

$$V_{2n} = o_P\left(\frac{1}{ng}\right). \tag{A.13}$$

Since calculations for the statistic

$$V_{3n} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_g(X_i - X_j)\delta(X_i)\delta(X_j)$$

are similar to those we have already done, we only state the estimates for its expectation and variance, which are

$$\mathrm{E}(V_{3n}) = O\left(h_1^4 + h_2^{2q} + \frac{1}{nh_1}\right), \tag{A.14}$$

$$\mathrm{var}(V_{3n}) = O\left(\frac{h_1^4 + h_2^{2q}}{nh_1} + \frac{1}{n^2 h_1^2}\right). \tag{A.15}$$

From (A.14) and (A.15) and Assumption 5 we obtain

$$V_{3n} = o_P\left(\frac{1}{ng}\right), \tag{A.16}$$

and the assertion of Theorem 3.1 follows from (A.1), (A.5), (A.13) and (A.16).

A.2. Proof of Theorem 3.2

If the regression is not additive we obtain a different decomposition of the residuals, that is,

$$\hat{e}_j = Y_j - \hat{m}_0(X_j) = \sigma(X_j)\varepsilon_j + \Delta(X_j) - \delta(X_j),$$

where $\delta = \hat{m}_0 - m_0$, $\Delta = m - P_0 m = m - m_0$. Therefore the corresponding decomposition of $T_{0,n}$ in (A.1) involves three additional terms,

$$T_{0,n} = V_{1n} - 2V_{2n} + V_{3n} + 2V_{4n} - 2V_{5n} + V_{6n}, \tag{A.17}$$

where V_{1n}, V_{2n}, V_{3n} are as defined in (A.2), (A.3), (A.4), respectively, and the remaining terms are given by

$$V_{4n} = \frac{1}{n(n-1)}\sum_{i\neq j} L_g(X_i - X_j)\Delta(X_j)\sigma(X_i)\varepsilon_i, \tag{A.18}$$

$$V_{5n} = \frac{1}{n(n-1)}\sum_{i\neq j} L_g(X_i - X_j)\Delta(X_j)\delta(X_i), \tag{A.19}$$

$$V_{6n} = \frac{1}{n(n-1)}\sum_{i\neq j} L_g(X_i - X_j)\Delta(X_i)\Delta(X_j). \tag{A.20}$$

From the proof of Theorem 3.1 and Assumption 5 (in the case $d = 2$) we have

$$\begin{aligned} V_{1n} &= O_P\left(\frac{1}{ng}\right) = o_P\left(\frac{1}{\sqrt{n}}\right), \\ V_{2n} &= o_P\left(\frac{1}{ng}\right) = o_P\left(\frac{1}{\sqrt{n}}\right), \\ V_{3n} &= o_P\left(\frac{1}{ng}\right) = o_P\left(\frac{1}{\sqrt{n}}\right), \end{aligned} \tag{A.21}$$

and it remains to discuss the asymptotic behaviour of the terms V_{4n}, V_{5n}, V_{6n}.

For the latter random variable we apply Lemma 3.1 in Zheng (1996) to the kernel $H(x, y) = L_g(x - y)\Delta(x)\Delta(y)$. A straightforward calculation and Assumption 5 (in the case $d = 2$) give

$$\mathrm{E}[H^2(X_1, X_2)] = O\left(\frac{1}{g^2}\right) = o(n),$$

which implies that

$$V_{6n} = \mathrm{E}[H(X_1, X_2) + \frac{2}{n}\sum_{i=1}^{n}\{\mathrm{E}(H(X_i, X_j)|X_i) - \mathrm{E}[H(X_i, X_j)]\} + o_P\left(\frac{1}{\sqrt{n}}\right). \tag{A.22}$$

Note that by Taylor expansion the first term in this expansion is given by

$$\mathrm{E}[H(X_1, X_2)] = \mathrm{E}[(\Delta^2 f)(X_1)] + (g^2). \tag{A.23}$$

In order to treat V_{4n} we introduce the notation

$$Z_i = \frac{1}{n(n-1)}\sum_{\substack{j=1\\ j\neq i}}^{n} L_g(X_i - X_j)\Delta(X_j),$$

and obtain by straightforward algebra

$$\mathrm{E}[(Z_i - \mathrm{E}[Z_i|X_i])^2] = o\left(\frac{1}{n^2}\right)$$

uniformly with respect to i. This shows that

$$\begin{aligned} V_{4n} &= \sum_{i=1}^{n}\sigma(X_i)\varepsilon_i\mathrm{E}[Z_i|X_i] + \sum_{i=1}^{n}\sigma(X_i)\varepsilon_i(Z_i - \mathrm{E}[Z_i|X_i]) \\ &= \sum_{i=1}^{n}\sigma(X_i)\varepsilon_i\mathrm{E}[Z_i|X_i] + o_P\left(\frac{1}{\sqrt{n}}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sigma(X_i)(\Delta f)(X_i)\varepsilon_i + o_P\left(\frac{1}{\sqrt{n}}\right), \end{aligned} \tag{A.24}$$

where the third estimate follows from a standard calculation of the conditional expectation $E[Z_i|X_i]$.

The estimation of the remaining term V_{5n} is more delicate. As we did in the analysis of the term V_{2n} in the proof of Theorem 3.1, we first decompose V_{5n} into

$$V_{5n} = V_{5n}^{(1)} + V_{5n}^{(2)} - V_{5n}^{(0)},$$

where

$$V_{5n}^{(0)} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_g(X_i - X_j)\Delta(X_j)\delta_0,$$

$$V_{5n}^{(r)} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_g(X_i - X_j)\Delta(X_j)\delta_r(X_{ir}), \qquad r = 1, 2,$$

and the functions δ_0, δ_1, δ_2 are defined in (A.6).

With this notation we obtain for $V_{5n}^{(1)}$,

$$V_{5n}^{(1)} = V_{5n}^{(1.1)} + V_{5n}^{(1.2)} + V_{5n}^{(1.3)},$$

where

$$V_{5n}^{(1.1)} = \frac{1}{n^3(n-1)} \sum_{i,l,k=1}^{n} \sum_{j \neq i} L_g(X_i - X_j)\Delta(X_j) w_{kl}^{(1)}(X_{i1})\sigma(X_l)\varepsilon_l,$$

$$V_{5n}^{(1.2)} = \frac{1}{n^3(n-1)} \sum_{i,k,l=1}^{n} \sum_{j \neq i} L_g(X_i - X_j)\Delta(X_j) w_{kl}^{(1)}(X_{i1})(m(X_{l1}, X_{l2}) - m(X_{i1}, X_{k2})),$$

$$V_{5n}^{(1.3)} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_g(X_i - X_j)\Delta(X_j) \left(\frac{1}{n} \sum_{k=1}^{n} m(X_{i1}, X_{k2}) - m_1(X_{i1}) \right)$$

and $w_{kl}^{(1)}$ is defined in (A.7).

The term $V_{5n}^{(1.1)}$ can be rewritten as

$$V_{5n}^{(1.1)} = \frac{1}{n} \sum_{l=1}^{n} \sigma(X_l)\varepsilon_l W_l,$$

where

$$W_l = \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} L_g(X_i - X_j)\Delta(X_j) w_{kl}^{(1)}(X_{i1}).$$

Now a Taylor expansion and (A.7) give, for $i, j, k \neq l$,

$$\begin{aligned} \mathrm{E}(W_l|X_l) &= \mathrm{E}(L_g(X_i - X_j)\Delta(X_j) w_{kl}^{(1)}(X_{i1})|X_l)(1 + o_P(1)) \\ &= \mathrm{E}\left(L_g(X_i - X_j)\Delta(X_j) \frac{K_{1,h_1}(X_{l1} - X_{i1}) K_{2,h_2}(X_{l2} - X_{k2})}{f(X_{i1}, X_{k2})} |X_l \right)(1 + o_P(1)) \\ &= \frac{f_2(X_{l2})}{f(X_{l1}, X_{l2})} \int (\Delta f^2)(X_{l1}, t_2) \mathrm{d}t_2 \cdot (1 + o_P(1)). \end{aligned} \tag{A.25}$$

Moreover, a tedious calculation shows that

$$\mathrm{E}[(W_l - E(W_l|X_l))^2] = o(1),$$

which implies that

$$V_{5n}^{(1.1)} = \frac{1}{n} \sum_{l=1}^{n} \sigma(X_l)\varepsilon_l \mathrm{E}(W_l|X_l) + o_P\left(\frac{1}{\sqrt{n}}\right). \tag{A.26}$$

For the term $V_{5n}^{(1.2)}$ we have

$$V_{5n}^{(1.2)} = \frac{1}{n^3(n-1)} \sum_{i,k,l=1}^{n} \sum_{j \neq i} H(X_i, X_j, X_k, X_l),$$

with the notation

$$H(X_i, X_j, X_k, X_l)$$
$$= L_g(X_i - X_j)\Delta(X_j)\frac{K_{1,h_1}(X_{l1} - X_{i1})K_{2,h_2}(X_{l2} - X_{k2})}{\hat{f}^{(1)}(X_{i1}, X_{k2})}(m(X_{l1}, X_{l2}) - m(X_{i1}, X_{k2})).$$

Computing the expectation of $V_{5n}^{(1.2)}$, we obtain, for pairwise different i, j, k, l,

$$\begin{aligned}
\mathrm{E}(V_{5n}^{(1.2)}) &= \mathrm{E}[H(X_i, X_j, X_k, X_l)] \cdot (1 + o(1)) \\
&= \mathrm{E}\Bigg[(\Delta f)(X_i)\mathrm{E}\Bigg(\frac{K_{1,h_1}(X_{l1} - X_{i1})K_{2,h_2}(X_{l2} - X_{k2})}{f(X_{i1}, X_{k2})} \\
&\quad \times (m(X_{l1}, X_{l2}) - m(X_{i1}, X_{k2}))|X_i\Bigg)\Bigg] \cdot (1 + o(1)) \\
&= \mathrm{E}[(\Delta f)(X_i) \cdot b_1(X_{i1})] \cdot h_1^2 + o(h_1^2) + O(h_2^q), \qquad \text{(A.27)}
\end{aligned}$$

where $b_1(x_1)$ is defined in (3.6). For the squared statistic we have

$$\left(V_{5n}^{(1.2)}\right)^2 = \frac{1}{n^6(n-1)^2}\sum_{i,i',k,k',l,l'=1}^{n}\sum_{j\neq i, j'\neq i'} H(X_i, X_j, X_k, X_l)H(X_{i'}, X_{j'}, X_{k'}, X_{l'}),$$

and observe that only terms with $\{i, j, k, l\} \cap \{i', j', k', l'\} \neq \varnothing$ contribute to the variance. All terms with more than one index in common give a contribution of order $o(1/n)$. The terms with exactly one index in common are all treated similary and we discuss by way of example the case $k' = k$. For this case we obtain

$$\mathrm{E}[H(X_i, X_j, X_k, X_l)H(X_{i'}, X_{j'}, X_k, X_{l'})] = \mathrm{E}[\mathrm{E}(H(X_i, X_j, X_k, X_l)|X_k)^2],$$

where the conditional expectation can be estimated as follows:

$$\begin{aligned}
&\mathrm{E}[H(X_i, X_j, X_k, X_l)|X_k] \\
&= \mathrm{E}\left[(\Delta f)(X_i)\frac{K_{1,h_1}(X_{k1} - X_{i1})K_{2,h_2}(X_{k2} - X_{l2})}{f(X_{i1}, X_{l2})}(m(X_{k1}, X_{k2}) - m(X_{i1}, X_{l2}))|X_k\right] + o(1) \\
&= \mathrm{E}\left[(\Delta f)(X_i)\frac{K_{1,h_1}(X_{k1} - X_{i1})f_2(X_{k2})}{f(X_{i1}, X_{k2})}(m(X_{k1}, X_{k2}) - m(X_{i1}, X_{k2}))|X_k\right] + o(1) \\
&= o(1).
\end{aligned}$$

Here the first equality follows by conditioning on X_i, X_k, X_l, the second by conditioning on X_k, X_i and the third by a direct integration. This implies that

$$\sqrt{n}(V_{5n}^{(1.2)} - \mathrm{E}(V_{5n}^{(1.2)})) = o_P(1). \qquad \text{(A.28)}$$

Finally,

$$V_{5n}^{(1.3)} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{E}[\Delta f(X_i)(m(X_{i1}, X_{k2}) - m_1(X_{i1}))|X_k] + o_P\left(\frac{1}{\sqrt{n}}\right)$$

$$= \frac{1}{n}\sum_{k=1}^{n} \mathrm{E}[\Delta f(X_i)(m(X_{i1}, X_{k2}))|X_k] - \mathrm{E}[\Delta f(X_i)(m(X_{i1}, X_{k2}))] + o_P\left(\frac{1}{\sqrt{n}}\right) \qquad \text{(A.29)}$$

which gives, by a combination of (A.25)–(A.29), and noting that $\mathrm{E}(V_{5n}^{(1.3)}) = O(1/n)$,

$$V_{5n}^{(1)} - \mathrm{E}(V_{5n}^{(1)}) = \frac{1}{n}\sum_{l=1}^{n} \sigma(X_l)\varepsilon_l \left[\frac{f_2(X_{l2})}{f(X_{l1}, X_{l2})}\int (\Delta f^2)(X_{l1}, t_2)\,\mathrm{d}t_2\right]$$

$$+ \frac{1}{n}\sum_{k=1}^{n}\{E[\Delta f(X_i)m(X_{i1}, X_{k2})|X_k]$$

$$- \mathrm{E}[\Delta f(X_i)m(X_{i1}, X_{k2})]\} + o_P\left(\frac{1}{\sqrt{n}}\right) \qquad \text{(A.30)}$$

and

$$\mathrm{E}(V_{5n}^{(1)}) = \mathrm{E}[(\Delta f)(X_i)\cdot b_1(X_{i1})]\cdot h_1^2 + o(h_1^2) + O(h_2^q), \qquad \text{(A.31)}$$

where b_1 is defined in (3.6).

The term $V_{5n}^{(2)}$ is treated exactly in the same way, showing that

$$V_{5n}^{(2)} - E(V_{5n}^{(2)}) = \frac{1}{n}\sum_{l=1}^{n} \sigma(X_l)\varepsilon_l \left[\frac{f_1(X_{l1})}{f(X_{l1}, X_{l2})}\int \Delta f^2(t_1, X_{l2})\,\mathrm{d}t_1\right]$$

$$+ \frac{1}{n}\sum_{k=1}^{n}\{\mathrm{E}[\Delta f(X_i)m(X_{k1}, X_{i2})|X_k]$$

$$- \mathrm{E}[\Delta f(X_i)m(X_{k1}, X_{i2})]\} + o_P\left(\frac{1}{\sqrt{n}}\right), \qquad \text{(A.32)}$$

where

$$\mathrm{E}(V_{5n}^{(2)}) = \mathrm{E}[\Delta f(X_i)\cdot b_2(X_{i2})]\cdot h_1^2 + o(h_1^2) + O(h_2^q) \qquad \text{(A.33)}$$

and $b_2(x_2)$ is given by in (3.6).

For the remaining term $V_{5n}^{(0)}$, we have

$$V_{5n}^{(0)} = \frac{1}{n}\sum_{k=1}^{n}(Y_k - c)\cdot\left\{\frac{1}{n(n-1)}\sum_{i\neq k}\sum_{j\neq i,k} L_g(X_i - X_j)\Delta(X_j)\right\} + O_P\left(\frac{1}{n}\right)$$

$$= \frac{1}{n}\sum_{k=1}^{n}(\sigma(X_k)\varepsilon_k + (m(X_k) - c))\cdot \mathrm{E}(\Delta f(X_1)) + o_P\left(\frac{1}{\sqrt{n}}\right)$$

$$= \frac{1}{n}\sum_{k=1}^{n}\{\sigma(X_k)\varepsilon_k \cdot \mathrm{E}(\Delta f(X_i)) + \mathrm{E}(\Delta f(X_i)\, m(X_k)|X_k) - \mathrm{E}(\Delta f(X_i)\, m(X_k))\} + o_P\left(\frac{1}{\sqrt{n}}\right). \tag{A.34}$$

A combination of the above results (A.22)–(A.24) and (A.30)–(A.34) gives

$$\sqrt{n}(T_{0,n} - \mathrm{E}(T_{0,n})) = A_n + B_n + C_n + o_P(1),$$

where $\mathrm{E}(T_{0,n})$ is defined in (3.5),

$$A_n = \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\{\mathrm{E}(H(X_i, X_j)|X_i) - \mathrm{E}[H(X_i, X_j)]\} = \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\{\Delta^2 f(X_i) - \mathrm{E}(\Delta^2 f(X_i))\} + o_P(1),$$

$$B_n = \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\sigma(X_i)\varepsilon_i\{(\Delta f)(X_i) - P_0^*(\Delta f)(X_i)\},$$

$$C_n = \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{E}(\Delta f(X_j)[m(X_{i1}, X_{i2}) - m(X_{j1}, X_{i2}) - m(X_{i1}, X_{j2})]|X_i)$$

$$- \mathrm{E}(\Delta f(X_j)[m(X_{i1}, X_{i2}) - m(X_{j1}, X_{i2}) - m(X_{i1}, X_{j2})])$$

and the mapping P_0^* is given by (3.8). The asymptotic normality now follows by a standard application of Lyapunov's theorem. The asymptotic variance is obtained by a routine calculation. We obtain

$$\mathrm{var}(A_n + C_n) = 4\,\mathrm{var}[\Delta^2 f(X_1) + \mathrm{E}(\Delta f(X_2)[m(X_{11}, X_{12}) - m(X_{21}, X_{12}) - m(X_{11}, X_{22})]|X_1)],$$

$$\mathrm{var}(B_n) = 4\mathrm{E}(\sigma^2(X_1)\{(I - P_0^*)(\Delta f)(X_1)\}^2)$$

and $\mathrm{cov}(A_n + C_n, B_n) = 0$, which yields the asymptotic variance in (3.7) for $\pi = 1$ and completes the proof of Theorem 3.2.

A.3. Proof of Lemma 3.6

From Jensen's inequality and Fubini's theorem we have

$$\int (K * K)^2(x)\,\mathrm{d}x = \int \left\{ \int K(x-u)K(u)\,\mathrm{d}u \right\}^2 \mathrm{d}x$$
$$\leqslant \iint K^2(x-u)K(u)\,\mathrm{d}u\,\mathrm{d}x = \int K^2(x)\,\mathrm{d}x,$$

which proves the left-hand side of (3.11). The remaining part is obtained by using the first part and the triangle inequality, that is,

$$\left\{ \int (2K - K * K)^2(x)\,\mathrm{d}x \right\}^{1/2} \geqslant 2\left\{ \int K^2(x)\,\mathrm{d}x \right\}^{1/2} - \left\{ \int (K * K)^2(x)\,\mathrm{d}x \right\}^{1/2}$$
$$\geqslant \left\{ \int K^2(x)\,\mathrm{d}x \right\}^{1/2}.$$

Acknowledgements

The authors are grateful to I. Gottschlich who typed parts of this paper with considerable technical expertise, and to S. Sperlich for very helpful discussions about the method of marginal integration. We also thank O. Linton for sending us an earlier version of Gozalo and Linton (2001), and L. Mattner for his help with the proof of Lemma 3.6. The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, Reduction of complexity in multivariate data structures) is gratefully acknowledged.

References

Alcalá, J.T., Christóbal, J.A. and González-Manteiga, W. (1999) Goodness-of-fit test for linear models based on local polynomials. *Statist. Probab. Lett.*, **42**, 39–46.

Andrews, D.W.K. and Whang, Y.-J. (1990) Additive interactive regression models: circumvention of the curse of dimensionality. *Econometric Theory*, **6**, 466–479.

Azzalini, A. and Bowman, A. (1993) On the use of nonparametric regression for checking linear relationships. *J. Roy. Statist. Soc. Ser. B*, **55**, 549–557.

Barry, D. (1993) Testing for additivity of a regression function. *Ann. Statist.*, **21**, 235–254.

Berger, J.O. and Delampady, M. (1987) Testing precise hypotheses. *Statist. Sci.*, **2**, 317–352.

Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive models. *Ann. Statist.*, **17**, 453–555.

Dette, H. (1999) A consistent test for the functional form of a regression based on a difference of variance estimators. *Ann. Statist.*, **27**, 1012–1040.

Dette, H. and Munk, A. (1998) Validation of linear regression models. *Ann. Statist.*, **26**, 778–800.

Eubank, R.L., Hart, J.D., Simpson, D.G. and Stefanski, L.A. (1995) Testing for additivity in nonparametric regression. *Ann. Statist.*, **23**, 1896–1920.

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

Friedman, J.H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Amer. Statist. Assoc*., **76**, 817–823.

González-Manteiga, W. and Cao, R. (1993) Testing the hypothesis of a general linear model using nonparametric regression estimation. *Test*, **2**, 161–188.

Gozalo, P.L. and Linton, O.B. (2001) Testing additivity in generalized nonparametric regression models with estimated parameters. *J. Econometrics*. To appear.

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

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

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models*. London: Chapman & Hall.

Hengartner, N.W. (1996) Rate optimal estimation of additive regression via the integration method in the presence of many covariates. Preprint, Department of Statistics, Yale University.

Hjellvik, V. and Tjøstheim, D. (1995) Nonparametric tests of linearity for time series. *Biometrika*, **82**, 351–368.

Li, K.-C. (1991) Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc*., **86**, 316–342.

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

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

Linton, O.B. (1997) Efficient estimation of additive nonparametric regression models. *Biometrika*, **84**, 469–473.

Linton, O.B. (2000) Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory*, **16**, 502–523.

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

Nadaraya, E.A. (1965) On non-parametric estimates of density functions and regression curves. *Theory Probab. Appl*., **10**, 186–190.

Nielsen, J.P. and Linton, O.B. (1998) An optimization interpretation of integration and back-fitting estimators for separable nonparametric models. *J. Roy. Statist. Soc. Ser. B*, **60**, 217–222.

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

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

Silverman, B.W. (1978) Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist*., **6**, 177–184.

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

Sperlich, S., Tjøstheim, D. and Yang, L. (1999) Nonparametric estimation and testing of interaction in additive models. *Econometric Theory*. To appear.

Staudte, R.S. and Sheather, S.J. (1990) *Robust Estimation and Testing*. New York: Wiley.

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

Tjøstheim, D. (1994) Nonlinear time series: a selective review. *Scand. J. Statist*., **21**, 97–130.

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

Tukey, J. (1949) One degree of freedom for non-additivity. *Biometrics*, **5**, 232–242.

Watson, G.S. (1964) Smooth regression analysis. *Sankhyā, Ser. A*, **26**, 359–372.

Wu, C.F.Y. (1986) Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist*., **14**, 1261–1295.

Zheng, J.X. (1996) A consistent test of functional form via nonparametric estimation techniques. *J. Econometrics*, **75**, 263–289.

Received November 1999 and revised January 2001