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

Holger Dette
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum
Germany
email: holger.dette@ruhr-uni-bochum.de

Carsten von Lieres und Wilkau
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum
Germany
email: carsten.von.lieres@ruhr-uni-bochum.de

July 20, 2000

Abstract

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 either based on the differences between a fit under the assumption of additivity and a fit in the general model or based 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 alternatives.

AMS Subject Classification: 62G07, 62G10
Keywords and Phrases: Additive models, dimension reduction, test of additivity, marginal integration estimate

1 Introduction

Consider the common nonparametric regression model

$$Y = m(X) + \sigma(X)\varepsilon \qquad (1.1)$$

where $X = (X_1, \ldots, X_d)^T$ 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 as kernel

or 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 e.g. Friedman and Stuetzle (1981), Li (1991)]. Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1990) promoted the additive regression model

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

where $k_1, \ldots, k_d$ are unknown smooth functions normalized by $E[k_\alpha(X_\alpha)] = 0$ and $x = (x_1, \ldots, x_d)^T$. 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, Hastie and Tibshirani (1989) proposed the 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 been recently analyzed by Opsomer and Ruppert (1997) and Linton, Mammen and Nielsen (1999). Because of the implicit definition of these estimates several authors have proposed a direct method that is based on marginal integration (see e.g. Tjøstheim and Auestad (1994), Tjøstheim (1994), Linton and Nielsen (1995) or Chen, Härdle, Linton and Severance-Lossin (1996)). This method does not require an iterative solution of a system of nonlinear equations and yields an alternative projection onto the subspace of additive functions which is not necessarily orthogonal. 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), the problem of testing additivity has only found recently interest in the literature [see e.g. Hastie and Tibshirani (1990), Barry (1993) or Eubank, Hart, Simpson and Stefanski (1995), Sperlich, Tjøstheim and Yang (1999), Linton and Gozalo (1999)].
As diverse as this literature appears all proposed methods have one thing in common: they all test what they actually should not, namely that the preassigned additive model is NOT valid. 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) or Staudte and Sheater (1990)]. These authors point out that often it is preferable to reformulate the hypothesis (1.2) in a testing problem, which allows the experimenter to show that m is "close" to additivity at a controlled error rate. In other words, if M^2 is a measure of additivity (i.e. $M^2 = 0$ if H_0 is valid) it is proposed to reformulate the hypothesis (1.2) into

$$H_\varepsilon : M^2 > \eta \quad H_1 : M^2 \leq \eta \tag{1.3}$$

where η is a given sufficiently small constant such that the experimenter agrees to analyze the data under the assumption of additivity, whenever $M^2 \leq \eta$. 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 from a marginal integration fit [see also Zheng (1996), who used a similar idea for testing a parametric form of the regression] and 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) and Gozalo and Linton (1999)]. The main difference between our approach and the work of the lastnamed authors is that we are able to find the asymptotic properties of the tests under any fixed alternative of non-additivity. We will demonstrate at the end of Section 3 that these results can be used for the estimation of the type II error of the test for the classical hypothesis (1.2) and for testing the precise hypotheses of the form (1.3). As a further 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. Finally, some of the proofs, which are rather cumbersome, are deferred to the appendix.

2 Marginal integration revisited

Let f denote the density of the explanatory variable $X = (X_1, \ldots, X_d)^T$ with marginal distributions 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 α-th coordinate from x, i.e. $x_{\underline{\alpha}} = (x_1, \ldots, x_{\alpha-1}, x_{\alpha+1}, \ldots, x_d)$. If L^2_{add} denotes the subspace of additive functions in the Hilbert space $L^2(f)$ we consider the projection P_0 from $L^2(f)$ onto L^2_{add} defined by

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

where

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

$$(2.3) \qquad c = \int m(t) f(t) dt.$$

Here we used the notation

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

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

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

and

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

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 easily verifies that it is an orthogonal projection in the case of independent predictors.

Unless it is not mentioned differently let $K_i(\cdot)$ $(i = 1, 2)$ denote one – and $(d-1)$ – dimensional Lipschitz – continuous kernels of order 2 and $q \geq d$, respectively, with compact support and define for a bandwidth $h_i > 0$, $t_1 \in I\!R$, $t_2 \in I\!R^{d-1}$

$$K_{i,h_i}(t_i) = \frac{1}{h_i} K_i(\frac{t_i}{h_i}) \quad i = 1, 2. \tag{2.4}$$

For an i.i.d. sample $(X_i, Y_i)_{i=1}^n, X_i = (X_{i1}, \ldots, X_{id})^T$ we consider the empirical counterparts of the components of m_0 in (2.1), i.e.

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

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

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 (1964) 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 $\hat{e}_j = Y_j - \hat{m}_0(X_j) \quad (j = 1, \dots, n)$. 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, $L_g(\cdot) = \frac{1}{g^d} L_g(\frac{\cdot}{g})$, $g > 0$ 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 (1999) in the problem of testing additivity in a more general context. A theoretical justification for the application of this statistic for 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) 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

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

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, that is

$$\begin{aligned} T_{1,n} &= \frac{1}{n} \sum_{i=1}^n [\hat{m}(X_i) - \hat{m}_0(X_i)]^2 \pi(X_i) \\ T_{2,n} &= \frac{1}{n} \sum_{i=1}^n \hat{e}_i [\hat{m}(X_i) - \hat{m}_0(X_i)] \pi(X_i) \\ T_{3,n} &= \frac{1}{n} \sum_{i=1}^n [\hat{e}_i^2 - \hat{d}_i^2] \pi(X_i) \end{aligned} \tag{2.13}$$

In (2.13) $\hat{m}$ 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 nonrestricted 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 (1999) 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 (1999) using different techniques in the proofs. It is the main purpose of the present paper to show that the asymptotic behaviour of the statistics $T_{0,n}$ - $T_{3,n}$ under fixed alternatives is rather different and to demonstrate potential applications of such results.

3 Main results and a comparison

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

(A1) 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 $q \geq d$.

(A2) $m \in C_b^q(Q)$, where $C_b^q(Q)$ denotes the class of bounded functions (defined on Q) with continuous partial derivatives of order q.

(A3) $\sigma \in C_b(Q)$ where $C_b(Q)$ denotes the class of bounded continuous functions (defined on Q).

(A4) The distribution of the error has a finite fourth moment, i.e. $E[\varepsilon^4] < \infty$.

(A5) The bandwidths $g, h_1, h_2 > 0$ satisfy (as $n \to \infty$)

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

Note that the optimal order for a two times continuously differentiable regression function $h_1 \sim n^{-1/5}$ in (A5) requires $q > d - 1$ in order to fulfill

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

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

Theorem 3.1. *If assumptions (A1) - (A5) and the hypothesis of additivity are satisfied, then the statistic $T_{0,n}$ defined in (2.10) is asymptotically normal distributed, i.e.*

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

where the asymptotic variance is given by

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

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 (1999) and provides a test for the hypothesis of additivity by rejecting H_0 for large values of $T_{0,n}$, i.e.

$$ng^{\frac{d}{2}}T_{0,n} > u_{1-\alpha}\hat{\lambda}_{0,n} \tag{3.3}$$

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 as given in Zheng (1996), i.e.

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

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 an 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 (A1) – (A5) are satisfied and the regression is not additive, i.e. $\Delta = m - P_0 m \neq 0$, then*

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

where

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

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

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

$c_2\left(K_1\right) = \int t_1^2 K_1(t_1)dt_1$ and the asymptotic variance is given by

$$\begin{aligned} \mu_0^2 &= 4E[\sigma^2(X_1)\{P_1(\Delta\pi^2 f)(X_1)\}^2] \\ &+ 4V\Big[(\Delta^2\pi^2 f)(X_1) - E\Big(\Delta\pi^2 f(X_2)\{\sum_{\alpha=1}^{d} m(X_{2\alpha}, X_{1\underline{\alpha}}) - (d-1)m(X_1)\} \mid X_1\Big)\Big] \end{aligned} \tag{3.7}$$

$P_1 m = m - P_0^* m$ and the mapping P_0^* is defined by

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

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 easily shows $P_0^* = P_0$. Moreover, if additionally the weight function is given by $\pi = \frac{1}{\sqrt{f}}$, the asymptotic variance in (3.7) simplifies to

$$\mu_0^2 = 4E[\sigma^2(X_1)\Delta^2(X_1)] + 4V[\Delta^2(X_1)]$$

where $\Delta = m - m_0$.

Remark 3.4. A careful analysis of the proof of Theorem 3.2 shows (see also Chen, Härdle, Linton, Severance–Lossin (1996)) that for a sufficiently smooth regression and kernels L and $K_i,\ i = 1, 2$ of sufficiently high order we have

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

where the term $M^2 := E[\Delta^2(X_1)(\pi^2 f)(X_1)]$ 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)

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

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

$$H_0 : M^2 > \eta \quad H_1 : M^2 \leq \eta$$

where η is a given sufficiently small constant for which the experimenter agrees to analyze the data in the additive model. An asymptotic level α test is given by rejecting the null-hypothesis $H_0 : M^2 > \eta$ if

$$\sqrt{n}(T_{0,n} - \eta) \leq u_\alpha \hat{\mu}_0,$$

where $\hat{\mu}_0^2$ is an appropriate estimator of the asymptotic variance in Theorem 3.2. This formulation allows to test that the model is “close” to additivity at a controlled error rate. We finally 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. *Assume that (A1) – (A5) are satisfied and $T_{1,n},\ T_{2,n},\ T_{3,n}$ are defined in (2.13).*

(i) Under the hypothesis of additivity we have

$$ng^{\frac{d}{2}}\{T_{j,n} - E_{H_0}[T_{j,n}]\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \lambda_j^2)\ ;\ j = 1, \ldots, 3$$

where

$$\begin{aligned}
B_1 = E_{H_0}[T_{1,n}] &= \frac{1}{ng^d}\int L^2(x)dx \int \sigma^2(x)\pi(x)dx + o(\frac{1}{ng^{\frac{d}{2}}}),\\
B_2 = E_{H_0}[T_{2,n}] &= \frac{1}{ng^d}L(0) \int \sigma^2(x)\pi(x)dx + o(\frac{1}{ng^{\frac{d}{2}}}),\\
B_3 = E_{H_0}[T_{3,n}] &= \frac{1}{ng^d}(2L(0) - \int L^2(x)dx) \int \sigma^2(x)\pi(x)dx + o(\frac{1}{ng^{\frac{d}{2}}})
\end{aligned}$$

and

$$\begin{aligned}
\lambda_1^2 &= 2\int \sigma^4(x)\pi(x)dx \int (L*L)^2(x)dx, \\
\lambda_2^2 &= 2\int \sigma^4(x)\pi(x)dx \int L^2(x)dx, \\
\lambda_3^2 &= 2\int \sigma^4(x)\pi(x)dx \int (2L-(L*L))^2(x)dx
\end{aligned}$$

where $f*g$ denotes the convolution of the functions f and g.

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

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

where

$$\begin{aligned}
E_{H_1}[T_{1,n}] &= B_1 + \nu_0 - 2\nu_1 + 2\nu_2, \\
E_{H_1}[T_{2,n}] &= B_2 + \nu_0 - 2\nu_1 + \ \nu_2, \\
E_{H_1}[T_{3,n}] &= B_3 + \nu_0 - 2\nu_1,
\end{aligned}$$

$$\begin{aligned}
\nu_0 &= E[(\Delta^2\pi)(X_1)], \\
\nu_1 &= E[(\Delta\pi)(X_1)b(X_1)]\cdot h_1^2 + o\left(h_1^2\right), \\
\nu_2 &= E[(\Delta\pi)(X_1)b_{NW}(X_1)]\cdot g^2 + o\left(g^2\right),
\end{aligned}$$

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

$$\begin{aligned}
\mu_j^2 &= 4E[\sigma^2(X_1)\{P_1(\Delta\pi)(X_1)\}^2] \\
&+ V\Big[(\Delta^2\pi)(X_1) - 2E\Big(\Delta\pi(X_2)\{\sum_{\alpha=1}^{d} m(X_{2\alpha}, X_{1\underline{\alpha}}) - (d-1)m(X_1)\} \mid X_1\Big)\Big]
\end{aligned}$$

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

In the remaining part of this section we will use Theorem 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

$$E_{H_1}[T_{j,n}] = M_j^2 + B_j + o(\frac{1}{\sqrt{n}}) \quad j=0,\ldots,3$$

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

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

In this case the probability of rejection is approximately given by

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

where μ_j, λ_j $(j=0,\dots,3)$ are defined in Theorem 3.1, 3.2 and 3.5, respectively. 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 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 $\frac{\pi}{\sqrt{f}}$ (in practice an estimator of f has to be used), which gives

$$M_j^2 = E[\Delta^2(X_1)\pi(X_1)] \quad (j=0,\dots,3)$$

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

$$(3.10) \qquad \mu_0^2 > \mu_j^2 \quad (j=1,\dots,3).$$

Looking at the dominating term in (3.9) we thus obtain that (asymptotically) tests based on the statistics $T_{j,n}$ $(j=1,\dots,3)$ 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 $(ng^{\frac{d}{2}})^{-\frac{1}{2}}$. 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 (1999) without a proof. A rigorous derivation will be given at the end of the appendix.

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

$$(3.11) \qquad \int (K*K)^2(x)dx \le \int K^2(x)dx \le \int (2K - K*K)^2(x)dx$$

or equivalently

$$\lambda_1^2 = \lambda_2^2 \le \lambda_0^2 \le \lambda_3^2$$

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.

Remark 3.7. Note that Gozalo and Linton (1999) study the asymptotic distribution of the statistics $T_{0,n} - T_{3,n}$ 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 did not formulate the results in full detail, but indicate the generalization of Theorem 3.1, 3.2 in the situation of a known link function as considered in Linton and Härdle (1996). In the nonparametric regression model

$$E[Y|X = x] = m(x)$$

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 straight-forward [see e.g. 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

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

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

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

denote an estimator of $\int G(m(x))f(x)dx$. Defining

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

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

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

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 e.g. Linton and Härdle (1996) or Gozalo and Linton (1999)] Theorem 3.1 remains valid. On the other hand, under a fixed alternative $\sqrt{n}(T_{0,n} - E[T_{0,n}])$ is asymptotically normal where the asymptotic variance is given by

$$\begin{aligned}\mu_0^2 &= 4E[\sigma^2(X_1)P_1^G(\Delta\pi^2)(X_1)] \\ &+ 4V\Big[(\Delta\pi)^2(X_1)f(X_1) - E\Big((\Delta\pi^2 f)(X_2)\{\sum_{\alpha=1}^{d} G(m(X_{2\alpha}, X_{1\underline{\alpha}})) - (d-1)G(m(X_1))\}|X_1\Big)\Big]\end{aligned}$$

where $\sigma^2(x) = V[Y|X = x]$ 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

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

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

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

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.

A Proofs

For the sake of a transparent notation we consider the case $d = 2$. In addition we use $\pi(x) \equiv 1$ as weight function; the general case is treated exactly in the same way. Because all results are essentially proved similary, 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

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

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

$$\delta_r(x_r) = \widehat{m}_r(x_r) - m_r(x_r), \ r = 1, 2 \ ; \ \delta_0 = \frac{1}{n}\sum_{k=1}^{n} Y_k - c \tag{A.6}$$

we derive the decomposition

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

where

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

and

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

At first we will show that

$$V_{2n}^{(r)} = O_P(\frac{1}{nh_1}) \quad ; r = 1, 2.$$

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

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

where

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

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

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

(by the law of the iterated logarithm) we get (note that $\frac{1}{n}\sum_{l=1}^{n} w_{kl}^{(1)}(x_1) = 1$)

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

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

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

Now, by definition (A.7)

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

where the last equality is obtained by the strong uniform consistency of the kernel density estimate $\widehat{f}^{(1)}$ [see e.g. Silverman (1978)]. For $k \neq i, j$ Taylorexpansion gives

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

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

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

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 the above sum have expectation zero exept for the case where

$$\begin{aligned} i' = i &\text{ and } l' = l \\ i' = l &\text{ and } i = l' \\ i = l &\text{ and } i' = l' \\ i' = i &= l' = l. \end{aligned}$$

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

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

which is of order $O\left(\frac{1}{n^2h_1^2}\right)$ by the same reasoning as above. The other cases are treated in the same way showing that $V(V_{2n}^{(1.1)}) = O(\frac{1}{n^2h_1^2})$. It follows by Chebyshev's inequality

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

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

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

In order to find the corresponding variance we note that

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

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

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

Using the strong uniform consistency of $\widehat{f}$ again and the assumption $\frac{\log n}{nh_1h_2} = o\left(h_1^2\right)$ we get by a lengthy argument

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

where the latter is asymptotically equal to

$$\begin{aligned}&\Big\{E\Big(\frac{K_{1,h_1}\left(X_{l1}-X_{j1}\right)f_2\left(X_{l2}\right)}{f\left(X_{j1},X_{l2}\right)}\left(m\left(X_{l1},X_{l2}\right)-m\left(X_{j1},X_{l2}\right)\right)\mid X_j\Big)+O\left(h_2^q\right)\Big\}\left(1+o\left(1\right)\right)\\ =\ &O\left(h_1^2\right)+O\left(h_2^q\right)\end{aligned}$$

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

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

where the last equality is a consequence of assumption (A5). The terms in the sum (A.10) with $i' = i$ and $l' = l$ (all other indices pairwise different) have expectation

$$\begin{aligned}
&\frac{1}{n^2} E\Big[\sigma^2(X_i) E\big(L_g(X_i - X_j) \\
&\times E\Big(w_{kl}^{(1)}(X_{j1})\left(m(X_{l1}, X_{l2}) - m(X_{j1}, X_{k2})\right) \mid X_i, X_j, X_l\Big) \mid X_i, X_l\Big)^2\Big] \\
= &\frac{1}{n^2} E\Big[\sigma^2(X_i) E\Big(L_g(X_i - X_j) \\
&\times K_{1,h_1}(X_{l1} - X_{j1})\big(\frac{f_2(X_{l2})}{f(X_{j1}, X_{l2})}\left(m(X_{l1}, X_{l2}) - m(X_{j1}, X_{l2})\right) + o(1)\Big) \mid X_i, X_l\Big)^2\Big] \\
= &O(\frac{1}{n^2 h_1^2})
\end{aligned}$$

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

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

Combining (A.9) and (A.12) we get

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

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

$$V_{2n} = O_P(\frac{1}{nh_1}\Big)$$

It follows from assumption (A5) that

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

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 already did, we only state the estimates for its expectation and variance, that is

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

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

From (A.14) and (A.15) and assumption (A5) we obtain

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

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, that is

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

where V_{1n}, V_{2n}, V_{3n} are 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 (A5) (in the case $d = 2$) we have

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

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 (A5) (in the case $d = 2$) give

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

which implies

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

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

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

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

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

(uniformly with respect to i). This shows

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

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 proof of Theorem 3.1 in the analysis of the term V_{2n} we first decompose V_{5n} into

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

where

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

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

and the functions $\delta_0,\ \delta_1,\ \delta_2$ 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

$$\begin{aligned} V_{5n}^{(1.1)} &= \frac{1}{n^3\,(n-1)} \sum_{i,l,k=1}^{n} \sum_{j\neq i} L_g\left(X_i - X_j\right) \Delta\left(X_j\right) w_{kl}^{(1)}\left(X_{i1}\right) \sigma\left(X_l\right) \varepsilon_l \\ V_{5n}^{(1.2)} &= \frac{1}{n^3\,(n-1)} \sum_{i,k,l=1}^{n} \sum_{j\neq i} L_g\left(X_i - X_j\right) \Delta\left(X_j\right) w_{kl}^{(1)}\left(X_{i1}\right) \left(m\left(X_{l1}, X_{l2}\right) - m\left(X_{i1}, X_{k2}\right)\right) \\ V_{5n}^{(1.3)} &= \frac{1}{n\,(n-1)} \sum_{i=1}^{n} \sum_{j\neq i} L_g\left(X_i - X_j\right) \Delta\left(X_j\right) \Big(\frac{1}{n}\sum_{k=1}^{n} m(X_{i1}, X_{k2}) - m_1(X_{i1})\Big) \end{aligned}$$

and $w_{kl}^{(1)}$ is defined in (A.7). The term $V_{5n}^{(1.1)}$ can be rewritten as

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

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 Taylorexpansion and (A.7) give for $i, j, k \neq l$

$$\begin{aligned}
E\,(W_l \mid X_l) &= E(L_g\,(X_i - X_j)\,\Delta\,(X_j)\,w_{kl}^{(1)}\,(X_{i1}) \mid X_l)\,(1 + o_P\,(1)) \\
&= E\Big(L_g\,(X_i - X_j)\,\Delta\,(X_j)\,\frac{K_{1,h_1}\,(X_{l1} - X_{i1})\,K_{2,h_2}\,(X_{l2} - X_{k2})}{f\,(X_{i1}, X_{k2})} \mid X_l\Big)\,(1 + o_P\,(1)) \\
&= \frac{f_2\,(X_{l2})}{f\,(X_{l1}, X_{l2})} \int (\Delta f^2)\,(X_{l1}, t_2)\,dt_2 \cdot (1 + o_P\,(1)) \qquad \text{(A.25)}
\end{aligned}$$

Moreover, a tedious calculation shows

$$E\left[(W_l - E\,(W_l \mid X_l))^2\right] = o\,(1)$$

which implies

$$V_{5n}^{(1.1)} = \frac{1}{n} \sum_{l=1}^{n} \sigma\,(X_l)\,\varepsilon_l E\,(W_l \mid X_l) + o_P(\frac{1}{\sqrt{n}}). \qquad \text{(A.26)}$$

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

$$\begin{aligned}
H\,(X_i, X_j, X_k, X_l) &= L_g\,(X_i - X_j)\,\Delta\,(X_j) \\
&\quad \times \frac{K_{1,h_1}\,(X_{l1} - X_{i1})\,K_{2,h_2}\,(X_{l2} - X_{k2})}{\widehat{f}^{(1)}\,(X_{i1}, X_{k2})}\,(m\,(X_{l1}, X_{l2}) - m\,(X_{i1}, X_{k2}))
\end{aligned}$$

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

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

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 $\{i,j,k,l\} \cap \{i',j',k',l'\} \neq \emptyset$ 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 exemplarily discuss the case $k'=k$. For this case we obtain

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

where the conditional expectation can be estimated as follows

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

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

$$\sqrt{n}\left(V_{5n}^{(1.2)} - E(V_{5n}^{(1.2)})\right) = o_P(1)\,. \tag{A.28}$$

Finally,

$$\begin{aligned}
V_{5n}^{(1.3)} &= \frac{1}{n}\sum_{k=1}^{n} E\left[\Delta f\left(X_i\right)\left(m(X_{i1},X_{k2})-m_1(X_{i1})\right)|X_k\right] + o_P(\frac{1}{\sqrt{n}}) \\
&= \frac{1}{n}\sum_{k=1}^{n} E\left[\Delta f\left(X_i\right)\left(m(X_{i1},X_{k2})\right)|X_k\right] - E\left[\Delta f\left(X_i\right)\left(m(X_{i1},X_{k2})\right)\right] + o_P(\frac{1}{\sqrt{n}})
\end{aligned} \tag{A.29}$$

which gives by a combination of (A.25) – (A.29) [Note that $E(V_{5n}^{(1.3)}) = O\left(\frac{1}{n}\right)$]

$$\begin{aligned}
V_{5n}^{(1)} - E(V_{5n}^{(1)}) &= \frac{1}{n}\sum_{l=1}^{n}\sigma\left(X_l\right)\varepsilon_l\Big[\frac{f_2\left(X_{l2}\right)}{f\left(X_{l1},X_{l2}\right)}\int(\Delta f^2)\left(X_{l1},t_2\right)dt_2\Big] \\
&+ \frac{1}{n}\sum_{k=1}^{n}\Big\{E\left[\Delta f\left(X_i\right)\left(m(X_{i1},X_{k2})\right)|X_k\right] \\
&\qquad -E\left[\Delta f\left(X_i\right)\left(m(X_{i1},X_{k2})\right)\right]\Big\} + o_P(\frac{1}{\sqrt{n}})
\end{aligned} \tag{A.30}$$

and

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

where b_1 is defined in (3.6). The term $V_{5n}^{(2)}$ is treated exactly in the same way showing that

$$V_{5n}^{(2)} - E(V_{5n}^{(2)}) = \frac{1}{n}\sum_{l=1}^{n}\sigma\left(X_l\right)\varepsilon_l\Big[\frac{f_1\left(X_{l1}\right)}{f\left(X_{l1},X_{l2}\right)}\int\Delta f^2\left(t_1,X_{l2}\right)dt_1\Big]$$

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

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

where

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

and $b_2(x_2)$ is given by in (3.6). For the remaining term $V_{5n}^{(0)}$ we have

$$\begin{aligned} V_{5n}^{(0)} &= \frac{1}{n}\sum_{k=1}^{n}(Y_k-c)\cdot\Big\{\frac{1}{n\left(n-1\right)}\sum_{i\neq k}\sum_{j\neq i,k}L_g\left(X_i-X_j\right)\Delta\left(X_j\right)\Big\} + O_P(\frac{1}{n}) \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(\sigma(X_k)\varepsilon_k + (m(X_k)-c)\right)\cdot E\left(\Delta f\left(X_1\right)\right) + o_P(\frac{1}{\sqrt{n}}) \\ &= \frac{1}{n}\sum_{k=1}^{n}\Big\{\sigma(X_k)\varepsilon_k\cdot E\left(\Delta f\left(X_i\right)\right) \\ &\qquad +E\left(\Delta f(X_i)\,m(X_k)|X_k\right) - E\left(\Delta f(X_i)\,m(X_k)\right)\Big\} + o_P(\frac{1}{\sqrt{n}}) \end{aligned} \tag{A.34}$$

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

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

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

$$\begin{aligned} A_n &= \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\{E(H(X_i,X_j)|X_i) - E[H(X_i,X_j)]\} = \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\{\Delta^2 f(X_i) - E\left(\Delta^2 f(X_i)\right)\} + o_P(1) \\ B_n &= \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\sigma\left(X_i\right)\varepsilon_i\left\{\left(\Delta f\right)\left(X_i\right) - P_0^*\left(\Delta f\right)\left(X_i\right)\right\} \\ C_n &= \frac{2}{\sqrt{n}}\sum_{i=1}^{n}E\Big(\Delta f\left(X_j\right)\left[m(X_{i1},X_{i2}) - m(X_{j1},X_{i2}) - m(X_{i1},X_{j2})\right]|X_i\Big) \\ &\qquad -E\Big(\Delta f\left(X_j\right)\left[m(X_{i1},X_{i2}) - m(X_{j1},X_{i2}) - m(X_{i1},X_{j2})\right]\Big) \end{aligned}$$

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

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

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

and $Cov\left(A_n+C_n,B_n\right)=0$ 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

$$\begin{aligned}\int (K*K)^2(x)dx &= \int \Big\{ \int K(x-u)K(u)du \Big\}^2 dx \\ &\leq \int \int K^2(x-u)K(u)dudx = \int K^2(x)dx\end{aligned}$$

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

$$\begin{aligned}\Big\{ \int (2K - K*K)^2(x)dx \Big\}^{\frac{1}{2}} &\geq 2\Big\{ \int K^2(x)dx \Big\}^{\frac{1}{2}} - \Big\{ \int (K*K)^2(x)dx \Big\}^{\frac{1}{2}} \\ &\geq \Big\{ \int K^2(x)dx \Big\}^{\frac{1}{2}}\end{aligned}$$

□

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 the unpublished work of Linton and Gozalo (1999) and L. Mattner for his help with the proof of Lemma 3.6. The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, Reduction fo complexity in multivariate data structures) is gratefully acknowledged.

References

A. Azzalani, A. Bowman (1993). On the use of nonparametric regression for checking linear relationships. J. Roy. Stat. Soc. Ser. B, 55, 549 - 557.

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

J.O. Berger, M. Delampady (1987). Testing precise hypotheses. Stat. Sci. 2, 317-352.

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

R. Chen, W. Härdle, O. Linton, E. Severance-Lossin (1996). Estimation and variable selection in additive nonparametric regression models. In Statistical Theory and Computational Aspects of Smoothing (W. Härdle and M. Schimek, eds.). Physika, Heidelberg.

H. Dette (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. Ann. Statist. to appear.

R.L. Eubank, J.D. Hart, D.G. Simpson, L.A. Stefanski (1995). Testing for additivity in nonparametric regression. Ann. Statist. 23 (6), 1896-1920.

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

W. González Manteiga, R. Cao (1993). Testing hypothesis of general linear model using nonparametric regression estimation. Test 2, 161-189.

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

P. Hall (1984). Central limit theorem for integrated square error of multivariate density estimators. J. Mult. Anal. 14, 1-16.

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

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

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

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

O.B. Linton, E. Mammen, J. Nielsen (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Ann. Statist., to appear.

O.B. Linton, P.L. Gozalo (1999). Testing additivity in generalized nonparametric regression models. Preprint.

E.A. Nadaraya (1964). On estimating regression. Theory Probab. Appl., 10, 186-190.

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

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

S. Sperlich, D. Tjøstheim, L. Yang (1999). Nonparametric estimation and testing in additive models. Preprint.

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

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

D. Tjøstheim (1994). Nonparametric specification procedures for time series. Scand. J. of Statistics 21, 97-130.

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

G.S. Watson (1964). Smooth regression analysis. Sankhya, Ser. A, 26, 359-372.

J.X. Zheng (1996). A consistent test of a functional form via nonparametric estimation techniques. J. of Economentrics, 75, 263-289.