# A comparison of different nonparametric methods for inference on additive models

**Holger Dette**
Ruhr-Universität Bochum
Fakultät für Mathematik
D - 44780 Bochum, Germany

**Carsten von Lieres und Wilkau**
Ruhr-Universität Bochum
Fakultät für Mathematik
D - 44780 Bochum, Germany

**Stefan Sperlich**
Universidad Carlos III de Madrid
Departamento de Estadística y Econometría
E - 28903 Getafe, Spain

May 10, 2001

**Abstract**

In this article we highlight the main differences of available methods for the analysis of regression functions that are probably additive separable. We first discuss definition and interpretation of the most common estimators in practice. This is done by explaining the different ideas of modeling behind each estimator as well as what the procedures are doing to the data. Computational aspects are mentioned explicitly. The illustrated discussion concludes with a simulation study on the mean squared error for different marginal integration approaches. Next, various test statistics for checking additive separability are introduced and accomplished with asymptotic theory. Based on the asymptotic results under hypothesis as well as under the alternative of non additivity we compare the tests in a brief discussion. For the various statistics, different smoothing and bootstrap methods we perform a detailed simulation study. A main focus in the reported results is directed on the (non-) reliability of the methods when the covariates are strongly correlated among themselves. Again, a further point are the computational aspects. We found that the most striking differences lie in the different pre-smoothers that are used, but less in the different constructions of test statistics. Moreover, although some of the observed differences are strong, they surprisingly can not be revealed by asymptotic theory.[1]

AMS Subject Classification: 62G07, 62G10
Keywords: marginal integration, additive models, test of additivity.

[1]Acknowledgements: This research was financially supported by the Spanish "Dirección General de Enseñanza Superior" (DGES), reference number PB98-0025 and the Deutsche Forschungsgemeinschaft (SFB 475: *Komplexitätsreduktion in multivariaten Datenstrukturen*, Teilprojekt A2; Sachbeihilfe: *Validierung von Hypothesen*, De 502/9-1). Parts of this paper were written while the first author was visiting Perdue University and this author would like to thank the Department of Statistics for its hospitality.

# 1 Introduction

In the last ten years additive models have attracted an increasing amount of interest in nonparametric statistics. Also in the econometric literature these methods have a long history and are widely used today in both, theoretical considerations and empirical research. Deaton and Müllbauer (1980) provided many examples in microeconomics where the additive structure follows from economic theory of separable decision making like two step budgeting or optimization. Furthermore, additivity is the natural structure when production processes have independent substitution rates for separable goods. In statistics, additivity leads to the circumvention of the curse of dimensionality (see Stone 1985) that usually affects multidimensional nonparametric regression.

The most common and best known nonparametric estimation approaches in these models can be divided into three main groups: the backfitting (see Buja, Hastie and Tibshirani 1989, or Hastie and Tibshirani 1990 for algorithms, and Opsomer and Ruppert 1997 or Mammen, Linton and Nielsen 1999), series estimators (see Andrews and Whang 1990 or Li 2000), and the marginal integration estimator (see Tjøstheim and Auestad 1994, Linton and Nielsen 1995, and also Kim, Linton, Hengartner 2000 for an important modification). Certainly, here we have mentioned only the main references respective basic ideas and theory. Among them, to our knowledge, the series estimator is so far not explored in practice, i.e. although straightforward implementation and good performance is declared, we could not find a simulation study or an application of this method. Moreover, usually hardly feasible assumptions are made on the series and its "smoothing parameters", e.g. reducing bias and variance simultaneously, but without giving a correct idea how to choose them in practice. The backfitting of Buja, Hastie and Tibshirani (1989) is maybe the most studied additive model estimator in practice, and algorithms are developed for various regression problems. However, the backfitting version of Mammen, Linton and Nielsen (1999), for which closed theory is provided but no Monte-Carlo studies, differs a lot in definition and implementation from that one. The marginal integration, finally, has experienced most extensions in theory but actually a quite different interpretation than the aforementioned estimators. This was first theoretically highlighted by Nielsen and Linton (1997) and empirically investigated by Sperlich, Linton and Härdle (1999) in a detailed simulation study. The main point is that backfitting, at least the version of Mammen et al. (1999), and series estimators are orthogonal projections of the regression into the additive space whereas the marginal integration estimator always estimates the marginal impact of the explanatory variables taking into account possible correlation among them. This led Pinske (2000) to the interpretation of the marginal integration estimator as a consistent estimator of weak separable components, which, in the case of additivity, coincide with the additive components. From this it can be expected that the distance between the real regression function and its estimate increases especially fast when the data generating regression function is not additive but estimated by the sum of component estimates obtained from marginal integration instead of backfitting or series estimates. A consequence could be to prefer marginal integration for the construction of additivity tests. Nevertheless, until now backfitting was not used for testing simply because of the lack of theory for the estimator.

Due to the mentioned econometric results and statistical advantages there is an increasing interest in testing the additive structure. Eubank, Hart, Simpson and Stefanski (1995) constructed such a test but used special series estimates that apply only on data observed on a grid. Gozalo and Linton (2000) as well as Sperlich, Tjøstheim and Yang (2000) introduced a bootstrap based additivity test applying the marginal integration. Here, Sperlich et al. (2000) concentrated on the analysis of particular interaction terms rather than on general separability. Finally, Dette and von Lieres (2000) have summarized the test statistics considered by Gozalo and Linton (2000) and compared them theoretically and also in a small simulation study. Their motivation for using the marginal integration was its direct definition which allows an asymptotic treatment of the test statistics using central limit theorems for degenerate $U$-statistics. They argued that such an approach based on backfitting seems to be intractable, because their asymptotic analysis does not require the asymptotic properties of the estimators as e.g. derived by Mammen, Linton and Nielson (1999) but an explicit representation of the residuals. Further, Dette and Munk (1998) pointed out several drawbacks in the application of Fourier series estimation for checking model assumptions. For these and the former mentioned reasons we do not consider series estimators for the construction of tests for additivity in this paper.

For the empirical researcher it would be of essential interest how the different methods perform in finite samples and which method should be preferred. Therefore the present article is mainly concerned about the practical performance of the different procedures and for a better understanding of some of the above mentioned problems in estimating and testing. Hereby, the main part studies performance, feasibility and technical differences of estimation respectively testing procedures based on different estimators. We concentrate especially on the differences caused by the use of different (pre-)smoothers in marginal integration, in particular on the classic approach of Linton and Nielsen (1995) and on the internalized Nadaraya–Watson estimator (Jones, Davies and Park 1994) as suggested by Kim, Linton and Hengartner (2000). Notice that this study is not thought as an illustration of the general statement of consistency and convergence. Our main interest is directed to the investigation and comparison of finite sample behavior of these procedures.

The marginal integration estimator becomes inefficient with increasing correlation in the regressors, see Linton (1997). He suggested to combine the marginal integration with a one step backfitting afterwards to reach efficiency. Unfortunately, this combination destroys any interpretability of the estimate when the additivity assumption is violated. The same loss of efficiency was also observed in a simulation study by Sperlich, Linton and Härdle (1999) for the backfitting estimator, although these results do not reflect the asymptotic theory. In their article it is further demonstrated that with increasing dimension the additive components are still estimated with a reasonable precision, whereas the estimation of the regression function becomes problematic. This fact could cause problems for prediction and for bootstrap tests. We will investigate and explain that the use of the internalized Nadaraya–Watson estimator for the marginal integration can partly ameliorate this problem. This is actually not based on theoretical results but more on numerical circumstances respective the handling of "poor data areas". Throughout this paper we will call the classical marginal integration estimator CMIE, and IMIE the one using the internalized Nadaraya–Watson estimator as multidimensional pre-smoother.

The rest of the paper is organized as follows. In Section 2 we give the definitions of the analyzed estimators and some more discussion about their advantages and disadvantages. Finally we provide some simulation results on the Cross-Validation mean squared errors for the different methods of estimation. In Section 3 we introduce various test statistics based in the IMIE to check the additivity assumption, present closed form asymptotic theory and a theoretical comparison. Notice that for the IMIE, at least for testing, little theory has been done until now and hardly empirical studies. Therefore we provide both in this work, an extensive simulation study but also a closed theory about the asymptotic properties for any new estimator and test we are considering. Section 4 finally is dedicated to an intensive simulation study for these test statistics, all using bootstrap methods. The proofs of the asymptotic results are cumbersome and deferred to the Appendix in Section 5.

## 2 Marginal Integration and Additive Models

Let us consider the general regression model

$$Y = m(X) + \sigma(X)\varepsilon \tag{2.1}$$

where $X = (X_1, \ldots, X_d)^T$ is a $d$-dimensional random variable with density $f$, $Y$ is the real valued response, and $\varepsilon$ the error, independent of $X$ with mean 0 and variance 1. Further, $m, \sigma$ are unknown (smooth) functions and the regression function $m(\cdot)$ has to be estimated nonparametrically. As indicated above the marginal integration estimator is constructed to catch the marginal impact of one or some regressors $X_\alpha \in I\!R^{d_\alpha}$, $d_\alpha < d$. For the ease of notation we will restrict ourselves to the case $d_\alpha = 1$ for all $\alpha$. Notice first that in case of additivity, i.e. there exist functions $m_\alpha$, $m_{-\alpha}$ such that

$$m(X) = m_\alpha(X_\alpha) + m_{-\alpha}(X_{-\alpha}) \tag{2.2}$$

with $X_{-\alpha}$ being the vector $X$ without the component $X_\alpha$, the marginal impact of $X_\alpha$ corresponds exactly to the additive component $m_\alpha$. For identification we set $E[m_\alpha(X_\alpha)] = 0$ and consequently $E[Y] = E[m_{-\alpha}(X_{-\alpha})] = c$. The marginal integration estimator is defined noting that

$$E_{X_{-\alpha}}[m(x_\alpha, X_{-\alpha})] = \int m(x_\alpha, x_{-\alpha}) f_{-\alpha}(x_{-\alpha}) dx_{-\alpha} \tag{2.3}$$

$$= E_{X_{-\alpha}}[m_{-\alpha}(X_{-\alpha}) + m_\alpha(x_\alpha)] = c + m_\alpha(x_\alpha), \tag{2.4}$$

where $f_{-\alpha}$ denotes the marginal density of $X_{-\alpha}$, and the second line follows from the first line in the case of additivity, see equation (2.2). So marginal integration yields the function $m_\alpha$ up to a constant that can easily be estimated by the average over the observations $Y_i$. We estimate the right hand side of equation (2.3) by replacing the expectation by an average and the unknown mutidimensional regression function $m$ by a pre-smoother $\tilde{m}$. Certainly, having a completely additive separable model of the form

$$m(X) = c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha), \tag{2.5}$$

this method can be applied to estimate all components $m_\alpha$, and finally the regression function $m$ is estimated by summing up an estimator $\hat{c}$ of $c$ with the estimates $\hat{m}_\alpha$.

### 2.1 Formal Definition

Although the pre-smoother $\tilde{m}$ could be calculated applying any smoothing method, theory has always been derived for kernel estimators [note that the same happened to the backfitting (Opsomer and Ruppert 1997, Mammen, Linton and Nielsen 1999)]. Therefore we will concentrate only on the kernel based definitions even though spline implementation is known to be computationally more advantageous. We first give the definition of the classic marginal integration method (CMIE). Let $K_i(\cdot)$ $(i = 1, 2)$ denote one - and $(d-1)$ - dimensional Lipschitz - continuous kernels of order $p$ and $q$, respectively, with compact support, and define for a bandwidth $h_i > 0$, $i = 1, 2$, $t_1 \in I\!R$, $t_2 \in I\!R^{d-1}$

$$K_{1,h_1}(t_1) = \frac{1}{h_1} K_1\left(\frac{t_1}{h_1}\right), \qquad K_{2,h_2}(t_2) = \frac{1}{h_2^{d-1}} K_2\left(\frac{t_2}{h_2}\right). \tag{2.6}$$

For the sample $(X_i, Y_i)_{i=1}^n, X_i = (X_{i1}, \ldots, X_{id})^T$ the CMIE is defined by

$$\hat{m}_\alpha(x_\alpha) = \frac{1}{n}\sum_{j=1}^n \tilde{m}(x_\alpha, X_{j,-\alpha}) = \frac{1}{n^2}\sum_{k=1}^n\sum_{j=1}^n \frac{K_{1,h_1}(X_{j\alpha} - x_\alpha)K_{2,h_2}(X_{j,-\alpha} - X_{k,-\alpha})Y_j}{\hat{f}(x_\alpha, X_{k,-\alpha})} \quad (2.7)$$

$$\hat{f}(x_\alpha, x_{-\alpha}) = \frac{1}{n}\sum_{i=1}^n K_{1,h_1}(X_{i,\alpha} - x_\alpha)K_{2,h_2}(X_{i,-\alpha} - x_{-\alpha}) \quad (2.8)$$

$$\hat{c} = \frac{1}{n}\sum_{j=1}^n Y_j \quad (2.9)$$

and $X_{i,-\alpha}$ denotes the vector $X_i$ without the component $X_{i\alpha}$. Note that $\hat{f}$ is an estimator of the joint density of $X$ and $\tilde{m}$ denotes the Nadaraya Watson estimator with kernel $K_{1,h_1} \cdot K_{2,h_2}$.

The modification giving us the internalized marginal integration estimate (IMIE) concerns the definition of $\hat{m}$, equation (2.7), where $\hat{f}(x_\alpha, X_{k,-\alpha})$ is substituted by $\hat{f}(X_{j\alpha}, X_{j,-\alpha})$, see Jones, Davies and Park (1994) or Kim, Linton and Hengartner (2000) for details. The resulting definition of the IMIE is

$$\hat{m}_\alpha^I(x_\alpha) = \frac{1}{n^2}\sum_{k=1}^n\sum_{j=1}^n \frac{K_{1,h_1}(X_{j\alpha} - x_\alpha)K_{2,h_2}(X_{j,-\alpha} - X_{k,-\alpha})Y_j}{\hat{f}(X_{j\alpha}, X_{j,-\alpha})} \quad (2.10)$$

$$= \frac{1}{n}\sum_{j=1}^n K_{1,h_1}(X_{j\alpha} - x_\alpha)\frac{\hat{f}_{-\alpha}(X_{j,-\alpha})}{\hat{f}(X_{j\alpha}, X_{j,-\alpha})}Y_j, \quad (2.11)$$

where $\hat{f}_{-\alpha}$ is an estimate of the marginal density $f_{-\alpha}$. Notice that the fraction before $Y_j$ in (2.11) is the inverse of the conditional density $f_{\alpha|-\alpha}(X_\alpha|X_{-\alpha})$. It is well known that under the hypothesis of an additive model $\hat{m}_\alpha$ and $\hat{m}_\alpha^I$ are consistent estimates of $m_\alpha \quad (\alpha = 1, \ldots, d)$ (see Tjøstheim and Auestad, 1994, and Kim, Linton and Hengartner, 2000).

## 2.2 On a Better Understanding of Marginal Integration

Although the papers of Nielsen and Linton (1997) and Sperlich, Linton and Härdle (1999) already emphasized the differences of backfitting and marginal integration, often they are still interpreted as competing estimators for the same aim. For a better understanding of the difference between orthogonal projection into the additive space (backfitting) and measuring the marginal impact (marginal integration) we give two more examples.

As has been explained in Stone (1994) and Sperlich, Tjøstheim and Yang (2000), any model can be written in the form

$$m(x) = c + \sum_{\alpha=1}^d m_\alpha(x_\alpha) + \sum_{1\le\alpha<\beta\le d} m_{\alpha\beta}(x_\alpha, x_\beta) + \sum_{1\le\alpha<\beta<\gamma\le d} m_{\alpha\beta\gamma}(x_\alpha, x_\beta, x_\gamma) + \cdots . \quad (2.12)$$

The latter mentioned article, even when they worked it out in detail only for second order interactions, showed that all these components can be identified and consistently estimated by marginal integration obtaining the optimal convergence rate in smoothing. The main reason for this nice property is, that definition, algorithm and thus the numerical results for the estimates do not differ whatever the chosen extension or the true model is. This certainly is different for an orthogonal projection. At first we note that so far model (2.12) can not be estimated by backfitting. Secondly, Stone (1994) gives (formal) algorithms and convergence rates for series estimators to estimate (2.12); but for each interaction added, the whole procedure changes, i.e. has to be redone and gives different numerical results.

Our second example is a simulation of a by far not additive model. We created a sample of $n = 250$ observations generated as

$$\begin{aligned} X &\sim N\left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.0 & 0.2 & 0.4 \\ 0.2 & 1.0 & 0.6 \\ 0.4 & 0.6 & 1.0 \end{pmatrix} \right), \\ Y &= X_1 \exp\left( \frac{X_2 + X_3}{4} \right) + X_1 \left| X_2 + X_3 \right| - 3X_3 \end{aligned} \tag{2.13}$$

and applied the backfitting as well as the marginal integration estimate (IMIE), always with quartic kernels and bandwidth 1.0 for all directions. Notice that we did not add any error term when creating the observation $Y$ and thus get quite fair projections for both procedures (Figure 1) highlighting some of the main differences of these estimators. Even if we fade out the boundary effects (in Figure 1 the estimates on the outer points are omitted) we see clearly that the slopes can even go into contrary directions although both procedures do what they should. However, summing up the estimated components, both estimators would give ridiculous predictors for such a non-additive regression model.

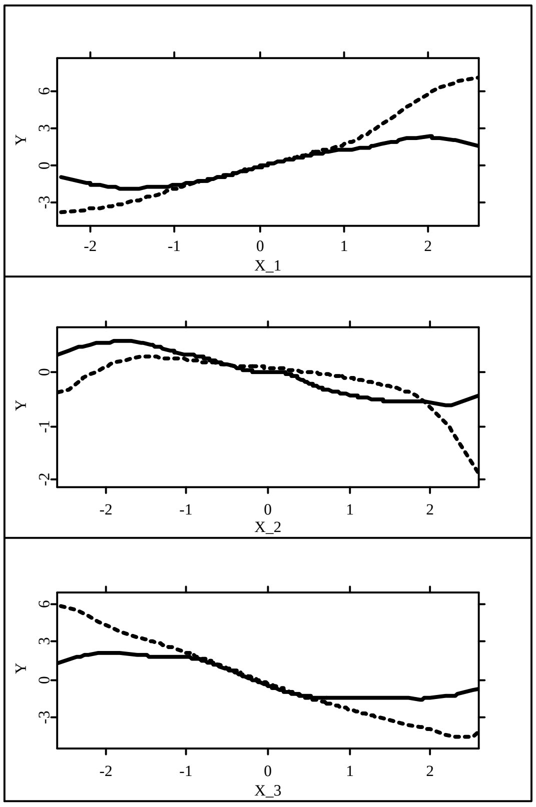


Figure 1: *Backfitting estimates (dashed) and IMIE (solid) using the same data generated according to equation (2.13).*

Recall now the definitions of the CMIE and the IMIE in Section 2.1 and let us discuss their differences.

Obvious advantages of the IMIE are the possibility of changing the sums and getting rid of the $x_\alpha$ in the density estimates, see (2.11). Camlong-Viot (2000) chose this for the simpler theoretical analysis of this estimator while Hengartner (1996) showed that the bandwidths conditions for the nuisance corrections depend only on the smoothness of the densities but not, as for the CMIE, on the smoothness of their component functions. Kim, Linton and Hengartner (2000), finally, underlined the possibility of fast implementation. Notice that calculating the regression at all points demands $O(n^3)$ steps for the CMIE whereas only $O(n^2)$ for the IMIE. Assuming smart implementation and large memory the IMIE can thus be pretty fast even in its kernel-smoothing version. This point is of special importance when statistics for testing additivity have to be calculated and bootstrap must be applied.

A main advantage of the CMIE is the straight forward extension to local polynomial smoothing. This is not only reducing the bias but also enables us to (efficient) estimation of derivatives, which is an important problem in economics (e.g. estimation of elasticities, return to scales, substitution rates, etc.). Such an extension is much harder to find for the IMIE. An important point is that the asymptotic expressions for both estimators are all the same. Thus, differences can only be found by a perfect understanding of what each procedure is doing to the data and intensive simulation studies.

So far not investigated are differences in the finite sample performance. The first question is certainly whether the simplification done in the definition of the IMIE is only negligible asymptotically. The second question refers to the bad performance of the CMIE when the covariates are strongly correlated, see discussion above. Let us consider Figure 2 to understand and explain why here the IMIE performs somehow better. On the left side the points, including the filled circles, represent a two dimensional normally distributed design with expectation zero, variance one and covariance 0.8, $n = 200$ observations. The circle in the upper left ($\circ$) is a combination of the $X_1$-value of the lower left with the $X_2$-value of the upper right filled circle ($\bullet$). The point we want to make here is that the marginal integration estimator tries to predict the pre-smoother ( $\tilde{m}$ ) on **all** combinations of $X_1$- and $X_2$- values, e.g. also for the point presented by the circle $\circ$, and not only on the sample data. Thus, the more correlated the covariates are, the more this means extrapolation as e.g. in the case indicated by the circles. Certainly, those extrapolations using smoothing methods often break down and this explains the bad performance. In the extreme case of having a design as in Figure 2 on the right hand side, no one of the so far developed additive estimators yields reasonable estimates at all. The IMIE now is less affected by this problem since the expression $\hat{f}(x_\alpha, X_{k-\alpha})$ has been substituted by $\hat{f}(X_{j\alpha}, X_{j-\alpha})$. Thus it does not estimate $f$ on data empty areas but only on sample data. This gives some hope that the IMIE might perform better when data are sparse, and in particular when covariates are correlated. A further trial to circumvent the problem would be to integrate (average) over $\tilde{m}$, equation (2.7), only inside the convex hull formed by the observations $\{X_i\}_{i=1}^n$. Let us call this modification the MMIE in the following. Under these considerations, the simulation results in the next section are surprising.

Finally, for the sake of completeness, we have to mention the inclusion of variance minimizing weights in the integration, introduced by Fan, Härdle, Mammen (1998). However, since those weights consist only of nontrivial unknown parametric and nonparametric expressions, these weights first have to be estimated appropriately. Moreover, their estimation algorithm already needs $O(n^4)$ calculation steps and is thus out of discussion for the construction of test statistics, bootstrap or detailed simulation studies.

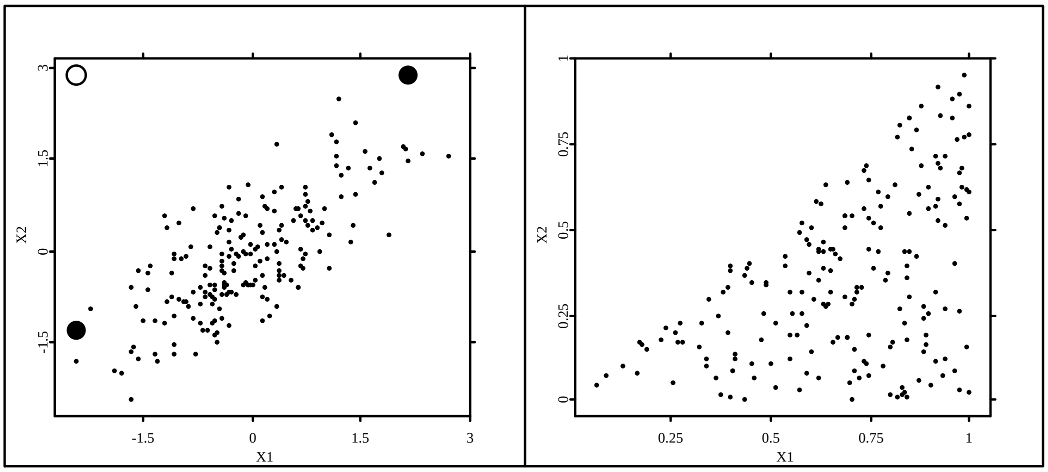


Figure 2: *Problematic designs for additive modeling. left side:* $\bullet$, $\cdot$ *are sample data, and* $\circ$ *is a cross combination of the two* $\bullet$.

### 2.3 Some Simulation Results

Since asymptotically all methods are consistent, differences and problems can better be observed for small samples. Consequently we did simulations only with $n = 100$ and 200 observations and report mainly the results for $n = 100$. The conclusions for $n = 200$ are all the same. In these samples the functional forms of the components can still be estimated with reasonable precision. The comparison was done by calculating the Cross-Validation value for all estimators over a wide range of bandwidths, where the CV-value was calculated on the whole support as well as on trimmed ones in order to get an idea about the importance of boundary effects. For a better comparison and to see the advantages of additive modeling we always give additionally the CV-values for the multi dimensional Nadaraya-Watson Smoother (NWS). We averaged the CV-values of 100 runs for all bandwidth combinations. The results in the tables below refer to the smallest (average) CV-value obtained for the particular estimation method. For each run we draw a new sample since it is demonstrated in Sperlich, Linton, Härdle (1999) that in small samples the mean squared error varies substantially with the design even if the design is drawn from the same distribution. Thus a comparison of methods based on just one fixed design could be biased by chance in favor of one of the considered methods. The bandwidths were chosen from $h = 0.25 std(X)$ to $1.6 std(X)$ where $std(X)$ is the vector of the empirical standard deviations of the particular design. Since the IMIE allows for oversmoothing in the nuisance directions we chose $h_1 = h$ but $h_2 = Nh$ with $N \in I\!N$ from 1 to 8 knowing that this can lead to suboptimal results. For the NWS, CMIE, and MMIE we set $h_1 = h_2 = h$. In all calculations the quartic kernel was applied.

As a first example consider the two dimensional model

$$Y = X_1^2 + 2\sin(0.5\pi X_2) + \varepsilon, \tag{2.14}$$

with $\varepsilon \sim N(0,1)$, $(X_1, X_2)^T \sim N\{0, \Sigma_\gamma\}$, $\gamma = 1, 2, 3$, where the covariance matrices are given by

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \ \Sigma_2 = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}, \ \Sigma_3 = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}. \tag{2.15}$$

The CV-values were calculated on the whole (tr0) and on the trimmed supports cutting at 1.96 (tr5), respectively 1.645 (tr10) in each direction. This corresponds approximately to a trimming of 5%, respectively 10% of the support. The results for $n = 100$ observations are given in Table 1.

Mean Squared Errors by Cross Validation, dimension $d = 2$

| used | $\Sigma_1$ | | | $\Sigma_2$ | | | $\Sigma_3$ | | |
|---|---|---|---|---|---|---|---|---|---|
| estimator | tr0 | tr5 | tr10 | tr0 | tr5 | tr10 | tr0 | tr5 | tr10 |
| NWS | 1.976 | 1.296 | 1.214 | 1.907 | 1.295 | 1.214 | 1.694 | 1.280 | 1.218 |
| CMIE | 1.715 | 1.196 | 1.147 | 1.701 | 1.205 | 1.150 | 1.678 | 1.254 | 1.185 |
| MMIE | 1.826 | 1.212 | 1.156 | 1.821 | 1.222 | 1.161 | 1.824 | 1.288 | 1.201 |
| IMIE | 1.893 | 1.219 | 1.173 | 1.857 | 1.223 | 1.177 | 1.751 | 1.276 | 1.211 |

Table 1: *Average CV-value of the different estimators over 100 runs for optimal (i.e. CV-minimizing) bandwidths. The data were drawn from model (2.14), $n = 100$, with covariances $\Sigma_\gamma$ from (2.15). The CV-values were calculated on the whole support (tr0), and on trimmed ranges ( 5% trimming: tr5, respectively 10%: tr10).*

Obviously, all estimators are very close. Surprisingly, the modification in the MMIE leads to a worse performance.

We increase now the data sparseness by increasing the dimension keeping $n = 100$ observations. Consider the model

$$Y = X_1 + X_2^2 + 2\sin(\pi X_3) + \varepsilon, \tag{2.16}$$

with $\varepsilon \sim N(0,1)$, $(X_1, X_2)^T \sim N\{0, \Sigma_\gamma\}$,

$$\Sigma_\gamma = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}, \tag{2.17}$$

where we set for $\gamma = 1$: $\rho_{12} = \rho_{13} = \rho_{23} = 0$, $\gamma = 2$: $\rho_{12} = 0.2$, $\rho_{13} = 0.4$, $\rho_{23} = 0.6$, and $\gamma = 3$: $\rho_{12} = 0.4$, $\rho_{13} = 0.6$, $\rho_{23} = 0.8$. For the calculation of the CV-values we used the same trimming at 1.96 (tr5), respectively 1.645 (tr10) in each direction. The results are given in Table 2.

It can be observed from these results that now all the above mentioned problems have appeared. Due to the sparseness of data the IMIE substantially outperforms the other methods, especially in the presence of high correlation. Moreover, in this case even the multidimensional NWS is better than CMIE and MMIE. Further, the boundary effects are substantial in both cases now (compare the trimmed with the untrimmed results, i.e. tr0 with tr5 or tr10). We can conclude that our heuristical arguments in favor of the the IMIE for increasing dimension and correlation in relatively small samples have been confirmed by this empirical study.

# 3 Testing Additivity

As discussed in the introduction, the additivity of the regression function is important in terms of interpretability and its ability to deliver fast rates of convergence in estimating the regression. For

Mean Squared Errors by Cross Validation, dimension $d = 3$

| used | $\Sigma_1$ | | | $\Sigma_2$ | | | $\Sigma_3$ | | |
|---|---|---|---|---|---|---|---|---|---|
| estimator | tr0 | tr5 | tr10 | tr0 | tr5 | tr10 | tr0 | tr5 | tr10 |
| NWS | 3.743 | 2.538 | 2.314 | 3.378 | 2.346 | 2.169 | 3.112 | 2.147 | 2.023 |
| CMIE | 2.641 | 1.781 | 1.699 | 2.781 | 1.994 | 1.881 | 3.158 | 2.378 | 2.242 |
| MMIE | 3.136 | 2.032 | 1.873 | 3.630 | 2.406 | 2.173 | 4.511 | 3.049 | 2.678 |
| IMIE | 2.477 | 1.606 | 1.522 | 2.540 | 1.777 | 1.667 | 2.477 | 1.606 | 1.522 |

Table 2: *Average CV-value of the different estimators over 100 runs for optimal (i.e. CV-minimizing) bandwidths. The data were drawn from model (2.16), $n = 100$, with covariances $\Sigma_\gamma$ from (2.17). The CV-values were calculated on the whole support (tr0), and on trimmed ranges ( 5% trimming: tr5, respectively 10%: tr10).*

these reasons the additive model should be accompanied by an adequate model check. Tests for the hypothesis of additivity have recently found considerable interest in the literature, see the references in Section 1. In this section we investigate several tests but will only concentrate on statistics based on residuals from an internal marginal integration fit. For asymptotic properties of tests based on residuals from the classical marginal integration fit we refer to Gozalo and Linton (2000) and Dette and von Lieres (2000). We prove asymptotic normality of the corresponding test statistics under the null hypothesis of additivity and fixed alternatives with different rates of convergence corresponding to both cases. Note that we are able to find the asymptotic properties of the tests under any fixed alternative of non-additivity. These results can be used for the calculation of the probability of the type II error of the corresponding tests and for the construction of tests for precise hypotheses as proposed in Berger and Delampady (1987) or Staudte and Sheather (1990). These authors point out that often it is preferable to reformulate the hypothesis

$$H_0 : m(x) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha), \tag{3.1}$$

into

$$H_\eta : M^2 > \eta \ , \qquad H_1 : M^2 \leq \eta \ , \tag{3.2}$$

where $M^2$ is a nonnegative measure of additivity, which is equal to 0, if and only if the hypothesis (3.1) holds. In equation (3.2) the quantity $\eta$ 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 (3.1) ($M^2 = 0$) but also at any point of the alternative ($M^2 > 0$).

The regression estimator based on the internalized marginal integration is defined by

$$\hat{m}_0^I(x) = \sum_{\alpha=1}^{d} \hat{m}_\alpha^I(x_\alpha) - (d-1)\hat{c}, \tag{3.3}$$

where $\hat{m}_\alpha^I$ is given in (2.11), and the residuals from this fit are denoted by $\hat{e}_j = Y_j - \hat{m}_0^I(X_j)$ $(j = 1, \ldots, n)$. The test statistics we consider are

$$T_{1n} = \frac{1}{n}\sum_i \left[\hat{m}^I(X_i) - \hat{m}_0^I(X_i)\right]^2 \tag{3.4}$$

$$T_{2n} = \frac{1}{n}\sum_i \hat{e}_i \left[\hat{m}^I(X_i) - \hat{m}_0^I(X_i)\right] \tag{3.5}$$

$$T_{3n} = \frac{1}{n}\sum_i \left[(\hat{e}_i)^2 - (\hat{u}_i)^2\right] \tag{3.6}$$

$$T_{4n} = \frac{1}{n(n-1)}\sum_{j\neq i} L_g(x_i - x_j)\hat{e}_i\hat{e}_j \tag{3.7}$$

where $L$ is a bounded $d$-dimensional symmetric kernel of order $r$ with compact support, $L_g(\cdot) = \frac{1}{g^d}L_g(\frac{\cdot}{g})$, $g > 0$. In (3.4) $\hat{m}^I$ is the internalized Nadaraya - Watson estimator with kernel $L$ and bandwidth $h$, and the random variables $\hat{u}_i = Y_i - \hat{m}^I(X_i)$ in (3.6) denote the corresponding residuals. The estimate $T_{1n}$ compares the 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. The statistic $T_{2n}$ was introduced by Gozalo and Linton (2000) and was motivated by Lagrange multiplier tests of classical statistics. $T_{3n}$ is essentially a 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 for parametric structures of the regression function. Statistics like $T_{4n}$ were originally introduced by Zheng (1996) and independently discussed by Fan and Li (1996, 1999) and Gozalo and Linton (2000) in the problem of testing additivity. In the following section we investigate the asymptotic behavior of these statistics under the null hypothesis and fixed alternatives. It is demonstrated that the asymptotic behavior of the statistics $T_{1n}$ to $T_{4n}$ under fixed alternatives is rather different and we will indicate potential applications of such results.

### 3.1 Theoretical Results

For the ease of notation we suppose that $h_1 = h_2 = h$, $K_1 = K$ and $K_2 = K^{\otimes(d-1)}$, where $K$ is a bounded Lipschitz - continuous kernel of order $r > d$ with compact support. We assume that the following assumptions are satisfied.

*(A1) The explanatory variable $X$ has a density $f$ supported on the cube $Q = [0,1]^d$. $f$ is bounded from below by a positive constant $c > 0$ and has continuous partial derivatives of order $r$.*

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

*(A3) $\sigma \in C_b(Q)$, where $C_b(Q)$ denotes the class of bounded continuous functions (defined on $Q$). Furthermore, $\sigma$ is bounded from below by a positive constant $c > 0$.*

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

*(A5) The bandwidths $g, h > 0$ satisfy*

$$\lim_{n\to\infty} nh^{2r} < \infty, \; \lim_{n\to\infty} \frac{\log n}{nh^{d+r}} = 0, \; \lim_{n\to\infty} \frac{1}{ng^d} = 0, \; \lim_{n\to\infty} ng^{2r} = 0, \; \lim_{n\to\infty} \frac{g^d}{h^2} = 0$$

Notice that with the last condition, it is sufficient to express all convergence rates only in terms of bandwidth $g$.

**Theorem 1** *Assume that assumptions (A1) - (A5) are satisfied and $T_{1n}, \ldots, T_{4n}$ are defined in (3.4) - (3.7) and define*

$$m_0(x) := \sum_{\alpha=1}^{d} \int m(x_\alpha, y_{-\alpha}) f_{-\alpha}(y_{-\alpha}) dy_{-\alpha} + (d-1) \int m(y) f(y) dy. \tag{3.8}$$

(i) *Under the null hypothesis of additivity, i.e. $m = m_0$, we have, as $n \to \infty$:*

$$n g^{\frac{d}{2}} (T_{jn} - \beta_{jn}) \xrightarrow{D} N\left(0, \lambda_j^2\right) \quad (j = 1, \ldots, 4) \tag{3.9}$$

*where*

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

*$\beta_{4n} = 0$ and the asymptotic variances are given by*

$$\begin{aligned}
\lambda_1^2 &= 2 \int \sigma^4(x) dx \int (L * L)^2(x) dx, \\
\lambda_2^2 &= 2 \int \sigma^4(x) dx \int L^2(x) dx, \\
\lambda_3^2 &= 2 \int \sigma^4(x) dx \int (2L - L * L)^2(x) dx \\
\lambda_4^2 &= 2 \int \sigma^4(x) f^2(x) dx \int L^2(x) dx.
\end{aligned} \tag{3.10}$$

(ii) *If the regression function is not additive, i.e. $\Delta = m - m_0 > 0$, then we have*

$$\sqrt{n} \left(T_{jn} - M_j^2 - \beta_{jn}\right) \xrightarrow{D} N\left(0, \mu_j^2\right) \quad (j = 1, \ldots, 4), \tag{3.11}$$

*where*

$$M_j^2 = \int (\Delta^2 f)(x) dx \quad (j = 1, \ldots, 3), \quad M_4^2 = \int (\Delta^2 f^2)(x) dx. \tag{3.12}$$

*The asymptotic variances are given by*

$$\begin{aligned}
\mu_j^2 &= 4 \int \sigma^2(x) \{\Delta(x) - p(\Delta)(x)\}^2 f(x) dx \\
&\quad + \int \{\Delta^2(x) - q_j(\Delta)(x)\}^2 f(x) dx - \left(\int \{\Delta^2(x) - q_j(\Delta)(x)\} f(x) dx\right)^2
\end{aligned}$$

*(j=1,...,3),*

$$\begin{aligned}
\mu_4^2 &= 4 \int \sigma^2(x) \{(\Delta f)(x) - p(\Delta f)(x)\}^2 f(x) dx \\
&\quad + \int \{2(\Delta^2 f)(x) - q_4(\Delta f)(x)\}^2 f(x) dx - \left(\int \{2(\Delta^2 f)(x) - q_4(\Delta f)(x)\} f(x) dx\right)^2,
\end{aligned}$$

*where $p$, $q_j$ $(j = 1, \ldots, 4)$ denote mappings defined by*

$$p(g)(x) \quad : \quad = \sum_{\alpha=1}^{d} f^{-1}(x) f_{-\alpha}(x_{-\alpha}) \int (gf)(x_\alpha, y_{-\alpha}) \, dy_{-\alpha} - (d-1) \int (gf)(y) \, dy, \tag{3.13}$$

$$q_j(g)(x) \quad : \quad = (2p(g)(x) - k_j \Delta(x)) m(x) + 2 \sum_{\alpha=1}^{d} \int (gf)(y) m(y_\alpha, x_{-\alpha}) dy$$

$(j = 1, \ldots, 4)$ *and the constants $k_j$ are given by $k_1 = 2,\ k_2 = 1,\ k_3 = 0$ and $k_4 = 0$, respectively.*

Note that the first part of Theorem 1 shows that a test of additivity with asymptotic level $\alpha$ can be obtained by rejecting the null hypothesis of additivity, if

$$ng^{\frac{d}{2}} \frac{T_{jn} - \beta_{jn}}{\lambda_j} > z_{1-\alpha}, \qquad j = 1, \ldots, 4$$

where $z_{1-\alpha}$ denotes the $(1-\alpha)$ quantile of the standard normal distribution and $\beta_{jn}$ $\lambda_j$ have to be replaced by appropriate estimators. It should be mentioned that all these tests serve in practice for small samples only when the critical values are determined by bootstrap methods. We will come back to this point in the next section. Note further that Gozalo and Linton (2000) and Dette and von Lieres (2000) considered weight functions in the definition of the corresponding test statistics based on residuals from the classical marginal integration fit. However, the latter authors found out that, if trimming is not necessary for numerical reasons, the optimal weights are the uniform ones. That is why we skipped them here. Again, a comparison with the asymptotic theory for these statistics when using the CMIE reveals that under the null hypothesis the asymptotics are all the same except the smoothness and bandwidths conditions (see Gozalo and Linton, 2000, or Dette and von Lieres, 2000). However, under fixed alternatives we find substantial differences in the asymptotic variances (see Dette and von Lieres, 2000).

What can we get out from Theorem 1 about the different quality of the proposed test statistics? Obviously it depends on many factors like the density of the covariates, kernel choice, error variance function, and the functional $\Delta = m - m_0$ which test has more power. We can mainly detect three points. Assuming a sufficiently smooth regression function $m$, so that we get under the alternative $H_1$ the bias

$$E_{[H_0]}[T_{jn}] = M_j^2 + \beta_{jn} + o\left(\frac{1}{\sqrt{n}}\right), \qquad j = 1, \ldots, 4 ,$$

the probability of rejection (if the hypothesis of additivity is not valid) is approximately equal to

$$\Phi\Big(\frac{\sqrt{n}}{\mu_j}\Big\{M_j^2 - \frac{z_{1-\alpha}\lambda_j}{n\sqrt{g^d}}\Big\}\Big) ,$$

where $\Phi$ is the cumulated standard normal distribution function, and $\mu_j, M_j, \lambda_j$ are defined in Theorem 1. Here, an appropriate weighting in $T_{4n}$ (see Dette and von Lieres 2000) would lead to $M_4 = M_j$, $j = 1, 2, 3$ with $\mu_4 > \mu_j$. But as that particular weighting is not optimal we can not conclude a uniform inferiority of $T_{4n}$ to the others. In contrast, if we next look at the biases $\beta_{jn},\ j = 1, \ldots, 4$ which can be rather influential in samples of common size, we notice that only $\beta_{4n}$ is equal to zero.

Consequently, $T_{4n}$ might give more reliable results in such cases. Finally, coming back to variance considerations: Since

$$\int (L*L)^2(x)dx \leq \int L^2(x)dx \leq \int (2L - L*L)^2(x)dx,$$

see Dette and von Lieres (2000), it can be seen from Theorem 1 that $\lambda_1^2 \leq \lambda_2^2 \leq \lambda_3^2$. But again, all in all there are to many factors that have to be taken into account for making a theory based statement about possible superiority of one of the considered tests. Therefore we will include them all in the simulation study of the next section.

# 4 Simulation Comparison of Additivity Tests

In this section we continue the considerations of the last part of Section 2 but extend them to the various (bootstrap) tests for checking additivity. We concentrate especially on the differences caused by the use of different pre-smoothers, i.e. we compare CMIE with IMIE, but certainly also consider differences between $T_{1n}$ to $T_{4n}$. Finally, we compare the difference in performance between tests using the bootstrap based on residuals taken from $Y - \hat{m}_0$ ($B_0$), as e.g. Gozalo and Linton (2000) or Härdle and Mammen (1993), versus bootstrap based on residuals taken from $Y - \hat{m}$ ($B_1$) as e.g. Dette and von Lieres (2000). For the sake of simplicity we omit the index $n$ in the definition of the statistics in this section and write $T_i = T_{in}$, $i = 1, \ldots, 4$. Notice that this section is not thought as an illustration of the general statement of consistency and convergence for the former presented tests. Our interest is directed to the investigation and comparison of feasibility and finite sample performance.

We took always the bandwidths minimizing the average of the CV values for trimming tr5 and covariance $\Sigma_2$. Again, we report simulations only for $n = 100$ and $n = 200$ observations when the functional forms of the additive components seem still to be estimated reasonably well. However, since now the estimation of the regression function is crucial the bootstrap tests can already give ridiculous results. Due to computational restrictions we did the simulations only for 500 bootstrap replications. The results refer to 1000 simulation runs with a randomly drawn design for each run, see above for explanation.

## 4.1 The case $d = 2$

As a first example we consider the two dimensional model (2.14), (2.15) but adding the interaction term $aX_1X_2$ with $a$ being a scalar, i.e.

$$Y = X_1^2 + 2\sin(0.5\pi X_2) + aX_1X_2 + \varepsilon. \tag{4.1}$$

Our bandwidth selection yields for the Nadaraya-Watson estimator a bandwidth $h = 0.85std(X)$, for the CMIE $h_1 = h_2 = 0.7std(X)$, and for the IMIE $h_1 = 0.85std(X)$. As for the IMIE the optimal $k = h_2/h_1$ was either 2 or 3, depending on the correlation of the regressors and trimming, we set $h_2 = 2.5 * h_1$. Finally, since results for the test statistic $T_4$ depend strongly on the choice of bandwidth $g$, we tried out various bandwidths and report the results for $0.1std(X)$ ($g_1$), and $0.2std(X)$ ($g_2$).

Note that for the ease of presentation all tables will have the same structure. We give the percentage of rejections for the one and the five percent level for all test statistics, without trimming ($tr0 = \infty$), and at the approximate 95% quantile ($tr5 = 1.96$), respectively the 90% ($tr10 = 1.645$). In the left part of each Table the results are given under the null hypothesis of additivity, i.e. for scalar $a = 0.0$; in the right part we present results under some alternative ($a = 1.0$). Tables for independent and correlated designs are separated. Our first finding is that tests using the bootstrap based on residuals $Y_i - \hat{m}_0(X_i)$ under the null hypothesis of additivity work much better than the application of residuals from the general model, i.e. $Y_i - \hat{m}(X_i)$. On the one hand these tests are more reliable with respect to the accuracy of the approximation of the level, on the other hand they yield the same power as the bootstrap test obtained from general residuals. For these reasons all results presented here and in the following are based on bootstrap taking residuals under the null hypothesis. In Table 3 we give the first results for the CMIE, independent design, in Table 4 the corresponding ones for the IMIE.

CMIE, classic marginal integration estimator, $cov(X_1, X_2) = 0.0$, $d = 2$

| | under the null model, $a = 0.0$ | | | | | under the alternative, $a = 1.0$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ |
| | without trimming: | | | | | | | | | |
| 5% | 0.05 | 0.082 | 0.04 | 0.03 | 0.026 | 0.098 | 0.29 | 0.562 | 0.53 | 0.462 |
| 1% | 0.002 | 0.008 | 0.008 | 0.006 | 0.002 | 0.008 | 0.052 | 0.304 | 0.232 | 0.146 |
| | with trimming at 1.96: | | | | | | | | | |
| 5% | 0.048 | 0.080 | 0.118 | 0.074 | 0.05 | 0.09 | 0.278 | 0.926 | 0.89 | 0.782 |
| 1% | 0.002 | 0.008 | 0.014 | 0.012 | 0.01 | 0.006 | 0.046 | 0.714 | 0.58 | 0.416 |
| | with trimming at 1.645: | | | | | | | | | |
| 5% | 0.046 | 0.07 | 0.126 | 0.084 | 0.062 | 0.084 | 0.224 | 0.944 | 0.9 | 0.784 |
| 1% | 0.002 | 0.008 | 0.018 | 0.012 | 0.004 | 0.006 | 0.038 | 0.774 | 0.648 | 0.438 |

Table 3: *Percentage of rejection of* 1000 *repetitions applying the various tests on model (4.1) with* $n = 100$ *and independent regressors.* $\alpha$ *gives the wanted significance level. Results are given for* 500 *bootstrap replications using the CMIE.*

We see that all methods seem to work, though if not perfectly for such small samples. Obviously the test based on $T_4$ has the worst power as could be expected from Theorem 1. A difference between the use of CMIE and IMIE can not be detected so far. Trimming has quite different effects for the various methods but does not uniformly improve the performance what might surprise thinking of the strong boundary effects in estimation. We now turn to the (highly) correlated design, when the covariance between the components $X_1$, $X_2$ is 0.8 keeping the variances at 1.0. The results are given in Tables 5, 6.

In general we can see that all methods work much worse for correlated designs. Especially bad performs the test based on $T_1$, and for trimmed statistics also $T_2$ whereas the test based on $T_3$ is the most reliable one. This is rather interesting since from the theoretical results it would have been expected the other way around. These findings hold independently from the used estimator CMIE or IMIE. They mainly differ in the sense that the IMIE produces more conservative tests, but it performs a little bit better when looking at the null model. Summarizing over the various situations none of the test statistics and estimators outperforms significantly all the other ones. However, it is notable that the IMIE is preferable due to its computational advantages being much (exponentially) faster than the CMIE.

IMIE, marginal integration using internalized pre-estimator, $cov(X_1, X_2) = 0.0$, $d = 2$

| $\alpha$ | under the null model, $a = 0.0$ | | | | | under the alternative, $a = 1.0$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ |
| | without trimming: | | | | | | | | | |
| 5% | 0.03 | 0.076 | 0.07 | 0.047 | 0.042 | 0.078 | 0.289 | 0.759 | 0.743 | 0.696 |
| 1% | 0.0 | 0.009 | 0.007 | 0.004 | 0.002 | 0.005 | 0.056 | 0.479 | 0.392 | 0.281 |
| | with trimming at 1.96: | | | | | | | | | |
| 5% | 0.028 | 0.071 | 0.086 | 0.068 | 0.056 | 0.073 | 0.272 | 0.929 | 0.886 | 0.781 |
| 1% | 0 | 0.009 | 0.014 | 0.008 | 0.004 | 0.005 | 0.051 | 0.727 | 0.587 | 0.401 |
| | with trimming at 1.645: | | | | | | | | | |
| 5% | 0.03 | 0.074 | 0.078 | 0.073 | 0.071 | 0.069 | 0.235 | 0.94 | 0.868 | 0.767 |
| 1% | 0 | 0.007 | 0.013 | 0.01 | 0.005 | 0.004 | 0.036 | 0.761 | 0.603 | 0.376 |

Table 4: *Percentage of rejection of* 1000 *repetitions applying the various tests on model (4.1) with* $n = 100$ *and independent regressors.* $\alpha$ *gives the wanted significance level. Results are given for* 500 *bootstrap replications using the IMIE.*

CMIE, classic marginal integration estimator, $cov(X_1, X_2) = .8$, $d = 2$

| $\alpha$ | under the null model, $a = 0.0$ | | | | | under the alternative, $a = 1.0$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ |
| | without trimming: | | | | | | | | | |
| 5% | 0.043 | 0.101 | 0.045 | 0.024 | 0.022 | 0.177 | 0.633 | 0.438 | 0.379 | 0.305 |
| 1% | 0 | 0.012 | 0.003 | 0.001 | 0.001 | 0.014 | 0.215 | 0.204 | 0.133 | 0.082 |
| | with trimming at 1.96: | | | | | | | | | |
| 5% | 0.041 | 0.096 | 0.174 | 0.115 | 0.076 | 0.179 | 0.638 | 0.952 | 0.917 | 0.856 |
| 1% | 0 | 0.008 | 0.029 | 0.018 | 0.012 | 0.013 | 0.213 | 0.773 | 0.654 | 0.498 |
| | with trimming at 1.645: | | | | | | | | | |
| 5% | 0.04 | 0.088 | 0.188 | 0.153 | 0.099 | 0.16 | 0.577 | 0.992 | 0.982 | 0.938 |
| 1% | 0 | 0.008 | 0.04 | 0.024 | 0.016 | 0.012 | 0.192 | 0.912 | 0.82 | 0.674 |

Table 5: *Percentage of rejection of* 1000 *repetitions applying the various tests on model (4.1) with* $n = 100$ *and covariance* $\Sigma_3$ *(2.15), i.e. correlated regressors with* $\delta = .8$ *.* $\alpha$ *gives the wanted significance level. Results are given for* 500 *bootstrap replications using the CMIE.*

## 4.2 The case $d = 3$

As for estimation, also for testing the results change significantly when we increase the dimension of the model. Indeed, even the increase from $d = 2$ to $d = 3$ changes things dramatically. In order to illustrate these effects we consider the model (2.16),(2.17), Section 2, but adding the interaction term

IMIE, marginal integration using internalized pre-estimator, $cov(X_1, X_2) = .8$, $d = 2$

| | under the null model, $a = 0.0$ | | | | | under the alternative, $a = 1.0$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ |
| | without trimming: | | | | | | | | | |
| 5% | 0.035 | 0.093 | 0.08 | 0.058 | 0.043 | 0.091 | 0.355 | 0.403 | 0.337 | 0.255 |
| 1% | 0.001 | 0.008 | 0.009 | 0.004 | 0 | 0.002 | 0.084 | 0.173 | 0.103 | 0.052 |
| | with trimming at 1.96: | | | | | | | | | |
| 5% | 0.035 | 0.089 | 0.158 | 0.092 | 0.065 | 0.085 | 0.347 | 0.881 | 0.802 | 0.65 |
| 1% | 0.001 | 0.007 | 0.017 | 0.009 | 0.005 | 0.004 | 0.075 | 0.626 | 0.422 | 0.204 |
| | with trimming at 1.645: | | | | | | | | | |
| 5% | 0.033 | 0.081 | 0.131 | 0.102 | 0.085 | 0.081 | 0.333 | 0.935 | 0.873 | 0.724 |
| 1% | 0.001 | 0.006 | 0.016 | 0.008 | 0.004 | 0.004 | 0.067 | 0.733 | 0.549 | 0.332 |

Table 6: *Percentage of rejection of* 1000 *repetitions applying the various tests on model (4.1) with* $n = 100$ *and covariance* $\Sigma_3$ *(2.15), i.e. correlated regressors with* $\delta = .8$ *.* $\alpha$ *gives the wanted significance level. Results are given for* 500 *bootstrap replications using the IMIE.*

$aX_2X_3$, i.e.

$$Y = X_1 + X_2^2 + 2\sin(\pi X_3) + aX_2X_3 + \varepsilon. \tag{4.2}$$

The Cross Validation yields for the Nadaraya-Watson an optimal bandwidth $h_1 = 0.9std(X)$, for the CMIE $h_1 = 0.85std(X)$, and for the IMIE $h = 0.7std(X)$, but $h_2 = 6.0 * h_1$ for the IMIE. Results for $T_4$ now refer to bandwidth $g_1 = 0.5std(X)$, and $g_2 = 1.0std(X)$. We skipped the presentation of the results under alternatives in the case of a correlated design because all methods fail for the highly correlated design already under the null when $a = 0.0$. Thus, a power statement or comparison would not make much sense. We will restrict ourselves on some remarks. In Table 7 we give our results for the CMIE, independent design, in Table 8 the corresponding ones for the IMIE.

Comparing the $H_0$ with the $H_1$ case, we see that all methods seem to work when using the IMIE (although the approximation of the nominal level is not too accurate), but they clearly fail applying the CMIE. Note that to emphasize our points we simulated here extreme situations of data sparseness due to dimensionality. Having in mind the size of the sample compared to the complexity of the model, it might be more surprising how well the IMIE works than the bad performance of the CMIE. When we tried the same simulation with $n = 200$ observations the results slightly improved but it seems to us that much bigger samples are needed to reach reliable results when using the CMIE. In this case one kicks in another finding from a computational point of view: Although we had implemented all methods using $(n \times n)$ weighting-matrices to avoid many loops, for sample sizes bigger than $n = 150$ the simulations with the CMIE took about 10 times longer than with the IMIE (measured in days). This is an almost striking argument in favor of the IMIE when it additionally even performs at least equally well. Finally, it is remarkable that the bandwidth $g$ in the test based on the statistic $T_4$ plays a really important rule. Since there does not really exist a practical rule how to choose the bandwidth, this is a crucial argument against its practical use.

We turn to highly correlated designs, i.e. using $\Sigma_3$. The results are given in Table 9. In general we

CMIE, classic marginal integration estimator, $\Sigma_1$, $d = 3$

| $\alpha$ | under the null model, $a = 0.0$ | | | | | under the alternative, $a = 2.0$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ |
| | without trimming: | | | | | | | | | |
| 5% | 0.561 | 0.635 | 0.002 | 0.001 | 0 | 0.242 | 0.286 | 0.246 | 0.128 | 0.051 |
| 1% | 0.185 | 0.245 | 0 | 0 | 0 | 0.066 | 0.082 | 0.058 | 0.014 | 0.001 |
| | with trimming at 1.96: | | | | | | | | | |
| 5% | 0.446 | 0.551 | 0.039 | 0.008 | 0.002 | 0.228 | 0.307 | 0.631 | 0.413 | 0.181 |
| 1% | 0.127 | 0.217 | 0.004 | 0 | 0 | 0.061 | 0.103 | 0.235 | 0.09 | 0.013 |
| | with trimming at 1.645: | | | | | | | | | |
| 5% | 0.323 | 0.479 | 0.073 | 0.014 | 0.003 | 0.196 | 0.308 | 0.777 | 0.521 | 0.242 |
| 1% | 0.086 | 0.167 | 0.007 | 0.001 | 0 | 0.052 | 0.111 | 0.38 | 0.137 | 0.028 |

Table 7: *Percentage of rejection of* 1000 *repetitions applying the various tests on model (4.2) with* $n = 100$ *and uncorrelated regressors.* $\alpha$ *gives the wanted significance level. Results are given for* 500 *bootstrap replications using the CMIE.*

IMIE, marginal integration using internalized pre-estimator, $\Sigma_1$, $d = 3$

| $\alpha$ | under the null model, $a = 0.0$ | | | | | under the alternative, $a = 2.0$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ |
| | without trimming: | | | | | | | | | |
| 5% | 0.078 | 0.125 | 0.023 | 0.01 | 0.006 | 0.125 | 0.176 | 0.844 | 0.826 | 0.768 |
| 1% | 0.011 | 0.031 | 0 | 0 | 0 | 0.034 | 0.054 | 0.365 | 0.259 | 0.135 |
| | with trimming at 1.96: | | | | | | | | | |
| 5% | 0.055 | 0.099 | 0.006 | 0.002 | 0 | 0.101 | 0.178 | 0.775 | 0.712 | 0.556 |
| 1% | 0.005 | 0.022 | 0 | 0 | 0 | 0.025 | 0.054 | 0.374 | 0.23 | 0.081 |
| | with trimming at 1.645: | | | | | | | | | |
| 5% | 0.044 | 0.091 | 0.012 | 0.002 | 0.002 | 0.096 | 0.222 | 0.69 | 0.599 | 0.407 |
| 1% | 0.009 | 0.018 | 0.001 | 0 | 0 | 0.016 | 0.071 | 0.365 | 0.188 | 0.051 |

Table 8: *Percentage of rejection of* 1000 *repetitions applying the various tests on model (4.2) with* $n = 100$ *and uncorrelated regressors.* $\alpha$ *gives the wanted significance level. Results are given for* 500 *bootstrap replications using the IMIE.*

can see that all methods do hardly work anymore. The tests based on the CMIE performs particulary bad. It is interesting and not easy to explain why the effects are contrary, i.e. the tests using the CMIE reject too often whereas the IMIE produces too conservative tests. In any case our results show that when data are sparse and correlation is high, these tests even under the use of bootstrap are little helpful.

Highly correlated design ($\Sigma_3$), $d = 3$

| | using the CMIE | | | | | using the IMIE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ |
| | without trimming: | | | | | | | | | |
| 5% | 0.88 | 0.917 | 0.18 | 0.097 | 0.035 | 0.03 | 0.054 | 0.005 | 0.001 | 0.001 |
| 1% | 0.562 | 0.623 | 0.025 | 0.009 | 0.003 | 0.004 | 0.005 | 0 | 0 | 0 |
| | with trimming at 1.96: | | | | | | | | | |
| 5% | 0.829 | 0.887 | 0.589 | 0.408 | 0.256 | 0.02 | 0.037 | 0.004 | 0.002 | 0.001 |
| 1% | 0.499 | 0.572 | 0.232 | 0.102 | 0.035 | 0.004 | 0.002 | 0 | 0 | 0 |
| | with trimming at 1.645: | | | | | | | | | |
| 5% | 0.754 | 0.856 | 0.655 | 0.496 | 0.297 | 0.016 | 0.027 | 0 | 0 | 0 |
| 1% | 0.403 | 0.507 | 0.29 | 0.151 | 0.056 | 0.001 | 0.005 | 0 | 0 | 0 |

Table 9: *Percentage of rejection of* 1000 *repetitions applying the various tests with both, IMIE and CMIE, on model (4.2) with* $n = 100$ *and covariance* $\Sigma_3$ *from (2.17).* $\alpha$ *gives the wanted significance level. Results are given for* 500 *bootstrap replications.*

Finally let us study the impact of increasing correlation when the explanatory variables are still in three dimensions but transformed to an almost unique cube partially ameliorating the data sparseness. For this purpose consider the model as above but transform the regressors to

$$X \to (atan(X) \cdot 2.4/\pi + 1.0) \cdot 0.5 \tag{4.3}$$

before they enter in the model (4.2), still with $\Sigma_\gamma$, $\gamma = 1, 2, 3$ defined as in (2.17). Notice that in this case all points are contained in a cube with values between $-0.1$ and $1.1$. Further we increased the sample size to $n = 150$. We again did first a CV study to find the optimal bandwidths. For the test based on $T_4$ we tried several values for $g$ and give results for $g_1 = 0.25std(X)$ and $g_2 = 0.5std(X)$. First, for $\Sigma_1$, i.e. uncorrelated design, we compare once again IMIE and CMIE in Table 10. The trimming boundaries correspond approximately to cutting the outer 5%, respectively 10% of the data.

Looking at Table 10 we see advantages also in performance (not only computational time) for the IMIE, even though not strong ones. An analysis of the power shows that both test behave similar but poor, especially the tests based on $T_4$, but also the tests obtained from $T_1$. In the last table, Table 11, we only give results for the better behaving IMIE to show how much the IMIE is affected by the correlation of the regressors if the data are not too sparse (due to transformation and having $n = 150$ what is still not much for the underlying problem). Again, a study of the power reveals that the tests based on $T_4$ and $T_1$ break down completely. The best performing test is always obtained from the statistic $T_3$ but this could maybe depend on our particular model even if it should not, see Theorem 1 with discussion.

Uncorrelated transformed designs, $d = 3$

| $\alpha$ | using the CMIE | | | | | using the IMIE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ |
| | without trimming: | | | | | | | | | |
| 5% | 0.082 | 0.09 | 0.044 | 0.03 | 0.026 | 0.066 | 0.066 | 0.046 | 0.058 | 0.052 |
| 1% | 0.006 | 0.016 | 0.006 | 0.006 | 0.002 | 0.016 | 0.014 | 0.012 | 0.006 | 0.004 |
| | trimming about 5%: | | | | | | | | | |
| 5% | 0.078 | 0.098 | 0.044 | 0.042 | 0.03 | 0.068 | 0.064 | 0.052 | 0.06 | 0.048 |
| 1% | 0.008 | 0.016 | 0.012 | 0.006 | 0.006 | 0.014 | 0.014 | 0.01 | 0.008 | 0.008 |
| | trimming about 10%: | | | | | | | | | |
| 5% | 0.07 | 0.072 | 0.046 | 0.046 | 0.032 | 0.062 | 0.062 | 0.05 | 0.078 | 0.074 |
| 1% | 0.008 | 0.016 | 0.006 | 0.008 | 0.006 | 0.018 | 0.02 | 0.008 | 0.006 | 0.004 |

Table 10: *Percentage of rejection of* 500 *repetitions applying the various tests with both, IMIE and CMIE, on the null model [$a = 0.0$ in (4.2)] with $n = 150$ and transformed (4.3), uncorrelated regressors. $\alpha$ gives the wanted significance level. Results are given for* 500 *bootstrap replications.*

Correlated transformed designs, IMIE, $d = 3$

| Covariance | $\Sigma_2$ | | | | | $\Sigma_3$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ | $T_4(g_1)$ | $T_4(g_2)$ | $T_1$ | $T_2$ | $T_3$ |
| | without trimming: | | | | | | | | | |
| 5% | 0.08 | 0.092 | 0.058 | 0.07 | 0.066 | 0.074 | 0.13 | 0.044 | 0.064 | 0.082 |
| 1% | 0.016 | 0.032 | 0.014 | 0.01 | 0.01 | 0.012 | 0.016 | 0.004 | 0.014 | 0.012 |
| | trimming about 5%: | | | | | | | | | |
| 5% | 0.07 | 0.068 | 0.056 | 0.072 | 0.064 | 0.07 | 0.084 | 0.036 | 0.058 | 0.078 |
| 1% | 0.018 | 0.018 | 0.014 | 0.01 | 0.012 | 0.01 | 0.008 | 0.006 | 0.008 | 0.008 |
| | trimming about 10%: | | | | | | | | | |
| 5% | 0.066 | 0.064 | 0.058 | 0.068 | 0.054 | 0.06 | 0.068 | 0.038 | 0.052 | 0.066 |
| 1% | 0.012 | 0.014 | 0.01 | 0.006 | 0.008 | 0.01 | 0.004 | 0.006 | 0.008 | 0.01 |

Table 11: *Percentage of rejection of* 500 *repetitions applying the various tests, using IMIE, on the null model [$a = 0.0$ in (4.2)] with $n = 150$. Regressors are correlated according $\Sigma_\gamma$, $\gamma = 1, 2$, see (2.17), and transformed according (4.3). $\alpha$ gives the wanted significance level. Results are given for* 500 *bootstrap replications.*

## 5 Appendix: Proof of Theorem 1

**Proof of Theorem 1 (i) [the case of the null hypothesis]:**

For the sake of brevity we restrict ourselves to a consideration of the statistic $T_{4n}$ and a two dimensional explanatory variable, i.e. $d = 2$. Throughout this proof the marginal densities of $X_1$ and $X_2$ are denoted by $f_1$ and $f_2$, respectively (i.e. $f_1 = f_{-2}$, $f_2 = f_{-1}$). Note that under the null hypothesis of additivity we have $m = m_0$, where $m_0$ is defined in (3.8). Introducing the notation

$$\delta(x) := \widehat{m}_0^I(x) - m(x) \tag{A.1}$$

we obtain $\widehat{e}_i = \sigma(X_i)\varepsilon_i - \delta(X_i)$. This yields the following decomposition of the test statistic $T_{4n}$:

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

where

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

The term $V_{1n}$ can be treated by similar arguments as given in Zheng (1996) and we obtain

$$n g^{\frac{d}{2}} V_{1n} \xrightarrow{D} N\left(0, \lambda_4^2\right) \tag{A.4}$$

as $n \to \infty$, where $\lambda_4^2$ is defined in (3.10). For the remaining terms in (A.2) we will prove

$$V_{2n} = o_P(n^{-1} g^{-\frac{d}{2}}), \qquad V_{3n} = o_P(n^{-1} g^{-\frac{d}{2}}),$$

which yields the assertion of Theorem 1 under the null hypothesis of additivity. For the estimation of the term $V_{2n}$ we introduce

$$\begin{aligned} \delta_\alpha(x_\alpha) &:= \widehat{m}_\alpha^I(x_\alpha) - \int m(x_\alpha, y_{-\alpha})\, f_{-\alpha}(y_{-\alpha})\, dy_{-\alpha}\,,\ \alpha = 1, \ldots, d\,, \\ \delta_0 &:= \widehat{c} - c, \end{aligned} \tag{A.5}$$

and obtain

$$\delta(x) = \sum_{\alpha=1}^{d} \delta_\alpha(x_\alpha) - (d-1)\,\delta_0,$$

which yields the decomposition

$$V_{2n} = \sum_{\alpha=1}^{d} V_{2n}^{(\alpha)} - (d-1)\, V_{2n}^{(0)}, \tag{A.6}$$

where

$$V_{2n}^{(\alpha)} := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_g(X_i - X_j)\,\sigma(X_i)\,\varepsilon_i\, \delta_\alpha(X_j)\ ,\ \alpha = 1, \ldots, d, \tag{A.7}$$

$$V_{2n}^{(0)} := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_g(X_i - X_j)\,\sigma(X_i)\,\varepsilon_i\, \delta_0. \tag{A.8}$$

We prove in a first step the estimate

$$V_{2n}^{(\alpha)} = o_P(n^{-1}g^{-\frac{d}{2}}), \ \alpha = 1, \dots, d \tag{A.9}$$

but obviously it is sufficient to consider the case $\alpha = 1$. Recall the definition (2.10) of $\widehat{m}_1^I$ for the case $d = 2$ with equal kernels and bandwidths, i.e.

$$\widehat{m}_1^I(x_1) = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{K_h(X_{l1}-x_1)K_h(X_{l2}-X_{k2})}{\widehat{f}(X_l)}Y_l$$

with

$$\widehat{f}(x) = \frac{1}{n}\sum_{s=1}^{n}K_h(X_{s1}-x_1)K_h(X_{s2}-x_2).$$

Then it follows that

$$\begin{aligned}
\delta_1(x_1) &= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}K_h(X_{l1}-x_1)K_h(X_{l2}-X_{k2})\widehat{f}^{-1}(X_l)\ \sigma(X_l)\varepsilon_l \\
&+\frac{1}{n^2}\sum_{k=1}^{n}\{\sum_{l=1}^{n}K_h(X_{l1}-x_1)K_h(X_{l2}-X_{k2})\widehat{f}^{-1}(X_l)\ m(X_l) - m(x_1, X_{k2})\} \\
&+\frac{1}{n}\sum_{k=1}^{n}m(x_1, X_{k2}) - \int m(x_1, x_2)f_2(x_2)\,dx_2.
\end{aligned}$$

Consequently, we obtain the following decomposition for the term $V_{2n}^{(1)}$:

$$V_{2n}^{(1)} = V_{2n}^{(1.1)} + V_{2n}^{(1.2)} + V_{2n}^{(1.3)}, \tag{A.10}$$

where

$$\begin{aligned}
V_{2n}^{(1.1)} &:= \frac{1}{n^3(n-1)}\sum_{i,k,l=1}^{n}\sum_{j\neq i}L_g(X_i-X_j)\ \sigma(X_i)\varepsilon_i \\
&\qquad \times K_h(X_{l1}-X_{j1})K_h(X_{l2}-X_{k2})\widehat{f}^{-1}(X_l)\ \sigma(X_l)\varepsilon_l, \\
V_{2n}^{(1.2)} &:= \frac{1}{n^3(n-1)}\sum_{i,k,l=1}^{n}\sum_{j\neq i}L_g(X_i-X_j)\ \sigma(X_i)\varepsilon_i \\
&\qquad \times \{\ K_h(X_{l1}-X_{j1})K_h(X_{l2}-X_{k2})\widehat{f}^{-1}(X_l)\,m(X_l) - m(X_{j1}, X_{k2})\ \}, \\
V_{2n}^{(1.3)} &:= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}L_g(X_i-X_j)\,\sigma(X_i)\,\varepsilon_i \\
&\qquad \times \{\ \frac{1}{n}\sum_{k=1}^{n}m(X_{j1}, X_{k2}) - \int m(X_{j1}, x_2)f_2(x_2)\,dx_2\ \}.
\end{aligned}$$

For the expectation of the first term on the right hand side of (A.10) this yields

$$E(V_{2n}^{(1.1)}) = \frac{1}{n^3(n-1)}\sum_{i,k=1}^{n}\sum_{j\neq i}E[L_g(X_i-X_j)K_h(X_{i1}-X_{j1})K_h(X_{i2}-X_{k2})\widehat{f}^{-1}(X_i)\ \sigma^2(X_i)].$$

The strong uniform convergence of the kernel density estimate $\widehat{f}$ gives (see Collomb and Härdle, 1986)

$$\sup_x \widehat{f}^{-1}(x) \leq C_1 \tag{A.11}$$

for some constant $C_1$. Observing that the kernel is bounded we therefore obtain

$$E(V_{2n}^{(1.1)}) = O(n^{-1}h^{-1} + n^{-2}h^{-2}) = O(n^{-1}h^{-1}). \tag{A.12}$$

In a second step we give an estimate of the variance of the statistic $V_{2n}^{(1.1)}$. To be precise we note that

$$\begin{aligned} Var(V_{2n}^{(1.1)}) \leq E[(V_{2n}^{(1.1)})^2] &= \frac{1}{n^6\,(n-1)^2} \sum_{i=1}^{n} \sum_{j\neq i} \sum_{k=1}^{n} \sum_{l=1}^{n} \sum_{i'=1}^{n} \sum_{j'\neq i'} \sum_{k'=1}^{n} \sum_{l'=1}^{n} \\ E\,[\,G_1\,(X_i, X_j, X_k, X_l)\,\widehat{f}^{-1}\,(X_l)\;&G_1(X_{i'}, X_{j'}, X_{k'}, X_{l'})\widehat{f}^{-1}(X_{l'})\sigma\,(X_i)\,\sigma(X_{i'})\sigma\,(X_l)\,\sigma(X_{l'})\varepsilon_i\varepsilon_{i'}\varepsilon_l\varepsilon_{l'}\,], \end{aligned} \tag{A.13}$$

where we introduce the notation

$$G_1\,(X_i, X_j, X_k, X_l) := L_g\,(X_i - X_j)\,K_h\,(X_{l1} - X_{j1})\,K_h\,(X_{l2} - X_{k2})\,.$$

Note that only summation over the pairs $(i' = i \neq l' = l)$, $(i' = l \neq l' = i)$, $(i = l \neq i' = l')$ or $(i' = i = l' = l)$ yields a non-negligible contribution to the expectation of $(V_{2n}^{(1.1)})^2$. For pairwise different indices $i, j, j', k, k', l$ it follows that

$$\begin{aligned} E\left(\left|G_1\,(X_i, X_j, X_k, X_l)\,G_1(X_i, X_{j'}, X_{k'}, X_l)\right| \mid X_i, X_l\right) &= E\,(|G_1\,(X_i, X_j, X_k, X_l)| \mid X_i, X_l)^2 \\ &\qquad\qquad\qquad\qquad\qquad\qquad (A.14) \\ = E\,(|L_g\,(X_i - X_j)\,K_h\,(X_{l1} - X_{j1})| \mid X_i, X_l)^2\,E\,(|K_h\,(X_{l2} - X_{k2})| \mid X_l)^2 &= O\left(h^{-2}\right) \end{aligned}$$

(uniformly with respect $i$ and $l$). Therefore the part of the sum with $i' = i,\ l' = l$ and pairwise different indices $i, j, j', k, k', l$ in (A.13) is of order $O\left(n^{-2}h^{-2}\right)$ (note (A.11) and (A.14)). The remaining part of the sum on the right hand side of (A.13) is treated similarly, which gives for the variance of $V_{2n}^{(1.1)}$

$$Var(V_{2n}^{(1.1)}) = O(n^{-2}h^{-2}). \tag{A.15}$$

Now a combination of (A.12) and (A.15) yields

$$V_{2n}^{(1.1)} = O_P(n^{-1}h^{-1}) = o_P(n^{-1}g^{-\frac{d}{2}}), \tag{A.16}$$

where the last equality is a consequence of the assumption(A5), that is $g^d = o\left(h^2\right)$.

For the second term in the decomposition (A.10) it obviously follows that $E(V_{2n}^{(1.2)}) = 0$. For the calculation of the variance we consider the sum

$$\begin{aligned} E[(V_{2n}^{(1.2)})^2] &= \frac{1}{n^6\,(n-1)^2} \sum_{i} \sum_{j\neq i} \sum_{j'\neq i} \sum_{k} \sum_{k'} \sum_{l} \sum_{l'} \\ &E\Big[L_g\,(X_i - X_j)\,L_g\left(X_i - X_{j'}\right)\sigma^2\,(X_i) \\ &\times \{\,K_h\,(X_{l1} - X_{j1})\,K_h\,(X_{l2} - X_{k2})\,\widehat{f}^{-1}\,(X_l)\,m\,(X_{l1}, X_{l2}) - m\,(X_{j1}, X_{k2})\,\} \\ &\times \{\,K_h(X_{l'1} - X_{j'1})K_h(X_{l'2} - X_{k'2})\widehat{f}^{-1}(X_{l'})m(X_{l'1}, X_{l'2}) - m(X_{j'1}, X_{k'2})\,\}\Big], \end{aligned} \tag{A.17}$$

where the expectation is determined by conditioning, that is

$$\begin{aligned} G_2\,(X_i, X_j, X_k, X_l) \;:=&\; E\,(\,L_g\,(X_i - X_j)\,\{\,K_h\,(X_{l1} - X_{j1})\,K_h\,(X_{l2} - X_{k2})\,\widehat{f}^{-1}\,(X_l)\,m\,(X_l) \\ &\; -m\,(X_{j1}, X_{k2})\,\}\ \mid X_i, X_j, X_k, X_l\,) \\ =&\; L_g\,(X_i - X_j)\,\{\,K_h\,(X_{l1} - X_{j1})\,K_h\,(X_{l2} - X_{k2})\,f^{-1}\,(X_l)\,m\,(X_l)\,(1 + O(h^r)) \\ &\; -m\,(X_{j1}, X_{k2})\,\}\,. \end{aligned}$$

Observing this result we obtain from (A.17)

$$E[(V_{2n}^{(1.2)})^2] \;=\; \frac{1}{n^6\,(n-1)^2}\sum_{i,k,k',l,l'}\sum_{j\neq i}\sum_{j'\neq i}E\Big[\sigma^2(X_i) \times E\big[G_2(X_i,X_j,X_k,X_l)G_2(X_i,X_{j'},X_{k'},X_{l'}) \mid X_i\big]\Big]. \tag{A.18}$$

If $i,j,k,l$ are pairwise different we have

$$\begin{aligned}
&E\left(G_2\left(X_i,X_j,X_k,X_l\right)\mid X_i,X_j,X_k\right)=L_g\left(X_i-X_j\right)\\
&\times\Big\{E\left(K_h\left(X_{l1}-X_{j1}\right)K_h\left(X_{l2}-X_{k2}\right)f^{-1}\left(X_l\right)m\left(X_l\right)\mid X_i,X_j,X_k\right)\left(1+O(h^r)\right)-m\left(X_{j1},X_{k2}\right)\Big\}\\
&=L_g\left(X_i-X_j\right)\{\int K_h\left(x_1-X_{j1}\right)K_h\left(x_2-X_{k2}\right)m\left(x\right)dx\left(1+O(h^r)\right)-m\left(X_{j1},X_{k2}\right)\}\\
&=L_g\left(X_i-X_j\right)\,O(h^r)
\end{aligned}$$

and integration with respect to $X_j$ and $X_k$ yields

$$E\left(G_2\left(X_i,X_j,X_k,X_l\right)\mid X_i\right)=O\left(h^r\right).$$

This implies for pairwise different indices $i,j,j',k,k',l,l'$

$$E\left[G_2\left(X_i,X_j,X_k,X_l\right)G_2(X_i,X_{j'},X_{k'},X_{l'})\mid X_i\right]=E\left[G_2\left(X_i,X_j,X_k,X_l\right)\mid X_i\right]^2=O\left(h^{2r}\right).$$

Consequently, the part of the sum with pairwise different indices in (A.18) is of order $O(n^{-1}h^{2r}) = o(n^{-2}g^{-d})$, where we used the assumption (A5) in the last step. The remaining part of the sum on the right hand side of (A.18) is treated similarly and also of order $o\left(n^{-2}g^{-d}\right)$, which yields

$$V_{2n}^{(1.2)}=o_P(n^{-1}g^{-\frac{d}{2}}). \tag{A.19}$$

For the term $V_{2n}^{(1.3)}$ we note that $E(V_{2n}^{(1.3)})=0$ and

$$E[(V_{2n}^{(1.3)})^2]=\frac{1}{n^2\,(n-1)^2}\sum_i\sum_{j\neq i}\sum_{j'\neq i}\sum_k\sum_{k'}E[\sigma^2\left(X_i\right)G_3\left(X_i,X_j,X_k\right)G_3(X_i,X_{j'},X_{k'})], \tag{A.20}$$

where

$$G_3\left(X_i,X_j,X_k\right):=L_g\left(X_i-X_j\right)\Big\{m\left(X_{j1},X_{k2}\right)-\int m\left(X_{j1},x_2\right)f_2\left(x_2\right)dx_2\Big\}.$$

If $i,j,j',k,k'$ are pairwise different we have

$$\begin{aligned}
&E\left(G_3\left(X_i,X_j,X_k\right)G_3(X_i,X_{j'},X_{k'})\mid X_i\right)=E\left(G_3\left(X_i,X_j,X_k\right)\mid X_i\right)^2\\
&=E\Big(L_g\left(X_i-X_j\right)E\Big(m\left(X_{j1},X_{k2}\right)-\int m\left(X_{j1},x_2\right)f_2\left(x_2\right)dx_2\mid X_j\Big)\mid X_i\Big)^2=0
\end{aligned}$$

and the corresponding terms in the sum (A.20) vanish. The remaining part of (A.20) is of order $O\left(n^{-2}\right)=o\left(n^{-2}g^{-d}\right)$, which follows by a straightforward argument. Consequently, the variance of $V_{2n}^{(1.3)}$ is of order $o(n^{-1}g^{d})$, which yields

$$V_{2n}^{(1.3)}=o_P(n^{-1}g^{-\frac{d}{2}}). \tag{A.21}$$

Observing (A.10), (A.16), (A.19) and (A.21) we obtain $V_{2n}^{(1)}=o_P\left(n^{-1}g^{-d/2}\right)$, that is (A.9) for $\alpha=1$. Finally, we introduce the decomposition

$$V_{2n}^{(0)}=V_{2n}^{(0.1)}+V_{2n}^{(0.2)}$$

where

$$V_{2n}^{(0.1)} \; : \; = \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} L_g(X_i - X_j)\, \sigma(X_i)\sigma(X_k)\varepsilon_i\varepsilon_k,$$

$$V_{2n}^{(0.2)} \; : \; = \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} L_g(X_i - X_j)\, \sigma(X_i)\varepsilon_i(m(X_k) - c).$$

Note that $E(V_{2n}^{(0.1)}) = O(n^{-1})$, $E(V_{2n}^{(0.2)}) = 0$, and that the second moments of these statistics can be estimated as follows

$$\begin{aligned} E[(V_{2n}^{(0.1)})^2] &= \frac{1}{n^4(n-1)^2} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} \sum_{i'=1}^{n} \sum_{j' \neq i'} \sum_{k'=1}^{n} \\ &\quad E[L_g(X_i - X_j)\, L_g(X_{i'} - X_{j'})\sigma(X_i)\sigma(X_{i'})\sigma(X_k)\sigma(X_{k'})\varepsilon_i\varepsilon_{i'}\varepsilon_k\varepsilon_{k'}] \\ &= O(n^{-2}), \end{aligned}$$

$$\begin{aligned} E[(V_{2n}^{(0.2)})^2] &= \frac{1}{n^4(n-1)^2} \sum_{k,i,i'=1}^{n} \sum_{j \neq i} \sum_{j' \neq i'} E[(m(X_k) - c)^2 L_g(X_i - X_j)\, L_g(X_{i'} - X_{j'})\sigma(X_i)\sigma(X_{i'})\varepsilon_i\varepsilon_{i'}] \\ &= O\left(n^{-2}\right), \end{aligned}$$

which yields

$$V_{2n}^{(0)} = O_P\left(n^{-1}\right) = o_P(n^{-1}g^{-d}).$$

Observing the decomposition (A.6) it follows that

$$V_{2n} = o_P(n^{-1}g^{-\frac{d}{2}}). \tag{A.22}$$

Using similar arguments we obtain for the term $V_{3n}$ in (A.3)

$$V_{3n} = o_P(n^{-1}g^{-\frac{d}{2}}), \tag{A.23}$$

which yields the assertion of Theorem 1 under the null hypothesis observing the equations (A.2), (A.4), (A.22) and (A.23).

**Proof of Theorem 1 (ii) [the case of fixed alternatives]:**

In the case of non-additivity we introduce the decomposition

$$\widehat{e}_i = \sigma(X_i)\,\varepsilon_i - \delta(X_i) + \Delta(X_i)$$

where $\Delta = m - m_0 \neq 0$ and $m_0$ and $\delta$ are defined in (3.8) and (A.1), respectively. This yields to a more detailed decomposition of the statistic $T_{4n}$, i.e.

$$T_{4n} = V_{1n} - 2V_{2n} + V_{3n} + 2V_{4n} - 2V_{5n} + V_{6n} \tag{A.24}$$

where $V_{1n}$, $V_{2n}$ and $V_{3n}$ are defined in (A.3) and

$$V_{4n} \; := \; \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_g(X_i - X_j)\, \Delta(X_j)\, \sigma(X_i)\, \varepsilon_i,$$

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

$$V_{6n} \; := \; \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_g(X_i - X_j)\, \Delta(X_i)\, \Delta(X_j).$$

Observing the arguments in the first part of the proof we have

$$V_{in} \quad = \quad o_P(n^{-\frac{1}{2}}), i = 1, 2, 3 \tag{A.25}$$

and it remains to consider the statistics $V_{4n},\ V_{5n}$ and $V_{6n}$. We prove the following three assertions:

**(a)** $E(V_{4n}) = 0$ and

$$V_{4n} = \frac{1}{n}\sum_{i=1}^{n} \sigma\,(X_i)\,\varepsilon_i\ (\Delta f)\,(X_i) + o_P(n^{-\frac{1}{2}}). \tag{A.26}$$

**(b)** $E\,(V_{5n}) = o(n^{-\frac{1}{2}})$ and

$$V_{5n} - E\,(V_{5n}) = \frac{1}{n}\sum_{i=1}^{n}\{\sigma\,(X_i)\,\varepsilon_i\, p(\Delta f)(X_i) + \frac{1}{2}q_4(\Delta f)(X_i) - E[\frac{1}{2}q_4(\Delta f)(X_i)]\} + o_P(n^{-\frac{1}{2}}),$$

where the mappings $p$ and $q_4$ are defined in (3.13).

**(c)** Finally,

$$V_{6n} - E\,(V_{6n}) = \frac{2}{n}\sum_{i=1}^{n}\left(\Delta^2 f\right)(X_i) - \int (\Delta f)^2\,(x)\,dx + o_P(n^{-\frac{1}{2}}) \tag{A.27}$$

and

$$E\,(V_{6n}) = \int (\Delta f)^2\,(x)\,dx + o(n^{-\frac{1}{2}}). \tag{A.28}$$

If (a), (b), (c) have been established, then a combination of these results with (A.24) and (A.25) yields

$$\begin{aligned} T_{4n} - E(T_{4n}) \quad &= \quad \frac{1}{n}\sum_{i=1}^{n}\Big\{\sigma\,(X_i)\,\varepsilon_i 2((\Delta f)(X_i) - p(\Delta f)\,(X_i)) + \Big[2(\Delta^2 f)(X_i) - q_4(\Delta f)\,(X_i) \\ &\qquad - E\big(2(\Delta^2 f)(X_i) - q_4(\Delta f)\,(X_i)\big)\Big]\Big\} + o_P(n^{-\frac{1}{2}}), \end{aligned}$$

and

$$E(T_{4n}) = \int (\Delta f)^2\,(x)\,dx\ + o(n^{\frac{-1}{2}}). \tag{A.29}$$

The assertion of part (ii) of Theorem 1 (for the statistic $T_{4n}$) now follows from (A.29) and an application of the central limit theorem.

*Proof of (a):* Obviously we have $E(V_{4n}) = 0$ and with the notation

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

we obtain the following representation for the statistic $V_{4n}$:

$$\begin{aligned} V_{4n} \quad &= \quad \frac{1}{n}\sum_{i=1}^{n}\sigma\,(X_i)\,\varepsilon_i\, Z_{in}^{(1)} \qquad (A.30) \\ &= \quad \frac{1}{n}\sum_{i=1}^{n}\sigma\,(X_i)\,\varepsilon_i\Big\{(\Delta f)(X_i) + \big[E(Z_{in}^{(1)}\mid X_i) - (\Delta f)(X_i)\big] + \big[Z_{in}^{(1)} - E(Z_{in}^{(1)}\mid X_i)\big]\Big\}. \end{aligned}$$

A straightforward Taylor expansion gives (a.s)

$$\sup_{1\le i\le n}\left[E\left(Z_{in}^{(1)}\mid X_i\right)-(\Delta f)(X_i)\right]=O\left(g^2\right)=o(1), \tag{A.31}$$

which yields

$$\begin{aligned}&E\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\sigma(X_i)\varepsilon_i\left[E(Z_{in}^{(1)}\mid X_i)-(\Delta f)(X_i)\right]\Big)^2\Big]\\&\le n^{-1}E\Big[\sup_{1\le i\le n}(\sigma^2(X_i)\left[E(Z_{in}^{(1)}\mid X_i)-(\Delta f)(X_i)\right]^2)\Big]=o(n^{-1}).\end{aligned} \tag{A.32}$$

The conditional variance of $Z_{in}^{(1)}$ given $X_i$ is estimated as follows

$$\begin{aligned}&\sup_{1\le i\le n}E\left[\left(Z_{in}^{(1)}-E\left(Z_{in}^{(1)}\mid X_i\right)\right)^2\mid X_i\right]=\sup_{1\le i\le n}\frac{1}{(n-1)^2}\sum_{j=1,\,j\neq i}^{n}[\,E\left(L_g^2(X_i-X_j)\Delta^2(X_j)\mid X_j\right)\\&-E\left(L_g(X_i-X_j)\Delta(X_j)\mid X_i\right)^2\,]=O(n^{-1}g^{-d})=o(1),\end{aligned}$$

and we obtain by the same arguments

$$\begin{aligned}&E\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\sigma(X_i)\varepsilon_i\left[Z_{in}^{(1)}-E\left(Z_{in}^{(1)}\mid X_i\right)\right]\Big)^2\Big]\\&=\frac{1}{n^2}\sum_{i=1}^{n}E[\sigma^2(X_i)\;E(\left[Z_{in}^{(1)}-E\left(Z_{in}^{(1)}\mid X_i\right)\right]^2\mid X_i)]=o\left(n^{-1}\right).\end{aligned} \tag{A.33}$$

A combination of (A.30), (A.32) and (A.33) gives the equation (A.26 ) and proves assertion (a).

*Proof of (b):* We introduce a similar decomposition as used for $V_{2n}$ in the first part of the proof:

$$V_{5n}=\sum_{\alpha=1}^{d}V_{5n}^{(\alpha)}-(d-1)V_{5n}^{(0)}, \tag{A.34}$$

where

$$V_{5n}^{(\alpha)}=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}L_g(X_i-X_j)\Delta(X_j)\delta_\alpha(X_{i\alpha}),\alpha=1,\ldots,d, \tag{A.35}$$

$$V_{5n}^{(0)}=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}L_g(X_i-X_j)\Delta(X_j)\delta_0, \tag{A.36}$$

and the functions $\delta_\alpha$ $(\alpha=0,\ldots,d)$ are defined in (A.5). In the following we prove

$$\begin{aligned}V_{5n}^{(\alpha)}-E(V_{5n}^{(\alpha)})&=\frac{1}{n}\sum_{i=1}^{n}\Big\{\left[\sigma(X_i)\varepsilon_i+m(X_i)\right]f^{-1}(X_i)f_{-\alpha}(X_{i,-\alpha})\int\left(\Delta f^2\right)(X_{i\alpha},y_{-\alpha})\,dy_{-\alpha}\\&+\int\left(\Delta f^2\right)(y)\;m(y_\alpha,X_{i,-\alpha})\,dy\\&-E\Big[m(X_i)f^{-1}(X_i)f_{-\alpha}(X_{i,-\alpha})\int\left(\Delta f^2\right)(X_{i\alpha},y_{-\alpha})\,dy_{-\alpha}\\&+\int\left(\Delta f^2\right)(y)\;m(y_\alpha,X_{i,-\alpha})\,dy\Big]\Big\}+o_P(n^{-\frac{1}{2}}),\end{aligned} \tag{A.37}$$

$$E(V_{5n}^{(\alpha)}) = O(g^r + h^r) = o(n^{-\frac{1}{2}}). \tag{A.38}$$

Note that it is again sufficient to consider only the case $\alpha = 1$. Further,

$$V_{5n}^{(0)} = \frac{1}{n}\sum_{i=1}^{n}\Big\{\sigma(X_i)\,\varepsilon_i + m(X_i) - E[m(X_i)]\Big\}\int(\Delta f^2)(y)dy + o_P(n^{-\frac{1}{2}}) \tag{A.39}$$

and

$$E(V_{5n}^{(0)}) = o(n^{-\frac{1}{2}}). \tag{A.40}$$

The assertion (b) then follows by a combination of (A.34) - (A.40) and the definition of the functions $p$ and $q_4$ in (3.13).

For the statistic $V_{5n}^{(1)}$ we obtain the decomposition

$$V_{5n}^{(1)} = V_{5n}^{(1.1)} + V_{5n}^{(1.2)} + V_{5n}^{(1.3)}, \tag{A.41}$$

where

$$\begin{aligned}
V_{5n}^{(1.1)} &= \frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n} L_g(X_i - X_j)\,\Delta(X_j) \\
&\quad \times K_h(X_{l1} - X_{i1})\,K_h(X_{l2} - X_{k2})\,\widehat{f}^{-1}(X_l)\sigma(X_l)\,\varepsilon_l, \\
V_{5n}^{(1.2)} &= \frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n} L_g(X_i - X_j)\,\Delta(X_j) \\
&\quad \times\{K_h(X_{l1} - X_{i1})\,K_h(X_{l2} - X_{k2})\,\widehat{f}^{-1}(X_l)m(X_{l1}, X_{l2}) - m(X_{i1}, X_{k2})\}, \\
V_{5n}^{(1.3)} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} L_g(X_i - X_j)\,\Delta(X_j) \\
&\quad \times\{\ \frac{1}{n}\sum_{k=1}^{n} m(X_{i1}, X_{k2}) - \int m(X_{i1}, y_2)\,f_2(y_2)\,dy_2\ \}.
\end{aligned}$$

Remember that we consider the case $d = 2$. At first we discuss the statistic $V_{5n}^{(1.1)}$ following the arguments given in the proof of part (a). Obviously, we have $E(V_{5n}^{(1.1)}) = 0$. With the notation

$$Z_{l\,n}^{(2)} = \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n} L_g(X_i - X_j)\,\Delta(X_j)\,K_h(X_{l1} - X_{i1})\,K_h(X_{l2} - X_{k2})\,\widehat{f}^{-1}(X_l)$$

we obtain

$$V_{5n}^{(1.1)} = \frac{1}{n}\sum_{l=1}^{n}\sigma(X_l)\,\varepsilon_l\,Z_{l\,n}^{(2)}.$$

For the conditional expectation of $Z_{l\,n}^{(2)}$ given $X_l$ we have

$$\begin{aligned}
E(Z_{l\,n}^{(2)} \mid X_l) &= E(L_g(X_i - X_j)\,\Delta(X_j)\,K_h(X_{l1} - X_{i1})\,K_h(X_{l2} - X_{k2})\,\widehat{f}^{-1}(X_l) \mid X_l) \\
&= f^{-1}(X_l)\,f_2(X_{l2})\int\left(\Delta f^2\right)(X_{l1}, y_2)\,dy_2\,(1 + o(1))
\end{aligned}$$

(uniformly with respect to $l$). A similar argument shows

$$E[(Z_{l\,n}^{(2)} - E(Z_{l\,n}^{(2)} \mid X_l))^2 \mid X_l] = o(1)$$

(uniformly with respect to $l$), which yields (using similar arguments as in the proof of part (a))

$$V_{5n}^{(1.1)} = \frac{1}{n} \sum_{l=1}^{n} \sigma(X_l) \varepsilon_l f^{-1}(X_l) f_2(X_{l2}) \int (\Delta f^2)(X_{l1}, y_2) dy_2 + o_P(n^{-\frac{1}{2}}). \quad \text{(A.42)}$$

For the estimation of the term $V_{5n}^{(1.2)}$ we introduce the notation

$$\begin{aligned} &\widehat{G}_4(X_i, X_j, X_k, X_l) := L_g(X_i - X_j) \Delta(X_j) \\ &\times\{K_h(X_{l1} - X_{i1}) K_h(X_{l2} - X_{k2}) \widehat{f}^{-1}(X_l) m(X_{l1}, X_{l2}) - m(X_{i1}, X_{k2})\}. \end{aligned}$$

and obtain the representation

$$\begin{aligned} V_{5n}^{(1.2)} &= \frac{1}{n^3(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} \sum_{l=1}^{n} \widehat{G}_4(X_i, X_j, X_k, X_l) \\ &= \frac{1}{n^3(n-1)} \sum_{l=1}^{n} \sum_{i \neq l} \sum_{j \neq i,l} \sum_{k \neq i,j,l} \widehat{G}_4(X_i, X_j, X_k, X_l) + o_{L^1(P)}(n^{-\frac{1}{2}}). \end{aligned}$$

Moreover, a straightforward calculation of the variance shows

$$\frac{1}{n^3(n-1)} \sum_{l=1}^{n} \sum_{i \neq l} \sum_{j \neq i,l} \sum_{k \neq i,j,l} \{ \widehat{G}_4(X_i, X_j, X_k, X_l) - E(\widehat{G}_4(X_i, X_j, X_k, X_l) \mid X_l) \} = o_{L^2(P)}(\frac{1}{n}), \quad \text{(A.43)}$$

which yields

$$V_{5n}^{(1.2)} = \frac{1}{n} \sum_{l=1}^{n} E(\widehat{G}_4(X_i, X_j, X_k, X_l) \mid X_l) + o_{L^1(P)}(n^{-\frac{1}{2}}) + o_{L^2(P)}(n^{-1}). \quad \text{(A.44)}$$

The calculation of the conditional expectation gives

$$\begin{aligned} &E(\widehat{G}_4(X_i, X_j, X_k, X_l) \mid X_l) \\ &= E(L_g(X_i - X_j) \Delta(X_j) \widehat{f}^{-1}(X_l) K_h(X_{l1} - X_{i1}) K_h(X_{l2} - X_{k2}) \mid X_l) m(X_l) \\ &- E(L_g(X_i - X_j) \Delta(X_j) m(X_{i1}, X_{k2})) \\ &= \{m(X_l) f^{-1}(X_l) f_2(X_{2l}) \int (\Delta f^2)(X_{l1}, y_2) dy_2 \\ &- \int (\Delta f^2)(x) m(x_1, y_2) f_2(y_2) dx dy_2\} (1 + O(h^r + g^r)) \end{aligned}$$

and (A.44) implies

$$E(V_{5n}^{(1.2)}) = O(h^r + g^r) + o(n^{-\frac{1}{2}}) = o(n^{-\frac{1}{2}}) \quad \text{(A.45)}$$

as well as

$$\begin{aligned} V_{5n}^{(1.2)} - E(V_{5n}^{(1.2)}) &= \frac{1}{n} \sum_{l=1}^{n} \{m(X_l) f^{-1}(X_l) f_2(X_{2l}) \int (\Delta f^2)(X_{l1}, y_2) dy_2 \quad \text{(A.46)} \\ &- E[m(X_l) f^{-1}(X_l) f_2(X_{2l}) \int (\Delta f^2)(X_{l1}, y_2) dy_2] \} + o_P(n^{-\frac{1}{2}}). \end{aligned}$$

We now consider the random variable $V_{5n}^{(1.3)}$ and introduce the notation

$$G_5\left(X_i, X_j, X_k\right) := L_g\left(X_i - X_j\right) \Delta\left(X_j\right) \left\{\, m\left(X_{i1}, X_{k2}\right) - \int m\left(X_{i1}, y_2\right) f_2\left(y_2\right) dy_2 \,\right\},$$

which yields the representation

$$V_{5n}^{(1.3)} = \frac{1}{n^2\,(n-1)} \sum_{k=1}^{n} E\left(G_5\left(X_i, X_j, X_k\right) \mid X_k\right) + o_{L^1(P)}(n^{-\frac{1}{2}}) + o_{L^2(P)}(n^{-1}). \tag{A.47}$$

Observing (A.47) and

$$\begin{aligned} E\left(G_5\left(X_i, X_j, X_k\right) \mid X_k\right) &= E(L_g\left(X_i - X_j\right) \Delta\left(X_j\right) \left\{\, m\left(X_{i1}, X_{k2}\right) - \textstyle\int m\left(X_{i1}, y_2\right) f_2\left(y_2\right) dy_2 \,\right\} \mid X_k) \\ &= \textstyle\int \left(\Delta f^2\right)(x) \left\{\, m\left(x_1, X_{k2}\right) - \int m\left(x_1, y_2\right) f_2\left(y_2\right) dy_2 \,\right\} dx \,\left(1 + O\left(g^r\right)\right) \end{aligned}$$

when $i, j, k$ are pairwise different, it follows that

$$\begin{aligned} V_{5n}^{(1.3)} &= \frac{1}{n} \sum_{k=1}^{n} \int \left(\Delta f^2\right)(x) \left\{\, m\left(x_1, X_{k2}\right) - \int m\left(x_1, y_2\right) f_2\left(y_2\right) dy_2 \,\right\} dx \\ &\quad + o_{L^1(P)}(n^{-\frac{1}{2}}) + o_{L^2(P)}(n^{-1}). \end{aligned} \tag{A.48}$$

Note that

$$E(V_{5n}^{(1.3)}) = o(n^{-\frac{1}{2}}), \tag{A.49}$$

then a combination of (A.41) - (A.42), (A.45) - (A.46) and (A.48) - (A.49) (for different indices) yields

$$\begin{aligned} V_{5n}^{(1)} - E(V_{5n}^{(1)}) &= \frac{1}{n} \sum_{i=1}^{n} \{[\sigma\left(X_i\right) \varepsilon_i + m(X_i)] f^{-1}\left(X_i\right) f_2\left(X_{i2}\right) \int \left(\Delta f^2\right)\left(X_{i1}, y_2\right) dy_2 \\ &+ \int \left(\Delta f^2\right)(x)\; m\left(x_1, X_{i2}\right) dx - E[m(X_i) f^{-1}\left(X_i\right) f_2\left(X_{i2}\right) \int \left(\Delta f^2\right)\left(X_{i1}, y_2\right) dy_2 \\ &+ \int \left(\Delta f^2\right)(x)\; m\left(x_1, X_{i2}\right) dx] \} + o_P(n^{-\frac{1}{2}}) \end{aligned} \tag{A.50}$$

and

$$E(V_{5n}^{(1)}) = O\left(h^r + g^r\right) = o(n^{-\frac{1}{2}}), \tag{A.51}$$

which proves (A.37) and (A.38) in the case $\alpha = 1$ and $d = 2$.

Finally, the remaining term $V_{5n}^{(0)}$ is calculated as follows:

$$\begin{aligned} V_{5n}^{(0)} &= \frac{1}{n} \sum_{k=1}^{n} \left(Y_k - E\left(m\left(X_k\right)\right)\right) \left\{ \frac{1}{n\,(n-1)} \sum_{i} \sum_{j \neq i} L_g\left(X_i - X_j\right) \Delta\left(X_j\right) \right\} \\ &= \frac{1}{n} \sum_{k=1}^{n} \left\{ \sigma\left(X_k\right) \varepsilon_k + m\left(X_k\right) - \int (mf)(y) dy \right\} \int (\Delta f^2)(x) dx \\ &\quad + o_{L^1(P)}(n^{-\frac{1}{2}}) + o_{L^2(P)}(n^{-1}), \end{aligned}$$

which implies (A.40) and completes the proof of part (b).

*Proof of (c):* The random variable

$$V_{6n} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L_g\left(X^i - X^j\right) \Delta\left(X^i\right) \Delta\left(X^j\right)$$

is a U-statistic with symmetric kernel $G_6(x,y) := L_g(x-y)\,\Delta(x)\,\Delta(y)$ . A straightforward calculation gives $E\left(G_6^2\left(X^i, X^j\right)\right) = O(g^{-d}) = o(n)$ and it follows from Lemma 3.1 in Zheng (1996)

$$V_{6n} - E(V_{6n}) = \frac{2}{n} \sum_{i=1}^{n} E\left(G_6(X_i, X_j) \mid X_i\right) - E\left(G_6(X_i, X_j)\right) + o_P(n^{-\frac{1}{2}}). \tag{A.52}$$

A Taylor expansion gives

$$E\left(G_6\left(X^i, X^j\right) \mid X^i\right) = \left(\Delta^2 f\right)\left(X^i\right) + O\left(g^r\right) \tag{A.53}$$

(uniformly with respect to $i$) and

$$E\left(G_6\left(X^i, X^j\right)\right) = \int (\Delta f)^2(x)\,dx + O\left(g^r\right). \tag{A.54}$$

A combination of (A.52) - (A.54) yields both assertions in (c) and completes the proof of the second part of Theorem 1 (for the statistic $T_{4n}$). □

# References


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

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

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

Camlong-Viot, C. (2000) Modele additif de regression sous des conditions de melange. *PhD thesis at the Universit Toulouse III - Paul Sabatier.*

Collomb, G. and W. Härdle (1986) Strong uniform convergence rates in robust nonparametric time series analysis and prediction: kernel regression estimation from dependent observations. *Stoch. Proc. Appl.,* **23**: 77-89.

Deaton, A. and J.Muellbauer (1980) *Economics and Consumer Behavior.* Cambridge University Press: Cambridge.

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 C. von Lieres und Wilkau (2000) Testing additivity by kernel based methods – what is a reasonable test? *forthcomming in Bernoulli*

http://www.ruhr-uni-bochum.de/mathematik3/preprint.htm

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

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

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

Fan, J. and Q. Li (1996) Consistent model specification test: Omitted variables and semiparametric forms. *Econometrica* **64**: 865-890.

Fan, J. and Q. Li (1999) Central limit theorem for degenerate U-Statistics of absolutely regular processes with applications to model specification testing. *Nonparametric Statistics* **10**: 245-271.

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

Gozalo, P.L. and O.B.Linton (2000) Testing additivity in generalized nonparametric regression models. *forthcoming in the J. Econometrics.*

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

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

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

Jones, M.C., S.J.Davies and B.U.Park (1994) Versions of kernel-type regression estimators. *J. Am. Statist. Assoc.*, **89**: 825-832.

Kim, W., O.B.Linton and N.Hengartner (2000) A Computationally Efficient Oracle Estimator of Additive Nonparametric Regression with Bootstrap Confidence Intervals. *J. Computational and Graphical Statist.,* forthcoming

Li, Qi (2000) Efficient Estimation of Additive Partially Linear Models. *International Economic Review,* **41**: 1073-1092.

Linton, O.B. (1997) Efficient estimation of additive nonparamteric regression models. *Biometrika*, **84**: 469-473.

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

Mack, Y.B., Silverman, B.W. (1982) Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verw. Gebiete*, **60**, 405-415.

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

NIELSEN, J.P. AND O.B.LINTON (1997) An optimization interpretation of integration and backfitting estimators for separable nonparametric models. *J. Royal Statist. Soc., Series B,* **60**: 217-222.

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

PINSKE, J. (2000) Feasible Multivariate Nonparametric Regression Estimation Using Weak Separability. *Preprint, University of British Columbia, Canada*

SPERLICH, S., O.B.LINTON AND W.HÄRDLE (1999) Integration and Backfitting methods in additive models: Finite sample properties and comparison. *Test*, **8**: 419-458.

SPERLICH, S., D.TJØSTHEIM AND L.YANG (2000) Nonparametric Estimation and Testing of Interaction in Additive Models. *forthcoming in Econometric Theory*, `http://halweb.uc3m.es/esp/personal/personas/stefan/publik.htm`

STAUDTE, R.G. AND S.J.SHEATHER (1990) *Robust estimation and testing.* Wiley, New York.

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

STONE, C.J. (1994) The Use of Polynomial Splines and their Tensor Products in Multivariate Function Estimation. *Ann. Statist.*, **22**: 118-184.

TJØSTHEIM, D. AND B.H.AUESTAD (1994) Nonparametric identification of nonlinear time series: projections. *J. Am. Statist. Assoc.,* **89**: 1398-1409.

J.X. ZHENG (1996) A consistent test of a functional form via nonparametric estimation techniques. *J. Econometrics,* **75**: 263-289.