

# DFG-SNF Research Group FOR916

Statistical Regularization and Qualitative Constraints

Enno Mammen

Stefan Sperlich

## Backfitting Tests in Additive Interaction Models

Preprint FOR916 10-21

Preprint-Series of the Research Group FOR916

# Backfitting Tests in Additive Interaction Models

Enno Mammen\*      Stefan Sperlich†

April 16, 2010

## Abstract

This paper introduces a new test procedure for additive separable models. The test is applied to check for interaction terms, and to do variable selection. Further possible applications are testing for pure additivity or for endogeneity. The test is based on optimal nonparametric data fits on the hypothesis and on the alternative. The fits make use of the smooth backfitting technique that was introduced by Mammen, Linton and Nielsen (1999), and it is the first test of this kind. A complete asymptotic theory for the test is developed. We further discuss different wild bootstrap procedures for our test and compare different possible implementations. A comparison study with simulation and data examples shows excellent performance and explains the difference to existing methods in practice.

*Key words:* Nonparametric Testing; Interactions; Additive Models; Smooth Backfitting.

*Journal of Economic Literature Classification:* C14

---

\*Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany. E-mail address: emammen@rumms.uni-mannheim.de. Tel. 0049 621 181 1927.

†Institut für Statistik und Ökonometrie, Georg-August Universität Göttingen, Platz der Göttinger Sieben 5, 37073 Göttingen, Germany. E-mail address: stefan.sperlich@wiwi.uni-goettingen.de. Both authors acknowledge support by the DFG project FOR916.

# 1 Introduction

Additive models are an important tool in nonparametric regression. They allow a flexible modeling, are easy to interpret and there exist stable and reliable estimators. They avoid a lot of problems that are present if one uses a full dimensional nonparametric regression model. In such models estimators are unstable even for moderate sample sizes and the fitted regression functions cannot be visualized and interpreted. On the other hand, additive models are easily understood and analyzed. Furthermore, the additive structure is rather reasonable with many important implications in economic theory, see e.g. Deaton and Muellbauer (1980).

However, even though additivity is a desirable property in many empirical research, imposing this structure without empirical evidence can cause much stronger errors in estimation than the curse of dimensionality. Consequently it will lead to poor data fits and to wrong conclusions. Therefore many efforts have been undertaken to develop testing procedures for additivity in parametric as well as in nonparametric models. For nonparametric models there does not exist a fully satisfactory test. As has been shown by Dette, von Lieres und Wilkau, and Sperlich (2005), almost all proposed procedures have difficulties in practice when the regressors are correlated or data are sparse. Albeit the use of bootstrap they pointed out that the so far developed tests often have rather poor power and problems to hold the error size.

In this paper, we propose new tests for significant interactions, variable selection, endogeneity and other structured alternatives in an else additive model. Our test procedure is based on the comparison of nonparametric estimators constructed in different nonparametric specifications. As a comparison criterion we use the  $L_2$ -norm. A similar approach has been used in a series of testing problems where nonparametric estimators have been compared with estimators of the parametric component in a parametric or semiparametric model. A point that needs a lot of care in such tests are the bias terms of the nonparametric estimators. If one does not correct for these terms one may arrive at an asymptotically linear tests that mainly checks for deviations from the parametric model in one direction, see e.g. Härdle and Mammen (1993) for an early discussion. There exist different ways to avoid these bias effects. One way is to estimate the bias terms and to plug these terms into the test statistic. In another approach one modifies the parametric estimator such that it has asymptotically the same bias term as the nonparametric estimator. Then in the comparison of the parametric and the nonparametric estimator the bias terms cancel out. In this paper we follow the second approach. But in our case we do not compare a nonparametric and a parametric estimator, in our testing problem we compare two nonparametric estimators that are constructed in two different

nonparametric models with different nonparametric complexity. This requires a careful construction of the two estimators. Another point that has been raised in the discussion of nonparametric  $L_2$ -tests is the slow convergence of the test statistic to its normal limit. For this reason bootstrap procedures have been proposed that leads to more reliable distributional approximations and more accurate critical values. In the case of a parametric or also a semiparametric hypothesis, this can be easily implemented because one can show that one can resample from the fitted parametric or semiparametric model of the hypothesis. We will show that the same also works in our testing problem. And in particular, this can be done without using any additional smoothing parameter in fitting the nonparametric hypothesis. This is a very nice feature compared to other related proposals that requires oversmoothing to get higher order properties of the nonparametric hypothesis asymptotically correctly. Instead of marginal integration we use the smooth backfitting (SB hereafter) method of Mammen, Linton and Nielsen (1999). This is done mainly for two reasons:

- (a) SB directly addresses the question if an additive model can approximately describe the data. It checks adequacy of the additive model that is the  $L_2$ -optimal additive approximation of the full dimensional regression function. In contrast, marginal integration estimates the marginal average impact of the explanatory variables whatever the real underlying model structure is, and it fits the additive model that is based on marginal average characteristics of single explanatory variables.
- (b) Unlike other estimators, the SB turned out to perform excellent and stable and to efficiently circumvent the curse of dimensionality, see Nielsen and Sperlich (2005) or Roca-Pardiñas and Sperlich (2007). We will argue below that the bad performance [in small samples] of the so far existing nonparametric tests for additivity is inherited from the poor performance of the applied estimators.
- (c) Marginal integration only works under more restrictive assumptions. More specifically, one has to assume higher order smoothness for the additive functions and to use higher order kernels. The order of the smoothness assumptions and of the order of the kernels grows linearly in the number of the additive components, see Fan, Härdle and Mammen (1998). The assumptions for smooth backfitting do not depend on the number of components and they are comparable to one-dimensional smoothing.

The first two points, (a) and (b), matter in particular when regressors are correlated or the sample has areas where data are sparse, see e.g. Sperlich, Linton, and Härdle (1999) and Linton (1997). Another problem occurs when other interactions

than the considered one are significant. In Sperlich, Linton, and Härdle (1999) it is shown that also the classical backfitting estimator of Hastie and Tibshirani (1990) runs into problems when the design is correlated. The performance of SB is well understood. The basic asymptotic theory is given in Mammen, Linton and Nielsen (1999). Bandwidth choice and practical implementations are discussed in Mammen and Park (2005) and Nielsen and Sperlich (2005). SB methods for generalized additive models were introduced in Yu, Park and Mammen (2008). Haag (2006a) discusses SB for nonparametric additive diffusion models, Schienle (2007) for nonstationary data. Finally, SB has been used to estimate generalized structured models, see Mammen and Nielsen (2003) and Roca-Pardiñas and Sperlich (2007). Additive regression is an example of a nonparametric model where the nonparametric function is given as a solution of an integral equation. This has been outlined in Linton and Mammen (2002), Carrasco, Florens and Renault (2006), and Mammen and Yu (2009), where also other examples of statistical integral equations are given. Examples are additive models where the additive components are linked as in Linton and Mammen (2005), or regression models where a transformation leads to an additive model, see Linton and Mammen (2006) and Linton, Sperlich, Van Keilegom (2008). Related additivity tests are considered in Fan and Jiang (2005) and Haag (2006b). But they consider the case where the hypothesis is parametric and where the alternative is an additive model. Their tests are based on the comparison of parametric fits with additive nonparametric fits using classical or SB, respectively. Related tests for variable selection are Hrdle, Sperlich, and Spokoiny (2001) which use wavelets in additive models, or Sperlich, Tjøstheim, and Yang (2002), Yang, Sperlich, and Härdle (2003), Hrdle, Huet, Mammen, and Sperlich (2004) which proposed bootstrap inference with marginal integration estimates in generalized partial linear additive models. Gozalo and Linton (2001) consider the hypothesis of a generalized additive model and propose to use an  $L_2$ -test statistic that compares a nonparametric additive estimator with a full dimensional kernel smoother. Their approach is also based on marginal integration and needs higher order smoothness assumptions with order linearly increasing with the number of additive components. In particular, their assumptions imply that the bias terms of the nonparametric estimators are asymptotically negligible.

The paper is organized as follows. In the next section we first study the estimation of additive models with interaction terms by SB and study their asymptotics. Section 3 introduces the testing procedure, i.e. the statistic and bootstrap procedures and discusses its asymptotic behavior. The simulations in Section 4 demonstrate an excellent performance of our test proposal, and the data example highlights the difference to existing methods in practice. Section 5 concludes. Assumptions and all technical proofs are deferred to appendices.

## 2 Estimating Interactions by SB

Consider the following general set up. For a vector  $X$  of explanatory variables and a scalar response  $Y$  we model the conditional expectation  $m(x) = \mathbb{E}[Y|X = x]$  as

$$m(x) = c + \sum_{j=1}^r m_j(x_j), \quad (1)$$

where  $x = (x_1^T, \dots, x_r^T)^T$ . The dimensions of  $x$ ,  $x_1, \dots, x_r$  are denoted by  $d, d_1, \dots, d_r$ , respectively, with  $d_1 + \dots + d_r = d$ . For identifiability we assume that the functions  $m_j$  are  $\mathbb{R}^{d_j} \rightarrow \mathbb{R}$  fulfill  $\mathbb{E}[m_j(X_j)] = 0$  for  $j = 1, \dots, r$ . We want to test the following hypotheses

$$H_0^1 : \quad m_1(x_1) = \sum_{j=1}^s m_{1j}(x_{1j}), \quad d_1, s > 1, \quad (2)$$

$$H_0^2 : \quad m_1 \equiv 0, \quad s \in \mathbb{N} \quad (3)$$

where  $x_1 = (x_{11}^T, \dots, x_{1s}^T)^T$  with dimensions  $d_{11}, \dots, d_{1s}$ , respectively,  $d_{11} + \dots + d_{1s} = d_1$ . Here,  $m_{1j}$  are additive functions from  $\mathbb{R}^{d_{1j}}$  to  $\mathbb{R}$  ( $j = 1, \dots, s$ ).

We assume that  $X_j$  lie in compact sets. Furthermore, for the ease of presentation, we make the assumption that all elements of  $X$  lie in  $[0, 1]$ , i.e.  $X \in [0, 1]^d$  and  $X_j \in [0, 1]^{d_j}$  for  $j = 1, \dots, r$ . The smoothing requires the choice of a  $d$ -dimensional bandwidth vector  $h$ . We decompose the vector  $h$  into  $(h_1^T, \dots, h_r^T)^T$  where the vectors  $h_1, \dots, h_r$  have dimensions  $d_1, \dots, d_r$ . The diagonal matrix that has the elements of  $h_j$  as diagonal element is denoted by  $H_j$ . Furthermore, we have to specify a kernel  $K : \mathbb{R} \rightarrow \mathbb{R}$  that will be used in the construction of the SB estimator. We need that the kernel integrates to one over the support of  $X$ , here  $[0, 1]$ . This is achieved by the following boundary modification for bandwidth  $g > 0$

$$K_g(u, v) = \begin{cases} \frac{K[(u-v)/g]}{\int_0^1 K[(w-v)/g] dw} & \text{if } u, v \in [0, 1], \\ 0 & \text{else.} \end{cases}$$

For multivariate  $u = (u_1, \dots, u_q)^T$  and  $v = (v_1, \dots, v_q)^T$  and bandwidth vector  $g = (g_1, \dots, g_q)^T$  we denote the multiplicative kernel  $\prod_{j=1}^q K_{g_j}(u_j, v_j)$  by  $K_g(u, v)$ . Then, for a sample  $\{X^i, Y^i\}_{i=1}^n$ , local linear SB estimators  $\hat{c}, \hat{m}_1, \dots, \hat{m}_r, \hat{m}^1, \dots, \hat{m}^r$  of  $c, m_1, \dots, m_r$  and its derivatives  $m'_1, \dots, m'_r$  are defined as minimizers of

$$\sum_{i=1}^n \int [Y^i - \hat{c} - \sum_{j=1}^r \hat{m}_j(u_j) - (X_j^i - u_j)^T \hat{m}^j(u_j)]^2 K_h(u, X^i) du. \quad (4)$$

The minimization runs under the constraint that

$$\int \hat{m}_j(u_j) \hat{p}_j(u_j) du_j = 0,$$

where  $\widehat{p}_j(u_j) = n^{-1} \sum_{i=1}^n K_{h_j}(u_j, X_j^i)$  is a kernel density estimator of the marginal density  $p_j$  of  $X_j$ .

By differentiation of (4) w.r.t.  $\widehat{m}_j(u_j) + \widehat{c}$  and  $\widehat{m}^j(u_j)$ , one gets the following linear equations for the estimators:

$$\begin{aligned} \widehat{M}_j(x_j) \begin{pmatrix} \widehat{m}_j(x_j) \\ \widehat{m}^j(x_j) \end{pmatrix} &= \\ &= \widehat{M}_j(x_j) \begin{pmatrix} \widetilde{m}_j(x_j) \\ \widetilde{m}^j(x_j) \end{pmatrix} - \widehat{c} \begin{pmatrix} \widehat{V}_{0,0}^j(x_j) \\ \widehat{V}_{j,0}^j(x_j) \end{pmatrix} - \sum_{l \neq j} \int \widehat{S}_{l,j}(x_l, x_j) \begin{pmatrix} \widetilde{m}_l(x_l) \\ \widetilde{m}^l(x_l) \end{pmatrix} dx_l, \end{aligned}$$

where  $(\widetilde{m}_j, \widetilde{m}^j)$  together with a constant  $\widehat{c}_j$  are the marginal local linear estimators. They are defined as minimizers of

$$\sum_{i=1}^n \int [Y^i - \widehat{c}_j - \widetilde{m}_j(u_j) - (X_j^i - u_j)^T \widetilde{m}^j(u_j)]^2 K_{h_j}(u_j, X_j^i) du_j$$

under the constraint  $\int \widetilde{m}_j(u_j) \widehat{p}_j(u_j) du_j = 0$ . Furthermore,

$$\begin{aligned} \widehat{M}_j(x_j) &= \begin{pmatrix} \widehat{V}_{0,0}^j(x_j) & \widehat{V}_{j,0}^j(x_j) \\ \widehat{V}_{j,0}^j(x_j) & \widehat{V}_{j,j}^j(x_j) \end{pmatrix} \\ &= \frac{1}{N} \sum_{i=1}^n K_{h_j}(x_j, X_j^i) \begin{pmatrix} 1 & [X_j^i - x_j]^T \\ X_j^i - x_j & [X_j^i - x_j][X_j^i - x_j]^T \end{pmatrix}, \\ \widehat{S}_{l,j}(x_l, x_j) &= \frac{1}{N} \sum_{i=1}^n K_{h_j}(x_j, X_j^i) K_{h_l}(x_l, X_l^i) \begin{pmatrix} 1 & [X_l^i - x_l]^T \\ X_j^i - x_j & [X_j^i - x_j][X_l^i - x_l]^T \end{pmatrix}, \\ \widehat{V}_{0,0}^j(x_j) &= \frac{1}{N} \sum_{i=1}^n K_{h_j}(x_j, X_j^i), \\ \widehat{V}_{j,0}^j(x_j) &= \frac{1}{N} \sum_{i=1}^n K_{h_j}(x_j, X_j^i) (X_j^i - x_j). \end{aligned}$$

Using these equations, the local linear SB estimator can be calculated by an iterative algorithm as described in Mammen, Linton and Nielsen (1999) for the special case  $d_1 = \dots = d_r = 1$ . There, the matrices were slightly differently defined with some elements scaled by bandwidths. This was done mostly for simplification of the notation in the proofs. Note finally, that we have used here local linear instead of local constant estimation to get rid of additional bias terms which would otherwise distort the test statistics in Section 3.

For completeness, we state a theorem on the asymptotic distribution of

**Theorem 2.1.** *Under assumptions (A1), (A2), (A3'), (A4), (A5), it holds that*

$$\sup_{u \in [0,1]^{d_1}} |\widehat{m}_1(u) - \widehat{m}_1^A(u) - \widehat{m}_1^B(u)| = o_P([n\pi(h_1)]^{-1/2} + \|h\|^2),$$

where  $\pi(h_1)$  denotes the product of the elements of  $h_1$  and where

$$\begin{aligned}\widehat{m}_1^A(u) &= \frac{\sum_{i=1}^n K_{h_1}(u, X_1^i) \varepsilon^i}{\sum_{i=1}^n K_{h_1}(u, X_1^i)}, \\ \widehat{m}_1^B(u) &= \begin{cases} \frac{1}{2} h_1^T m_1''(u) h_1 \int v^2 K(v) dv & \text{if } h_1 \leq u \leq 1 - h_1 \text{ componentwise,} \\ O_P(\|h_1\|^2) & \text{else,} \end{cases}\end{aligned}$$

with  $\varepsilon^i = Y^i - m(X^i)$ .

While this second theorem is of its own interest for nonparametric regression, Theorem B.1 arms us with the adequate pre-estimators for our testing purposes.

### 3 Testing in Additive Interaction Models

We now introduce our test procedures. As variable selection, testing for endogeneity, etc. are special applications of our more general statistic, we start with testing hypothesis (2). The test statistic is based on a comparison of a nonparametric fit in model (1) with an additive estimate under the additional constraint (2). The fit under the additional assumption of (2) can be achieved by projection of  $\widehat{m}_1$ :

$$(\widehat{m}_{11}, \dots, \widehat{m}_{1s}) = \underset{\bar{m}_{11}, \dots, \bar{m}_{1s}}{\operatorname{argmin}} \int \left\{ \widehat{m}_1(x_1) - \sum_{j=1}^s \bar{m}_{1j}(x_{1j}) \right\}^2 \widehat{p}_1(x_1) dx_1$$

Alternatively, the components  $m_{11}, \dots, m_{1s}$  could be directly estimated by SB in model (1) with  $m_1(x_1)$  replaced by  $m_{11}(x_{11}) + \dots + m_{1s}(x_{1s})$ . However, this approach leads to some complications because the resulting bias structure turns out to be much more complicated.

Like for SB, the projection can be calculated by an iterative algorithm. In this algorithm the left hand side of the following equation is iteratively updated:

$$\widehat{m}_{1j}(x_{1j}) = \int \left\{ \widehat{m}_1(x_1) - \sum_{l \neq j} \widehat{m}_{1l}(x_{1l}) \right\} \frac{\widehat{p}_1(x_1)}{\widehat{p}_{1j}(x_{1j})} dx_{11} \cdots dx_{1j-1} dx_{1j+1} \cdots dx_{1s}.$$

The adequacy of additivity as in (2) can be measured by the squared difference

$$T_n = \int \left\{ \widehat{m}_1(x_1) - \sum_{j=1}^s \widehat{m}_{1j}(x_{1j}) \right\}^2 \widehat{p}_1(x_1) dx_1. \quad (5)$$

We propose to use  $T_n$  as test statistic for hypothesis (2). The asymptotic distribution of  $T_n$  is given in Theorem 3.1. It is formulated for neighbored alternative regression functions of the type

$$m(x) = c + \sum_{l=1}^s m_{1l}(x_{1l}) + \sum_{j=2}^r m_j(x_j) + n^{-1/2} \pi(h_1)^{-1/4} \Delta_1(x_1), \quad (6)$$

where  $\Delta_1$  is a function that is orthogonal to all additive functions in  $L_2(p_1)$ , i.e.

$$\int \Delta_1(x_1)p_1(x_1)dx_{11} \cdot \dots \cdot dx_{1j-1}dx_{1j+1} \cdot \dots \cdot dx_{1s} = 0 \quad (7)$$

for  $j = 1, \dots, s$ . Then, the first theorem determines the asymptotic power of our test:

**Theorem 3.1.** *Under assumptions (A1)-(A5),(A7)-(A9), the statistic*

$$n\pi(h_1)^{1/2} \left[ T_n - n^{-1}\pi(h_1)^{-1}K^{(2)}(0)^{d_1} \int_{[0,1]^{d_1}} \sigma^2(x_1) dx_1 \right]$$

*has a limiting normal distribution with mean  $\int \Delta_1^2(x_1)dx_1$  and variance*

$$2 \int_{[0,1]^{d_1}} \sigma^4(x_1) dx_1 K^{(4)}(0)^{d_1} \quad . \quad (8)$$

*Here,  $\sigma^2(x_1)$  is the conditional variance of  $\varepsilon_1$  given  $X_1 = x_1$ . Furthermore,  $K^{(j)}$  denotes the  $j$ -times convolution product of  $K$  (for  $j \geq 1$ ).*

The normal approximation in Theorem 3.1 is useful for an asymptotic description of the performance of the test but the accuracy is too poor for reliable finite sample approximations of critical values. Basically in the background of the theorem there acts a central limit theorem with  $O(\pi(h_1)^{-1})$  independent observations. This is the order of the number of non overlapping intervals in the smoothing. This number is very small in finite samples and thus the normal approximation is relatively crude. This problem has already been shown for a simpler test problem in Härdle and Mammen (1993). In that paper the wild bootstrap method was introduced for nonparametric testing, motivated by finding approximative critical values. We follow this suggestion and have tried four different versions:

1. Sample  $Y^{i,*} = \sum_{l=1}^s \widehat{m}_{1l}(X_{1l}^i) + \sum_{j=2}^r \widehat{m}_j(X_j^i) + \varepsilon^{i,*}$  where the  $\varepsilon^{i,*}$  are generated by using wild bootstrap with residuals from the Null hypothesis:  $\varepsilon^{i,*} = \widehat{\varepsilon}^i \eta^i$  where  $\eta_i$  is an i.i.d. sequence with mean 0, second and third moment equal to 1, and  $\widehat{\varepsilon}^i = Y^i - \sum_{l=1}^s \widehat{m}_{1l}(X_{1l}^i) - \sum_{j=2}^r \widehat{m}_j(X_j^i)$ .
2. Like in 1, but with residuals taken from the alternative,  $\widehat{\varepsilon}^i = Y^i - \sum_{j=1}^r \widehat{m}_j(X_j^i)$ .
3. Take  $Y^{i,*} = \varepsilon^{i,*}$ , where the  $\varepsilon^{i,*}$  are generated as in 1.
4. Same as 3., but with residuals calculated under the alternative.

In all resampling schemes the bootstrap test statistics  $T_n^*$  are constructed from the samples  $\{X^i, Y^{i,*}\}_{i=1}^n$ . Note that all these wild bootstrap implementations do not require the choice of an additional smoothing parameter. All four wild bootstrap methods work asymptotically, and give a consistent estimate of critical values.

**Theorem 3.2.** *Under assumptions (A1)-(A5), (A8)-(A10), conditionally given the sample, the (normalized) wild bootstrap statistic*

$$n\pi(h_1)^{1/2} \left[ T_n^* - n^{-1}\pi(h_1)^{-1}K^{(2)}(0)^{d_1} \int_{[0,1]^{d_1}} \sigma^2(x_1) dx_1 \right]$$

*has a limiting normal distribution with mean 0 and variance (8).*

Note that in our simulations, method 1 was the most reliable approach. Methods 3 and 4 neglect higher order effects of bias terms and this may be the reason why the fitted critical values were not very accurate. Method 2 leads to tests that are too liberal. The reason is that in practice, often the residuals of the alternatives are much smaller than for the null model, even if the null is true.

In case one is interested in testing whether one has a pure additive model in terms of  $r = d$  in model (1), there surely exist various ways how to proceed with our test statistic. The simplest thing to do is to set  $x_1 = x$  and  $s = d$  in (2). While the disadvantages have been discussed in the introduction (we refer once more to Dette, von Lieres und Wilkau, and Sperlich, 2005), the advantage is also obvious: there is no additional effort needed to control for the all over error rate. This is different when one follows the idea of testing additivity stepwise by checking one by one for certain significant interaction terms. Here, a statistical advantage is the circumvention of the curse of dimensionality, and the econometric advantage is the possibility to allow only for interactions which make sense, and in case of rejection to know which are significant and which are not. Strategies to control the over all significance level in stepwise multiple testing are still discussed in the literature, see for example Romano and Wolf (2005a, 2005b, 2007) for a recent discussion.

Little more effort is necessary to sketch the idea of variable selection. Here we can directly use our theorem from above and reformulate them as corollaries for the case of when hypothesis (3) is tested. In order to do so, let us denote the squared deviation from  $H_0^2$  by

$$S_n = \int \widehat{m}_1^2(x_1) \widehat{p}_1(x_1) dx_1. \quad (9)$$

We propose to use  $S_n$  as a test statistic for hypothesis (3), and denote its bootstrap analogue by  $S_n^*$ . Here, the neighbored alternative regression functions we are thinking of are of the type

$$m(x) = c + \sum_{j=2}^r m_j(x_j) + n^{-1/2}\pi(h_1)^{-1/4}\Delta_2(x_1), \quad \int \Delta_2(x_1)p_1(x_1)dx_1 \neq 0. \quad (10)$$

**Corollary 3.1.** *Under assumptions (A1)-(A5), (A8)-(A9), and (A7) with  $\Delta_2$  substituted for  $\Delta_1$ , the statistic*

$$n\pi(h_1)^{1/2} \left[ S_n - n^{-1}\pi(h_1)^{-1}K^{(2)}(0)^{d_1} \int_{[0,1]^{d_1}} \sigma^2(x_1) dx_1 \right]$$

*has a limiting normal distribution with mean  $\int \Delta_2^2(x_1)dx_1$  and variance (8).*

*Further, under assumptions (A1)-(A5), (A8)-(A10), conditionally given the sample, the (normalized) wild bootstrap statistic*

$$n\pi(h_1)^{1/2} \left[ S_n^* - n^{-1}\pi(h_1)^{-1}K^{(2)}(0)^{d_1} \int_{[0,1]^{d_1}} \sigma^2(x_1) dx_1 \right]$$

*has a limiting normal distribution with mean 0 and variance (8).*

Concerning the problem of whether  $X_1$  should include all covariates in question at once or step by step, we refer to the discussion on multiple testing above.

To overcome problems of endogeneity, Newey, Powell, and Vella (1999), and Moral-Arce, Rodriguez-Po, and Sperlich (2007) use in their additive separable models non-parametric instruments and additive control functions. When we abstract from the censoring problem in Moral-Arce, Rodriguez-Po, and Sperlich (2007), then both consider the following setup:

$$Y = m(X, Z) + \varepsilon, \quad E[\varepsilon|Z] = 0, \quad (11)$$

$$X = g(W, Z) + e, \quad E[e|W, Z] = 0, \quad E[\varepsilon|Z] = 0, \quad (12)$$

but  $E[\varepsilon|e, Z] = E[\varepsilon|e]$  possibly not being equal to zero. Then, denoting  $E[\varepsilon|e]$  as  $m_1(e)$ , our function of interest,  $m(x, z)$ , can be estimated by nonparametrically regressing the additive model, see (11) and (12),

$$E[Y|X, Z, W] = m(X, Z) + E[\varepsilon|X, Z] = m(X, Z) + m_1(e) = m_{aux}(X, Z, e) .$$

In practice, the procedure is to first regress the possibly endogenous variable  $X$  on instruments  $(W, Z)$ , and afterwards regress  $Y$  on  $(X, Z, \hat{e})$  with  $\hat{e} = X - \hat{g}(W, Z)$ . Independently from the first regression, in the second one we propose to apply SB estimation to obtain  $\hat{m}_1$  and  $\hat{m}$ . Now, apply  $S_n$  to check  $H_0^2$  with  $e = x_1$  in (3). Note that this is equivalent to test whether  $X$  is endogenous in (11).

## 4 An Empirical Study of the Test

We first study our test comparing it to other tests by simulations and real data applications. The comparison is based on results given in and Dette, von Lieres

und Wilkau, and Sperlich (2005), and in Sperlich, Tjøstheim, and Yang (2002). At the end of this section we also comment on the alternative bootstrap implementations and summarize our findings. As already explained in the introduction, it suffices to restrict the simulations to the problem of testing additivity, respectively for significant interactions.

In all simulations we dealt with  $s = 2, r = 2$  and  $d_{11} = d_{12} = d_2 = 1$ , such that the regression function is  $m(x_{11}, x_{1,2}, x_2) = c + m_{11}(x_{11}) + m_{12}(x_{12}) + m_2(x_2)$  under the null hypothesis, and  $m(x_{11}, x_{12}, x_2) = c + m_1(x_{11}, x_{12}) + m_2(x_2)$  under the alternative. In this set up it is more natural to use a different way of indexing, writing  $m(x_1, x_2, x_3) = c + m_1(x_1) + m_2(x_2) + m_3(x_3)$  for the null hypothesis and  $m(x_1, x_2, x_3) = c + m_{1,2}(x_1, x_2) + m_3(x_3)$  for the alternative.

Like in the above mentioned papers, we used quartic product kernel throughout. The bandwidths  $h_1, h_2$  needed for kernel density estimation of  $p_{12}, p_1, p_2$  and for estimation of  $m_1, m_2$  out of  $m_{1,2}$  were chosen as  $h_j \approx \text{std}(X_j)n^{-1/5}$ , where  $\text{std}(X_j)$  stands for the empirical standard deviation of  $X_j$ . In the simulation study we have replaced  $\text{std}(X_j)$  by the interquartile range of  $X_j$  divided by 1.34.

## 4.1 The simulation study

In Dette, von Lieres und Wilkau, and Sperlich (2005) the following additive interaction model is considered:

$$m(x) = E(Y|X = x) = c + \sum_{j=1}^3 m_j(x_j) + \eta_{12}(x_1, x_2) \quad (13)$$

such that  $m_{1,2}(x_1, x_2) = m_1(x_1) + m_2(x_2) + \eta_{12}(x_1, x_2)$ , where

$$\begin{aligned} m_1(u) &= 2 \sin(\pi u) & , & & m_2(u) &= u^2, \\ m_3(u) &= u & \text{and} & & \eta_{12}(u, v) &= auv. \end{aligned} \quad (14)$$

Further,  $\varepsilon \sim N(0, 1)$ ,  $(X_1, X_2, X_3)^T \sim N\{0, \Sigma_\gamma\}$ , with

$$\Sigma_\gamma = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}, \quad (15)$$

where they set for  $\gamma = 1$ :  $\rho_{12} = \rho_{13} = \rho_{23} = 0$ , for  $\gamma = 2$ :  $\rho_{12} = 0.2$ ,  $\rho_{13} = 0.4$ ,  $\rho_{23} = 0.6$ , and for  $\gamma = 3$ :  $\rho_{12} = 0.4$ ,  $\rho_{13} = 0.6$ ,  $\rho_{23} = 0.8$ . They generated samples of size  $n = 100$  to study four different tests statistics:

$$\begin{aligned} T_1 &= \frac{1}{n} \sum_i [\hat{m}^I(X^i) - \hat{m}_0^I(X^i)]^2, & T_2 &= \frac{1}{n} \sum_i \hat{\varepsilon}_i [\hat{m}^I(X^i) - \hat{m}_0^I(X^i)] \\ T_3 &= \frac{1}{n} \sum_i [(\hat{\varepsilon}_i)^2 - (\hat{u}_i)^2], & T_4 &= \frac{1}{n(n-1)} \sum_{j \neq i} L_g(X^i - X^j) \hat{\varepsilon}_i \hat{\varepsilon}_j. \end{aligned} \quad (16)$$

In statistic  $T_4$ , the function  $L$  denotes a bounded  $d$ -dimensional symmetric kernel with compact support,  $L_g(\cdot) = \frac{1}{g^d}L_g(\frac{\cdot}{g})$ ,  $g > 0$ . Dette et al. (2005) used (product) quartic kernels and tried out different bandwidths  $g$ . In  $T_1$  and  $T_2$ ,  $\hat{m}^I$  is the multidimensional internalized Nadaraya-Watson smoother with kernel  $L$  and bandwidth  $h$ , whereas  $\hat{m}_0^I$  indicates the internalized marginal integration estimator under the null hypothesis. Further,  $\hat{\varepsilon}_i$  are the residuals under the null hypothesis, and  $\hat{u}_i = Y^i - \hat{m}^I(X^i)$  in  $T_3$  denotes the  $i$ th residual under the alternative. They chose optimal cross validation bandwidths for both, the null and the alternative model. All given results were calculated with 500 bootstrap replications and 1000 simulation runs. Note that for the correlated cases ( $\gamma = 2$ ,  $\gamma = 3$ ) they found that all tests failed to work. Even when they had used the transformation  $X \rightarrow (\arctan(X) \cdot 2.4/\pi + 1.0) \cdot 0.5$ , to map the covariates into the interval  $[-0.1, 1.1]$ , and for increased sample size ( $n = 150$ ) the tests failed to work, for example none of the four hold the level.

We compare our test with the implementation that had the best performance in their study. This is the untrimmed version with using internalized kernel smoothing and  $g = 0.5 \text{ std}(X)$  for  $T_4$ . For the implementation of our test we have chosen bandwidths that lead to virtually smooth estimates:  $h_j = c_j \text{std}(X_j) \left(\frac{100}{n}\right)^{1/6}$ , for  $j = 1, 2$  (the directions of interest), and  $h_3 = c_3 \text{std}(X_3) \left(\frac{100}{n}\right)^{1/5}$  with  $c_1 = 0.8$ ,  $c_2 = 1.2$ , and  $c_3 = 1.4$ .

$\alpha$	hypothesis, $a = 0.0$					alternative, $a = 2.0$				
	$T_0$	$T_1$	$T_2$	$T_3$	$T_4$	$T_0$	$T_1$	$T_2$	$T_3$	$T_4$
5%	.024	.023	.010	.006	.078	.900	.844	.826	.768	.125
1%	.002	.000	.000	.000	.011	.752	.365	.259	.135	.034

Table 1: Percentages of rejection in model (14)-(15) with test statistics  $T_0$  ( $= T_n$ , as defined in (5)) and  $T_1, \dots, T_4$  (see (16)), uncorrelated regressors ( $\gamma = 1$ ), and with untransformed regressors. The results for  $T_1, \dots, T_4$  are taken from Dette et al. (2005).

In the simulations reported in Table 1, our test clearly outperforms the others; it further always holds the level, but is a little bit conservative. For  $\Sigma_2$  the rejections under  $H_0$  are 0.2 and 3.2% for the 1, respectively 5% nominal level. For  $\Sigma_3$  we got 0.01 and 2.2% respectively. Note that this is also due to the ad hoc choice of bandwidths (for each alternative and significance level there exist an “optimal” bandwidth). Unfortunately, for  $\Sigma_2$  and  $\Sigma_3$  the alternative that has been chosen in Dette et al. (2005) lies very near to the hypothesis, such that their tests had at most trivial power, sometimes the percentage of rejections even went down compared to  $H_0$ . They argued that this may happen due to the problems marginal integration has with correlated designs. On the other hand, the internalized version they used

should not share this handicap to that extreme. In contrast to these bad news, our new test shows nontrivial power: we found for  $\Sigma_3$  a power of 12.5% for the significance level  $\alpha = 5\%$ .

In summary, our test performs quite well and stable, and it outperforms all tests that were considered by Dette et al. (2005). Next, we turn to a comparison with the tests introduced in Sperlich, Tjøstheim and Yang (2002).

In their simulation study they draw samples of size  $n = 150$  from model (13) with

$$\begin{aligned} m_1(u) &= 2u & , & & m_2(u) &= 1.5 \sin(-1.5u) & , & & (17) \\ m_3(u) &= -u^2 + E(X_3^2) & \text{ and } & & \eta_{12}(u, v) &= auv & . \end{aligned}$$

The input variables  $X_j$ ,  $j = 1, 2, 3$ , were i.i.d. uniform on  $[-2, 2]$ . To generate the response  $Y$  they added normally distributed error terms with standard deviation  $\sigma_\varepsilon = 0.5$  to the regression function  $m(x)$ . They determined the critical value on the base of 249 bootstrap replications for the following statistics

$$\frac{1}{n} \sum_{i=1}^n \widehat{\eta}_{12}^2(X_1^i, X_2^i) \mathbb{1}\{|X_k^i| \leq 1.6 \text{ for } k = 1, 2\} \quad (18)$$

$$\frac{1}{n} \sum_{i=1}^n (\widehat{\eta}_{12}^{(1,1)})^2(X_1^i, X_2^i) \mathbb{1}\{|X_k^i| \leq 1.6 \text{ for } k = 1, 2\}. \quad (19)$$

Here  $\widehat{\eta}_{12}^{(1,1)}$  is an estimator of the mixed derivative  $(\partial^2 \eta_{1,2}) / (\partial x_1 \partial x_2)$ . Table 2 gives the errors of the first kind for statistics (18) and (19) as reported in Sperlich, Tjøstheim and Yang (2002), and of our test  $T_0$  with bandwidth  $h = 0.75$ . For a fair comparison we restrict our test statistic on the trimmed support  $|X_\alpha| \leq 1.6$ ,  $\alpha = 1, 2$ , as in Sperlich, Tjøstheim and Yang (2002). In additional simulations it turned out that with full support the size of  $T_0$  (under  $H_0$ ) decreased whereas the power increased a lot. For  $T_0$  values are calculated from 100 simulation runs.

From Table 2 we see that our test always holds the level. Further, Figure 1 shows that it also has strongest power. This is somewhat surprising because we have uncorrelated covariates here so that along the arguments of Dette et al. (2005) we would have expected more power for tests (18) and (19). Obviously, the excellent performance of SB compensates the advantages of using marginal integration even when having uncorrelated designs. Already for  $a = 0.4$  the alternative is almost surely detected by  $T_0$  for any level  $\alpha \geq 1\%$ .

In Section 3 we discussed modifications of the bootstrap method. In the first modification instead of taking the residuals from the null hypothesis model, one calculates the residuals under the alternative model. This was proposed in Härdle and Mammen (1993) because typically it will improve the power of the bootstrap test. However, at least for moderate sample sizes the procedure then becomes quite sensitive

significance level in %	1	5	10	15
test statistic (18)	.030	.060	.127	.173
test statistic (19)	.005	.045	.114	.144
test statistic $T_0$	.006	.032	.054	.120

Table 2: Percentage of rejections under  $H_0$  in model (17). The results for the statistics (18) and (19) are taken from Sperlich, Tjøstheim and Yang (2002).

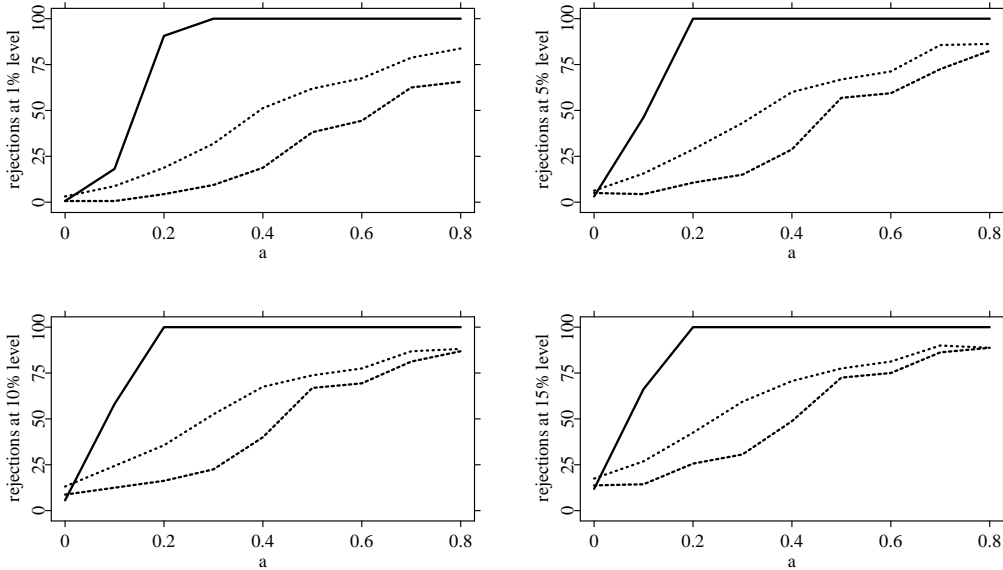


Figure 1: Power functions at the 1, 5, 10 and 15% significance levels for the three test statistics  $T_0$  (solid, highest) (18) (dotted, second highest), and (19) (dashed, lowest).

to the bandwidth choice. In particular one needs a very careful choice of the bandwidth that is used when fitting the regression function on the null hypothesis. This function is used in the wild bootstrap as true regression function. To guarantee that the test holds the level one has to choose a larger bandwidth, see Härdle and Marron (1991). We tried this in different simulation settings. But for almost all bandwidth combinations we then got a (much) too liberal test. The same problem appeared when we generated bootstrap samples under the null by just taking  $Y^* = \hat{\varepsilon}^*$ , no matter whether we took residuals from the null or the alternative model. One would expect that this problem disappears for large samples, but still for  $n = 500$  we got no satisfactory results.

## 4.2 Testing for Additivity in Farm Production

We finally consider the question, if the typically assumed additive separability in log-production functions is tenable. The data set has been studied in Sperlich,

Tjøstheim and Yang (2002), and by Severance-Lossin and Sperlich (1999). We will discuss how our approach compares to the procedure of Sperlich et al. (2002). A set of  $n = 250$  observations of Wisconsin farms was used, collected by the Farm Credit Service of St. Paul, Minnesota in 1987. The data consist of farm level inputs and outputs measured in dollars. The output  $Y$  in this analysis is livestock; the five input variables are family labor  $X_1$ , hired labor  $X_2$ , miscellaneous inputs (e.g. repairs, rent, custom hiring, supplies, insurance, gas, oil, and utilities)  $X_3$ , animal inputs (purchased feed, breeding, and veterinary services)  $X_4$ , and intermediate run assets (assets with a useful life of one to ten years)  $X_5$ . For more details see Severance-Lossin and Sperlich (1999).

Sperlich et al. (2002) used the marginal integration to estimate the model

$$\ln(y) = c + \sum_{\alpha=1}^5 m_{\alpha} \{\ln(x_{\alpha})\} + \sum_{1 \leq \alpha < \beta \leq 5} \eta_{\alpha\beta} \{\ln(x_{\alpha}), \ln(x_{\beta})\}, \quad (20)$$

so they allowed for nonparametric second order interaction terms. Quartic kernel and various bandwidth combinations were used. To test the different interaction terms for significance, they used test statistics (18) and (19) in an iterative model selection procedure: in a first step the p-values for each interaction term  $\eta_{\alpha\beta}$  were calculated with all other functions  $m_{\gamma}$ ,  $1 \leq \gamma \leq d$  and  $\eta_{\gamma\delta}$ ,  $1 \leq \gamma < \delta \leq d$ ,  $(\gamma, d) \neq (\alpha, \beta)$  included. The function  $\eta_{\alpha\beta}$  with the highest p-value was removed, and the p-values for the remaining interaction terms were determined as before. Stepwise eliminating the interaction terms with the highest p-value, the algorithm ended with the most significant ones.

Applying statistic (18), the term  $\eta_{13}$  (family labor and miscellaneous inputs) turned out to be significant with a p-value of about 2%. From the other terms,  $\eta_{35}$  and  $\eta_{15}$  had the smallest p-values, but both  $\gg 5\%$ . For statistic (19),  $\eta_{15}$  (family labor and intermediate run assets) and  $\eta_{35}$  (miscellaneous inputs and intermediate run assets) had the lowest p-values. The observed critical value of  $\eta_{15}$  was smaller than 1%. This suggests that the joint input of  $X_1$ ,  $X_3$  and  $X_5$  is not well approximated by the additive sum of the estimates of their marginal impacts. The different outcomes are not a contradiction as (18) looks more for smooth alternatives, whereas (19) more easily detects local deviations like a single peak or bump.

In the application of our test procedure, we started with an additive model and then tested the interaction terms one by one for significance. This is done because given the definition of backfitting estimates (4), it is not that clear how to estimate overlapping second order interaction terms simultaneously. Thus, in the first step

the test problem is a purely additive model versus

$$\ln(y) = c + \sum_{\alpha=1}^5 m_{\alpha} \{\ln(x_{\alpha})\} + \eta_{1,2} \{\ln(x_2), \ln(x_2)\}. \quad (21)$$

We tried several bandwidths and evaluated the test statistic with and without trimming. For the trimming we skipped about 10% of the range in each direction. For all reasonable bandwidths the p-values turned out to be bigger than 10% for all interactions. Without trimming the test only rejected for very small bandwidths the hypothesis of no interaction  $\eta_{2,3}$  and  $\eta_{3,4}$  with observed p-values between 7.5 and 10%. These rejections were caused by boundary peaks that are of minor importance for the interpretation of the data.

When we compare the definitions of marginal integration and backfitting (Nielsen and Linton, 1998, or Sperlich, Linton, and Hrdle, 1999) with the discrepancy measures described by our test statistics, we see that the SB estimator is congruent with our test statistic whereas this is not that clear for marginal integration and (18), (19). Now the seemingly contradicting outcomes for  $T_n$  versus (18), and (19) respectively, become more clear. When taking the sum of the one dimensional marginal impacts estimated via marginal integration, the inclusion of possible interactions between  $X_1$ ,  $X_3$  and  $X_5$  changes significantly the log-production function. However, the sum of the one dimensional additive SB-components gives such a good data fit that the inclusion of further interaction terms would not significantly improve this fit. We conclude that the decision about whether to include interactions or not depends very much on the estimators used. Recalling Nielsen and Linton (1998) this does not surprise that much. Finally, another possible explanation for the different testing outcomes might come from the bandwidths. Sperlich et al. (2002) had the problem that they had to choose oversmoothing bandwidths for the bootstrap, compare discussion above. While they could control this choice in their simulation study, it is not perfectly clear what has happened in this application. Maybe the bandwidth choice for the bootstrap has caused too liberal tests. But as we speak of real data it is impossible to know this for sure.

## 5 Conclusions

In this paper we have proposed a new test procedure for a general class of testing problems in additive models. We have shown that SB can be used to construct a stable and reliable test. We argue that the test outperforms other approaches used so far in additive models. There are several reasons. First, SB has been shown to outperform any other additive model estimator (Nielsen and Sperlich, 2005),

especially for higher dimensions and in case of regions with sparse data like it uses to happen for correlated designs. In those cases for example, marginal integration and classical backfitting estimates often show poor performance, compare Sperlich et al. (1999). This property then is inherited by testing procedures based on these estimators. Secondly, SB fits the best additive approximation. Marginal integration for example, fits average effects of (tuples of) covariates. If the model is not additive, then our test checks for the significance of interactions by measuring the distance to the nearest additive model. This coherence between our statistic and the used estimators makes the test so powerful, and its outcomes easily interpretable.

Our testing approach could be used for model choice. In our general class of test hypotheses we consider different specifications of the additive model. In each specification functions of tuples of regressors enter (or not) additively into the model. The models differ by the way how the regressors are divided into different tuples. The same idea applies to the problem of variable selection. Finally, we also have shown how to test for endogeneity in a semiparametric additive interaction model.

Our results are given with complete asymptotic theory. Bootstrap strategies, implementation and bandwidth choice has been discussed. To conclude, we have introduced the class of test procedures treating both aspects, theory and practice.

## A Assumptions

(A1) The kernel  $K$  is bounded, has compact support ( $[-1, 1]$ , say), is symmetric about zero, and is Lipschitz continuous, i.e. there exists a positive finite constant  $C$  such that  $|K(t_1) - K(t_2)| \leq C|t_1 - t_2|$ .

(A2) The  $d$ -dimensional vector  $X^i = (X_1^i, \dots, X_r^i)^T$  has compact support. Without loss of generality we assume that the support of  $X_j^i$  is equal to  $[0, 1]^{d_j}$ ,  $j = 1, \dots, r$ ;  $d = d_1 + \dots + d_r$ . The density  $p$  of  $X^i$  is bounded away from zero and infinity on  $[0, 1]^d$ . The tuples  $(X^i, \varepsilon^i)$  are i.i.d.

(A3) Given  $X^i$  the error variable  $\varepsilon^i$  has conditional zero mean and it holds for some  $\gamma > 4$  and  $C' < \infty$ , that

$$E \left[ |\varepsilon^i|^\gamma \middle| X^i \right] < C' \quad \text{a.s..}$$

(A3') Condition (A3) holds for some  $\gamma > 5/2$ .

- (A4) The (matrix/vector-valued) functions  $m_j'', p_j'$  and  $(\partial/\partial x_j)p_{jk}(x_j, x_k)$  ( $1 \leq j, k \leq r$ ) exist and are continuous. Here,  $p_j$  denotes the density of  $X_j$ . The joint density of  $X_j$  and  $X_k$  is denoted by  $p_{jk}$ .
- (A5) For a  $\delta > 0$  it holds that  $n^{1-\delta}\pi(h_j)\pi(h_k) \rightarrow \infty$  for  $1 \leq j < k \leq r$ . Here  $\pi(h_j)$  denotes the product of the elements of the vector  $h_j$ .
- (A6) It holds that  $\|m_j''(u) - m_j''(v)\| \leq C\|u - v\|^{\rho_j}$  for some constants  $\rho_j \geq 0$ .
- (A7) The regression function  $m$  is defined as in (6) with  $\Delta_1$  fulfilling (7). The function  $\Delta_1$  is continuous on  $[0, 1]^{d_1}$ .
- (A8) Assumption (A6) holds with constants  $\rho_j$  such that  $\|h_j\|^{4+2\rho_j} = o(n^{-1}\pi(h_1)^{-1/2})$  if  $\rho_j > 0$  and such that  $\|h_j\|^4 = O(n^{-1}\pi(h_1)^{-1/2})$  if  $\rho_j = 0$ .
- (A9) The conditional variance  $\sigma^2(x_1) = E[\varepsilon^2|X_1 = x_1]$  is continuous on  $[0, 1]^{d_1}$  and strictly positive.
- (A10) The regression function  $m$  fulfills (1) with (2). The wild bootstrap multipliers  $\eta_i$  have subexponential tails.  $E[\exp(u\eta_i)] < \infty$  for  $|u|$  small enough.

## B Proofs

### B.1 Proof of Theorem 2.1

One can proceed with similar arguments as in the proof of Theorem 4' in Mammen, Linton and Nielsen (1999) where the statement was proved for the case  $d_1 = \dots = d_r = 1$ .

### B.2 Proof of Theorem 3.1

In the first theorem we state a result that gives a uniform expansion for the additive fits. In the formulation of the theorem for a function  $g : \mathbb{R}^q \rightarrow \mathbb{R}$  we denote by  $g'(x)$  the  $q$ -dimensional vector of partial derivatives of order one and by  $g''(x)$  the  $q \times q$  matrix of all second order partial derivatives.

**Proposition B.1.** *Under assumptions (A1)-(A6) for some constants  $\rho_j \geq 0$ , see Appendix A, there exist bounded functions  $r : [0, 1]^{d_1+d} \rightarrow \mathbb{R}$  with*

$$\sup_{u', u \in [0, 1]^{d_1}, x, x' \in [0, 1]^d} |r(u, x) - r(u', x)| \leq C[\|u - u'\| + \|x - x'\|]$$

for a constant  $C$  with the following property:

$$\sup_{u \in [0,1]^{d_1}} |\widehat{m}_1(u) - m_1(u) - \widehat{m}_1^A(u) - \widehat{m}_1^B(u) - \widehat{m}_1^C(u)| \leq R_n(X) + o_P(n^{-1/2}),$$

where  $\widehat{m}_1^A$  and  $\widehat{m}_1^B$  were defined in Theorem 2.1 and where

$$\widehat{m}_1^C(u) = n^{-1} \sum_{i=1}^n r(u, X^i) \varepsilon^i.$$

Furthermore,  $R_n(X)$  is a random variable that only depends on  $X^1, \dots, X^n$  (and, in particular, not on  $\varepsilon^1, \dots, \varepsilon^n$ ) fulfilling

$$R_n(X) = \sum_{j:\rho_j>0} O_P(\|h_j\|^{2+\rho_j}) + \sum_{j:\rho_j=0} o_P(\|h_j\|^2).$$

The constants  $\rho_j$  are the orders of Hölder continuity of the functions  $m_j''$ , see (A6).

Note that in Proposition B.1 an expansion of order  $o_P(n^{-1/2})$  is given. We need this accuracy in the next section for the asymptotic discussion of our test statistic. If one is only interested in the asymptotic distribution of the estimator  $\widehat{m}_1$ , a less accurate expansion suffices. Such an expansion is given in the next theorem, and holds under slightly weaker conditions.

*Proof of Proposition B.1.*

In the proof one can proceed similarly as in the proofs of Theorems 6.1 and 6.2 in Mammen and Park (2005) where a similar statement was proved for the case  $d_1 = \dots = d_r = 1$ . Condition (A5) implies that the kernel density estimator of the joint density of  $(X_j, X_k)$  approximates the true density uniformly with rate  $o_P(n^{-\eta})$  for  $\eta > 0$  small enough.

We have to show that

$$\begin{aligned} n\pi(h_1)^{1/2} \left[ T_n - n^{-1} \pi(h_1)^{-1} K^{(2)}(0)^{d_1} \int_{[0,1]^{d_1}} \sigma^2(x_1) dx_1 \right] &= \int \Delta_1(x_1)^2 p_1(x_1) dx_1 \\ &+ n\pi(h_1)^{1/2} \int \widehat{m}_1^A(x_1)^2 p_1(x_1) dx_1 + o_P(1). \end{aligned}$$

Then the statement of the theorem follows by applying a central limit theorem for the U-statistic  $\int \widehat{m}_1^A(x_1)^2 p_1(x_1) dx_1$  as has been done e.g. in the proof of Theorem 1 in Härdle and Mammen (1993). For the proof of (22) we apply the expansion of  $\widehat{m}_1$  given in Theorem B.1. This can be written as  $\widehat{m}_1 = m_1 + \widehat{m}_1^A + \dots + \widehat{m}_1^E$  with  $\widehat{m}_1^A, \dots, \widehat{m}_1^C$  defined as in Theorem B.1 and with  $\sup_{u \in \mathbb{R}^{d_1}} |\widehat{m}_1^D(u)| = o_P(n^{-1/2})$  and where  $\widehat{m}_1^E$  is a function that only depends on  $X^1, \dots, X^n$  and fulfills  $\sup_{u \in \mathbb{R}^{d_1}} |\widehat{m}_1^G(u)| \leq R_n(X)$ . Note that because of (A8) it holds that  $R_n(X) =$

$o_P(n^{-1/2}\pi(h_1)^{-1/4})$ . Similar expansions can be proved for  $\widehat{m}_{11}, \dots, \widehat{m}_{1s}$ . It holds that  $\widehat{m}_{1j} = m_{1j} + \widehat{m}_{1j}^A + \dots + \widehat{m}_{1j}^E$  with  $\sup_{u \in \mathbb{R}^{d_{1j}}} |\widehat{m}_{1j}^D(u)| = o_P(n^{-1/2})$  and  $\sup_{u \in \mathbb{R}^{d_{1j}}} |\widehat{m}_{1j}^E(u)| = o_P(n^{-1/2}\pi(h_1)^{-1/4})$ . Again  $m_{1j}^E$  only depends on  $X^1, \dots, X^n$ . Further,  $m_{1j}^A, \dots, m_{1j}^C$  are defined as

$$\begin{aligned}\widehat{m}_{1j}^A(u) &= \frac{\sum_{i=1}^n K_{h_{1,j}}(u, X_{1j}^i) \varepsilon^i}{\sum_{i=1}^n K_{h_{1,j}}(u, X_{1j}^i)}, \\ \widehat{m}_{1j}^B(u) &= \begin{cases} \frac{1}{2} h_{1,j}^T m_{1j}''(u) h_{1,j} \int v^2 K(v) dv & \text{if } h_{1,j} \leq u \leq 1 - h_{1,j} \text{ componentwise,} \\ O_P(\|h_{1,j}\|^2) & \text{else} \end{cases} \\ \widehat{m}_1^C(u) &= n^{-1} \sum_{i=1}^n r_j(u, X^i) \varepsilon^i\end{aligned}$$

for a function  $r_j$  with the same smoothness conditions as that stated for  $r$  in the theorem. Note that on the hypothesis (2) it holds that  $\widehat{m}_1^B = \sum_{j=1}^s \widehat{m}_{1j}^B$ . Using this fact we get that

$$\widehat{m}_1 - \sum_{j=1}^s \widehat{m}_{1j} = n^{-1/2} \pi(h_1)^{1/4} \Delta_1 + \widehat{m}_1^A - \sum_{j=1}^s \widehat{m}_{1j}^A + \widehat{m}_1^F + \widehat{m}_1^G,$$

where  $\widehat{m}_1^F$  and  $\widehat{m}_1^G$  fulfill:  $\sup_{u \in \mathbb{R}^{d_1}} |\widehat{m}_1^F(u)| = o_P(n^{-1/2})$ , and where  $\widehat{m}_1^G$  is a function that only depends on  $X^1, \dots, X^n$  and fulfills  $\sup_{u \in \mathbb{R}^{d_1}} |\widehat{m}_1^G(u)| = o_P(n^{-1/2}\pi(h_1)^{-1/4})$ .

Claim (22) now follows from

$$\begin{aligned}n\pi(h_1)^{1/2} \int \left[ \widehat{m}_1^A(x_1) - \sum_{j=1}^s \widehat{m}_{1j}^A(x_{1j}) \right]^2 p_1(x_1) dx_1 &= n\pi(h_1)^{1/2} \int \widehat{m}_1^A(x_1)^2 p_1(x_1) dx_1 \\ &\quad + o_P(1), \\ \int \left[ \widehat{m}_1^A(x_1) - \sum_{j=1}^s \widehat{m}_{1j}^A(x_{1j}) \right] R_n(x_1) p_1(x_1) dx_1 &= o_P(n^{-1/2}),\end{aligned}$$

for random functions  $R_n$  that depend only on  $X^1, \dots, X^n$  and fulfill  $\sup_{u \in \mathbb{R}^{d_1}} |R_n(u)| = o_P(1)$ .

### B.3 Proof of Theorem 3.2

The proof for the third and fourth wild bootstrap method follows by a direct application of a central limit theorem for U-statistics. For the first two procedures the proof is slightly more complicated. Here the bootstrap responses are  $Y^{i,*} = \sum_{l=1}^s \widehat{m}_{1l}(X_{1l}^i) + \sum_{j=2}^r \widehat{m}_j(X_j^i) + \varepsilon^{i,*}$ . With the notations of the last proof this can be written as  $Y^{i,*} = \sum_{l=1}^s m_{1l}(X_{1l}^i) + \sum_{j=2}^r m_j(X_j^i) + \sum_{l=1}^s \widehat{m}_{1l}^A(X_{1l}^i) + \sum_{j=2}^r \widehat{m}_j^A(X_j^i) + \dots + \sum_{l=1}^s \widehat{m}_{1l}^F(X_{1l}^i) + \sum_{j=2}^r \widehat{m}_j^F(X_j^i) + \varepsilon^{i,*}$ . Most of these terms can

be discussed as in the proof of the last theorem. Only the discussion for the terms with superscript A and with superscript B differs. For the A terms one can show that the smooth backfitting fit to the "observations"  $\sum_{l=1}^s \widehat{m}_{1l}^A(X_{1l}^i) + \sum_{j=2}^r \widehat{m}_j^A(X_j^i)$  is equal to  $\sum_{l=1}^s \widehat{m}_{1l}^A(x_{1l}) + \sum_{j=2}^r \widehat{m}_j^A(x_j)$  plus terms of order  $O_P(n^{-1/2})$  that can be treated as the terms with superscript C in the last proof. For the discussion of the B terms one uses the fact that the order of the terms does not increase after projection and the smoothness conditions on the second derivatives that are available for the terms with  $\rho_j > 0$ .

## References

- Carrasco, M., J.P. Florens, and E. Renault, 2006, Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization. In *Handbook of Econometrics*, Ed. J. Heckman und E. Leamer, Vol. 6. North Holland.
- Deaton, A. and J. Muellbauer, 1980, *Economics and Consumer Behavior*. Cambridge University Press, New York.
- Dette, H., C. von Lieres und Wilkau, and S. Sperlich, 2005, A comparison of different nonparametric methods for inference on additive models. *Journal of Nonparametric Statistics*, 17, 57-81.
- Fan, J., Härdle, W. and E. Mammen, 1998, Direct estimation of low dimensional components in additive models. *Ann. Statist.*, 26, 943 - 971.
- Fan, J. and J. Jiang, 2005, Nonparametric inference for additive models. *Journal of the American Statistical Association* 100, 890-907.
- Gozalo, P. and O. Linton, A Nonparametric Test of Additivity in Generalized Nonparametric Regression with estimated parameters. *Journal of Econometrics*, 104, 1-48.
- Haag, B., 2006a, Nonparametric estimation of additive multivariate diffusion processes. *Working paper*, University of Mannheim.
- Haag, B., 2006b, Nonparametric regression tests using dimension reduction techniques. *Working paper*, University of Mannheim.

- Härdle, W., S. Huet, E. Mammen, and S. Sperlich, 2004, Bootstrap Inference in Semiparametric Generalized Additive Models. *Econometric Theory*, 20, 265-300.
- Härdle, W. and E. Mammen, 1993, Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, 21, 1926-1947.
- Härdle, W. and J.S. Marron, 1991, Bootstrap Simultaneous Error Bars for Nonparametric Regression. *Annals of Statistics*, 19, 778-796.
- Härdle, W., S. Sperlich, and V. Spokoiny, 2001, Structural tests in additive regression. *Journal of the American Statistical Association*, 96, 1333-1347.
- Hastie, T.J. and R.J. Tibshirani, 1990, *Generalized Additive Models*. Chapman and Hall, London
- Linton, O.B., 1997, Efficient Estimation of Additive Nonparametric Regression Models. *Biometrika*, 84, 469-473.
- Linton, O., and E. Mammen, 2002, Nonparametric smoothing methods for a class of non-standard curve estimation problems. In: M.G. Akritas and D.N. Politis, (Eds.), *Recent Advances and Trends in Nonparametric Statistics*, Elsevier, Amsterdam.
- Linton, O., and E. Mammen, 2005, Estimating semiparametric ARCH ( $\infty$ ) models by kernel smoothing methods. *Econometrica* 73, 771-836.
- Linton, O., S. Sperlich, S., and I. Van Keilegom, 2008, Estimation of a Semiparametric Transformation Model. *Annals of Statistics*, 36, 686-718.
- Mammen, E., O.B. Linton, and J.P. Nielsen, 1999, The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, 27, 1443-1490.
- Mammen, E., and B. Park, 2005, Bandwidth selection for smooth backfitting in additive models. *Annals of Statistics*, 33, 1260-1294.
- Mammen, E., and K. Yu, 2009, Nonparametric estimation of noisy integral equations of the second kind. (with discussion) *J. Korean Statist. Soc.*, to appear.
- Moral-Arce, I., J.M. Rodriguez-Po, and S. Sperlich, 2007, Feasible Semiparametric Estimation of Censored Expenditure Equations, *Preprints ZfS, Centre for*

*Statistics Gttingen, ZfS-2007-13.*

- Newey, W.K., J.L. Powell, and F. Vella, 1999, Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67, 565-603.
- Nielsen, J.P., and O. Linton, 1998, An optimization interpretation of integration and back-fitting estimators for separable nonparametric models. *Journal of the Royal Statistical Society, B*, 60, 217-222.
- Nielsen, J.P., and S. Sperlich, 2005, Smooth backfitting in practice. *Journal of the Royal Statistical Society, B*, 67, 43-61.
- Romano, J.P., and M. Wolf, 2005a, Exact and approximate step-down methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100, 94-108.
- Romano, J.P., and M. Wolf, 2005b, Stepwise multiple testing as formalized data snooping. *Econometrica*, 73, 1237-1282.
- Romano, J.P., and M. Wolf, 2007, Control of Generalized Error Rates in Multiple Testing. *Annals of Statistics*, 35, 1387-1408.
- Severance-Lossin, E., and S. Sperlich, 1999, Estimation of derivatives for additive separable models. *Statistics*, 33, 241-265.
- Sperlich, S., D. Tjøstheim, and L. Yang, 2002, Nonparametric Estimation and Testing of Interaction in Additive Models. *Econometric Theory*, 18, 197-251.
- 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.
- Yang, L., S. Sperlich, and W. Härdle, 2003, Derivative Estimation and Testing in Generalized Additive Models. *Journal of Statistical Planning and Inference*, 115/2, 521-542.
- Yu, K., B. Park and E. Mammen, 2008, Smooth backfitting in generalized additive models. *Annals of Statistics*, 36, 228-260.