Structural tests in additive regression

Wolfgang Härdle, Stefan Sperlich, and Vladimir Spokoiny*

February 11, 2000

Keywords: Additive models; Component analysis, Haar basis; Hypothesis testing; Nonparametric alternative; Regression

AMS 1996 Classification: 62H25; Secondary 62G10

*Wolfgang Härdle is Professor, Humboldt-Universität zu Berlin, Institut für Statistik und Ökonometrie, Spandauer Str. 1, D - 10178 Berlin. Stefan Sperlich is Assistent Professor, Universidad Carlos III de Madrid, Departamento de Estadística y Econometría, c/ Madrid 126, E - 28903 Getafe - Madrid. Vladimir Spokoiny is Professor, Weierstrass-Institute, Mohrenstr. 39, 10117, Berlin, Germany.
This project was supported by the Deutsche Forschungsgemeinschaft, SFB 373, Berlin, Germany and the Spanish "Dirección General de Enseñanza Superior" (DGES), reference number PB98-0025.

Abstract

We consider the component analysis problem for a regression model with an additive structure. The problem is to test if some of the additive components is of polynomial structure, e.g. linear, without specifying the structure of the remaining components. A particular case is the problem of selecting the significant covariates. The presented method is based on the wavelet transform using the Haar basis, which allows for applications under mild conditions on the design and smoothness of the regression function. The results demonstrate that each component of the model can be tested with the rate corresponding to the case if all the remaining components were known. The proposed procedure is also computationally straightforward. Simulation results and a real data example about female labor supply demonstrate the good performance of the test.

1 Introduction

In *multivariate regression problems* we study the structural relationship between the response variable Y and the vector of covariates $X = (X_1, \dots, X_d)^\top$ via the regression curve

$$F(x) = E(Y|X = x)$$

with $x = (x_1, \dots, x_d)^\top$. Purely nonparametric models do not make any assumption about the form of the d-variate function $F(x)$. The problem is then to fit a d-dimensional surface to the observed data $\{(X_i, Y_i) : i = 1, \dots, n\}$. The obvious approach is to generalize the univariate smoothing techniques based on local 'averaging' to this multivariate situation. A serious problem arising here is that we need much more data material in higher dimensions in order to have enough data points in a local neighborhood of each point. Several approaches for dimensionality reduction have been proposed to deal with this so-called *curse of dimensionality.* A promising one is *additive modeling* as in economic theory it is a favorite structure anyway, see e.g. Deaton and Muellbauer (1980).

Such a nonparametric additive regression model has the form

$$y = F(x) + \xi, \qquad x = (x_1, \dots, x_d) \in \mathbb{R}^d, \tag{1.1}$$
$$F(x) = f_1(x_1) + \dots + f_d(x_d), \tag{1.2}$$

where y is a scalar variable, $\{f_m\}_{m=1}^d$ is a set of unknown component functions and ξ is a random error.

This class of models has been shown to be useful in statistical practice: it generalizes linear regression in a natural way and allows interpretation of marginal changes i.e. the effect of one variable on the mean function F holding all else constant. Additive models were considered first by Leontief (1947) for input-output analysis speaking of *separable* models. In the statistical literature the nonparametric additive regression has been introduced in the eighties, see Buja, Hastie and Tibshirani (1989) for a survey. An advantage of additive models is that they combine flexible modelling of many variables with statistical precision that is typical for just one explanatory variable, see Stone (1985, 1986). Algorithmic aspects of additive modelling by backfitting are discussed in Hastie and Tibshirani (1990) or Venables and Ripley (1994). Tjøstheim, Auestad (1994) and Linton, Nielsen (1995) proposed a method of analysis based on marginal integration.

An essential advantage in additive models is that they allow component wise inferences. Important problems of component analysis in economics are the question of significance as well as of linearity, since nonlinearities often rise serious problems e.g. of identification in equation or economic equilibrium systems. In nonparametrics, among others Hastie and Tibshirani (1990) or Härdle and Korostelev (1996) considered also the problem of selection of significant covariates. In this paper we focus on the much more general problem of testing for component f_m the hypothesis of being of polynomial form, e.g. constant or linear.

Theory for nonparametric hypothesis testing is well developed. So the problem of testing a simple null hypothesis versus a univariate nonparametric alternative is studied in detail, see e.g. Ingster (1993), Härdle and Mammen (1993), Hart (1997), Stute (1997) for historical background and further references. Many tests have been shown to be sensitive against every directional local alternative, e.g. Bierens (1982), Eubank and Hart (1992), Stute (1997) and references therein. Unfortunately, the power of every particular test cannot be uniform w.r.t. the "direction" in the function space as shown in Burnashev (1979) or Ingster (1982). This particularly means that the finite sample power of every test could be better for some local alternatives and worse for the others. The same arguments apply to the so called "intermediate" efficiency approach of Inglot and Ledwina (1996).

Ingster (1982, 1993) has shown that a test could be uniformly consistent against a smooth alternative only if this alternative deviates from the null with the distance of order $n^{-2s/(4s+1)}$ with s being the degree of smoothness. The structure of proposed rate-optimal tests also essentially rely on the smoothness properties of the underlying function though such kind of prior information about the underlying function is typically lacking in practical applications. Spokoiny (1996) offered an adaptive data-driven testing procedure which does not require knowledge of smoothness properties of the tested function and allow for a near optimal testing rate up to a log log n factor. The latter can be viewed as the price for adaptation. Horowitz and Spokoiny (1999) proposed a similar test based on kernel smoothers with different bandwidths and shown that it is simultaneously consistent against any local "directional" alternative which approaches the null hypothesis at the rate $(n/\log\log n)^{-1/2}$.

It is worth noting that the adaptive testing procedure from Spokoiny (1996) is essentially a theoretical device rather than a practically applicable method since it is developed for the idealized "signal + white noise" model, simple null, known noise variance etc. Practically relevant procedures should address numerous issues raising in particular applications. In the context of multidimensional additive modelling, an additional challenge comes from the fact that the considered component f_1, even being completely specified, does not specify the whole model since nothing is assumed about the other components, $f_2, \ldots, f_d$ which can be viewed as an infinite-dimensional nuisance parameter. This particularly creates a serious problem with evaluating the critical value of the proposed test statistics which provides the prescribed type I level.

Therefore, the task is to develop a procedure which, independently on the functional form of the 'nuisance' components $f_2, \ldots, f_d$, leads to the given type I error α if f_1 is linear and is sensitive against a smooth alternative with unknown degree of smoothness. In view of practical applications (see Section 4) we proceed with a deterministic non-regular design allowing discrete components and with unknown noise variance.

In this paper we apply a Haar decomposition which is a particular and non-regular case of the wavelet transform. Nevertheless, for the hypothesis testing framework the application of the Haar basis leads not only to the desired optimal testing rate but also provides a test which is more stable w.r.t. the design non regularity. This is important for practical applications allowing to relax and to simplify the conditions on the design, reduces computational burdens and more!

Our approach is based on the simultaneous approximation of all components $f_1, \ldots, f_d$ by Haar sums: we first estimate the Haar coefficients for all components and then analyze the coefficients corresponding to the first one. The testing problem is formulated in the next section, the procedure is described in Section 2. The asymptotic properties are discussed in Section 3. We compare the sensitivity of our procedure with the ideal one designed for the case as if the other components and all smoothness properties were known. The results demonstrate asymptotic optimality of the proposed procedure and they are stated under mild conditions on the design. The performance of the procedure for the finite sample size case is examined by simulations and an application in Section 4. Extensions to more general problems including model check of additivity and multiple testing of several components simultaneously are shortly discussed in Section 3.1 and the proofs are postponed to Section 5.

1.1 Model and testing problem

We are given data $(X_i, Y_i), i = 1, \ldots, n$, with $X_i \in I\!R^d$, $Y_i \in I\!R^1$, obeying the regression equation

$$Y_i = F(X_i) + \xi_i \tag{1.3}$$

where F is an unknown regression function with the additive structure

$$F(x) = f_1(x_1) + \ldots + f_d(x_d), \tag{1.4}$$

and ξ_i are normal random errors with zero mean and known variance σ^2. We allow for deterministic non-regular design $X_1, \ldots, X_n$ with possible replications. It is only supposed that the design is rescaled to the unit cube $[0,1]^d$, that is, $X_{i,\ell} \in [0,1]$ for all $i \le n$ and $\ell \le d$.

Our aim is to analyze each component f_m, $m = 1, \ldots, d$. For simplicity we present the procedure focusing on the first component f_1, and on the problem of testing linearity, i.e. the hypothesis $H_0 : f_1(t) = a_1 + b_1 t$ for some constants a_1, b_1.

Let ϕ be a test, a measurable function of observations with values 0 (accept) and 1 (reject). Denote by $\boldsymbol{P}_F$ the distribution of the data $Y_1, \ldots, Y_n$ for a fixed model function F, see (1.3) and (1.4). Let now F_0 be a function with a linear first component. The type I error probability is the probability under F_0 to reject the hypothesis: $\alpha_{F_0}(\phi) = \boldsymbol{P}_{F_0}(\phi = 1)$. Similarly one defines the error probability $\beta_F(\phi)$ of the second type. If the first component f_1 is not linear, then $\beta_F(\phi) = \boldsymbol{P}_F(\phi = 0)$. Given $\alpha > 0$, we wish to construct such a test ϕ that $\alpha_{F_0}(\phi) \le \alpha$ for all F_0 with a linear first component and, in addition, it is sensitive against a large class of alternatives F.

2 Testing procedure

In order to illustrate the main ideas, we begin with the univariate case i.e. $d = 1$.

2.1 The case of $d = 1$

Consider the univariate regression model

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{2.5}$$

which corresponds to (1.3) with $d = 1$. We write here f instead of f_1 to minimize the notation. The problem consists in testing the hypothesis that the function f is linear.

Eubank and Hart (1992) nicely pointed out a common feature of the most of procedures for model checking. Let $\mathcal{F}_0$ be the set of regression functions considered under the null hypothesis (here the linear functions). Then f is written as $f(x,\theta_0)+\sum_i \theta_j\psi_j(x)$ with $f(x,\theta_0)$ a member of $\mathcal{F}_0$ and $\{\psi_j\}$ an orthonormal system. The testing problem reduces now to testing $\theta_j=0$ for all j, cf. also Stute (1997).

The procedure proposed here follows this idea and relies on a special piecewise constant approximation (the Haar decomposition) of the function f.

Denote by I the multi-index $I=(j,k)$ with $j=1,2,\ldots$ and $k=0,1,\ldots,2^j-1$, and by $\mathcal{I}$, the set of all such multi-indices. Let now the function $\psi(t)$ (the mother wavelet) be defined by

$$\psi(t)=\begin{cases}0, & t<0,\, t\geq 1,\\ 1, & 0\leq t<1/2,\\ -1, & 1/2\leq t<1.\end{cases}$$

For every $I=(j,k)$ with $j\geq 0$ and $k=0,\ldots,2^j-1$ set

$$h_I(t)=\psi(2^j t-k).$$

Clearly the function ψ_I with $I=(j,k)$ is supported on the interval $A_I=[2^{-j}k,2^{-j}(k+1)]$. Denote also by $\mathcal{I}_j$ the index subset corresponding to the j-th resolution level:

$$\mathcal{I}_j=\{I=(j,k),\, k=0,1,\ldots,2^j-1\}\qquad j\geq 0.$$

The idea of the test is to estimate from the data the coefficients c_I of the approximation of the unknown regression function f by the sum

$$c_0+c_1x+\sum_{\ell=0}^{j}\sum_{I\in\mathcal{I}_\ell}c_I h_I(x)$$

and then to check whether some of estimated coefficients c_I differ significantly from zero.

For a formal description, define with $I=(j,k)\in\mathcal{I}$

$$\begin{aligned}\mu_I^2 &= \sum_{i=1}^{n}h_I^2(X_i),\\ \psi_I(X_i) &= \mu_I^{-1}h_I(X_i).\end{aligned}$$

Clearly μ_I^2 is the number of design points in A_I, that is, $\mu_I^2=\#\{i: X_i\in A_I\}$, $I\in\mathcal{I}$.

We also define two functions $\psi_0\equiv\mu_0^{-1}$ and $\psi_1(t)=\mu_1^{-1}t$ with $\mu_0^2=n$ and $\mu_1^2=\sum_{i=1}^n X_i^2$ and introduce the index set

$$\mathcal{I}(j)=\{0,1\}+\bigcup_{\ell=0}^{j}\mathcal{I}_\ell. \tag{2.6}$$

By $N(j)$ we denote the number of indices in $\mathcal{I}(j)$. Obviously $N(j) = 2^{j+1} + 1$. Let $\boldsymbol{\theta}(j)$ denote a vector in $I\!R^{N(j)}$ with entries θ_I, $I \in \mathcal{I}(j)$. Define the vector $\widehat{\boldsymbol{\theta}}(j)$ as solution to the quadratic problem

$$\widehat{\boldsymbol{\theta}}(j) = \operatorname*{arginf}_{\boldsymbol{\theta}(j) \in I\!R^{N(j)}} \sum_{i=1}^{n} \left(Y_i - \sum_{I \in \mathcal{I}(j)} \theta_I \psi_I(X_i) \right)^2$$

To get an explicit expression for $\widehat{\boldsymbol{\theta}}(j)$ we introduce vector notation. Let g be a function observed at point $X_1, \dots, X_n$. We identify every such function with the column-vector $\boldsymbol{g}$ in $I\!R^n$ with the entries $g(X_i)$ and define $\|\boldsymbol{g}\|_n$ by $\|\boldsymbol{g}\|_n^2 = \sum_{i=1}^n g^2(X_i)$. Let also $\boldsymbol{Y}$ stand for the column vector $(Y_1, \dots, Y_n)^\top$. Introduce a $n \times N(j)$-matrix $\Psi(j)$ with entries $\psi_I(X_I)$:

$$\Psi(j) = \Big(\psi_I(X_i),\ i = 1, \dots, n,\ I \in \mathcal{I}(j) \Big).$$

Then

$$\begin{aligned} \widehat{\boldsymbol{\theta}}(j) &= \operatorname*{arginf}_{\boldsymbol{\theta}(j) \in I\!R^{N(j)}} \|\boldsymbol{Y} - \Psi(j)\boldsymbol{\theta}(j)\|_n^2 \\ &= V(j)\Psi(j)^\top \boldsymbol{Y} = W(j)^\top \boldsymbol{Y}, \end{aligned}$$

where $V(j)$ is the pseudo-inverse of $\Psi(j)^\top \Psi(j)$, that is, $V(j) = \left\{ \Psi(j)^\top \Psi(j) \right\}^-$ and $W(j) = \Psi(j)V(j)$ is a $n \times N(j)$-matrix.

Since the errors ξ_i are normal $\mathcal{N}(0, \sigma^2)$, we obtain by (2.5) that $\widehat{\boldsymbol{\theta}}(j)$ is a Gaussian vector with the mean $\boldsymbol{\theta}^*(j) = W(j)^\top \boldsymbol{f} = V(j)\Psi(j)^\top \boldsymbol{f}$ and the covariance matrix $\sigma^2 V(j)$,

$$\widehat{\boldsymbol{\theta}}(j) \sim \mathcal{N}\left\{ \boldsymbol{\theta}^*(j), \sigma^2 V(j) \right\}.$$

The entries of the matrix $V(j)$ (resp. $W(j)$) will be denoted by $v_{I,I'}$ (resp. $w_{i,I}$) where $I, I' \in \mathcal{I}$ and $i = 1, \dots, n$. All these values depend on j, but do not indicate this dependence explicitly to simplify the notation.

By $\widehat{\boldsymbol{\theta}}_j$ we denote the part of the vector $\widehat{\boldsymbol{\theta}}(j)$ corresponding to j-th resolution level: $\widehat{\boldsymbol{\theta}}_j = (\widehat{\theta}_I, I \in \mathcal{I}_j)^\top$, so that $\widehat{\boldsymbol{\theta}}_j \in R^{2^j}$. Obviously $\widehat{\boldsymbol{\theta}}_j = W_j^\top \boldsymbol{Y}$ where W_j is the $n \times 2^j$-submatrix of $W(j)$ corresponding to the index set $\mathcal{I}_j$: $W_j = (w_{i,I}, i = 1, \dots, n,\ I \in \mathcal{I}_j)$. Similarly we define the 2^j-vector $\boldsymbol{\theta}_j^*$ and $2^j \times 2^j$-submatrix V_j of $V(j)$:

$$\boldsymbol{\theta}_j^* = (\theta_I^*, I \in \mathcal{I}_j), \qquad V_j = (v_{I,I'}, I, I' \in \mathcal{I}_j).$$

Clearly $\widehat{\boldsymbol{\theta}}_j \sim \mathcal{N}(\boldsymbol{\theta}_j^*, \sigma^2 V_j)$ and $V_j = W_j^\top W_j$.

2.2 Level test statistic for $d = 1$

The proposed testing procedure is based on the fact that for f linear, all the empirical coefficients $\widehat{\theta}_I$, $I \neq 0, 1$, are zero mean Gaussian r.v.'s. We build for every j one test statistic corresponding to the hypothesis $\boldsymbol{\theta}_j^* = 0$.

By definition $\widehat{\boldsymbol{\theta}}_j = W_j^\top \boldsymbol{Y}$ which yields $\widehat{\boldsymbol{\theta}}_j \sim \mathcal{N}(\boldsymbol{\theta}_j^*, \sigma^2 V_j)$ with $V_j = W_j^\top W_j$. This naturally leads to the likelihood-based statistic $S_j = \widehat{\boldsymbol{\theta}}_j^\top V_j^- \widehat{\boldsymbol{\theta}}_j$ where V_j^- means the pseudo-inverse of V_j. Under the null hypothesis (that is, for a linear function f), it clearly holds $\boldsymbol{\theta}_j^* = 0$ and $\widehat{\boldsymbol{\theta}}_j = W_j^\top \boldsymbol{\xi}$, and hence,

$$S_j = \boldsymbol{\xi}^\top W_j V_j^- W_j^\top \boldsymbol{\xi} = \boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{\xi} \tag{2.7}$$

where $\mathcal{R}_j = W_j V_j^- W_j^\top = W_j \left(W_j^\top W_j\right)^- W_j^\top$ is a projector in the space $I\!R^n$ (that is, $\mathcal{R}_j^2 = \mathcal{R}_j$). By N_j we denote the rank of $\mathcal{R}_j$. By definition $N_j \le 2^j$. The definition (2.7) particularly yields that $\sigma^{-2} S_j$ follows the χ^2-distribution with N_j degree of freedom.

The level test statistic T_j is defined via centering and standardization of S_j. The following simple properties are useful here:

$$\boldsymbol{E} S_j = \boldsymbol{E} \boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{\xi} = \sigma^2 \operatorname{tr} \mathcal{R}_j = \sigma^2 N_j, \qquad \boldsymbol{E}\left(S_j - \sigma^2 N_j\right)^2 = 2\sigma^4 N_j.$$

Since the noise variance σ^2 is usually unknown, we replace it by a pilot estimate $\widehat{\sigma}^2$, see Section 2.5 below. This leads to the test statistic T_j of the form:

$$T_j = \frac{\widehat{\boldsymbol{\theta}}_j^\top V_j^- \widehat{\boldsymbol{\theta}}_j - \widehat{\sigma}^2 N_j}{\widehat{\sigma}^2 \sqrt{2N_j}} \tag{2.8}$$

An important feature of this statistic is that under the null hypothesis, it has a nondegenerated distribution (which approaches the standard normal law as N_j grows). Moreover, this distribution is known (see Section 2.6 for a closed form expression) which allows to precisely evaluate the corresponding $(1-\alpha)$-quantile $t_{j,\alpha}$ defined by

$$\boldsymbol{P}_0\left(T_j > t_{j,\alpha}\right) = \alpha \tag{2.9}$$

where $\boldsymbol{P}_0$ means the distribution of T_j under the null hypothesis.

2.3 A multiscale test for $d = 1$

The proposed test analyzes all statistics T_j for different j simultaneously. Similar ideas are intensively discussed in the literature. Eubank and Hart (1992) proposed the so called "order selection" test using a modified Mallows' criterion (Mallows, 1973) for selecting the number of considered terms of an orthogonal series expansion for the deviation of the underlying function f from the null hypothesis, see also Aerts, Claeskens and Hart (1999). This method leads to the maximum of $\sigma^{-2}S_j - (1+C_n)N_j$ where $(1+C_n)N_j$ is the penalty term for going to a more complicated model. A similar test, called the data-driven Neyman's smooth test is proposed in Ledwina (1994) and Kallenberg and Ledwina (1995). Fan (1996), Spokoiny (1996) and Fan and Huang (1998) considered the tests based on the maximum of centered and standardized sums like T_j. Such a test has strong appeal: the results from Ingster (1982, 1993) show that the test T_j with a special choice of the index j is rate optimal against a smooth alternative with a smoothness degree s. The test based on the maximum of T_j is adaptive in the sense that it is near optimal against a smooth alternative with unknown degree of smoothness.

Here we apply the method based on the multiscaling idea (see Section 2.3 for more discussion) which is close to the proposal from Fan (1996) and Spokoiny (1996): the test statistic T^* is defined as the maximum of $T_j - t_{j,\alpha}$ over all considered levels j with $t_{j,\alpha}$ from (2.9). Namely, we consider all j from $j=0$ until the finest resolution level j_n defined as $j_n = [\log_2(n/3)]$ where $[a]$ means the integer part of a. We now set

$$T^* = \max_{0 \le j \le j_n} (T_j - t_{j,\alpha}).$$

A choice of the critical value for this test is discussed in Section 2.6.

2.4 A multiscale test for $d > 1$

The basic idea of testing is similar to the univariate case and it is based on the approximation of each component f_m from (1.3) by the sum

$$c_{1,m}x_m + \sum_{j=0}^{j_n} \sum_{I \in \mathcal{I}_j} c_{I,m} h_I(x_m), \qquad m = 1, \dots, d.$$

(We skip here the constant term to provide identifiability of each component.) Let us fix a level j for the first component and a level j_n for the remaining ones, and let $\mathcal{I}(j)$ be due to (2.6), $\mathcal{I}(j) = \{0,1\} + \bigcup_{0 \le \ell \le j} \mathcal{I}_\ell$. We also define $\mathcal{I}'(j) = \{1\} + \bigcup_{0 \le \ell \le j} \mathcal{I}_\ell$. To define the level test, we approximate $F(x)$ by

$$\sum_{I \in \mathcal{I}(j)} c_{I,1} h_I(x_1) + \sum_{m=2}^{d} \sum_{I \in \mathcal{I}'(j_n)} c_{I,m} h_I(x_m).$$

Here $N = 2^{j_n+1}$ coefficients are used for each component f_m, $m \ge 2$, and, assuming that $j \le j_n$, the total number of coefficients is at most $Nd+1$. We modify now the definition of j_n from the one-dimensional case to provide $Nd + 1 \le 2n/3$ that leads to the choice

$$j_n = \left[\log_2\left(\frac{n}{3d}\right)\right]. \tag{2.10}$$

To define the test, we first standardize each basis function:

$$\psi_{I,m}(t) = \mu_{I,m}^{-1} h_I(t) \text{ with } \mu_{I,m}^2 = \sum_{i=1}^{n} h_I^2(X_{i,m})$$

Here $(X_{i,1}, \ldots, X_{i,d})$ is the coordinate representation of X_i.

Let now some $j \le j_n$ be fixed. Denote by $\mathcal{I}(d,j)$ the index set

$$\mathcal{I}(d,j) = \Big\{(I,1),\, I \in \mathcal{I}(j)\Big\} \times \prod_{m=2}^{d} \Big\{(I,m),\, I \in \mathcal{I}'(j_n)\Big\}$$

and let

$$N(d,j) = N(j) + (d-1)N = 2^{j+1} + (d-1)2^{j_n+1} + 1$$

be the number of elements in $\mathcal{I}(d,j)$.

Set $\Psi(d,j)$ for the $n \times N(d,j)$ matrix with entries $\psi_{I,m}(X_i) = \mu_{I,m}^{-1} h_I(X_{i,m})$, $i = 1, \ldots, n$, $(I,m) \in \mathcal{I}(d,j)$, and define the vector $\widehat{\boldsymbol{\theta}}(d,j)$ in $I\!R^{N(d,j)}$ as a solution to the quadratic problem:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}(d,j) &= \operatorname*{arginf}_{\boldsymbol{\theta}(d,j) \in I\!R^{N(d,j)}} \|\boldsymbol{Y} - \Psi(d,j)\boldsymbol{\theta}(d,j)\|_n^2 \\
&= \operatorname*{arginf}_{\boldsymbol{\theta}(d,j) \in I\!R^{N(d,j)}} \sum_{i=1}^{n} \Bigg(Y_i - \sum_{I \in \mathcal{I}(j)} \theta_{I,1}\psi_{I,1}(X_{i,1}) - \sum_{m=2}^{d} \sum_{I \in \mathcal{I}'(j_n)} \theta_{I,m}\psi_{I,m}(X_{i,m}) \Bigg)^2 .
\end{aligned}$$

As in the univariate case, we derive

$$\widehat{\boldsymbol{\theta}}(d,j) = V(d,j)\Psi(d,j)^\top \boldsymbol{Y} = W(d,j)^\top \boldsymbol{Y} \tag{2.11}$$

where the matrix $V(d,j)$ is the pseudo-inverse of $\Psi(d,j)^\top\Psi(d,j)$, i.e. $V(d,j) = \big\{\Psi(d,j)^\top\Psi(d,j)\big\}^-$ and $W(d,j) = \Psi(d,j)V(d,j)$. The entries of the matrix $V(d,j)$ (resp. $W(d,j)$) will be denoted by $v_{(I,m),(I',m')}$ (resp. $w_{i,(I,m)}$).

Similarly to the univariate case, we define the level test making use of the subvector $\widehat{\boldsymbol{\theta}}_j = (\widehat{\theta}_{I,1},\, I \in \mathcal{I}_j)$ and the submatrix $V_j = (v_{(I,1),(I',1)},\, I, I' \in \mathcal{I}_j)$ of the covariance matrix $V(d,j)$. Let W_j again denote the submatrix of $W(d,j)$ corresponding the level j of the first component: $W_j = (w_{i,(I,1)}, i = 1, \ldots, n,\, I \in \mathcal{I}_j)$. Then clearly $\widehat{\boldsymbol{\theta}}_j = W_j^\top \boldsymbol{Y}$ and $V_j = W_j^\top W_j$. The test statistic T_j is defined as follows, cf. (2.8):

$$T_j = \frac{\widehat{\boldsymbol{\theta}}_j^\top V_j^- \widehat{\boldsymbol{\theta}}_j - \widehat{\sigma}^2 N_j}{\widehat{\sigma}^2\sqrt{2N_j}} = \frac{\boldsymbol{Y}^\top \mathcal{R}_j \boldsymbol{Y} - \widehat{\sigma}^2 N_j}{\widehat{\sigma}^2\sqrt{2N_j}},$$

where $\mathcal{R}_j = W_j V_j^- W_j^\top$ and N_j is the rank of $\mathcal{R}_j$ (or equivalently of V_j), $N_j \le 2^j$. With $t_{j,\alpha}$ fulfilling (2.9), the final test statistic is again of the form:

$$T^* = \max_{0 \le j \le j_n} (T_j - t_{j,\alpha}). \tag{2.12}$$

Remark 1 In some practical applications, see e.g. our example in Section 4, one or more explanatory variables X_m can be discrete with only few possible values, say two or three. In that case the corresponding component function f_m is completely determined by its values at these points and it can be precisely expanded by a finite Haar sum with very few Haar levels. Of course, for such situation it is not reasonable to consider all j_n Haar levels for those components and the required number of levels for every particular component f_m should be determined by the identifiability reasons, see Section 4 for an example.

2.5 Estimation of the noise variance

Here we indicate how the noise variance σ^2 can be estimated from the data. One may apply two different approaches for variance estimation. One way is based on residuals from locally polynomial fitting, see e.g. Rice (1984) or Gasser et al (1986) for the univariate case or Hall et al. (1991) and Spokoiny (1999b) and references therein for a detailed discussion of the multivariate case. Another approach is to retrieve the residuals from the same orthogonal series expansion which is used for model checking. Here we follow the last proposal.

Let j_n be defined in (2.10). Due to this definition we have $n/3 \le d2^{j_n+1} \le 2n/3$.

Let $\widehat{\boldsymbol{\theta}}(d, j_n)$ be the least square estimator from (2.11) with $j = j_n$, that is, the maximal number of Haar coefficients are used for all components f_m. This vector is Gaussian with the mean $\boldsymbol{\theta}^*(d, j_n) = W(d, j_n)\boldsymbol{F}$ and the covariance matrix $\sigma^2 V(d, j_n)$. Moreover, $\Psi(d, j_n)\widehat{\boldsymbol{\theta}}(d, j_n) = \Pi_n Y$ where

$$\Pi_n = \Psi(d, j_n)\Big(\Psi(d, j_n)^\top \Psi(d, j_n)\Big)^{-} \Psi^T(d, j_n)$$

is the projector in $\mathbb{R}^n$ on the subspace generated by additive functions of the form

$$\theta_0 + \sum_{m=1}^{d} \sum_{I \in \mathcal{I}'(j_n)} \theta_{I,m}\psi_{I,m}(x_m).$$

One can easily check that

$$\begin{aligned}
\boldsymbol{E}\|\boldsymbol{Y} - \Psi(d, j_n)\widehat{\boldsymbol{\theta}}(d, j_n)\|_n^2 &= \|\boldsymbol{F} - \Pi_n \boldsymbol{F}\|_n^2 + \boldsymbol{E}\|\boldsymbol{\xi} - \Pi_n \boldsymbol{\xi}\|_2^2 \\
&= \|\boldsymbol{F} - \Pi_n \boldsymbol{F}\|_n^2 + \sigma^2 \operatorname{tr}(I_n - \Pi_n) \\
&= \|\boldsymbol{F} - \Pi_n \boldsymbol{F}\|_n^2 + \sigma^2 (n - r_n)
\end{aligned}$$

where I_n denotes the identity $n \times n$-matrix and r_n is the rank of Π_n. By definition $r_n \le 2n/3$.

Under usual regularity condition on the function F, see e.g. Lemma 1 in the next section, the accuracy of approximating F by such expansion tends to zero as n tends to infinity in the sense that

$$n^{-1}\|\boldsymbol{F}-\Pi_n\boldsymbol{F}\|_n^2 \to 0, \qquad n\to\infty.$$

This consideration prompts to use the value

$$\widehat{\sigma}^2 = \frac{1}{n-r_n}\|\boldsymbol{Y}-\Psi(d,j_n)\widehat{\boldsymbol{\theta}}(d,j_n)\|_n^2$$

for estimating σ^2. It is important to mention that if $F\equiv 0$, then $(n-r_n)\widehat{\sigma}^2 = \|\boldsymbol{\xi}-\Pi_n\boldsymbol{\xi}\|_n^2$ follows the χ^2-distribution with $n-r_n$ degree of freedom and $\widehat{\sigma}^2$ and $\widehat{\boldsymbol{\theta}}(d,j_n)$ are independent.

2.6 Critical level of the test

First we again discuss the univariate situation with $d=1$. In that case the function F coincides with the first component f_1 and its structure is known under the null hypothesis. Moreover, in view of the method of approximation, the linear trend in f_1 has no influence on the remaining coefficients and we may suppose that the function f_1 is exactly zero. The same applies for the variance estimate $\widehat{\sigma}^2$. This reduces the linear hypothesis to the case of a simple null hypothesis $f_1\equiv 0$, that is, the observations Y_i coincide with the noise ξ_i. In this situation one has $S_j=\boldsymbol{\xi}^\top\mathcal{R}_j\boldsymbol{\xi}$, $\widehat{\sigma}^2=(n-r_n)^{-1}\|(\mathbf{1}-\Pi_n)\boldsymbol{\xi}\|_n^2=(n-r_n)^{-1}\boldsymbol{\xi}(\mathbf{1}-\Pi_n)\boldsymbol{\xi}$ where $\mathcal{R}_j=W_jV_j^-W_j^\top$ and $\mathbf{1}$ denotes the unit operator in $I\!R^n$ and the test statistics T_j can be represented in the form

$$T_j \;=\; \frac{\boldsymbol{\xi}^\top\mathcal{R}_j\boldsymbol{\xi}-\widehat{\sigma}^2N_j}{\widehat{\sigma}^2\sqrt{2N_j}} = \frac{\boldsymbol{\xi}^\top\mathcal{R}_j\boldsymbol{\xi}(n-r_n)}{\boldsymbol{\xi}(\mathbf{1}-\Pi_n)\boldsymbol{\xi}\sqrt{2N_j}} - \sqrt{N_j/2}\,. \tag{2.13}$$

Therefore, each T_j is the ratio of two quadratic forms of ξ_i's and as a consequence, it does not depend on the noise variance and its distribution can be precisely described via the Fisher distribution $F_{N_j,n-r_n}$ with N_j and $n-r_n$ degree of freedom. The values $t_{j,\alpha}$ defined in (2.9) can therefore be calculated using the proper quantile of this Fisher distribution and they depend only on N_j, $n-r_n$ and α. Since all the T_j's are constructed on the base of the same data, they are dependent in a rather complicated way and hence, the closed form expression for the distribution of the maximum $T^*=\max_{j\le j_n}(T_j-t_{j,\alpha})$ is difficult to obtain and some Monte-Carlo experiments can be used for finding a proper quantile λ satisfying $\boldsymbol{P}_0\,(T^*>\lambda)=\alpha$ where $\boldsymbol{P}_0$ means that each T_j follows (2.13) with a standard Gaussian vector $\boldsymbol{\xi}$. Having done this, we define the test ϕ^* as

$$\phi^* = \mathbf{1}(T^*>\lambda). \tag{2.14}$$

For the general multivariate case one can show that under some regularity conditions (see Condition (D) in the next section) the influence of the remaining components $f_2, \dots, f_m$ on the test statistic T^* is asymptotically negligible and we therefore determine the critical value λ in the same way using simulated data from the no response d-variate model with standard Gaussian errors. For further discussion and details respective this Monte-Carlo method, see Section 4 or Spokoiny (1999a).

Remark 2 Note that the adjustment of T_j by $t_{j,\alpha}$ is more of theoretical nature allowing for the unified exposition. Indeed, all the test statistics T_j have non-degenerate distribution with the variance 1 and moreover, for large j, this distribution is close to the standard normal CDF. This means that all the $t_{j,\alpha}$'s are of the same order and the effect of this adjustment is inessential. This issue is confirmed by our simulation results, see Section 4.

3 Main results

In this section we present asymptotic properties of the proposed testing procedure. We state the results on the type one and type two error probabilities separately since we evaluate them under different assumptions on the design variables. The result on the type I error probabilities $\alpha_{F_0}(\phi^*)$ is valid under mild assumptions on the design. But for high sensitivity of the test, we need slightly stronger regularity conditions on the design variables.

When testing the first component of the function F from (1.4), the remaining components $f_2, \dots, f_d$ can be viewed as a nonparametrically specified nuisance parameter which are to be estimated by a pilot estimator. In order to ensure the required accuracy of estimation, we need some conditions on the rate of approximation of each function f_m with $2 \le m \le d$ by the Haar series. We formulate these conditions exactly in the required form. Later we show that these conditions are met, for instance, under mild conditions on smoothness of f_m and on the design $X_1, \dots, X_n$.

Recall that we identify every function g on $I\!R^d$ with the vector $\boldsymbol{g} = (g(X_i), i = 1, \dots, n)^\top$ in $I\!R^n$. In particular, each component f_m is identified with the vector $\boldsymbol{f}_m = (f_m(X_{i,m}), i = 1, \dots, n)^\top$ and $\boldsymbol{\psi}_{I,m}$ is understood as the vector with the elements $\psi_{I,m}(X_{i,m}) = \lambda_{I,m}^{-1} h_I(X_{i,m})$. Recall also the notation $\|\boldsymbol{g}\|_n^2 = \sum_{i=1}^n g^2(X_i)$.

Denote by $\mathcal{L}_m(j)$ the linear subspace in $I\!R^n$ generated by the functions (vectors) $\{\boldsymbol{\psi}_{I,m}\}$, $I \in \mathcal{I}_\ell, 0 \le \ell \le j$,

$$\mathcal{L}_m(j) = \left\{ \theta_{0,m} + \theta_{1,m}\boldsymbol{\psi}_{1,m} + \sum_{\ell=0}^{j} \sum_{I \in \mathcal{I}_\ell} \theta_{I,m}\boldsymbol{\psi}_{I,m} \right\}.$$

Clearly all the functions (or vectors) from $\mathcal{L}_m(j)$ depend only on m-th coordinates $X_{i,m}$ of design points X_i, $i = 1, \ldots, n$. By $\Pi_{m,n}\boldsymbol{f}_m$ we denote the projection of $\boldsymbol{f}_m$ onto $\mathcal{L}_m(j_n)$ w.r.t. the distance $\|\cdot\|_n$,

$$\Pi_{m,n}\boldsymbol{f}_m = \operatorname*{arginf}_{\boldsymbol{g}\in\mathcal{L}_m(j_n)} \|\boldsymbol{f}_m - \boldsymbol{g}\|_n = \operatorname*{arginf}_{\boldsymbol{g}\in\mathcal{L}_m(j_n)} \sum_{i=1}^{n} |f_m(X_{i,m}) - g(X_{i,m})|^2.$$

In our results we impose the following condition:

Condition (D) *For some fixed constant C and n large*

$$\sum_{m=1}^{d} \|\boldsymbol{f}_m - \Pi_{m,n}\boldsymbol{f}_m\|_n \le C\sigma n^{-1/2}.$$

The following lemma shows that condition (D) is satisfied under mild smoothness conditions on each component f_m.

Lemma 1 *Let $\mu_{n,m}$ be the m-th marginal of the empirical design measure μ_n,*

$$\mu_{n,m}(A) = n^{-1}\sum_{i=1}^{n} \mathbf{1}(X_{i,m} \in A), \qquad m = 1, \ldots, d.$$

Let further C_1 be a constant such that for every $0 \le a < b \le 1$ with $b - a > 1/n$, it holds

$$\mu_{n,m}[a,b] \le C_1(b-a).$$

If each f_m, $m = 2, \ldots, d$, is a Lipschitz function i.e.

$$|f_m(x) - f_m(x')| \le C_2|x - x'|, \qquad \forall x, x' \in [0,1],$$

then condition (D) is fulfilled with C depending on σ, C_1 and C_2 only.

Another situation in which the difference $\|\boldsymbol{f}_m - \Pi_{m,n}\boldsymbol{f}_m\|_n$ can be easily controlled, is the case of a discrete m-th component (i.e. when all $X_{i,m}$ belong to some finite set). In that case, the value $\|\boldsymbol{f}_m - \Pi_{m,n}\boldsymbol{f}_m\|_n$ is zero provided that n is large enough.

Let ϕ^* be the test introduced above in (2.14).

Theorem 1 *Suppose that the observations (X_i, Y_i), $i = 1, \ldots, n$, obey the regression model (1.3) and (1.4), and let condition (D) hold. If the first component f_1 of the function F is linear, then*

$$\boldsymbol{P}_F(\phi^* = 1) \le \alpha + \delta_1(n),$$

where $\delta_1(n) \to 0$ as $n \to \infty$ and it depends on n and constant C arising in condition (D) only.

The proof of the theorem is given in Section 5.

We now turn to the results concerning the sensitivity of test ϕ^*. The first assertion shows under which conditions we reject an alternative with a high probability.

Theorem 2 *Let the function F in model (1.3) be of the form (1.4). Let also $\boldsymbol{\theta}_j^* = (\theta_{I,1}^*, I \in \mathcal{I}_j)$ be the subvector of the vector $\boldsymbol{\theta}^*(d,j)$ corresponding to j-th resolution level of the first component and let $V_j = (v_{(I,1),(I',1)}, I, I' \in \mathcal{I}_j)$ be the covariance submatrix corresponding this index set. If, for some $j \le j_n$, $\beta > 0$ and $c > 0$, it holds*

$$T_j^* \equiv (2N_j)^{-1/2} \sigma^{-2} \boldsymbol{\theta}_j^{*\top} V_j^- \boldsymbol{\theta}_j^* > t_{j,\alpha} + \lambda + t'_{j,\beta} + c\sqrt{\log j_n},$$

with $t'_{j,\beta}$ defined by the equality $\boldsymbol{P}(T_j^0 < -t'_{j,\beta}) = \beta$, then

$$\boldsymbol{P}_F(\phi^* = 0) \le \beta + \delta_1(n)$$

where $\delta_1(n)$ depends on α, β and c only and $\delta_1(n) \to 0$, for $n \to \infty$.

Remark 3 This result claims that the test ϕ^* rejects with a probability close to one any alternative for which at least one of T_j^* exceeds $c'\sqrt{\log j_n}$ with some fixed constant c'. Therefore, we may suppose that the error of the second type may occur only if

$$T_j^* \le c'\sqrt{\log j_n}, \qquad 0 \le j \le j_n. \tag{3.15}$$

Now we discuss how this statement can be transformed into the result about the uniform rate of testing. Following Ingster (1982, 1993) we consider the class of alternatives with the first component f_1 separated from the null (the set of the linear functions) with distance at least ϱ,

$$\inf_{a,b} \| f_1(\cdot) - a - b\cdot \| \ge \varrho$$

where $\|\cdot\|$ means the usual L_2-norm, and in addition we assume that f_1 is smooth in the sense that f_1 belongs to some class of functions $\mathcal{F}$. Ingster (1982) established his results assuming that the underlying function belongs to a Hölder or L_2-Sobolev ball $\mathcal{F}$, Spokoiny (1998) studied the case of a more general L_p-Sobolev ball with any $p \ge 1$.

We are interested in a minimal separation distance ρ which still allows for a uniform testing. To state the result we need some regularity conditions on the design and smoothness conditions on the first component f_1. The reason why stronger conditions on the design are required can be explained by the fact that a degenerate design leads to an identification problem: the components cannot be separated and therefore it is impossible to make any inference about them. Set

$$u_*(j) = \inf_{I \in \mathcal{I}_j} 2^j M_I / n, \qquad u^*(j) = \sup_{I \in \mathcal{I}_j} 2^j M_I / n,$$

where, given $I=(\ell,k)$, the value M_I stands for the number of design points X_i whose first component belongs to the interval $A_I=[k2^{-\ell},(k+1)2^{-\ell})$, that is, $M_I=\#\{i: X_{i,1}\in A_I\}$. *Design regularity* means in particular that $u_*(j)$ is bounded away from zero i.e. each interval A_I contains enough design points $X_{i,1}$, cf. the condition in Lemma 1.

Recall the notation $V_j=\left(v_{(I,1),(I',1)}\,,\, I,I'\in\mathcal{I}_j\right)$ and N_j denotes its rank, $N_j\le 2^j$. Set

$$v^*(j)=\|V_j\|.$$

Here, the norm $\|A\|$ of a matrix A is understood as the maximal eigenvalue of this matrix. We understand design regularity in the sense that V_j is non-degenerate and all the $v^*(j)$'s are bounded.

Finally, given an integer s, suppose that the function f_1 is s times differentiable and the value

$$r_s^2=\int_0^1|f_1^{(s)}(x)|^2dx$$

is finite. $f_1^{(s)}$ means the s-th derivative of f_1.

Theorem 3 *Let condition (D) hold. Suppose there exists an integer s and for some $j\le j_n$, the first component f_1 of the model function F satisfies the following inequality*

$$\inf_{a,b}\|f_1-a-b\psi_{1,1}\|_n^2\ge C_1\,r_s^2\,n\,2^{-js}+C_2\frac{u^*(j)}{u_*(j)}\,v^*(j)\,2^{j/2}\sigma^2\sqrt{\log j_n} \qquad (3.16)$$

with $\psi_{1,1}(x)=x_1$ and constants C_1 and C_2 depending on s only, then

$$\boldsymbol{P}_F(\phi^*=0)\le\delta_1(n)\to 0,\qquad n\to\infty,$$

with $\delta_1(n)$ as in Theorem 2.

The proof of this assertion is based only on (3.15) and can be found in Härdle, Sperlich, Spokoiny (1997) or Spokoiny (1999a).

Remark 4 By minimizing the sum of the form $B_1n2^{-js}+B_22^{j/2}\sigma^2\sqrt{\log\log n}$ with fixed B_1 and B_2 with respect to j we find that a smooth alternative will be rejected with a high probability if

$$\inf_{a,b}n^{-1}\|f_1-a-b\psi_{1,1}\|_n^2\ge B_3\left(\frac{n}{\sigma^2\sqrt{\log\log n}}\right)^{-\frac{2s}{4s+1}}$$

for a constant B_3 depending on B_1 and B_2 only. Spokoiny (1996) has shown that this rate is optimal in the problem of testing against a smooth alternative with an unknown degree of smoothness s.

3.1 Extensions

Here we briefly discussed possible extensions of the test which we introduced previously.

3.1.1 Testing additivity

Though our test was constructed for testing functional forms of the additive components, it can also be useful when the presence of interaction is at question. Often, the additive structure is given or wanted by the economic theory the particular model is based on, see e.g. Deaton and Muellbauer (1980) or also our application in Section 4 However, not only from a statistical point of view it is interesting to scrutinize this assumption in some cases. Several approaches of testing additivity are discussed in Hart (1997), but nonparametric theory for this problem is quite recent, see e.g. Sperlich, Tjøstheim and Yang (1999), also for more references.

As said at the beginning, our procedure can test significance of a component at all. Thus, for testing of no interaction one can proceed as follows. Introduce artificial covariates $X_{m,m'} = X_m X_{m'}$ for $m \neq m'$. No interaction between X_m and $X_{m'}$ means that the covariate $X_{m,m'}$ has no response which is a particular case of the problem we considered before.

3.1.2 Non-Gaussian errors

In our results we suppose Gaussian homoskedastic noise with unknown dispersion σ^2. This assumption allows to simplify the calculations and highlight the main ideas skipping a lot of technical details which appear when considering non-Gaussian noise. However, the results from Section 3 apply for i.i.d. errors with unknown distribution under some moment conditions. We refer to Spokoiny (1999a) for the analysis of non-Gaussian noise in the univariate case. An extension to the multivariate situation is straightforward.

3.1.3 Multiple testing

The above test was developed for testing one component of an additive model. In statistical practice one would be interested to test all the components of the model simultaneously. This leads to a multiple testing problem which requires a more careful evaluation of the corresponding critical values. Following the rule proposed in Section 2 one can construct for every component f_m the corresponding test statistic T_m^* and calculate the corresponding critical level λ_m. Now we apply the same idea of multiscale testing as one used for construction of every component test. Namely, to provide a prescribed nominal level α of the multiple test, which checks all components f_m simultaneously, all these critical values λ_m should be slightly increased, e.g. by the same value $\Delta\lambda$ such that

$$\boldsymbol{P}_0 \left(\max_{m=1,\ldots,d} (T_m^* - \lambda_m) > \Delta\lambda \right) \leq \alpha$$

where $\boldsymbol{P}_0$ means the distribution on the space of observations under the no-response model $F \equiv 0$ with the standard Gaussian errors (that is, Y_i are i.i.d. standard Gaussian).

3.1.4 Local test

In parallel to the test T^* based on the maximum of some quadratic forms of the empirical Haar coefficients $\widehat{\theta}_I$, one may consider another test which is called the "local" test in Härdle et al. (1997). This test is based on the maximum of the standardized empirical coefficients $\widehat{\theta}_I$ over all $I \in \mathcal{I}_j$. More precisely, for every $j \leq j_n$, we define

$$T_{j,\text{loc}} = \max_{I \in \mathcal{I}_j} \frac{\widehat{\theta}_I^2}{\widehat{\sigma}_I^2} - \tau_j$$

where $\widehat{\sigma}_I^2 = \widehat{\sigma}^2 v_{(I,1),(I,1)}$ and τ_j are such that

$$\boldsymbol{P}_0 \left(T_{j,\text{loc}} > \tau_j \right) = \alpha_{\text{loc}}$$

with $\boldsymbol{P}_0$ being again the distribution under the no-response model with standard normal errors. The multilevel "local" test ϕ_{loc}^* is defined by

$$\phi_{\text{loc}}^* = \mathbf{1} \left(\max_{j \leq j_n} T_{j,\text{loc}} > \tau^* \right)$$

where τ^* fulfills

$$\boldsymbol{P}_0 \left(\max_{j \leq j_n} T_{j,\text{loc}} > \tau^* \right) = \alpha_{\text{loc}}.$$

For applications one can use an approximation $\tau_j \approx 2 \log N_j - 2 \log\log N_j + 2 \log \alpha_{\text{loc}}^{-1}$. Such defined "local" test has been shown to be sensitive against a "non-smooth" alternative (e.g. an alternative with jumps), see Haerdle et al. (1997). In practical applications one would be willing to apply both tests T^* and T_{loc}^* simultaneously which requires some additional adjustment of the critical levels for both tests. Taking into account the specific structure of the test ϕ_{loc}^*, our recommendation is to perform this "local" test at a very small significance level, e.g. $\alpha_{\text{loc}} = 0.005$ or even smaller which does not require an additional adjustment of the test ϕ^*.

Also the theoretical properties of such defined test are presented and discussed in there.

4 Simulation studies and an application

The performance of the suggested test procedure for finite samples has been examined in a simulation study and was applied for the analysis of female labor supply data.

4.1 Some simulated examples

We considered 3-dimensional regression problems having additive components of the following form:

$$f_1(x) = 2\sin(\pi x), \quad f_2(x) = 2\sin(3\pi x), \quad f_3(x) = x^2 - \boldsymbol{E}[x^2]$$

and thus three components of quite different smoothness.

In our simulations we consider a uniformly distributed design on the cube $[-2,2]^3$ and standard normal errors. We do not assume to know the standard deviation of the error terms but estimate σ as suggested in Section 2.5 and apply our test procedure for every component f_m, $m = 1,2,3$ independently. The sample size is set to $n = 150$. The sensitivity of the test is investigated by calculating the power functions for different significance levels. We also compare the performance of our procedure with the one for ideal ("oracle") parametric t-test (or Neymann-Pearson NP), see below, for the sample sizes 150 and 300.

Let be f_1 the function of interest for a moment. The exactness and power of the test we calculated by considering the data generating process

$$Y = (1-v)x_1 + vf_1(x_1) + f_2(x_2) + f_3(x_3) + \xi$$

with v running from zero to one. This parameter v has the same meaning as the separation distance between the null and the alternative. The t-test means testing the hypothesis $H_0:\ \beta_2 = 0$ in the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 f_1(x_1) + \beta_3 f_2(x_2) + \beta_4 f_3(x_3) + \xi$$

with known f_1, f_2, f_3.

For getting the critical values we did 249 Monte Carlo simulations from a no response model. To estimate the standard deviation we used $j_{150} = 4$ along the rules presented in Section 2.5 and got, as expected, only slightly overestimated $\widehat{\sigma}_\xi$ (5 to 15%).

The results for the number of rejections after 500 replications with $n = 150$ can be found in the Table 1, together with the average over the resolution levels at which our procedure rejected. As discussed in Remark 2, we have to decide how to choose $t_{j,\alpha}$. We present results for the two most natural choices. First, we set $t_{j,\alpha}$ equal to the $F_{N_j, n-r_n}(\alpha)-$quantile with $\alpha = .01$, $.05$, $.1$ being the significance level (in tables indicated by $F(\alpha)$); second, we tried our procedure with simply $t_{j,\alpha} \equiv 0$ (in tables indicated by $'0'$).

Table 1: Percentage of rejections for functions $(1-v)x_m + vf_m(x_m)$, $m = 1, 2$, and average of active resolution level j_1 (in parenthesis) for two tests with different $t_{j,\alpha}$, i.e. $F(\alpha)$ and $'0'$.

$v\backslash\alpha$	f_1						f_2					
	.01		.05		.10		.01		.05		.10	
	$F(\alpha)$	$'0'$	$F(\alpha)$	$'0'$	$F(\alpha)$	$'0'$	$F(\alpha)$	$'0'$	$F(\alpha)$	$'0'$	$F(\alpha)$	$'0'$
0.0	.018	.004	.060	.054	.108	.104	.010	.004	.050	.048	.102	.106
	(3.22)	(3.00)	(3.30)	(2.56)	(2.90)	(2.44)	(3.40)	(2.00)	(3.08)	(1.88)	(2.49)	(2.11)
0.1	.016	.020	.0660	.074	.110	.122	.006	.004	.060	.058	.104	.110
	(1.25)	(0.70)	(2.03)	(0.973)	(1.82)	(1.26)	(2.00)	(0.00)	(2.67)	(2.14)	(2.44)	(2.04)
0.2	.038	.100	.172	.228	.266	.320	.024	.028	.128	.128	.232	.224
	(0.95)	(0.18)	(0.94)	(0.54)	(0.90)	(0.65)	(2.83)	(2.07)	(2.58)	(1.72)	(2.36)	(1.97)
0.3	.130	.282	.386	.500	.558	.648	.060	.066	.290	.246	.396	.372
	(0.51)	(0.14)	(0.46)	(0.25)	(0.56)	(0.36)	(2.77)	(2.18)	(2.72)	(2.35)	(2.73)	(2.42)
0.4	.358	.588	.720	.810	.828	.860	.236	.210	.502	.464	.642	.620
	(0.15)	(0.08)	(0.20)	(0.16)	(0.23)	(0.17)	(2.80)	(2.26)	(2.69)	(2.33)	(2.63)	(2.41)
0.5	.566	.774	.866	.918	.940	.964	.454	.398	.738	.712	.838	.820
	(0.15)	(0.11)	(0.21)	(0.15)	(0.19)	(0.16)	(2.89)	(2.34)	(2.75)	(2.40)	(2.63)	(2.42)
0.6	.806	.930	.958	.988	.988	.994	.692	.596	.912	.878	.958	.950
	(0.12)	(0.07)	(0.12)	(0.10)	(0.12)	(0.10)	(2.94)	(2.64)	(2.82)	(2.67)	(2.77)	(2.68)
0.7	.948	.982	.990	1.00	1.00	1.00	.878	.832	.974	.970	.986	.988
	(0.06)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(2.97)	(2.79)	(2.89)	(2.77)	(2.85)	(2.77)
0.8	.976	.990	1.00	1.00	1.00	1.00	.946	.912	.990	.990	.998	.996
	(0.06)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(2.94)	(2.72)	(2.85)	(2.68)	(2.78)	(2.68)

We see that the test performs quite well. The difference in power for f_1, f_2 and f_3 corresponds to the different smoothness of the underlying function. Also the information at which resolution level j_1 the procedure rejects, clearly depends on the smoothness of our functions as well as on the distance between null and alternative. It can further be seen, that only looking at one special level would reduce a lot the power of our procedure.

Figure 1 demonstrates how fast the power of our procedure increases and the separation distance between the null and the alternative decreases with the number of observations giving the power functions of our test with $t_{j,\alpha} \equiv 0$ for significant levels being 5%. In parallel we present the results for the ideal t-test.

4.2 Applications

We now turn to an application to demonstrate the performance on real data. The data set is a subsample of the Socio Economic Panel of Germany from 1992. To study the female

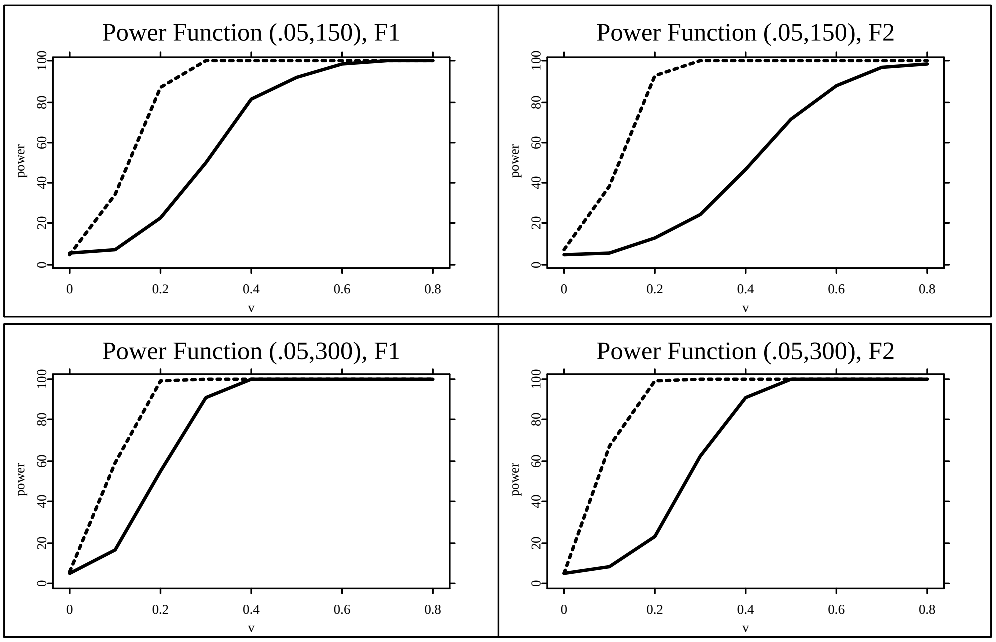

Figure 1: The power functions for $n = 150$ (upper) and $n = 300$ (lower), testing linearity of $(1-v)x_m + vf_m(x_m)$, for $m = 1$ on the left, $m = 2$ on the right with 5% significance level. Solid line is for using wavelets with $t_{j,\alpha} = 0$, dotted line if using t-Test with all information about functional forms.

labor supply in East Germany, 607 women with job and living together with a partner in East Germany have been asked for their weekly number of working hours. The following observations have been chosen as explanatory variables: the age of the woman X_1, her earnings per hour X_2, the prestige index of her kind of profession X_3 (called "Treimann Prestige Index" , see Treimann, 1978), the monthly rent or redemption X_4 for their flat or house, the monthly net income of her partner (in most cases her husband) X_5, her education X_6 measured in years, the unemployment rate X_7 of the particular country of the Federal Republic of Germany where the woman is living in and the number of children younger than 16 years, X_8. These data have already been analyzed nonparametrically in Sperlich (1998), see e.g. for functional forms and further discussion.

The study of female labor supply is quite common in economic empirical research and usually done with (log-) linear additive models using indicators for which the above mentioned set of variables is typical, see e.g. Mroz (1987), Blundell, Duncan, Meghir (1998), Strøm and Wagenhals (1991) or any Handbook of Labor Economics. Among them, especially Mroz (1987) investigated the sensitivity against model specification in this context and found tremendous differences in results depending on the particular specification. A first natural generalization would be to allow the additive components to be non-(log-)linear. Exactly this we now want to investigate. Later, we will additionally have a look on the additivity assumption.

Since some of these explanatory variables are not only discrete but have even less then 10 different realizations observed (e.g. for children and unemployment rate of country - there are only five countries in East Germany), we have to choose respectively low j_n to avoid overparametrization in this component, see Remark 3. Therefore we chose $j_6(n) = 2$ for X_6, $j_7(n) = 2$ for X_7, and for X_8 (number of children) only $j_8(n) = 1$. For X_1, X_2, X_3, X_4 and X_5 we chose $j_n = 4$. These are also the functions of interest we want to analyze.

In Figure 2 we have displayed the wavelet coefficient estimates used for the test statistics. They are standardized, i.e. divided by $v_I\widehat{\sigma}$, but not corrected for the correlation inside the levels $j_1 = 0, 1, 2, 3, 4$. The length is indicating their absolute value.

Often, the earnings per hour (X_2) are modeled log-linear by some reasoning from economic theory as well as it turned out, that linearity did not fit well. So in a second run we also want to test the influence of $\ln(X_2)$ instead of using X_2 against linearity and gave the coefficients in the lower right.

Looking at the construction of T_j, T^*, Figure 2 gives some ideas where we would expect the test to reject: e.g. for "earnings per hour" at $j_1 = 0$, "prestige" at $j_1 = 4$, and "log(earnings per hour)" at $j_1 = 2$, whereas it is not that clear for "age", "income of partner" or "rent/redemption". For the latter one we even would guess that there is definitely no significance in the coefficients.

Actually, the linearity hypothesis was rejected for X_1 at only 10% (at $j_1 = 2$), X_2 at 1% (at $j_1 = 0$), for X_3 at (almost) 5% (p-value≈ 0.052, at $j_1 = 4$), and for $\ln(X_2)$ also at the 1% significance level (at $j_1 = 2$).

Though the additivity assumption is not that of interest for us, we finally also looked for possible second order interactions between the regressors. We applied the procedure described in Section 3.1 to all combinations X_jX_k, $k \neq j$, $j, k = 1, \ldots 8$. As before, we chose $j_6(n) = j_7(n) = 2$, $j_8(n) = 1$. It turned out, that the null hypothesis *no interaction* between "age" and "prestige" is rejected at 1%, "prestige" and "years of education" exactly at 5%, and between "age" and "u-rate" and "earnings of husband" and "u-rate" at the 10% level. Hence, our testing procedure demonstrates that, at least for the underlying data, the standard additive structural assumption in female labor supply model is by far not fulfilled.

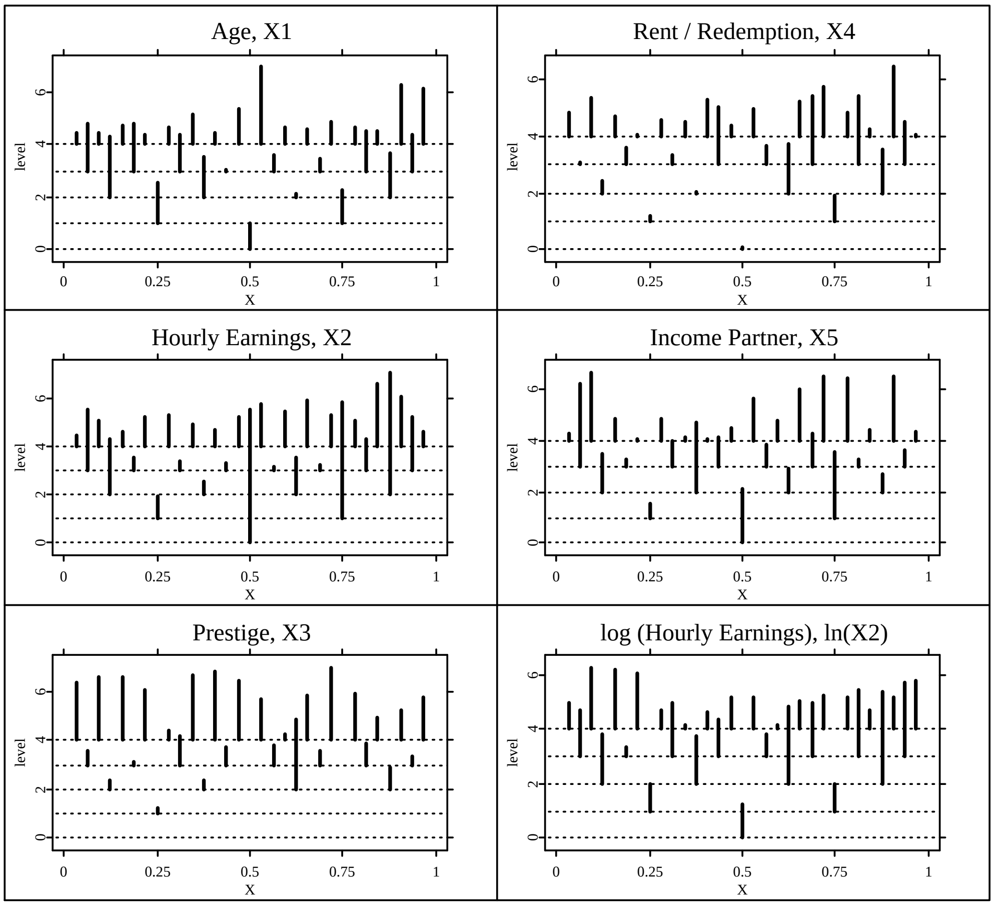

Figure 2: The estimated wavelet coefficients for some additive component functions. The coefficients θ_I are first standardized by $v_I\widehat{\sigma}$; the length indicates size. Coefficients with no length are zero. Resolution levels j_1 are given at vertical axes. The range of the regressors is normed to $[0,1]$ in which the coefficients are positioned in the center of the support of its corresponding mother wavelet ψ, e.g. for $j=0$ on 0.5. Not given are the coefficients for the constant nor for the linear term since they do not enter to the test statistic.

5 Proofs

In this section we collect the proofs of Theorems 1 through 2 and of the other statements presented in Section 3.

5.1 Proof of Lemma 1

By definition of j_n it holds $3dn^{-1}/2 \le 2^{-j_n-1} \le 3dn^{-1}$. Next, define $\Pi'_{m,n}$ as the projector on the space of piecewise constant functions of the m-th component x_m with the piece length 2^{-j_n-1}. Since $\Pi_{m,n}$ projects on a larger space generated by piecewise constant functions and the linear function $\psi_{1,m}(x) = x_m$, it clearly holds $\|\boldsymbol{f}_m - \Pi_{m,n}\boldsymbol{f}_m\|_n \le \|\boldsymbol{f}_m - \Pi'_{m,n}\boldsymbol{f}_m\|_n$.

Let A be an interval of the form $A = [k2^{-j_n-1}, (k+1)2^{-j_n-1})$ and let $N_{A,m}$ denote the number of design points X_i with $X_{i,m} \in A$. The condition of the lemma on the marginals $\mu_{m,n}$ of the empirical measure implies that $N_{A,m} \le C_1 2^{-j_n-1} n$. Denote also by $f_{m,A}$ the arithmetic mean of the values $f_m(X_i)$ over all X_i with $X_{i,m} \in A$. Then $\Pi'_{m,n} f_m(X_i) = f_{m,A}$ and the Lipschitz condition on the component functions f_m yields $|f_m(X_{i,m}) - f_{m,A}| \le C_2 2^{-j_n-1}$ for $X_{i,m} \in A$ and hence

$$\sum_{i:X_{i,m}\in A} |f_m(X_i) - f_{m,A}|^2 \le N_{A,m} \left|C_2 2^{-j_n-1}\right|^2 \le C_1 C_2^2 n 2^{-3j_n-3}.$$

We have 2^{j_n+1} such intervals and therefore

$$\|\boldsymbol{f}_m - \Pi_{m,n}\boldsymbol{f}_m\|_n^2 \le C_1 C_2^2 n 2^{-2j_n-2} \le C_1 C_2^2 d^2 n^{-1}$$

and the assertion follows.

5.2 Some properties of the variance estimate

It is well known that under mild regularity conditions, the unknown variance σ^2 can be estimated at the rate $n^{-1/2}$. We now show that the proposed estimate $\widehat{\sigma}^2$ is also root-n consistent under the condition (D).

The estimate $\widehat{\sigma}^2$ can be represented in the form

$$\begin{aligned} \widehat{\sigma}^2 &= (n - r_n)^{-1} \boldsymbol{Y}^\top (\mathbf{1} - \Pi_n) \boldsymbol{Y} \\ &= (n - r_n)^{-1} (\boldsymbol{\xi} + \boldsymbol{F})^\top (\mathbf{1} - \Pi_n)(\boldsymbol{\xi} + \boldsymbol{F}). \end{aligned}$$

where r_n was the rank of Π_n. Condition (D) provides $\|(\mathbf{1} - \Pi_n)\boldsymbol{F}\|_n \le C\sigma n^{-1/2}$, see the proof of Lemma 1.

Lemma 2 *Under the condition* (D) *it holds*

$$\boldsymbol{P}\left(\left|\sigma^{-2}\widehat{\sigma}^2 - 1\right| > \sqrt{(n-r_n)^{-1}\log n}\right) = o_n(1)$$

where $o_n(1)$ *denotes a numerical sequence tending to zero as* $n \to \infty$. *Moreover,*

$$\boldsymbol{P}\left(\sigma^{-2}\left|\widehat{\sigma}^2 - \widehat{\sigma}_0^2\right| > n^{-1}\right) = o_n(1)$$

where

$$\widehat{\sigma}_0^2 = (n-r_n)^{-1}\boldsymbol{\xi}^\top(\mathbf{1}-\Pi_n)\boldsymbol{\xi}.$$

Proof. By definition

$$\sigma^{-2}(n-r_n)\left(\widehat{\sigma}^2 - \widehat{\sigma}_0^2\right) = \sigma^{-2}\boldsymbol{F}^\top(\mathbf{1}-\Pi_n)\boldsymbol{F} + 2\sigma^{-2}\boldsymbol{\xi}^\top(\mathbf{1}-\Pi_n)\boldsymbol{F}.$$

Condition (D) provides

$$\boldsymbol{F}^\top(\mathbf{1}-\Pi_n)\boldsymbol{F} = \|(\mathbf{1}-\Pi_n)\boldsymbol{F}\|_n^2 \le C^2 n^{-1}.$$

Next, since $\sigma^{-2}\boldsymbol{\xi}^\top(\mathbf{1}-\Pi_n)\boldsymbol{F}$ is the linear combination of Gaussian errors ξ_i, it is also a Gaussian random variable with zero mean and the variance

$$\begin{aligned}
\sigma^{-4}\boldsymbol{E}\left|\boldsymbol{\xi}^\top(\mathbf{1}-\Pi_n)\boldsymbol{F}\right|^2 &= \sigma^{-4}\boldsymbol{E}\boldsymbol{F}^\top(\mathbf{1}-\Pi_n)\boldsymbol{\xi}\boldsymbol{\xi}^\top(\mathbf{1}-\Pi_n)\boldsymbol{F} \\
&= \sigma^{-2}\boldsymbol{F}^\top(\mathbf{1}-\Pi_n)\boldsymbol{F} \le C^2 n^{-1}.
\end{aligned}$$

This implies for every $t \ge 1$

$$\boldsymbol{P}\left(\sigma^{-2}\left(\widehat{\sigma}^2 - \widehat{\sigma}_0^2\right)\frac{C^2}{n(n-r_n)} + \frac{Cn^{-1/2}t}{n-r_n}\right) \le e^{-t^2/2}$$

and the second assertion of the lemma follows in view of $n - r_n \le n/3$.

For the first one, it remains to estimate $\sigma^{-2}\widehat{\sigma}_0^2 - 1 = \sigma^{-2}(n-r_n)^{-1}\boldsymbol{\xi}^\top(\mathbf{1}-\Pi_n)\boldsymbol{\xi}$. Since $\sigma^{-2}\boldsymbol{\xi}^\top(\mathbf{1}-\Pi_n)\boldsymbol{\xi}$ follows the χ^2-distribution with $n-r_n$ degree of freedom, the expression $\sqrt{(n-r_n)/2}\left(\sigma^{-2}\widehat{\sigma}_0^2 - 1\right)$ is asymptotically standard normal and the required assertion follows. ∎

5.3 Proof of Theorem 1

Let j_n be due to (2.10), then with condition (D),

$$\sum_{m=2}^{d}\|\boldsymbol{f}_m - \Pi_{m,n}\boldsymbol{f}_m\|_n \le C\sigma n^{-1/2}$$

and each $\Pi_{m,n}\boldsymbol{f}_m$ can be represented in the form

$$\Pi_{m,n}\boldsymbol{f}_m = \sum_{I\in\mathcal{I}'(j_n)} \theta_{I,m}\boldsymbol{\psi}_{I,m}, \qquad m = 2,\ldots,d,$$

with some coefficients $\theta_{I,m}$, $I \in \mathcal{I} \in \mathcal{I}'(j_n)$.

Under the null hypothesis, $\boldsymbol{f}_1 = \theta_{0,1} + \theta_{1,1}\boldsymbol{\psi}_{1,1}$. Define

$$\boldsymbol{F}' = \boldsymbol{F} - \boldsymbol{f}_1 - \Pi_{2,n}\boldsymbol{f}_2 - \ldots - \Pi_{d,n}\boldsymbol{f}_d.$$

Then the above bound yields

$$\|\boldsymbol{F}'\|_n = \|\boldsymbol{F} - \boldsymbol{f}_1 - \Pi_{2,n}\boldsymbol{f}_2 - \ldots - \Pi_{d,n}\boldsymbol{f}_d\|_n \le C\sigma n^{-1/2}. \tag{5.17}$$

Now we show that the original regression function F can be replaced by F'.

Lemma 3 *The change F for F' does not affect the test statistics $T^* = \max_{j\le j_n} T_j$.*

Proof. Let some $j \le j_n$ be fixed. Denote by $\boldsymbol{\theta}(d,j) = (\theta_{I,m}, (I,m) \in \mathcal{I}(d,j))$ the vector with $\theta_{I,1} = 0$ for $I \in \mathcal{I}_j$, $0 \le j \le j$, and with the above defined $\theta_{I,m}$ for $m \ge 2$. Then $\boldsymbol{F}' = \boldsymbol{F} - \Psi(d,j)\boldsymbol{\theta}(d,j)$ and the inequality in (5.17) can be rewritten in the form $\|\boldsymbol{F} - \Psi(d,j)\boldsymbol{\theta}(d,j)\|_n \le \sigma n^{-1/4}$. Define also $\boldsymbol{\theta}^*(d,j) = W(d,j)^\top \boldsymbol{F}$ and let $\boldsymbol{\theta}_j^*$ be the subvector of $\boldsymbol{\theta}^*(d,j)$ corresponding to the j th resolution level of the first component. This vector can be written in the form $\boldsymbol{\theta}_j^* = \mathcal{E}_j\boldsymbol{\theta}^*(d,j)$ with $\mathcal{E}_j$ being the projector from $I\!R^{N(d,j)}$ onto $I\!R^{2^j}$ keeping the entries $\theta_{I,1}^*$, $I \in \mathcal{I}_j$, of the vector $\boldsymbol{\theta}^*(d,j)$ corresponding to the j th resolution level of the first component. Then it holds

$$\boldsymbol{\theta}_j^* = \mathcal{E}_j W(d,j)^\top \boldsymbol{F} = W_j^\top \boldsymbol{F}.$$

Since the test statistic T^* is calculated via the estimates $\widehat{\boldsymbol{\theta}}_j = W_j^\top \boldsymbol{Y}$ for $j \le j_n$, and since $W_j^\top \boldsymbol{Y} = W_j^\top \boldsymbol{\xi} + W_j^\top \boldsymbol{F}$, it only remains to check that $W_j^\top \Psi(d,j)\boldsymbol{\theta}(d,j) = 0$ for all $j \le j_n$. The definition of $\boldsymbol{\theta}(d,j)$ provides $\mathcal{E}_j\boldsymbol{\theta}(d,j) = 0$, and hence,

$$\begin{aligned} W_j^\top \Psi(d,j)\boldsymbol{\theta}(d,j) &= \mathcal{E}_j W(d,j)^\top \Psi(d,j)\boldsymbol{\theta}(d,j) \\ &= \mathcal{E}_j \left(\Psi(d,j)^\top \Psi(d,j)\right)^- \Psi(d,j)^\top \Psi(d,j)\boldsymbol{\theta}(d,j) = 0 \end{aligned}$$

as required. ∎

This lemma allows to reduce the statement of the theorem to the case with $\|\boldsymbol{F}\|_n \le C\sigma n^{-1/2}$.

Recall that the critical value of the test is evaluated under the condition $F \equiv 0$. Now we intend to show that $\boldsymbol{P}_F(\phi^* = 1) = \alpha + o_n(1)$ for every regression function F satisfying $\|\boldsymbol{F}\|_n \le C\sigma n^{-1/2}$. The test ϕ^* is based on the test statistic $T^* = \max_{j\le j_n}(T_j - t_{j,\alpha})$

with

$$T_j = \frac{\boldsymbol{Y}^\top W_j V_j^- W_j^\top \boldsymbol{Y}}{\widehat{\sigma}^2 \sqrt{2N_j}} - \sqrt{N_j/2} = \frac{\boldsymbol{Y}^\top \mathcal{R}_j \boldsymbol{Y}}{\widehat{\sigma}^2 \sqrt{2N_j}} - \sqrt{N_j/2}.$$

Here W_j is the submatrix of the matrix $W(d,j)$ corresponding to the j th resolution level of the first component, $W_j = \mathcal{E}_j W(d,j)$, and $V_j = W_j^\top W_j$, so that $\mathcal{R}_j = W_j V_j^- W_j^\top$ is a projector in $I\!R^n$ on the N_j-dimensional subspace. The model $\boldsymbol{Y} = \boldsymbol{F} + \boldsymbol{\xi}$ implies

$$T_j = \frac{\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{\xi}}{\widehat{\sigma}^2 \sqrt{2N_j}} - \sqrt{N_j/2} + \frac{2\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{F} + \boldsymbol{F}^\top \mathcal{R}_j \boldsymbol{F}}{\widehat{\sigma}^2 \sqrt{2N_j}}.$$

Define

$$T_j^0 = \frac{\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{\xi}}{\widehat{\sigma}_0^2 \sqrt{2N_j}} - \sqrt{N_j/2}. \tag{5.18}$$

We intend to bound the difference $T_j - T_j^0$.

Lemma 4 *Let condition* (D) *be fulfilled and the component* f_1 *be a linear functions. Then it holds*

$$\sum_{j=0}^{j_n} \boldsymbol{P}\left(|T_j - T_j^0| > \epsilon_j\right) = o_n(1). \tag{5.19}$$

where $\epsilon_j = 3C\sqrt{\frac{\log j_n}{nN_j}}$.

Proof. Clearly we have

$$T_j - T_j^0 = \frac{\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{\xi}}{\sqrt{2N_j}} \left(\widehat{\sigma}^{-2} - \widehat{\sigma}_0^{-2}\right) + \frac{2\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{F} + \boldsymbol{F}^\top \mathcal{R}_j \boldsymbol{F}}{\widehat{\sigma}^2 \sqrt{2N_j}}.$$

Similarly to the proof of Lemma 2 one can show that

$$\frac{\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{\xi}}{\sqrt{2N_j}} \left(\widehat{\sigma}^{-2} - \widehat{\sigma}_0^{-2}\right) = o(n^{-1})$$

and for every $t \geq 1$,

$$\boldsymbol{P}\left(\frac{2\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{F}}{\sigma^2 \sqrt{2N_j}} > \frac{2\|\mathcal{R}_j \boldsymbol{F}\|_n t}{\sigma^2 \sqrt{2N_j}}\right) \leq e^{-t^2/2}. \tag{5.20}$$

Since $\|\mathcal{R}_j \boldsymbol{F}\|_n \leq \|\boldsymbol{F}\|_n \leq Cn^{-1/2}$, this inequality applied with $t = \sqrt{3\log j_n}$ yields

$$\sum_{j=0}^{j_n} \boldsymbol{P}\left(\frac{2\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{F} + \boldsymbol{F}^\top \mathcal{R}_j \boldsymbol{F}}{\sigma^2 \sqrt{2N_j}} > Cn^{-1} + 2C\sqrt{\frac{3\log j_n}{2nN_j}}\right)$$
$$\leq (1 + j_n) e^{3/2 \log j_n} = o_n(1).$$

Now the required assertion follows in view of the first statement of Lemma 2. ∎

By construction, it holds

$$\boldsymbol{P}\left(\max_{0\le j\le j_n}(T_j^0 - t_{j,\alpha}) > \lambda\right) = \alpha.$$

The idea is to show that this equality remains valid in the asymptotic sense if we replace here T_j^0 by T_j. Assertion (5.19) yields

$$\begin{aligned}\boldsymbol{P}\left(\max_{0\le j\le j_n}(T_j^0 - t_{j,\alpha} - \epsilon_j) > \lambda\right) - o_n(1) &\le \boldsymbol{P}\left(\max_{0\le j\le j_n}(T_j - t_{j,\alpha}) > \lambda\right)\\ &\le \boldsymbol{P}\left(\max_{0\le j\le j_n}(T_j^0 - t_{j,\alpha} + \epsilon_j) > \lambda\right) + o_n(1).\end{aligned}$$

Now it suffices to check that

$$\sum_{j=0}^{j_n}\boldsymbol{P}\left(|T_j^0 - t_{j,\alpha} - \lambda| \le \epsilon_j\right) = o_n(1).$$

The distribution of T_j^0 is precisely known and for sufficiently large n it is very close to the centered and standardized χ^2-distribution with N_j degrees of freedom. This particularly yields that the density of this distribution with respect to the Lebesgue measure is bounded by 1 and therefore,

$$\sum_{j=0}^{j_n}\boldsymbol{P}\left(|T_j^0 - t_{j,\alpha} - \lambda| \le 3C\sqrt{\frac{\log j_n}{nN_j}}\right) \le \sum_{j=0}^{j_n} 6C\sqrt{\frac{\log j_n}{nN_j}} = o_n(1)$$

and the theorem is proved.

5.4 Proof of Theorem 2

The proof utilizes the following technical

Lemma 5 *For sufficiently large n, it holds*

$$\max_{0\le j\le j_n} t_{j,\alpha} + \lambda \le 2\sqrt{\log j_n}.$$

Proof. The statement obviously follows from the fact that

$$\sum_{j=0}^{j_n}\boldsymbol{P}\left(T_j^0 > 2\sqrt{\log j_n}\right) = o_n(1)$$

where every T_j^0 is defined by centering and standardization of a χ^2-sum with N_j degree of freedom, see Spokoiny (1999a) for more details. ∎

Let, for some $j \le j_n$, it holds

$$T_j^* = (2N_j)^{-1/2}\sigma^{-2}\boldsymbol{\theta}_j^{*T} V_j^- \boldsymbol{\theta}_j^* \ge (t_{j,\alpha} + \lambda + t'_{j,\beta}) + c\sqrt{\log j_n}$$

with some $c > 0$. This inequality can be represented in the form

$$\frac{\boldsymbol{F}^\top \mathcal{R}_j \boldsymbol{F}}{\sigma^2\sqrt{2N_j}} > 2t_{j,\alpha} + \lambda + t'_{j,\beta} + a\sqrt{\log j_n}. \tag{5.21}$$

We will show that under the above assumption,

$$\boldsymbol{P}_F(T_j < t_{j,\alpha} + \lambda) \le \beta + o_n(1),$$

which obviously implies the assertion.

Similarly to the proof of Theorem 1 we derive

$$\begin{aligned} T_j - T_j^0 &= \frac{\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{\xi}}{\sqrt{2N_j}}\left(\frac{1}{\widehat{\sigma}^2} - \frac{1}{\widehat{\sigma}_0^2}\right) + \frac{2\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{F} + \boldsymbol{F}^\top \mathcal{R}_j \boldsymbol{F}}{\widehat{\sigma}^2\sqrt{2N_j}} \\ &= o(n^{-1}) + \frac{2\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{F} + \boldsymbol{F}^\top \mathcal{R}_j \boldsymbol{F}}{\widehat{\sigma}^2\sqrt{2N_j}} \\ &= o(n^{-1}) + \frac{2\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{F} + \boldsymbol{F}^\top \mathcal{R}_j \boldsymbol{F}}{\sigma^2\sqrt{2N_j}} \end{aligned}$$

with T_j^0 from (5.18). The definition of $t'_{j,\beta}$ provides

$$\boldsymbol{P}\left(-T_j^0 < -t'_{j,\beta}\right) = \beta.$$

Now (5.20) implies

$$\boldsymbol{P}_F(T_j < t_{j,\alpha} + \lambda) \le \boldsymbol{P}\left(T_j^0 < -t'_{j,\beta}\right) + \boldsymbol{P}\left(-\gamma_j > T_j^* - t_{j,\alpha} - \lambda - t'_{j,\beta}\right) + o_n(1)$$

where $\gamma_j = \frac{2\boldsymbol{\xi}^\top \mathcal{R}_j \boldsymbol{F}}{\sigma^2\sqrt{2N_j}}$ is a Gaussian r.v. with zero mean and $\boldsymbol{E}\gamma_j^2 = 4T_j^*(2N_j)^{-1/2}$. It remains to check that condition (5.21) and Lemma 5 imply

$$\frac{4T_j^*(2N_j)^{-1/2}}{(T_j^* - t_{j,\alpha} - \lambda - t'_{j,\beta})^2} = o_n(1).$$

References

[1] Aerts, M., Claesken, G. and Hart, J.D., (1999), "Testing the fit of a parametric function," *J. Amer. Statis. Ass.*, 94, 869–879.

[2] Bierens, H.J., (1982), "Consistent model specification tests," *J. Econometrics*, 20, 105–134.

[3] Blundell, R., A.Duncan and C.Meghir (1998), "Estimating Labor Supply Responses using Tax Reforms," *Econometrica*, 4, 827–861.

[4] Buja, A., Hastie, T.J., and Tibshirani, R.J., (1989), "Linear smoothers and additive models," *Ann. Statist.*, 17, no. 2, 453–555.

[5] Burnashev, M.V., (1979), "On the minimax detection of an inaccurately known signal in a white Gaussian noise background," *Theory Probab. Appl.*, 24, 107–119.

[6] Deaton, A. and Muellbauer, J. (1980), *Economics and Consumer Behavior,* Cambridge University Press, New York.

[7] Eubank, R.L., and Hart, J.D. (1992), "Testing Goodness of Fit in Regression via Order Selection Criteria," *Annals of Statisitcs*, 20, 1412–1425.

[8] Fan, J., (1996), "Test of significance based on wavelet thresholding and Neyman's truncation," *J. Amer. Statist. Ass.*, 91, 674–688.

[9] Fan, J. and Huang, L., (1998), "Goodness-of-fit test for parametric regression models," Preprint, University of North Caroline.

[10] Gasser, T. and Sroka, L. and Jennen-Steinmetz, C., (1986), "Residual variance and residual pattern in nonlinear regression," *Biometrika*, 73, 625–633.

[11] Hall, P., Kay, J.W. and Titterington, D.M. (1991), "On estimation of noise variance in two-dimensional signal processing," *Adv. Appl. Probab.*, 23, 476–495.

[12] Härdle, W., and Korostelev, A., (1996), "Search of significant variables in nonparametric additive regression," *Biometrika*, 83, 541–549.

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

[14] Härdle, W., Sperlich, S., and Spokoiny, V., (1997), Component analysis for additive models", *SFB Discussion Paper*, **52**, Humboldt University, Berlin.

[15] Hart, J.D. (1997), *Nonparametric Smoothing and Lack-of-Fit Tests* New York, Berlin, Heidelberg: Springer.

[16] Hastie, T.J., and Tibshirani, R.J., (1990), *Generalized additive models.* Chapman & Hall.

[17] Horowitz, J., and Spokoiny, V., (1999), "On adaptive, rate-optimal test of a parametric model against a nonparametric alternative," *Econometrica*, tentatively accepted.

[18] Inglot, T., and Ledwina, T., (1996), Asymptotic optimality of data-driven Neyman's tests for uniformity, *Ann. Statist.*, 24, no. 5, 1982–2019.

[19] Ingster, Yu.I., (1982), "Minimax nonparametric detection of signals in white Gaussian noise," *Problems Inform. Transmission*, 18, 130–140.

[20] Ingster, Yu.I., (1993), "Asymptotically minimax hypothesis testing for nonparametric alternatives," I–III. *Math. Methods of Statist.*, 2, 85–114; 3, 171–189; 4, 249–268.

[21] Kallenberg, W.C.M. and Ledwina, T. (1995), "Consistency and Monte-Carlo simulatioins of a data driven version of smooth goodness-of-fit tests," *Ann. Statist.*, 23, no. 5, 1594–1608.

[22] Ledwina, T. (1994), "Data-driven version of Neyman's smooth test of fit," *J. Amer. Statis. Ass.*, 89, no. 427, 1000–1005.

[23] Leontief, W., (1947), "Introduction to a theory of the internal structure of functional relationships," *Econometrica*, 15, 361–373.

[24] Linton, O.B., and Nielsen, J.P., (1995), "A kernel method of estimating structured nonparametric regression based on marginal integration," *Biometrika*, 82, no. 1, 93–100.

[25] Mallows, C.L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

[26] Mroz, T.A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–800.

[27] Rice, J., (1984), "Bandwidth choice for nonparametric regression," *Ann. Statis.*, 12, 1215–1230.

[28] Sperlich, S., (1998), *Additive modelling and Testing Model Specification.* Shaker, Aachen.

[29] Sperlich, S., D.Tøstheim and L.Yang (1999), "Nonparametric Estimation and Testing of Interaction in Additive Models," *Preprint, Carlos III de Madrid, Spain*

[30] Spokoiny, V., (1996), "Adaptive hypothesis testing using wavelets," *Ann. Statist.*, 24, no.6, 2477–2498.

[31] Spokoiny, V., (1998), "Adaptive and spatially adaptive testing of a nonparametric hypothesis," *Math. Methods of Statist.*, 7, no. 3, 245–273.

[32] Spokoiny, V., (1999a), "Data-driven testing the fit of linear models," *Preprint 471*, Weierstrass-Institute, Berlin.

[33] Spokoiny, V., (1999b), "Variance estimation for high-dimensional regression models," *Preprint 503*, Weierstrass-Institute, Berlin.

[34] Stone, C.J., (1985), "Additive regression and other nonparametric models," *Ann. Statist.*, 13, no. 2, 689-705.

[35] Stone, C.J., (1986), "The dimensionality reduction principle for generalized additive models," *Ann. Statist.*, 14, 590-606.

[36] Strøm, St. and G.Wagenhals (1991), "Female Labour Supply in the Federal Republik," *Jahrbuch für Nationalökonomie und Statistik*, 208(6), 575–595.

[37] Stute, W., (1997), "Nonparametric model checks for regression," *Ann. Statist.* 25, no. 2, 613–641.

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

[39] Treiman, D.J., (1978), "Probleme der Begriffsbildung und Operationalisierung in der international vergleichenden Mobilitätsforschung," in *Sozialstrukturanalysen mit Umfragedaten*, edit. by F.U. Pappi, Athenäum, Kronberg im Taunus.

[40] Venables, W.N., and Ripley, B., (1994), *Modern applied statistics with S-Plus*, Springer, N.Y.