COMPONENT ANALYSIS FOR ADDITIVE MODELS

HÄRDLE, W.[1], SPERLICH, S.[1], AND SPOKOINY, V.[2]

[1] *Humboldt-Universität zu Berlin, Institut für Statistik & Ökonometrie, Spandauer Str. 1, 10178 Berlin*

[2] *Weierstrass-Institut, Mohrenstr. 39, 10117 Berlin*

ABSTRACT. We consider the component analysis problem for a regression model with an additive structure. The problem is to check the hypothesis of linearity for each component without specifying the structure of the remaining components. In this paper we show that under mild conditions on the design and smoothness of the regression function, each component can be tested with the rate corresponding to the case if all the remaining components were known. The proposed procedure is based on the Haar transform and it is computationally straightforward.

1. Introduction

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

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

with $x = (x_1, \ldots, x_d)^T$. 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 $\{(X_i, Y_i) : i = 1, \ldots, n\}$. 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 neighbourhood 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*

1991 *Mathematics Subject Classification.* 62H25; Secondary 62G10.

Key words and phrases. additive model, component analysis, Haar basis, hypothesis of linearity, nonparametric alternative, regression .

This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation ökonomischer Prozesse", Berlin, Germany .

modelling which has been used in many applications and for which software is easily accessible.

A nonparametric additive regression model has the form

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

$$F(x) = f_1(x_1) + \ldots + f_d(x_d), \tag{1.2}$$

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 practice: additive models naturally generalize the linear regression models and allow interpretation of marginal changes i.e. the effect of one variable on the mean function F holding all else constant. They are interesting from a theoretical point of view since they combine flexible nonparametric modelling of many variables with statistical precision that is typical for just one explanatory variable. To our knowledge model (1.2) has been first considered in the context of input-output analysis by Leontief (1947) who called it *additive separable*. In the statistical literature the additive regression model has been introduced in the eighties, see Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1990). Stone (1985, 1986) proved that model (1.2) can be estimated with a one-dimensional rate of convergence typical for estimating a single function f of one regressor only.

Algorithmic aspects of additive modelling by e.g. the backfitting or the Gauss-Seidel algorithm are discussed in Venables and Ripley (1994). Linton and Nielsen (1995) proposed a method of analysis of additive models based on marginal integration. Linton and Härdle (1996) extended this approach to generalized additive models.

It is of basic interest in additive modelling to analyze the components further. Härdle and Tsybakov (1995) proposed a selection procedure to determine which covariates should be included in an additive regression model. The general problem of finding all the significant components can be regarded as a family of testing problems: for each component f_m, we test the null hypothesis $f_m \equiv 0$ or, in the other words, f_m is not significant. Härdle and Korostelev (1996) considered a similar problem of testing a single component f_m of the regression function F with a criteria based on large deviation asymptotics for probabilities of errors.

In this paper we concentrate on the more general problem of testing linearity for each single component, say f_1. Therefore we test the null hypothesis $H_0 : f_1$ 'is linear' versus the nonparametric alternative $H_1 : f_1$ 'is not linear'.

The nonparametric theory of hypothesis testing is now well developed. The problem of testing a simple null hypothesis versus a univariate nonparametric alternative is studied in details, see Ingster (1993) or Lepski and Spokoiny (1995) for a historical background and further references. Härdle and Mammen (1993) and Härdle, Spokoiny and Sperlich (1997)

considered the problem of testing a parametric null hypothesis versus nonparametric or semiparametric alternative in the case of a multivariate regression function.

These ealier results show that the optimal quality or rate of testing and the structure of rate-optimal tests essentially depend on the smoothness properties of the underlying function. An adaptive testing procedure which does not require knowledge of smoothness properties of the tested function, has been proposed in Spokoiny (1996a). This adaptive method achieves a "near optimal rate" of testing which is worse than optimal only by some log log-factor. The latter can be viewed as the price for adaptation.

In this paper we develop further this approach and apply it to multidimensional situation with an additive model function. A direct application of the original procedure from Spokoiny (1996a) is impossible for several reasons. First, we consider now a special testing problem. The additional difficulty comes from the fact that the function f_1, even being completely specified, does not specify the whole model since nothing is saying about the other components, $f_2, \ldots, f_d$. These functions can be viewed as an infinite-dimensional nuisance parameter. Therefore, we aim to develop a procedure, which is adaptive both to unknown smoothness properties of the first components f_1 and to the presence of the nuisance parameters.

Second, the original procedure was developed for the so called 'signal + white noise' model which is an idealization of the regression model with the equispaced design. In view of real applications which we address in Section 5 we try to relax these assumptions and to deal with an arbitrary fixed design.

Third, the above mentioned testing procedure from Spokoiny (1996a) applies a wavelet transform with a regular wavelet basis. Here we apply a Haar decomposition which is a particular (and non-regular) case of the wavelet transform. This choice of basis allows to relax and to simplify the conditions on the design and also to reduce the computational difficulties. As we will see, the loss of regularity of the basis does not necessarily result in a loss of sensitivity of testing.

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. Our testing procedure is described in Section 3. The asymptotic properties of this procedure are discussed in Section 4. We compare the sensitivity of our procedure with the optimal one designed for the ideal situation when the other components and smoothness properties of the tested component were known. The results show that our procedure achieves a nearly optimal rate by some log-factor. The results are stated under very mild conditions on the design which can be instructively checked in practical applications. In our results we suppose Gaussian homoskedastic noise with known 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 heteroskedastic non-Gaussian noise. We indicate where necessary how the procedure can be applied in this situation too by plugging in a pilot estimate of the noise variance in place of the σ^2.

Some simulation results and applications are shown in Section 5. The proofs are postponed until Section 6.

2. Model and testing problem

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

$$Y_i = F(X_i) + \xi_i \tag{2.1}$$

where F is an unknown regression function with the additive structure

$$F(x) = f_1(x_1) + \ldots + f_d(x_d), \tag{2.2}$$

and ξ_i are normal random errors with zero mean and known variance σ^2. The design $X_1, \ldots, X_n$ is not assumed random or regular (for example, equidistant) because we want to keep things as general as possible.

Our aim is to analyze each component f_m, $m = 1, \ldots, d$. For simplicity of presentation we concentrate on the first component f_1. More specifically, we wish to test the hypothesis of linearity $H_0 : f_1$ 'is linear' that means that $f_1(t) = a_1 + b_1 t$ for some constants a_1, b_1.

Let ϕ be a test i.e. 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 (2.1) and (2.2). Let now F_0 be a function with a linear first component. The error probability of the first kind 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 kind. If the first component f_1 of a function F is not linear, then

$$\beta_F(\phi) = \boldsymbol{P}_F(\phi = 0).$$

Given $\alpha_0 > 0$, we wish to construct such a test ϕ that $\alpha_{F_0}(\phi) \le \alpha_0$ for all F_0 with a linear first component and, in addition, it is sensitive against a large class of alternatives F. Obviously no test can be sensitive against all alternatives. Following Ingster (1982, 1993) we consider therefore the class of alternatives with the first component f_1 which is separated from the set of linear functions with distance at least ϱ,

$$\inf_{a,b} \| f_1(\cdot) - a - b \cdot \| \ge \varrho \tag{2.3}$$

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 $\mathcal{F}$. The reason for introducing smoothness assumptions is that if the underlying function is very irregular, then it is impossible to distinguish between the noise and the systematic component, Burnashev (1979). Ingster (1982) established his results assuming that the underlying function belongs to a Hölder or L_2-Sobolev ball $\mathcal{F}$, Lepski and Spokoiny (1995), Spokoiny (1996b) studied the case of a L_p-Sobolev ball with $p < 2$ which corresponds to alternatives with heterogeneous smoothness properties. We use a slightly different form of smoothness assumptions supposing that f_1 is approximated with a certain rate by piecewise polynomials, see Section 4.

In general, the testing problem can be formulated as follows. Let some class $\mathcal{F}$ of univariate functions be fixed. We test the null hypothesis $H_0 : f_1(x_1) = a + bx_1$ against the alternative $H_1(\varrho) : f_1 \in \mathcal{F}$ and f_1 obeys (2.3). Given positive numbers $\alpha_0 < 1$ and $\beta_0 < 1$, we consider tests ϕ such that

$$P_{F_0}(\phi = 1) \le \alpha_0$$
$$P_F(\phi = 0) \le \beta_0$$

for all F_0 with a linear first component and for all F from the alternative set $H_1(\varrho)$. We characterize the sensitivity of each test ϕ by the minimal distance ϱ for which the above constraints on the probabilities of errors are satisfied. A test ϕ^* which leads to the minimal distance ϱ^* among all feasible tests is called *optimal*.

Further we consider the asymptotic set-up assuming that the number of observations tends to infinity. Increasing the number n of observations results in higher sensitivity. We let therefore the value ϱ^* depend on n, $\varrho^* = \varrho^*(n)$. This sequence $\varrho^*(n)$ determines "the optimal rate" of testing as n increases to infinity.

Our final goal is to construct tests $\phi(n)$ with a "near optimal" sensitivity $\varrho(n)$ in the sense that the ratio $\varrho(n)/\varrho^*(n)$ grows at most logarithmically.

3. Testing procedure

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

3.1. The case of $d = 1$

Consider the univariate regression model

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \dots, n. \tag{3.1}$$

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 versus a nonparametric alternative.

The proposed procedure makes use of the Haar decomposition of the function f.

3.1.1. **The Haar transform**

Let us recall the construction and the main properties of the Haar transform. By I we denote the multi-index $I = (j,k)$ with $j = 0,1,2,\ldots$ 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

$$\psi(t) = \begin{cases} 0, & t < 0,\ t \geq 1, \\ 1, & 0 \leq t < 1/2, \\ -1, & 1/2 \leq t < 1. \end{cases} \tag{3.2}$$

For every $I = (j,k)$, set

$$\psi_I(t) = 2^{j/2}\psi(2^j t - k). \tag{3.3}$$

Clearly the function ψ_I is supported by the interval $[2^{-j}k, 2^{-j}(k+1)]$. Each measurable function f can be decomposed in the following way, see e.g. Alexits (1961, p.57),

$$f(t) = c_0 + \sum_{I \in \mathcal{I}} c_I \psi_I(t). \tag{3.4}$$

Hence the problem of recovering the function f in (3.1) can be reformulated as the problem of estimating the coefficients c_I from given data. Since we have only n observations, we restrict the total number of considered levels. Fix the level j_1 such that

$$2^{j_1+1} \leq n$$

and set

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

for the index set corresponding to j-th level. Now we approximate the infinite decomposition (3.4) by the finite sum $\sum_{I \in \mathcal{I}(j_1)} c_I \psi_I(t)$ where the index set $\mathcal{I}(j_1)$ contains all level sets $\mathcal{I}_j$ with $j \leq j_1$. Taking into account the structure of the null hypothesis, we complement the set of functions $(\psi_I,\ I \in \mathcal{I}_j)$, $j \leq j_1$, with two functions $\psi_0 \equiv 1$ and $\psi_1(t) = t$, and set

$$\mathcal{I}(j_1) = \{0,1\} + \bigcup_{j=0}^{j_1} \mathcal{I}_j. \tag{3.5}$$

The idea of the proposed procedure is first to estimate all the coefficients $\{c_I,\ I \in \mathcal{I}(j_1)\}$ from the data and then to test that all the coefficients c_I for $I \neq 0,1$ are zero.

Before we begin with our procedure, let us note that the functions ψ_0 and ψ_I, $I \in \mathcal{I}$, form an ortonormal basis in $L_2[0,1]$ with respect to Lebesgue measure on $[0,1]$. When dealing with real data, it may occur that the functions ψ_I are no more orthonormal and

are not orthogonal to each other in $L_2(\mu_n)$, where μ_n is the empirical design measure, $\mu_n(A) = \sum_{i=1}^n \mathbf{1}(X_i \in A)$. To cope with this, we replace the functions ψ_I by its standardized versions ψ'_I: for $I = (j,k)$,

$$\psi'_I(t) = \lambda_I^{-1}\psi(2^j t - k), \tag{3.6}$$

where ψ is defined in (3.2) and the normalizer λ_I satisfies

$$\lambda_I^2 = \sum_{i=1}^n |\psi(2^j X_i - k)|^2. \tag{3.7}$$

(Recall that $d = 1$ and hence X_i takes values in the interval $[0,1]$.) Particularly, $\lambda_0^2 = n$, $\lambda_1^2 = (X_1^2 + \ldots + X_n^2)$, and

$$\lambda_I^2 = M_I = \#\{i : X_i \in [2^{-j}k, 2^{-j}(k+1))\}, \qquad I \in \mathcal{I}.$$

In the sequel, we approximate the function f by linear combinations of the functions ψ'_I, $I \in \mathcal{I}(j_1)$. Let g be a function observed at point $X_1, \ldots, X_n$. Define $\|g\|_n$ by

$$\|g\|_n^2 = \sum_{i=1}^n g^2(X_i).$$

Determine a column-vector $\boldsymbol{\theta}^*(j_1) = (\theta_I^*, I \in \mathcal{I}(j_1))$ as a minimizer of the error of approximation,

$$\boldsymbol{\theta}^*(j_1) = \operatorname*{arginf}_{\boldsymbol{\theta}(j_1)} \|f - \sum_{I \in \mathcal{I}(j_1)} \theta_I \psi'_I\|_n. \tag{3.8}$$

Such a vector always exists but may not be unique.

To get an explicit representation of $\boldsymbol{\theta}^*(j_1)$, we introduce matrix notation. First of all, we make an agreement to identify every function g on $\mathbb{R}^d$ with the vector $(g(X_i), i = 1, \ldots, n)$ in $\mathbb{R}^n$. Particularly, the model function f is identified with the vector $(f(X_i), i = 1, \ldots, n)$. Define also Y as the column vector $(Y_1, \ldots, Y_n)^T$, where the sign ^{T} means transposition.

Denote by N_j the number of elements in each level j,

$$N_j = \#(\mathcal{I}_j) = 2^j$$

and let $N(j_1)$ be the total number of elements in the set $\mathcal{I}(j_1)$,

$$N(j_1) = 2 + \sum_{j=0}^{j_1} N_j = 1 + 2^{j_1+1}. \tag{3.9}$$

Introduce $n \times N(j_1)$-matrix $\Psi(j_1) = (\psi_{i,I}, i = 1, \ldots, n, I \in \mathcal{I}(j_1))$ with elements

$$\psi_{i,I} = \psi'_I(X_i) = \lambda_I^{-1}\psi(2^j X_i - k), \qquad I \in \mathcal{I}(j_1), i = 1, \ldots, n. \tag{3.10}$$

Now the approximation problem (3.8) can be rewritten in the form

$$\boldsymbol{\theta}^*(j_1) = \operatorname*{arginf}_{\boldsymbol{\theta}(j_1)} \|f - \Psi(j_1)\boldsymbol{\theta}(j_1)\|_n^2.$$

The solution to this quadratic problem can be represented as

$$\boldsymbol{\theta}^*(j_1) = \left\{\Psi^T(j_1)\Psi(j_1)\right\}^{-1}\Psi^T(j_1)f. \tag{3.11}$$

Strictly speaking, this representation is valid only if the matrix $\Psi^T(j_1)\Psi(j_1)$ is not degenerate. More generally one may use the same expression for $\boldsymbol{\theta}^*(j_1)$ when understanding $\left\{\Psi^T(j_1)\Psi(j_1)\right\}^{-1}$ as a pseudo-inverse matrix.

We begin our testing procedure by estimating the coefficients $(\theta_I^*,\, I \in \mathcal{I}(j_1))$ by the least square method.

3.1.2. **Estimating the Haar coefficients**

The least squares estimator $\hat{\boldsymbol{\theta}}(j_1)$ of the vector $\boldsymbol{\theta}^*(j_1)$ is defined by minimization of the residual sum of squares,

$$\hat{\boldsymbol{\theta}}(j_1) = \underset{\boldsymbol{\theta}(j_1)}{\operatorname{arginf}} \|Y - \Psi(j_1)\boldsymbol{\theta}(j_1)\|_n^2 = \underset{\{\theta_I \in \mathcal{I}(j_1)\}}{\operatorname{arginf}} \sum_{i=1}^{n}\left(Y_i - \sum_{I\in\mathcal{I}(j_1)} \theta_I \psi_I'(X_i)\right)^2. \tag{3.12}$$

Let $V(j_1)$ be the pseudo-inverse of $\Psi^T(j_1)\Psi(j_1)$,

$$V(j_1) = \left\{\Psi^T(j_1)\Psi(j_1)\right\}^{-}.$$

Then

$$\hat{\boldsymbol{\theta}}(j_1) = V(j_1)\Psi^T(j_1)Y. \tag{3.13}$$

Since the errors ξ_i are normal $\mathcal{N}(0,\sigma^2)$, we obtain via (3.1) that $\hat{\boldsymbol{\theta}}(j_1)$ is a Gaussian vector with the mean $\boldsymbol{\theta}^*(j_1)$ and the covariance matrix $\sigma^2 V(j_1)$,

$$\hat{\boldsymbol{\theta}}(j_1) \sim \mathcal{N}\left\{\boldsymbol{\theta}^*(j_1), \sigma^2 V(j_1)\right\}. \tag{3.14}$$

3.1.3. **Tests**

The proposed testing procedure is based on the fact that for a linear function f, all the coefficients θ_I^*, $I \neq 0, 1$, are zero and therefore, the corresponding estimates $\hat{\theta}_I$ are Gaussian zero mean random variables.

We execute the procedure recursively starting from $j_1 = 0$ until the finest resolution level $j(n)$ defined as

$$j(n) = \lfloor \log_2(n) - 1 \rfloor. \tag{3.15}$$

For each $j_1 \leq j(n)$, let $\hat{\boldsymbol{\theta}}(j_1)$ be defined by (3.12). Denote by $\hat{\boldsymbol{\theta}}_{j_1}$ the part of the vector $\hat{\boldsymbol{\theta}}(j_1)$ corresponding to the level j_1,

$$\hat{\boldsymbol{\theta}}_{j_1} = (\hat{\theta}_I,\, I \in \mathcal{I}_{j_1}).$$

At the step j_1, we analyze the subvector $\hat{\boldsymbol{\theta}}_{j_1}$ only.

Following Spokoiny (1996a), we introduce two kinds of tests: the first one, so called a 'local test', analyses each term $\hat\theta_I$, $I \in \mathcal{I}_{j_1}$, separately; the second one is levelwise, i.e. all the estimates $\hat\theta_I$, $I \in \mathcal{I}_{j_1}$, are used for calculating the corresponding test statistic.

Let $v_{I,I'}$, $I, I' \in \mathcal{I}(j_1)$, be the elements of the matrix $V(j_1) = \{\Psi^T(j_1)\Psi(j_1)\}^-$. Due to (3.14), we have under the null hypothesis $\hat\theta_I \sim \mathcal{N}(0, v_{I,I})$ and hence each variable $v_{I,I}^{-1/2}\hat\theta_I$ is standard normal (if $v_{I,I} > 0$). The local test rejects the null hypothesis if at least one such value exceeds a certain logarithmic level,

$$\phi_{\mathrm{loc}}(j_1) = \mathbf{1}\left(\max_{I\in\mathcal{I}_{j_1}} \sigma^{-1} v_{I,I}^{-1/2} |\hat\theta_I| > \lambda_n\right) \tag{3.16}$$

where

$$\lambda_n = 1 + 2\sqrt{\log n}. \tag{3.17}$$

In the definition (3.16) we use the fact that $v_{I,I} = 0$ implies $\hat\theta_I = 0$, see (3.13), and we assume $0/0 = 0$. Note that both $v_{I,I}$ and θ_I depend on j_1. The local test ϕ_{loc} is very sensitive to functions f containing local fluctuations like jumps or jumps of derivatives.

The next test, which was called a χ^2-test in Ingster (1993) and a L_2-test in Spokoiny (1996b), allows us to detect very small but systematic components. It is based on the standardized sum of squares of $\hat\theta_I$, $I \in \mathcal{I}_{j_1}$. Let V_{j_1} be the submatrix of the matrix $V(j_1)$ corresponding to the level j_1, i.e. $V_{j_1} = (v_{I,I'}, I, I' \in \mathcal{I}_{j_1})$. In view of (3.14), under the null hypothesis, the vector $\hat{\boldsymbol{\theta}}_{j_1}$ is Gaussian zero mean with covariance matrix $\sigma^2 V_{j_1}$. First consider the case when V_{j_1} is non-degenerate, necessary corrections of the procedure for the degenerate case will be discussed later. If $\det V_{j_1} \neq 0$, then under the null, the vector $\boldsymbol{\eta}_{j_1} = (\eta_I, I \in \mathcal{I}_{j_1})$, defined as the standardization of $\hat{\boldsymbol{\theta}}_{j_1}$,

$$\boldsymbol{\eta}_{j_1} = \sigma^{-1} V_{j_1}^{-1/2} \hat{\boldsymbol{\theta}}_{j_1}, \tag{3.18}$$

is standard normal. Define χ^2-type statistics

$$S_{j_1} = \|\boldsymbol{\eta}_{j_1}\|^2 = \sum_{I\in\mathcal{I}_{j_1}} \eta_I^2. \tag{3.19}$$

For each $f \in \mathcal{F}_0$ (i.e. for a linear f), the distribution of S_{j_1} does not depend on f and we denote by $\boldsymbol{E}_0$ and $\boldsymbol{D}_0$ the corresponding expectation and variance. Clearly

$$\begin{aligned} \boldsymbol{E}_0 S_{j_1} &= N_{j_1}, \\ \boldsymbol{D}_0 S_{j_1} &= \boldsymbol{E}_0 (S_{j_1} - \boldsymbol{E}_0 S_{j_1})^2 = 2N_{j_1}. \end{aligned}$$

This leads finally to the test statistic T_{j_1}

$$T_{j_1} = \frac{S_{j_1} - \boldsymbol{E}_0 S_{j_1}}{\sqrt{\boldsymbol{D}_0 S_{j_1}}} = (2N_{j_1})^{-1/2}(S_{j_1} - N_{j_1}). \tag{3.20}$$

Each T_{j_1} has under the null the normalized χ^2-distribution with N_{j_1} degrees of freedom which is for large N_{j_1} approximated by the standard normal distribution. We define

therefore

$$\phi_{\chi^2}(j_1) = \mathbf{1}\left(|T_{j_1}| > \lambda_n\right) \tag{3.21}$$

with the same λ_n as above.

In the case when $\det V_{j_1} = 0$, denote by $V_{j_1}^-$ the pseudo-inverse of V_{j_1} and set

$$S_{j_1} = \sigma^{-2}\hat{\boldsymbol{\theta}}_{j_1}^T V_{j_1}^- \hat{\boldsymbol{\theta}}_{j_1}, \tag{3.22}$$

$$N'_{j_1} = \mathrm{tr}(V_{j_1}^- V_{j_1}). \tag{3.23}$$

Then S_{j_1} is again a χ^2-statistic, but now with N'_{j_1} degrees of freedom. We take therefore the test statistic again of the form (3.20) with N'_{j_1} in place of N_{j_1},

$$T_{j_1} = (2N'_{j_1})^{-1/2}(S_{j_1} - N'_{j_1}). \tag{3.24}$$

The representation (3.21) for the test $\phi_{\chi^2}(j_1)$ remains valid with such defined T_{j_1}. Finally we reject the linear hypothesis H_0 if one of $\phi_{\mathrm{loc}}(j_1)$ or $\phi_{\chi^2}(j_1)$ does,

$$\phi^* = \max_{0 \le j_1 \le j(n)} \max\{\phi_{\mathrm{loc}}(j_1), \phi_{\chi^2}(j_1)\}. \tag{3.25}$$

3.2. **Test procedure for** $d > 1$

The method of testing is essentially the same as in the univariate case and it is based on the decomposition of each component f_m from (2.1) by the Haar basis,

$$f_m(t) = \sum_{I \in \mathcal{I}} c_{I,m}\psi_I(t), \qquad m = 1, \ldots, d.$$

For the additive model (2.2) this gives for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$

$$F(x) = \sum_{m=1}^{d} \sum_{I \in \mathcal{I}} c_{I,m}\psi_I(x_m).$$

We proceed as above for the univariate case by replacing the infinite decomposition by a finite approximation. Let us fix a level j_1 for the first component and a level $j(n)$ for the remaining ones, and let $\mathcal{I}(j_1)$ be due to (3.5), $\mathcal{I}(j_1) = \{0,1\} + \bigcup_{j \le j_1} \mathcal{I}_j$. We approximate $F(x)$ by

$$\sum_{I \in \mathcal{I}(j_1)} c_{I,1}\psi_I(x_1) + \sum_{m=2}^{d} \sum_{j=0}^{j(n)} \sum_{I \in \mathcal{I}_j} c_{I,m}\psi_I(x_m).$$

We use here $N = 2^{j(n)+1} - 1$ coefficients for each component f_m, $m \ge 2$, and, assuming that $j_1 \le j(n)$, the total number $N(d, j_1)$ of coefficients is at most $Nd + 2$. Modify now the definition of $j(n)$ from the one-dimensional case to provide $N(d, j_1) \le n$ that leads to the choice

$$j(n) = \lfloor \log_2(n/d) - 1 \rfloor. \tag{3.26}$$

Let now some $j_1 \le j(n)$ be fixed. Denote by $\mathcal{I}(d, j_1)$ the index set

$$\mathcal{I}(d, j_1) = \{(I, 1), I \in \mathcal{I}(j_1), (I, m), I \in \mathcal{I}_j, 0 \le j \le j(n), m = 2, \dots, d\}$$

and let

$$N(d, j_1) = N(j_1) + (d-1)N = 2^{j_1+1} + (d-1)2^{j(n)+1} - d + 2$$

be the number of elements in $\mathcal{I}(d, j_1)$.

The next step is to renormalize the basis functions ψ_I. Define $\lambda_{I,m}$ by

$$\lambda_{I,m}^2 = \sum_{i=1}^{n} |\psi_I(X_{i,m})|^2, \qquad I \in \mathcal{I},\ m = 1, \dots, d, \tag{3.27}$$

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

Set $\Psi(d, j_1)$ for the $n \times N(d, j_1)$ matrix with elements $\psi'_{i,(I,m)} = \lambda_{I,m}^{-1} \psi_I(X_{i,m})$, $i = 1, \dots, n$, $(I, m) \in \mathcal{I}(d, j_1)$, and define the vector $\boldsymbol{\theta}^*(d, j_1)$ with elements $\theta^*_{I,m}$, $(I, m) \in \mathcal{I}^{(d)}(j_1)$ as a solution to the quadratic problem

$$\boldsymbol{\theta}^*(d, j_1) = \operatorname*{arginf}_{\boldsymbol{\theta}(d,j_1)} \|\boldsymbol{F} - \Psi(d, j_1)\boldsymbol{\theta}(d, j_1)\|_n^2. \tag{3.28}$$

This leads to the same representation for $\boldsymbol{\theta}^*(d, j_1)$ as in the univariate case, $\boldsymbol{\theta}^*(d, j_1) = V(d, j_1)\Psi^T(d, j_1)\boldsymbol{F}$. Again the matrix $V(d, j_1)$ is to be understood as the pseudo-inverse of $\Psi^T(d, j_1)\Psi(d, j_1)$, $V(d, j_1) = \{\Psi^T(d, j_1)\Psi(d, j_1)\}^-$.

Given data $Y = (Y_1, \dots, Y_n)$, we estimate $\boldsymbol{\theta}^*(d, j_1)$ by the least squares method,

$$\hat{\boldsymbol{\theta}}(d, j_1) = V(d, j_1)\Psi^T(d, j_1)Y. \tag{3.29}$$

Next, for testing the first component f_1, we proceed in the same line as for the univariate case making use of the estimates $\hat{\boldsymbol{\theta}}_{j_1} = (\hat{\theta}_{I,1}, I \in \mathcal{I}_{j_1})$ and the submatrix $V_{j_1} = (v_{(I,1),(I',1)}, I, I' \in \mathcal{I}_{j_1})$ of the covariance matrix $V(d, j_1)$. For the level j_1, the local test $\phi_{\mathrm{loc}}(j_1)$ is defined by

$$\phi_{\mathrm{loc}}(j_1) = \mathbf{1}\left(\max_{I \in \mathcal{I}_{j_1}} \sigma^{-1} v_{(I,1),(I,1)}^{-1/2} |\hat{\theta}_{I,1}| > \lambda_n\right)$$

and the χ^2-test $\phi_{\chi^2}(j_1)$ has the form (3.21),

$$\phi_{\chi^2}(j_1) = \mathbf{1}\left(|T_{j_1}| > \lambda_n\right)$$

with T_{j_1} due to (3.22) through (3.24). The test ϕ^* is again the combination of all local and χ^2-tests for $j_1 = 0, 1, \dots, j(n)$,

$$\phi^* = \max_{0 \le j_1 \le j(n)} \max\{\phi_{\mathrm{loc}}(j_1), \phi_{\chi^2}(j_1)\}. \tag{3.30}$$

3.3. Level dependent thresholds

The above described procedure is essencially levelwise. However we apply for each level j and for both subtests $\phi_{\mathrm{loc}}(j)$ and $\phi_{\chi^2}(j)$ the same test threshold λ_n. This was done for the sake of simplicity of presentation. In general it is possible to apply different thresholds for different subtests. Moreover, the simulation studies show (see Section 5) that the proposed test with the universal threshold is too conservative and an application of different thresholds is reasonable.

The following threshold values can be used:

$$\lambda_{\mathrm{loc}}(j) = 1 + \sqrt{2\log(N_j) + a_{\mathrm{loc}}\log\log(n)},$$
$$\lambda_{\chi^2}(j) = 2^{-1/2} + \sqrt{N_j/2n} + \sqrt{a_{\chi^2}\log\log(n)}$$

with arbitrary constants $a_{\mathrm{loc}} \geq 2$ and $a_{\chi^2} \geq 2$. Hence we set

$$\phi_{\mathrm{loc}}(j_1) = \mathbf{1}\left(\max_{I\in\mathcal{I}_{j_1}} \sigma^{-1} v^{-1/2}_{(I,1),(I,1)} |\hat\theta_{I,1}| > \lambda_{\mathrm{loc}}(j_1)\right),$$
$$\phi_{\chi^2}(j_1) = \mathbf{1}\left(|T_{j_1}| > \lambda_{\chi^2}(j_1)\right)$$

with the above defined test statistics $\hat{\boldsymbol{\theta}}$ and T_{j_1} and then apply the combined test (3.30). It can be seen by inspecting the proofs that all the results formulated for the original test procedure remain valid for this modified test.

3.4. The case of unknown variance of errors

In the above procedure we assumed that the variance σ^2 of errors ε_i is unknown. Here we shortly indicate the necessary modification of the procedure when σ^2 is unknown. We apply the standard approach by making use of a pilot estimator of σ^2.

Let $j(n)$ be defined in (3.26). Due to this definition we have $n/2 \leq d2^{j(n)+1} \leq n$. We suppose that $n - d2^{j(n)+1} \geq n/4$, otherwise the value $j(n)$ can be reduced to $j(n) - 1$.

Let $\hat{\boldsymbol{\theta}}_n$ be the least square estimator from (3.29) of the vector $\boldsymbol{\theta}^*_n = \boldsymbol{\theta}^*(d, j_1)$ with $j_1 = j(n)$. Denote also $V_n = V(d, j(n))$ and $\Psi_n = \Psi(d, j(n))$. Then $V_n = \left(\Psi_n^T\Psi_n\right)^-$ and

$$\hat{\boldsymbol{\theta}}_n = V_n\Psi_n^T Y.$$

We know that the vector $\hat{\boldsymbol{\theta}}_n$ has the mean $\boldsymbol{\theta}^*_n$ and the covariance matrix $\sigma^2 V_n$. Moreover, the vector $\boldsymbol{\theta}^*_n$ is the solution of the optimization problem (3.28) and hence $\Psi_n\boldsymbol{\theta}^*_n$ corresponds to the best approximation of the regression function F by the Haar sum with the highest level $j(n)$. Under usual regularity condition on this function F the accuracy of approximation tends to zero as $j(n)$ tends to infinity in the sense that

$$n^{-1}\|F - \Psi_n\boldsymbol{\theta}^*_n\| \to 0, \qquad n \to \infty.$$

This consideration prompts to use the value

$$\Sigma_n = \|Y - \Psi_n \hat{\boldsymbol{\theta}}_n\|_n^2$$

for estimating σ^2. We have

$$\begin{aligned} E\Sigma_n &= E\|Y - \Psi_n V_n \Psi_n^T Y\|_n^2 \\ &= E\|(I_n - \Psi_n V_n \Psi_n^T)F + (I_n - \Psi_n V_n \Psi_n^T)\varepsilon\|_n^2 \\ &= \|F - \Psi_n \boldsymbol{\theta}_n^*\|_n^2 + E\|(I_n - \Psi_n V_n \Psi_n^T)\varepsilon\|_n^2 \end{aligned}$$

where I_n denotes the identity $n \times n$-matrix. Next, since the errors ε_i are independent zero mean random variables with $E\varepsilon_i^2 = \sigma^2$, we obtain by straightforward calculation

$$E\|(I_n - \Psi_n V_n \Psi_n^T)\varepsilon\|_n^2 = \sigma^2(n - \operatorname{tr} V_n V_n^-).$$

The rank of the matrix V_n is at most $d2^{j(n)+1}$ and due to our assumption it holds $n - d2^{j(n)+1} \geq n/4$ that gives $n - \operatorname{tr} V_n V_n^- \geq n/4$. We set

$$\hat{\sigma}^2 = \Sigma_n/(n - \operatorname{tr} V_n V_n^-).$$

It can be shown that $\hat{\sigma}^2$ is a consistent and even root-n consistent estimator of σ^2.

Using this estimator, we define the test procedure in the same line as before replacing σ by its estimate $\hat{\sigma}$.

4. Main results

In this section we study asymptotic properties of our testing procedure. We state the results on the error probabilities of the first and of the second kind separately since we evaluate them under different assumptions on the design variables. The result on the error probabilities $\alpha_{F_0}(\phi^*)$ of the first kind is valid under mild assumptions on the design. 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 (2.2), the remaining components $f_2, \ldots, f_d$ 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 $2 \leq m \leq d$ 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 $X_1, \ldots, X_n$.

Recall that we identify every function g on $\mathbb{R}^d$ with the vector $(g(X_i), i = 1, \ldots, n)$ in $\mathbb{R}^n$. In particular, each f_m is identified with $(f_m(X_{i,m}), i = 1, \ldots, n)$ and $\psi'_{I,m}$ is understood as the vector with the elements $\lambda_{I,m}^{-1}\psi_I(X_{i,m})$. Recall also the notation $\|g\|_n^2 = \sum_{i=1}^n g^2(X_i)$.

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

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

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 $\Pi_m(j+1) f_m$ we denote the projection of f_m onto $\mathcal{L}_m(j)$ w.r.t. the distance $\|\cdot\|_n$,

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

We write also $\Pi_{m,n}$ for $\Pi_m(j(n)+1)$.

In our results we suppose the following condition to be fulfilled:

Condition (D) **:** *For n sufficiently large*

$$\sum_{m=2}^{d} \|f_m - \Pi_{m,n} f_m\|_n \le \sigma n^{-1/4}.$$

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

Lemma 4.1. *Let $\mu_{n,m}$ be the m-th marginal of the empirical design measure μ_n,*

$$\mu_{n,m}(A) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(X_{i,m} \in A), \qquad m = 2, \ldots, d.$$

Let also there be a constant C_1 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

$$\|f_m - \Pi_{m,n} f_m\|_n \le C n^{-1/2}$$

with C depending on C_1 and C_2 only and condition (D) is fulfilled for n large enough.

Another situation when the difference $\|f_m - \Pi_{m,n} f_m\|_n$ can be easily estimated, is in the case of a discrete m-th component (i.e. when all $X_{i,m}$ belong to some finite set).

Let ϕ^* be the test introduced above in (3.30).

Theorem 4.1. *Suppose that the observations* (X_i, Y_i), $i = 1, \dots, n$, *obey the regression model (2.1) and (2.2), and let condition* (D) *hold. If the first component* f_1 *of the function* F *is linear, then*

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

where $\delta_1(n)$ *depends on* n *only and* $\delta_1(n) \to 0$ *as* $n \to \infty$.

The proof of the theorem is given in Section 6.

Now we state the results concerning the sensitivity of the proposed test ϕ^*. The first assertion shows under which conditions we detect an alternative with a high probability. Then we discuss how these conditions can be transferred into precise statement about the rate of testing.

Proposition 4.1. *Let the function* F *in model (2.1) be of the form (2.2). Let also* $\boldsymbol{\theta}_j^* = (\theta_{I,1}^*,\, I \in \mathcal{I}_j)$ *be the subvector of the vector* $\boldsymbol{\theta}^*(d,j)$ *from (3.28) 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 corresponding covariance submatrix. If, for some* $j \le j(n)$, *it holds*

$$T_j^* \equiv 2^{-(j+1)/2}\sigma^{-2}\boldsymbol{\theta}_j^{*T} V_j^{-1}\boldsymbol{\theta}_j^* > 2\lambda_n,$$

with $\lambda_n = 1 + 2\sqrt{\log n}$, *then*

$$\boldsymbol{P}_F\{\phi_{\chi^2}(j) = 0\} \le \delta(n) \to 0, \qquad n \to \infty,$$

where $\delta(n)$ *depends on* n *only.*

If, for some $j \le j(n)$, *it holds*

$$T_{j,\infty}^* \equiv \max_{I \in \mathcal{I}_j} \sigma^{-1} v_{(I,1),(I',1)}^{-1/2} |\theta_{I,1}^*| > 2\lambda_n,$$

then

$$\boldsymbol{P}_F\{\phi_{\mathrm{loc}}(j) = 0\} \le \delta(n) \to 0,$$

with the same $\delta(n)$.

This proposition says that the test ϕ^* detects with a probability close to one any alternative for which at least one from the corresponding values T_j^* and $T_{j,\infty}^*$ exceeds the level $2\lambda_n$. Therefore, we may suppose that the error of the second kind may occur only if

$$T_j^* \le 2\lambda_n, \qquad 0 \le j \le j(n), \tag{4.1}$$

$$T_{j,\infty}^* \le 2\lambda_n, \qquad 0 \le j \le j(n). \tag{4.2}$$

It remains to understand what follows for the first component f_1 of the function F from these inequalities. To this end we impose some regularity conditions on the design and smoothness conditions on the first component f_1 of the function F.

The reason why we need stronger conditions on the design 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 one of them.

Smoothness (or regularity) conditions on a function f can be formulated in a different forms. We find it convenient to define them in terms of accuracy of approximation of this function by piecewise polynomials of certain degree. Given $j \le j(n)$, denote by $\{A_I, I \in \mathcal{I}_j\}$ the partition of the interval $[0,1]$ into intervals of the length 2^{-j}: if $I=(j,k)$ then $A_I = [k2^{-j}, (k+1)2^{-j})$. Next, for an integer s, define $\mathcal{P}_s(j)$ as the set of piecewise polynomials of degree $s-1$ on the partition $\{A_I\}$ i.e. every function g from $\mathcal{P}_s(j)$ coincides on each A_I with a polynomial $a_0 + a_1 x + \ldots + a_{s-1}x^{s-1}$ where the coefficients $a_0, \ldots, a_{s-1}$ may depend on I. Now the condition that a function f has regularity s can be understood in the sense that this function is approximated by functions from $\mathcal{P}_s(j)$ with the rate 2^{-js}, or, in other words, the distance from the function f_1 to the linear space $\mathcal{P}_s(j)$ can be bounded by $C2^{-js}$ with some positive constant C depending on s only.

Let now a function F with the structure from (2.2) be fixed and let f_1 be the first component. Let also j_0 be such that $2^{j_0-1} \ge s$. Set for $j \ge j_0$

$$r_s(j) = \inf_{g \in \mathcal{P}_s(j-j_0)} \|f_1 - g\|_n = \inf_{g \in \mathcal{P}_s(j-j_0)} \left\{ \sum_{i=1}^{n} |f_1(X_{i,1}) - g(X_{i,1})|^2 \right\}^{1/2}. \tag{4.3}$$

The quantity $r_s(j)$ characterizes the accuracy of approximation of f_1 by piecewise polynomials. In our procedure, we use the Haar approximation which corresponds to the case of a locally constant approximation with $s=1$.

In order to state our next results we need to define regularity characteristics of the design $X_1, \ldots, X_n$. Set

$$u_*(j) = \inf_{I \in \mathcal{I}_j} 2^j M_I/n, \tag{4.4}$$

$$u^*(j) = \sup_{I \in \mathcal{I}_j} 2^j M_I/n, \tag{4.5}$$

with $M_I = \#\{i : X_{i,1} \in A_I\}$. *Design regularity* means in particular that $u_*(j)$ is bounded away from zero i.e. each interval A_I contains enough design points $X_{i,1}$ and this satisfies the condition in Lemma 4.1.

Let also V_j be the submatrix of $V(d,j) = \{\Psi^T(d,j)\Psi(d,j)\}^-$ corresponding to the first component, see Subsection 3.2, $V_j = \left(v_{(I,1),(I',1)}, I, I' \in \mathcal{I}_j\right)$. Clearly V_j is a $N_j \times N_j$-matrix, $N_j = 2^j$. Set

$$v^*(j) = \|V_j\|, \tag{4.6}$$

Here the norm $\|A\|$ of a matrix A is understood as the maximal eigenvalue of this matrix or equivalently, $\|A\| = \sup_{\gamma:\|\gamma\|=1} \|A\gamma\|$ where the sup is taken over $\gamma \in R^{N_j}$ and $\|\gamma\|^2 = \gamma_1^2 + \ldots + \gamma_{N(j)}^2$. We may therefore define $v^*(j)$ as the maximal eigenvalues

of V_j. We shall understand design regularity in the sense that V_j is non-degenerate and $v^*(j)$ are bounded for sufficiently large j. Note that the values $v^*(j)$, $u_*(j)$ and $u^*(j)$ are related to each other: the regularity conditions in terms of $v^*(j)$ are stronger than in terms of $u_*(j)$ and $u^*(j)$. Indeed, $u_*(j)$ and $u^*(j)$ characterize only the properties of the first marginal of the design, whereas $v^*(j)$ tells us additionally about identifiability of the first component.

Theorem 4.2. *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*

$$\inf_{a,b} \|f_1 - a - b\psi_{1,1}\|_n^2 \ge C_1 r_s^2(j) + C_2 \frac{u^*(j)}{u_*(j)} v^*(j) 2^{j/2} \sigma^2 \lambda_n \tag{4.7}$$

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

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

where $\delta(n)$ is from Proposition 4.1.

Remark 4.1. It is of interest to compare this Theorem with other results on the rate of hypothesis testing. It was shown in Ingster (1982) that if f belongs to a Sobolev ball $W_s(1)$ with

$$W_s(1) = \left\{ f : \int_0^1 |f^{(s)}(x)|^2 dx \le 1 \right\},$$

$f^{(s)}$ being s-th derivative of f, then the optimal rate of testing is $n^{-2s/(4s+1)}$ and it is achieved by a testing procedure which makes explicit use of knowledge of s.

Our procedure is adaptive i.e. we do not need to know s. Next, the condition $f_1 \in W_s(1)$ yields $n^{-1/2} r_s(j) \le C 2^{-js}$ and, if the design is regular (that means that all $v^*(j)$ are bounded), then the optimization over j in the right hand-side of (4.7) gives the rate $\left(\frac{n}{\sqrt{\log n}}\right)^{-2s/(4s+1)}$ for the deviation of the function f_1 from the space of linear functions. Therefore, our procedure provides the near optimal rate of testing by the logarithmic factor $(\log n)^{s/(4s+1)}$. It was shown in Spokoiny (1996a) that the optimal adaptive rate differs from the non-adaptive one by the factor $(\log \log n)^{s/(4s+1)}$.

Remark 4.2. The result of Theorem 4.2 helps to understand what happens in the case when our design is not regular and, for instance, $u_*(j) = 0$ for all large j. It was already mentioned that the procedure can be applied in this situation too and the probability of the error of the first kind is very small. Concerning the error probability of the second kind, the inspection of the proof shows that design irregularity decreases the sensitivity of our procedure in the following sense: there exist smooth alternatives with probably large L_2-norm which are not detected. Such an alternative f_1 deviates from the best linear approximation only in the domain where there are very few design points or where an identification problem between f_1 and the remaining components is met.

Remark 4.3. It is also worth speaking about the relation between smoothness conditions on the additive model components $f_1, \ldots, f_d$ and regularity conditions on the design. Namely, the first term in the right hand side of (4.7) depends only on smoothness properties of the function f_1 and the second one depends only on design regularity. Therefore, if we would like to keep the same sensitivity for a less regular component function we need more regular design and vice versa.

The result of Theorem 4.2 is formulated for the case when smoothness properties of the function f_1 are measured in the L_2-norm. An inspection of the proof shows that in this situation it suffices to apply only the test ϕ_{χ^2} which exactly corresponds to testing in L_2-norm. Lepski and Spokoiny (1995) and Spokoiny (1996b) have shown that if smoothness properties of the tested function are measured in some norm L_p with $p < 2$, then the testing procedure has to be modified to attain the optimal rate of testing which is $n^{-(2sp-1+p/2)/(2sp-1+p)}$ in this situation. The latter case of $p < 2$ corresponds to a function f with heterogeneous smoothness properties, in particular, when this function has jumps or jumps of derivatives. We conclude by stating one more result related to this situation.

Given $j \le j(n)$, let $\boldsymbol{\theta}^*(d,j) = (\theta^*_{I,m}, (I,m) \in \mathcal{I}(d,j))$ be due to (3.28) and let $\boldsymbol{\theta}^*_j = (\theta^*_{I,1}, I \in \mathcal{I}_j)$ be the subvector tested at j-th step. The test $\phi_{\chi^2}(j)$ is sensitive when $\|\boldsymbol{\theta}^*_j\|^2 \ge C\lambda_n 2^j$, see Proposition 4.1 and Lemma 6.2 below. At the same time, the test $\phi_{\mathrm{loc}}(j)$ is sensitive in the situation when at least one coefficient $\theta^*_{I,1}$ is greater than $C'\lambda_n$. This means that it is reasonable to apply the test ϕ_{loc} when the majority of coefficients from level j are small and a few of them are of order λ_n. This corresponds exactly to the case of a function with nonhomogeneous smoothness properties, e.g. to a function with jumps.

Set

$$r(j,t) = \sum_{I \in \mathcal{I}_j} |\theta^*_{I,1}|^2 \mathbf{1}(|\theta^*_{I,1}| \le t). \tag{4.8}$$

We exploit the fact that under some regularity conditions on the first component f_1 of F, the value $r(j,t)$ is small for j large enough.

Theorem 4.3. *Let condition (D) be satisfied. If, for some integer s and some $j_1 \le j_2 \le j(n)$, the first component f_1 of the model function F satisfies the following inequality*

$$\inf_{a,b} \|f_1 - a - b\psi_{1,1}\|_n^2 \ge C_1 r_s^2(j_2) + \frac{v^*(j_2)u^*(j_2)}{u_*(j_2)} \left\{ C_2 2^{j_1/2}\sigma^2\lambda_n + C_3 \sum_{j=j_1}^{j_2} r(j,t_j) \right\}$$

with $t_j = 2\sigma\lambda_n\sqrt{v^*(j)}$ and some positive constants C_1, C_2 and C_3 depending on s only, then

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

with the same $\delta(n)$ *as in Proposition 4.1.*

As a corollary of the last result, we show that our testing procedure provides a near optimal rate of testing over Sobolev balls $W_{s,p}(1)$ with $p < 2$ and $s \leq 1$,

$$W_{s,p}(1) = \left\{ f : \int_0^1 |f^{(s)}(x)|^p dx \leq 1 \right\}.$$

Indeed, it is well known, see Triebel (1992), that if $f \in W_{s,p}(1)$ for $s \leq 1$ and if the design is regular, then

$$n^{-1} \sum_{i=1}^{n} |f_j(X_i)|^p \leq C 2^{-jsp}.$$

This gives for a regular design

$$r(j, t_j) \leq C n 2^{-jsp} \left(n^{-1} 2^j 4\sigma^2 \lambda_n^2 \right)^{1-p/2} = C' n^{p/2} \lambda_n^{2-p} 2^{-j(sp-1+p/2)}.$$

Since $sp - 1 + p/2 > 0$ for every integer s and $p \geq 1$, we obtain

$$\sum_{j=j_1}^{j_2} r(j, t_j) \leq C'' n^{p/2} \lambda_n^{2-p} 2^{-j_1(sp-1+p/2)}.$$

Now we select j_1 to minimize the sum $C_2 2^{j_1/2} \lambda_n + C'' n^{p/2} \lambda_n^{2-p} 2^{-j_1(sp-1+p/2)}$ that leads to the rate $n^{-(2sp-1+p/2)/(2sp-1+p)} \lambda_n^{2(p-1)/(2sp-1+p)}$ which is near optimal by the logarithmic factor $\lambda_n^{2(p-1)/(2sp-1+p)}$.

5. Simulation studies and applications

The behaviour of the suggested test procedure for finite samples has been examined in a small simulation study and then the test was applied for analysis of econometric data.

5.1. A simulated example

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

$$\begin{aligned} f_a(x) &= 2\sin(-2x), \\ f_b(x) &= 2\sin(-5x), \\ f_c(x) &= x + 4\ \mathbb{1}\{x \in (-0.15, 0)\} - 4\ \mathbb{1}\{x \in (0, 0.15)\}, \\ f_d(x) &= x^2 - E[x^2]. \end{aligned}$$

The models we chose were

- Model 1 : $F_1(x) = f_a(x_1) + f_c(x_2) + f_d(x_3)$
- Model 2 : $F_2(x) = f_b(x_1) + f_c(x_2) + f_d(x_3)$
- Model 3 : $F_3(x) = f_c(x_1) + f_a(x_2) + f_d(x_3)$,

see also Figure 1. Here f_a and f_d are fairly smooth functions that could also be approximated by a polynomial approach, whereas for f_b and f_c application of wavelets seems to be reasonable.

In our simulations we assumed the uniform random design on the cube $[-2,2]^3$ and standard normal errors. We do not assume to know the standard deviation of the error terms and therefore estimate σ as suggested in section 3. 4.

We apply the modified test procedure from Section 3.4 with the level-dependent thresholds $\lambda_{loc}(j) = 1 + \sqrt{2\log(N_j) + a_{\rm loc}\log\log(n)}$ and $\lambda_{\chi^2,j} = \frac{1}{\sqrt{2}} + \sqrt{\frac{N_j}{2n}} + \sqrt{a_{\chi^2}\log\log(n)}$, where we took $a_{\rm loc} = 2$ and $a_{\chi^2} = 4$.

First we compared the performance of the test for different numbers of observations $n = 200$ and $n = 400$. Second, we investigated the relative efficiency of our test by calculating the power functions of the competing optimal parametric Likelihood Ratio Test (LR–test) and our nonparametric test procedure. In practice the parametric LR–test can only be applied if we already knew the functional forms of the additive components.

We further observed at what level j_1 the test rejects the hypothesis H_0. Finally we compared the partial tests $\varphi_{\rm loc}$ and φ_{χ^2}.

In Figure 1 we have displayed the functions f_a through f_d, the responses (an example with $n = 400$) and estimates. The solid lines are the data generating functions, the points are the Y's and the dashed lines are wavelet estimates, using the Haar basis and the highest possible level $j(n) = 5$. This value was calculated due to the rule $j(n) = [\log_2(n/d)] - 1$ under the constraint $n - d2^{j(n)+1} \geq n/4$ which leads for $d = 3$ and $n = 400$ exact to $j(n) = 5$.

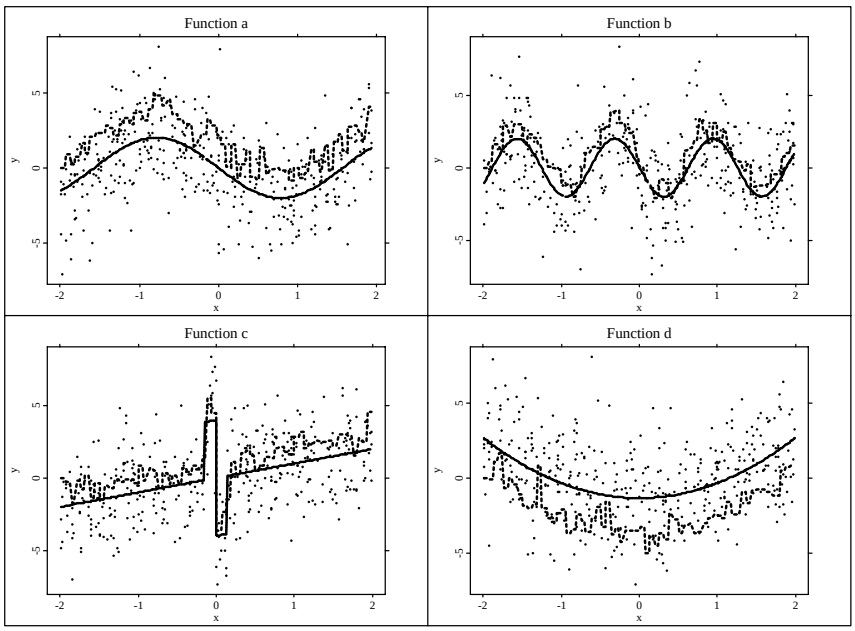

Figure 1: *The functions and the estimates.*

Note that these wavelet estimates are not to be seen as smoothers in this context. But our goal is in the structural analysis of the components rather than their estimation.

To get power functions for our test procedure we considered the following data generating process. For Model 1

$$Y = (1-v)x_1 + v f_a(x_1) + f_c(x_2) + f_d(x_3) + \varepsilon, \quad v \in [0,1] \tag{5.1}$$

where ε means a standard normal error. We compare the sensitivity of our test procedure with the LR–test designed for this specific form of the alternative: the parametric hypothesis $H_0 : v = 0$ is tested versus the parametric alternative $H_1 : v \neq 0$. For the case of Gaussian errors this test is based on the statistics T_n of the form $T_n = n^{-1/2} \sum_i \left\{ |Y_i - X_{i,1} - f_c(X_{i,2}) - f(X_{i,3})|^2 - 1 \right\}$: we reject the null hypothesis if T_n exceeds the proper quantile of the standard normal law. For each v we pick the threshold value for LR-test to provide the same value of the error probability of the first kind and compare the error probabilities of the second kind. Similarly we proceeded for Models 2 and 3.

In Figure 2 we see the power functions for Model 1 and Model 2 for a sample size of $n = 200$. The solid lines are for the LR–test, the dashed ones for the nonparametric procedure. For $n = 200$ we have $j(n) = 5$.

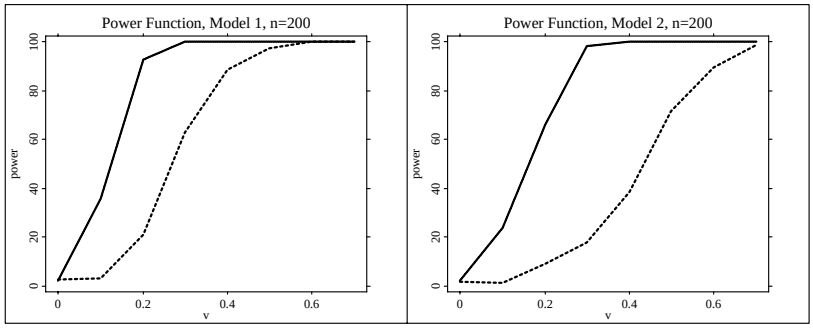

Figure 2: *The power functions for $n = 200$ observations.*

In Figure 3 we display the corresponding power functions for a sample size $n = 400$. Looking at the scale of v we recognize that for $n = 400$ the power functions are steeper than for $n = 200$. This fact is not surprising but it can also be seen that the relative efficiency of the nonparametric procedure stays almost the same. As we expected the efficiency in Model 1 is very high whereas in Model 2, where we consider a sin function with high frequency, the distance between parametric and nonparametric power fits is even for $n = 400$ rather spread. This fact is in agreement with the general results on nonparametric hypothesis testing, see e.g. Ingster (1982, 1993) where it has been shown that optimal nonparametric rate of testing is $n^{-2s/(4s+1)}$ is worse than the parametric rate $n^{-1/2}$.

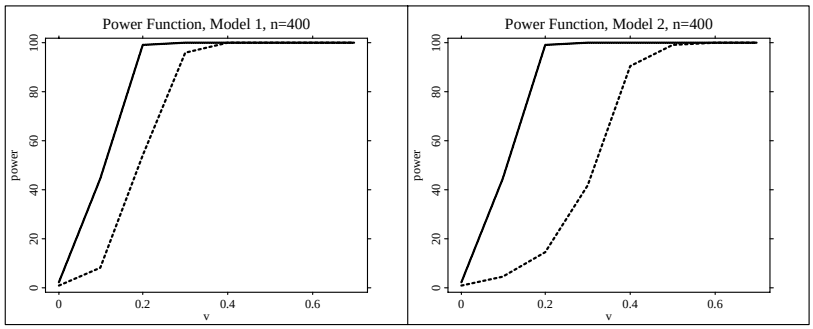

Figure 3: *The power functions for* $n = 400$ *observations.*

Our test procedure is levelwise and it is natural to expect that different levels play different roles for different model functions. The next two tables show at which level our test detects firstly the alternative for the Models 1 and 2 and different values of v in (5.1). We see in Table 1 that the very smooth but nonlinear function f_a is detected typically at lowest level with $j = 0$. Similar simulation results for the case $f_1 = f_b$ are presented in Table 2. Here we recognize that higher levels become more important. The simulation were carried our for $n = 400$ observations. The last column of each table is displayed in Figure 3.

		level j						
		0	1	2	3	4	5	sum
	0.0	0.33	0.33	0.00	0.00	0.33	0.00	1.0
	0.1	6.67	0.00	0.00	0.00	0.67	0.67	8.0
	0.2	52.67	0.00	1.33	0.00	0.00	0.00	54.0
v	0.3	95.33	0.00	0.67	0.00	0.00	0.00	96.0
	0.4	100.00	0.00	0.00	0.00	0.00	0.00	100.0
	0.5	100.00	0.00	0.00	0.00	0.00	0.00	100.0
	0.6	100.00	0.00	0.00	0.00	0.00	0.00	100.0

TABLE 1: Percentage of rejection in Model 1

		level j						
		0	1	2	3	4	5	sum
	0.0	0.00	0.00	0.00	0.33	0.67	0.00	1.0
	0.1	2.00	0.67	1.33	0.00	0.00	0.67	4.7
	0.2	3.33	0.67	6.00	2.67	0.67	1.33	14.7
	0.3	4.67	4.00	28.00	4.67	0.67	0.00	42.0
v	0.4	14.00	10.00	60.00	6.00	0.00	0.67	90.7
	0.5	23.33	24.00	48.67	3.33	0.00	0.00	99.3
	0.6	36.67	28.00	35.33	0.00	0.00	0.00	100.0
	0.7	46.67	28.67	24.67	0.00	0.00	0.00	100.0
	0.8	62.67	22.00	15.33	0.00	0.00	0.00	100.0

TABLE 2: Percentage of rejection in Model 2

Finally we have done the same simulation study for the Model 3 whose first component has jumps. But now we investigate separately the tests ϕ_{loc} and ϕ_{χ^2}. The results are given in Table 3.

	φ_{loc}							φ_{χ^2}						
level	0	1	2	3	4	5	sum	0	1	2	3	4	5	sum
0.0	0.0	0.7	0.7	0.0	0.0	0.0	1.0	0.0	0.3	0.0	0.0	0.0	0.0	0.3
0.1	1.3	0.7	0.0	0.7	0.0	0.0	2.7	0.7	0.7	0.0	0.0	0.0	0.0	1.4
0.2	13.3	1.3	0.0	0.7	0.0	0.0	15.3	8.7	0.0	0.7	0.0	0.0	0.0	9.4
0.3	42.0	1.3	1.3	1.3	0.0	0.0	46.0	37.3	0.0	0.7	1.3	0.0	0.0	39.3
v 0.4	64.0	0.0	1.3	1.3	1.3	0.7	68.7	58.7	0.0	2.0	0.0	0.7	0.0	61.4
0.5	91.3	0.7	0.0	0.0	0.7	0.0	92.7	88.7	0.7	0.7	1.3	0.0	0.0	91.4
0.6	94.7	0.7	0.0	0.0	2.7	0.0	98.0	94.0	0.7	0.7	0.7	0.7	0.0	96.7
0.7	98.0	0.0	0.0	0.0	2.0	0.0	100	97.3	0.0	0.0	1.3	0.0	0.0	98.6
0.8	100	0.0	0.0	0.0	0.0	0.0	100	99.3	0.0	0.0	0.7	0.0	0.0	100

TABLE 3: Percentage of rejection in Model 3 for φ_{loc} and φ_{χ^2} separately

Here we can see that the local test has better performance than the test ϕ_{χ^2}. The special role of the fourth level for the test ϕ_{loc} is due to the fact that the length of support of the corresponding wavelet function is of the same order as the length of the bump.

5.2. **Applications**

We now turn to an application to demonstrate the performace of our procedure on real data. The data set is a subsample of the Socio Economic Panel of Germany from 1992. To study the female labour supply in East Germany, 607 women with job and living together with a partner in East Gemany 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 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.

Since the realizations of these explanatory variables took very different numbers of values, e. g. they have maximal 5 children but there are more than 500 different wages in this sample, it made no sense to take one $j(n)$ equal for all components. We chose $j_6(n) = 2$ for X_6, $j_7(n) = 2$ for X_7 and $j(n) = 5$ for X_1, X_2, X_3, X_4 and X_5. For X_8 (number of children) we tried $j_8(n) = 1$ (first run) as well as $j_8(n) = 0$ (second run) to avoid overparametrization in that direction.

We apply again the modified tets procedure from Section 3.3 with the parameters $a_{\mathrm{loc}} = a_{\chi^2} = 2$. This choice is the smallest possible one and it leads to the least conservative test.

In Figure 4 we have displayed the functional forms for the additive components except for X_7 and X_8 as we had estimated them by using wavelets. These plots indicate that the influences of X_1 (age), X_2 (hourly earnings) and X_3 (prestige) may be not linear whereas X_4, X_5 and X_6 look rather linear. Step by step we tested each component against linearity.

In the first run linearity was rejected for X_2 and X_3. For the first component X_1 the linearity was not rejected but the values of the test statistics have been very close to the rejection boundary at almost all levels. For instance, for $j_1 = 4$ we have $T_4 = 3.487$ and $\lambda_{\chi^2}(4) = 3.548$.

In the second run the test procedure rejected linearity for X_1 and X_2 but this time not for X_3. Looking at the values of the test statistics and rejection boundaries we notice that they are pretty close at almost all levels. E. g. for $j_1 = 2$ we have $T_2 = 3.440$ and $\lambda_{\chi^2,2} = 3.490$.

Though there is a difference between the first and second run, the related results are very close and the performance of the proposed test is quite satisfactory.

In Figure 4 we have displayed the wavelet estimates of the components. Again we see that the quality of estimation is not good because of undersmoothing but some qualitative analysis of each component (in our case it is testing of linearity) is still possible. From this pictures we can recognize that the dependence from the age is rather parabolic then linear, and that the Hourly-Earnings component is not considered as a linear function because of the flat part between 5 and 15.

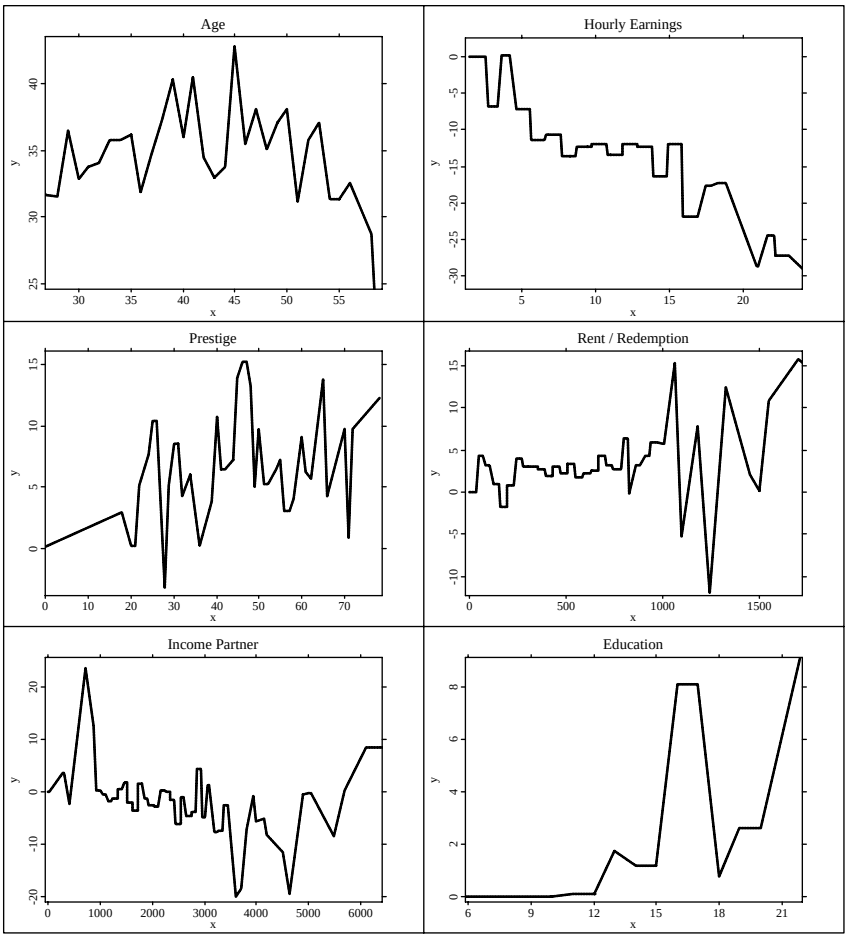

Figure 4: *The estimated functions with a part of the response variables.*

6. **Proofs**

In this section we prove Theorems 4.1 through 4.2.

6.1. **Proof of Theorem 4.1**

In a first step, we reduce the multidimensional problem with a first linear component f_1, to a univariate problem with a regression function f' satisfying the condition

$$\|f'\|_n \leq \sigma n^{-1/4}. \tag{6.1}$$

Let $j(n)$ be due to (3.26), let some $j_1 \le j(n)$ be fixed and let $\boldsymbol{\theta}^*(d, j_1)$ be introduced in (3.28). Due to condition (D),

$$\sum_{m=2}^{d} \|f_m - \Pi_{m,n} f_m\|_n \le \sigma n^{-1/4}$$

and each $\Pi_{m,n} f_m$, $m = 2, \ldots, d$, can be represented in the form

$$\Pi_{m,n} f_m = \sum_{j=0}^{j(n)} \sum_{I \in \mathcal{I}_j} \theta_{I,m} \psi'_{I,m}, \qquad m = 2, \ldots, d,$$

with some coefficients $\theta_{I,m}$, $I \in \mathcal{I}$.

Next, under the null hypothesis, the first component f_1 of F is linear, $f_1 = \theta_{0,1} + \theta_{1,1}\psi_{1,1}$. This and the above bound yield

$$\|F - f_1 - \Pi_{2,n} f_2 - \ldots - \Pi_{d,n} f_d\|_n \le \sigma n^{-1/4}. \tag{6.2}$$

Denote by $\boldsymbol{\theta}(d, j_1) = (\theta_{I,m}, (I,m) \in \mathcal{I}^{(d)}(j_1))$ the vector with $\theta_{I,1} = 0$ for $I \in \mathcal{I}_j$, $0 \le j \le j_1$, and with the above defined $\theta_{I,m}$ for $m \ge 2$. Then the inequality in (6.2) can be rewritten in the form $\|F - \Psi(d, j_1)\boldsymbol{\theta}(d, j_1)\|_n \le \sigma n^{-1/4}$. the vector $(g(X_i), i = 1, \ldots, n)$.) Set

$$F' = F - f_1 - \Pi_{2,n} f_2 - \ldots - \Pi_{d,n} f_d, \tag{6.3}$$

$$\boldsymbol{\theta}'(d, j_1) = \boldsymbol{\theta}^*(d, j_1) - \boldsymbol{\theta}(d, j_1). \tag{6.4}$$

Then obviously $\boldsymbol{\theta}'(d, j_1) = V(d, j_1)\Psi(d, j_1)F'$ and $\|F - \Psi(d, j_1)\boldsymbol{\theta}^*(d, j_1)\|_n = \|F' - \Psi(d, j_1)\boldsymbol{\theta}'(d, j_1)\|_n$. At the same time, the vectors $\boldsymbol{\theta}^*(d, j_1)$ and $\boldsymbol{\theta}'(d, j_1)$ have the same subvector $\boldsymbol{\theta}^*_{j_1}$. Taking into account the model equation (2.2) we conclude that when considering test statistics based on $\hat{\boldsymbol{\theta}}_{j_1}$, the regression function F can be changed by F' without any influence on their behavior, although $\|F'\|_n \le \sigma n^{-1/4}$ due to (6.2). Moreover, in further calculations, we operate only with the subvector $\hat{\boldsymbol{\theta}}_{j_1}$ and the corresponding covariance matrix V_{j_1} and therefore the multi-dimensional structure of the model is not important. We will use only the latter bound. This allows to reduce the original problem to the univariate case with the model function f' satisfying (6.1).

At the next step, we evaluate the error probabilities of the first kind for the tests ϕ_{loc} and ϕ_{χ^2}. We use the following technical assertion.

Lemma 6.1. *Let numbers $a_1, \ldots, a_n$ be such that the sum $a_1\xi_1 + \ldots + a_n\xi_n$ is standard normal, where $\xi_1, \ldots, \xi_n$ are independent normal $\mathcal{N}(0, \sigma^2)$ random variables. Then*

$$\left| \sum_{i=1}^{n} a_i f'(X_i) \right|^2 \le \sum_{i=1}^{n} |f'(X_i)|^2 = \sigma^{-2} \|f'\|_n^2.$$

Proof. It suffices to note that the standard normality of $a_1\xi_1 + \ldots + a_n\xi_n$ implies $1 = \boldsymbol{E}(a_1\xi_1 + \ldots + a_n\xi_n)^2 = \sigma^2(a_1^2 + \ldots + a_n^2)$ and the assertion follows by application of the Cauchy-Schwarz inequality. □

Given $j \le j(n)$, let $\hat\theta_I$, $I \in \mathcal{I}_j$, be the elements of the vector $\hat{\boldsymbol\theta}_j$ and let $V_j = (v_{I,I'}, I, I' \in \mathcal{I}_j)$ be the corresponding covariance matrix. The local test $\phi_{\mathrm{loc}}(j)$ is based on the statistics $T_I = \sigma^{-1} v_{I,I}^{-1/2} \hat\theta_I$, and

$$\boldsymbol{P}\{\phi_{\mathrm{loc}}(j) = 1\} \le \sum_{I \in \mathcal{I}_j} \boldsymbol{P}(|T_I| > \lambda_n).$$

Obviously T_I can be represented in the form $T_I = a_1 Y_1 + \ldots + a_n Y_n$ with some coefficients a_i depending on I and on the design $X_1, \ldots, X_n$. Using the model equation (2.1) with f' in place of F, we get

$$T_I = \sum_{i=1}^n a_i f'(X_i) + \sum_{i=1}^n a_i \xi_i = b_I + \zeta_I.$$

Recall that the choice of normalizer $\sigma^{-1} v_{I,I}^{-1/2}$ for $\hat\theta_I$ was made to provide standard normality of the stochastic term $\zeta_I = a_1\xi_1 + \ldots + a_n\xi_n$. Now we obtain by (6.1) and Lemma 6.1 for the deterministic term $b_I = a_1 f'(X_1) + \ldots + a_n f'(X_n)$

$$|b_I| \le \sigma^{-1} \|f'\|_n \le n^{-1/4}.$$

Since ζ_I is standard normal, we get

$$\begin{aligned} \boldsymbol{P}(|T_I| > \lambda_n) &= \boldsymbol{P}(|b_I + \zeta_I| > \lambda_n) \le \boldsymbol{P}(|\zeta_I| > \lambda_n - |b_I|) \\ &\le 2\exp\{-\left(\lambda_n - n^{-1/4}\right)^2/2\}. \end{aligned}$$

This yields

$$\begin{aligned} \boldsymbol{P}\{\phi_{\mathrm{loc}}(j) = 1\} &= \boldsymbol{P}\left(\max_{I \in \mathcal{I}_j} |T_I| > \lambda_n\right) \le \sum_{I \in \mathcal{I}_j} \boldsymbol{P}(|T_I| > \lambda_n) \\ &\le 2^{j+1} \exp\{-\left(\lambda_n - n^{-1/4}\right)^2/2\} \end{aligned}$$

and

$$\boldsymbol{P}\{\phi_{\mathrm{loc}} = 1\} \le \sum_{j=0}^{j(n)} \boldsymbol{P}(\phi_{\mathrm{loc}}(j) = 1) \le 2^{j(n)+2} \exp\{-\left(\lambda_n - n^{-1/4}\right)^2/2\}.$$

Recall that the definition of $j(n)$ implies $2^{j(n)+1} \le n$. If now $n \ge 2$, then $n^{-1/4} \le 1$, and $\boldsymbol{P}(\phi_{\mathrm{loc}} = 1) \le 2n \exp\{-2\log n\} = o_n(1)$.

Next we consider the test ϕ_{χ^2}. Let us fix again some level $j \le j(n)$. We suppose for simplicity that the matrix V_j is of the full rank. The general case can be studied in the similar way.

The subtest $\phi_{\chi^2}(j)$ is based on the statistic $S_j = \|\boldsymbol\eta_j\|^2 = \sigma^{-2}\|V_j^{-1/2}\hat{\boldsymbol\theta}_j\|^2$. Again we can represent $\boldsymbol\eta_j = \sigma^{-1} V_j^{-1/2} \hat{\boldsymbol\theta}_j$ in the form

$$\boldsymbol\eta_j = A(Y) = A(f') + A(\boldsymbol\xi) = \boldsymbol{b}_j + \boldsymbol\zeta_j$$

where A is a linear operator from $\mathbb{R}^n$ into $\mathbb{R}^{N_j}$, $\boldsymbol{b}_j$ is a constant vector in $\mathbb{R}^{N_j}$ and $\boldsymbol{\zeta}_j$ is a standard normal vector in $\mathbb{R}^{N_j}$. Applying Lemma 6.1 to each component b_I of the vector $\boldsymbol{b}_j$, $I \in \mathcal{I}_j$, we get $b_I^2 \le \sigma^{-2}\|f'\|_n^2$ and hence by (6.1)

$$\|\boldsymbol{b}_j\|^2 \le N_j \sigma^{-2}\|f'\|_n^2 \le N_j n^{-1/2}. \tag{6.5}$$

Denote

$$\gamma_j = \|\boldsymbol{b}_j\|^{-1} \sum_{I \in \mathcal{I}_j} b_I \zeta_I.$$

Clearly γ_j is a standard Gaussian random variable and we can decompose $S_j = \|\boldsymbol{b}_j + \boldsymbol{\zeta}_j\|^2 = \|\boldsymbol{b}_j\|^2 + \|\boldsymbol{\zeta}_j\|^2 + 2\|\boldsymbol{b}_j\|\gamma_j$. Now by (6.5)

$$\begin{aligned}
\boldsymbol{P}\{\phi_{\chi^2}(j) = 1\} &= \boldsymbol{P}\left\{|S_j - N_j| > \sqrt{2N_j}\lambda_n\right\} \\
&\le \boldsymbol{P}\left(\left|\|\boldsymbol{\zeta}_j\|^2 - N_j\right| > \sqrt{2N_j}(\lambda_n - 2^{-1/2}) - \|\boldsymbol{b}_j\|^2\right) + \boldsymbol{P}\left(2\|\boldsymbol{b}_j\|\,|\gamma_j| > \sqrt{N_j}\right) \\
&\le \boldsymbol{P}\left\{\left|\|\boldsymbol{\zeta}_j\|^2 - N_j\right| > \sqrt{2N_j}\left(\lambda_n - 2^{-1/2} - \sqrt{N_j/(2n)}\right)\right\} + \boldsymbol{P}(|\gamma_j| > n^{1/4}/2).
\end{aligned}$$

Notice that $N_j \le n/2$ for all $j \le j(n)$ and we have for all $n \ge 2$ that $\lambda_n - 2^{-1/2} - \sqrt{N_j/(2n)} \ge \sqrt{2\log n}$. Next, see Petrov (1975),

$$\boldsymbol{P}\left(\frac{\left|\|\boldsymbol{\zeta}_j\|^2 - N_j\right|}{\sqrt{2N_j}} > \sqrt{2\log n}\right) \le \exp\{-\log n\} = n^{-1},$$

and $\boldsymbol{P}(|\gamma_j| > n^{1/4}/2) \le \exp\{-n^{1/2}/8\}$. Therefore,

$$\boldsymbol{P}(\phi_{\chi^2}(j) = 1) \le n^{-1} + \exp\{-n^{1/2}/8\}.$$

Summing up over all j from zero to $j(n)$ we conclude that

$$\boldsymbol{P}(\phi_{\chi^2} = 1) \le \sum_{j=0}^{j(n)} \boldsymbol{P}(\phi_{\chi^2}(j) = 1) \le \left\{n^{-1} + \exp(-n^{1/2}/8)\right\}\log n \to 0,$$

as $n \to \infty$. This completes the proof of Theorem 4.1.

6.2. **Proof of Proposition 4.1**

For nonational simplicity we write θ_I^* resp. $v_{I,I}$ instead of $\theta_{I,1}^*$ resp. $v_{(I,1),(I,1)}$. We suppose also that the matrix V_j is non-degenerate.

Let, for some $j \le j(n)$ and some $I \in \mathcal{I}_j$,

$$\sigma^{-1}|v_{I,I}^{-1/2}\theta_I^*| > 2\lambda_n.$$

We use the decomposition $\sigma^{-1} v_{I,I}^{-1/2} \hat{\theta}_I = \sigma^{-1} v_{I,I}^{-1/2} \theta_I^* + \zeta_I$ where ζ_I is standard normal. Clearly

$$\begin{aligned} \boldsymbol{P}_F\{\phi_{\mathrm{loc}}(j) = 0\} &\le \boldsymbol{P}_F\left(\sigma^{-1}|v_{I,I}^{-1/2}\hat{\theta}_I| < \lambda_n\right) \\ &= \boldsymbol{P}_F\left(\sigma^{-1}|v_{I,I}^{-1/2}\theta_I^* + \zeta_I| < \lambda_n\right) \\ &\le \boldsymbol{P}_F\left(|\zeta_I| > \lambda_n\right) \le e^{-\lambda_n^2/2} \to 0, \qquad n \to \infty, \end{aligned}$$

as required.

Next we consider the situation when

$$T_j^* = 2^{-(j+1)/2}\sigma^{-2}\boldsymbol{\theta}_j^{*T} V_j^{-1} \boldsymbol{\theta}_j^* > 2\lambda_n. \tag{6.6}$$

We will show that under the above assumption,

$$\boldsymbol{P}_F(T_j < \lambda_n) \le \delta(n) \to n, \qquad n \to \infty, \tag{6.7}$$

that obviously implies the assertion.

Recall that in the case when $\det V_j \neq 0$, one has $T_j = 2^{-(j+1)/2}(S_j - 2^j)$ where $S_j = \sigma^{-2}\|V_j^{-1/2}\hat{\boldsymbol{\theta}}_j\|^2$. By construction, we can represent the vector $\sigma^{-1}V_j^{-1/2}\hat{\boldsymbol{\theta}}_j$ in the form

$$\sigma^{-1}V_j^{-1/2}\hat{\boldsymbol{\theta}}_j = \boldsymbol{b}_j + \boldsymbol{\zeta}_j$$

where $\boldsymbol{b}_j = \sigma^{-1}V_j^{-1/2}\boldsymbol{\theta}_j^*$ and $\boldsymbol{\zeta}_j$ is a standard Gaussian vector. Notice that

$$\|\boldsymbol{b}_j\|^2 = \sigma^{-2}\boldsymbol{\theta}_j^{*T}V_j^{-1}\boldsymbol{\theta}_j^* = 2^{(j+1)/2}T_j^*.$$

Denote

$$\gamma_j = \|\boldsymbol{b}_j\|^{-1}\sum_{I\in\mathcal{I}_j} b_I\zeta_I.$$

Clearly γ_j is a standard Gaussian random variable and we can decompose

$$S_j = \|\boldsymbol{b}_j + \boldsymbol{\zeta}_j\|^2 = \|\boldsymbol{b}_j\|^2 + \|\boldsymbol{\zeta}_j\|^2 + 2\|\boldsymbol{b}_j\|\gamma_j.$$

Now we have

$$\begin{aligned} \boldsymbol{P}_F(|T_j| < \lambda_n) &= \boldsymbol{P}\left(\left|\|\boldsymbol{b}_j\|^2 + \|\boldsymbol{\zeta}_j\|^2 - 2^j + 2\|\boldsymbol{b}_j\|\gamma_j\right| < \lambda_n 2^{(j+1)/2}\right) \\ &\le \boldsymbol{P}\left(2^{-(j+1)/2}|\|\boldsymbol{\zeta}_j\|^2 - 2^j| > \tfrac{3}{4}T_j^* - \lambda_n\right) + \boldsymbol{P}(|\gamma_j| > \tfrac{1}{4}T_j^{*1/2}). \end{aligned}$$

It remains to note that $\frac{3}{4}T_j^* - \lambda_n \ge \lambda_n/2$ in view of (6.6) and we end up using the arguments from the proof of Theorem 4.1.

6.3. **Proof of Theorem 4.2**

We begin again by reduction of the multi-dimensional problem to a univariate one. Let the functions $\Pi_{m,n} f_m$ for $m = 2, \ldots, d$ be defined as above, see condition (D), and let $F'' = F - \Pi_{2,n} f_2 - \ldots - \Pi_{d,n} f_n$. Define coefficients $\theta_{0,1}, \theta_{1,1}$ by

$$(\theta_{0,1}, \theta_{1,1}) = \operatorname*{arginf}_{(a,b)} \|F'' - a - b\psi_{1,1}\|_n = \operatorname*{arginf}_{(a,b)} \sum_{i=1}^{n} \{F''(X_{i,1}) - a - bX_{i,1}\}^2.$$

We set

$$F' = F'' - \theta_{0,1} - \theta_{1,1}\psi_{1,1} = F - \theta_{0,1} - \theta_{1,1}\psi_{1,1} - \Pi_{2,n} f_2 - \ldots - \Pi_{d,n} f_n.$$

Similarly to the proof of Theorem 4.1, we change F by F'. With this change, the vectors $\boldsymbol{\theta}^*(d,j)$ will be transformed into $\boldsymbol{\theta}'(d,j)$, having the same subvectors $\boldsymbol{\theta}_j^*$, $j \geq 0$. At the same time, by the triangle inequality and condition (D), for all a, b,

$$\begin{aligned} \|F' - a - b\psi_{1,1}\|_n &\geq \|f_1 - (a - \theta_{0,1}) - (b - \theta_{1,1})\psi_{1,1}\|_n - \sum_{m=2}^{d} \|f_m - \Pi_{m,n} f_m\|_n \\ &\geq \varrho(n) - n^{-1/4}. \end{aligned}$$

Here we have set

$$\varrho(n) = \inf_{a,b} \|f_1 - a - b\psi_{1,1}\|_n.$$

Similarly we transform smoothness properties of the first component f_1 into the accuracy of approximation of F' by piecewise-polynomial functions of x_1,

$$\inf_{g \in \mathcal{P}_s(j)} \|F' - g\|_n \leq r_s'(j) = r_s(j) + n^{-1/4}.$$

From this point on, we may treat the multi-dimensional structure of our model as if we are given univariate data corresponding to a univariate function f in place of F'. We omit therefore the second subindex m in our notation. About this function f we know that

$$\|f\|_n = \inf_{a,b} \|f - a - b\psi_1\|_n \geq \varrho'(n) = \varrho(n) - n^{-1/4}, \tag{6.8}$$

$$\inf_{g \in \mathcal{P}_s(j)} \|f - g\|_n \leq r_s'(j) = r_s(j) + n^{-1/4}, \tag{6.9}$$

$$\boldsymbol{\theta}^*(j) = \operatorname*{arginf}_{\boldsymbol{\theta}(j)} \|f - \Psi(j)\boldsymbol{\theta}(j)\|_n, \tag{6.10}$$

for all j from zero to $j(n)$.

Now we turn directly to the proof of the theorem using the result of Proposition 4.1. We show that condition (4.7) of the theorem with sufficiently large C_1 and C_2 along with (6.8) and (6.9) contradict to the constraints from (4.1).

First we rewrite the latter constraints in term of $\|\boldsymbol{\theta}_j^*\|$. Recall that $\boldsymbol{\theta}_j^*$ is the subvector of $\boldsymbol{\theta}^*(j)$ corresponding to j-th level, and V_j is the corresponding covariance submatrix of $V(j)$.

Lemma 6.2. *If* $T_j^* = 2^{-(j+1)/2}\sigma^{-2}\boldsymbol{\theta}_j^{*T}V_j^{-1}\boldsymbol{\theta}_j^* \le 2\lambda_n$, *then*

$$\|\boldsymbol{\theta}_j^*\|^2 \le 2^{(j+3)/2}\sigma^2\lambda_n v^*(j). \tag{6.11}$$

Similarly, the inequality $T_{j,\infty}^* = \max_{I\in\mathcal{I}_j}\sigma^{-1}v_{I,I}^{-1/2}|\theta_I^*| \le 2\lambda_n$ *implies*

$$\max_{I\in\mathcal{I}_j}|\theta_I^*| \le 2\sigma\lambda_n\sqrt{v^*(j)}. \tag{6.12}$$

Proof. Both statements are a direct consequence of the definition of the norm of a matrix. Indeed, let $\boldsymbol{\eta}_j = \sigma^{-1}V_j^{-1/2}\boldsymbol{\theta}_j^*$. Then $T_j^* = \|\boldsymbol{\eta}_j\|^2$ and $\boldsymbol{\theta}_j^* = \sigma V_j^{1/2}\boldsymbol{\eta}_j$. Next, obviously $\|V_j^{1/2}\| = \sqrt{\|V_j\|} = \sqrt{v^*(j)}$. Particularly this yields that

$$\|\boldsymbol{\theta}_j^*\|^2 = \sigma^2\|V_j^{1/2}\boldsymbol{\eta}_j\|^2 \le \sigma^2\left(\|V_j^{1/2}\|\,\|\boldsymbol{\eta}_j\|\right)^2 \le \sigma^2 v^*(j)T_j^*,$$

and (4.1) implies (6.11).

Similarly $v_{I,I} \le \|V_j\|$ for all $I \in \mathcal{I}_j$ and hence

$$|\theta_I^*| = v_{I,I}^{1/2}|v_{I,I}^{-1/2}\theta_I^*| \le \sqrt{v^*(j)}\sigma T_{j,\infty}^*.$$

□

Recall the notation $\mathcal{L}(j)$ for the linear space generated by functions ψ_I, $I \in \mathcal{I}_{j'}$, with $0 \le j' < j$ and $\Pi(j)f$ for the projection of f onto the space $\mathcal{L}(j)$ with respect to the norm $\|\cdot\|_n$,

$$\Pi(j)f = \operatorname*{arginf}_{h\in\mathcal{L}(j)}\|f-h\|_n.$$

Particularly, $\Pi(0)f$ denotes the projection of f on the space of linear functions and by (6.8) $\Pi(0)f = 0$.

Lemma 6.3. *For each* $j \le j(n)$,

$$\|\Pi(j+1)f\|_n \le \|\Pi(j)f\|_n + \|\boldsymbol{\theta}_j^*\|.$$

Proof. Since $\mathcal{L}(j-1) \subseteq \mathcal{L}(j)$, then

$$\Pi(j)f = \Pi(j)\Pi(j+1)f.$$

If $f(j+1) = \Pi(j+1)f$, then $\Pi(j)f = \Pi(j)f(j+1)$ and we have to show that

$$\|\Pi(j)f(j+1)\|_n \ge \|f(j+1)\|_n - \|\boldsymbol{\theta}_j^*\|.$$

In view of (6.10)

$$f(j+1) = \sum_{I\in\mathcal{I}(j)}\theta_I^*\psi_I'.$$

Denote by f_j the part of this sum corresponding to the last level $\mathcal{I}_j$ in $\mathcal{I}(j)$,

$$f_j = \sum_{I\in\mathcal{I}_j}\theta_I^*\psi_I'.$$

By construction, the functions ψ'_I, $I \in \mathcal{I}_j$, are orthonormal w.r.t. to the inner product $\|\cdot\|_n$ and in particular

$$\|f_j\|_n^2 = \sum_{I \in \mathcal{I}_j} |\theta^*_I|^2 = \|\boldsymbol{\theta}^*_j\|^2.$$

Next, obviously $f(j+1) - f_j \in \mathcal{L}(j)$, and by definition of $\Pi(j)$,

$$\|f(j+1) - \Pi(j) f(j+1)\|_n \le \|f(j+1) - \{f(j+1) - f_j\}\|_n = \|f_j\|_n = \|\boldsymbol{\theta}^*_j\|.$$

Now the assertion follows from the triangle inequality. □

Lemma 6.4. *Given $j_1 \le j(n)$, let (4.1) hold true for all $j \le j_1$. Then*

$$\|\Pi(j_1) f\|_n^2 \le \varkappa_1 2^{j_1/2} \sigma^2 \lambda_n v^*(j_1) \tag{6.13}$$

with $\varkappa_1 = 2^{3/2}(2^{1/4} - 1)^{-2}$.

Proof. Recursive application of Lemma 6.3 gives

$$\|\Pi(j_1) f\|_n \le \sum_{j=0}^{j_1 - 1} \|\boldsymbol{\theta}^*_j\|.$$

Here we have used that $\Pi(0) f = 0$. Since the norm $v^*(j)$ obviously increases with j, then this result along with the bound (6.11) yields

$$\|\Pi(j_1) f\|_n \le \sum_{0}^{j_1 - 1} \left\{ 2^{(j+3)/2} \sigma^2 \lambda_n v^*(j_1) \right\}^{1/2} = \left\{ 2^{3/2} \sigma^2 \lambda_n v^*(j_1) \right\}^{1/2} \sum_{j=0}^{j_1 - 1} 2^{j/4}$$

and the assertion follows by straightforward calculation. □

Lemma 6.5. *There is a constant $\varkappa_2 > 0$ depending on s only such that for each $j \le j(n)$*

$$\|\Pi(j) f\|_n \ge \varkappa_2 u^*(j)/u_*(j) \left(\|f\|_n - r'_s(j) \right).$$

Proof. Let $g \in \mathcal{P}_s(j - j_0)$ be such that $\|f - g\|_n \le r'_s(j)$. Then

$$\begin{aligned} \|\Pi(j) f\|_n &= \|\Pi(j) g + \Pi(j)(f - g)\|_n \ge \|\Pi(j) g\|_n - \|\Pi(j)(f - g)\|_n \\ &\ge \|\Pi(j) g\|_n - r'_s(j). \end{aligned}$$

Recall that g is a piecewise polynomial function on the partition A_I, $I \in \mathcal{I}_{j - j_0}$ and the projection $\Pi(j) g$ means the approximation of each polynomial on interval A_I of length $2^{-(j - j_0)}$ by piecewise constant functions with piece length 2^{-j}. Therefore, it suffices to prove that for each piece A_I and every polynomial $P(x) = a_0 + a_1 x + \ldots + a_{s-1} x^{s-1}$, it holds

$$\sum_{A_I} \{\Pi(j) P(X_i)\}^2 \ge \varkappa u^*(j)/u_*(j) \sum_{A_I} P^2(X_i)$$

where the constant $\varkappa$ depends on s only. This fact is a consequence of the next general statement. □

Lemma 6.6. *Let* $u_* \le 1$, $u^* \ge 1$ *and let* μ *be a measure on* $[0,1]$ *such that*

$$u_* 2^{-j_0} \le \mu(A_k) \le u^* 2^{-j_0} \tag{6.14}$$

for all intervals $A_k = [k2^{-j_0}, (k+1)2^{-j_0})$, $k = 0, 1, \ldots, 2^{j_0}$. *Then there exists a constant* $\varkappa$ *depending on* s *only and such that for every polynomial* $P(x) = a_0 + a_1 x + \ldots + a_{s-1}x^{s-1}$

$$\sum_{k=0}^{2^{j_0}-1} \left\{ \int_{A_k} P(x)\mu(dx) \right\}^2 \ge \varkappa u_*/u^* \int_0^1 P^2(x)\mu(dx).$$

Proof. The similar fact with integration instead of summation over A_I was stated in Ingster (1993) and we present here only a sketch of the proof for our situation.

We begin by reducing the case of arbitrary u_* and u^* to the case when $u_* = u^* = 1$. Define a measure μ' on $[0,1]$ by $d\mu'/d\mu(x) = u_* 2^{-j_0}/\mu(A_k)$ if $x \in A_k$. Due to (6.14), $d\mu'/d\mu \le 1$ and obviously $\mu'(A_k) = u_* 2^{-j_0}$. Next, similarly $d\mu/d\mu' = u_*^{-1} 2^{j_0}\mu(A_k) \le u_*^{-1}u^*$. Now

$$\begin{aligned} \sum_{k=0}^{2^{j_0}-1} \left\{ \int_{A_k} P(x)\mu(dx) \right\}^2 &\ge \sum_{k=0}^{2^{j_0}-1} \left\{ \int_{A_k} P(x)\mu'(dx) \right\}^2, \\ \int_0^1 P^2(x)\mu'(dx) &\ge u_*^{-1}u^* \int_0^1 P^2(x)\mu(dx). \end{aligned}$$

Therefore, it suffices to show that

$$\sum_{k=0}^{2^{j_0}-1} \left\{ \int_{A_k} P(x)\mu'(dx) \right\}^2 \ge \varkappa \int_0^1 P^2(x)\mu'(dx),$$

or, equivalently to consider the original problem with $\mu(A_k) = 2^{-j_0}$ for all $k = 0, \ldots, 2^{j_0} - 1$.

Let $a = (a_0, \ldots, a_{s-1})$ be the vector of coefficients of P. Then obviously

$$\int_0^1 P^2(x)\mu(dx) \le C\|a\|^2$$

where $\|a\|^2 = a_0^2 + \ldots + a_{s-1}^2$. Introduce a matrix M with elements $\mu_{k,l} = \int_{A_k} x^l \mu(dx)$, $k = 0, \ldots, 2^{j_0} - 1$, $l = 0, \ldots, s-1)$. Then Ma is a vector in the space $\mathbb{R}^{2^{j_0}}$ and

$$\sum_{k=0}^{2^{j_0}-1} \left\{ \int_{A_k} P(x)\mu(dx) \right\}^2 = \|Ma\|^2.$$

Now we use that $\|Ma\|^2 = a^T M^T M a \ge \|a\|^2/\|(M^T M)^{-1}\|$. It remains to note that the conditions $s < 2^{j_0-1}$ and $\mu(A_k) = 2^{-j_0}$ yield that $\|(M^T M)^{-1}\| \le C$ for some constant C depending on s only. □

Summing up the results of Lemma 6.2 through 6.5 we see that the inequality $\|f\|_n \ge r'_s(j) + C\sqrt{2^{j/2}\lambda_n v^*(j)u^*(j)/u_*(j)}$ for sufficiently large C contradicts to constraints (4.1) and the theorem is proved.

6.4. **Proof of Theorem 4.3**

We proceed in the same line as in the proof of Theorem 4.2. The difference is only in evaluating the norm $\|\Pi(j)f\|_n$, see Lemmas 6.3 and 6.4. Similarly to Lemma 6.3 we can show that

$$\|\Pi(j+1)f\|_n \le \|\Pi(j)f\|_n + \|\boldsymbol{\theta}_j^*\|.$$

Next, in view of the constraints from (4.2) and Lemma 6.2, one has

$$\|\boldsymbol{\theta}_j^*\|^2 = \sum_{I\in\mathcal{I}_j} |\theta_I^*|^2 = \sum_{I\in\mathcal{I}_j} |\theta_I^*|^2 \mathbf{1}(|\theta_I^*| \le t_j) = r(j,t_j)$$

where $t_j = 2\sigma\lambda_n\sqrt{v^*(j)}$. Using this bound for j between j_1 and j_2 and the bound from Lemma 6.3 for $j < j_1$, we estimate

$$\|\Pi(j_2)f\|_n \le \varkappa_1 2^{j_1}\sigma^2\lambda_n v^*(j_1) + \sum_{j=j_1}^{j_2} r(j,t_j).$$

This allows to complete the proof by the same arguments as for Theorem 4.2.

6.5. **Proof of Lemma 4.1**

By definition of $j(n)$ it holds $n^{-1} \le 2^{-j(n)-1} \le dn^{-1}$. Next, it is easy to see that $\Pi_{m,n}$ is the projection of the function f_m on the space of piecewise constant functions with the piece length $2^{-j(n)-1}$. Let A be one from these intervals and let for $m \le d$, $N_{A,m}$ 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 imlies 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

$$\|f_m - \Pi_{m,n} f_m\|_n^2 \le C_1 C_2^2 n 2^{-2j(n)-2} \le C_1 C_2^2 d^2 n^{-1}$$

as required.

References

[1] Alexits, G. (1961). *Convergence problem of orthogonal series.* Akademiai Kiado, Budapest.

[2] 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.

[3] Buja, Hastie and Tibshirani (1989). Linear smoothers and additive models. *Ann. Statist.*, **17**, no. 2, 453–555.

[4] Härdle and Korostelev (1996). Search of significant variables in nonparametric additive regression. *Biometrika* **83** 541–549.

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

[6] Härdle, Spokoiny and Sperlich (1997). Semiparametric single index versus fixed link function modelling. *Ann. Statist.* **27**, no. 1, 212–243.

[7] Härdle and Tsybakov (1995). Additive nonparametric regression on principle components. *J. Nonparametric Statistics* **5** 157–184.

[8] Hastie and Tibshirani (1990). *Generalized additive models.* Chapman & Hall.

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

[10] 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.

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

[12] Lepski, O.V. and Spokoiny, V.G. (1995). Minimax Nonparametric Hypothesis Testing: The Case of an Inhomogeneous Alternative. *Bernoulli*, tentatively accepted.

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

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

[15] Petrov, V.V. (1975). *Sums of Independent Random Variables.* Springer, New York.

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

[17] Spokoiny, V. (1996b) Adaptive and spatially adaptive testing of a nonparametric hypothesis. *Math. Methods of Statist.*, submitted.

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

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

[20] Triebel, H. (1992). *Theory of function spaces.* Birkhäuser, Basel.

[21] Venables and Ripley (1994). *Modern applied statistics with S-Plus.* Springer. New York.