ESAIM: Probability and Statistics Will be set by the publisher
URL: `http://www.emath.fr/ps/`

e-mail: `spokoiny@wias-berlin.de`

DATA-DRIVEN TESTING THE FIT OF LINEAR MODELS *

SPOKOINY, V.[1]

Abstract. The paper is concerned with the problem of testing a linear hypothesis about regression function. We propose a new testing procedure based on the Haar transform which is adaptive to unknown smoothness properties of the underlying function. The results describe optimality properties of this procedure under mild conditions on the model.

AMS Subject Classification. 62H25; Secondary 62G10.

October, 8, 1999.

1. Introduction

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

$$Y_i = f(X_i) + \xi_i \tag{1.1}$$

where f is an unknown regression function and ξ_i are zero mean random errors. Statistical analysis for such models may focus on the qualitative features of the underlying function f. Particularly, no-response model corresponds to testing the simple zero hypothesis that f is a constant function. Another typical example is connected to the hypothesis of linearity. More generally one may consider a parametric type hypothesis about f. In this paper, we restrict ourselves to the case of the hypothesis of linearity. Using the hypothesis testing framework, we test the null hypothesis $H_0 : f$ 'is linear', that is, $f(x) = a + bx$ for some constants a, b, versus the alternative $H_1 : f$ 'is not linear'.

The problem of testing a simple or parametrically specified hypothesis is one of the classical in statistical inference, see e.g. Neyman (1937), Mann and Wald (1942), Lehmann (1957). Let ϕ be a test i.e. a measurable function of the observations $Y_1, \ldots, Y_n$ with two values $0, 1$. As usual, the event $\{\phi = 0\}$ is treated as accepting the hypothesis and $\phi = 1$ means that the hypothesis is rejected. The quality of a test ϕ is described in terms of the corresponding error probabilities of the first and second kinds. Let $\boldsymbol{P}_f$ denote the distribution of the data $Y_1, \ldots, Y_n$ for a fixed model function f, see (1.1). If f coincides with a linear function f_0, then the error probability of the first kind at the point f_0 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 function f is not linear, then

$$\beta_f(\phi) = \boldsymbol{P}_f(\phi = 0).$$

Keywords and phrases: data-driven test, Haar basis, linear hypothesis, nonparametric alternative, regression model

* *The author is grateful to the anonymous Referee for helpful remarks and critics.*

[1] Weierstrass-Institute for Applied Analysis and Statistics, Mohrenstr. 39, 10117 Berlin Germany

Typically one aims to construct a test φ of the prescribed level α_0, that is, satisfying for a given $\alpha_0 > 0$ the condition $\alpha_{f_0}(\phi) \le \alpha_0$ which also has a nontrivial power $1 - \beta_f(\varphi) > 0$ against a possibly large class of alternatives f. A large number of proposals for constructing such tests can be found in the literature. We refer to Hart (1997) where the reader can find historical remarks and further references. Note meanwhile, that the majority of results in this domain is concentrated either only on verifying the condition $\alpha_{f_0}(\phi) \le \alpha_0$ or on studying asymptotic properties of the power function $1 - \beta_f(\varphi)$ for a fixed or local alternative. The local alternative approach assumes that the hypothesis is tested versus alternatives approaching the null hypothesis from a specific direction. Many tests have been shown to have nontrivial asymptotic power against every such local alternative, see e.g. Bierens (1982, 1990), Eubank and Spiegelman (1990), Andrews (1997), Stute (1997) among others. However, it turns out that the finite sample power of the proposed tests is not uniform with respect to alternative directions: some of directional alternatives can be detected easily, the others require a huge sample size. Moreover, Burnashev (1979) and Ingster (1982) have shown that no test can be uniformly powerful against all the local alternatives. This leads to considering the uniform power of the test over a large class $\mathcal{F}$ of alternatives, so that $\beta_f(\varphi) \le \beta_0$ with some $\beta_0 < 1 - \alpha_0$ uniformly over $f \in \mathcal{F}$. Following Ingster (1982, 1993), we consider the class $\mathcal{F}(\rho)$ consisting of smooth (in some sense) alternatives which are also separated from the set of linear functions with the distance ρ, that is,

$$\inf_{a,b} \|f(\cdot) - a - b \cdot\| \ge \rho, \tag{1.2}$$

$\|\cdot\|$ being the usual L_2-norm. Then the quality of a test φ of the level α_0 can be measured by a minimal *separation* distance ρ such that $\beta_f(\phi) \le \beta_0$ for all f from $\mathcal{F}(\rho)$. A test ϕ^* with the level α_0 is *optimal* if it minimizes the corresponding value ρ. Under this approach, the goal is both to evaluate the minimal possible separation distance ρ and to describe the corresponding optimal tests.

It turns out that the structure of optimal tests and the corresponding separation distance strongly depend on the smoothness class $\mathcal{F}$ we consider. Ingster (1982, 1993) described the optimal rate of decay of the separation distance ρ to zero as the sample size n tends to infinity for Hölder and Sobolev function classes, the case of Besov classes is considered in Lepski and Spokoiny (1998) and Spokoiny (1998). Sharp optimal asymptotic results can be found in Ermakov (1990), Lepski and Tsybakov (1996), Ingster and Suslina (1998).

Unfortunately all the mentioned procedures hardly apply in practice since the information about smoothness properties of the underlying function f is typically lacking. Some adaptive (data-driven) smooth tests are proposed in Eubank and Hart (1992), Ledwina (1994), Fan (1996), Hart (1997) where the reader can found further references. Spokoiny (1996, 1998) considered the problem of adaptive testing against a smooth alternative and constructed an adaptive test which is near optimal by a *log log* multiple for a wide range of smoothness classes. Moreover, the test is rate optimal in the class of adaptive tests, that is, this *log log* factor is an unavoidable payment for the adaptive property. The inconvenience for practical applications is that this procedure is designed for an idealized 'signal + white noise' model and only the case of a simple null is considered.

The aim of this paper is to develop an adaptive testing method which allows for a non-regular design, non-Gaussian errors with an unknown distribution and a non-simple null, and which is computationally simple and stable w.r.t. the design non-regularity. The latter property is achieved by making use of the simplest wavelet basis, namely the Haar transform. It is worth mentioning that the Haar basis is not often used for estimating the regression function f from (1.1) because of its non-regularity: the corresponding estimator is based on the piecewise constant approximation of the underlying function and it is only rate suboptimal. Nevertheless, Ingster (1993) has constructed a χ^2-test (also based on a piecewise constant approximation) which provides the optimal testing rate in the 'signal + white noise' framework. Here his construction is extended to the case of testing the linear hypothesis for regression with unknown smoothness properties and with a deterministic non-regular design.

Another remark concerns the assumption on the errors ξ_i. Assuming i.i.d. errors with a known distribution, one can easily select a critical level for any test statistic using the Monte-Carlo or other resampling technique.

For practical applications, this approach needs to be justified since the underlying error distribution is typically unknown. The problem becomes even more complicated if a data-driven test basing on the maximum of different test statistics is used. We establish some general results on the approximation of quadratic forms of independent random variables by similar quadratic forms of Gaussian random variables which help to justify the following recipe: if the critical level of the considered test statistic is calculated for Gaussian errors, then it applies, at least asymptotically, as the sample size grows, for an arbitrary errors distribution with bounded 6 moments.

The paper is organized as follows. Section 2 contains the description of the proposed testing procedure. The properties of this procedure are discussed in Section 3. The proofs are postponed to Section 4. In the Appendix we collect some general results for quadratic forms.

2. Testing procedure

We consider the univariate regression model

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

with additive homogeneous noise, that is, the errors ξ_i are independent identically distributed with zero mean and the variance σ^2: $\boldsymbol{E}\xi_i = 0$ and $\boldsymbol{E}\xi_i^2 = \sigma^2$. The design points $X_1, \ldots, X_n$ are assumed to be rescaled to the interval $[0,1]$, that is, $X_i \in [0,1]$ for all $i = 1, \ldots, n$.

The proposed test makes use of the Haar transform. We first recall some useful facts about the Haar decomposition and then explain the idea of the method.

2.1. Preliminaries

Hereafter we denote by I the multi-index $I = (j,k)$ with $j = 0,1,2,\ldots$ and $k = 0,1,\ldots,2^j - 1$, and let $\mathcal{I}$ be the set of all such multi-indices. We also set

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

for the index set corresponding to j-th level. Let now the function $\psi(t)$ be defined by

$$\psi(t) = \begin{cases} 0 & t < 0, \\ 1 & 0 \le t < 1/2, \\ -1 & 1/2 \le t < 1, \\ 0 & t > 1. \end{cases} \tag{2.2}$$

For every $I = (j,k)$, define the Haar basis function h_I by

$$h_I(t) = 2^{j/2}\psi(2^j t - k). \tag{2.3}$$

Clearly the function h_I is supported on the interval $A_I = [2^{-j}k, 2^{-j}(k+1)]$. It is well known that each measurable function f on $[0,1]$ can be decomposed in the following way

$$f(t) = c_0 + \sum_{I \in \mathcal{I}} c_I h_I(t) = c_0 + \sum_{j=0}^{\infty} \sum_{I \in \mathcal{I}_j} c_I h_I(t). \tag{2.4}$$

This means that the problem of recovering the function f can be transformed into the problem of estimating the coefficients c_I by given data. Since we have only n observations, it makes no sense to estimate more (in order) than n coefficients. We restrict therefore the total number of considered levels j. Let some j be fixed such that $2^{j+1} < n$. We also introduce the rescaled basis functions ψ_I to provide $\sum_i |\psi_I(X_i)|^2 = 1$, that is,

$$\psi_I(X_i) = \mu_I^{-1} h_I(X_i),$$

with $\mu_I^2 = \sum_{i=1}^n h_I^2(X_i)$. Next we replace the infinite decomposition (2.4) by the finite approximation $\sum_{I\in\mathcal{I}(j)} c_I \psi_I(t)$ where the index set $\mathcal{I}(j)$ contains all level sets $\mathcal{I}_\ell$ with $\ell \le j$. Taking into account the structure of the null hypothesis, we complement the set of functions $(\psi_I,\, I \in \mathcal{I}_\ell)$, $\ell \le j$, with two functions $\psi_0 \equiv 1$ and $\psi_1(t) = t$, that is, we consider the set of indices

$$\mathcal{I}(j) = \{0,1\} + \bigcup_{\ell=0}^{j} \mathcal{I}_\ell. \tag{2.5}$$

The idea of the proposed procedure is to estimate all the coefficients $(c_I,\, I \in \mathcal{I}(j))$ from the data $Y_1, \ldots, Y_n$ and then to test that all the coefficients c_I for $I \ne 0,1$ are zero.

For a function g, define $\|g\|_n$ by

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

Define also the column-vector $\boldsymbol{\theta}^*(j) = (\theta_I^*,\, I \in \mathcal{I}(j))$ as a minimizer of the error of approximating f by a linear combination of ψ_I, $I \in \mathcal{I}(j)$:

$$\boldsymbol{\theta}^*(j) = \operatorname*{arginf}_{\boldsymbol{\theta}(j)} \|f - \sum_{I\in\mathcal{I}(j)} \theta_I \psi_I\|_n^2. \tag{2.6}$$

This is a quadratic optimization problem with respect to the coefficients $\{\theta_I\,, I \in \mathcal{I}(j)\}$. Therefore, the solution $\boldsymbol{\theta}^*$ always exists but it is probably non unique. To get an explicit representation for $\boldsymbol{\theta}^*$ we introduce matrix notation.

First of all, we make an agreement to identify every function g with the vector $(g(X_i),\, i = 1, \ldots, n)^\top$ in $\mathbb{R}^n$ where the symbol $^\top$ means transposition. Particularly, the model function f is identified with the vector $(f(X_i),\, i = 1, \ldots, n)^\top$.

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

$$N_j = \#(\mathcal{I}_j) = 2^j, \qquad j = 0, 1, \ldots, j$$

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

$$N(j) = 2 + \sum_{\ell=0}^{j} N_\ell = 1 + 2^{j+1}. \tag{2.7}$$

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

$$\psi_{i,I} = \psi_I(X_i) = \psi_I(X_i), \qquad I \in \mathcal{I}(j),\, i = 1, \ldots, n. \tag{2.8}$$

Clearly $\psi_I(X_i) = \pm 1/\sqrt{M_I}$ where M_I is the number of design points in the interval A_I corresponding to the index I, and also $\psi_{i,0} = n^{-1/2}$ and $\psi_{i,1} = X_i \left(\sum_{\ell=1}^n X_\ell^2\right)^{-1/2}$. Now the approximation problem (2.6) can be rewritten in the form

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

The solution to this quadratic problem can be represented as

$$\boldsymbol{\theta}^*(j) = \left(\Psi(j)^\top \Psi(j)\right)^{-1} \Psi(j)^\top f. \tag{2.9}$$

Strictly speaking, this representation is valid only if the matrix $\Psi(j)^\top \Psi(j)$ is not degenerate. In the general case, one may use the similar expression for $\boldsymbol{\theta}^*(j)$ when understanding $\left(\Psi(j)^\top \Psi(j)\right)^{-1}$ as a pseudo-inverse matrix.

If the function f is linear, that is, $f(x) = \theta_0 + \theta_1 x$, we clearly get $\theta_0^* = \theta_0$, $\theta_1^* = \theta_1$ and $\theta_I^* = 0$ for all $I = (\ell, k)$ with $\ell \geq 0$ and $k \geq 0$. For a non-linear function f, the sum $\sum_{\ell=0}^{j} \sum_{I \in \mathcal{I}_\ell} |\theta_I^*|^2$ can be used to characterize the deviation of f from the space of linear functions.

Since the function f is observed with a noise, we cannot calculate directly the coefficients θ_I^* and we consider the least squares estimator $\widehat{\boldsymbol{\theta}}(j)$ of the vector $\boldsymbol{\theta}^*(j)$ which is defined by minimization of the sum of residuals squared:

$$\widehat{\boldsymbol{\theta}}(j) = \operatorname*{arginf}_{\boldsymbol{\theta}(j)} \|Y - \Psi(j)\boldsymbol{\theta}(j)\|_n^2 = \operatorname*{arginf}_{\{\theta_I \in \mathcal{I}(j)\}} \sum_{i=1}^{n} \left(Y_i - \sum_{I \in \mathcal{I}(j)} \theta_I \psi_I(X_i) \right)^2. \tag{2.10}$$

Here $\boldsymbol{Y}$ means the column-vector with elements $Y_i, i = 1, \ldots, n$.

Define $V(j)$ as the pseudo-inverse of $\Psi(j)^\top \Psi(j)$, $V(j) = \left(\Psi(j)^\top \Psi(j)\right)^-$ It is a symmetric $N(j) \times N(j)$ matrix (by $v_{I,I'}$ we denote its elements, $I, I' \in \mathcal{I}(j)$) and

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

The proposed test is based on the centered and standardized sum of empirical coefficients squares: $\sum_{\ell=0}^{j} \sum_{I \in \mathcal{I}_\ell} |\widehat{\theta}_I|^2$ for some j. This idea goes back to Neyman (1937) 'smooth' test. Ingster (1982, 1993) suggested the special choice of j depending on the smoothness properties of the function f which allows for a rate optimal testing. Spokoiny (1996) extended the method of Ingster (1993) to adaptive testing by considering all such tests for different j simultaneously. Here we slightly modify that approach by considering the family of levelwise tests, that is, for every level j, we construct a test statistic based only on the empirical Haar coefficients $\widehat{\theta}_I$ for $I \in \mathcal{I}_j$, and the resulting test is defined as the maximum of all levelwise ones.

Let some number $j(n)$ be fixed such that $2^{j(n)+1} < n$ and let, for every $j \leq j(n)$, the estimate $\widehat{\boldsymbol{\theta}}(j)$ be defined by (2.10). Denote by $\widehat{\boldsymbol{\theta}}_j$ the part of the vector $\widehat{\boldsymbol{\theta}}(j)$ corresponding to the level j,

$$\widehat{\boldsymbol{\theta}}_j = (\widehat{\theta}_I, I \in \mathcal{I}_j).$$

We analyze every such vector separately for all $j \leq j(n)$. Namely, for every $j \leq j(n)$, we use the statistic based on the sum $\sum_{I \in \mathcal{I}_j} |\widehat{\theta}_I|^2$ corresponding to jth resolution level.

To define our test, we have to study the properties of such sums under the null hypothesis, i.e. when the function f is linear: $f(x) = \theta_0 + \theta_1 x$. We have already mentioned that in this situation $f = \Psi(j)\boldsymbol{\theta}^*$ where $\theta_0^* = \theta_0$, $\theta_1^* = \theta_1$ and all remaining coefficients θ_I^* vanish. Therefore, using the model equation $\boldsymbol{Y} = f + \boldsymbol{\xi}$, we obtain

$$\begin{aligned}\widehat{\boldsymbol{\theta}}(j) &= V(j)\Psi(j)^\top(f+\boldsymbol{\xi}) \\ &= V(j)\Psi(j)^\top\Psi(j)\boldsymbol{\theta}^* + V(j)\Psi(j)^\top\boldsymbol{\xi} \\ &= \boldsymbol{\theta}^* + V(j)\Psi(j)^\top\boldsymbol{\xi}. \end{aligned} \tag{2.12}$$

Obviously $\boldsymbol{\zeta}(j) = V(j)\Psi(j)^\top\boldsymbol{\xi}$ is a random vector in $\mathbb{R}^{N(j)}$ with zero mean. Moreover, it holds for its covariance matrix

$$\begin{aligned}\boldsymbol{E}\boldsymbol{\zeta}(j)\boldsymbol{\zeta}(j)^\top &= V(j)\Psi(j)^\top\boldsymbol{E}\boldsymbol{\xi}\boldsymbol{\xi}^\top\Psi(j)V(j) \\ &= \sigma^2 V(j)\Psi(j)^\top\Psi(j)V(j) = \sigma^2 V(j). \end{aligned} \tag{2.13}$$

Due to (2.12), the subvector $\widehat{\boldsymbol{\theta}}_j$ of $\widehat{\boldsymbol{\theta}}(j)$ coincides under the null with the corresponding subvector $\boldsymbol{\zeta}_j$ of the vector $\boldsymbol{\zeta}(j)$, and it holds under the null in view of (2.13)

$$\begin{aligned}\boldsymbol{E}\widehat{\boldsymbol{\theta}}_j &= \boldsymbol{E}\boldsymbol{\zeta}_j = 0, \\ \boldsymbol{E}\widehat{\boldsymbol{\theta}}_j\widehat{\boldsymbol{\theta}}_j^\top &= \boldsymbol{E}\boldsymbol{\zeta}_j\boldsymbol{\zeta}_j^\top = \sigma^2 V_j\end{aligned}$$

where V_j is the submatrix of $V(j)$ corresponding to the index subset $\mathcal{I}_j$: $V_j = (v_{I,I'}\, I, I' \in \mathcal{I}_j)$. This particularly implies

$$\boldsymbol{E}\sum_{I\in\mathcal{I}_j}|\widehat{\theta}_I|^2 = \boldsymbol{E}\sum_{I\in\mathcal{I}_j}|\zeta_I|^2 = \sigma^2\operatorname{tr}V_j$$

where $\operatorname{tr}A$ denotes the trace of a matrix A. Moreover, for the case of Gaussian errors ξ_i in (1.1), the estimates $\widehat{\theta}_I$ are also Gaussian random variables, and it holds

$$\begin{aligned}\operatorname{Var}\left(\sum_{I\in\mathcal{I}_j}|\widehat{\theta}_I|^2\right) &= \boldsymbol{E}\left(\sum_{I\in\mathcal{I}_j}|\widehat{\theta}_I|^2 - \sigma^2\operatorname{tr}V_j\right)^2 \\ &= \boldsymbol{E}\left(\sum_{I\in\mathcal{I}_j}|\zeta_I|^2 - \sigma^2\operatorname{tr}V_j\right)^2 = 2\sigma^4\operatorname{tr}V_j^2, \end{aligned} \tag{2.14}$$

see (2.13). This leads to the obvious idea to use the centered and standardized sum

$$T_j = \frac{1}{\sqrt{2\sigma^4\operatorname{tr}V_j^2}}\left(\sum_{I\in\mathcal{I}_j}|\widehat{\theta}_I|^2 - \sigma^2\operatorname{tr}V_j\right)$$

as a test statistic. To define our testing procedure, we simply take the maximum of all such statistics over the set of all considered Haar levels j.

2.2. Testing procedure

First we define the finest considered resolution level $j(n)$ which has to satisfy $2^{j(n)+1} < n$ and $n2^{-j(n)} \to \infty$, e.g.

$$j(n) = [\log_2 n - \log_2\log_2 n]. \tag{2.15}$$

where $[a]$ denotes the integer part of a. For each $j \le j(n)$, let $\widehat{\boldsymbol{\theta}}(j)$ be defined by (2.11). Denote by $\widehat{\boldsymbol{\theta}}_j$ the part of the vector $\widehat{\boldsymbol{\theta}}(j)$ corresponding to the level j,

$$\widehat{\boldsymbol{\theta}}_j = (\widehat{\theta}_I,\, I \in \mathcal{I}_j)$$

and let V_j be the submatrix of the matrix $V(j) = \left(\Psi(j)^\top \Psi(j)\right)^-$ corresponding to the level j, i.e. $V_j = (v_{I,I'},\, I, I' \in \mathcal{I}_j)$. We consider χ^2-type statistics

$$S_j = \|\widehat{\boldsymbol{\theta}}_j\|^2 = \sum_{I \in \mathcal{I}_j} \widehat{\theta}_I^2. \tag{2.16}$$

and define test statistics T_j by centering and Studentization of S_j:

$$T_j = \frac{1}{\sqrt{2\widehat{\sigma}^4 \operatorname{tr} V_j^2}} \left(\sum_{I \in \mathcal{I}_j} |\widehat{\theta}_I|^2 - \widehat{\sigma}^2 \operatorname{tr} V_j \right) \tag{2.17}$$

where $\widehat{\sigma}$ is the estimate of the error standard deviation defined in the next subsection. The proposed test rejects the null hypothesis, if at least one such statistic is significantly large, that is,

$$\phi^* = \mathbf{1}\,(T^* > \lambda) \qquad \text{with} \qquad T^* = \max_{j=0,\dots,j(n)} |T_j| \tag{2.18}$$

where λ is a critical value. The choice of λ is discussed in Section 2.4.

2.3. **Estimation of** σ^2

Recall that we assume a homogeneous additive noise in the model (1.1), that is, the errors ξ_i are independent identically distributed random variables fulfilling $\boldsymbol{E}\xi_i = 0$ and $\boldsymbol{E}\xi_i^2 = \sigma^2$. The variance σ^2 is typically unknown in practical applications but this value is important for the definition of our test procedure. Below we discuss how it can be estimated from the data $Y_1, \dots, Y_n$. We suppose for simplicity that the design points are ordered in a way that $X_1 \le \ldots \le X_n$. There are several proposals for variance estimation. One possibility is to estimate σ^2 by the expression of the form $\frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2$, see Rice (1984). We follow the proposal from Gasser et al. (1986) see also Hart (1997, Section 5.3) which provides an unbiased estimate of the variance under the linear null hypothesis.

Define for $i = 2, \dots, n-1$ *pseudo-residuals*

$$\widehat{e}_i = \frac{(X_{i+1} - X_i)}{(X_{i+1} - X_{i-1})} Y_{i-1} + \frac{(X_i - X_{i-1})}{(X_{i+1} - X_{i-1})} Y_{i+1} - Y_i = a_i Y_{i-1} + b_i Y_{i+1} - Y_i$$

which are the result of joining Y_{i+1} and Y_{i-1} by a straight line and taking the difference between this line and Y_i. A variance estimate based on these pseudo-residuals is

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=2}^{n-1} \frac{\widehat{e}_i^2}{a_i^2 + b_i^2 + 1}. \tag{2.19}$$

It is easy to check that $\boldsymbol{E}\widehat{\sigma}^2 = \sigma^2$ if f is a linear function. Some other properties of this estimates are listed in Lemmas 4.1, 4.2 and 4.9 below.

2.4. **Critical value λ**

Here we discuss how to select the critical value λ to provide, at least asymptotically for large n, the condition $\alpha_{f_0}(\phi^*) \le \alpha_0$ for all linear functions f_0. We apply a Monte-Carlo procedure by resampling from the no-response model (which is a particular case of a linear model) with standard normal errors

$$Y^*_{i,m} = \xi^*_{i,m}, \qquad i = 1, \dots, n,$$

for $m = 1, \dots, M$, where the design points $X_1, \dots, X_n$ are the same as for the original model (1.1), $\xi^*_1, \dots, \xi^*_n$ are i.i.d. standard normal random variables and M is the considered number of Monte-Carlo samples.

For every Monte-Carlo sample $Y^*_{1,m}, \dots, Y^*_{n,m}$, we recalculate the test statistic T^*_m from this sample using the previous procedure (including the step of variance estimation). Finally we define the critical value λ as the α_0-level for the set $\{T^*_m, m = 1, \dots, M\}$:

$$\lambda = \min \left\{ t : M^{-1} \sum_{m=1}^{M} \mathbf{1}(T^*_m > t) \le \alpha_0 \right\}.$$

3. Main results

In this section we present the results describing asymptotic properties of the proposed testing procedure. We first discuss the properties of the test under the null and then we consider the power of the test.

3.1. **Behavior under the null**

Let ϕ^* be the test introduced above. Our first result concerns the case of Gaussian errors ξ_i in the model (1.1). In this situation, independently of the design, the nominal level of the test ϕ^* is close to α_0 provided that the number M of Monte-Carlo samples is sufficiently large.

Theorem 3.1. *Let observations Y_i, X_i, $i = 1, \dots, n$, obey the regression model (1.1) with a deterministic design $X_1, \dots, X_n$ and with i.i.d Gaussian errors $\xi_i \sim \mathcal{N}(0, \sigma^2)$. If the function f is linear, $f(x) = \theta_0 + \theta_1 x$, then the value $\alpha_f(\phi^*) = \boldsymbol{P}_f(\phi^* = 1)$ does not depend on the coefficients θ_0 and θ_1 and*

$$\alpha_f(\phi^*) \to \alpha_0 \qquad M \to \infty.$$

Our next result deals with a more general situation when the errors ξ_i are i.i.d. with 6 finite moments. In this case we need some mild regularity conditions on the design.

Recall the notation $A_I = [2^{-j}k, 2^{-j}(k+1)]$ and let M_I stand for the number of design points in A_I: $M_I = \#\{i : X_i \in A_I\}$. Design regularity particularly means that each interval A_I contains enough design points X_i.

(D) (i) It holds for some positive constants C_* and C^* and all $j \le j(n)$

$$\inf_{I \in \mathcal{I}_j} 2^j M_I / n \ge C_*,$$
$$\sup_{I \in \mathcal{I}_j} 2^j M_I / n \le C^*; \tag{3.1}$$

(ii) For some fixed constant C_D and all $j \le j(n)$

$$\operatorname{tr} V_j^2 \ge C_D 2^j;$$

(iii) For some fixed constant C_V and all $j \le j(n)$

$$\|V(j)\| \le C_V$$

Here the norm $\|A\|$ of a symmetric matrix A is understood as the maximal eigenvalue of this matrix;

(iv) For some $D > 0$ and all i, it holds $X_{i+1} - X_i \le Dn^{-1}$.

Condition (D) is trivially fulfilled with $C_* = C^* = C_D = C_V = D = 1$ for the case of the deterministic equidistant design when $V(j)$ is the unit matrix.

Theorem 3.2. *Let observations Y_i, X_i, $i = 1, \ldots, n$, obey the regression model (1.1) with a deterministic design $X_1, \ldots, X_n$ satisfying (D) and with i.i.d. errors ξ_i satisfying $\boldsymbol{E}\xi_i = 0$, $\boldsymbol{E}\xi_i^2 = \sigma^2$ and $\boldsymbol{E}|\xi_i^2 - \sigma^2|^3 \le \sigma^6 C_6$ where C_6 is a fixed constant. If the function f is linear, $f(x) = \theta_0 + \theta_1 x$, then*

$$\alpha_f(\phi^*) \equiv \boldsymbol{P}_f(\phi^* = 1) \le \alpha_0 + \delta_1(n),$$

where $\delta_1(n)$ depends on n, C_6 and the constants C_*, C^*, C_D, C_V from condition (D) only and $\delta_1(n) \to 0$ as $n \to \infty$.

3.2. Sensitivity of the test

Now we state the results concerning the sensitivity of the proposed test ϕ^*. The first assertion presents sufficient conditions for detecting an alternative with a high probability. Next we demonstrate how these conditions can be transferred into a more usual form about the rate of testing against a smooth alternative.

Proposition 3.1. *Let the design $X_1, \ldots, X_n$ obey (D) and the errors $\xi_1, \ldots, \xi_n$ fulfill the conditions of Theorem 3.2. Let then the regression function f be two times continuously differentiable and the second derivative f'' fulfill the condition:*

$$\int_0^1 |f''(x)|^2 \le L^2 \tag{3.2}$$

with some constant L satisfying $8D^3L^2 \le \sigma^2 n^3$. Let also $\boldsymbol{\theta}_j^* = (\theta_I^*, I \in \mathcal{I}_j)$ be the subvector of the vector $\boldsymbol{\theta}^*(j)$ from (2.9) corresponding to j th resolution level and let V_j be the corresponding covariance submatrix, $j = 1, \ldots, j(n)$. If, for some $j \le j(n)$, it holds

$$T_j^* \equiv \frac{\|\boldsymbol{\theta}_j^*\|^2}{\sigma^2\sqrt{2\,\mathrm{tr}\,V_j^2}} \ge 3(\lambda_n^{1/2} + 1)^2, \tag{3.3}$$

with $\lambda_n = \max\{\lambda, 2\sqrt{\log j(n)}\}$, then

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

where $\delta(n)$ depends on n and the constants C_6, C_*, C^*, C_D, C_V only.

We shall show, see Lemma 4.2 that, at least for sufficiently large n, it holds $\lambda \le 2\sqrt{\log j(n)}\,(1 + o_n(1))$. Hence, the result of Proposition 3.1 means that the test ϕ^* detects with a probability close to one any alternative for which at least one from the corresponding values T_j^* exceeds $6\sqrt{\log j(n)}\,(1 + o_n(1))$. Therefore, the error of the second kind may occur with a significant probability only if

$$T_j^* \le 6\sqrt{\log j(n)}\,(1 + o_n(1)), \qquad 0 \le j \le j(n). \tag{3.4}$$

It remains to understand what follows for the function f from these inequalities.

3.3. Power against a smooth alternative

To formulate the results on the power of the test against a smooth alternative, we have to introduce some smoothness conditions on the function f. This can be done in different ways. We choose one based on the accuracy of approximating 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 length 2^{-j}: if $I = (j,k)$, then $A_I = [k2^{-j}, (k+1)2^{-j})$. Next, for a natural number 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 $\mathcal{P}_s(j)$ at the rate 2^{-js}, or, more precisely,

$$\inf_{g \in \mathcal{P}_s(j)} \left[\int_0^1 |f(t) - g(t)|^2 dt \right]^{1/2} \le C_s 2^{-js}$$

where a positive constant C_s depends on s only.

In our conditions we change the integral by summation over observation points. This helps to present the results in a more readable form without changing the sense of required conditions. It can be easily seen that if the design is regular, then the both forms are equivalent up to a constant factor.

Let now a function f be fixed. Let also j_0 be such that $2^{j_0 - 1} \ge s$. Set for $j \ge j_0$

$$r_s(j) = \inf_{g \in \mathcal{P}_s(j-j_0)} \|f - g\|_n = \inf_{g \in \mathcal{P}_s(j-j_0)} \left[\sum_{i=1}^n |f(X_i) - g(X_i)|^2 \right]^{1/2} . \tag{3.5}$$

The quantity $r_s(j)$ characterizes the accuracy of approximation of f by piecewise polynomials. In particular, the Haar approximation corresponds to the case with $s = 1$.

Theorem 3.3. *Let condition (D) hold, the errors $\xi_1, \ldots, \xi_n$ fulfill the conditions of Theorem 3.2, and the regression function f obey (3.2) with a constant L satisfying $8D^3L^2 \le \sigma^2 n^3$. There exist a constant κ depending on the values C_V, C_D, C_*, C^* only, such that if, for some $j \le j(n)$, the following inequality holds true:*

$$\inf_{a,b} \|f - a - b\psi_1\|_n \ge \kappa \left(r_s(j) + \sqrt{2^{j/2}\lambda_n} \right) \tag{3.6}$$

with $\psi_1(x) = x$, then

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

where $\delta(n)$ is shown in Proposition 3.1.

Remark 3.1. It is of interest to compare this result with the existing results on the rate of hypothesis testing. For instance, it was shown in Ingster (1982, 1993) 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 separation rate between the simple null $f \equiv 0$ and a smooth alternative from $W_s(1)$ is $n^{-2s/(4s+1)}$.

For our procedure, the following result is a straightforward corollary of Theorem 3.3 which for the sake of simplicity is formulated for the equidistant design only.

Corollary 3.1. *Let the design $X_1, \ldots, X_n$ be equidistant. (so that condition (D) holds automatically), the errors $\xi_1, \ldots, \xi_n$ fulfill the conditions of Theorem 3.2, and the underlying function f belong to a Sobolev ball $W_s(1)$ and f'' fulfill (3.2) with a constant L satisfying $8L^2 \le \sigma^2 n^3$. Then there exists a constant $C_s > 0$ depending on s only and such that, for n large enough, the inequality*

$$\inf_{a,b} \|f - a - b\psi_1\|_n^2 \ge C_s (n/\lambda_n)^{-\frac{2s}{4s+1}} \tag{3.7}$$

implies

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

where $\delta(n)$ depends on the distribution of the errors ε_i only.

Indeed, under the equidistant design, it holds $r_s(j) \le C n^{1/2} 2^{-js}$ for every function f from $W_s(1)$ with a fixed constant C depending on s only. Now the right hand-side of (3.7) arises via minimization of the sum $C n^{1/2} 2^{-js} + \sqrt{2^{j/2}\lambda_n}$ with respect to j.

By comparison to the mentioned result of Ingster (1982) we observe that the proposed method leads to a near optimal rate up to a log-log multiple in the class of all tests. Moreover, Spokoiny (1996) has shown (for the 'signal + white noise' model) that this separation rate is optimal in the class of all *adaptive* tests. The latter result allows for a straightforward extension to Gaussian regression using the general asymptotic equivalence result, Brown and Low (1996). The additional smoothness condition (3.2) with $L^2 \le \sigma^2 n^3/8$ is required for ensuring a good quality of the pilot estimate of the unknown variance. This assumption is not restrictive since the constant L may rapidly grow with n. In particular, the low bound results from Ingster (1993) and Spokoiny (1996) allow for a straightforward extension under this constraint. We therefore, resume that the proposed test is rate optimal among all adaptive tests (at least for the case of the equidistant design).

Remark 3.2. The result of Theorem 3.3 helps to understand what happens in the case when the design is not regular and, for instance, if there exist some intervals I with $M_I = 0$. It was already mentioned that the procedure applies in this situation as well and the error probability of the first kind is about α_0 at least for n sufficiently large and for Gaussian errors ξ_i. 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. This may occur e.g. in the situation when f is deviated from the best linear approximation only in the domain with very few design points inside.

4. Proofs

In this section we first prove Theorems 3.1 and 3.3 for the case of Gaussian errors ξ_i and then discuss the generalization to the general case.

4.1. Proof of Theorem 3.1

It suffices to check that the distribution of the test statistic T^* based on the Monte-Carlo sample $Y_1^*, \ldots, Y_n^*$ is the same as for the original sample $Y_1, \ldots, Y_n$. The difference between these two samples is only in the linear trend (which can be nontrivial for the original sample but does not appear in the Monte-Carlo one) and in the noise variance (we resample with the error variance 1 instead of σ^2). Note however that the linear trend in the regression function makes no influence on the considered test statistics T_j. Indeed, the numerator of this statistic is defined as the centered sum over $\mathcal{I}_j$ of the the empirical Haar coefficients $\widehat{\theta}_I$ squared, so that the linear trend is removed automatically from the test statistics, see the proof of Theorem 3.3 for more details. Similarly, the estimate $\widehat{\sigma}^2$ of the noise variance σ^2 is based on the pseudo-residuals $\widehat{e}_i$ which are defined in a way that the linear trend in the regression function cancels out, see Lemma 4.1.

Further, for the case of zero trend, both numerator and denominator of each T_j is some quadratic forms of the errors ξ_i which can be represented as $\xi_i = \sigma\widetilde{\xi}_i$ with i.i.d. standard normal variables $\widetilde{\xi}_i$, $i = 1, \ldots, n$. This yields, see (2.19), that the distribution of each test statistic T_j does not depend on σ. The same is obviously true for the maximum T^* and the assertion follows.

4.2. Properties of the estimate $\widehat{\sigma}^2$

Here we discuss the properties of the estimate $\widehat{\sigma}^2$ of the noise variance σ^2. We present two results. The first one describes the properties under the null, and the second one applies under a smooth alternative as well. The results are stated under the Gaussian errors ξ_i. For the extension, see Section 4.5.

Lemma 4.1. *Let the regression function f be linear. Then* $\boldsymbol{E}\widehat{\sigma}^2 = \sigma^2$

$$\boldsymbol{E}\left(\widehat{\sigma}^2 - \sigma^2\right)^2 \le \frac{7\sigma^4}{2(n-2)}.$$

Proof. For the case of a linear function $f(x) = \theta_0 + \theta_1 x$, one easily gets with the coefficients $a_i = \frac{(X_{i+1}-X_i)}{(X_{i+1}-X_{i-1})}$, $b_i = \frac{(X_i - X_{i-1})}{(X_{i+1}-X_{i-1})}$

$$a_i f(X_{i-1}) + b_i f(X_{i+1}) - f(X_i) = 0.$$

Now the model equation (1.1) implies

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=2}^{n-1}|\eta_i|^2$$

with

$$\eta_i = \frac{a_i\xi_{i-1} + b_i\xi_{i+1} - \xi_i}{\sqrt{a_i^2 + b_i^2 + 1}}.$$

To estimate the difference $|\widehat{\sigma}^2 - \sigma^2|$, we apply Proposition 5.1 from the Appendix. Let $\boldsymbol{\eta}$ denote the vector $(\eta_2, \ldots, \eta_{n-1})^\top$. Obviously $\boldsymbol{E}\boldsymbol{\eta} = 0$. Define $\Sigma = \boldsymbol{E}\boldsymbol{\eta}\boldsymbol{\eta}^\top$. Observe first that

$$\frac{1}{n-2}\operatorname{tr}\Sigma = \frac{1}{n-2}\sum_{i=2}^{n-1}\frac{\sigma^2(a_i^2 + b_i^2 + 1)}{(a_i^2 + b_i^2 + 1)} = \sigma^2$$

which implies the equality $\boldsymbol{E}\widehat{\sigma}^2 = \sigma^2$ by Proposition 5.1.

Next, it is easy to check that $2\max\{a_i^2, b_i^2\} \le a_i^2 + b_i^2 + 1$. Now, it obviously holds:

$$\begin{aligned}
\boldsymbol{E}\eta_i^2 &= \sigma^2, \\
|\boldsymbol{E}\eta_i\eta_{i+1}| &\le \sqrt{\boldsymbol{E}\eta_i^2\boldsymbol{E}\eta_{i+1}^2} = \sigma^2, \\
|\boldsymbol{E}\eta_{i-1}\eta_{i+1}| &= \frac{\sigma^2 b_{i-1}a_{i+1}}{\sqrt{(a_{i-1}^2 + b_{i-1}^2 + 1)(a_{i+1}^2 + b_{i+1}^2 + 1)}} \le \sigma^2/2, \\
\boldsymbol{E}\eta_i\eta_{i'} &= 0, \qquad |i' - i| > 2,
\end{aligned}$$

This allows to estimate $\operatorname{tr}\Sigma^2$ as follows:

$$\begin{aligned}
\frac{1}{(n-2)^2}\operatorname{tr}\Sigma^2 &= \frac{1}{(n-2)^2}\sum_{i=2}^{n-1}\sum_{j=2}^{n-1}(\boldsymbol{E}\eta_i\eta_j)^2 \\
&= \frac{1}{(n-2)^2}\sum_{i=2}^{n-1}\left[(\boldsymbol{E}\eta_{i-1}\eta_i)^2+(\boldsymbol{E}\eta_i^2)^2+(\boldsymbol{E}\eta_i\eta_{i+1})^2+(\boldsymbol{E}\eta_i\eta_{i-2})^2+(\boldsymbol{E}\eta_i\eta_{i+2})^2\right] \\
&\le \frac{\sigma^4}{(n-2)^2}\sum_{i=2}^{n-1}(1+1+1+1/4+1/4) \\
&= \frac{7\sigma^4}{2(n-2)}
\end{aligned}$$

which implies the second assertion of the lemma by Proposition 5.1. □

Next we show that $\widehat{\sigma}^2$ estimate the true value σ^2 at the rate $n^{-1/2}$ under a mild assumption on the regression function f and the design $X_1,\ldots,X_n$. We again assume that the design points are renumbered to provide $X_1 \le X_2 \le \ldots \le X_n$.

Lemma 4.2. *Let the design $X_1,\ldots,X_n$ fulfill $X_{i+1}-X_i \le Dn^{-1}$ with some constant D. Let next the regression function f from (1.1) fulfills the condition*

$$\int_0^1 |f''(x)|^2 dx \le L^2$$

for some $L \ge 0$ satisfying $8L^2D^3n^{-3} \le \sigma^2$. Then

$$\boldsymbol{E}(\widehat{\sigma}^2-\sigma^2)^2 \le 9\sigma^4(n-2)^{-1}.$$

Proof. The definition of the coefficients a_i and b_i, see Section 2.3, provides for any linear function $\ell(x)$ the identity $a_i\ell(X_{i-1})+b_i\ell(X_{i+1})-\ell(X_i)=0$. The application of $\ell(x)=f'(X_i)(x-X_i)$ yields

$$\begin{aligned}
&|a_if(X_{i-1})+b_if(X_{i+1})-f(X_i)| \\
&\quad\le a_i\,|f(X_{i-1})-f(X_i)-f'(X_i)(X_{i-1}-X_i)| + b_i\,|f(X_{i+1})-f(X_i)-f'(X_i)(X_{i+1}-X_i)|.
\end{aligned}$$

Let $f(X_{i-1})-f(X_i)=(X_{i-1}-X_i)f'(u)$ for some $u\in[X_{i-1},X_i]$. Then, by the Cauchy-Schwarz inequality and the condition $X_i-X_{i-1}\le Dn^{-1}$,

$$\begin{aligned}
&|f(X_{i-1})-f(X_i)-f'(X_i)(X_{i-1}-X_i)| \le (X_i-X_{i-1})\left|\int_u^{X_i} f''(s)ds\right| \\
&\quad\le (X_i-X_{i-1})\int_{X_{i-1}}^{X_i}|f''(x)|dx \le (X_i-X_{i-1})^{3/2}\left(\int_{X_{i-1}}^{X_i}|f''(x)|^2dx\right)^{1/2} \le (Dn^{-1})^{3/2}L_i
\end{aligned}$$

with $L_i^2=\int_{X_{i-1}}^{X_i}|f''(x)|^2dx$, and similarly for $|f(X_{i+1})-f(X_i)-f'(X_i)(X_{i+1}-X_i)|$. These two bounds imply

$$|a_if(X_{i-1})+b_if(X_{i+1})-f(X_i)| \le (a_iL_i+b_iL_{i+1})(Dn^{-1})^{3/2}. \tag{4.1}$$

Next, define

$$\eta_i = \frac{a_i \xi_{i-1} + b_i \xi_{i+1} - \xi_i}{\sqrt{a_i^2 + b_i^2 + 1}}, \qquad \Delta_i = \frac{a_i f(X_{i-1}) + b_i f(X_{i+1}) - f(X_i)}{\sqrt{a_i^2 + b_i^2 + 1}}.$$

Then

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=2}^{n-1} |\Delta_i + \eta_i|^2 .$$

To estimate the difference $\widehat{\sigma}^2 - \sigma^2$, we apply Proposition 5.2 from the Appendix. Let $\eta = (\eta_2, \ldots, \eta_{n-1})^\top$. We know, see the proof of Lemma 4.1, that $\boldsymbol{E}\boldsymbol{\eta} = 0$ and the matrix $\Sigma = \boldsymbol{E}\boldsymbol{\eta}\boldsymbol{\eta}^\top$ fulfills

$$\frac{1}{n-2} \operatorname{tr} \Sigma = \sigma^2, \qquad \frac{1}{(n-2)^2} \operatorname{tr} \Sigma^2 \le \frac{7\sigma^4}{2(n-2)}.$$

The inequality $2\max\{a_i^2, b_i^2\} \le 1 + a_i^2 + b_i^2$ and (4.1) provide

$$\begin{aligned}
\|\Delta\|^2 &= \sum_{i=2}^{n-1} \Delta_i^2 \le D^3 n^{-3} \sum_{i=2}^{n-1} \frac{(a_i L_i + b_i L_{i+1})^2}{a_i^2 + b_i^2 + 1} \\
&\le D^3 n^{-3} \sum_{i=2}^{n-1} (L_i^2 + L_{i+1}^2) \le 2D^3 n^{-3} \int_0^1 |f''(x)|^2 dx \le 2D^3 n^{-3} L^2
\end{aligned}$$

The application of Proposition 5.2 from the Appendix with $\boldsymbol{c} = \frac{\Delta}{\sqrt{n-2}}$ and $\boldsymbol{\varepsilon} = \frac{\boldsymbol{\eta}}{\sqrt{n-2}}$ yields

$$\boldsymbol{E}\left(\widehat{\sigma}^2 - \sigma^2\right)^2 = \boldsymbol{E}\left(\|\boldsymbol{c} + \boldsymbol{\varepsilon}\|^2 - \operatorname{tr} V\right)^2 \le \|\boldsymbol{c}\|^4 + 4\|\boldsymbol{c}\|^2 \sqrt{\operatorname{tr} V^2} + 2 \operatorname{tr} V^2.$$

where $V = (n-2)^{-1}\Sigma$. This along with the inequalities $\operatorname{tr} V^2 \le \frac{7\sigma^4}{2(n-2)}$ and $4\|\boldsymbol{c}\|^2 \le 8L^2 D^3 n^{-3} (n-2)^{-1} \le \sigma^2 (n-2)^{-1}$ imply the required assertion. □

Lemma 4.3. *Let $N_j = 2^j$ denote the number of elements in the set $\mathcal{I}_j$. It holds*

$$\frac{\operatorname{tr} V_j}{\sqrt{2 \operatorname{tr} V_j^2}} \le \sqrt{N_j/2}.$$

Proof. Clearly

$$\operatorname{tr} V_j^2 = \sum_{I \in \mathcal{I}_j} \sum_{I' \in \mathcal{I}_j} v_{I,I'}^2 \ge \sum_{I \in \mathcal{I}_j} v_{I,I}^2 .$$

Next, the Cauchy-Schwarz inequality implies

$$N_j^{-1} \operatorname{tr} V_j = N_j^{-1} \sum_{I \in \mathcal{I}_j} v_{I,I} \le \left(N_j^{-1} \sum_{I \in \mathcal{I}_j} v_{I,I}^2 \right)^{1/2}$$

and the assertion follows. □

Lemma 4.4. *Let λ be the critical value of the test selected by the testing procedure. If design $X_1, \ldots, X_n$ fulfills (D), then, for n sufficiently large,*

$$\lambda \le 2\sqrt{\log j(n)}\,(1 + o_n(1)) .$$

Proof. Recall that the critical value λ corresponds to the $1-\alpha_0$-quantile of the distribution of the test statistic $T^* = \max_{j \le j(n)} T_j$ under the no-response model $f(x) \equiv 0$ and under the assumption of standard normal errors ξ_i, $i = 1, \ldots, n$. In such a situation, the subvector $\widehat{\theta}_j$ of $\widehat{\theta}(j)$ coincides with the Gaussian vector $\boldsymbol{\zeta}_j \sim \mathcal{N}(0, V_j)$, see Section 2.1, and hence the corresponding statistic T_j can be represented in the form

$$T_j = \frac{\|\boldsymbol{\zeta}_j\|^2 - \widehat{\sigma}^2 \operatorname{tr} V_j}{\widehat{\sigma}^2 \sqrt{2 \operatorname{tr} V_j^2}}.$$

and it suffices to show that

$$\boldsymbol{P}\left(\max_{j \le j(n)} T_j > 2\sqrt{\log j(n)}\bigl(1 + \delta_1(n)\bigr)\right) \le \delta_2(n)$$

with two numeric sequences $\delta_1(n) \to 0$ and $\delta_2(n) \to 0$.

Now, for every $z \ge 1$ and $a \in (0,1)$,

$$\begin{aligned}
\left\{T_j > \frac{z+1}{a}\right\} &= \left\{\frac{\|\boldsymbol{\zeta}_j\|^2 - \widehat{\sigma}^2 \operatorname{tr} V_j}{\sigma^2\sqrt{2 \operatorname{tr} V_j^2}} > \frac{(z+1)\widehat{\sigma}^2}{a\sigma^2}\right\} \\
&\subseteq \left\{\frac{\|\boldsymbol{\zeta}_j\|^2 - \sigma^2 \operatorname{tr} V_j}{\sigma^2\sqrt{2 \operatorname{tr} V_j^2}} > z\right\} \cup \left\{\frac{(\widehat{\sigma}^2 - \sigma^2) \operatorname{tr} V_j}{\sigma^2\sqrt{2 \operatorname{tr} V_j^2}} > 1\right\} \cup \left\{\frac{\widehat{\sigma}^2}{\sigma^2} < a\right\}.
\end{aligned}$$

This clearly yields in view of Lemma 4.3

$$\begin{aligned}
&\boldsymbol{P}\left(\max_{j \le j(n)} T_j > \frac{z+1}{a}\right) \\
&\quad \le \boldsymbol{P}\left(\frac{\widehat{\sigma}^2}{\sigma^2} < a\right) + \boldsymbol{P}\left(\frac{\widehat{\sigma}^2}{\sigma^2} - 1 > \frac{1}{\sqrt{N_{j(n)}/2}}\right) + \sum_{j=0}^{j(n)} \boldsymbol{P}\left(\frac{\|\boldsymbol{\zeta}_j\|^2 - \sigma^2 \operatorname{tr} V_j}{\sigma^2\sqrt{2 \operatorname{tr} V_j^2}} > z\right).
\end{aligned}$$

We apply this bound with $z = 1 + v_n$ and $a = 1 - v_n^{-1}$ where $v_n = 2\sqrt{\log j(n)}$. Let j_1 be the minimal integer satisfying $C_D 2^{j_1} \ge 2C_V^2 v_n^2$. It follows from condition (D) that $v_n \le \|V_j\|^{-1}\sqrt{\operatorname{tr} V_j^2/2}$ for all $j \ge j_1$. An application of Proposition 5.1 from the Appendix with $\gamma = v_n$ and $t = 1$ for $j \ge j_1$ and with $\gamma = 1$ and $t = v_n$ allows to bound

$$\boldsymbol{P}\left(\frac{\|\boldsymbol{\zeta}_j\|^2 - \sigma^2 \operatorname{tr} V_j}{\sigma^2\sqrt{2 \operatorname{tr} V_j^2}} > v_n + 1\right) \le \begin{cases} e^{-v_n^2/4 - v_n/2} & j \ge j_1, \\ e^{-v_n/2} & \text{otherwise.} \end{cases}$$

Lemma 4.1 and the Chebyshev inequality provide

$$\begin{aligned}
&\boldsymbol{P}\left(\frac{\widehat{\sigma}^2}{\sigma^2} < 1 - v_n^{-1}\right) + \boldsymbol{P}\left(\frac{\widehat{\sigma}^2}{\sigma^2} - 1 > \frac{1}{\sqrt{N_{j(n)}/2}}\right) \\
&\qquad \le v_n^2 \sigma^{-4} \boldsymbol{E}\left(\widehat{\sigma}^2 - \sigma^2\right)^2 + \frac{2\boldsymbol{E}\left(\widehat{\sigma}^2 - \sigma^2\right)^2}{\sigma^4 N_{j(n)}} \le \frac{7v_n^2}{2(n-2)} + \frac{N_{j(n)}}{2(n-2)} = \delta_3(n) \to 0,
\end{aligned}$$

since, by definition of $j(n)$, it holds $n/N_{j(n)} \to \infty$. Therefore,

$$\begin{aligned}
\boldsymbol{P}\left(T^* > \frac{2+v_n}{1-v_n^{-1}}\right) &\le \delta_3(n) + \sum_{j=0}^{j_1-1} e^{-v_n/2} + \sum_{j=j_1}^{j(n)} e^{-v_n^2/4 - v_n/2} \\
&\le \delta_3(n) + \log_2(2C_V^2 v_n^2/C_D) e^{-v_n/2} + \tfrac{1+j(n)}{j(n)} e^{-v_n/2} \to 0, \quad n \to \infty.
\end{aligned}$$

□

4.3. Proof of Proposition 3.1

We again restrict ourselves to the case of Gaussian errors ξ_i in (1.1). Recall that the vector $\widehat{\boldsymbol{\theta}}_j$ is defined as the subvector of $\widehat{\boldsymbol{\theta}}(j) = \left(\Psi(j)^\top \Psi(j)\right)^{-1} \Psi(j)^\top \boldsymbol{Y}$, $j \le j(n)$. The model equation (1.1) yields

$$\widehat{\boldsymbol{\theta}}(j) = \left(\Psi(j)^\top \Psi(j)\right)^{-1} \Psi(j)^\top (f + \boldsymbol{\xi}) = \boldsymbol{\theta}^*(j) + \boldsymbol{\zeta}(j)$$

with $\boldsymbol{\theta}^*(j) = V(j)\Psi(j)^\top f$ and $\boldsymbol{\zeta}(j) = V(j)\Psi(j)^\top \boldsymbol{\xi}$ where $V(j) = \left(\Psi(j)^\top \Psi(j)\right)^{-1}$. Hence $\widehat{\boldsymbol{\theta}}_j = \boldsymbol{\theta}_j + \boldsymbol{\zeta}_j$ where $\boldsymbol{\theta}_j^*$ (resp. $\boldsymbol{\zeta}_j$) is the subvector of $\boldsymbol{\theta}^*(j)$ (resp. of $\boldsymbol{\zeta}(j)$) corresponding to the jth resolution level. This particularly implies that $\boldsymbol{\zeta}_j$ is a zero mean random vector with the covariance matrix V_j which is the submatrix of the matrix $V(j) = \left(\Psi(j)^\top \Psi(j)\right)^{-1}$. Moreover, if the errors ξ_i in (1.1) are Gaussian, then ζ_j is a Gaussian random vector with parameters $(0, V_j)$ for each $j \le j(n)$.

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

$$T_j^* = \frac{\|\boldsymbol{\theta}_j^*\|^2}{\sigma^2 \sqrt{2 \operatorname{tr} V_j^2}} \ge 3(\lambda_n^{1/2} + 1)^2 \tag{4.2}$$

with $\lambda_n = \max\{\lambda, 2\sqrt{\log j(n)}\}$. We shall show that under this condition

$$\boldsymbol{P}_f(T_j < \lambda) \le \delta(n) \to 0, \qquad n \to \infty, \tag{4.3}$$

which obviously implies the assertion.

Observe first that

$$\begin{aligned}
\boldsymbol{P}(T_j < \lambda) &= \boldsymbol{P}\left(\|\boldsymbol{\theta}_j^* + \boldsymbol{\zeta}_j\|^2 - \widehat{\sigma}^2 \operatorname{tr} V_j < \lambda \widehat{\sigma}^2 \sqrt{2 \operatorname{tr} V_j^2}\right) \\
&\le \boldsymbol{P}\left(\|\boldsymbol{\theta}_j^* + \boldsymbol{\zeta}_j\|^2 - \sigma^2 \operatorname{tr} V_j < \lambda \sigma^2 \sqrt{2 \operatorname{tr} V_j^2} + (\widehat{\sigma}^2 - \sigma^2)\left(\lambda \sqrt{2 \operatorname{tr} V_j^2} + \operatorname{tr} V_j\right)\right) \\
&\le \boldsymbol{P}\left(\|\boldsymbol{\theta}_j^* + \boldsymbol{\zeta}_j\|^2 - \sigma^2 \operatorname{tr} V_j - \|\boldsymbol{\theta}_j^*\|^2 < (\lambda + \lambda_n^{1/2})\sigma^2 \sqrt{2 \operatorname{tr} V_j^2} - \|\boldsymbol{\theta}_j^*\|^2\right) \\
&\qquad + \boldsymbol{P}\left((\widehat{\sigma}^2 - \sigma^2)\left(\lambda \sqrt{2 \operatorname{tr} V_j^2} + \operatorname{tr} V_j\right) < -\sigma^2 \lambda_n^{1/2} \sqrt{2 \operatorname{tr} V_j^2}\right).
\end{aligned}$$

By Lemma 4.3 $\operatorname{tr} V_j\,(2\operatorname{tr} V_j^2)^{-1/2} \le \sqrt{N_j/2} \le \sqrt{N_{j(n)}/2}$ for all $j \le j(n)$. Further, by Lemma 4.2

$$
\begin{aligned}
&\boldsymbol{P}\left(\widehat{\sigma}^2 - \sigma^2 < -\frac{\sigma^2 \lambda_n^{1/2} \sqrt{2\operatorname{tr} V_j^2}}{\lambda\sqrt{2\operatorname{tr} V_j^2} + \operatorname{tr} V_j}\right) \\
&\qquad \le \frac{\left(\lambda\sqrt{2\operatorname{tr} V_j^2} + \operatorname{tr} V_j\right)^2}{\sigma^4 \lambda_n 2\operatorname{tr} V_j^2} \boldsymbol{E}\left(\widehat{\sigma}^2 - \sigma^2\right)^2 \le \frac{9(\lambda + \sqrt{N_{j(n)}/2})^2}{\lambda_n (n-2)} = \delta_4(n), \qquad n \to \infty
\end{aligned}
$$

since $n/N_{j(n)} = n 2^{-j(n)} \to \infty$.

Next, for every positive u, the inequality $\|\boldsymbol{\theta}\| \ge 3u$ implies $\|\boldsymbol{\theta}\|^2 - 2u\|\boldsymbol{\theta}\| - 3u^2 \ge 0$. Coupled with (4.2), this ensures with $u = 3^{-1/2}(\lambda_n^{1/2} + 1)\tau_j$ and $\tau_j = \sigma(2\operatorname{tr} V_j^2)^{1/4}$ that

$$
\begin{aligned}
\|\boldsymbol{\theta}_j^*\|^2 &\ge \sqrt{4/3}\|\boldsymbol{\theta}_j^*\|(\lambda_n^{1/2}+1)\tau_j + (\lambda_n^{1/2}+1)^2\tau_j^2 \\
&\ge \|\boldsymbol{\theta}_j^*\|(\lambda_n^{1/2}+1)\tau_j + (\lambda_n + 2\lambda_n^{1/2}+1)\tau_j^2.
\end{aligned}
$$

Now Proposition 5.2 from the Appendix with $\gamma = 1$ and $t = \lambda_n^{1/2}$ implies

$$
\begin{aligned}
&\boldsymbol{P}\left(T_j < \lambda\right) \\
&\quad \le \boldsymbol{P}\left(\|\boldsymbol{\theta}_j^* + \boldsymbol{\zeta}_j\|^2 - \sigma^2\operatorname{tr} V_j - \|\boldsymbol{\theta}_j^*\|^2 < -(\lambda_n^{1/2}+1)\|\boldsymbol{\theta}_j^*\|\tau_j - (\lambda_n^{1/2}+1)\tau_j^2\right) + \delta_4(n) \\
&\quad \le 2e^{-\lambda_n^{1/2}/2} + \delta_4(n) \to 0, \qquad n \to \infty
\end{aligned}
$$

as required.

4.4. **Proof of Theorem 3.3**

For the proof, we use the result of Proposition 3.1. Namely we show that the condition (3.6) of the theorem with κ large enough contradict to the constraints

$$
T_j^* \le t_n, \qquad j \le j(n), \tag{4.4}
$$

with $t_n = 3\left(1 + \lambda_n^{1/2}\right)^2$ and $\lambda_n = \max\{\lambda, 2\sqrt{\log j(n)}\}$.

We begin by reduction of the problem of testing a linear hypothesis to the problem with a simple null hypothesis. Define coefficients θ_0, θ_1 by

$$
(\theta_0, \theta_1) = \operatorname*{arginf}_{(a,b)} \|f - a - b\psi_1\|_n = \operatorname*{arginf}_{(a,b)} \sum_{i=1}^n (f(X_i) - a - bX_i)^2.
$$

and set

$$
f_0 = f - \theta_0 - \theta_1\psi_1.
$$

Note that for all $j \ge 0$, the vectors $\boldsymbol{\theta}^*(j) = V(j)\Psi(j)f$ and $\boldsymbol{\theta}(j) = V(j)\Psi(j)f_0$ have the same components except the first two. Obviously the smoothness properties of f and f_0 also coincide and

$$
\inf_{a,b} \|f - a - b\psi_1\|_n = \inf_{a,b} \|f_0 - a - b\psi_1\|_n
$$

Recall also, that the linear trend in the regression function has no influence on our variance estimator $\widehat{\sigma}^2$. Hence, replacing f by f_0 changes nothing in the test behaviour and we may suppose from the beginning that the coefficients θ_0^* and θ_1^* of the vector $\boldsymbol{\theta}^*(j)$ vanish.

About this new function f we know that

$$\|f\|_n = \inf_{a,b} \|f - a - b\psi_1\|_n \geq \varrho(n),$$

$$\inf_{g \in \mathcal{P}_s(j)} \|f - g\|_n = r_s(j), \tag{4.5}$$

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

Next we rewrite the constraints from (4.4) in term of the vectors $\|\boldsymbol{\theta}_j^*\|$, $j \leq j(n)$. Recall that $\boldsymbol{\theta}_j^*$ is the subvector of $\boldsymbol{\theta}^*(j)$ corresponding to j th level, and V_j is the corresponding submatrix of $V(j)$.

Let $\mathcal{L}(j)$ stand for the linear space generated by functions ψ_I, $I \in \mathcal{I}(j)$. We denote also by $\Pi(j)f$ 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 onto the space of linear functions (and hence, $\Pi(0)f = 0$) and, by definition of $\boldsymbol{\theta}(j)$,

$$\Pi(j)f = \sum_{I \in \mathcal{I}(j)} \theta_I^* \psi_I \tag{4.6}$$

where θ_I 's are the coefficients of the vector $\boldsymbol{\theta}^*(j)$.

Lemma 4.5. *For each* $1 \leq j \leq j(n)$,

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

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

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

When denoting $f(j) = \Pi(j)f$, one has $\Pi(j-1)f = \Pi(j-1)f(j)$ and we have to show that

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

In view of (4.6)

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

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 ortonormal w.r.t. to the inner product $\|\cdot\|_n$ and particularly

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

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

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

and the assertion follows by the triangle inequality. $\square$

Lemma 4.6. *Given* $j \le j(n)$, *let (4.4) hold true for all* $\ell \le j$. *Then*

$$\|\Pi(j)f\|_n^2 \le \kappa_1 \sigma C_V 2^{j/2} t_n \tag{4.7}$$

with $\kappa_1 = 2^{1/2}(2^{1/4}-1)^{-2}$.

Proof. Recursive application of Lemma 4.5 gives

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

Here we have used that $\Pi(0)f = 0$. Now (4.4) and $(D.iii)$ yield

$$\|\boldsymbol{\theta}_\ell^*\|^2 \le \sigma^2 t_n \sqrt{2\,\mathrm{tr}\, V_\ell^2} \le \sigma^2 t_n \sqrt{C_V^2 2^{\ell+1}}$$

and thus,

$$\|\Pi(j)f\|_n \le \sum_{\ell=1}^{j} \sigma \left(2^{\ell/2} t_n C_V\right)^{1/2} = \sigma (C_V t_n)^{1/2} \sum_{\ell=1}^{j} 2^{\ell/4}$$

and the assertion follows by simple algebra. $\square$

Let now j_0 fulfill $2^{j_0} > s$ and $\mathcal{P}_s(j-j_0)$ denote the space of piecewise polynomials with piece length $2^{-(j-j_0)}$. Let now some $j \le j(n)$ be fixed and let $g \in \mathcal{P}_s(j-j_0)$ be such that

$$\|f - g\|_n \le r_s(j).$$

Lemma 4.7. *There is a constant* $\kappa_2 > 0$ *depending on* C_*, C^* *and* s *only and such that for each* j *with* $j_0 \le j \le j(n)$

$$\|f\|_n \le \kappa_2 \left\{\|\Pi(j)f\|_n + r_s(j)\right\}.$$

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

$$\|f\|_n \le \|g\|_n + r_s(j)$$

and, since $\Pi(j)$ is a projector,

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

and the assertion follows from

$$\|g\|_n^2 \le \kappa_3 \|\Pi(j)g\|_n^2.$$

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

$$\sum_{A_I} [\Pi(j)P(X_i)]^2 \ge \kappa_3 \sum_{A_I} P^2(X_i)$$

where the constant κ_3 depends on C_*, C^* and s only. The similar fact with integration instead of summation over the design points in A_I has been stated in Ingster (1993) and we present here only a sketch of the proof for our situation.

The key idea of the proof can be formulated as a separate statement.

Lemma 4.8. *Let $P(x)$ be a polynomial of degree $s-1$ and let m be an integer with $m > s$. Define $A_k = [(k-1)/m, k/m)$ for $k = 1, \ldots, m$. Then for every measure μ on $[0,1]$ with $0 < C_* \le \mu(A_k) \le C^* > 0$ for all $k \le m$,*

$$\sum_{k=1}^{m} \left[\int_{A_k} P(x)\mu(dx) \right]^2 \ge \kappa_3 \int_0^1 P^2(x)\mu(dx).$$

with a positive number κ_3 depending on C_*, C^* and s only.

Proof. Let $a = (a_0, \ldots, a_{s-1})$ be the vector of coefficients of P. Without loss of generality, we may assume that $\|a\|_\infty = \max_{j=0,\ldots,s-1}\{|a_j|\} \le 1$. Obviously, both

$$\begin{aligned} \|a\|_{\mu,1}^2 &= \sum_{k=1}^{m} \left(\int_{A_k} P(x)\mu(dx) \right)^2, \\ \|a\|_{\mu,2}^2 &= \int_0^1 P^2(x)\mu(dx) \end{aligned}$$

are some Euclidean norms in the space $\mathbb{R}^s$. Next, $\|a\|_{\mu,2} = 0$ only if $a = 0$ i.e. $P(x) \equiv 0$ and the same applies for $\|a\|_{\mu,1}$, since $P(x)$ has at most $s-1$ roots and μ is supported on $m > s$ disjoint intervals. Note also that $\|a\|_{\mu,1}$ and $\|a\|_{\mu,2}$ are continuous functionals of a and μ and the space $\mathcal{M}_m(C_*, C^*)$ of measures μ on $[0,1]$ satisfying the condition of the lemma is compact in the weak topology. Hence,

$$\sup_{a:\|a\|_\infty \le 1} \sup_{\mu \in \mathcal{M}_m(C_*,C^*)} \frac{\|a\|_{\mu,2}}{\|a\|_{\mu,1}} = \kappa_3 < \infty$$

as required. $\square$

Application of this result to each interval A_I, $I \in \mathcal{I}_{j-j_0}$ yields the desirable assertion. $\square$

The results of Lemma 4.5 through 4.7 yield the inequality

$$\|f\|_n \le \kappa_2 \left(r_s(j) + \sqrt{\kappa_1 C_V 2^{j/2} \lambda_n} \right)$$

which contradicts to the constraints $\|f\|_n \ge \kappa \left(r_s(j) + \sqrt{2^{j/2}\lambda_n} \right)$ if $\kappa > \kappa_2$, and the theorem is proved.

4.5. **Proof of Theorem 3.2**

Now we disregard the assumption that the errors ξ_i in (1.1) are normally distributed and assume only that they have 6 finite moments. We outline the proof of Theorem 3.2 only. Proposition 3.1 can be considered similarly.

Lemma 4.9. *Let the errors ξ_i in (1.1) be i.i.d. and satisfy $\boldsymbol{E}\xi_i = 0$, $\xi_i^2 = \sigma^2$ and $\boldsymbol{E}\left|\xi_i^2 - \sigma^2\right|^3 \le C_6\sigma^6$. Define $s_4^2 = 2\sigma^{-4}\boldsymbol{E}(\xi_1^2 - \sigma^2)^2$. If the regression function f is linear then*

$$\boldsymbol{E}(\widehat{\sigma}^2 - \sigma^2)^2 \le \frac{(s_4 + 1/2)\sigma^4}{n-2}.$$

Proof. Similarly to the Gaussian case discussed in Section 4.2, it suffices to consider the case of the no-response model with the vanishing regression function. In this case, the variance estimate $\widehat{\sigma}^2$ is a quadratic form of the errors ξ_i which allows for the following representation:

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=2}^{n-1}\frac{(a_i\xi_{i-1} + b_i\xi_{i+1} - \xi_i)^2}{a_i^2 + b_i^2 + 1}$$

where $a_i = \frac{(X_{i+1}-X_i)}{(X_{i+1}-X_{i-1})}$, $b_i = \frac{(X_i - X_{i-1})}{(X_{i+1}-X_{i-1})}$, $i = 1, \ldots, n$. Simple algebra yields

$$\begin{aligned}
&(n-2)(\widehat{\sigma}^2 - \sigma^2) \\
&= \sum_{i=1}^{n}\frac{a_i^2(\xi_{i-1}^2 - \sigma^2) + b_i^2(\xi_{i+1}^2 - \sigma^2) + (\xi_i^2 - \sigma^2)}{a_i^2 + b_i^2 + 1} + \sum_{i=2}^{n-1}\frac{a_ib_i\xi_{i-1}\xi_{i+1} - a_i\xi_{i-1}\xi_i - b_i\xi_i\xi_{i+1}}{a_i^2 + b_i^2 + 1} \\
&= \frac{1}{n-2}\sum_{i=2}^{n-1}\left(\frac{a_{i+1}^2}{a_{i+1}^2 + b_{i+1}^2 + 1} + \frac{1}{a_i^2 + b_i^2 + 1} + \frac{b_{i-1}^2}{a_{i-1}^2 + b_{i-1}^2 + 1}\right)(\xi_i^2 - \sigma^2) \\
&\quad + \frac{2}{n-2}\sum_{i=2}^{n-1}\frac{a_ib_i}{a_i^2 + b_i^2 + 1}\xi_{i-1}\xi_{i+1} - \frac{2}{n-2}\sum_{i=2}^{n-1}\left(\frac{a_i}{a_i^2 + b_i^2 + 1} + \frac{b_{i-1}}{a_{i-1}^2 + b_{i-1}^2 + 1}\right)\xi_{i-1}\xi_i \\
&= \sum_{i=1}^{n}\alpha_{ii}(\xi_i^2 - \sigma^2) + \sum_{i\neq j}\alpha_{ij}\xi_i\xi_j
\end{aligned}$$

where $a_j = b_j = 0$ for $j = 0, 1, n, n+1$ and α_{ij} are some coefficients. This clearly implies $\boldsymbol{E}\widehat{\sigma}^2 = \sigma^2$. It is also easy to see that

$$(n-2)^2\boldsymbol{E}(\widehat{\sigma}^2 - \sigma^2)^2 = \sigma^4\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}^2 + \sigma^4(s_4 - 3)\sum_{i=1}^{n}\alpha_{ii}^2$$

where $s_4 = \sigma^{-4}\boldsymbol{E}(\xi_i^2 - \sigma^2)^2$.

One can easily check that the matrix A with the entries α_{ij} fulfills

$$\sigma^4 \operatorname{tr} A^\top A = \sigma^4\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{ij}^2 = \operatorname{tr}\Sigma^2$$

with the matrix Σ defined in Lemma 4.1 and hence, $\sigma^4 \operatorname{tr} A^\top A \le \frac{7}{2}\sigma^4(n-2)$. Since $\sum_{i=1}^{n} \alpha_{ii} = n-2$ and $\alpha_{ii} \le 2$ for all i, we derive

$$\frac{1}{n-2}\sum_{i=1}^{n} \alpha_{ii}^2 \le \frac{\max_{i=1,\dots,n} \alpha_{ii}}{n-2}\sum_{i=1}^{n} \alpha_{ii} \le 2$$

and

$$\boldsymbol{E}(\widehat{\sigma}^2 - \sigma^2)^2 \le \frac{7\sigma^4}{2(n-2)} + (s_4 - 3)\sigma^4 \frac{2}{n-2} \le \frac{(s_4 + 1/2)\sigma^4}{n-2}.$$

$\square$

In the same way one can extend the result of Lemma 4.2 to the non-Gaussian case: $\widehat{\sigma}^2$ estimates the true variance σ^2 at the rate $n^{-1/2}$ provided that f is sufficiently smooth.

Now we turn to Theorem 3.2. It obviously suffices to show that the distribution of the test statistic T^* can be approximated by a similar distribution corresponding to the case of Gaussian errors. Then the result follows from Theorem 3.1.

As in the proof of Theorem 3.1, the general case can be reduced to the no-response model with the vanishing regression function. Further, since the difference $\widehat{\sigma}^2 - \sigma^2$ is of order $n^{-1/2}$, it suffices to consider the expressions T_j', $j \le j(n)$, defined by

$$T_j' = \frac{1}{\sqrt{2\sigma^4 \operatorname{tr} V_j^2}} \left(\sum_{I \in \mathcal{I}_j} |\widehat{\theta}_I|^2 - \sigma^2 \operatorname{tr} V_j \right) = \frac{S_j - \sigma^2 \operatorname{tr} V_j}{\sqrt{2\sigma^4 \operatorname{tr} V_j^2}}$$

where $\widehat{\boldsymbol{\theta}}_I$ are elements of the vector $\widehat{\boldsymbol{\theta}}(j)$, cf. the proof of Lemma 4.4. Under the no-response hypothesis, this vector admits the representation: $\widehat{\boldsymbol{\theta}}(j) = W(j)\boldsymbol{\xi}$ with $W(j) = \left(\Psi(j)^\top \Psi(j)\right)^{-1} \Psi(j)^\top$, see (2.12). If E_j denotes the projector from $\mathcal{I}(j)$ onto $\mathcal{I}_j$ keeping the coordinates x_I with $I \in \mathcal{I}_j$, then $\widehat{\boldsymbol{\theta}}_j = E_j \widehat{\boldsymbol{\theta}}(j) = E_j W(j)\boldsymbol{\xi}$ and

$$S_j = \|\widehat{\boldsymbol{\theta}}_j\|^2 = \boldsymbol{\xi}^\top W(j)^\top E_j^\top E_j W(j)\boldsymbol{\xi} = \boldsymbol{\xi}^\top A_j \boldsymbol{\xi}$$

with $A_j = W(j)^\top E_j^\top E_j W(j)$, so that S_j is a quadratic form of the errors ξ_i. We also know that $V_j = E_j W(j) W(j)^\top E_j^\top$, and $\boldsymbol{E} S_j = \sigma^2 \operatorname{tr} A_j = \sigma^2 \operatorname{tr} V_j$. This form in its turn can be represented as a sum of a diagonal form $T_j^{(1)}$ and a quadratic form $T_j^{(2)}$ with vanishing diagonal terms. We first show that the impact of diagonal terms is negligible and then apply Corollary 5.2 to $T_j^{(2)}$'s.

Let o_i denote the i-th basis vector in $\mathbb{R}^n$. Then the i-th diagonal element a_{ii} of A_j is equal to $o_i^\top A_j o_i$:

$$\begin{aligned} a_{ii} &= o_i^\top A_j o_i \\ &= o_i^\top \Psi(j)^\top \left(\Psi(j)^\top \Psi(j)\right)^{-1} E_j^\top E_j \left(\Psi(j)^\top \Psi(j)\right)^{-1} E_j^\top E_j \left(\Psi(j)^\top \Psi(j)\right)^{-1} \Psi(j) o_i. \end{aligned}$$

Clearly

$$\left\| \left(\Psi(j)^\top \Psi(j)\right)^{-1} E_j^\top E_j \left(\Psi(j)^\top \Psi(j)\right)^{-1} \right\| \le \left\| \left(\Psi(j)^\top \Psi(j)\right)^{-2} \right\| = \|V(j)^2\| \le C_V^2.$$

Next, for every Haar level $\ell \le j$, there exists only one index $I \in \mathcal{I}_\ell$ such that $\psi_I(X_i) \ne 0$. More precisely, for this index I, it holds $\psi_I(X_i) = \pm 1/\sqrt{M_I}$ where M_I is the number of design points in the interval A_I

corresponding to the index I. Condition $(D.i)$ implies $M_I \ge C_* n 2^{-\ell}$ for every $I \in \mathcal{I}_\ell$. Also $\psi_0(X_i) = n^{-1/2}$ and $\psi_1(X_i) = X_i \left(\sum_{i'=1}^n X_{i'}^2\right)^{-1/2}$. Hence, the definition of the matrix $\Psi(j)$ and condition $(D.i)$ provide

$$|\Psi(j)o_i| \le n^{-1/2} + \left(\sum_{i'=1}^n X_{i'}^2\right)^{-1/2} + \sum_{\ell=0}^j \sqrt{\frac{2^\ell}{nC_*}} < 3C_*^{-1/2} 2^{j/2} n^{-1/2}. \tag{4.8}$$

Therefore,

$$a_{ii} \le |\Psi(j)o_i|^2 \left\| \left(\Psi(j)^\top \Psi(j)\right)^{-1} E_j^\top E_j \left(\Psi(j)^\top \Psi(j)\right)^{-1} \right\| \le 9C_*^{-1} 2^j n^{-1} C_V^2.$$

Define $G_j^2 = 2\sigma^4 \operatorname{tr} A_j^2$. Note

$$\begin{aligned} \operatorname{tr} A_j^2 &= \operatorname{tr} W(j)^\top E_j^\top E_j W(j) W(j)^\top E_j^\top E_j W(j) \\ &= \operatorname{tr} E_j W(j) W(j)^\top E_j^\top E_j W(j) W(j)^\top E_j^\top = \operatorname{tr} V_j^2 \end{aligned}$$

so that $T_j^{(1)} = G_j^{-1} \sum_{i=1}^n a_{ii}(\xi_i^2 - \sigma^2)$. In view of condition $(D.ii)$ it holds $\operatorname{tr} A_j^2 \ge C_D 2^j$. Now, for every $\delta > 0$,

$$\begin{aligned} \boldsymbol{P}\left(\max_{j=0,\dots,j(n)} T_j^{(1)} > \delta\right) &\le \sum_{j=0}^{j(n)} \boldsymbol{P}\left(T_j^{(1)} > \delta\right) \\ &\le \delta^{-2} \sum_{j=0}^{j(n)} \boldsymbol{E}\left|T_j^{(1)}\right|^2 \\ &\le \delta^{-2} \sum_{j=0}^{j(n)} G_j^{-2} \sigma^4 s_4 \sum_{i=1}^n a_{ii}^2 \\ &\le \delta^{-2} \sum_{j=0}^{j(n)} 2^{-1} C_D^{-1} 2^{-j} s_4 n \left(9C_*^{-1} 2^j n^{-1} C_V^2\right)^2 \\ &\le C\delta^{-2} n^{-1} 2^{j(n)+1} \to 0, \qquad n \to \infty. \end{aligned}$$

Next we consider $T_j^{(2)}$ which is obtained from T_j' by removing the diagonal terms. This quadratic form can be approximated (in distribution) by a similar one with Gaussian errors $\widetilde{\xi}_i$ at a reasonable rate provided that the corresponding value C_A, defined as n times the ratio of the maximal diagonal element of the matrix $\sigma^4 A_j^2$ to $G_j^2 = \sigma^4 \operatorname{tr} A_j^2$, see (5.2) and Remark 5.1 in the Appendix, remains bounded.

The i-th diagonal element d_i of A_j^2 is equal to $o_i^\top A_j^2 o_i$:

$$\begin{aligned} d_i &= o_i^\top A_j^2 o_i \\ &= o_i^\top \left\{ \Psi(j) \left(\Psi(j)^\top \Psi(j)\right)^{-1} E_j^\top E_j \left(\Psi(j)^\top \Psi(j)\right)^{-1} \Psi(j)^\top \right\}^2 o_i \\ &= o_i^\top \Psi(j)^\top \left(\Psi(j)^\top \Psi(j)\right)^{-1} E_j^\top E_j \left(\Psi(j)^\top \Psi(j)\right)^{-1} E_j^\top E_j \left(\Psi(j)^\top \Psi(j)\right)^{-1} \Psi(j) o_i. \end{aligned}$$

Clearly

$$\begin{aligned}&\left\|\left(\Psi(j)^\top\Psi(j)\right)^{-1}E_j^\top E_j\left(\Psi(j)^\top\Psi(j)\right)^{-1}E_j^\top E_j\left(\Psi(j)^\top\Psi(j)\right)^{-1}\right\|\\&\quad\le\left\|\left(\Psi(j)^\top\Psi(j)\right)^{-3}\right\|=\|V(j)^3\|\le C_V^3.\end{aligned}$$

The use of (4.8) provides

$$\begin{aligned}d_i&\le|\Psi(j)o_i|^2\left\|\left(\Psi(j)^\top\Psi(j)\right)^{-1}E_j^\top E_j\left(\Psi(j)^\top\Psi(j)\right)^{-1}E_j^\top E_j\left(\Psi(j)^\top\Psi(j)\right)^{-1}\right\|\\&\le 9C_*^{-1}2^jn^{-1}C_V^3\end{aligned}$$

and

$$C_A\le\frac{9C_*^{-1}C_V^3 2^j}{C_D2^j}=\frac{9C_*^{-1}C_V^3}{C_D}$$

that is, the value C_A is bounded by a fixed constant depending on design regularity only.

By Corollary 5.2, the joint distribution of $T_j^{(2)}$, $j\le j(n)$, and the distribution of their maximum, can be approximated by the distribution of similar quadratic forms of Gaussian r.v.'s which implies the required assertion.

5. Appendix

Here we briefly discuss some general properties of quadratic forms of random variables. We first consider the case when the underlying random variables are Gaussian and establish an exponential bound for deviations of such forms over certain level. Next we show how an arbitrary quadratic form of independent random variables can be approximated (in distribution) by a similar quadratic form of Gaussian random variables.

5.1. Deviation probabilities for quadratic forms of Gaussian random variables

Let $\varepsilon_1,\ldots,\varepsilon_N$ be Gaussian random variables with zero mean and the covariance $N\times N$ matrix V, i.e. $V=\boldsymbol{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top$ where $\boldsymbol{\varepsilon}$ denotes the vector $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_N)^\top$.

We first present the following general results about quadratic forms of Gaussian random variables.

Proposition 5.1. *Let $\varepsilon_1,\ldots,\varepsilon_N$ be Gaussian random variables with zero mean and the covariance matrix $V:=\boldsymbol{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top$. Then*

$$\begin{aligned}\boldsymbol{E}\|\boldsymbol{\varepsilon}\|^2&:=\boldsymbol{E}\left(\varepsilon_1^2+\ldots+\boldsymbol{E}\varepsilon_N^2\right)=\operatorname{tr}V,\\\boldsymbol{E}\left(\|\boldsymbol{\varepsilon}\|^2-\operatorname{tr}V\right)^2&=2\operatorname{tr}V^2.\end{aligned}$$

Moreover, for $\gamma\le\|V\|^{-1}\sqrt{\operatorname{tr}V^2/2}$ and each $t\ge0$,

$$\boldsymbol{P}\left(\pm(\|\boldsymbol{\varepsilon}\|^2-\operatorname{tr}V)>(\gamma+t)\sqrt{2\operatorname{tr}V^2}\right)\le e^{-\gamma t/2-\gamma^2/4}.$$

Proof. Let $V=U^\top\Lambda U$ be a diagonal representation of V with a diagonal matrix $\Lambda=\operatorname{diag}\{\lambda_1,\ldots,\lambda_N\}$ and an ortonormal matrix U. It is well known that $\zeta=\Lambda^{-1/2}U\boldsymbol{\varepsilon}$ is a standard Gaussian vector and $\|\boldsymbol{\varepsilon}\|^2=\zeta^\top\Lambda\zeta$. Also it holds $\operatorname{tr}V=\lambda_1+\ldots+\lambda_N$, $\operatorname{tr}V^2=\lambda_1^2+\ldots+\lambda_N^2$ and $\|V\|=\max\{\lambda_1,\ldots,\lambda_N\}$. To bound the expression $\|\varepsilon\|^2 - \operatorname{tr} V$, we apply the exponential Chebyshev inequality: with each $\mu \ge 0$ satisfying $2\mu\lambda_i < 1$ and every z

$$\begin{aligned}
\boldsymbol{P}\left(\|\varepsilon\|^2 - \operatorname{tr} V > z\right) &\le e^{-\mu z}\boldsymbol{E}\exp\left\{\mu(\|\varepsilon\|^2 - \operatorname{tr} V)\right\} = e^{-\mu z}\boldsymbol{E}\exp\left\{\mu\sum_{i=1}^{N}\lambda_i(\zeta_i^2 - 1)\right\} \\
&= e^{-\mu z}\prod_{i=1}^{N}\boldsymbol{E}\exp\left\{\mu\lambda_i(\zeta_i^2 - 1)\right\} = \exp\left\{-\mu z - \mu\sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{N}\frac{1}{2}\log(1 - 2\mu\lambda_i)\right\}.
\end{aligned}$$

We now set $\mu = \frac{\gamma}{2\sqrt{2\operatorname{tr} V^2}}$ so that $2\mu\lambda_i = \frac{\gamma\lambda_i}{\sqrt{2\operatorname{tr} V^2}} < 1/2$ and use that $-\log(1-u) \le u + u^2$ for $0 \le u \le 1/2$. This yields

$$\begin{aligned}
\boldsymbol{P}\left(\|\varepsilon\|^2 - \operatorname{tr} V > (\gamma + t)\sqrt{2\operatorname{tr} V^2}\right) &\le \exp\left(-\frac{\gamma(\gamma+t)}{2} + \frac{\gamma^2}{4\operatorname{tr} V^2}\sum_{i=1}^{N}\lambda_i^2\right) \\
&= \exp\left(-\gamma t/2 - \gamma^2/4\right)
\end{aligned}$$

as required. The bound for $-(\|\varepsilon\|^2 - \operatorname{tr} V)$ is proved in the same line. □

Further, for a deterministic vector $\boldsymbol{c} = (c_1, \ldots, c_N)^\top$ from $\mathbb{R}^N$, we consider quadratic forms of type

$$\|\boldsymbol{c} + \varepsilon\|^2 = \sum_{j=1}^{N}|c_j + \varepsilon_j|^2.$$

Proposition 5.2. *Let $\varepsilon_1, \ldots, \varepsilon_N$ be Gaussian random variables with zero mean and the covariance matrix V. Then it holds for any vector $\boldsymbol{c} = (c_1, \ldots, c_N)^\top$ in $\mathbb{R}^N$*

$$\begin{aligned}
&\boldsymbol{E}\|\boldsymbol{c} + \varepsilon\|^2 = \|\boldsymbol{c}\|^2 + \operatorname{tr} V, \\
&\operatorname{Var}\|\boldsymbol{c} + \varepsilon\|^2 := \boldsymbol{E}\left(\|\boldsymbol{c} + \varepsilon\|^2 - \|\boldsymbol{c}\|^2 - \operatorname{tr} V\right)^2 = 4\boldsymbol{c}^\top V\boldsymbol{c} + 2\operatorname{tr} V^2, \\
&\boldsymbol{E}\left(\|\boldsymbol{c} + \varepsilon\|^2 - \operatorname{tr} V\right)^2 = \|\boldsymbol{c}\|^4 + 4\boldsymbol{c}^\top V\boldsymbol{c} + 2\operatorname{tr} V^2 \le \|\boldsymbol{c}\|^4 + 4\|\boldsymbol{c}\|^2\sqrt{\operatorname{tr} V^2}\boldsymbol{c} + 2\operatorname{tr} V^2.
\end{aligned}$$

Moreover, for every positive γ with $\gamma \le \|V\|^{-1}\sqrt{\operatorname{tr} V^2/2}$ and every $t \ge 0$

$$\boldsymbol{P}\left(\pm(\|\boldsymbol{c} + \varepsilon\|^2 - \|\boldsymbol{c}\|^2 - \operatorname{tr} V) > \gamma\|\boldsymbol{c}\|(2\operatorname{tr} V^2)^{1/4} + (\gamma + t)\sqrt{2\operatorname{tr} V^2}\right) \le 2e^{-\gamma^2/4 - \gamma t/2}.$$

Proof. With vector notation, the studied quadratic form can be rewritten as $\|\boldsymbol{c} + \varepsilon\|^2 = (\boldsymbol{c} + \varepsilon)^\top(\boldsymbol{c} + \varepsilon)$. Now, since $\boldsymbol{E}\varepsilon_i = 0$, it holds

$$\boldsymbol{E}\|\boldsymbol{c} + \varepsilon\|^2 = \boldsymbol{E}\left(\|\boldsymbol{c}\|^2 + 2\boldsymbol{c}^\top\varepsilon + \|\varepsilon\|^2\right) = \|\boldsymbol{c}\|^2 + \boldsymbol{E}\|\varepsilon\|^2 = \|\boldsymbol{c}\|^2 + \operatorname{tr} V.$$

Next,

$$\begin{aligned}
\operatorname{Var}\|\boldsymbol{c} + \varepsilon\|^2 &= \boldsymbol{E}\left(\|\boldsymbol{c} + \varepsilon\|^2 - \boldsymbol{E}\|\boldsymbol{c} + \varepsilon\|^2\right)^2 \\
&= \boldsymbol{E}\left(2\boldsymbol{c}^\top\varepsilon + \|\varepsilon\|^2 - \operatorname{tr} V\right)^2 \\
&= 4\boldsymbol{E}|\boldsymbol{c}^\top\varepsilon|^2 + 4\boldsymbol{E}\boldsymbol{c}^\top\varepsilon\left(\|\varepsilon\|^2 - \operatorname{tr} V\right) + \boldsymbol{E}\left(\|\varepsilon\|^2 - \operatorname{tr} V\right)^2.
\end{aligned}$$

The Gaussian vector $\varepsilon \sim \mathcal{N}(0, V)$ fulfills

$$\boldsymbol{E}\,\varepsilon\left(\|\varepsilon\|^2 - \operatorname{tr} V\right) = 0,$$
$$\boldsymbol{E}|\boldsymbol{c}^\top \varepsilon|^2 = \boldsymbol{c}^\top(\boldsymbol{E}\varepsilon\varepsilon^\top)\boldsymbol{c} = \boldsymbol{c}^\top V \boldsymbol{c}$$

so that in view of Proposition 5.1 $\operatorname{Var}\|\boldsymbol{c}+\varepsilon\|^2 = 4\boldsymbol{c}^\top V\boldsymbol{c} + 2\operatorname{tr} V^2$ as required. Similarly one obtains

$$\boldsymbol{E}\left(\|\boldsymbol{c}+\varepsilon\|^2 - \operatorname{tr} V\right)^2 = \|\boldsymbol{c}\|^4 + 4\boldsymbol{c}^\top V\boldsymbol{c} + 2\operatorname{tr} V^2$$

and by the Cauchy-Schwarz inequality $\boldsymbol{c}^\top V\boldsymbol{c} \le \|\boldsymbol{c}\|^2\sqrt{\operatorname{tr} V^2}$.

Let now $\gamma \ge 1$ be fixed such that $\gamma \le \|V\|^{-1}\sqrt{\operatorname{tr} V^2/2}$. This particularly means that $\|V\| \le \sqrt{\operatorname{tr} V^2/2}$. Note that the scalar product $\boldsymbol{c}^\top\varepsilon$ is a linear combination of the Gaussian zero mean random variables and it is therefore Gaussian as well with $\boldsymbol{E}\boldsymbol{c}^\top\varepsilon = 0$ and $\boldsymbol{E}|\boldsymbol{c}^\top\varepsilon|^2 = \boldsymbol{c}^\top V\boldsymbol{c}$. This yields for every $\gamma \ge 1$

$$\boldsymbol{P}\left(\boldsymbol{c}^\top\varepsilon > \gamma\sqrt{\boldsymbol{c}^\top V\boldsymbol{c}}\right) \le e^{-\gamma^2/2}.$$

The condition $\|V\| \le \sqrt{\operatorname{tr} V^2/2}$ provides $\boldsymbol{c}^\top V\boldsymbol{c} \le \|\boldsymbol{c}\|^2\|V\| \le \|\boldsymbol{c}\|^2\sqrt{\operatorname{tr} V^2/2}$. Combining this inequality with the previous one implies

$$\boldsymbol{P}\left(2\boldsymbol{c}^\top\varepsilon > (\gamma+t)\|\boldsymbol{c}\|(2\operatorname{tr} V^2)^{1/4}\right) \le e^{-(\gamma+t)^2/4}.$$

Next, by Proposition 5.1

$$\boldsymbol{P}\left(\|\varepsilon\|^2 - \operatorname{tr} V > (\gamma+t)\sqrt{2\operatorname{tr} V^2}\right) \le e^{-\gamma^2/4-\gamma t/2}.$$

Summing up the previous estimates, we obtain

$$\begin{aligned}
&\boldsymbol{P}\left(\sum_{j=1}^N |c_j+\varepsilon_j|^2 - \operatorname{tr} V > \|\boldsymbol{c}\|^2 + (\gamma+t)\|\boldsymbol{c}\|(2\operatorname{tr} V^2)^{1/4} + (\gamma+t)\sqrt{2\operatorname{tr} V^2}\right) \\
&\quad = \boldsymbol{P}\left(2\boldsymbol{c}^\top\varepsilon + \|\varepsilon\|^2 - \operatorname{tr} V > (\gamma+t)\|\boldsymbol{c}\|(2\operatorname{tr} V^2)^{1/4} + (\gamma+t)\sqrt{2\operatorname{tr} V^2}\right) \\
&\quad \le \boldsymbol{P}\left(2\boldsymbol{c}^\top\varepsilon > (\gamma+t)\|\boldsymbol{c}\|(2\operatorname{tr} V^2)^{1/4}\right) + \boldsymbol{P}\left(\|\varepsilon\|^2 - \operatorname{tr} V > (\gamma+t)\sqrt{2\operatorname{tr} V^2}\right) \\
&\quad \le 2e^{-\gamma^2/4-\gamma t/2}
\end{aligned}$$

as required. □

5.2. Gaussian approximation for quadratic forms

In what follows we consider quadratic forms $\sum_{i=1}^n\sum_{\ell=1}^n a_{i\ell}\xi_i\xi_\ell$ of independent but not necessarily normal random variables $\xi_1,\ldots\xi_n$ with vanishing diagonal coefficients, i.e. $a_{ii} = 0$. We aim to show that, under moment conditions on ξ_i's and mild assumptions on the coefficients of the quadratic form, the asymptotic distribution of this quadratic form only weakly depends on the particular distribution of ξ_i's and, as a consequence, it can be approximated by a distribution of a similar quadratic form of Gaussian r.v.'s with the same first and second moments.

Let $A = (a_{i\ell}\,, i,j = 1,\ldots,n)$ be a $n\times n$ symmetric matrix with $a_{ii} = 0$ for all i, and let $\xi_1,\ldots,\xi_n$ be independent zero mean r.v.'s with $\boldsymbol{E}\xi_i^4 < \infty$ for all i. Define $\sigma_i^2 = \boldsymbol{E}\xi_i^2$. We study some properties of the quadratic form $\sum_{i=1}^n\sum_{j=1}^n a_{i\ell}\xi_i\xi_\ell$.

Lemma 5.1. *It holds*

$$\boldsymbol{E}\sum_{i=1}^{n}\sum_{\ell=1}^{n} a_{i\ell}\xi_i\xi_\ell = \sum_{i=1}^{n} a_{ii}\sigma_i^2 = 0,$$

$$\boldsymbol{E}\left\{\sum_{i=1}^{n}\sum_{\ell=1}^{n} a_{i\ell}\xi_i\xi_\ell\right\}^2 = 2\sum_{i=1}^{n}\sum_{\ell\neq i} a_{i\ell}^2\sigma_i^2\sigma_\ell^2. \tag{5.1}$$

Proof. Obvious. Here it is only important that the diagonal elements a_{ii} vanish. □

By $A(\xi_1,\dots,\xi_n)$ we denote the corresponding quadratic form, that is

$$A(\xi_1,\dots,\xi_n) = \sum_{i=1}^{n}\sum_{\ell\neq i}^{n} a_{i\ell}\xi_i\xi_\ell.$$

Let also $\widetilde{\xi}_1,\dots,\widetilde{\xi}_n$ be a sequence of independent Gaussian r.v.'s with $\boldsymbol{E}\widetilde{\xi}_i = 0$ and $\boldsymbol{E}\widetilde{\xi}_i^2 = \sigma_i^2$, $i=1,\dots,n$. Define another quadratic form

$$A(\widetilde{\xi}_1,\dots,\widetilde{\xi}_n) = \sum_{i=1}^{n}\sum_{\ell\neq i} a_{i\ell}\widetilde{\xi}_i\widetilde{\xi}_\ell$$

Clearly $\boldsymbol{E}A(\widetilde{\xi}_1,\dots,\widetilde{\xi}_n) = 0$ and $\boldsymbol{E}|A(\widetilde{\xi}_1,\dots,\widetilde{\xi}_n)|^2 = \boldsymbol{E}|A(\xi_1,\dots,\xi_n)|^2$.

Proposition 5.3. *Let $\boldsymbol{E}\xi_i^4 \le C_4\sigma_i^4$ for some fixed constant $C_4 \ge 3$. Let, for a symmetric matrix A with $a_{ii}=0$ for $i=1,\dots,n$, and for a normalizing constant G, the numbers C_A be defined by*

$$C_A = \max_{i=1,\dots,n} nG^{-2}\sum_{\ell=1}^{n} a_{i\ell}^2\sigma_i^2\sigma_\ell^2. \tag{5.2}$$

Then, for every three times continuously differentiable function f, it holds

$$\left|\boldsymbol{E}f\left(G^{-1}A(\xi_1,\dots,\xi_n)\right) - \boldsymbol{E}f\left(G^{-1}A(\widetilde{\xi}_1,\dots,\widetilde{\xi}_n)\right)\right| \le \frac{8}{3}f_3(C_4C_A)^{3/2}n^{-1/2}$$

where f_3 means the maximum of the absolute value of the third derivative of f, that is, $f_3 = \sup_x|f'''(x)|$.

Remark 5.1. The value C_A can be easily evaluated for the case of an homogeneous noise when all σ_i^2 coincide with some σ^2. Clearly each sum $d_i = \sum_{\ell=1}^{n} a_{i\ell}^2$ is i-th diagonal element of A^2 and $C_A \le G^{-2}\max_{i=1,\dots,n}\{nd_i\}$.

Remark 5.2. The conditions of Proposition 5.3 do not guarantee that the distribution of $G^{-1}A(\xi_1,\dots,\xi_n)$ is close to some normal distribution. A typical example which just meets in hypothesis testing framework corresponds to the quadratic form $A(\xi_1,\dots,\xi_n) = (\xi_1+\dots+\xi_n)^2$, which, even with normal ξ_i's, is χ_1^2-distributed.

Proof. The change ξ_i for ξ_i/σ_i and $a_{i\ell}$ for $a_{i\ell}\sigma_i\sigma_\ell$ allows to reduce the general case to the situation with $\sigma_i = 1$ for all i. Hence, for the sake of notation simplicity, we suppose that $\xi_i^2 = 1$, $i=1,\dots,n$.

We use the following obvious inequality

$$\begin{aligned}&\left|\boldsymbol{E}f\Big(G^{-1}A(\xi_1,\dots,\xi_n)\Big) - \boldsymbol{E}f\left(G^{-1}A(\widetilde{\xi}_1,\dots,\widetilde{\xi}_n)\right)\right|\\ &\le \sum_{i=1}^{n}\left|\boldsymbol{E}f\left(G^{-1}A(\xi_1,\dots,\xi_i,\widetilde{\xi}_{i+1},\dots,\widetilde{\xi}_n)\right) - \boldsymbol{E}f\left(G^{-1}A(\xi_1,\dots,\xi_{i-1},\widetilde{\xi}_i,\dots,\widetilde{\xi}_n)\right)\right|\end{aligned}$$

where we assume $\xi_0 = \widetilde{\xi}_{n+1} = 0$. We evaluate the last summand here, the other can be bounded in the same way. Denote

$$\begin{aligned} u_{n-1} &= G^{-1} \sum_{i=1}^{n-1} \sum_{\ell \neq i}^{n-1} a_{i\ell} \xi_i \xi_\ell, \\ \Delta_n &= G^{-1} A(\xi_1, \ldots, \xi_n) - u_{n-1} = 2G^{-1} \xi_n \sum_{i=1}^{n-1} a_{in} \xi_i \,, \\ \widetilde{\Delta}_n &= G^{-1} A(\xi_1, \ldots, \xi_{n-1}, \widetilde{\xi}_n) - u_{n-1} = 2G^{-1} \widetilde{\xi}_n \sum_{i=1}^{n-1} a_{in} \xi_i \,. \end{aligned}$$

The Taylor expansion yields

$$\begin{aligned} &\left| \boldsymbol{E} f \left(G^{-1} A(\xi_1, \ldots, \xi_n)\right) - \boldsymbol{E} f \left(G^{-1} A(\xi_1, \ldots, \xi_{n-1}, \widetilde{\xi}_n)\right) \right| \\ &\le \left| \boldsymbol{E} f'(u_{n-1})(\Delta_n - \widetilde{\Delta}_n) \right| + \frac{1}{2} \left| \boldsymbol{E} f''(u_{n-1})(\Delta_n^2 - \widetilde{\Delta}_n^2) \right| + \frac{f_3}{6} (\boldsymbol{E} \, |\Delta_n|^3 + \boldsymbol{E} |\widetilde{\Delta}_n|^3). \end{aligned} \tag{5.3}$$

Since ξ_n and $\widetilde{\xi}_n$ are independent of $\xi_1, \ldots, \xi_{n-1}$ and since $\boldsymbol{E}\xi_n = \boldsymbol{E}\widetilde{\xi}_n = 0$, $\boldsymbol{E}\xi_n^2 = \boldsymbol{E}\widetilde{\xi}_n^2 = 1$, taking the conditional expectation given $\xi_1, \ldots, \xi_{n-1}$, we obtain

$$\boldsymbol{E} \left(\Delta_n - \widetilde{\Delta}_n \mid \xi_1, \ldots, \xi_{n-1} \right) = 0, \qquad \boldsymbol{E} \left(\Delta_n^2 - \widetilde{\Delta}_n^2 \mid \xi_1, \ldots, \xi_{n-1} \right) = 0. \tag{5.4}$$

Further we evaluate $\boldsymbol{E}|\Delta_n|^3$ and $\boldsymbol{E}|\widetilde{\Delta}_n|^3$. Note first that, since $\boldsymbol{E}\xi_n^4 \le C_4$ with $C_4 \ge 3$,

$$\begin{aligned} \boldsymbol{E} \left(\sum_{i=1}^{n-1} a_{in} \xi_i \right)^4 &= \sum_{i=1}^{n-1} a_{in}^4 \boldsymbol{E} \xi_i^4 + 3 \sum_{\ell \neq i}^{n-1} a_{in}^2 a_{\ell n}^2 \\ &\le \sum_{i=1}^{n-1} a_{in}^4 (C_4 - 3) + 3 \left(\sum_{i=1}^{n-1} a_{in}^2 \right)^2 \le C_4 \left(\sum_{i=1}^{n-1} a_{in}^2 \right)^2 . \end{aligned}$$

Now the Hölder inequality yields in view of $\boldsymbol{E}|\xi_n|^3 \le C_4^{3/4}$

$$\begin{aligned} G^3 \boldsymbol{E} |\Delta_n|^3 &= \boldsymbol{E} |\xi_n|^3 \boldsymbol{E} \left| 2 \sum_{i=1}^{n-1} a_{in} \xi_i \right|^3 \\ &\le 8 C_4^{3/4} \left\{ \boldsymbol{E} \left(\sum_{i=1}^{n-1} a_{in} \xi_i \right)^4 \right\}^{3/4} \le 8 C_4^{3/2} \left(\sum_{i=1}^{n} a_{in}^2 \right)^{3/2} \end{aligned}$$

and the condition $G^{-2} \sum_{i=1}^n a_{in}^2 \le n^{-1} C_A$ provides

$$\boldsymbol{E} |\Delta_n|^3 \le 8 (C_4 C_A)^{3/2} n^{-3/2}. \tag{5.5}$$

For the Gaussian r.v. $s_n \widetilde{\xi}_n$, the similar bound applies:

$$\boldsymbol{E} |\widetilde{\Delta}_n|^3 \le 8 (C_4 C_A)^{3/2} n^{-3/2}. \tag{5.6}$$

Substituting these estimates as well as (5.4) in (5.3) implies

$$\left|\boldsymbol{E} f\left(\frac{A(\xi_1,\ldots,\xi_n)}{G}\right) - \boldsymbol{E} f\left(\frac{A(\xi_1,\ldots,\xi_{n-1},\widetilde{\xi}_n)}{G}\right)\right| \le \frac{16}{6} f_3 (C_4 C_A)^{3/2} n^{-3/2}.$$

Similar bounds hold for the other summands in (5.3). Summing them out, we obtain

$$\left|\boldsymbol{E} f\left(G^{-1}A(\xi_1,\ldots,\xi_n)\right) - \boldsymbol{E} f\left(G^{-1}A(\widetilde{\xi}_1,\ldots,\widetilde{\xi}_n)\right)\right| \le \frac{8}{3} f_3 (C_4 C_A)^{3/2} n^{-1/2}$$

as required. □

Corollary 5.1. *Under the conditions of Proposition 5.3, for each* $\delta > 0$ *and every* x

$$\boldsymbol{P}\left(G^{-1}A(\xi_1,\ldots,\xi_n) > x\right) \le \boldsymbol{P}\left(G^{-1}A(\widetilde{\xi}_1,\ldots,\widetilde{\xi}_n) > x - \delta\right) + Const. C_A^{3/2} n^{-1/2}\delta^{-3}$$

with a constant Const. depending on C_4 *only. If, in addition,* $G^2 \ge \boldsymbol{E}|A(\xi_1,\ldots,\xi_n)|^2$ *, then*

$$\boldsymbol{P}\left(G^{-1}A(\xi_1,\ldots,\xi_n) > x\right) \le \boldsymbol{P}\left(G^{-1}A(\widetilde{\xi}_1,\ldots,\widetilde{\xi}_n) > x\right) + Const. C_A^{3/2} n^{-1/2}\delta^{-3} + \delta.$$

Proof. Let a smooth function f fulfill $f(u) = 0$ for $u \le -1$ and $f(u) = 1$ for $u \ge 0$. Define $C_f = \sup_u |f'''(u)|$. Now, given x and $\delta > 0$, set $f_{x,\delta}(u) = f(\delta^{-1}(u - x))$. Obviously $f_{x,\delta}(u) = 0$ for $u \le x - \delta$ and $f_{x,\delta}(u) = 1$ for $u \ge x$ and also $|f'''_{x,\delta}(u)| \le C_f \delta^{-3}$.

Next, by Proposition 5.3

$$\begin{aligned}\boldsymbol{P}\left(G^{-1}A(\xi_1,\ldots,\xi_n) > x\right) &\le \boldsymbol{E} f_{x,\delta}\left(G^{-1}A(\xi_1,\ldots,\xi_n)\right)\\ &\le \boldsymbol{E} f_{x,\delta}\left(G^{-1}A(\widetilde{\xi}_1,\ldots,\widetilde{\xi}_n)\right) + \frac{8}{3}(C_A C_4)^{3/2} C_f \delta^{-3} n^{-1/2}.\end{aligned}$$

It remains to note that

$$\boldsymbol{E} f_{x,\delta}\left(G^{-1}A(\widetilde{\xi}_1,\ldots,\widetilde{\xi}_n)\right) \le \boldsymbol{P}\left(G^{-1}A(\widetilde{\xi}_1,\ldots,\widetilde{\xi}_n) > x - \delta\right)$$

The last statement of the corollary follows from the obvious fact that the density of $G^{-1}A(\widetilde{\xi}_1,\ldots,\widetilde{\xi}_n)$ is bounded by 1 for every G with $G^2 \ge \boldsymbol{E}|A(\widetilde{\xi}_1,\ldots,\widetilde{\xi}_n)|^2$. □

5.3. A family of quadratic forms

Here we briefly discuss the situation arising in adaptive testing problem when the maximum of a family of quadratic forms of ξ_i 's is considered. We again aim to show that the joint distribution of this family (and thus the distribution of the maximum) can be well approximated by the similar distribution for quadratic forms of Gaussian random variables.

Let $A_1,\ldots,A_M$ be a collection of symmetric $n\times n$-matrices with vanishing diagonal elements. We analyze the joint distribution of the standardized quadratic forms $G_m^{-1}A_m(\xi_1,\ldots,\xi_n)$ with independent random variables ξ_i satisfying $\boldsymbol{E}\xi_i = 0$, $\boldsymbol{E}\xi_i^2 = \sigma_i^2$ and $\boldsymbol{E}\xi_i^4 < \infty$, and some constants G_m, $m = 1,\ldots,M$. More precisely, we intend to show that the distribution of this family is close to the distribution of the family $\{G_m^{-1}A_m(\widetilde{\xi}_1,\ldots,\widetilde{\xi}_n), m = 1,\ldots,M\}$ with Gaussian variables $\widetilde{\xi}_i \sim \mathcal{N}(0,\sigma_i^2)$.

Proposition 5.4. *Let the variables ξ_i fulfill $\boldsymbol{E}\xi_i^4 \le C_E\sigma_i^4$ and let every matrix A_m satisfy the conditions of Proposition 5.3 with the same constant C_A, $m = 1,\dots,M$. Then, for every three times continuously differentiable function f in the space $\mathbb{R}^M$, it holds*

$$\left|\boldsymbol{E}f\left(G^{-1}A(\xi_1,\dots,\xi_n)\right) - \boldsymbol{E}f\left(G^{-1}A(\widetilde{\xi}_1,\dots,\widetilde{\xi}_n)\right)\right| \le \frac{8}{3} f_3 M^3 (C_4 C_A)^{3/2} n^{-1/2}$$

where $G^{-1}A$ denotes the vector with elements $G_m^{-1}A_m$ and f_3 means the maximum of the absolute value of the third derivative of f, that is,

$$f_3 = \sup_{x\in\mathbb{R}^M} \max_{i,j,k=1,\dots,M} \left|\frac{\partial^3 f(x)}{\partial x_i \partial x_j \partial x_k}\right|.$$

Proof. The proof follows the same line as in the case of one quadratic forms when understanding $G^{-1}A$, u_{n-1}, $f'(u_{n-1})$ and Δ_n as vectors in $\mathbb{R}^M$ and $f''(u_{n-1})$ as the $M\times M$-matrix of the second derivatives of f at u_{n-1}. The only difference is that we apply the bound $\boldsymbol{E}|\Delta_n|^3 \le M^3 8(C_4C_A)^{3/2}n^{-3/2}$ for the norm of Δ_n which is M^3 times larger than in the case of $M=1$, cf. (5.5). The details are left to the reader. □

A straightforward corollary of this results concerns the maximum of $G_m^{-1}A_m$'s.

Corollary 5.2. *Let the conditions of Proposition 5.4 be fulfilled. Then*

$$\boldsymbol{P}\left(\max_{m\le M} G_m^{-1}A_m(\xi_1,\dots,\xi_n)\le x\right) - \boldsymbol{P}\left(\max_{m\le M} G_m^{-1}A_m(\widetilde{\xi}_1,\dots,\widetilde{\xi}_n)\le x-\delta\right)$$
$$\le Const.M^3 C_A^{3/2} n^{-1/2}\delta^{-3}$$

with a constant Const. depending on C_4 only. If, in addition, $G_m^2 \ge \boldsymbol{E}|A_m(\xi_1,\dots,\xi_n)|^2$ for all $m\le M$, then

$$\boldsymbol{P}\left(\max_{m\le M} G_m^{-1}A_m(\xi_1,\dots,\xi_n)\le x\right) - \boldsymbol{P}\left(\max_{m\le M} G_m^{-1}A_m(\widetilde{\xi}_1,\dots,\widetilde{\xi}_n)\le x\right)$$
$$\le Const.M^3 C_A^{3/2} n^{-1/2}\delta^{-3} + M\delta.$$

Proof. The first statement can be checked exactly as for the case of $M=1$, see the proof of Corollary 5.1. As regard to the second statement, it suffices to mention that the density of each $G_m^{-1}A_m(\widetilde{\xi}_1,\dots,\widetilde{\xi}_n)$ is bounded by 1 and hence the density of the maximum of $G_m^{-1}A_m(\widetilde{\xi}_1,\dots,\widetilde{\xi}_n)$'s is bounded by M. □

Remark 5.3. If M is not too large in the sense that $M^3n^{-1/2}$ is small, then, selecting a proper δ, we can derive from this statement that the distribution of the maximum of $G_m^{-1}A_m(\xi_1,\dots,\xi_n)$'s is approximated by the similar distributions for $G_m^{-1}A_m(\widetilde{\xi}_1,\dots,\widetilde{\xi}_n)$'s.

References

[1] Andrews, D.W.K. (1997). A conditional Kolmogorov test. *Econometrika* **58** 1097–1128.

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

[3] Bierens, H.J. (1990). A consistent conditional moment test of functional form. *Econometrika* **58** 1443–1458.

[4] Brown, L.D. and Low, M.G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, **24**, 2384–2398.

[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] Eubank, R.L. and Spiegelman, C.H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Amer. Stat. Ass.* **85** 387–392.
[7] Eubank, R.L. and Hart, J.D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Ann. Statist.* **20** no. 3, 1424–1425.
[8] Ermakov, M.S. (1990). Minimax detection of a signal in a white Gaussian noise. *Theory Probab. Appl.*, **35**, 667–679.
[9] Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Ass.* **91** 674–688.
[10] Gasser, T. and Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633.
[11] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **4** 1926–1947.
[12] Hart, J. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests* New York, Berlin, Heidelberg: Springer.
[13] Ingster, Yu.I. (1982). Minimax nonparametric detection of signals in white Gaussian noise *Problems Inform. Transmission* **18** 130–140.
[14] Ingster, Yu.I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I–III. *Math. Methods Statist.* **2** 85–114, **3** 171–189, **4** 249–268.
[15] Ingster, Yu.I. and Suslina, I. (1998). Minimax nonparametric hypothesis testing for ellipsoids and Besov bodies. *Problems Inform. Transmission* **34** no. 1, 56–68.
[16] 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.
[17] Klein, R.L. and Spady, R.H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61** 387–421.
[18] Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *J. Amer. Stat. Ass.* **89** no. 427, 1000–1005.
[19] Lehmann, E.L. (1959). *Testing Statistical Hypothesis.* Wiley, New York.
[20] Lepski, O.V. and Spokoiny, V.G. (1999). Minimax Nonparametric Hypothesis Testing: The Case of an Inhomogeneous Alternative. *Bernoulli*, **5** (1999) no. 2, 333–358.
[21] Lepski, O. and Tsybakov, A. (1996). Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. SFB Discussion Paper **91**, Humboldt University, Berlin.
[22] Mann, H.B. and Wald, A. (1942). On the choice of the number of intervals in the application of the chi-square test. *Ann. Math. Stat.* **13** 306–317.
[23] Neyman, J. (1937). "Smooth test" for goodness of fit. *Scand. Aktuarietidskr.* **20** 149–199.
[24] Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.
[25] Spokoiny, V. (1996). Adaptive hypothesis testing using wavelets. *Annals of Statistics* **24** no. 6, 2477–2498.
[26] Spokoiny, V. (1998). Adaptive and spatially adaptive testing a nonparametric hypothesis. *Math. Methods of Stat.* **7** no. 3, 245–273.
[27] Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.* **25** no. 2, 613–641.