MINIMAX NONPARAMETRIC HYPOTHESIS TESTING: THE CASE OF AN INHOMOGENEOUS ALTERNATIVE

LEPSKI, O.V. AND SPOKOINY, V.G.*

Humboldt University, SFB 373, Spandauer Str. 1, 10178 Berlin

Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117 Berlin

Running title: Testing a hypothesis vs. an inhomogeneous alternative.

Abstract. We study the problem of testing a simple hypothesis for a nonparametric "signal + white noise" model. It is assumed under the null hypothesis that the "signal" is completely specified, for example, that no signal is present. This hypothesis is tested against a composite alternative of the following form: the underlying function (the signal) is separated away from the null in the L_2-norm and in addition, it possesses some smoothness properties. We focus on the case of a non-homogeneous alternative when the smoothness properties of the signal are measured in a L_p-norm with $p < 2$. We consider tests whose errors have probabilities which do not exceed prescribed values and we measure the quality of testing by the minimal distance between the null and the alternative set for which such testing is still possible. We evaluate the optimal rate of decay of this distance to zero as the noise level tends to zero. Then a rate-optimal test is proposed which essentially uses a pointwise-adaptive estimation procedure.

1. Introduction

This study is motivated by two sources. First of all, we issue from the series of results by Y. Ingster (1982–1993) on the problem of testing a nonparametric hypothesis. They showed an essential difference between estimation and testing in a nonparametric context. In particular, the minimax rate is different in these two problems. The other source is connected with the recent progress in nonparametric estimation which is now referred to as "spatially adaptive estimation". This direction in nonparametrics was initiated in the pioneering paper by Nemirovski (1985) and followed by a series of articles by D.

1991 *Mathematics Subject Classification.* 62G10; Secondary 62G20.

Key words and phrases. bandwidth selection, error probabilities, minimax hypothesis testing, nonparametric alternative, pointwise adaptive estimation, signal detection .

The authors thank A.Nemirovski, A. Juditsky and M. Neumann for helpful remarks and discussion. The authors were partially supported by INTAS, grant N93/1585.

Donoho, I. Johnstone, G. Kerkyacharian and D. Picard on wavelet estimation. It was shown that the classical linear methods of nonparametric estimation do not provide the optimal rate of convergence when functions with non-homogeneous smoothness properties are considered. To be rate-optimal, a method of estimation has to be locally adaptive ("spatially adaptive") and hence nonlinear. As an alternative to linear methods, a nonlinear wavelet procedure was proposed which turned out to be efficient for a wide range of criteria, see Donoho and Johnstone (1992, 1994), Kerkyacharian and Picard (1993), Delyon and Juditski (1994), Donoho et al. (1995).

Another "spatially adaptive" procedure was proposed in Lepski et al.(1994). This is a kernel estimator with a variable data-driven bandwidth. It turned out that this estimator retains most of the optimal properties possessed by the wavelet one. Lepski and Spokoiny (1995) enlarged on this result and proved that a slightly modified version of the initial procedure is asymptotically sharp-optimal for the problem of adaptive estimation at a point. This paper presents one more application of the idea of pointwise adaption: we apply it to the problem of hypothesis testing.

We are unable to describe in details the historical background of this problem. We mention only a few early pertinent results by Neyman (1937), Mann and Wald (1942), Huber (1956) among others. For more information see Ingster (1993). It was Ingster (1982) who initiated the study of the problem of testing a hypothesis from the modern minimax nonparametric point of view. True, some closely related considerations appeared in earlier papers by Burnashev (1979) and Ibragimov and Khasminskii (1977). Further progress in this direction was mostly due to the St.-Petersburg school, see Ingster (1993) for the detailed description of these results. We mention here only a few points which are important for our further exposition.

Typically, a null hypothesis corresponds to our belief that the observed data are organized in a relatively simple way, which means that the structure of the underlying model is completely specified. Therefore, when considering a goodness-of-fit problem of such a sort, it is natural to measure the quality of any test by its sensitivity to perturbations or contaminations of this model. The optimal test has to be sensitive for as large a set of alternatives as possible.

Below, we consider the "signal + noise" model when the observed process X is described by the stochastic differential equation

$$dX(t) = f(t)dt + \varepsilon dW(t)$$

where ε is the noise level and W denotes a standard white noise process. The null corresponds to the case when the signal is identically zero or, in the other words, no signal is present. The corresponding testing can be viewed as a problem of signal detection. A set of alternatives can be naturally defined in the following way. Let f be "true". We say that this function belongs to the alternative set if it is separated away from the "null" in some integral L_r-norm:

$$\|f\|_r := \left[\int |f(t)|^r dt\right]^{1/r} \geq \varrho. \tag{1.1}$$

The radius ϱ characterizes the sensitivity of testing. For a small noise level ε, we may expect that ϱ can be also small. Hence, it is assumed that ϱ depends on ε, $\varrho = \varrho_\varepsilon$, and our aim is to describe the optimal rate of decay $\varrho_\varepsilon \to 0$ under which testing with prescribed probabilities of errors is still possible. Note, however, that the assumption (1.1) is not sufficient for a consistent testing, see Ibragimov and Khasminskii (1977), Burnashev (1979) or Ingster (1982): with no assumption on the regularity of the signal, it is impossible to distinguish between this signal and a noise. Typically, one additionally assumes that the underlying function f possesses some smoothness properties. A standard assumption here is that f belongs to some function class $\mathcal{F}$, for instance, to some Hölder or Sobolev ball. The recent works done by Donoho, Johnstone, Kerkyacharian and Picard on wavelet methods in statistical inference showed that the formalism of the more general Besov function classes provides a useful technical tool for nonparametric statistical considerations. For a statistical analysis, the following factors appear to be of the greatest importance: the smoothness degree s, the relation between the norm power p, in which we measure smoothness properties of the function f and the norm power r, in which we measure errors of estimation or the distance between the null and the alternative set.

For the problem of estimation, the case $p \geq r$ is classical and here the linear methods provide the optimal rate of convergence which is equal to $\varepsilon^{2s/(2s+1)}$. If $p < r$, which corresponds to functions with non-homogeneous smoothness properties, the minimax rate of estimation is the same, but it cannot be acieved by any linear method, see Nemirovski (1985), Donoho and Johnstone (1992).

The situation significantly changes for the problem of hypothesis testing. If $p = r = 2$, then the optimal rate of testing is $\varepsilon^{4s/(4s+1)}$, see Ingster (1982), which is better than the rate of estimation. For this case not only the optimal rate is described, but also asymptotically optimal (up to an exact constant) test procedures were constructed, see Ermakov (1990) and Ingster (1993). The same rate is optimal for $r \leq 2$ and $p \geq 2$, Ingster (1986). Another unexpected feature of the testing problem is that for $p = r > 2$ the rate of testing depends on p. More precisely, see Ingster (1986), it is $\varepsilon^{2s/(2s+1-1/p)}$. However, the case of $p < r$, which corresponds to a set of alternatives with inhomogeneous smoothness properties, was not studied. At the same time, just as in the estimation problem, it is of essential importance both from the theoretical point of view and for applications.

Here we focus on the situation when $r = 2$ and $p < 2$ which admits of relatively simple and evident description and the proofs are clearer. The cases when $p < r$ and $r > 2$ or when $r > p \geq 2$ and $r < 2, p < 2$ are more involved and lead to new phenomena. Any further discussion of the problem of testing in a L_r-norm with $r \neq 2$ lies beyond the scope of this paper.

In the next section we formulate our problem and state the main results pertaining to the optimal rate of testing. In Section 3 we present tests which achieve the optimal rate of testing. Next we mention some possible directions for further developments and we postpone the proofs until the last section.

2. The minimax rate of testing

In this section we specify the problem of hypothesis testing and state the main results.

2.1. The model

Suppose that we are given the observations $X(t)$, $t \in [0,1]$ described by the stochastic differential equation

$$dX(t) = f(t)dt + \varepsilon dW(t), \qquad 0 \le t \le 1, \tag{2.1}$$

where $W(t)$ is a Brownian motion and f is an unknown function.

2.2. The null and the alternative

Our aim is to test the null hypothesis H_0 that the regression function (signal) f is identically zero,

$$H_0 : \qquad f \equiv 0.$$

Under the alternative, we assume that f is separated away from the null in the L_2-norm and belongs to some smoothness class $\mathcal{F}$. Below, we assume that $\mathcal{F}$ is a Besov ball $B^s_{p,q}(M)$ with

$$B^s_{p,q} = \{f : \|f\|_{B^s_{p,q}} \le M\}$$

where, see Triebel (1992),

$$\|f\|_{B^s_{p,q}} = \begin{cases} \|f\|_p + \left[\int\limits_0^1 h^{-sq} \|\mathrm{osc}\, f(\cdot,h)\|_p^q \frac{dh}{h}\right]^{1/q}, & \text{if } r < \infty, \\ \|f\|_p + \sup\limits_{0 \le h \le 1} h^{-s} \|\mathrm{osc}\, f(\cdot,h)\|_p, & \text{if } q = +\infty. \end{cases}$$

Here $\|f\|_p$ is the L_p-norm, $\|f\|_p^p = \int_0^1 |f|^p$, and the local oscillation $\mathrm{osc}\, f(x,h)$ of f is defined as

$$\mathrm{osc}\, f(x,h) = \inf \sup_{|y-x| \le h} |f(y) - P(y)|.$$

The infimum here is taken over all polynomials of order m, which is the maximal integer smaller than s , and the supremum with respect to x, y is restricted to the interval $[0,1]$. The parameters s, p, q, M are such that $p, q \geq 1$, $s, M > 0$ and $sp > 1$.

We arrive at a set of alternatives $\mathcal{F}(\varrho_\varepsilon)$ of the form

$$H_1: \qquad \mathcal{F}(\varrho_\varepsilon) = \{f \in B^s_{p,q}(M),\ \|f\| \geq \varrho_\varepsilon\}.$$

Here $\|f\|$ means the usual L_2 -norm, $\|f\|^2 = \int_0^1 f^2(t)dt$.

Remark 2.1. For an integer s , one may consider instead of the Besov norm $\|\cdot\|_{B^s_{p,q}}$ the Sobolev seminorm $\|\cdot\|_{W^s_p}$ with

$$\|f\|_{W^s_p} = \left(\int |f^{(s)}(t)|^p dt\right)^{1/p}$$

and $f^{(s)}(t)$ stands for the s -th generalized derivative of the function f .

The values of s, p, q, M entering into the definition of the alternative set are assumed fixed and known. Note, however, that only s and p are important for the results and the construction of the optimal tests. The problem of adaptive testing when the smoothness parameters are unknown is briefly discussed in Section 4.

2.3. The problem of hypothesis testing

A test ϕ_ε is a rule to accept or to reject the null hypothesis by means of the observed process $X(t), 0 \leq t \leq 1$, therefore, it is a measurable function of observations taking values in the two-point set $\{0,1\}$. The value $\phi_\varepsilon = 0$ is treated as accepting H_0 , and $\phi_\varepsilon = 1$ means that the test rejects H_0 .

The quality of any test is measured by the probabilities of the corresponding errors. The probability $\alpha_0(\phi_\varepsilon)$ of error of the first kind is the probability under the null to reject the hypothesis,

$$\alpha_0(\phi_\varepsilon) = P_0(\phi_\varepsilon = 1)$$

where P_0 is the distribution on the space of observations corresponding to H_0 .

The probability of error of the second kind can be viewed as the probability to accept H_0 if f belongs to the alternative set H_1 . We denote it by $\alpha_1(\phi_\varepsilon, \varrho_\varepsilon)$ taking into account that H_1 is a composite alternative:

$$\alpha_1(\phi_\varepsilon, \varrho_\varepsilon) = \sup_{f \in \mathcal{F}(\varrho_\varepsilon)} P_f(\phi_\varepsilon = 0)$$

where P_f is the distribution corresponding to a particular function f .

We are studying the asymptotic behavior of these probabilities as the noise level ε tends to zero. We are interested in describing the fastest rate of decay to zero of such a radius ϱ_ε for which it is still possible to construct a test ϕ_ε such that at least for a small level noise ε the probabilities $\alpha_0(\phi_\varepsilon)$ and $\alpha_1(\phi_\varepsilon, \varrho_\varepsilon)$ do not exceed some prescribed values α_0 and α_1 respectively.

2.4. **The main results**

Ingster (1982–1993) studied the above problem for function classes of the Sobolev type and for $p \geq 2$. The optimal rate ϱ_ε in this case is

$$\varrho_\varepsilon = \varepsilon^{\frac{4s}{4s+1}}.$$

Here we are concentrating on the situation when $p < 2$. As usual, we distinguish the results obtained for the lower and upper bounds. The first of these, related to the lower bound, describes the rate for ϱ_ε which cannot be improved by any test.

Theorem 2.1. *Let $p \leq 2$, $sp > 1$ and*

$$\varrho_\varepsilon = \varepsilon^{\frac{4s''}{4s''+1}} \tag{2.2}$$

where

$$s'' = s - \frac{1}{2p} + \frac{1}{4}.$$

Then, for any sequence ϱ'_ε with $\varrho'_\varepsilon / \varrho_\varepsilon \to 0$ as $\varepsilon \to 0$ and for any tests ϕ_ε

$$\liminf_{\varepsilon \to 0} [\alpha_0(\phi_\varepsilon) + \alpha_1(\phi_\varepsilon, \varrho'_\varepsilon)] \geq 1.$$

The second result for the upper bound claims that there exist tests ϕ^*_ε which provide the rate of testing ϱ_ε described in Theorem 2.1. Their structure is explained in the next section.

Theorem 2.2. *Let s,p and ϱ_ε be the same as in Theorem 2.1. For each positive α_0 and α_1, there exist a constant c_1, depending on s, p, q, M, and tests ϕ^*_ε such that*

$$\lim_{\varepsilon \to 0} \alpha_0(\phi^*_\varepsilon) \leq \alpha_0 \tag{2.3}$$

and

$$\lim_{\varepsilon \to 0} \alpha_1(\phi^*_\varepsilon, c_1 \varrho_\varepsilon) \leq \alpha_1. \tag{2.4}$$

3. **A test procedure**

In this section we explain the structure of the tests ϕ^*_ε mentioned in Theorem 2.2. We start with some preliminary discussion.

3.1. Preliminaries

Below we give some heuristic explanation of the results and the proposed test procedures.

We shall consider test statistics based on kernel smoothers. Let K be a kernel satisfying standard conditions, see $(K1)$ – $(K5)$ below. Let us also fix a bandwidth value $h \in [0,1]$ (which we specify later) and consider a kernel estimator

$$\widetilde{f}_h(t) = \frac{1}{h}\int K\left(\frac{t-s}{h}\right) dX(s), \qquad t \in [0,1].$$

This can be decomposed in a standard way into a deterministic and a stochastic part,

$$\widetilde{f}_h(t) = f_h(t) + \xi_h(t)$$

where

$$\begin{aligned} f_h(t) &= \frac{1}{h}\int K\left(\frac{t-s}{h}\right) f(s)ds, \\ \xi_h(t) &= \frac{\varepsilon}{h}\int K\left(\frac{t-s}{h}\right) dW(s). \end{aligned}$$

It is natural to use the value

$$T_h = \|\widetilde{f}_h\|^2 = \int_0^1 \widetilde{f}_h^2(t)dt$$

for testing the null hypothesis $H_0 : \|f\| = 0$ against the alternative $H_1 : \|f\| \geq \varrho_\varepsilon$. Under the null, one has $\widetilde{f}_h = \xi_h$ and

$$T_h = T_h^0 = \int_0^1 \xi_h^2(t)dt.$$

It is not difficult to derive that

$$\begin{aligned} ET_h^0 &= \frac{\varepsilon^2\|K\|^2}{h}, \\ DT_h^0 &= \frac{\varepsilon^4 d^2}{h}, \end{aligned}$$

where $d = d(K)$ is some constant depending only on the kernel K. Moreover, if

$$\eta_h^0 = \frac{T_h^0 - ET_h^0}{\sqrt{DT_h^0}},$$

then

$$\mathcal{L}\left(\eta_h^0 \mid P_0\right) \xrightarrow{w} \mathcal{N}(0,1).$$

This leads to the test of the form

$$\phi_h = \mathbf{1}(\eta_h > \chi_{\alpha_0})$$

where

$$\eta_h = \frac{T_h - ET_h^0}{\sqrt{DT_h^0}},$$

and χ_α is the $(1-\alpha_0)$-quantile of the standard normal law.

Under the alternative, for some $f \in H_1$, we have

$$T_h = \int_0^1 (f_h(t) + \xi_h(t))^2 dt = \|f_h\|^2 + T_h^0 + \text{cross term}.$$

It is easy to show that the "cross term" is relatively small. Hence,

$$\eta_h \asymp \eta_h^0 + \|f_h\|^2 (d\varepsilon^2 h^{-1/2})^{-1}$$

where "$\asymp$" means asymptotic equivalence. But $\|f_h\|^2 \geq \|f\|^2/2 - \|f-f_h\|^2$ and therefore, the test ϕ_h detects a signal f from the alternative set if

$$\|f\|^2 \geq C(\varepsilon^2\sqrt{h} + \|f - f_h\|^2) \tag{3.1}$$

for some sufficiently large C.

The value $\|f - f_h\|$ can be estimated by using of the smoothness condition $f \in B_{p,q}^s$. If $p \geq 2$, then, see Triebel (1992), $\|f - f_h\| \leq O(h^s)$. In this case, minimization over h in (3.1) leads to the bandwidth choice of $h = O(\varepsilon^{\frac{4}{4s+1}})$ and (3.1) is met for $\|f\| \geq C_1 \varepsilon^{\frac{4s}{4s+1}}$ with some $C_1 > 0$. Note that a value of $h = O(\varepsilon^{\frac{2}{2s+1}})$, which is typical for the estimation problem, leads to the rate of testing $\varepsilon^{\frac{2s}{2s+1}}$. This rate is usual for estimation but it is relatively poor for testing.

If $p < 2$, then the estimate $\|f - f_h\| \leq O(h^s)$ is no longer true. The condition $f \in B_{p,q}^s$ ensures only that $\|f - f_h\|_p = O(h^s)$. To get a bound for the L_2-norm, we can apply the embedding theorem for Besov classes, see Triebel (1992): $\|f - f_h\| \leq O(h^{s'})$ where $s' = s - 1/p + 1/2$. This approach obviously leads to the bandwidth choice $h' = O(\varepsilon^{\frac{4}{4s'+1}})$ and to the corresponding rate of testing $\varepsilon^{\frac{4s'}{4s'+1}}$. Since $s'' > s'$, it is worse than the optimal rate shown in Theorem 2.2. The situation here is similar to that is met for the estimation problem. Tests of type ϕ_h are analogous to linear methods in the estimation theory. It is known Nemirovski (1985), Donoho and Johnstone (1992), that for $p < 2$ linear methods, can only achieve the rate with $\varepsilon^{2s'/(2s'+1)}$ instead of $\varepsilon^{2s/2s+1}$. An improvement can be accomplished by nonlinear methods possessing some "spatially adaptive" properties, see Donoho et al. (1995) or Lepski et al. (1994). Below, this idea is extended onto the problem of hypothesis testing. Following Lepski et al. (1994), we apply a nonlinear pointwise-adaptive procedure which can be regarded as the described above kernel method with a variable data-driven bandwidth. In essence, this method allows to control the differences $|f_h(t) - f_{h/2}(t)|$ for different h from a diadic geometric grid. If such a difference is for some h and some $t \in [0,1]$ so large , that it cannot be explained by the noise flictuation, then we detect the signal. Otherwise we have a bound of the form $|f_h(t) - f_{h/2}(t)| \leq \lambda_h$ with some λ_h which allows to estimate

$$\|f_h - f_{h/2}\|^2 \leq \|f_h - f_{h/2}\|_p^p \|f_h - f_{h/2}\|_\infty^{2-p} \leq C h^{sp} \lambda_h^{2-p}.$$

Our calculatons are based exactly on this idea.

Some more information about the difference between the cases of $p \geq 2$ and $p < 2$ can be extracted from the structure of the least favorable prior distributions for the problem of detecting a random signal. Ingster (1982, 1986) showed that for $p \geq 2$ such a random signal is wiggling and uniformly small with the altitude of order $\varepsilon^{(2s+1)/(4s+1)}$ which is essentially smaller than the noise level. Note that the L_p-norm of this signal for any $p \geq 1$ is of the same order and depends on p very weakly.

By inspecting the proof of Theorem 2.1 one can see that for $p < 2$, the structure of the least favorable priors is entirely different. Namely, the corresponding random signal is almost everywhere zero with $N = \varepsilon^{-2/(2s+1-1/p)}$ peaks. Such a structure is caused by the extremal problem of maximizing over the given Besov class the L_p-norm of a function when the L_2-norm is fixed. In particular, the ratio $\|f\|_p/\|f\|_2$ for such signals tends to infinity as ε tends to zero. This explains why the rate of testing depends on p and justifies the using of the notion of an alternative with inhomogeneous smoothness properties.

3.2. A data-driven bandwidth selector

For the construction of tests we need in splitting the observed data $X(\cdot)$ from (2.1) into two independent parts. For a model with discrete time, the usual way of doing this is in splitting the observations into even and odd points. For the continuous-time model (2.1), the following method can be used. Let W' be a white Gaussian noise independent of W. Define two processes $\widetilde{X}$ and $\tilde{\tilde{X}}$ by

$$\begin{aligned} \widetilde{X}(t) &= X(t) + \varepsilon W'(t), \\ \tilde{\tilde{X}}(t) &= X(t) - \varepsilon W'(t). \end{aligned}$$

Obviously, $\widetilde{X}$ and $\tilde{\tilde{X}}$ obey the equations

$$\begin{aligned} d\widetilde{X}(t) &= f(t)dt + \varepsilon\sqrt{2}\, d\widetilde{W}(t), \\ d\tilde{\tilde{X}}(t) &= f(t)dt + \varepsilon\sqrt{2}\, d\tilde{\tilde{W}}(t), \end{aligned}$$

where

$$\begin{aligned} \widetilde{W}(t) &= 2^{-1/2}[W(t) + W'(t)], \\ \tilde{\tilde{W}}(t) &= 2^{-1/2}[W(t) - W'(t)], \end{aligned}$$

are two independent white Gaussian noises. We treat $\widetilde{X}$ and $\tilde{\tilde{X}}$ as two independent data sets. One part provided by $\widetilde{X}$ will be used for a pointwise bandwidth selection and the other one, for constructing the kernel-type test statistics with the plugged-in bandwidth. This splitting procedure obviously leads to some loss of efficiency which is manifested by an increase in the noise level (by $\sqrt{2}$) for the process $\tilde{\tilde{X}}$. This factor $\sqrt{2}$ can be viewed as a payment for the pointwise adaptation.

Now we introduce a family of kernel estimators with a kernel K satisfying usual regularity conditions. Let $m = \lfloor s \rfloor$, the largest possible integer smaller than s. Let now $K(u)$ be a function defined on the real axis such that

$(K1)$ it is symmetric, $K(u) = K(-u)$, $u \in R^1$;

$(K2)$ it is compactly supported i.e. $K(u) = 0$ for $|u| > b$ for some $b > 0$;

$(K3)$ it is continuous;

$(K4)$ $\int K(u)du = 1$;

$(K5)$ $\int K(u)u^i du = 0, \qquad i = 1, \ldots, m$.

In what follows, we omit the integration limit if the integration is taken over the whole real line.

Denote, for given $h > 0$ and $t \in [0, 1]$,

$$\begin{aligned} \widetilde{f}_h(t) &= \frac{1}{h} \int K\left(\frac{t-s}{h}\right) d\widetilde{X}(s), \\ \tilde{\tilde{f}}_h(t) &= \frac{1}{h} \int K\left(\frac{t-s}{h}\right) d\tilde{\tilde{X}}(s). \end{aligned}$$

Remark 3.1. These definitions should be corrected near the end points $t = 0$ and $t = 1$ which might be done in a standard way by replacing the kernel near these points by special boundary one-sided kernels. Therefore, we actually need three kernels: one (symmetric) for application inside the interval $(0, 1)$; another one (right-sided with a support of the form $[0, b]$) for applying near the point 0; and the third one (left-sided with a support of the form $[-b, 0]$), near 1. All the three kernels should satisfy the above-mentioned conditions $(K1)$ through $(K5)$. For more details see, for instance, Lepski et al.(1994). To simplify the exposition, we retain the notation K for the boundary corrected kernel.

Now we describe a pointwise bandwidth selector introduced in Lepski et al.(1994), see also Lepski and Spokoiny (1995). We begin by introducing a set $\mathcal{H}$. Our pointwise bandwidth takes its values in this set. Denote

$$h^* = \varepsilon^{\frac{4}{4s''+1}} = \varepsilon^{\frac{2}{2s+1-1/p}} \tag{3.2}$$

and set

$$\mathcal{H} = \{h = h^* 2^{-k},\ k = 0, 1, 2, \ldots, h \geq \varepsilon^2\}.$$

In particular, h^* is the largest considered bandwidth value. We also apply h^* to define the boundary corrected kernel: the symmetric kernel K is to be replaced by the right-sided kernel in the interval $[0, bh^*]$ and by the left-sided kernel in $[1 - bh^*, 1]$.

Given η, h from $\mathcal{H}$ with $\eta < h$ and $c = \eta/h$, set

$$\sigma^2(\eta, h) = \frac{2\varepsilon^2}{\eta} \int |K(u) - cK(uc)|^2 du \tag{3.3}$$

and

$$\psi(\eta, h) = \sigma(\eta, h)\sqrt{2\ln(h^*/\eta)}.$$

Denote also

$$C(K) = \sup_{0\le c\le 1} \int |K(u) - cK(uc)|^2 du$$

and

$$\psi(h) = \frac{\varepsilon\sqrt{2}C(K)}{\sqrt{h}}\sqrt{\max\{2\ln(h^*/h), 1\}}. \tag{3.4}$$

Note that the values $C(K), \psi(\eta, h)$ and $\psi(h)$ depend on t via the boundary corrected kernel K.

Given $t \in [0, 1]$, define the pointwise data-driven bandwidth $\widehat{h}(t)$ by

$$\widehat{h}(t) = \max\left\{h \in \mathcal{H} : |\tilde{\tilde{f}}_\eta(t) - \tilde{\tilde{f}}_h(t)| \le \psi(\eta, h) + 2\psi(h), \quad \forall \eta \in \mathcal{H}, \eta < h\right\}.$$

3.3. A test

First we define an estimator $\widehat{f}(t)$ which is the kernel estimator $\widetilde{f}_h(t)$ with the plugged-in bandwidth $\widehat{h}$,

$$\widehat{f}(t) = \widetilde{f}_{\widehat{h}(t)}(t), \qquad t \in [0, 1].$$

Denote for $h \in \mathcal{H}$

$$B(h) = \frac{2\varepsilon^2\|K\|^2}{h},$$

where $\|K\|^2 = \int K^2(u)du$ and introduce statistics T_ε,

$$T_\varepsilon = \varepsilon^{-2}\sqrt{h^*}\int_0^1 \left[\widehat{f}^2(t) - B(\widehat{h}(t))\right] dt. \tag{3.5}$$

Below we will show that under the null the T_ε's are asymptotically normal $\mathcal{N}(0, d^2)$ with some $d > 0$ and in particular,

$$\lim_{\varepsilon\to 0} E_0 T_\varepsilon^2 = d^2. \tag{3.6}$$

The test ϕ_ε^* is based exactly on these statistics T_ε: we reject the null hypothesis if T_ε is large enough. More precisely,

$$\phi_\varepsilon^* = \mathbf{1}(T_\varepsilon/d > \chi_{\alpha_0})$$

where χ_α is defined for $\alpha \in (0, 1)$ by $\Phi(\chi_\alpha) = 1 - \alpha$, Φ being the Laplace function.

Remark 3.2. It follows from (3.6) that the value d is determined only by the behavior of the test statistic T_ε under the null hypothesis. Therefore, for numerical calculations it is not necessary to derive this value issuing from its theoretical expression; it can be calculated by the Monte-Carlo method for model (2.1) with $f \equiv 0$.

4. Some further developments

4.1. Other nonparametric models

In this study we restrict ourselves to the “ideal” (and convenient from the technical point of view) “signal + white noise” model. We would expect that the main results remain valid for more realistic statistical models such as the probability density model, the regression model etc. (perhaps under additional assumptions). We indicate here the relevant results by Ingster (1984a, 1984b, 1986, 1993) on minimax hypothesis testing for the density and spectral density models and the results by Brown and Low (1996), Nussbaum (1996) on the asymptotic equivalence between the regression (resp. density) model and the “signal + white noise” model.

4.2. Parametric versus nonparametric fits

This study focuses on the simple null hypothesis. Note, however, that a parametric null hypothesis with unknown values of parameters is more typical in practical applications. This means that the null hypothesis H_0 is of the form $f \in \{f_\theta, \theta \in \Theta\}$ where Θ is an open subset of the Euclidean space R^k. The alternative is again smooth and separated away from this parametric family $\{f_\theta\}$:

$$\inf_{\theta \in \Theta} \|f - f_\theta\| \geq \varrho_\varepsilon .$$

But such a testing problem can be reduced to the above considered one with a simple null using the following method. First a pilot parametric estimator $\widetilde{\theta}$ of the parameter θ is constructed; this can be typically done ε-consistently. Then the corresponding “parametric” estimator $f_{\widetilde{\theta}}$ can be subtracted from the observed data and we arrive at the situation with the simple null hypothesis. The crucial point here is that the rate of parametric estimation is higher than that of nonparametric testing. An example of such calculations can be found in Härdle and Mammen (1993).

4.3. Adaptive testing

One aspect of the problem of hypothesis testing in the nonparametric set-up is of special importance for practical applications, namely, that the structure of the proposed test depend critically on the smoothness parameters s,p whose prior knowledge is typically lacking. In our procedure, the value of the largest applied bandwidth h^* depends on s and p. An inspection of the proof shows that a whong choice of this value leads to an essentially worse rate of testing. This fact raises an important issue such as ‘Can this parameter be selected in an adaptive (data-based) way without any loss of sensitivity?’ A recent result Spokoiny (1996) shows that an adaptive testing is indeed possible with a loss of power by a negligible log log-factor.

5. Proof of Theorem 2.1 and 2.2

5.1. Proof of Theorem 2.1

We follow Ingster (1993). Let ϱ_ε be the same as in Theorem 2.1 and suppose that ϱ'_ε is such that $c_\varepsilon = \varrho'_\varepsilon/\varrho_\varepsilon \to 0$. We show that for any tests ϕ_ε

$$\liminf_{\varepsilon\to 0} \left[\alpha_0(\phi_\varepsilon) + \alpha_1(\phi_\varepsilon, \varrho'_\varepsilon)\right] \geq 1. \tag{5.1}$$

The idea of the method is standard: in essence, the minimax problem is replaced by a Bayes one. Let π_ε be (prior) measures on the alternative set $\mathcal{F}(\varrho'_\varepsilon) = \{f \in B^s_{p,q}(M) : \|f\| \geq \varrho'_\varepsilon\}$. Denote by P_{π_ε} the corresponding Bayes measure for model (2.1), $P_{\pi_\varepsilon} = \int P_f \pi_\varepsilon(df)$. Let also

$$Z_{\pi_\varepsilon} = dP_{\pi_\varepsilon}/dP_0,$$

where the measure P_0 corresponds to the null hypothesis. It is well known that (5.1) follows from

$$Z_{\pi_\varepsilon} \xrightarrow{w} 1, \tag{5.2}$$

see, for instance, Ingster (1993, II, p.171). He also showed in the same place that it is not necessary for the priors π_ε to be supported on $\mathcal{F}(\varrho'_\varepsilon)$, it is sufficient that

$$\pi_\varepsilon(\mathcal{F}(\varrho'_\varepsilon)) \to 1. \tag{5.3}$$

For the construction of the priors π_ε satisfying (5.2) and (5.3) we use the method described in Ingster (1993, Section 4.3). Let G be a smooth function supported on $[-1,1]$. Assume also that a parameter h is small enough; we specify its choice later. Denote by $\mathcal{I}$ the partition of the interval $[-1,1]$ into intervals of length $2h$ with N being their number. Without loss of generality we assume that

$$Nh = 1. \tag{5.4}$$

Denote by t_I the center of an interval I from $\mathcal{I}$ and introduce the family of functions $\varphi_I(\cdot)$, $I \in \mathcal{I}$, on $[-1,1]$ with

$$\varphi_I(t) = \frac{1}{\sqrt{h}\|G\|} G\left(\frac{t-t_I}{h}\right)$$

where $\|G\|^2 = \int G^2(t)dt$. It is easy to see that these functions form an orthonormal set of functions on $[-1,1]$.

Consider now the random signal

$$f(t) = \varepsilon c_\varepsilon \sum_{I\in\mathcal{I}} \xi_I \varphi_I(t)$$

where $c_\varepsilon = \varrho'_\varepsilon/\varrho_\varepsilon$, ξ_I, $I \in \mathcal{I}$, are independent identically distributed random variables with values in the three-point set $\{-1, 0, 1\}$ having the distribution

$$P(\xi_I = 0) = 1 - \sqrt{h}, \qquad P(\xi_I = \pm 1) = \sqrt{h}/2, \qquad I \in \mathcal{I}. \tag{5.5}$$

Let a prior measure π_ε correspond to the distribution of such random signal f. Ingster (1993, II, p.176) established (5.2) for such priors with arbitrary $h = h_\varepsilon \to 0$ as $\varepsilon \to 0$. To prove (5.3) we need to specify the choice of h. Let us take

$$h = h_\varepsilon = \varepsilon^{\frac{2}{2s+1-1/p}}. \tag{5.6}$$

We use the following technical assertion.

Lemma 5.1. *For any s, p, q, M satisfying the conditions of the theorem and any set $(\xi_I, I \in \mathcal{I})$ one has*

$$\|f\|^p_{B^s_{p,q}} \le C(G) c^p_\varepsilon \sqrt{h} \sum_{I \in \mathcal{I}} |\xi_I|. \tag{5.7}$$

Proof. We present only a sketch of the proof for the Sobolev seminorm $\|f\|_{W^s_p} = \left(\int |f^{(s)}(t)|^p dt\right)^{1/p}$ where $f^{(s)}(t)$ means the s-th generalized derivative of the function f. The arbitrary Besov norm can be handled in a similar way using a standard technique of the approximation theory, see Triebel (1992). Obviously,

$$\begin{aligned} \int |f^{(s)}(t)|^p dt &= (\varepsilon c_\varepsilon \|G\|^{-1} h^{-1/2})^p \sum_{I \in \mathcal{I}} |\xi_I|^p \int |h^{-s} G^{(s)}(h^{-1} \cdot)|^p = \\ &= C(G)(\varepsilon c_\varepsilon)^p h^{-sp-p/2+1} \sum_{I \in \mathcal{I}} |\xi_I| \end{aligned} \tag{5.8}$$

where $C(G) = \|G\|^{-p} \int |G^{(s)}(\cdot)|^p$. This, coupled with (5.6), yields the assertion. □

Lemma 5.2. *Let ξ_I, $I \in \mathcal{I}$, be independent identically distributed random variables with distribution (5.5). Then*

$$\sqrt{h} \sum_{I \in \mathcal{I}} |\xi_I| \xrightarrow{P} 1.$$

Proof. This statement is simply the law of large numbers for a sample of independent random variables with the distribution (5.5); for more details see Ingster (1993, Section 4.3). □

Since $c_\varepsilon \to 0$, the above lemmas guarantee that, with a high probability, the function f lies in the ball $B^s_{p,q}(M)$. Now, similarly

$$\|f\|^2 = (\varepsilon c_\varepsilon)^2 \sum_{I \in \mathcal{I}} |\xi_I|^2 \approx (\varepsilon c_\varepsilon)^2 h^{-1/2} = c^2_\varepsilon \varrho^2_\varepsilon = \varrho'^2_\varepsilon$$

which completes the proof of (5.3).

5.2. **Proof of Theorem 2.2**

We begin by decomposing the test statistics T_ε from (3.5) using the standard decomposition of the kernel estimator $\widetilde{f}_h(t)$ into a deterministic and a stochastic term. Namely, for each $h > 0$ and any $t \in [0,1]$, we have

$$\widetilde{f}_h(t) = f_h(t) + \xi_h(t), \tag{5.9}$$

where

$$\begin{aligned} f_h(t) &= \frac{1}{h}\int K\left(\frac{t-s}{h}\right) f(s)ds, \\ \xi_h(t) &= \frac{\varepsilon\sqrt{2}}{h}\int K\left(\frac{t-s}{h}\right) d\widetilde{W}(s). \end{aligned}$$

A similar decomposition holds true for $\tilde{\tilde{f}}_h$ with $\tilde{\tilde{W}}$ in place of $\widetilde{W}$.

Now we note that, by (3.2) and (2.2),

$$\varepsilon^2/\sqrt{h^*} = \varrho_\varepsilon^2.$$

Next, obviously

$$|\widetilde{f}_h(t)|^2 = |f_h(t)|^2 + 2f_h(t)\xi_h(t) + |\xi_h(t)|^2, \tag{5.10}$$

and, in view of (3.5),

$$T_\varepsilon = \varrho_\varepsilon^{-2}\int_0^1 \left[|\widetilde{f}_{\widehat{h}(t)}(t)|^2 - B(\widehat{h}(t))\right] dt = \varrho_\varepsilon^{-2}[\widehat{S} + 2\gamma_\varepsilon] + R_\varepsilon$$

where

$$\begin{aligned} \widehat{S} &= \int_0^1 f^2_{\widehat{h}(t)}(t)dt, & (5.11)\\ \gamma_\varepsilon &= \int_0^1 f_{\widehat{h}(t)}(t)\xi_{\widehat{h}(t)}(t)dt, & (5.12)\\ R_\varepsilon &= \frac{\sqrt{h^*}}{\varepsilon^2}\int_0^1 \left[\xi^2_{\widehat{h}(t)}(t) - \frac{2\varepsilon^2\|K\|^2}{\widehat{h}(t)}\right] dt = \\ &= 2\|K\|^2\sqrt{h^*}\int_0^1 \widehat{h}(t)^{-1}\left[\zeta^2_{\widehat{h}(t)}(t) - 1\right] dt & (5.13) \end{aligned}$$

with

$$\zeta_h(t) = \frac{\sqrt{h}}{\varepsilon\sqrt{2}\|K\|}\xi_h(t) = \frac{1}{\|K\|\sqrt{h}}\int K\left(\frac{t-s}{h}\right) d\widetilde{W}(s). \tag{5.14}$$

The idea of the proof is as follows. To show (2.3) we note that under the null the terms $\widehat{S}$ and γ_ε vanish and it remains to check that R_ε is asymptotically normal with zero mean and a finite variance d^2.

Let now f be an arbitrary function from $B^s_{p,q}(M)$. First we check that the "stochastic" term R_ε is bounded in probability uniformly in $f \in B^s_{p,q}(M)$; more precisely, for a small enough ε and a large enough z_1,

$$\sup_{f\in B^s_{p,q}(M)} P_f(R_\varepsilon > z_1) \le \alpha_1/2. \tag{5.15}$$

The next step is to show that the cross term γ_ε is relatively small; for each $\delta > 0$

$$P(2\gamma_\varepsilon > \delta(\widehat{S} + \varrho_\varepsilon^2)) = o_\varepsilon(1). \tag{5.16}$$

(Here and in what follows $o_\varepsilon(1)$ denotes any sequence depending on ε only and vanishing as $\varepsilon \to 0$. If there is no risk of confusion, we also omit the index f in P_f).

Note then that for each $h \in \mathcal{H}$ and any t

$$f_h^2(t) \ge \frac{1}{2} f^2(t) - |f(t) - f_h(t)|^2; \tag{5.17}$$

hence, by (5.11)

$$\widehat{S} \ge 1/2 \int_0^1 f^2(t)dt - \int_0^1 |f(t) - f_{\widehat{h}(t)}(t)|^2 dt. \tag{5.18}$$

Denote

$$Q_\varepsilon = \varrho_\varepsilon^{-2} \int_0^1 |f(t) - f_{\widehat{h}(t)}(t)|^2 dt.$$

We shall prove later that Q_ε is bounded in probability in the same sense as R_ε:

$$\sup_{f\in B^s_{p,q}(M)} P_f(Q_\varepsilon > z_2) \le \alpha_1/2 \tag{5.19}$$

if ε is small enough and z_2 is sufficiently large. Now we are showing how statement (2.4) of the theorem follows from (5.15), (5.19) and (5.16). In fact, making use of (5.16) and (5.18), one has for $\delta \le 1/3$ and any $f \in B^s_{p,q}(M)$

$$\begin{aligned} P(T_\varepsilon > z) &= P(\varrho_\varepsilon^{-2}(\widehat{S} + 2\gamma_\varepsilon) + R_\varepsilon > z) \ge \\ &\ge P(\varrho_\varepsilon^{-2}\widehat{S}(1-\delta) - \delta + R_\varepsilon > z) - o_\varepsilon(1) \ge \\ &\ge P(\varrho_\varepsilon^{-2}\|f\|^2/3 - \tfrac{2}{3}Q_\varepsilon + R_\varepsilon > z + 1/3) - o_\varepsilon(1). \end{aligned}$$

Let $z = d\chi_{\alpha_0}$ and suppose that z_1 and z_2 are the same as in (5.15) and (5.19) respectively. If f is such that $\|f\|^2 > 3\varrho_\varepsilon^2(z + 1/3 + z_1 + 2z_2/3)$, then

$$P(\phi^* = 1) = P(T_\varepsilon > z) \ge 1 - P(R_\varepsilon > z_1) - P(Q_\varepsilon > z_2) - o_\varepsilon(1) \ge 1 - \alpha_1 - o_\varepsilon(1)$$

as required in (2.4).

Therefore, to prove the theorem it suffices to show the asymptotic normality of R_ε under the null hypothesis and to check (5.15), (5.19) and (5.16). We begin by estimating R_ε. Denote by $\widetilde{\mathcal{G}}$ and $\tilde{\tilde{\mathcal{G}}}$ the σ-algebras generated by the random processes $\widetilde{W}$ and $\tilde{\tilde{W}}$ respectively. Since $\widetilde{W}$ and $\tilde{\tilde{W}}$ are independent, these algebras are also independent.

By definition, for each $h \in \mathcal{H}$ and any $t \in [0,1]$, the random variables $\widetilde{f}_h(t)$ are $\widetilde{\mathcal{G}}$-measurable, but $\tilde{\tilde{f}}_h(t)$ are $\tilde{\tilde{\mathcal{G}}}$-measurable, and so are $\widehat{h}(t)$. Therefore, the processes $\widetilde{f}_h(\cdot)$ and $\widehat{h}(\cdot)$ are independent. It is convenient to denote by $\widehat{E}$ the conditional expectation w.r.t. the σ-algebra $\tilde{\tilde{\mathcal{G}}}$, or, in the other words, the conditional expectation given $\widehat{h}(\cdot)$. Clearly $ER_\varepsilon^2 = E(\widehat{E}R_\varepsilon^2)$. To estimate $\widehat{E}R_\varepsilon^2$ we apply representation (5.13) and make use of some simple properties of the random variables $\zeta_h(t)$ from (5.14) collected in the next lemma.

Lemma 5.3. *Let $\zeta_h(t)$ be defined by (5.14). Then*

(i) *The random variables $\zeta_h(t)$ are standard normal and, in particular,* $E\zeta_h(t) = 0$, $E\zeta_h^2(t) = 1$.

(ii) *If $\eta, h \in \mathcal{H}$ and $|t-s| > b(\eta+h)$, then the random variables $\zeta_\eta(s)$ and $\zeta_h(t)$ are independent and, in particular,*

$$E\,\zeta_\eta(s)\zeta_h(t) = 0,$$
$$E(\zeta_\eta^2(s)-1)(\zeta_h^2(t)-1) = 0.$$

(iii) *If $\eta < h$, then for any s,t*

$$|E\zeta_\eta(s)\zeta_h(t)| \le C_1(K)\sqrt{\eta/h},$$
$$|E(\zeta_\eta^2(s)-1)(\zeta_h^2(t)-1)| \le C_2(K)\,\eta/h,$$

where $C_1(K)$ and $C_2(K)$ are some absolute constants depending only on the kernel K.

Proof. The first statement follows directly from (5.14). The second one holds because the supports of the functions $K((t-u)/h)$ and $K((s-u)/\eta)$ do not intersect for s,t with $|s-t| > b(\eta+h)$, and because the white noise $\widetilde{W}$ has independent increments .

Next, it follows directly from (5.14) that

$$|E\zeta_\eta(s)\zeta_h(t)| = \frac{1}{\sqrt{\eta h}\|K\|^2}\left|\int K\left(\frac{s-u}{\eta}\right)K\left(\frac{t-u}{h}\right)du\right| \le$$
$$\le \frac{\|K\|_\infty}{\|K\|^2}\sqrt{\eta/h}\int\left|K\left(\frac{s-u}{\eta}\right)\right|d\frac{u}{\eta} \le C_1(K)\sqrt{\eta/h}$$

where $\|K\|_\infty = \sup_u|K(u)|$ and $C_1(K) = \|K\|_\infty\|K\|_1\|K\|^{-2}$ with $\|K\|_1 = \int|K(u)|du$. This implies the first statement in (iii). Now, since $\zeta_\eta(s), \zeta_h(t)$ are standard normal, straightforward calculations provide

$$E(\zeta_\eta^2(s)-1)(\zeta_h^2(t)-1) = 2|E\zeta_\eta(s)\zeta_h(t)|^2.$$

Thus the second assertion in (iii) follows. □

Denote

$$V_\varepsilon(t) = \frac{\sqrt{h^*}}{\widehat{h}(t)}(\zeta_{\widehat{h}(t)}^2 - 1). \tag{5.20}$$

Since $\widehat{h}(t)$ takes values in $\mathcal{H}$, one may also use the following representation

$$V_\varepsilon(t) = \sqrt{h^*} \sum_{h \in \mathcal{H}} h^{-1}(\zeta_h^2 - 1)\, \mathbf{1}(\widehat{h}(t) = h).$$

Now, applying (ii) and (iii) of Lemma 5.3, we obtain

$$\begin{aligned}
&|\widehat{E}\, V_\varepsilon(t) V_\varepsilon(s)| = \\
&= h^* \left| \sum_{h\in\mathcal{H}} \sum_{\eta\in\mathcal{H}} (\eta h)^{-1} E(\zeta_\eta^2(s) - 1)(\zeta_h^2(t) - 1)\, \mathbf{1}(\widehat{h}(t) = h, \widehat{h}(s) = \eta) \right| \le \\
&\le 2C_2(K) h^* \sum_{h\in\mathcal{H}} \sum_{\eta\in\mathcal{H}, \eta<h} h^{-2} \mathbf{1}(|t-s| \le b(\eta + h))\, \mathbf{1}(\widehat{h}(t) = h, \widehat{h}(s) = \eta) \le \\
&\le 2C_2(K) h^* \sum_{h\in\mathcal{H}} h^{-2} \mathbf{1}(|t-s| \le 2bh)\, \mathbf{1}(\widehat{h}(s) \le h).
\end{aligned}$$

Hence

$$\begin{aligned}
\widehat{E} R_\varepsilon^2 &= \widehat{E} \left(\int_0^1 V_\varepsilon(t) dt \right)^2 = \\
&= \int_0^1 \int_0^1 \widehat{E} V_\varepsilon(t) V_\varepsilon(s) dt\, ds \le \\
&\le 2C_2(K) h^* \sum_{h\in\mathcal{H}} h^{-2} \int_0^1 \int_0^1 \mathbf{1}(|t-s| \le 2bh)\, \mathbf{1}(\widehat{h}(s) \le h) dt\, ds = \\
&= 8bC_2(K) \sum_{h\in\mathcal{H}} h^*/h \int_0^1 \mathbf{1}(\widehat{h}(s) \le h) ds. \qquad (5.21)
\end{aligned}$$

This immediately gives

$$E\, R_\varepsilon^2 \le 8bC_2(K) \sum_{h\in\mathcal{H}} h^*/h \int_0^1 P(\widehat{h}(s) \le h) ds. \qquad (5.22)$$

Note, that the above calculations are valid for any arbitrary function f. Now we analyze the last sum supposing that $f \equiv 0$. In this case the estimators $\tilde{\tilde{f}}_h(t)$ consist only of the stochastic term coinciding in distribution with $\xi_h(t)$. Hence, by definition of $\widehat{h}(t)$ we obtain for each $h_1 \in \mathcal{H}$

$$P(\widehat{h}(t) \le h_1) \le \sum_{h\in\mathcal{H}, h\le 2h_1} \sum_{\eta\in\mathcal{H}, \eta<h} P(|\xi_\eta(t) - \xi_h(t)| > \psi(\eta, h)).$$

The difference $|\xi_\eta(t) - \xi_h(t)|$ is a Gaussian random variable with the variance $\sigma^2(\eta, h)$, see (3.3), and

$$\begin{aligned}
P(|\xi_\eta(t) - \xi_h(t)| > \psi(\eta, h)) &= P\left(|\zeta| > 2\sqrt{\ln(h^*/\eta)} \right) \le 2\exp\{-2\ln(h^*/\eta)\} \\
&= 2(\eta/h^*)^2. \qquad (5.23)
\end{aligned}$$

Here ζ denotes a standard normal random variable. Making use of the definition of the set $\mathcal{H}$ as a dyadic series, we conclude that

$$P(\widehat{h}(t) \le h_1) \le \sum_{h\in\mathcal{H}, h\le 2h_1} \sum_{\eta\in\mathcal{H},\eta<h} 2(\eta/h^*)^2 \le 2(2h_1/h^*)^2.$$

By (5.22) this yields

$$E\, R_\varepsilon^2 \le 64bC_2(K) \sum_{h\in\mathcal{H}} (h/h^*) \le 128bC_2(K). \tag{5.24}$$

Note that this bound is sufficient to prove (5.15) with $f \equiv 0$. But we need to prove the asymptotic normality of R_ε under H_0. Let $V_\varepsilon(t)$ be given by (5.20). Define the process $U_\varepsilon(u)$ by

$$U_\varepsilon(u) = \sqrt{h^*} V_\varepsilon(uh^*), \qquad 0 \le u \le 1/h^*.$$

With this notation we obtain from (5.13)

$$R_\varepsilon = 2\|K\|^2 \sqrt{h^*} \int_0^{1/h^*} U_\varepsilon(u) du. \tag{5.25}$$

It is easy to see that the process $U_\varepsilon(u)$ is stationary under H_0 in the interval $u \in [b, \frac{1}{h^*} - b]$ because this holds true for the processes $\xi_h(\cdot)$ and $\widehat{h}(\cdot)$. Non-stationarity in the subintervals $[0, b]$ and $[\frac{1}{h^*} - b, \frac{1}{h^*}]$ is caused by the correction of the kernel at the end points. Next, statement (ii) of Lemma 5.3 shows that the process U_ε is mixing and finite-dependent, which means that $U_\varepsilon(u)$ and $U_\varepsilon(u')$ are independent if $|u - u'| > 2b$. Moreover, an easy analysis proves that the distribution of U_ε does not depend on ε. These facts along with (5.24) allow us to apply the central limit theorem to the integral of U_ε over the interval from b to $\frac{1}{h^*} - b$, see e.g. Ibragimov and Linnik (1965, Section XVIII.7). This clearly leads to an asymptotic normality of R_ε, compare (5.25).

We turn now to studying the behavior of the term R_ε for an arbitrary function $f \in B^s_{p,q}$. In contrast with the above case, the process $\widehat{h}(t)$ is not stationary anymore, because it describes local smoothness properties of the function f which, generally speaking, vary from point to point. The same is true for the above defined processes V_ε and U_ε. But estimate (5.22) remains valid and we show that it leads to (5.15). Namely we are verifying that

$$\sup_{f\in B^s_{p,q}(M)} E\, R_\varepsilon^2 \le C' \tag{5.26}$$

with some constant C' depending possibly on the parameters s, p, q, M. This yields (5.15) by the Chebyshev inequality. For this purpose we introduce a useful pointwise characteristic of the function f which reflects the local smoothness properties of this function in a small vicinity of each point. This notion in a slightly modified form was used in Lepski et al.(1994) and Lepski and Spokoiny (1995).

Given $t \in [0,1]$ and $h > 0$, let

$$\Delta_f(h,t) = \max_{\eta \in \mathcal{H}, \eta \le h} |f(t) - f_\eta(t)|.$$

Also set

$$h_f(t) = \max\{h \in \mathcal{H} : \Delta_f(h,t) \le \psi(h)\}, \tag{5.27}$$

$\psi(h)$ being defined in (3.4). Obviously

$$\psi(2h) \ge \psi(h)/\sqrt{3}$$

and definition (5.27) yields

$$|f(t) - f_h(t)| \le \psi(h_f(t)), \qquad \forall h \in \mathcal{H},\ h \le h_f(t), \tag{5.28}$$

$$|f(t) - f_{2h_f(t)}(t)| > \psi(2h_f(t)) > \psi(h_f(t))/\sqrt{3}, \qquad \text{if } h_f(t) < h^*. \tag{5.29}$$

Now we note that

$$P(\widehat{h}(t) \le h) \le \mathbf{1}(h_f(t) \le h) + P(\widehat{h}(t) \le h, h < h_f(t)). \tag{5.30}$$

The second term in the left side can be easily estimated.

Lemma 5.4. *For each* $t \in [0,1]$

$$P(\widehat{h}(t) \le h, h < h_f(t)) \le 2(h/h^*)^2.$$

Proof. Let us fix some $t \in [0,1]$ and set $h_1 = h_f(t)$. By the definition of $\widehat{h}(t)$

$$\begin{aligned} & P(\widehat{h}(t) \le h, h < h_1) \le \\ & \le \sum_{h \in \mathcal{H},\, h \le h_1} \sum_{\eta \in \mathcal{H},\, \eta < h} P(|\tilde{\tilde{f}}_\eta(t) - \tilde{\tilde{f}}_h(t)| > \psi(\eta, h) + 2\psi(h)). \end{aligned}$$

Now, decomposition (5.9) and properties (5.28) and (5.23) imply

$$\begin{aligned} & P\left(|\tilde{\tilde{f}}_\eta(t) - \tilde{\tilde{f}}_h(t)| > \psi(\eta, h) + 2\psi(h)\right) \le \\ & \le P\left(|\xi_\eta(t) - \xi_h(t)| + |f(t) - f_\eta(t)| + |f(t) - f_h(t)| > \psi(\eta, h) + 2\psi(h)\right) \le \\ & \le P\left(|\xi_\eta(t) - \xi_h(t)| > \psi(\eta, h)\right) \\ & \le (\eta/h^*)^2. \end{aligned}$$

We end up by the same arguments as in the proof of (5.24). □

Using this lemma we get

$$\sum_{h \in \mathcal{H}} h^*/h \int_0^1 P(\widehat{h}(t) \le h, h < h_f(t))dt \le \sum_{h \in \mathcal{H}} 2h/h^* \le 4.$$

In view of (5.22) and (5.30), statement (5.26) can now be reduced to

$$\sup_{f \in B^s_{p,q}(M)} R_f \le C''$$

where

$$R_f = \sum_{h\in\mathcal{H}} h^*/h \int_0^1 \mathbf{1}(h_f(t) \le h)dt.$$

Note that for each t

$$\sum_{h\in\mathcal{H},\, h\ge h_f(t)} h^*/h \le 2h^*/h_f(t),$$

so that we obtain

$$R_f = \int_0^1 \left(\sum_{h\in\mathcal{H},\, h\ge h_f(t)} h^*/h \right) dt \le 2\int_0^1 (h^*/h_f(t))dt.$$

By definition (3.4) we have $h^*/h \le \psi^2(h)/\psi^2(h^*)$ and it suffices to prove that

$$\sup_{f\in B^s_{p,q}(M)} R'_f \le C'''$$

with

$$R'_f = \int_0^1 \left| \frac{\psi(h_f(t))}{\psi(h^*)} \right|^2 dt.$$

By (5.29)

$$\begin{aligned} R'_f &\le 1 + \sum_{h\in\mathcal{H},\, h<h^*} \int_0^1 \left| \frac{\psi(h)}{\psi(h^*)} \right|^2 \mathbf{1}(h_f(t) = h)dt \le \\ &\le 1 + \sum_{h\in\mathcal{H},\, h<h^*} \psi^{-2}(h^*)|\psi(h)|^{2-p} \int_0^1 3^{p/2}\Delta^p_f(2h,t)\mathbf{1}(h_f(t) = h)dt \le \\ &\le 1 + 3^{p/2}\psi^{-2}(h^*) \sum_{h\in\mathcal{H},\, h<h^*} |\psi(h)|^{2-p} \int_0^1 \Delta^p_f(2h,t)dt. \end{aligned}$$

The properties of the Besov class $B^s_{p,q}(M)$ imply the following bound, see Lepski et al.(1994, formula (5.9)),

$$\sup_{f\in B^s_{p,q}(M)} \int_0^1 \Delta^p_f(h,t)dt \le Lh^{sp}$$

with some constant $L = L(s,p,q,M)$. This gives

$$\begin{aligned} R'_f &\le 1 + 3^{p/2}\psi^{-2}(h^*) \sum_{h\in\mathcal{H},\, h<h^*} |\psi(h)|^{2-p}\, L(2h)^{sp} \le \\ &\le 1 + 3^{p/2}2^{sp}L\varepsilon^2/h^* \sum_{h\in\mathcal{H},\, h<h^*} \varepsilon^{2-p}h^{sp-1+p/2}\ln(h^*/h). \end{aligned}$$

Since $sp - 1 + p/2 > 0$ for $sp > 1/2$, the latter expression is estimated as follows:

$$R'_f \le 1 + const.\, \varepsilon^{-p}{h^*}^{-1}{h^*}^{sp-1+p/2};$$

by substituting h^* from (3.2) we get

$$R'_f \le 1 + const.\, \varepsilon^{1/(2s+1-1/p)} = 1 + o_\varepsilon(1)$$

which completes the proof of (5.26) and hence of (5.15).

Now we verify (5.19) by means of the same method as the one applied above for estimating R_ε. Let $t \in [0,1]$ and let $h_f(t)$ be defined by (5.27). We consider separately the cases when $\widehat{h}(t) \le h_f$ and $\widehat{h}(t) > h_f(t)$.

For the sake of simplicity we write below $\widehat{h}$ and h_f instead of $\widehat{h}(t)$ and $h_f(t)$ respectively. Also set $h_+ = 2h_f = 2h_f(t)$. The definition of h_f yields

$$|f(t) - f_{\widehat{h}}(t)|^2 \mathbf{1}(\widehat{h} \le h_f) \le \Delta_f^2(h_f, t)\mathbf{1}(\widehat{h} \le h_f) \le |\psi(h_f)|^{2-p}\Delta_f^p(h_f, t). \tag{5.31}$$

Next, for the inverse case of $\widehat{h} > h_f$ we apply decomposition (5.9) and the definition of $\widehat{h}$ getting

$$\begin{aligned}
&|f(t) - f_{\widehat{h}}(t)|^2 \mathbf{1}(\widehat{h} > h_f) = \\
&= |f(t) - f_{h_f}(t) + \widetilde{f}_{h_f}(t) - \widetilde{f}_{\widehat{h}}(t) - (\xi_{h_f}(t) - \xi_{\widehat{h}}(t))|^2 \mathbf{1}(\widehat{h} > h_f) \le \\
&\le |\psi(h_f) + \psi(h_f, \widehat{h}) + \psi(\widehat{h}) + |\xi_{h_f}(t) - \xi_{\widehat{h}}(t)||^2 \mathbf{1}(\widehat{h} > h_f).
\end{aligned}$$

Now, for $\widehat{h} > h_f$,

$$\begin{aligned}
\psi(h_f, \widehat{h}) &\le \psi(h_f), \\
\psi(\widehat{h}) &\le \psi(h_f).
\end{aligned}$$

Since $\xi_h(\cdot)$ and $\widehat{h}(\cdot)$ are independent, we have

$$\widehat{E}\,|\xi_{h_f}(t) - \xi_{\widehat{h}}(t)|^2 = \sigma^2(h_f, \widehat{h}) \le \psi^2(h_f, \widehat{h}) \le \psi^2(h_f).$$

Hence

$$\begin{aligned}
|f(t) - f_{\widehat{h}}(t)|^2 \mathbf{1}(\widehat{h} > h_f) &\le E\left[3\psi(h_f) + |\xi_{h_f}(t) - \xi_{\widehat{h}}(t)|\right]^2 \le \\
&\le 18\psi^2(h_f) + 2E|\xi_{h_f}(t) - \xi_{\widehat{h}}(t)|^2 \le 20\psi^2(h_f).
\end{aligned}$$

The event $\{\widehat{h} > h_f\}$ implies that $h_f < h^*$ and, by (5.29),

$$\psi^2(h_f) \le 3\psi^2(h_+) \le 3|\psi(h_+)|^{2-p}|\Delta_f(h_+, t)|^p.$$

This inequality along with (5.31) allows us to conclude that

$$\begin{aligned}
E|f(t) - f_{\widehat{h}}(t)|^2 &\le const.\left[|\psi(h_f)|^{2-p}\Delta_f^p(h_f, t) + |\psi(h_+)|^{2-p}|\Delta_f(h_+, t)|^p\right] \le \\
&\le const. \sum_{h \in \mathcal{H}} |\psi(h)|^{2-p}\Delta_f^p(h, t)
\end{aligned}$$

and thus that

$$\begin{aligned}
E\,Q_\varepsilon &= \varrho_\varepsilon^{-2}\int_0^1 E|f(t) - f_{\widehat{h}}(t)|^2 dt \le \\
&\le const.\, \varrho_\varepsilon^{-2} \sum_{h \in \mathcal{H}} |\psi(h)|^{2-p} \int_0^1 \Delta_f^p(h, t) dt \le \\
&\le const.\, \varrho_\varepsilon^{-2} \sum_{h \in \mathcal{H}} |\psi(h)|^{2-p} h^{sp}.
\end{aligned}$$

Similarly to the above,

$$const.\, \varrho_\varepsilon^{-2} \sum_{h\in\mathcal{H}} |\psi(h_f)|^{2-p} h^{sp} \le const.\, \varrho_\varepsilon^{-2} |\psi(h^*)|^{2-p} h^{*sp} =$$
$$const. (\varepsilon^2/\sqrt{h^*})^{-1} (\varepsilon/\sqrt{h^*})^{2-p} h^{*sp} = const.$$

The last inequality obviously yields (5.19) and it remains to check (5.16). We proceed in the same way as we did when estimating R_ε. We have

$$\begin{aligned}\gamma_\varepsilon &= \sum_{h\in\mathcal{H}} \int_0^1 f_h(t)\xi_h(t)\mathbf{1}(\widehat{h}(t)=h)dt = \\ &= \varepsilon\sqrt{2}\|K\| \sum_{h\in\mathcal{H}} h^{-1/2} \int_0^1 f_h(t)\zeta_h(t)\mathbf{1}(\widehat{h}(t)=h)dt.\end{aligned}$$

Once more making use of Lemma 5.3 we obtain

$$\begin{aligned}\widehat{E}\,\gamma_\varepsilon^2 &= 2\varepsilon^2\|K\|^2 \sum_{h\in\mathcal{H}} \sum_{\eta\in\mathcal{H}} (h\eta)^{-1/2} \times \\ & \int_0^1\int_0^1 f_h(t)f_\eta(s)E[\zeta_h(t)\zeta_\eta(s)]\,\mathbf{1}(\widehat{h}(t)=h,\,\widehat{h}(s)=\eta)dt\,ds \le \\ &\le 4C_1(K)\varepsilon^2\|K\|^2 \sum_{h\in\mathcal{H}} \sum_{\eta\in\mathcal{H},\,\eta\le h} h^{-1} \times \\ & \int_0^1\int_0^1 f_h(t)f_\eta(s)\,\mathbf{1}(|t-s|\le 2bh)\,\mathbf{1}(\widehat{h}(t)=h,\,\widehat{h}(s)=\eta)dt\,ds. \qquad (5.32)\end{aligned}$$

The elementary inequality $ab \le (a^2+b^2)/2$ leads to

$$\begin{aligned}&\int_0^1\int_0^1 f_h(t)f_\eta(s)\,\mathbf{1}(|t-s|\le 2bh)\,\mathbf{1}(\widehat{h}(t)=h)\,\mathbf{1}(\widehat{h}(s)=\eta)dt\,ds \le \\ &\le \frac{1}{2}\int_0^1\int_0^1 f_h^2(t)\mathbf{1}(|t-s|\le 2bh)\,\mathbf{1}(\widehat{h}(t)=h)\,dt\,ds + \\ &\frac{1}{2}\int_0^1\int_0^1 f_\eta^2(s)\,\mathbf{1}(|t-s|\le 2bh)\,\mathbf{1}(\widehat{h}(s)=\eta)dt\,ds \le \\ &\le 2bh\int_0^1 f_h^2(t)\mathbf{1}(\widehat{h}(t)=h)\,dt + 2bh\int_0^1 f_\eta^2(s)\,\mathbf{1}(\widehat{h}(s)=\eta)ds.\end{aligned}$$

By (5.32) we arrive easily at

$$\begin{aligned}\widehat{E}\,\gamma_\varepsilon^2 &\le 4C_1(K)\|K\|^2\varepsilon^2 2r \times \\ & \sum_{h\in\mathcal{H}} \sum_{\eta\in\mathcal{H},\,\eta\le h} \left[\int_0^1 f_h^2(t)\mathbf{1}(\widehat{h}(t)=h)\,dt + \int_0^1 f_\eta^2(s)\,\mathbf{1}(\widehat{h}(s)=\eta)ds\right] \le \\ &\le 4C_1(K)\|K\|^2\varepsilon^2 2r\,\#\mathcal{H} \sum_{h\in\mathcal{H}} \int_0^1 f_h^2(t)\mathbf{1}(\widehat{h}(t)=h)\,dt \le \\ &\le const.\,\varepsilon^2\,\widehat{S}\,\#\mathcal{H}\end{aligned}$$

where $\#\mathcal{H}$ is the number of points in the grid $\mathcal{H}$. Clearly $\#\mathcal{H} \le 2\ln\varepsilon^{-1}$ and we get for any $\delta > 0$,

$$\begin{aligned} & P(\gamma_\varepsilon > \delta(\widehat{S} + \varrho_\varepsilon^2)) \le P(\gamma_\varepsilon > 2\delta\varrho_\varepsilon\sqrt{\widehat{S}}) \le \\ & \le \frac{\widehat{E}\,\gamma_\varepsilon^2}{4\delta^2\widehat{S}\varrho_\varepsilon^2} \le \frac{const.\,\varepsilon^2\ln\varepsilon^{-1}}{\delta^2\varrho_\varepsilon^2} \to 0, \qquad \varepsilon \to 0. \end{aligned}$$

The theorem is proved.

References

[1] 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.

[2] Brown, L.D. and Low, M.G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Annals of Stat.*, **24**, no.6.

[3] Delyon, B. and Juditski, A. (1994). Wavelet estimators, global error measures, revisited. *Technical Report*, IRISA, Rennes.

[4] Donoho,D.L. and Johnstone,I.M. (1992). Minimax estimation via wavelet shrinkage. *Technical Report 402*. Dep. of Statistics, Stanford University.

[5] Donoho,D.L. and Johnstone,I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

[6] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. Royal Statist. Soc.*, Ser.B, **57**, 301–369.

[7] Ermakov, M.S. (1990). Minimax detection of a signal in a white Gaussian noise. *Theory Probab. Appl.*, **35**, 667–679.

[8] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Stat.*, **21**, 1926–1947.

[9] Huber, P.J. (1956). A robust version of the probability ratio test. *Ann. Math. Stat.*, **36**, 1753–1766.

[10] Ibragimov,I.A. and Khasminskii,R.Z. (1977). One problem of statistical estimation in Gaussian white noise. *Soviet Math. Dokl.*, **236**, no.4, 1351–1354.

[11] Ibragimov, I.A. and Linnik, Yu.V. (1965). *Independent and stationary connected random variables.* Nauka, Moskva.

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

[13] Ingster, Yu.I. (1984a). Asymptotic minimax testing of nonparametric hypothesis on the distribution density of an independent sample. *Zapiski Nauchn. Seminar. LOMI*, **136**, 74–96 (In Russian).

[14] Ingster, Yu.I. (1984b). An asymptotic minimax test of nonparametric hypothesis about spectral density. *Theory Probab. Appl.*, **29**, 846–847.

[15] Ingster, Yu.I. (1986). Minimax testing of nonparametric hypothesis about a distribution density in L_p-metrics. *Theory Probab. Appl.*, **32**, 333–337.

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

[17] Kerkyacharian,G. and Picard,D. (1993). Density estimation by kernel and wavelet method, optimality in Besov space. *Stat. and Probab. Letters*, **18**, 327–336.

[18] Lepski, O.V., Mammen, E. and Spokoiny, V.G. (1994). Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. *Annals of Stat.*, to appear.

[19] Lepski, O.V. and Spokoiny, V.G. (1995). Optimal Pointwise Adaptive Methods in Nonparametric Estimation. *Annals of Stat.*, to appear.

[20] 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.

[21] Nemirovski, A. (1985). On nonparametric estimation of smooth regression function. *Sov. J. Comput. Syst. Sci.*, **23**, no. 6, 1–11.

[22] Neyman, J. (1937). “Smooth test” for goodness of fit. *Scand. Aktuarietidskr.*, **20**, 149–199.

[23] Nussbaum, M. (1996). Asymptotic equivalence of density estimation and white noise. *Annals of Stat.*, **24**, no.6., 2399–2430.

[24] Suslina, I.A. (1993). Minimax detection of a signal for l_q-ellipsoids with a removed l_p-ball. *Zapiski Nauchn. Seminar. SPOMI*, **207**, 127–137. (In Russian).

[25] Spokoiny, V. (1996). Adaptive hypothesis testing using wavelets. *Annals of Stat.*, **24**, no.6. 2477–2498.

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