Electronic Journal of Statistics
Vol. 11 (2017) 2321–2357
ISSN: 1935-7524
DOI: 10.1214/17-EJS1275

Adaptive variable selection in nonparametric sparse additive models

Cristina Butucea

LAMA(UMR 8050), UPEM, UPEC, CNRS, F-77454, Marne-la-Vallée, France
CREST, ENSAE, Université Paris-Saclay, 3, ave. P. Larousse 92245 Malakoff Cedex, France
e-mail: cristina.butucea@ensae.fr

and

Natalia Stepanova*

School of Mathematics and Statistics, Carleton University
1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada
e-mail: nstep@math.carleton.ca

Abstract: We consider the problem of recovery of an unknown multivariate signal f observed in a d-dimensional Gaussian white noise model of intensity ε. We assume that f belongs to a class of smooth functions in $L_2([0,1]^d)$ and has an additive sparse structure determined by the parameter s, the number of non-zero univariate components contributing to f. We are interested in the case when $d = d_\varepsilon \to \infty$ as $\varepsilon \to 0$ and the parameter s stays "small" relative to d. With these assumptions, the recovery problem in hand becomes that of determining which sparse additive components are non-zero.

Attempting to reconstruct most, but not all, non-zero components of f, we arrive at the problem of almost full variable selection in high-dimensional regression. For two different choices of a class of smooth functions, we establish conditions under which almost full variable selection is possible, and provide a procedure that achieves this goal. Our procedure is the best possible (in the asymptotically minimax sense) for selecting most non-zero components of f. Moreover, it is adaptive in the parameter s. In addition to that, we complement the findings of [17] by obtaining an adaptive exact selector for the class of infinitely-smooth functions. Our theoretical results are illustrated with numerical experiments.

MSC 2010 subject classifications: Primary 62G08; secondary 62G20.
Keywords and phrases: High-dimensional nonparametric regression, sparse additive signals, adaptive variable selection, exact and almost full selectors.

Received April 2016.

Contents

*Research supported by an NSERC grant.

1. Introduction

In recent years, there has been much work done on providing methods for variable selection in high-dimensional settings; refer, for example, to [4, 5, 11, 13, 24] and references therein. Among a variety of methods proposed, the lasso has become an important tool for dealing with sparse high-dimensional regression problems. Motivated by the fact that finding the lasso solutions is computationally demanding, Genovese et al. [11] studied the relative statistical performance of the lasso and marginal regression, which is also known as simple thresholding [7] and sure screening [9], for sparse high-dimensional regression problems.

Marginal regression is a simple method for variable selection using componentwise regression. It shrinks the full model down to a submodel of a smaller dimension by thresholding the marginal regression coefficients using a tuning parameter. Namely, let

$$Y = Z\gamma + z,$$

be a sparse linear regression model, where $Y = (Y_1, \dots, Y_n)^\top$ is a vector of responses, Z is an $n \times d$ design matrix with $d \gg n$, $\gamma = (\gamma_1, \dots, \gamma_d)^\top$ is a vector of coefficients with many components γ_j equal to zero, and $z = (z_1, \dots, z_n)^\top$ is a vector of noisy variables. The problem of variable selection in this context consists of determining which components of γ are non-zero. Under certain (more or less standard) conditions on the model, including the normality of z, and assuming that Z has been standardized, Genovese et al. [11] proposed the variable selection procedure $\hat{\gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_d)^\top$ where

$$\hat{\gamma}_j = \hat{\alpha}_j \mathbb{I}\{|\hat{\alpha}_j| \geq t\}, \quad j = 1, \dots, d,$$

and $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_d)^\top = Z^\top Y$. By setting the tuning parameter at level $t = C\sqrt{\log d}$, with a carefully chosen constant C depending on the number of non-zero components of γ, Genovese et al. [11] found, theoretically and numerically, that marginal regression provides a good alternative to the lasso and concluded that their procedure merited further study. Numerical results on the relative performance of marginal regression and the lasso for variable selection can be found in Genovese et al. [11] and Ji and Jin [21].

In this work, we are interested in obtaining asymptotically minimax exact and almost full recovery results for a continuous-time regression model. Our setting is that of a sparse additive Gaussian white noise model

$$X_\varepsilon = f + \varepsilon W, \tag{1}$$

where W is a d-dimensional Gaussian white noise on $[0,1]^d$, $\varepsilon > 0$ is the noise parameter, and $f \in \mathcal{F}^d \subset L_2([0,1]^d)$ is an unknown function with a certain additive structure that has to be recovered.

For a number $s \in \{1, \ldots, d\}$, which is referred to as the *sparsity parameter*, define the set

$$\mathcal{H}_{d,s} = \{\eta = (\eta_1, \ldots, \eta_d) : \eta_j \in \{0,1\}, 1 \le j \le d, \sum_{j=1}^d \eta_j = s\},$$

and assume that the regression function, or signal, f in model (1) has the form

$$f(\mathbf{x}) = \sum_{j=1}^d \eta_j f_j(x_j), \quad \mathbf{x} = (x_1, \ldots, x_d) \in [0,1]^d, \quad \eta = (\eta_1, \ldots, \eta_d) \in \mathcal{H}_{d,s},$$

where, for all $j = 1, \ldots, d$, $\int_0^1 f_j(x)\,dx = 0$ and f_j belongs to space $\mathcal{F}_\sigma \subset L_2[0,1]$ of smooth functions with a known smoothness parameter $\sigma > 0$. In the present paper, two examples of the space $\mathcal{F}_\sigma$ will be considered: (i) the *Sobolev class* of σ-smooth functions and (ii) the *class of analytic functions* containing periodic functions that can be continued analytically in some strip on the complex plane of width 2σ.

Thus, the class of s-sparse multivariate signals of interest is

$$\mathcal{F}_{s,\sigma}^d = \left\{ f : f(\mathbf{x}) = \sum_{j=1}^d \eta_j f_j(x_j),\ f_j \in \mathcal{F}_\sigma, \right. \\ \left. \int_0^1 f_j(x)\,dx = 0,\ 1 \le j \le d,\ \eta = (\eta_j) \in \mathcal{H}_{d,s} \right\},$$

where the components satisfy the side condition that guarantees uniqueness, and the signal recovery problem becomes that of determining which sparse additive components are non-zero.

We are interested in the regime where both the sparsity parameter s and the dimension d grow to infinity as $\varepsilon \to 0$ in such a way that $s/d \to 0$. This

assumption includes, as a particular case, a popular choice of s in the form $s = d^{1-\beta}$ for some $\beta \in (0,1)$. In general, the smaller the value of s is, the harder the problem of identifying the non-zero components of f is.

Let us define a *selector* as any measurable function $\eta^* = \eta^*(X_\varepsilon)$ taking values on $\{0,1\}^d$. Following [11] and [17], we judge the quality of a selector $\eta^* = (\eta_1^*, \dots, \eta_d^*)$ of a vector $\eta = (\eta_1, \dots, \eta_d) \in \mathcal{H}_{d,s}$ by using the *Hamming distance* on $\{0,1\}^d$, which counts the number of positions at which η^* and η differ:

$$|\eta^* - \eta| = \sum_{j=1}^{d} |\eta_j^* - \eta_j|.$$

As in [11], we distinguish between exact and almost full recovery, and define the risk $R_{f,\eta}(\eta^*)$ of a selector η^* to be $\mathbf{E}_{f,\eta}|\eta^* - \eta|$ and $s^{-1}\mathbf{E}_{f,\eta}|\eta^* - \eta|$, respectively.

The goal of this paper is three-fold. First, we establish *sharp selection boundaries* that allow us to separate detectable components of a signal $f \in \mathcal{F}_{s,\sigma}^d$ from non-detectable ones. Next, assuming that all active (non-zero) components f_j are detectable and that s belongs to some set $\mathcal{S}_d$, which puts mild restrictions on the range of s, we construct an adaptive (free of s) selector $\eta^* = \eta^*(X_\varepsilon)$ with the property

$$\sup_{s \in \mathcal{S}_d} \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{f \in \mathcal{F}_{s,\sigma}^d} R_{f,\eta}(\eta^*) \to 0, \qquad \text{as } \varepsilon \to 0. \tag{2}$$

Finally, we show that if at least one of the f_j's is undetectable, then

$$\liminf_{\varepsilon \to 0} \inf_{\tilde{\eta}} \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{f \in \mathcal{F}_{s,\sigma}^d} R_{f,\eta}(\eta^*) > 0, \tag{3}$$

that is, exact or almost full recovery is impossible, according to the risk function used.

Depending on whether $R_{f,\eta}(\eta^*) = \mathbf{E}_{f,\eta}|\eta^* - \eta|$ or $R_{f,\eta}(\eta^*) = s^{-1}\mathbf{E}_{f,\eta}|\eta^* - \eta|$, the selector η^* that satisfies (2) is said to provide asymptotically *exact recovery* or *almost full recovery* of a signal $f \in \mathcal{F}_{s,\sigma}^d$ in the continuous-time regression model (1). If, in addition, inequality (3) holds true, then the respective selection procedure based on η^* is called an *asymptotically minimax exact selector* or *asymptotically minimax almost full selector*, according to the risk function considered. The notion of optimality that we use here is taken from the minimax hypothesis testing theory.

It is well known that the continuous-time regression model (1) serves as a good approximation to a more realistic equidistant sampling scheme with discrete Gaussian white noise. In such an approximation, ε^{-2} roughly corresponds to the number n of observations per unit cube. In case of a sparse additive regression function f, the continuous model in hand is closely related to a sparse additive model (SpAM), as studied in Ravikumar et al. [25]:

$$y_i = \sum_{j=1}^{d} \eta_j f_j(x_{ij}) + \varepsilon_i, \quad i = 1, \dots, n,$$

in which y_i is the response for the ith observation unit and x_{ij} is the value of the jth explanatory variable for the ith observational unit. When d is very large but it is assumed that only a small subset of the explanatory variables truly contributes to the model, the objective is to identify those variables as precise as possible. The SpAM method of Ravikumar et al. [25] aims at estimating a function by using the so-called general additive models (GAMs).

Note that the selection of methods for variable selection in GAMs is rather broad. Meier et al. [23] proposed a penalized least-squares estimator for variable selection and estimation in GAMs and provided conditions under which, with probability approaching 1, their procedure selects a set of f_js containing the additive components whose distance from zero in a certain metric is greater than a specified threshold. However, they do not establish model-selection consistency of their procedure. In the approach of Ravikumar et al. [25], the penalty is imposed on the l_2-norm of the nonparametric components, as well as the mean value of the components to ensure identifiability. To be valid, however, their theoretical results require the condition that the "design matrix" formed from the basis functions for non-zero components be bounded away from zero and infinity, and it is not clear whether this condition holds in general. Under less complicated conditions, the adaptive group lasso applied in Huang et al. [13] for selecting non-zero components in GAMs was shown to be consistent. More recent generalized additive model selection (GAMSEL) in Chouldechova and Hastie [4], a penalized likelihood procedure for fitting sparse GAMs that scales to high-dimensional data, was numerically compared to several similar procedures from the literature, including the SpAM method, and was numerically found to perform very well, especially when some of the underlying effects are linear. For a brief overview of the topic of variable selection for GAMs, we refer to Huang et al. [13], Chouldechova and Hastie [4], and references therein.

In order to better place the present work in the current literature, let us go back and consider a simpler model

$$X_j = \eta_j m_j + \xi_j, \quad j = 1, \dots, d, \tag{4}$$

where $\xi_1, \dots, \xi_d$ are iid normal $\mathcal{N}(0,1)$ random variables, $\eta = (\eta_1, \dots, \eta_d)$ belongs to $\mathcal{H}_{d,s}$ and $m = (m_1, \dots, m_d)$ is an unknown vector of parameters with $m_j \geq a$ for some $a = a_d > 0$.

In terms of this model, the results in Section 4 of Genovese et al. [11] obtained for a Bayesian marginal regression setup and the stronger necessary and sufficient conditions for the possibility of successful variable selection established in Butucea et al. [3] can be stated as follows. For some $\delta > 0$ possessing the property $\delta \to 0$ and $\delta \log d \to \infty$ as $d \to \infty$, let $\tau^* = (\tau_j^*)$ be the selector with the components $\tau_j^* = \mathbb{I}\left(X_j > \sqrt{(2+\delta)\log d}\right)$ for $j = 1, \dots, d$. If the parameter $a = a_d$ is such that

$$\liminf_{d\to\infty} \frac{a}{\sqrt{2\log d} + \sqrt{2\log s}} > 1, \tag{5}$$

then τ^* provides exact variable selection. Also, if

$$\limsup_{d\to\infty} \frac{a}{\sqrt{2\log d}+\sqrt{2\log s}} < 1, \tag{6}$$

then exact variable selection is impossible, and thus τ^* is the *asymptotically minimax exact selector*.

For the same $\delta > 0$ as above, let $\hat\tau = (\hat\tau_j)$ be the selector with the components $\hat\tau_j = \mathbb{I}\left(X_j > \sqrt{2\log(d/s) + \delta\log d}\right)$ for $j = 1, \dots, d$. If

$$\liminf_{d\to\infty} \frac{a}{\sqrt{2\log(d/s)}} > 1, \tag{7}$$

then almost full variable selection is achieved by means of $\hat\tau$. At the same time, if

$$\limsup_{d\to\infty} \frac{a}{\sqrt{2\log(d/s)}} < 1, \tag{8}$$

then almost full selection is impossible, and hence $\hat\tau$ is the *asymptotically minimax almost full selector*. Relations (5)–(8) yield a partition of the set of possible values of (s, a) into three regions where (i) exact variable selection is possible (asymptotically), (ii) almost full variable selection is possible (asymptotically), and (iii) successful variable selection is impossible. Such a "phase diagram", in somewhat different terms, has been first obtained in Genovese et al. [11]. Later on, the same phase diagram appeared in Ji and Jin [21] in the regression settings that are somewhat different but, overall, less general than the ones considered in this work.

Back to the continuous-time regression model (1), the first question of interest is how to measure the signal strength, that is, what quantity should be considered in place of the value a in the Gaussian vector model (4), and also what statistic should be used instead of the observation X_j in the definition of the selectors τ_j^* and $\hat\tau_j$? A natural candidate for the signal strength is the total energy, that is, the L_2-norm of a signal.

A close correspondence between sharp detection boundaries in the nonparametric problem of signal detection and sharp selection boundaries in the problem of variable selection in model (1) has been for the first time established in Ingster and Stepanova [17]. Handling the problem of exact recovery for Sobolev classes $\mathcal{F}_\sigma$ of smoothness $\sigma > 0$, they constructed an asymptotically exact selector that compares a certain chi-square type statistic with a threshold $\sqrt{(2+\delta)\log d}$, the same as in τ_j^*. The sharp selection boundaries turned out to be strongly connected to the respective sharp detection boundaries for smooth nonparametric alternatives with an L_2-ball removed. Indeed, only those components f_j of a signal f in $\mathcal{F}_{s,d}^d$ that are detectable are also significant in the problem of variable selection.

Ingster and Stepanova [17] obtained an adaptive procedure that provides asymptotically minimax exact reconstruction of a σ-smooth signal $f \in \mathcal{F}_{s,\sigma}^d$

observed in a d-dimensional Gaussian white noise model. Their adaptive procedure is based on the idea of aggregation of tests, which is often used in adaptive testing.

A similar result for the space of infinitely-smooth functions is stated in this paper in Section 3 (see Theorems 1 and 2). Although the selector in Section 3 is based on somewhat different statistics as compared to the one in [17], both selectors have one common feature that their thresholds do not depend the sparsity parameter s. Therefore the aggregation of selectors does provide an adaptive method for exact variable selection.

Almost full selectors that are introduced in Section 4 are also based on certain chi-square type statistics. These selectors make a decision on whether or not the jth component of f is active by comparing the respective statistic to the threshold $\sqrt{2\log(d/s)+\delta\log d}$, the same as for the selector $\hat{\tau}_j$ in the Gaussian vector model (4). In this paper, we first consider the case of known parameter s, and then modify the obtained procedure for the case of unknown s. The asymptotically minimax almost full selector proposed in Section 4.1 has both the statistic and the threshold defining the selector dependent on s, and hence it does not solve the more intricate problem of almost full recovery for unknown s. A natural step to take next would be to aggregate the selectors. Unfortunately, in the case of high sparsity with very few active signals f_j, this approach fails to give an optimal (asymptotically minimax) procedure. Therefore, in Section 4.2 we propose a new selection procedure that is similar to the Lepski method of adaptive estimation of smooth signals, as introduced in Lepski [22]. Our procedure provides asymptotically minimax almost full selectors for both spaces of smooth functions under consideration.

The paper is organized as follows. In Section 2, we first translate the problem to an equivalent problem in terms of the Fourier coefficients, then describe two function spaces of our interest, and also present briefly the results of nonparametric hypothesis testing theory that are required to construct our selectors. In Section 3, in order to complete the picture of exact variable selection, we introduce an adaptive selection procedure that gives exact reconstruction for the space of analytic functions. In Section 4, we obtain almost full selectors for a known sparsity parameter s for both function spaces in hand. Then, we construct a more involved adaptive almost full selector and state our main results. The main results, as stated in Sections 3 and 4, are proved in Section 7. In Section 5, a good (optimal) performance of the newly proposed selectors is illustrated numerically. Section 6 contains concluding remarks and outlines possible extensions of the present study.

2. Building blocks

Before stating and proving our main results, we translate the statement of the problem to that for the equivalent model of sparse additive Gaussian sequences, and define the smoothness classes $\mathcal{F}_\sigma$, $\sigma > 0$, that we consider here. Then, we shall briefly present some important tools of minimax hypothesis testing

that will be used in the subsequence sections. For a complete exposition of the subject, see [18] and the review papers [14, 15, 16].

2.1. Sparse additive Gaussian sequence model

A Gaussian sequence model is equivalent to the corresponding Gaussian white noise model but is more convenient to deal with as it is written in terms of the Fourier coefficients. In what follows, $\{\phi_k(x)\}_{k\in\mathbb{Z}}$ is the orthonormal basis of $L_2[0,1]$ given by

$$\phi_0(x)=1,\quad \phi_k(x)=\sqrt{2}\cos(2\pi kx),\quad \phi_{-k}(x)=\sqrt{2}\sin(2\pi kx),\quad k>0. \tag{9}$$

For the index $l\in\mathbb{Z}^d$ whose jth component is equal to k and the other components are equal to zero, define the functions

$$\phi_{j,k}(\mathbf{x})=\phi_l(\mathbf{x})=\phi_k(x_j),\quad \mathbf{x}=(x_1,\dots,x_d)\in[0,1]^d,\quad 1\le j\le d,\quad k\in\mathbb{Z},$$

and denote by $\theta_{j,k}=(f,\phi_{j,k})=\int_0^1\phi_k(x)f_j(x)\,dx$ the kth Fourier coefficient of the jth component f_j. Then, the sequence space model that corresponds to model (1) takes the form

$$X_{j,k}=\eta_j\theta_{j,k}+\varepsilon\xi_{j,k},\quad \xi_{j,k}\overset{\text{iid}}{\sim}\mathcal{N}(0,1),\quad 1\le j\le d,\quad k\in\mathbb{Z}, \tag{10}$$

where $X_{j,k}=X_\varepsilon(\phi_{j,k})$ are the empirical Fourier coefficients and the collection $(\eta_1\theta_1,\dots,\eta_d\theta_d)$ consists of sequences $\eta_j\theta_j=\{\eta_j\theta_{j,k}\}_{k\in\mathbb{Z}}$ such that the index variables η_j sum up to s and that $\{\theta_{j,k}\}_{k\in\mathbb{Z}}$ belong to $l_2(\mathbb{Z})$.

From the theory of communication viewpoint, representation (10) corresponds to the transmission of a message f_j, $1\le j\le d$, by converting it into an infinite series of coefficients $\theta_{j,k}$ which are translated by linearly modulated signals with a Gaussian white noise. In this paper we have chosen to deal with model (10), which is technically more convenient. Although the set of θ_js involves an orthogonal system in $L_2([0,1]^d)$, the results on minimax errors and risks do not depend on the choice of this orthogonal system because the random variables $X_{j,k}$, which generate a sufficient σ-algebra for $f\in\mathcal{F}^d_{s,\sigma}$, are independent normal $\mathcal{N}(\eta_j\theta_{j,k},\varepsilon^2)$. Thus the distribution of $\{X_{j,k}\}$ depends on the Fourier coefficients $\theta_{j,k}$ of f with respect to the system $\{\phi_{j,k}\}$ but not on the choice of $\{\phi_{j,k}\}$. Using a suitable finite collection of the random variables $X_{j,k}$ as defined in (10), we wish to construct asymptotically minimax selection procedures that are adaptive in s. The main results presented in Sections 3 and 4 are formulated in terms of the observations $X_{j,k}$ as given in (10).

2.2. Two examples of the function space $\mathcal{F}_\sigma$

The following two examples of a smooth function space $\mathcal{F}_\sigma$ are very common in the literature on nonparametric estimation and hypothesis testing; we use these examples of $\mathcal{F}_\sigma$ to handle the problem of variable selection in nonparametric sparse regression. As before, $\{\phi_k(x)\}_{k\in\mathbb{Z}}$ is an orthonormal basis in $L_2[0,1]$ given by (9).

Example 1. Let $\mathcal{F}_\sigma$ with $\sigma > 0$ denote the Sobolev class of σ-smooth 1-periodic functions on $\mathbb{R}$. Define the norm $\|\cdot\|_\sigma$ on $\mathcal{F}_\sigma$ by the formula

$$\|f\|_\sigma^2 = \sum_{k\in\mathbb{Z}} \theta_k^2 c_k^2, \quad c_k^2 = c_k^2(\sigma) = (2\pi|k|)^{2\sigma}, \tag{11}$$

where θ_k is the kth Fourier coefficient of f with respect to $\{\phi_k(x)\}_{k\in\mathbb{Z}}$. If σ is an integer, then under the periodic constraints (when the function admits 1-periodic $[\sigma]$-smooth extension on the real line) the norm as in (11) corresponds to

$$\|f\|_\sigma^2 = \int_0^1 \left(f^{(\sigma)}(x)\right)^2 dx.$$

Here, $\mathcal{F}_\sigma$ is the space of 1-periodic functions f in $L_2[0,1]$ having $\|f\|_\sigma < \infty$.

The next example of $\mathcal{F}_\sigma$ is also well known in the context of minimax estimation and hypothesis testing.

Example 2. Let $\mathcal{F}_\sigma$ with $\sigma > 0$ be the class of analytic 1-periodic functions f on $\mathbb{R}$ admitting a continuation to the strip $S_\sigma = \{z = x + iy : |y| \le \sigma\} \subset \mathbb{C}$ such that $f(x+iy)$ is analytic on the interior of S_σ, bounded on S_σ and

$$\int_0^1 |f(x \pm i\sigma)|^2\, dx < \infty.$$

Let the norm $\|\cdot\|_{1,\sigma}$ on $\mathcal{F}_\sigma$ be given by (see, for example, [12])

$$\|f\|_{1,\sigma}^2 = \int_0^1 (\mathrm{Re} f(x+i\sigma))^2\, dx.$$

In terms of the Fourier coefficients, the squared norm $\|f\|_{1,\sigma}^2$ takes the form

$$\|f\|_{1,\sigma}^2 = \sum_{k\in\mathbb{Z}} \theta_k^2 c_k^2, \quad c_k^2 = c_k^2(\sigma) = \cosh^2(2\pi\sigma k).$$

In view of the relations

$$\exp(|x|) \le 2\cosh(x) \le 2\exp(|x|), \quad x \in \mathbb{R},$$

we may also consider an equivalent norm $\|\cdot\|_\sigma$ defined as

$$\|f\|_\sigma^2 = \sum_{k\in\mathbb{Z}} \theta_k^2 c_k^2, \quad c_k = c_k(\sigma) = \exp(2\pi\sigma|k|). \tag{12}$$

We have chosen to deal with the norm given by (12) rather than with the norm $\|f\|_{1,\sigma}$ as it is easier to study. Here, $\mathcal{F}_\sigma$ is the space of 1-periodic functions f in $L_2[0,1]$ having $\|f\|_\sigma < \infty$.

2.3. Minimax hypothesis testing for nonparametric alternatives

As demonstrated in [17], a natural condition that accounts for the signal strength is obtained by connecting the problem in hand to that of hypothesis testing. This is done as follows.

For each component f_j of a signal $f \in \mathcal{F}^d_{s,\sigma}$, consider testing the hypothesis of no signal $H_{0j} : f_j = 0$ versus the alternative $H_{1j} : f_j \in \mathcal{F}_\sigma(r_\varepsilon)$, where for a positive family $r_\varepsilon \to 0$

$$\mathcal{F}_\sigma(r_\varepsilon) = \{g \in \mathcal{F}_\sigma : \|g\|_\sigma \le 1,\ \|g\|_2 \ge r_\varepsilon\}, \tag{13}$$

and $\|\cdot\|_\sigma$ is a norm on $\mathcal{F}_\sigma$. In this problem, a precise demarcation between the signals that can be detected with error probabilities tending to 0 and the signals that cannot be detected is given in terms of a *detection boundary*, or *separation rate*, $r^*_\varepsilon \to 0$ as $\varepsilon \to 0$. For various function classes that are frequently used in minimax hypothesis testing, sharp asymptotics for r^*_ε are available (see, for example, [14]). The hypotheses H_{0j} and H_{1j} *separate asymptotically* (that is, the minimax error probability tends to zero) if $r_\varepsilon/r^*_\varepsilon \to \infty$ as $\varepsilon \to 0$. The hypotheses H_{0j} and H_{1j} *merge asymptotically* (that is, the minimax error probability tends to one) if $r_\varepsilon/r^*_\varepsilon \to 0$ as $\varepsilon \to 0$.

When H_{0j} and H_{1j} separate asymptotically, we say that f_j is *detectable*. If the hypotheses H_{0j} and H_{1j} separate (merge) asymptotically when

$$\liminf_{\varepsilon\to 0} r_\varepsilon/r^*_\varepsilon > 1 \quad (\limsup_{\varepsilon\to 0} r_\varepsilon/r^*_\varepsilon < 1),$$

the detection boundary r^*_ε is said to be *sharp*. The knowledge of a sharp detection boundary r^*_ε allows us to have a meaningful problem of testing $H_{0j} : f_j = 0$ versus $H_{1j} : f_j \in \mathcal{F}_\sigma(r_\varepsilon)$ by choosing r_ε so that $\liminf_{\varepsilon\to 0} r_\varepsilon/r^*_\varepsilon > 1$. Otherwise, the function f_j will be too "small" to be noticeable.

The quantity that is crucial for establishing *sharp selection boundaries* turns out to be exactly the quantity that defines sharp detection boundaries (in the previous testing problem). Therefore, below we provide details on the extremal problem whose value yields the required quantity. For the two examples of ellipsoids under study below, more details on the solution of this problem will be presented in the next section.

Let $\{\phi_k(x)\}_{k\in\mathbb{Z}}$ be the orthonormal basis in $L_2[0,1]$ given by (9). If $g \in L_2[0,1]$, then $g(x) = \sum_{k\in\mathbb{Z}} \theta_k\phi_k(x)$, where $\theta_k = (g,\phi_k)$ is the kth Fourier coefficient of g, and $\|g\|_2^2 = \sum_{k\in\mathbb{Z}} \theta_k^2$. Let $\mathcal{F}_\sigma$ be a function space depending on a parameter $\sigma > 0$ that is a subset of $L_2[0,1]$. Suppose that $g \in \mathcal{F}_\sigma \subset L_2[0,1]$ is observed in a univariate Gaussian white noise of intensity ε, and we wish to test the null hypothesis $H_0 : g = 0$ versus the alternative (more precisely, a family of alternatives) $H_1 : g \in \mathcal{F}_\sigma(r_\varepsilon)$, where the set $\mathcal{F}_\sigma(r_\varepsilon)$ is given by (13). For the two function spaces of our interest, the norm of an element g is expressed as $\|g\|_\sigma^2 = \sum_{k\in\mathbb{Z}} c_k^2\theta_k^2$ with the coefficients $c_k^2 = c_k^2(\sigma)$ specified by formulas (11) and (12). In the sequence space of Fourier coefficients, the set $\mathcal{F}_\sigma(r_\varepsilon)$ corresponds to the ellipsoid in the space $l_2(\mathbb{Z})$ with semi-axes $c_k = c_k(\sigma)$ and a small

neighbourhood of the point $\theta = 0$ removed:

$$\Theta_\sigma(r_\varepsilon) = \left\{\theta = (\theta_k)_{k\in\mathbb{Z}} \in l_2(\mathbb{Z}) : \sum_{k\in\mathbb{Z}} c_k^2\theta_k^2 \le 1, \sum_{k\in\mathbb{Z}} \theta_k^2 \ge r_\varepsilon^2\right\}. \tag{14}$$

Consider the problem of minimizing $(2\varepsilon^4)^{-1}\sum_{k\in\mathbb{Z}}\theta_k^4$ over all $\theta \in \Theta_\sigma(r_\varepsilon)$. Denote by $\theta^*(r_\varepsilon) = (\theta_k^*(r_\varepsilon))_{k\in\mathbb{Z}}$ the solution to this extreme problem:

$$\frac{1}{2\varepsilon^4}\sum_{k\in\mathbb{Z}}(\theta_k^*(r_\varepsilon))^4 = \inf_{\theta\in\Theta_\sigma(r_\varepsilon)}\frac{1}{2\varepsilon^4}\sum_{k\in\mathbb{Z}}\theta_k^4, \tag{15}$$

and let $u_\varepsilon^2(r_\varepsilon) = u_\varepsilon^2(\Theta_\sigma(r_\varepsilon))$ be the value of the problem, that is,

$$u_\varepsilon^2(r_\varepsilon) = \frac{1}{2\varepsilon^4}\sum_{k\in\mathbb{Z}}(\theta_k^*(r_\varepsilon))^4. \tag{16}$$

The function $u_\varepsilon^2(r_\varepsilon)$ plays a key role in the minimax theory of hypothesis testing. It controls the minimax total error probability and is used to set a cut-off point of the asymptotically minimax test procedure. The detection boundary r_ε^* in the problem of testing $H_0 : \theta = 0$ versus $H_1 : \theta \in \Theta_\sigma(r_\varepsilon)$ is determined by the relation $u_\varepsilon(r_\varepsilon^*) \asymp 1$. The function $u_\varepsilon(r_\varepsilon)$ is a non-decreasing function of the argument r_ε which possesses a kind of 'continuity' property. Namely, for any $\epsilon > 0$ there exist $\Delta > 0$ and $\varepsilon_0 > 0$ such that for any $\delta \in (0, \Delta)$ and $\varepsilon \in (0, \varepsilon_0)$,

$$u_\varepsilon(r_\varepsilon) \le u_\varepsilon((1+\delta)r_\varepsilon) \le (1+\epsilon)u_\varepsilon(r_\varepsilon). \tag{17}$$

These and some other facts about $u_\varepsilon^2(r_\varepsilon)$ can be found in [14, Sec. 3.2] and [18, Sec. 5.2.3]. In the context of variable selection, the knowledge of $u_\varepsilon(r_\varepsilon)$ makes it possible to establish the necessary and sufficient conditions for the possibility of variable selection in the exact and almost full regimes (see Section 4 for details).

For some standard function spaces with the norm $\|g\|_\sigma$ defined (under the periodic constraints) in terms of the Fourier coefficients as $\|g\|_\sigma^2 = \sum_{k\in\mathbb{Z}}\theta_k^2c_k^2$, the form of the extremal sequence $(\theta_k^*(r_\varepsilon)_{k\in\mathbb{Z}}$ in problem (15) as well as the sharp asymptotics for $u_\varepsilon(r_\varepsilon)$ are available. For the standard spaces of smooth functions, $\theta_k^*(r_\varepsilon)$ vanishes when k exceeds K_ε, where K_ε grows to infinity and depends on the function space under study.

2.4. Sharp testing constants for two examples of the space $\mathcal{F}_\sigma$

Below we detail asymptotic equivalents of the solutions of the extremal problem (15) for the Sobolev space of periodic σ-smooth function on $\mathbb{R}$ (see Example 1 in Section 2.2) and the space of periodic functions on $\mathbb{R}$ that admit an analytic continuation to the strip of width 2σ around the real line (see Example 2 in Section 2.2). They are used to construct our selectors.

First, we consider the function space of Example 1. Let $\mathcal{F}_\sigma$ with $\sigma > 0$ be the Sobolev space of σ-smooth 1-periodic functions, as introduced in Section 2.2.

For a function $f \in \mathcal{F}_\sigma$ consider testing the hypothesis $H_0 : f = 0$ versus the alternative $H_1 : f \in \mathcal{F}_\sigma(r_\varepsilon)$, where for a positive family $r_\varepsilon \to 0$

$$\mathcal{F}_\sigma(r_\varepsilon) = \{f \in \mathcal{F}_\sigma : \|f\|_\sigma \le 1,\ \|f\|_2 \ge r_\varepsilon\}.$$

Switching from Sobolev balls $\{f \in \mathcal{F}_\sigma : \|f\|_\sigma \le 1\}$ to Sobolev ellipsoids $\{\theta \in l_2(\mathbb{Z}) : \sum_{k\in\mathbb{Z}} c_k^2\theta_k^2 \le 1\}$ leads to the problem of testing $H_0 : \theta = 0$ versus $H_1 : \theta \in \Theta_\sigma(r_\varepsilon)$. The test procedure that does the best in distinguishing between the two latter hypotheses is obtained by solving the extreme problem (15) with the semi-axes c_k defined as in (11); see Section 3 of [14] for details. The extremal sequence $(\theta_k^*(r_\varepsilon)_{k\in\mathbb{Z}}$ satisfies (see, for example, [14, § 3.2] and Theorem 2 in [19]):

$$(\theta_k^*(r_\varepsilon))^2 \sim \frac{\pi(1+2\sigma)}{2\sigma(1+4\sigma)^{1/(2\sigma)}} r_\varepsilon^{2+1/\sigma} \left(1 - (|k|/K_\varepsilon)^{2\sigma}\right)_+, \tag{18}$$

where the notation $a_\varepsilon \sim b_\varepsilon$ means $\lim_{\varepsilon\to 0} a_\varepsilon/b_\varepsilon = 1$, $x_+ = \max(x, 0)$, and

$$K_\varepsilon = \lfloor (4\sigma+1)^{1/(2\sigma)} (2\pi)^{-1} r_\varepsilon^{-1/\sigma} \rfloor. \tag{19}$$

The sharp asymptotics for $u_\varepsilon(r_\varepsilon)$ defined by formula (16) are of the form (see [18, § 4.3.2] and Theorems 2 and 4 in [19])

$$u_\varepsilon(r_\varepsilon) \sim C(\sigma) r_\varepsilon^{2+1/(2\sigma)} \varepsilon^{-2}, \quad \varepsilon \to 0, \tag{20}$$

where (see, for example, p. 707 of [8] and p. 104 of [14])

$$C^2(\sigma) = \pi(1+2\sigma)(1+4\sigma)^{-1-1/(2\sigma)}.$$

Similar results are available for the class $\mathcal{F}_\sigma$ of 1-periodic analytic functions, as introduced in Example 2 of Section 2.2. In this case, the ball $\{f \in \mathcal{F}_\sigma : \|f\|_\sigma \le 1\}$ corresponds to the ellipsoid $\{\theta \in l_2(\mathbb{Z}) : \sum_{k\in\mathbb{Z}} c_k^2\theta_k^2 \le 1\}$ with the semi-axes c_k defined as in (12). Thus, translating the problem of testing $H_0 : f = 0$ versus $H_1 : f \in \mathcal{F}_\sigma(r_\varepsilon)$ to the one in terms of Fourier coefficients brings us to testing $H_0 : \theta = 0$ versus $H_1 : \theta \in \Theta_\sigma(r_\varepsilon)$. The asymptotically minimax test procedure that distinguishes between these two hypotheses is obtained by solving the extreme problem (15) with the semi-axes c_k defined as in (12). The elements of the extremal sequence $(\theta_k^*(r_\varepsilon)_{k\in\mathbb{Z}}$ in problem (15) with the semi-axis c_k as above may be taken as constants (independent of k) satisfying as $\varepsilon \to 0$ (see, for example, p. 707 of [8] and p. 104 of [14])

$$\theta_k^*(r_\varepsilon) \sim \sqrt{\pi\sigma}\, r_\varepsilon \log^{-1/2}(r_\varepsilon^{-1}) \left(1 - \exp(4\pi(|k| - K))\right)_+, \tag{21}$$

where

$$K_\varepsilon = \lfloor (2\pi\sigma)^{-1} \log(r_\varepsilon^{-1}) \rfloor, \tag{22}$$

and for the function $u_\varepsilon(r_\varepsilon)$ defined in (16) we have

$$u_\varepsilon(r_\varepsilon) \sim \left(\frac{r_\varepsilon}{\varepsilon}\right)^2 \frac{(\pi\sigma/2)^{1/2}}{\log^{1/2}(r_\varepsilon^{-1})}. \tag{23}$$

Formulas (21)–(23), as well as formulas (18)–(20), will be employed to construct exact and almost full selectors for the two examples of $\mathcal{F}_\sigma$ under study.

3. Adaptive exact variable selection for analytic functions

The problem of adaptive reconstruction of sparse additive functions in the Gaussian white noise model was studied in the only case of Sobolev σ-smooth functions, see [17]. Before handling the problem of almost full variable selection in adaptive settings, we complement the findings in [17] by presenting an adaptive exact selector for the space of analytic functions. The strategy is similar to the one suggested in [17] for σ-smooth functions, but the parameters of our statistics and the condition on the growth of the dimension d are different.

Consider the sequence space model that corresponds to the Gaussian white noise model with f from the class of analytic functions $\mathcal{F}_\sigma$ as defined in Section 2.2. Let $1 < s_1 < s_1 < \ldots < s_M < d$ be the grid of points as in (34). For any $m = 1, \ldots, M$, let the parameter $r^*_{\varepsilon,m} > 0$ be determined by the equation

$$\frac{u_\varepsilon(r^*_{\varepsilon,m})}{\sqrt{2\log d} + \sqrt{2\log s_m}} = 1.$$

Consider weighted chi-square type statistics

$$t_{j,m} = \sum_{1 \le |k| \le K_\varepsilon} \omega_k(r^*_{\varepsilon,m}) \left[\left(\frac{X_{j,k}}{\varepsilon} \right)^2 - 1 \right], \quad j = 1, \ldots, d, \quad m = 1, \ldots, M,$$

with weight functions

$$\omega_k(r^*_{\varepsilon,m}) = \frac{1}{2\varepsilon^2} \frac{(\theta^*_k(r^*_{\varepsilon,m}))^2}{u_\varepsilon(r^*_{\varepsilon,m})}$$

obeying the normalization condition $\sum_{k\in\mathbb{Z}} \omega_k^2(r^*_{\varepsilon,m}) = 1/2$. Next, for all $j = 1, \ldots, d$ and $m = 1, \ldots, M$, set

$$\eta_{j,m} = \mathbb{I}\left(t_{j,m} > \sqrt{(2+\delta)(\log d + \log M)} \right),$$

and define an *adaptive exact selector* η^{**} of a vector $\eta \in \mathcal{H}_{d,s}$ by the formula (see formula (18) in [17])

$$\eta^{**} = (\eta^{**}_1, \ldots, \eta^{**}_d), \quad \eta^{**}_j = \max_{1 \le m \le M} \eta_{j,m}, \quad j = 1, \ldots, d. \tag{24}$$

The idea behind the selector η^{**} is as follows. The jth component of a signal is viewed active if at least one of the statistics $t_{j,m}$, $m = 1, \ldots, M$, detects it. Therefore, thinking of $\eta_{j,m}$ and η^{**}_j as test functions, we get that the probability of having θ_j incorrectly undetected does not exceed the respective probability with the $\eta_{j,m}$ test, where s_m is close to the true (but unknown) value of s. Furthermore, the probability that η^{**}_j incorrectly detects θ_j is less than the sum of the respective probabilities for the $\eta_{j,m}$ tests over all $m = 1, \ldots, M$, and is small by the choice of a threshold.

Let the set $\Theta_{\sigma,d}(r_\varepsilon)$ be as in (38) with the coefficients c_k given by (12). The following two theorems, whose proofs are similar to those of Theorems 3 and 4 in [17], hold true.

Theorem 1. *Let $s \in \{1, \dots, d\}$ be such that $s = o(d)$. Assume that $\log d = o(\log \varepsilon^{-1})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\liminf_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2 \log d} + \sqrt{2 \log s}} > 1. \tag{25}$$

Then as $\varepsilon \to 0$

$$\sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} \mathbf{E}_{\eta,\theta} |\eta - \eta^{**}| \to 0,$$

where η^{**} is the selector of vector η as defined in (24).

Theorem 2. *Let $s \in \{1, \dots, d\}$ be such that $s = o(d)$. Assume that $\log d = o(\log \varepsilon^{-1})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\limsup_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2 \log d} + \sqrt{2 \log s}} < 1. \tag{26}$$

Then

$$\liminf_{\varepsilon \to 0} \inf_{\tilde{\eta}} \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} \mathbf{E}_{\eta,\theta} |\eta - \tilde{\eta}| > 0,$$

where the infimum is over all selectors $\tilde{\eta}$ of a vector η in model (10).

Remark 1. Let us comment on the statements of Theorems 1 and 2. Inequalities (25) and (26) describe the sharp *exact selection boundary* that indicates whether or not we are in a position to proceed with exact variable selection; this boundary is determined in terms of the function $u_\varepsilon(r_\varepsilon)$ whose sharp asymptotics for the two examples of $\mathcal{F}_\sigma$ are given by (20) and (23). The use of $u_\varepsilon(r_\varepsilon)$ instead of r_ε makes it easier to build a bridge between variable selection in Gaussian white noise setting and variable selection in regression setting as studied in Section 4 of [11]. Indeed, the comparison of inequalities (25) and (26) with the respective inequalities (5) and (6) shows that $u_\varepsilon(r_\varepsilon)$ does the same job in the space of Fourier coefficients as the parameter a does in a Gaussian vector model: namely, it controls the strength of a signal. In addition, using $u_\varepsilon(r_\varepsilon)$ instead of r_ε makes the statement of detectability condition precise. By 'continuity' of $u_\varepsilon(r_\varepsilon)$ as cited in (17), the conditions (25) and (26) that separate detectable components from undetectable ones can be written in the usual form $\liminf_{\varepsilon \to 0} r_\varepsilon / r_\varepsilon^* > 1$ paired with $\limsup_{\varepsilon \to 0} r_\varepsilon / r_\varepsilon^* < 1$, where for Sobolev ellipsoids the sharp detection boundary r_ε^* is found explicitly from (20), and for the ellipsoids of analytic functions it is obtained from (23). A similar remark applies to Theorems 3–6 stated in Section 4.2. In this case, we can see the close correspondence between inequalities (39) and (40) that establish the sharp *almost full selection boundary* and inequalities (7) and (8). Thus, in the nonparametric case, inequalities (25)–(26) and (39)–(40) altogether partition the parameter space of the problem into the same three regions of (i) exact variable selection, (ii) almost full variable selection, and (iii) no variable selection, as inequalities (5)–(8) do in the parametric case.

4. Almost full variable selection

In this section, we first discuss almost full variable selection for *known* sparsity s. Next, when s is *unknown*, we introduce the almost full selector that is adaptive in s and prove that, with a suitable choice of the parameters, both adaptive and non-adaptive selectors are asymptotically minimax over the corresponding smoothness classes.

4.1. Non-adaptive almost full variable selection

We first consider a non-adaptive setup when the sparsity parameter s is known. When dealing with the problem of variable selection in model (10), we make use of the statistics, cf. asymptotically minimax test statistics in Section 3.1 of [14],

$$t_j = t_j(s) = \sum_{1\le|k|\le K_\varepsilon} \omega_k(r_\varepsilon^*(s)) \left[\left(\frac{X_{j,k}}{\varepsilon} \right)^2 - 1 \right], \quad j = 1, \dots, d, \tag{27}$$

where for any $r_\varepsilon > 0$ the weight functions $\omega_k(r_\varepsilon)$ are given by the formula

$$\omega_k(r_\varepsilon) = \frac{1}{2\varepsilon^2} \frac{(\theta_k^*(r_\varepsilon))^2}{u_\varepsilon(r_\varepsilon)}, \quad 1 \le |k| \le K_\varepsilon,$$

and the number $r_\varepsilon^*(s) > 0$ is the solution of the equation

$$\frac{u_\varepsilon(r_\varepsilon^*(s))}{\sqrt{2\log(d/s)}} = 1. \tag{28}$$

For both function spaces of interest, the quantities K_ε, $\theta_k^*(r_\varepsilon)$, and $u_\varepsilon(r_\varepsilon)$ in formula (27) are specified in Section 2.4. The sparsity parameter $s \in \{1, 2, \dots, d\}$ is assumed to be small relative to d, that is, $s = o(d)$. Note that the weights $\omega_k(r_\varepsilon)$ are normalized to have $\sum_{1\le|k|\le K_\varepsilon} \omega_k^2(r_\varepsilon) = 1/2$.

Now we define a non-adaptive *almost full selector* to be

$$\check\eta = (\check\eta_1, \dots, \check\eta_d), \quad \check\eta_j = \mathbb{I}\left(t_j > \sqrt{2\log(d/s) + \delta \log d} \right), \quad j = 1, \dots, d, \tag{29}$$

where $\delta = \delta_\varepsilon > 0$ satisfies

$$\delta \to 0 \quad \text{and} \quad \delta \log d \to \infty, \quad \text{as } \varepsilon \to 0. \tag{30}$$

The arguments as in the proof of Theorem 1 show that for Sobolev ellipsoids, under the conditions, cf. (32),

$$\log d = o(\varepsilon^{-2/(2\sigma+1)}), \quad \liminf_{\varepsilon\to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log(d/s)}} > 1,$$

the selector $\check\eta$ reconstructs almost all relevant components of a vector $\eta \in \mathcal{H}_{d,s}$, and hence asymptotically provides almost full recovery of a signal $f \in \mathcal{F}_{s,\sigma}^d$ in model (1).

To illustrate the difference between exact and almost full reconstruction in adaptive settings, assume that $\mathcal{F}_\sigma$ is the Sobolev space. In this case, a selector (see Section 3.1 of [17] with s in place of $d^{1-\beta}$)

$$\eta^* = (\eta_1^*, \dots, \eta_d^*), \quad \eta_j^* = \mathbb{I}\left(t_j^* > \sqrt{(2+\delta)\log d}\right), \quad j = 1, \dots, d, \tag{31}$$

where the statistics t_j^* are defined similar to the t_js as in (27) with the relation

$$\frac{u_\varepsilon(r_\varepsilon^*(s))}{\sqrt{2\log d} + \sqrt{2\log s}} = 1$$

instead of (28), turns out to be a non-adaptive *exact selector* as long as

$$\log d = o(\varepsilon^{-2/(2\sigma+1)}) \quad \text{and} \quad \liminf_{\varepsilon\to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log d} + \sqrt{2\log s}} > 1. \tag{32}$$

Under the above conditions, the procedure based on η^* selects correctly all non-zero components of a vector $\eta \in \mathcal{H}_{d,s}$, and hence provides exact recovery of a signal $f \in \mathcal{F}_{s,\sigma}^d$ in model (1).

Contrasting with formula (31), the threshold in (29) is set at a lower level and is dependent on the parameter s. The latter fact makes the idea of adaptation suggested in [17] for the exact recovery case not applicable in the case of almost full recovery.

4.2. Adaptive almost full variable selection

In this section, we consider a more realistic problem when the sparsity parameter s is *unknown*. We derive conditions under which almost full variable selection is possible, and construct a selector for which the Hamming distance is much smaller than the number of relevant components (see Theorems 3 and 5). Our selector is adaptive in the sparsity parameter s and is unimprovable in the asymptotically minimax sense (see Theorems 4 and 6). It is obtained by using Lepski's method.

4.2.1. Construction of the almost full selector in the adaptive case

In this subsection, the selector $\check{\eta}$ as in (29) will be used to obtain the corresponding adaptive procedure. To avoid losses due to adaptation, we will have to limit the range of the possible values of s. Namely, we assume that for some constants $0 < c < C < 1$

$$c \le \liminf_{d\to\infty}(\log s/\log d) \le \limsup_{d\to\infty}(\log s/\log d) \le C, \tag{33}$$

and define the set

$$\mathcal{S}_d = \{s \in \{1, \dots, d\} \text{ is such that condition (33) holds}\}$$

over which the adaptive selector that we propose yields almost full selection. The restriction on s as in (33) is relatively mild. For instance, any $s = d^{1-\beta}$ with $\beta \in [b, B]$ for some constants $0 < b < B < 1$ belongs to $\mathcal{S}_d$.

To construct the desired selector, for some $\Delta = \Delta_d > 0$ and $M = \lceil (C - c)/\Delta \rceil + 1$, pick grid points over the interval $(1, d)$:

$$s_1 = d^c, \quad s_m = s_{m-1} d^{\Delta} = s_1 d^{(m-1)\Delta}, \quad 2 \le m \le M, \tag{34}$$

and assume that

$$\Delta \to 0, \quad \Delta \log d \to 0, \quad \text{as } \ d \to \infty, \tag{35}$$

yielding $d^{\Delta} \le$ const for all large enough d. For each $m = 1, \dots, M$, let the parameter $r^*_\varepsilon(s_m) > 0$ be determined by the equation, cf. (28),

$$\frac{u_\varepsilon(r^*_\varepsilon(s_m))}{\sqrt{2\log(d/s_m)}} = 1,$$

where, depending on a type of the ellipsoid $\Theta_\sigma(r_\varepsilon)$ we are dealing with, the function $u_\varepsilon(r_\varepsilon)$ satisfies either (20) or (23).

Similar to the case of known s, consider weighted chi-square type statistics, cf. (27),

$$t_j(s_m) = \sum_{1 \le |k| \le K_\varepsilon} \omega_k(r^*_\varepsilon(s_m)) \left[\left(\frac{X_{j,k}}{\varepsilon} \right)^2 - 1 \right], \quad j = 1, \dots, d, \quad m = 1, \dots, M,$$

with weight functions

$$\omega_k(r^*_\varepsilon(s_m)) = \frac{1}{2\varepsilon^2} \frac{(\theta^*_k(r^*_\varepsilon(s_m)))^2}{u_\varepsilon(r^*_\varepsilon(s_m))}, \quad 1 \le |k| \le K_\varepsilon,$$

possessing the property $\sum_{1 \le |k| \le K_\varepsilon} \omega_k^2(r^*_\varepsilon(s_m)) = 1/2$. The values of $\theta^*_k(r^*_\varepsilon(s))$ and K_ε depend on the function space under consideration. For the Sobolev space in hand, $\theta^*_k(r^*_\varepsilon(s))$ and K_ε are as in (18) and (19); for the space of analytic functions, $\theta^*_k(r^*_\varepsilon(s))$ and K_ε are as in (21) and (22).

Next, for all $j = 1, \dots, d$ and $m = 1, \dots, M$, set

$$\widehat{\eta}_j(s_m) = \mathbb{I}\left(t_j(s_m) > \sqrt{2\log(d/s_m) + \delta \log d} \right),$$

where $\delta = \delta_\varepsilon > 0$ satisfies (30), and define an adaptive selector of a vector $\eta \in \mathcal{H}_{d,s}$ by the formula

$$\widehat{\eta}(s_{\widehat{m}}) = (\widehat{\eta}_1(s_{\widehat{m}}), \dots, \widehat{\eta}_d(s_{\widehat{m}})), \tag{36}$$

where $\widehat{m}$ is chosen by Lepski's method (see Section 2 of [22]) as follows:

$$\widehat{m} = \min \left\{ 1 \le m \le M \ : \ |\widehat{\eta}(s_m) - \widehat{\eta}(s_i)| \le v_i \ \text{ for all } \ i \ge m \right\}, \tag{37}$$

and $\widehat{m} = M$ if the set in (37) is empty. Here the quantities $v_i = v_{i,d}$ are set to be

$$v_i = s_i/\tau_d, \quad m \le i \le M,$$

with a sequence of numbers $\tau_d \to \infty$ satisfying (recall that $d = d_\varepsilon \to \infty$ and $\delta \log d \to \infty$ as $\varepsilon \to 0$)

$$\tau_d = o\left(\min(\log d, d^{\delta/2})\right), \quad \text{as } \varepsilon \to 0.$$

The idea behind Lepski's procedure is as follows. Each estimator $\widehat{\eta}(s_i)$ is associated with the interval of a size proportional to its standard deviation. When, in the process of iteration, for some index m an estimator $\widehat{\eta}(s_m)$ is found to be far enough from the previous ones, this means that we have detected a bias change, indicating that we are probably close to the true value of s.

Algorithmically, Lepski's procedure for choosing $\hat{m}$ works as follows. We start by setting $\hat{m} = M$ and attempt to decrease the value of $\hat{m}$ from M to $M-1$. If $|\widehat{\eta}(s_{M-1}) - \widehat{\eta}(s_M)| \le v_M$, we set $\hat{m} = M-1$; otherwise, we keep $\hat{m}$ equal to M. In case $\hat{m}$ is decreased to $M-1$, we continue the process attempting to decrease it further. If $|\widehat{\eta}(s_{M-2}) - \widehat{\eta}(s_{M-1})| \le v_{M-1}$ and $|\widehat{\eta}(s_{M-2}) - \widehat{\eta}(s_M)| \le v_M$, we set $\hat{m} = M-2$; otherwise, we keep $\hat{m}$ equal to $M-1$; and so on. Notice that by construction $v_M \ge v_{M-1} \ge \ldots \ge v_1$.

In connection with formula (37), let us note that v_i, $m \le i \le M$, are real numbers such that $v_i = o(s_i) = o(s_M)$ and

$$v_i = s_i/\tau_d \ge s_1 \max\left((\log d)^{-1}, d^{-\delta/2}\right) \ge d^{c-\delta/2} \to \infty$$

as $d \to \infty$. As the random variables $|\widehat{\eta}(s_m) - \widehat{\eta}(s_i)| = \sum_{j=1}^d |\widehat{\eta}_j(s_m) - \widehat{\eta}_j(s_i)|$, $m \le i \le M$, take on integer values, the use of the integer part $[v_i]$ instead of v_i would also be possible yet more complicated in terms of notation. As we shall see in the next theorems, our selector $\hat{\eta}(s_{\hat{m}})$ with $\hat{m}$ as in (37) achieves almost full selection adaptively.

4.2.2. Sobolev smooth functions

Consider the set $\Theta_\sigma(r_\varepsilon)$ as in (14) with the coefficients c_k given by (11), and define the set

$$\Theta_{\sigma,d}(r_\varepsilon) = \Big\{\theta = (\theta_j) : \theta_j = (\theta_{j,k}) \in l_2(\mathbb{Z}),\ \sum_{k\in\mathbb{Z}} c_k^2\theta_{j,k}^2 \le 1, \\ \sum_{k\in\mathbb{Z}} \theta_{j,k}^2 \ge r_\varepsilon^2,\ 1 \le j \le d\Big\}. \quad (38)$$

Let $\widehat{\eta}(s_{\hat{m}})$ be the selector given by (36) based on the statistics $t_j(s_m)$ as in (27), where the quantities $\theta_k^*(r_\varepsilon)$, K_ε, and $u_\varepsilon(r_\varepsilon)$ are specified by formulas (18), (19), and (20), respectively. The following theorem holds.

Theorem 3. *Let $s \in \{1,\dots,d\}$ be such that (33) holds true. Assume that $\log d = o(\varepsilon^{-2/(2\sigma+1)})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\liminf_{\varepsilon\to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log(d/s)}} > 1. \tag{39}$$

Then as $\varepsilon \to 0$

$$\sup_{s\in\mathcal{S}_d} \sup_{\eta\in\mathcal{H}_{d,s}} \sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\widehat{\eta}(s_{\widehat{m}}) - \eta| \to 0.$$

Theorem 3 says that if all the hypotheses $H_{0j} : \theta_j \equiv 0$ and $H_{1j} : \theta_j \in \Theta_\sigma(r_\varepsilon)$, $j = 1,\dots,d$, separate asymptotically, then the selection procedure based on $\widehat{\eta}(s_{\widehat{m}})$ reconstructs almost all non-zero components of a vector $\eta \in \mathcal{H}_{d,s}$, and thus provides almost full recovery of $(\eta_1\theta_1,\dots,\eta_d\theta_d)$, uniformly in $\mathcal{S}_d$, $\mathcal{H}_{d,s}$, and $\Theta_{\sigma,d}(r_\varepsilon)$.

The next result shows that if the detectability condition (39) is not met, almost full selection is impossible.

Theorem 4. *Let $s \in \{1,\dots,d\}$ be such that $s = o(d)$. Assume that $\log d = o(\varepsilon^{-2/(2\sigma+1)})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\limsup_{\varepsilon\to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log(d/s)}} < 1. \tag{40}$$

Then

$$\liminf_{\varepsilon\to 0} \inf_{\tilde{\eta}} \sup_{\eta\in\mathcal{H}_{d,s}} \sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\eta - \tilde{\eta}| > 0,$$

where the infimum is over all selectors $\tilde{\eta}$ of a vector η in model (10).

4.2.3. Analytic functions

The results similar to Theorems 3 and 4 hold true for the space of analytic functions. Namely, consider the sets $\Theta_\sigma(r_\varepsilon)$ and $\Theta_{\sigma,d}(r_\varepsilon)$ as in (14) and (38) with the coefficients c_k given by (12). Again, let $\widehat{\eta}(s_{\widehat{m}})$ be the selector defined by (36) based on the statistics $t_j(s_m)$ as in (27), but the quantities $\theta_k^*(r_\varepsilon)$, K_ε, and $u_\varepsilon(r_\varepsilon)$ are now as in (21), (22), and (23), respectively. The following results hold true.

Theorem 5. *Let $s \in \{1,\dots,d\}$ be such that (33) holds true. Assume that $\log d = o(\log \varepsilon^{-1})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\liminf_{\varepsilon\to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log(d/s)}} > 1.$$

Then as $\varepsilon \to 0$

$$\sup_{s\in\mathcal{S}_d} \sup_{\eta\in\mathcal{H}_{d,s}} \sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\widehat{\eta}(s_{\widehat{m}}) - \eta| \to 0.$$

Theorem 6. *Let $s \in \{1, \dots, d\}$ be such that $s = o(d)$. Assume that $\log d = o(\log \varepsilon^{-1})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\limsup_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2 \log(d/s)}} < 1.$$

Then

$$\liminf_{\varepsilon \to 0} \inf_{\tilde\eta} \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} s^{-1} \mathbf{E}_{\eta,\theta} |\eta - \tilde\eta| > 0,$$

where the infimum is over all selectors $\tilde\eta$ of a vector η in model (10).

Remark 2. We should remark that the best selection procedure yields exact variable selection only if the condition $\liminf_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2 \log d} + \sqrt{2 \log s}} > 1$ holds; at the same time, the best selection procedure gives almost full variable selection if a *milder* condition $\liminf_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2 \log(d/s)}} > 1$ is met.

5. Numerical implementation

In this section, we illustrate numerically the behaviour of the proposed almost full selector, as defined in (36), for Sobolev smooth functions with $\sigma = 1$. Note that this setup is among the least favorable setups, as the rates are faster for larger values of σ and for analytic functions.

Let us note that our thresholding procedure is invariant with respect to any permutation of the component functions. Therefore, without loss of generality, we choose the first s components to contain a signal and let the other components be empty.

Let $s = 5$ and let d belong to $\{50, 100, 500, 1000, 5000, 10000, 50000\}$, $\varepsilon = 0.01$, $\sigma = 1$. Then the ratio s/d belongs to $\{0.1, 0.05, 0.01, 5 \cdot 10^{-3}, 10^{-3}, 5 \cdot 10^{-4}, 10^{-4}\}$. The condition $\log d < \varepsilon^{-2/(2\sigma+1)}$ in Theorem 3 is verified.

We pick five active component functions, all defined on $[0, 1]$, as follows:

$$\begin{aligned}
f_1(x) &= x^2 \left(2^{x-1} - (x - 0.5)^2\right) e^x - 0.4752, \\
f_2(x) &= x^2 \left(2^{x-1} - (x - 1)^5\right) - 0.4494, \\
f_3(x) &= 15\, x^2\, 2^{x-1} \cos(15x) - 0.5068, \\
f_4(x) &= x - 0.5, \\
f_5(x) &= 5(x - 0.7)^3 + 0.29,
\end{aligned}$$

where, up to four decimal places, $\int_0^1 f_j(x)dx = 0$ for all $j = 1, \dots, 5$. These functions are plotted on Figure 1. Their L_2-norms (total energies) are evaluated numerically:

$$\|f_j\|_2,\ j = 1, \dots, 5: \quad 0.6110 \quad 0.2409 \quad 3.7886 \quad 0.2887 \quad 0.4601.$$

First, we implement the procedure $\widehat\eta(s)$ that utilizes the true value of s. Then, we run the procedure with the estimated sparsity $s_{\widehat m}$, where $s_{\widehat m}$ is one of the grid points in (34) and the index $\widehat m$ is chosen in accordance with (37):

$$s_{\widehat m} = s_1 = d^{0.15}, \quad \Delta = 0.1 \text{ and } M = 7.$$

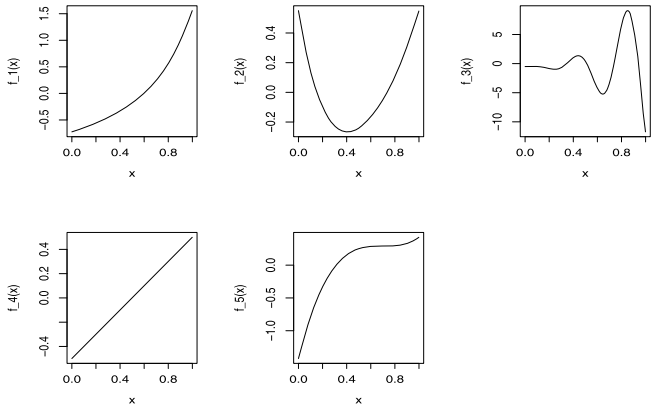

FIG 1. *Plots of the signal functions*

For example, for different values of d, the grid points s_m, $m = 1, \dots, 7$, are:

d	s_1	s_2	s_3	s_4	s_5	s_6	s_7
50	1	2	3	5	8	12	18
100	1	3	5	7	12	19	31
500	2	4	8	16	30	56	105
1000	2	5	11	22	44	89	177
5000	3	8	19	46	108	253	594
10000	3	10	25	63	158	398	1000
50000	5	14	44	130	384	1133	3343

Then, we use formula (36) to compute $\widehat{\eta}(s_{\widehat{m}})$ with the obtained value of $s_{\widehat{m}}$.

The study is based on $K = 500$ independent cycles of simulations. We estimate the Hamming error $s^{-1}\mathbf{E}(|\widehat{\eta}(s_{\widehat{m}}) - \eta|)$ of almost full recovery by means of the selector $\hat{\eta}(s_{\widehat{m}})$, which is free of s, by the quantity

$$err = \frac{1}{sK} \sum_{k=1}^{K} |\widehat{\eta}(s_{\widehat{m}}^{(k)}) - \eta|,$$

where $s_{\widehat{m}}^{(k)}$ is the value of $s_{\widehat{m}}$ obtained for the kth repetition of the experiment, $k = 1, \dots, K$. Due to the equality $\|\hat{\eta}(s_{\widehat{m}}) - \eta\| = |\widehat{\eta}(s_{\widehat{m}}) - \eta|$, the quantity err also estimates the risk $s^{-1}\mathbf{E}\|\hat{\eta}(s_{\widehat{m}}) - \eta\|$.

The adaptive procedure that produces our simulation results is amazingly fast and the obtained results are most encouraging, as seen from Table 5. The procedure never detects a signal if there is none.

As seen from Table 5, although the estimated error gets slightly larger as d increases, it is remarkably stable even for very small values of the ratio s/d.

The next table, Table 5, reports the values of the probability $\mathbf{P}(s_{\widehat{m}} \le s)$ of underestimating the true sparsity parameter s. With high probability $s_{\widehat{m}}$, with

s/d	0.1	0.05	0.01	$5\cdot10^{-3}$	10^{-3}	$5\cdot10^{-4}$	10^{-4}
$err(\widehat{\eta}(s_{\widehat{m}}))$	0.1272	0.1420	0.1616	0.1856	0.2156	0.2248	0.2340

TABLE 1
Estimated error of almost full selection as s/d tends to zero, $s = 5$.

$s/d = 5/d$	0.1	0.05	0.01	$5\cdot10^{-3}$	10^{-3}	$5\cdot10^{-4}$	10^{-4}
$\sum_{k=1}^{K}\mathbb{I}(s_{\widehat{m}}^{(k)} \le s)/K$	0.886	0.896	0.918	0.840	0.838	0.904	0.866

TABLE 2
Estimated probability of underestimating s for almost full selection as s/d tends to zero, $s = 5$.

$\widehat{m}$ as in (37), underestimates the true value of s, and the situation remains the same across different values of d.

Now we numerically study the question of how the signal strength effects the global risk of our selector. To this end, we modify the function $f_5(x)$ as follows:

$$f_{5,l}(x) = l\left((x-0.7)^3 + 0.058\right), \quad l \in \{0.01, 0.5, 1, 2, 3, 4, 5\}.$$

The function $f_{5,l}(x)$ has the L_2-norm that is proportional to l. For the adaptive selector in hand, Table 5 provides the norm of $f_{5,l}(x)$ and the estimated error of almost full recovery for different values of d. We note that, while being very stable, the error decreases when l is between 3 and 4 (or when $\|f_{5,l}\|_2$ is between 0.2761 and 0.3681). This is expected by the theory because, as $\|f_{5,l}\|_2$ increases with l, the upper boundary for selection is achieved and hence variable selection becomes easier. Moreover, starting from l equal to 4, the error of selection of the 5th component decays to zero very fast, which explains why starting from this value the selection error is stabilized.

l	0.01	0.5	1	2
$\|f_{5,l}\|_2$	0.0009	0.0460	0.0920	0.1841
$err(\widehat{\eta}(s_{\widehat{m}})), d = 1000$	0.3856	0.3684	0.3892	0.3924
$err(\widehat{\eta}(s_{\widehat{m}})), d = 5000$	0.3992	0.3948	0.4092	0.3916

l	3	4	5	6	7
$\|f_{5,l}\|_2$	0.2761	0.3681	0.4601	0.5522	0.6442
$err(\widehat{\eta}(s_{\widehat{m}})), d = 1000$	0.3620	0.1844	0.1848	0.1856	0.1684
$err(\widehat{\eta}(s_{\widehat{m}})), d = 5000$	0.3976	0.2064	0.2044	0.1992	0.1948

TABLE 3
Estimated error of the adaptive almost full selector for different choices of l, $s = 5$.

Next, we take $s = 10$ and the same values of d, ε, and σ as before. When

$s/d = 10/d$	0.2	0.1	0.02	0.01	$2\cdot10^{-3}$	10^{-3}	$2\cdot10^{-4}$
$err(\hat{\eta}(s_{\hat{m}}))$	0.1320	0.1456	0.1562	0.1890	0.2054	0.2254	0.2298

TABLE 4
Estimated error for almost full selection as s/d tends to zero, $s = 10$.

comparing the values of the estimated error of almost full selection in Table 5 with those in Table 5, we conclude that for fixed s/d these values are smaller for larger s and d, whereas for fixed d they are slightly larger for larger s.

Finally, let us examine how the risk varies depending on the L_2-norm of the same component function $f_{5,l}(x)$. Computed values of the estimated error are

l	0.01	0.5	1	2
$err(\widehat{\eta}(s_{\widehat{m}})), d = 1000$	0.2850	0.2690	0.2798	0.2772
$err(\widehat{\eta}(s_{\widehat{m}})), d = 5000$	0.2946	0.2942	0.3038	0.2932

l	3	4	5	6	7
$err(\widehat{\eta}(s_{\widehat{m}})), d = 1000$	0.2626	0.1798	0.1768	0.1850	0.1690
$err(\widehat{\eta}(s_{\widehat{m}})), d = 5000$	0.2960	0.1974	0.2024	0.1946	0.1942

TABLE 5

Estimated error of the adaptive almost full selector for different choices of l, $s = 10$.

presented in Table 5 ($s = 10$). These values are smaller than the corresponding values in Table 5 ($s = 5$). This can be explained by the fact that, in the case of $s = 10$ active component functions, changes that occur in the 5th component have less effect on the risk than in the case of $s = 5$ active components.

6. Concluding remarks

The present work has been in part motivated by the results of Genovese et al. [11] and Ingster and Stepanova [17], as outlined in the Introduction, who examined the effectiveness of simple thresholding, as opposed to the lasso, in high-dimensional problems of variable selection. In the context of variable selection in high dimensions, in both regression and white noise settings, simple thresholding provides plausible alternative to the lasso for a large range of problems. As a statistical tool, thresholding strategy is simple in nature and is not as computationally demanding as the lasso, especially in very high dimensional problems. At the same time, it is capable of doing at least as good as the lasso, or even better (see our Theorems 1 to 6, Theorems 9 to 11 in [11], and Theorems 1 and 2 in [17] for details). In light of these facts, we support the viewpoint of Genovese et al. [11] that for sparse high-dimensional regression problems a simple thresholding procedure merits further investigation.

In most publications on variable selection in high-dimensional settings, selection procedures are first proposed using certain heuristic arguments and their properties are then investigated. The approach employed in this paper is different. We first connect the problem of variable selection to that of signal detection. Then, we use some fundamental results of the minimax hypothesis testing theory, such as asymptotically exact conditions of distinguishability for alternatives described by function sets of certain degree of smoothness with some neighbourhood of the null hypothesis removed, to construct optimal selection procedures.

The use of the hypothesis testing methodology makes it possible to obtain sharp selection boundaries. These boundaries describe the variable selection

problem as a whole rather than a particular method. Therefore they can serve as benchmarks for practitioners who need to design experiments. The conditions for the possibility of successful variable selection involve the notion of a detection boundary and are stated in a way that makes a close correspondence between the obtained selection boundaries and the ones available for a linear regression model (see Remark 1 in Section 3 for details). Finally, to make the obtained selectors adaptive to the amount of signal contained in the data, we apply Lepski's method of adaptation. We do that because, unlike the case of exact variable selection (see [17] and Section 3 of this work), a common strategy of achieving adaptation by means of aggregation of test procedures fails to work in the almost full recovery regime when the data are highly sparse. The method of Lepski was originally introduced and used for problems of nonparametric estimation; its use here in a variable selection context turns out to be very effective.

To conclude our study, we point out possible directions for extending the results obtained in this paper. For the two examples of the function classes $\mathcal{F}_\sigma$ at hand, it might be of interest to produce asymptotically exact and almost full selectors in very high dimensional settings when the conditions $\log d = o\left(\varepsilon^{-2/(2\sigma+1)}\right)$ and $\log d = o(\log \varepsilon^{-1})$ on the growth of d as a function of ε are not valid.

In the present work, we have studied the problem of variable selection in a sparse additive model when an unknown regression function belongs to "dense" ellipsoids (corresponding to the Sobolev classes and the classes of analytic functions). It is also interesting to consider functions from "sparse" ellipsoids such as, for example, L_q-balls with $0 < q \le 1$. Another problem of interest is that of detecting signals with significant (large enough) L_p-norms when the signals belong to Besov classes $B^\sigma_{r,q}$ of σ-smooth functions with $1 \le p, r, q \le \infty$ and $(2\sigma+1)r \le p$ ("sparse" case). We expect that for such kind of signals, i.e., the signals that admit sparse representations in some bases (e.g., wavelet bases), our selection procedure, with proper modifications, will also work.

Furthermore, handling the problem of variable selection in a sequence space model, general ellipsoids $\{\theta \in l_2(\mathbb{Z}) : \sum_{k\in\mathbb{Z}} c_k^2\theta_k^2 \le 1\}$ in $l_2(\mathbb{Z})$, with semi-axes c_k decreasing fast enough, could be studied. A more complicated model, in which a d-variate regression function f admits a decomposition to a sum of s-variate components, with $2 \le s \le d$ and only a small number of these components being non-zero, also deserves some attention. The corresponding signal detection problem was solved in [20]. Specifically, the results on sharp detection boundaries obtained in [20] may serve as a basis for solving a more intricate problem of variable selection which extends the problem in hand in the following way. For $1 \le k \le d$, let $u = (j_1, \dots, j_k) \subset \{1, \dots, d\}$, $1 \le j_1 < \dots < j_k \le d$, $\mathcal{U}_{k,d} = \{u : u \subset \{1, \dots, d\}, |u| = k\}$, and $\mathbf{x}_u = (x_{j_1}, \dots, x_{j_k}) \in \mathbb{R}^k$. Rather than dealing with a sparse additive signal $f(\mathbf{x}) = \sum_{j=1}^d \eta_j f_j(x_j)$, $\mathbf{x} \in [0,1]^d$, with s active components f_j, we may instead assume that

$$f(\mathbf{x}) = \sum_{u\in\mathcal{U}_{s,d}} \eta_u f_u(\mathbf{x}_u),$$

where $s = s_d \to \infty$ and $s = o(d)$, $\eta_u = 1$ for some $u \in \mathcal{U}_{s,d}$ and zero otherwise, in which case the role of d is taken by $|\mathcal{U}_{s,d}| = \binom{d}{s}$. Technically, this problem is more challenging, especially in case of an unknown parameter s, and will be treated elsewhere.

To pursue more practical goals, one can try to translate the results obtained for an additive s-sparse Gaussian white noise model to the related discrete regression model for which the corresponding signal detection problem was solved in [1].

7. Proofs

We first consider the question of determining the conditions on d as a function of ε under which almost full variable selection is possible. Violation of these conditions will lead to entirely different selection strategies.

7.1. Conditions for almost full variable selection

In the sequence space of Fourier coefficients, consider testing the null hypothesis $H_{0j} : \theta_j = 0$ versus the alternative $H_{1j} : \theta_j \in \Theta_\sigma(r_\varepsilon)$, where the set $\Theta_\sigma(r_\varepsilon)$ is given by (14). It is easy to see that, under the null hypothesis H_{0j}, we have (see, for example, Section 4.1 of [17])

$$\mathbf{E}_0(t_j) = 0, \quad \mathbf{Var}_0(t_j) = 1,$$

while under the alternative H_{1j}, where for all sufficiently small ε a small parameter $r_\varepsilon > 0$ satisfies $r_\varepsilon / r_\varepsilon^*(s) > 1$,

$$\begin{aligned} \mathbf{E}_{\theta_j}(t_j) &= \varepsilon^{-2} \sum_{1 \le |k| \le K_\varepsilon} \omega_k(r_\varepsilon^*(s))\theta_{j,k}^2 \ge u_\varepsilon(r_\varepsilon^*(s)), \\ \mathbf{Var}_{\theta_j}(t_j) &= 1 + O(\mathbf{E}_{\theta_j}(t_j) \max_{1 \le |k| \le K_\varepsilon} \omega_k(r_\varepsilon^*(s))). \end{aligned} \tag{41}$$

Furthermore, under the above restrictions on r_ε and the detection boundary $r_\varepsilon^*(s)$, the following result holds (in case of Sobolev spaces, see Proposition 7.1 in [10] and Lemma 1 in [17]; in case of the space $\mathcal{F}_\sigma$ of analytic functions, the proof is similar to that for Sobolev spaces).

Let the quantity $T = T_\varepsilon \to -\infty$ and the weight functions $\omega_k(r_\varepsilon^*(s))$ as in (27) be such that as $\varepsilon \to 0$

$$T \max_{1 \le |k| \le K_\varepsilon} \omega_k(r_\varepsilon^*(s)) \to 0 \quad \text{and} \quad \mathbf{E}_{\theta_j}(t_j(s)) \max_{1 \le |k| \le K_\varepsilon} \omega_k(r_\varepsilon^*(s)) \to 0. \tag{42}$$

Then as $\varepsilon \to 0$

$$\mathbf{P}_0(t_j \le T) \le \exp\left(-\frac{T^2}{2}(1 + o(1))\right), \tag{43}$$

and for all $j = 1, \dots, d$, uniformly in $\theta_j \in \Theta_\sigma(r_\varepsilon)$,

$$\mathbf{P}_{\theta_j}(t_j - \mathbf{E}_{\theta_j}(t_j) \le T) \le \exp\left(-\frac{T^2}{2}(1 + o(1))\right). \tag{44}$$

For both function classes $\mathcal{F}_\sigma$ of our interest, the exponential bounds (43) and (44) will be applied below to the quantity $T = T_\varepsilon \to -\infty$ of order $O(\log^{1/2} d)$. This observation, together with (28) and (41), transforms requirement (42) into

$$\log^{1/2} d \max_{1 \le |k| \le K_\varepsilon} \omega_k(r_\varepsilon^*(s)) \to 0, \quad \varepsilon \to 0, \tag{45}$$

Condition (45) gives a restriction on the growth of $d = d_\varepsilon$ ensuring that the selection procedure works as designed. Indeed, as shown in Section 4.1 in [17], for the Sobolev space of σ-smooth functions, one has

$$\omega_k(r_\varepsilon) \asymp r_\varepsilon^{1/(2\sigma)} \quad \text{for} \quad 1 \le |k| \le K_\varepsilon,$$

and

$$r_\varepsilon^*(s) \asymp (\varepsilon \log d)^{\sigma/(4\sigma+1)}.$$

Here and below, the notation $a_\varepsilon \asymp b_\varepsilon$ means that $0 < \liminf_{\varepsilon \to 0}(a_\varepsilon / b_\varepsilon) \le \limsup_{\varepsilon \to 0}(a_\varepsilon / b_\varepsilon) < \infty$. Therefore, condition (45) is fulfilled when

$$\log d = o(\varepsilon^{-2/(2\sigma+1)}) \tag{46}$$

In case of the class $\mathcal{F}_\sigma$ of analytic functions, one has

$$\omega_k(r_\varepsilon) \asymp \log^{-1/2}(r_\varepsilon^{-1}) \quad \text{for} \quad 1 \le |k| \le K_\varepsilon, \tag{47}$$

and, in view of (23) and (28), the quantity $r_\varepsilon^*(s)$ satisfies

$$\log^{1/2} d \asymp \left(\frac{r_\varepsilon^*(s)}{\varepsilon}\right)^2 \log^{-1/2}\left((r_\varepsilon^*(s))^{-1}\right),$$

implying

$$r_\varepsilon^*(s) \asymp \varepsilon \log^{1/4}(d) \log^{1/4}((r_\varepsilon^*(s))^{-1}).$$

Therefore $\log\left((r_\varepsilon^*(s))^{-1}\right) \sim \log(\varepsilon^{-1})$, and (see (47))

$$\omega_k(r_\varepsilon^*(s)) \asymp \log^{-1/2}(\varepsilon^{-1}).$$

From this, the technical condition (45) holds true when, cf. formula (46),

$$\log d = o(\log(\varepsilon^{-1})), \quad \varepsilon \to 0. \tag{48}$$

7.2. Proofs of theorems

In this section, we prove Theorems 3 and 4. The proofs of Theorems 5 and 6 go along the same lines and therefore are omitted. Throughout the proof, the exponential bounds (43) and (44) on the tail probabilities of the statistics $t_j(s)$ will frequently be used.

Proof of Theorem 3. Let $m_0 \in \{2, \ldots, M\}$ be such that

$$s_{m_0-1} \leq s < s_{m_0},$$

which implies that $s_{m_0}/s < d^{\Delta}$. Then, using the definition of the selector $\hat{\eta}(s_{\hat{m}})$, we can write

$$\begin{aligned}
&\sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} s^{-1} \mathbf{E}_{\eta,\theta} |\hat{\eta}(s_{\hat{m}}) - \eta| \\
&\leq \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} s^{-1} \mathbf{E}_{\eta,\theta} \left(|\hat{\eta}(s_{\hat{m}}) - \eta| | \hat{m} < m_0 \right) \mathbf{P}_{\eta,\theta} \left(\hat{m} < m_0 \right) \\
&+ \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} s^{-1} \mathbf{E}_{\eta,\theta} \left(|\hat{\eta}(s_{\hat{m}}) - \eta| | \hat{m} \geq m_0 \right) \mathbf{P}_{\eta,\theta} \left(\hat{m} \geq m_0 \right) \\
&\leq \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} s^{-1} \mathbf{E}_{\eta,\theta} \left(|\hat{\eta}(s_{\hat{m}}) - \eta| | \hat{m} < m_0 \right) \mathbf{P}_{\eta,\theta} \left(\hat{m} < m_0 \right) \\
&\qquad + \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} (d/s) \mathbf{P}_{\eta,\theta} \left(\hat{m} \geq m_0 \right) =: I_1 + I_2. \qquad (49)
\end{aligned}$$

To complete the proof, we need to show that I_1 and I_2 are both negligibly small when ε is small.

Consider the term I_1 and observe that for all $\eta \in \mathcal{H}_{d,s}$ and $\theta \in \Theta_{\sigma,d}(r_\varepsilon)$,

$$\begin{aligned}
&s^{-1} \mathbf{E}_{\eta,\theta} \left(|\hat{\eta}(s_{\hat{m}}) - \eta| | \hat{m} < m_0 \right) \mathbf{P}_{\eta,\theta} \left(\hat{m} < m_0 \right) \\
&\leq s^{-1} \mathbf{E}_{\eta,\theta} \left(|\hat{\eta}(s_{\hat{m}}) - \hat{\eta}(s_{m_0})| \ \hat{m} < m_0 \right) \\
&+ s^{-1} \mathbf{E}_{\eta,\theta} \left(|\hat{\eta}(s_{m_0}) - \eta| | \hat{m} < m_0 \right) \mathbf{P}_{\eta,\theta} \left(\hat{m} < m_0 \right) \\
&\leq s^{-1} v_{m_0} + s^{-1} \mathbf{E}_{\eta,\theta} |\hat{\eta}(s_{m_0}) - \eta|,
\end{aligned}$$

where by (35) and the choice of the sequences τ_d and $\Delta = \Delta_d$

$$s^{-1} v_{m_0} = \tau_d^{-1} (s_{m_0}/s) < {\tau_d}^{-1} d^{\Delta} = o(1).$$

Next, by definition of the set $\mathcal{H}_{d,s}$ of s-sparse d-dimensional vectors η, we have

$$\begin{aligned}
&\sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} s^{-1} \mathbf{E}_{\eta,\theta} |\hat{\eta}(s_{m_0}) - \eta| \\
&\leq (d/s) \mathbf{P}_0 \left(t_1(s_{m_0}) > \sqrt{2 \log(d/s_{m_0}) + \delta \log d} \right) \\
&+ \sup_{\theta_1 \in \Theta_\sigma(r_\varepsilon)} \mathbf{P}_{\theta_1} \left(t_1(s_{m_0}) \leq \sqrt{2 \log(d/s_{m_0}) + \delta \log d} \right) \qquad (50)
\end{aligned}$$

where by (43) the first summand in the above expression satisfies

$$\begin{aligned}&(d/s)\mathbf{P}_0\left(t_1(s_{m_0}) > \sqrt{2\log(d/s_{m_0}) + \delta\log d}\right)\\ &\le (d/s)\exp\left(-\left(\log(d/s_{m_0}) + (\delta/2)\log d\right)(1+o(1))\right)\\ &= O\left((s_{m_0}/s)d^{-\delta/2}\right) = O\left(d^{\Delta-\delta/2}\right) = o(1),\end{aligned}$$

and the last equality is due to (30) and (35).

To treat the second term on the right side of (50), recall that $1 < s_{m_0}/s < d^{\Delta}$. Then, by the assumption on the parameter $r_\varepsilon = r_\varepsilon(s)$ and the 'continuity' of the function $u_\varepsilon(r_\varepsilon)$ as stated in (17), using the fact that $\Delta\log d \to 0$ as $d\to\infty$, one can find a constant $\delta_1 > 0$ such that for all sufficiently small ε

$$r_\varepsilon \ge r^*_\varepsilon(s_{m_0})(1+\delta_1).$$

From this, using Proposition 4.1 in [10] and recalling formula (41),

$$\begin{aligned}\inf_{\theta_1\in\Theta_\sigma(r_\varepsilon)} &\mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right) \ge \inf_{\theta_1\in\Theta_\sigma(r^*_\varepsilon(s_{m_0})(1+\delta_1))}\mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right)\\ &\ge (1+\delta_1)^2\inf_{\theta_1\in\Theta_\sigma(r^*_\varepsilon(s_{m_0}))}\mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right) \ge (1+\delta_1)^2 u_\varepsilon(r^*_\varepsilon(s_{m_0}))\\ &= (1+\delta_1)^2\sqrt{2\log(d/s_{m_0})} > \sqrt{2\log(d/s_{m_0}) + \delta\log d},\end{aligned} \tag{51}$$

where the last inequality follows from the fact that $d^c \le s_{m_0} < d^C$, which implies $\delta\log d = o(\log(d/s_{m_0}))$. Thus as $\varepsilon\to 0$

$$\sqrt{2\log(d/s_{m_0}) + \delta\log d} - \inf_{\theta_1\in\Theta_\sigma(r_\varepsilon)}\mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right) \to -\infty. \tag{52}$$

Now (44) in combination with (51) and (52) gives, uniformly in $\theta_1\in\Theta_\sigma(r_\varepsilon)$,

$$\begin{aligned}&\mathbf{P}_{\theta_1}\left(t_1(s_{m_0}) \le \sqrt{2\log(d/s_{m_0}) + \delta\log d}\right)\\ \le\ &\mathbf{P}_{\theta_1}\Big(t_1(s_{m_0}) - \mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right) \le \sqrt{2\log(d/s_{m_0}) + \delta\log d}\\ &\quad - \inf_{\theta_1\in\Theta_\sigma(r_\varepsilon)}\mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right)\Big)\\ \le\ &\mathbf{P}_{\theta_1}\left(t_1(s_{m_0}) - \mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right) \le -\sqrt{2\log(d/s_{m_0})}\left[(1+\delta_1)^2 - 1 + o(1)\right]\right)\\ \le\ &\exp\left(-\log(d/s_{m_0})\left[(1+\delta_1)^2 - 1 + o(1)\right]^2(1+o(1))\right)\\ =\ &O\left((s_{m_0}/d)^{\left[(1+\delta_1)^2-1\right]^2}\right) = o(1).\end{aligned}$$

Putting everything together, we conclude that the first term on the right side of (49) satisfies

$$I_1 = o(1), \quad \varepsilon\to 0. \tag{53}$$

Let us now show that as $\varepsilon \to 0$

$$I_2 = \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} (d/s) \mathbf{P}_{\eta,\theta} \left(\hat{m} \geq m_0 \right) = o(1).$$

By definition of $\hat{m}$, for all $\eta \in \mathcal{H}_{d,s}$ and all $\theta \in \Theta_{\sigma,d}(r_\varepsilon)$,

$$\begin{aligned}
\mathbf{P}_{\eta,\theta} \left(\hat{m} \geq m_0 \right) &= \sum_{k=m_0}^{M} \mathbf{P}_{\eta,\theta} \left(\hat{m} = k \right) \\
&= \sum_{k=m_0}^{M} \mathbf{P}_{\eta,\theta} \left(\exists\, i \in \{k, \ldots, M\} : |\hat{\eta}(s_{k-1}) - \hat{\eta}(s_i)| > v_i \right) \\
&\leq \sum_{k=m_0}^{M} \sum_{i=k}^{M} \mathbf{P}_{\eta,\theta} \left(|\hat{\eta}(s_{k-1}) - \hat{\eta}(s_i)| > v_i \right) \\
&= \sum_{k=m_0}^{M} \sum_{i=k}^{M} \mathbf{P}_{\eta,\theta} \left(\sum_{j=1}^{d} |\hat{\eta}_j(s_{k-1}) - \hat{\eta}_j(s_i)| > v_i \right).
\end{aligned}$$

Now, we introduce independent events

$$A_j(s) = \left\{ t_j(s) \leq \sqrt{2 \log(d/s) + \delta \log d} \right\}, \quad j = 1, \ldots, d,$$

and denote by $\overline{A_j(s)}$ the complement of $A_j(s)$. Observing that for all $m_0 \leq k \leq i \leq M$ the quantity $|\hat{\eta}_j(s_{k-1}) - \hat{\eta}_j(s_i)|$ is non-zero only if either $\overline{A_j(s_{k-1})} \cap A_j(s_i)$ or $A_j(s_{k-1}) \cap \overline{A_j(s_i)}$ occurs, we may continue

$$\begin{aligned}
&\mathbf{P}_{\eta,\theta} \left(\hat{m} \geq m_0 \right) \\
&\leq \sum_{k=m_0}^{M} \sum_{i=k}^{M} \mathbf{P}_{\eta,\theta} \left(\sum_{j=1}^{d} \left[\mathbb{I} \left(\overline{A_j(s_{k-1})} \cap A_j(s_i) \right) + \mathbb{I} \left(A_j(s_{k-1}) \cap \overline{A_j(s_i)} \right) \right] > v_i \right).
\end{aligned}$$

To bound this sum, we apply Bernstein's inequality saying that if $\mathbb{X}_1, \ldots, \mathbb{X}_d$ are independent random variables such that for all $j = 1, \ldots, d$ and for some $H > 0$

$$\mathbf{E}(\mathbb{X}_j) = 0 \quad \text{and} \quad \left| \mathbf{E}(\mathbb{X}_j^m) \right| \leq \frac{\mathbf{E}(\mathbb{X}_j^2)}{2} H^{m-2} m! < \infty, \quad m = 2, 3, \ldots, \tag{54}$$

then (see, for example, pp. 164–165 of [2])

$$\max \left\{ \mathbf{P} \left(\mathbb{S}_d \geq t \right), \mathbf{P} \left(\mathbb{S}_d \leq -t \right) \right\} \leq \begin{cases} \exp \left(-t^2 / 4B_d^2 \right) & \text{if} \quad 0 \leq t \leq B_d^2 / H, \\ \exp \left(-t / 4H \right) & \text{if} \quad t \geq B_d^2 / H, \end{cases} \tag{55}$$

where $\mathbb{S}_d = \sum_{j=1}^{d} \mathbb{X}_j$ and $B_d^2 = \sum_{j=1}^{d} \mathbf{E}(\mathbb{X}_j^2)$. Observe that for independent random variables $\mathbb{X}_1, \ldots, \mathbb{X}_d$ with the property

$$\mathbf{E}(\mathbb{X}_j) = 0 \quad \text{and} \quad |\mathbb{X}_j| \leq M, \quad j = 1, \ldots, d,$$

for some $M > 0$, the Bernstein condition (54) holds with $H = M/3$. Below we will use Bernstein's inequality in the case of $t \geq B_d^2/H$.

To do this, let us introduce random variables $\mathbb{X}_j = \mathbb{X}_j(s_{k-1}, s_i)$, $1 \leq j \leq d$, $m_0 \leq k \leq M$, $k \leq i \leq M$, by the formula

$$\begin{aligned}\mathbb{X}_j &= \mathbb{I}\left(\overline{A_j(s_{k-1})} \cap A_j(s_i)\right) + \mathbb{I}\left(A_j(s_{k-1}) \cap \overline{A_j(s_i)}\right) \\ &\quad - \left[\mathbf{P}_{\eta,\theta}\left(\overline{A_j(s_{k-1})} \cap A_j(s_i)\right) + \mathbf{P}_{\eta,\theta}\left(A_j(s_{k-1}) \cap \overline{A_j(s_i)}\right)\right],\end{aligned}$$

and observe that $|\mathbb{X}_j| \leq 4$, $j = 1, \ldots, d$, and for all $\eta \in \mathcal{H}_{d,s}$ and $\theta \in \Theta_{\sigma,d}(r_\varepsilon)$

$$\mathbf{E}_{\eta,\theta}(\mathbb{X}_j) = 0, \quad j = 1, \ldots, d.$$

Before applying Bernstein's inequality, we show that for all $\eta \in \mathcal{H}_{d,s}$ and $\theta \in \Theta_{\sigma,d}(r_\varepsilon)$, and for all $m_0 \leq k \leq M$ and $k \leq i \leq M$

$$\sum_{j=1}^{d}\left[\mathbf{P}_{\eta,\theta}\left(\overline{A_j(s_{k-1})} \cap A_j(s_i)\right) + \mathbf{P}_{\eta,\theta}\left(A_j(s_{k-1}) \cap \overline{A_j(s_i)}\right)\right] = o(v_i). \tag{56}$$

We have, for all η in $\mathcal{H}_{d,s}$ and all θ in $\Theta_{\sigma,d}(r_\varepsilon)$:

$$\begin{aligned}&\sum_{j=1}^{d}\left[\mathbf{P}_{\eta,\theta}\left(\overline{A_j(s_{k-1})} \cap A_j(s_i)\right) + \mathbf{P}_{\eta,\theta}\left(A_j(s_{k-1}) \cap \overline{A_j(s_i)}\right)\right] \\ &= (d-s)\left[\mathbf{P}_0\left(\overline{A_1(s_{k-1})} \cap A_1(s_i)\right) + \mathbf{P}_0\left(A_1(s_{k-1}) \cap \overline{A_1(s_i)}\right)\right] \\ &\quad + s \sup_{\theta_1 \in \Theta_\sigma(r_\varepsilon)}\left[\mathbf{P}_{\theta_1}\left(\overline{A_1(s_{k-1})} \cap A_1(s_i)\right) + \mathbf{P}_{\theta_1}\left(A_1(s_{k-1}) \cap \overline{A_1(s_i)}\right)\right] \\ &\leq d\left[\mathbf{P}_0\left(t_1(s_{k-1}) > \sqrt{2\log(d/s_{k-1}) + \delta\log d}\right)\right. \\ &\quad \left. + \mathbf{P}_0\left(t_1(s_i) > \sqrt{2\log(d/s_i) + \delta\log d}\right)\right] \\ &\quad + s \sup_{\theta_1 \in \Theta_\sigma(r_\varepsilon)}\left[\mathbf{P}_{\theta_1}\left(t_1(s_{k-1}) \leq \sqrt{2\log(d/s_{k-1}) + \delta\log d}\right)\right. \\ &\quad \left. + \mathbf{P}_{\theta_1}\left(t_1(s_i) \leq \sqrt{2\log(d/s_i) + \delta\log d}\right)\right] \\ &=: J_1(s_{k-1}, s_i) + J_2(s_{k-1}, s_i).\end{aligned} \tag{57}$$

Recalling (43) and the relation $\tau_d d^{-\delta/2} \to 0$ as $d \to \infty$, we get

$$\begin{aligned}&d\,\mathbf{P}_0\left(t_1(s_i) > \sqrt{2\log(d/s_i) + \delta\log d}\right) \\ &\leq d\exp\left(-\left(\log(d/s_i) + (\delta/2)\log d\right)(1 + o(1))\right) \\ &= O\left(s_i d^{-\delta/2}\right) = O\left(v_i \tau_d d^{-\delta/2}\right) = o(v_i).\end{aligned}$$

Similarly, using the fact that $v_{k-1} < v_i$ when $k \leq i \leq M$, we obtain

$$d\,\mathbf{P}_0\left(t_1(s_{k-1}) > \sqrt{2\log(d/s_{k-1}) + \delta\log d}\right) = o(v_{k-1}) = o(v_i).$$

Therefore for all $m_0 \le k \le M$ and $k \le i \le M$

$$J_1(s_{k-1}, s_i) = o(v_i). \tag{58}$$

Consider the second term on the right side of (57), $J_2(s_{k-1}, s_i)$. First, note that for all $m_0 \le k \le M$ and $k \le i \le M$,

$$s < s_i \quad \text{and} \quad s < s_{k-1}, \quad k \ne m_0.$$

and for $k = m_0$ one has $s_{k-1} = s_{m_0-1} \le s$, which implies $s/s_{m_0-1} < d^{\Delta}$. Therefore, by the assumption on $r_\varepsilon = r_\varepsilon(s)$ and the 'continuity' of the function $u_\varepsilon(r_\varepsilon)$ as cited in (17), using the fact that $\Delta \log d \to 0$ as $d \to \infty$, one can find constants $\delta_2 > 0$ and $\delta_3 > 0$ such that for all sufficiently small ε

$$r_\varepsilon \ge r_\varepsilon^*(s_i)(1+\delta_2) \quad \text{and} \quad r_\varepsilon \ge r_\varepsilon^*(s_{k-1})(1+\delta_3)$$

when $m_0 \le k \le M$ and $k \le i \le M$. From this, for all sufficiently small ε, cf. (51),

$$\inf_{\theta_1 \in \Theta_\sigma(r_\varepsilon)} \mathbf{E}_{\theta_1}\left(t_1(s_i)\right) \ge (1+\delta_2)^2 \sqrt{2\log(d/s_i)} > \sqrt{2\log(d/s_i) + \delta \log d}, \tag{59}$$

and hence as $\varepsilon \to 0$

$$\sqrt{2\log(d/s_i) + \delta \log d} - \inf_{\theta_1 \in \Theta_\sigma(r_\varepsilon)} \mathbf{E}_{\theta_1}\left(t_1(s_i)\right) \to -\infty. \tag{60}$$

It now follows from (44), (59), and (60) that, uniformly in $\theta_1 \in \Theta_\sigma(r_\varepsilon)$,

$$\begin{aligned}
&s\mathbf{P}_{\theta_1}\left(t_1(s_i) \le \sqrt{2\log(d/s_i) + \delta \log d}\right) \\
&\le s\mathbf{P}_{\theta_1}\left(t_1(s_i) - \mathbf{E}_{\theta_1}\left(t_1(s_i)\right) \le \sqrt{2\log(d/s_i) + \delta \log d} - \inf_{\theta_1 \in \Theta_\sigma(r_\varepsilon)} \mathbf{E}_{\theta_1}\left(t_1(s_i)\right)\right) \\
&\le s\mathbf{P}_{\theta_1}\left(t_1(s_i) - \mathbf{E}_{\theta_1}\left(t_1(s_i)\right) \le -\sqrt{2\log(d/s_i)}\left[(1+\delta_2)^2 - 1 + o(1)\right]\right) \\
&\le s\exp\left(-\log(d/s_i)\left[(1+\delta_2)^2 - 1 + o(1)\right]^2 (1+o(1))\right) \\
&= O\left(s(s_i/d)^{[(1+\delta_2)^2-1]^2}\right) = O\left(s_i(s_i/d)^{[(1+\delta_2)-1]^2}\right) = o(v_i),
\end{aligned}$$

where the last equality is due to restriction (33) imposed on s. Also, as relation (60) continues to hold with s_{k-1}, $m_0 \le k \le M$, instead of s_i, similar arguments yield

$$s\mathbf{P}_{\theta_1}\left(t_1(s_{k-1}) \le \sqrt{2\log(d/s_{k-1}) + \delta \log d}\right) = o(v_{k-1}) = o(v_i),$$

which implies

$$J_2(s_{k-1}, s_i) = o(v_i). \tag{61}$$

Combining (57), (58) and (61), we arrive at (56). We see then by (56) that

$$\sum_{j=1}^{d} \mathbf{E}_{\eta,\theta}(\mathbb{X}_j^2) = \left(\sum_{j=1}^{d} \left[\mathbf{P}_{\eta,\theta} \left(\overline{A_j(s_{k-1})} \cap A_j(s_i) \right) + \mathbf{P}_{\eta,\theta} \left(A_j(s_{k-1}) \cap \overline{A_j(s_i)} \right) \right] \right) (1 + o(1)) = o(v_i).$$

Therefore, we use Bernstein's inequality as in (55) for the case of $t \geq B_d^2/H$ with $H = 4/3$ and get as $\varepsilon \to 0$

$$I_2 = \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} (d/s) \mathbf{P}_{\eta,\theta} (\hat{m} \geq m_0)$$

$$\leq \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} (d/s) \sum_{k=m_0}^{M} \sum_{i=k}^{M} \mathbf{P}_{\eta,\theta} \left(\sum_{j=1}^{d} \mathbb{X}_j > v_i(1 + o(1)) \right)$$

$$\leq (d/s) \sum_{k=m_0}^{M} \sum_{i=k}^{M} \exp\left(-(3v_i/16)(1 + o(1))\right) = O\left(M^2 (d/s) \exp\left(-(3/16) v_{m_0}\right)\right)$$

$$= O\left(M^2 (d/s) \exp\left(-(3d^c/16\tau_d)\right)\right) = o(1).$$

This in combination with (49) and (53) completes the proof of Theorem 3. □

Proof of Theorem 4. To prove the theorem, we first pick good prior distributions on $\eta = (\eta_j)$ and $\theta = (\theta_j)$. Having done this, we bound the normalized minimax risk by the normalized Bayes risk and show that the latter risk is strictly positive. The first part of the proof, up to relation (66), goes along the lines of that of Theorem 2 in [17], with $p = s/d$ instead of $p = d^{-\beta}$.

Let $\theta_j^* = (\theta_{j,k}^*)_{k \in \mathbb{Z}}$ be the extremal sequence in the problem (the same for all $j = 1, \ldots, d$) of minimizing $(2\varepsilon^4)^{-1} \sum_{k \in \mathbb{Z}} \theta_k^4$ over $\Theta_\sigma(r_\varepsilon)$:

$$\frac{1}{2\varepsilon^4} \sum_{k \in \mathbb{Z}} (\theta_{j,k}^*)^4 = \inf_{\theta_j \in \Theta_\sigma(r_\varepsilon)} \frac{1}{2\varepsilon^4} \sum_{k \in \mathbb{Z}} \theta_{j,k}^4.$$

Let the prior distribution of a 'vector' $\theta = (\theta_1, \ldots, \theta_d) \in \Theta_{\sigma,d}(r_\varepsilon)$ be of the form

$$\pi_\theta(d\theta) = \prod_{j=1}^{d} \pi_{\theta_j}(d\theta_j), \quad \pi_{\theta_j}(d\theta_j) = \prod_{1 \leq |k| \leq K_\varepsilon} \left(\frac{\delta_{-\theta_{j,k}^*} + \delta_{\theta_{j,k}^*}}{2} \right) (d\theta_{j,k}),$$

where δ_x is the Dirac measure at point x. Denote by

$$p = s/d$$

the portion of non-zero components of a vector $\eta = (\eta_1, \ldots, \eta_d) \in \mathcal{H}_{d,s}$. The prior distribution of η is naturally defined to be

$$\pi_\eta(d\eta) = \prod_{j=1}^{d} \pi_{\eta_j}(d\eta_j), \quad \pi_{\eta_j}(d\eta_j) = ((1-p)\delta_0 + p\delta_1)(d\eta_j).$$

Then, assuming that $\theta = (\theta_j)$ and $\eta = (\eta_j)$ are independent, we get

$$\begin{aligned} R_\varepsilon &:= \inf_{\tilde\eta} \sup_{\eta\in\mathcal{H}_{d,s}} \sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\eta-\tilde\eta| \geq s^{-1}\inf_{\tilde\eta}\mathbf{E}_{\pi_\eta}\mathbf{E}_{\pi_\theta}\mathbf{E}_{\eta,\theta}|\eta-\tilde\eta| \\ &= s^{-1}\inf_{\tilde\eta}\mathbf{E}_{\pi_\eta}\mathbf{E}_{\pi_\theta}\mathbf{E}_{\eta,\theta}\sum_{j=1}^d|\eta_j-\tilde\eta_j| = s^{-1}\inf_{\tilde\eta}\sum_{j=1}^d\mathbf{E}_{\pi_{\eta_j}}\mathbf{E}_{\pi_{\theta_j}}\mathbf{E}_{\eta_j\theta_j}|\eta_j-\tilde\eta_j|, \end{aligned}$$

where the infimum is over all selectors $\tilde\eta = (\tilde\eta_j)$ and $\mathbf{E}_{\eta_j\theta_j}$ is the expected value that corresponds to the measure $\mathbf{P}_{\eta_j\theta_j}$ induced by the observation $X_j = (X_{j,k})_{1\leq|k|\leq K_\varepsilon}$ consisting of independent random variables $X_{j,k}$ that follow normal distributions $\mathcal{N}(\eta_j\theta_{j,k},\varepsilon^2)$.

Consider the mixture of distributions given by the formula

$$\begin{aligned} \mathbf{P}_{\pi,\eta_j}(dX_j) &= \mathbf{E}_{\pi_{\theta_j}}\mathbf{P}_{\eta_j\theta_j}(dX_{j,k}) \\ &= \prod_{1\leq|k|\leq K_\varepsilon}\left(\frac{\mathcal{N}(-\eta_j\theta^*_{j,k},\varepsilon^2)+\mathcal{N}(\eta_j\theta^*_{j,k},\varepsilon^2)}{2}\right)(dX_{j,k}). \quad (62) \end{aligned}$$

In particular, when $\eta_j = 0$, $\mathbf{P}_{\pi,0}(dX_j) = \prod_{1\leq|k|\leq K_\varepsilon}\mathcal{N}(0,\varepsilon^2)(dX_{j,k})$. Using the notation

$$v^*_{j,k} = \frac{\theta^*_{j,k}}{\varepsilon}, \quad (63)$$

we obtain with respect to the probability measure $\mathbf{P}_{\pi,\eta_j}$

$$Y_{j,k} := \frac{X_{j,k}}{\varepsilon} = \eta_j v^*_{j,k} + \xi_{j,k} \overset{\text{ind.}}{\sim} \mathcal{N}(\eta_j v^*_{j,k},1), \quad 1\leq j\leq d, \quad 1\leq|k|\leq K_\varepsilon.$$

Next, denoting $Y_j = (Y_{j,k})_{1\leq|k|\leq K_\varepsilon}$, we may rewrite the likelihood ratio in the form

$$\frac{d\mathbf{P}_{\pi,\eta_j}}{d\mathbf{P}_{\pi,0}}(Y_j) = \prod_{1\leq|k|\leq K_\varepsilon}\exp\left(-\frac{\eta_j(v^*_{j,k})^2}{2}\right)\cosh\left(\eta_j v^*_{j,k}Y_{j,k}\right). \quad (64)$$

From this, using the fact that each η_j takes on only two values, zero and one, with respective probabilities $(1-p)$ and p, we may continue

$$\begin{aligned} R_\varepsilon &\geq s^{-1}\sum_{j=1}^d\inf_{\tilde\eta_j}\mathbf{E}_{\pi_{\eta_j}}\mathbf{E}_{\pi,\eta_j}|\eta_j-\tilde\eta_j| \\ &= s^{-1}\sum_{j=1}^d\inf_{\tilde\eta_j}\left[(1-p)\mathbf{E}_{\pi,0}(\tilde\eta_j)+p\mathbf{E}_{\pi,1}(1-\tilde\eta_j)\right], \quad (65) \end{aligned}$$

where $\inf_{\tilde\eta_j}(1-p)\mathbf{E}_{\pi,0}(\tilde\eta_j)+p\mathbf{E}_{\pi,1}(1-\tilde\eta_j)$ is the Bayes risk in the problem of testing two simple hypotheses

$$H_0 : \mathbf{P} = \mathbf{P}_{\pi,0} \quad \text{vs.} \quad H_1 : \mathbf{P} = \mathbf{P}_{\pi,1},$$

with the probability measures $\mathbf{P}_{\pi,0}$ and $\mathbf{P}_{\pi,1}$ defined according to (62). In particular, under the null hypothesis, the vector $Y_j = (Y_{j,k})_{1\le|k|\le K_\varepsilon}$ has a normal distribution with density function $p_{\pi,0}(t) = \prod_{1\le|k|\le K_\varepsilon} (2\pi)^{-1/2}\exp(-t_k^2/2)$, $t = (t_k)_{1\le|k|\le K_\varepsilon}$. By (64) the likelihood ratio in this problem becomes

$$\Lambda_\pi(Y_j) := \frac{d\mathbf{P}_{\pi,1}}{d\mathbf{P}_{\pi,0}}(Y_j) = \prod_{1\le|k|\le K_\varepsilon} \exp\left(-\frac{(v^*_{j,k})^2}{2}\right)\cosh\left(v^*_{j,k}Y_{j,k}\right),$$

and the optimal (Bayes) test η_B that minimizes the Bayes risk in hand has the form (see, for example, [6, Sec. 8.11])

$$\eta_B(Y_j) = \mathbb{I}\left(\Lambda_\pi(Y_j) \ge \frac{1-p}{p}\right).$$

Using this, we infer from (65) that

$$\begin{aligned} R_\varepsilon &= \inf_{\tilde\eta}\sup_{\eta\in\mathcal{H}_{d,s}}\sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\eta-\tilde\eta| \\ &\ge (d/s)\mathbf{P}_{\pi,0}\left(\Lambda_\pi(Y_1)\ge\frac{1-p}{p}\right) + \mathbf{P}_{\pi,1}\left(\Lambda_\pi(Y_1)<\frac{1-p}{p}\right) \\ &=: A_\varepsilon + B_\varepsilon. \end{aligned} \tag{66}$$

where, under $\mathbf{P}_{\pi,\eta_1}$-probability with $\eta_1\in\{0,1\}$, the vector $Y_1 = (Y_{1,k})_{1\le|k|\le K_\varepsilon}$ has independent normal components:

$$Y_{1,k} = \eta_1 v^*_{1,k} + \xi_{1,k} \sim \mathcal{N}(\eta_1 v^*_{1,k}, 1), \quad 1\le|k|\le K_\varepsilon.$$

It now follows from (66) that the minimax risk R_ε is positive if at least one of the terms, A_ε or B_ε, is positive. Let us prove that for all sufficiently small ε the probability B_ε is separated from zero.

Recall that $d = d_\varepsilon\to\infty$ and $s = s_d = o(d)$ as $\varepsilon\to 0$. Put

$$H = H_\varepsilon = \log\left(\frac{1-p}{p}\right) \sim \log(d/s),$$

and introduce the random variable

$$\lambda_\pi(Y_1) := \log\Lambda_\pi(Y_1).$$

Using the notation $\mathbf{P}_0$ for $\mathbf{P}_{\pi,0}$, consider the probability measure $\mathbf{P}_h$, depending on a positive parameter $h = h_\varepsilon$, that is defined by the formula

$$\frac{d\mathbf{P}_h}{d\mathbf{P}_0}(Y_1) := \frac{\exp(h\lambda_\pi(Y_1))}{\Psi(h)}, \quad \Psi(h) = \mathbf{E}_{\mathbf{P}_0}\exp(h\lambda_\pi(Y_1)).$$

With the parameter $h>0$ chosen to satisfy

$$\mathbf{E}_{\mathbf{P}_h}\lambda_\pi(Y_1) = H,$$

we have (see Lemma 2 in [17])

$$h \sim \frac{1}{2} + \frac{H}{u_\varepsilon^2} = O(1), \tag{67}$$

and (see formula (45) in [17])

$$\Psi(h) = \exp\left(\frac{h^2 - h}{2}\, u_\varepsilon^2 (1 + o(1))\right), \tag{68}$$

where for notational simplicity we write u_ε^2 for $u_\varepsilon^2(r_\varepsilon)$.

We have

$$\begin{aligned} B_\varepsilon &= \mathbf{E}_{\pi,1}\left(\mathbb{I}\left\{\lambda_\pi(Y_1) < H\right\}\right) = \mathbf{E}_{\pi,0}\left(\exp(\lambda_\pi(Y_1)) \cdot \mathbb{I}\left\{\lambda_\pi(Y_1) < H\right\}\right) \\ &= \mathbf{E}_h\left(\frac{d\mathbf{P}_0}{d\mathbf{P}_h}(Y_1) \cdot \exp(\lambda_\pi(Y_1) \cdot \mathbb{I}\left\{\lambda_\pi(Y_1) < H\right\}\right) \\ &= \Psi(h)\mathbf{E}_h\left(\exp[(1-h)\lambda_\pi(Y_1)] \cdot \mathbb{I}\left\{\lambda_\pi(Y_1) < H\right\}\right). \end{aligned} \tag{69}$$

By Lemma 3 in [17], the standardized random variable

$$Z_h := \frac{\lambda_\pi(Y_1) - \mu_h}{\sigma_h},$$

where

$$\mu_h = \mathbf{E}_{\mathbf{P}_h}(\lambda_\pi(Y_1)) = u_\varepsilon^2(h - \frac{1}{2})(1+o(1)),\ \sigma_h^2 = \mathbf{Var}_{\mathbf{P}_h}(\lambda_\pi(Y_1)) = u_\varepsilon^2(1+o(1)),$$

converges in $\mathbf{P}_h$-distribution to an $\mathcal{N}(0,1)$. Therefore the statistic $\lambda_\pi(Y_1)$ on the right side of (69) is nearly a normal $\mathcal{N}(H, u_\varepsilon^2)$ random variable.

Next, by assumption and the 'continuity' of u_ε as stated in (17), for some constant $\delta_4 > 0$

$$u_\varepsilon / \sqrt{\log(d/s)} \le \sqrt{2}(1 - \delta_4),$$

provided ε is small enough. This and formula (67) give the inequality $1 - h < 0$, which implies for all $y \in \mathbb{R}^{2K_\varepsilon}$ and all sufficiently small ε

$$\exp[(1-h)\lambda_\pi(y)] \cdot \mathbb{I}\left\{\lambda_\pi(y) < H\right\} < \exp\left[(1-h)H\right] \sim (d/s)^{1-h} \le \text{const}.$$

Then, by the dominant convergence theorem, the replacement of $\lambda_\pi(Y_1)$ by an $\mathcal{N}(H, u_\varepsilon^2)$ on the right side (69) and the use of (67) and (68) yield for all sufficiently small ε

$$\begin{aligned} B_\varepsilon &\sim \exp\left(\frac{h^2 - h}{2} u_\varepsilon^2\right) \int_{-\infty}^{H} \exp\left[(1-h)x\right] \frac{1}{\sqrt{2\pi}u_\varepsilon} \exp\left(-\frac{(x-H)^2}{2u_\varepsilon^2}\right) dx \\ &= \exp\left(\frac{h^2 - h}{2} u_\varepsilon^2 + H(1-h) + \frac{(1-h)^2 u_\varepsilon^2}{2}\right) \end{aligned}$$

$$\times \int_{-\infty}^{H} \frac{1}{\sqrt{2\pi}u_\varepsilon} \exp\left(-\frac{\big(x-(H+(1-h)u_\varepsilon^2)\big)^2}{2u_\varepsilon^2}\right) dx$$

$$\sim \exp(0) \int_{-\infty}^{H} \frac{1}{\sqrt{2\pi}u_\varepsilon} \exp\left(-\frac{\big(x-(H+(1-h)u_\varepsilon^2)\big)^2}{2u_\varepsilon^2}\right) dx$$

$$\geq \int_{-\infty}^{H+(1-h)u_\varepsilon^2} \frac{1}{\sqrt{2\pi}u_\varepsilon} \exp\left(-\frac{\big(x-(H+(1-h)u_\varepsilon^2)\big)^2}{2u_\varepsilon^2}\right) dx = 1/2.$$

From this

$$\liminf_{\varepsilon\to 0} R_\varepsilon \geq \liminf_{\varepsilon\to 0} B_\varepsilon \geq 1/2 > 0,$$

and the proof of Theorem 4 is complete. □

Acknowledgment

The authors would like to thank the Associate Editor and anonymous referee for their helpful comments.

References

[1] Abramovich, F., De Feis, I. and Sapatinas, T. (2009). Optimal testing for additivity in multiple nonparametric regression. *Annals of the Insitute of Statistical Mathematics* **61** (3) 691–714. MR2529971

[2] Bernstein S. N. (1946). *Probability Theory*. OGIZ, Moscow–Leningrad. In Russian.

[3] Butucea, C., Stepanova, N. A. and Tsybakov, A. B. (2017). Variable selection with Hamming loss. *Annals of Statistics*, *to appear*.

[4] Chouldechova, A. and Hastie, T. (2015). Generalized additive model selection. http://arxiv.org/abs/1506.03850.

[5] Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Annals of Statistics* **40** (5) 2667–2696. MR3097616

[6] DeGroot, M. (1970). *Optimal Statistical Decisions*. McGraw-Hill Book Company, New York. MR0356303

[7] Donoho, D. (2006). For most large underdetermined systems of linear equations the minimal l^1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics* **59** (7) 907–934. MR2222440

[8] Ermakov, M. S. (1990). Minimax detection of a signal in a Gaussian white noise. *Theory of Probability and Its Applications* **35** (4) 667–679. MR1090496

[9] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of Royal Statistical Society, Series B* **70** 849–911. MR2530322

[10] Gayraud, G. and Ingster, Yu. I. (2012). Detection of sparse additive functions. *Electronic Journal of Statistics* **6** 1409–1448. MR2988453

[11] Genovese, C. R., Jin, J., Wasserman, L. and Yao, Z. (2012). A comparison of the lasso and marginal regression. *Journal of Machine Learning Research* **13** 2107–2143. MR2956354

[12] Golubev, Y. K. and Levit, B. Y. (1996). Asymptotically efficient estimation for analytic distributions. *Mathematical Methods of Statisics* **3** 357–368. MR1417678

[13] Huang. J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics* **38** 2282–2313. MR2676890

[14] Ingster, Yu. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Mathematical Methods of Statistics* **2** (2) 85–114. MR1257978

[15] Ingster, Yu. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. II. *Mathematical Methods of Statistics* **2** (3) 171–189. MR1257983

[16] Ingster, Yu. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. III. *Mathematical Methods of Statistics* **2** (4) 249–268. MR1259685

[17] Ingster, Yu. I. and Stepanova, N. A. (2014). Adaptive variable selection in nonparametric sparse regression. *Journal of Mathematical Sciences* **199** (2) 184–201. MR3032218

[18] Ingster, Yu. I. and Suslina, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Lecture Notes in Statistics, Vol. **169**, Springer-Verlag, New York. MR1991446

[19] Ingster, Yu. I. and Suslina, I. A. (2005). On estimation and detection of smooth function of many variables. *Mathematical Methods of Statistics* **14** 299–331. MR2195328

[20] Ingster, Yu. I. and Suslina, I. A. (2015). Detection of a sparse variable function. *Journal of Mathematical Sciences* **206** (2) 181–196. MR3373876

[21] Ji, O. and Jin, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Annals of Statistics* **40** (1) 73–103. MR3013180

[22] Lepski, O. V. (1991). One problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications* **35** (3), 454–466. MR1091202

[23] Meier, L., van de Geer, S. and Bühlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics* **37** 3779–3821. MR2572443

[24] Raskutti, G., Wainwright, M. J. and Yu, B. (2012). Minimax-optimal rates for high-dimensional sparse additive models over kernel classes. *Journal of Machine Learning Research* **13** 281–319. MR2913704

[25] Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2007). SpAM: sparse additive models. In: *Advances in Neural Information Processing Systems*, Vol. 20 (eds. J. C. Platt, D. Koller, Y. Singer, and S. Roweis), pp. 1202–1208, Cambridge, MA: MIT Press.