Tight conditions for consistency of variable selection in the context of high dimensionality

Laëtitia Comminges, Arnak S. Dalalyan

TIGHT CONDITIONS FOR CONSISTENCY OF VARIABLE SELECTION IN THE CONTEXT OF HIGH DIMENSIONALITY

By Laëtitia Comminges and Arnak S. Dalalyan*

Université Paris Est/ ENPC and ENSAE-CREST

We address the issue of variable selection in the regression model with very high ambient dimension, *i.e.*, when the number of variables is very large. The main focus is on the situation where the number of relevant variables, called intrinsic dimension and denoted by d^*, is much smaller than the ambient dimension d. Without assuming any parametric form of the underlying regression function, we get tight conditions making it possible to consistently estimate the set of relevant variables. These conditions relate the intrinsic dimension to the ambient dimension and to the sample size. The procedure that is provably consistent under these tight conditions is based on comparing quadratic functionals of the empirical Fourier coefficients with appropriately chosen threshold values.

The asymptotic analysis reveals the presence of two quite different regimes. The first regime is when d^* is fixed. In this case the situation in nonparametric regression is the same as in linear regression, *i.e.*, consistent variable selection is possible if and only if $\log d$ is small compared to the sample size n. The picture is different in the second regime, $d^* \to \infty$ as $n \to \infty$, where we prove that consistent variable selection in nonparametric set-up is possible only if $d^* + \log\log d$ is small compared to $\log n$. We apply these results to derive minimax separation rates for the problem of variable selection.

1. Introduction.   Real-world data such as those obtained from neuroscience, chemometrics, data mining, or sensor-rich environments are often extremely high-dimensional, severely underconstrained (few data samples compared to the dimensionality of the data), and interspersed with a large number of irrelevant or redundant features. Furthermore, in most situations the data is contaminated by noise making it even more difficult to retrieve useful information from the data. Relevant variable selection is a compelling approach for addressing statistical issues in the scenario of high-dimensional and noisy data with small sample size. Starting from Mallows [29], Akaike [1], Schwarz [36] who in-

troduced respectively the famous criteria C_p, AIC and BIC, the problem of variable selection was extensively studied in the statistical and machine learning literature both from the theoretical and algorithmic viewpoints. It appears, however, that the theoretical limits of performing variable selection in the context of nonparametric regression are still poorly understood, especially when the number of variables, denoted by d and referred to as ambient dimension, is much larger than the sample size n. The purpose of the present work is to explore this setting under the assumption that the number of relevant variables, hereafter called intrinsic dimension and denoted by d^*, may grow with the sample size but remains much smaller than d.

In the important particular case of linear regression, the latter scenario was the subject of a number of recent studies. Many of them rely on ℓ_1-norm penalization [38, 46, 31] and constitute an attractive alternative to iterative variable selection procedures [2, 45] and to marginal regression or correlation screening [42, 18]. Promising results for feature selection are also obtained by conformal prediction [20], (minimax) concave penalties [16, 17, 44], Bayesian approach [37] and higher criticism [15]. Extensions to other settings including logistic regression, generalized linear model and Ising model were carried out in [8, 34, 18], respectively. Variable selection in the context of groups of variables with disjoint or overlapping groups was studied by [43, 24, 28, 32, 21]. Hierarchical procedures for selection of relevant variables were proposed by [3, 5, 47].

It is now well understood that in the Gaussian sequence model and in the high-dimensional linear regression with a Gram matrix satisfying some variant of irrepresentable condition, consistent estimation of the pattern of relevant variables—also called the sparsity pattern—is possible under the condition $d^* \log(d/d^*) = o(n)$ as $n \to \infty$ [41]. Furthermore, it is well known that if $(d^* \log(d/d^*))/n$ remains bounded from below by some positive constant when $n \to \infty$, then it is impossible to consistently recover the sparsity pattern [40]. Thus, a tight condition exists that describes in an exhaustive manner the interplay between the quantities d^*, d and n that guarantees the existence of consistent estimators. The situation is very different in the case of non-linear regression, since, to our knowledge, there is no result providing tight conditions for consistent estimation of the sparsity pattern.

The papers [26] and [4], closely related to the present work, considered the problem of variable selection in nonparametric Gaussian regression model. They proved the consistency of the proposed procedures under some assumptions that—in the light of the present work—turn out to be suboptimal. More precisely, Lafferty and Wasserman [26] assumed the unknown regression function to be four times continuously differentiable with bounded derivatives. The algo-

rithm they proposed, termed Rodeo, is a greedy procedure performing simultaneously local bandwidth choice and variable selection. Rodeo is shown to converge when the ambient dimension d is $O(\log n/\log\log n)$ while the intrinsic dimension d^* does not increase with n. On the other hand, Bertin and Lecué [4] proposed a procedure based on the ℓ_1-penalization of local polynomial estimators and proved its consistency when $d^* = O(1)$ but d is allowed to be as large as $\log n$, up to a constant. They also have a weaker assumption on the regression function merely assumed to belong to the Holder class with smoothness $\beta > 1$. To complete the picture, let us mention that estimation and hypotheses testing problems for high-dimensional nonparametric regression under sparse additive modeling were recently addressed in [25, 33, 19].

This brief review of the literature reveals that there is an important gap in consistency conditions for the linear regression and for the non-linear one. For instance, if the intrinsic dimension d^* is fixed, then the condition guaranteeing consistent estimation of the sparsity pattern is $(\log d)/n \to 0$ in linear regression whereas it is $d = O(\log n)$ in the nonparametric case. While it is undeniable that the nonparametric regression is much more complex than the linear one, it is however not easy to find a justification to such an important gap between two conditions. The situation is even worse in the case where $d^* \to \infty$. In fact, for the linear model with at most polynomially increasing ambient dimension $d = O(n^k)$, it is possible to estimate the sparsity pattern for intrinsic dimensions d^* as large as $n^{1-\epsilon}$, for some $\epsilon > 0$. In other words, the sparsity index can be almost on the same order as the sample size. In contrast, in nonparametric regression, there is no procedure that is proved to converge to the true sparsity pattern when both n and d^* tend to infinity, even if d^* grows extremely slowly.

In the present work, we fill this gap by introducing a simple variable selection procedure that selects the relevant variables by comparing some quadratic functionals of empirical Fourier coefficients to prescribed significance levels. Consistency of this procedure is established under some conditions on the triplet (d^*, d, n) and the tightness of these conditions is proved. The main take-away messages deduced from our results are the following:

- When the number of relevant variables d^* is fixed and the sample size n tends to infinity, there exist positive real numbers c_* and c^* such that (a) if $(\log d)/n \leq c_*$ the estimator proposed in Section 3 is consistent and (b) no estimator of the sparsity pattern may be consistent if $(\log d)/n \geq c^*$.
- When the number of relevant variables d^* tends to infinity with $n \to \infty$, then there exist real numbers $\underline{c}_i$ and $\bar{c}_i$, $i = 1, 2$ such that $\underline{c}_1 > 0$, $\bar{c}_1 > 0$ and (a) if $\underline{c}_1 d^* + \log\log(d/d^*) - \log n < \underline{c}_2$ the estimator proposed in Section 3 is consistent and (b) no estimator of the sparsity pattern may be consistent

if $\bar{c}_1 d^* + \log\log(d/d^*) - \log n > \bar{c}_2$.

- In particular, if d grows not faster than a polynomial in n, then there exist positive real numbers c_0 and c^0 such that (a) if $d^* \leq c_0 \log n$ the estimator proposed in Section 3 is consistent and (b) no estimator of the sparsity pattern may be consistent if $d^* \geq c^0 \log n$.

In the regime of a growing intrinsic dimension $d^* \to \infty$ and a moderately large ambient dimension $d = O(n^C)$, for some $C > 0$, we make a concentrated effort to get the constant c_0 as close as possible to the constant c^0. This goal is reached for the model of Gaussian white noise and, very surprisingly, it required from us to apply some tools from complex analysis, such as the Jacobi θ-function and the saddle point method, in order to evaluate the number of lattice points lying in a ball of an Euclidean space with increasing dimension.

The rest of the paper is organized as follows. The notation and assumptions necessary for stating our main results are presented in Section 2. In Section 3, an estimator of the set of relevant variables is introduced and its consistency is established, in the case where the data come from the Gaussian white noise model. The main condition required in the consistency result involves the number of lattice points in a ball of a high-dimensional Euclidean space. An asymptotic equivalent for this number is presented in Section 4. Results on impossibility of consistent estimation of the sparsity pattern are derived in Section 5. Section 6 is devoted to exploring adaptation to the unknown parameters (smoothness and degree of significance) and recovering minimax rates of separation. Then, in Section 7, we show that some of our results can be extended to the model of nonparametric regression. The relations between consistency and inconsistency results are discussed in Section 8. The technical parts of the proofs are postponed to the Appendix.

2. The problem formulation and the assumptions. We are interested in the variable selection task (also known as model selection, feature selection, sparsity pattern estimation) in the context of high-dimensional non-linear regression. Let $f : [0,1]^d \to \mathbb{R}$ denote the unknown regression function. We assume that the number of variables d is very large, possibly much larger than the sample size n, but only a small number of these variables contribute to the fluctuations of the regression function f.

To be more precise, we assume that for some small subset J of the index set $\{1, \ldots, d\}$ satisfying $\mathrm{Card}(J) \leq d^*$, there is a function $\bar{f} : \mathbb{R}^{\mathrm{Card}(J)} \to \mathbb{R}$ such that

$$f(\boldsymbol{x}) = \bar{f}(\boldsymbol{x}_J), \qquad \forall \boldsymbol{x} \in \mathbb{R}^d,$$

where $\boldsymbol{x}_J$ stands for the subvector of $\boldsymbol{x}$ obtained by removing from $\boldsymbol{x}$ all the coordinates with indices lying outside J. In what follows, we allow d and d^* to

depend on n but we will not always indicate this dependence in notation. Note also that the genuine intrinsic dimension is $\mathrm{Card}(J)$; d^* is merely a known upper bound on the intrinsic dimension. In what follows, we use the standard notation for the vector and sequence norms:

$$\|\boldsymbol{x}\|_0 = \sum_j \mathbf{1}(x_j \neq 0), \qquad \|\boldsymbol{x}\|_p^p = \sum_j |x_j|^p, \ \forall p \in [1,\infty), \qquad \|\boldsymbol{x}\|_\infty = \sup_j |x_j|,$$

for every $\boldsymbol{x} \in \mathbb{R}^d$ or $\boldsymbol{x} \in \mathbb{R}^\mathbb{N}$.

Let us stress right away that the primary aim of this work is to understand when it is possible to estimate the sparsity pattern J (with theoretical guarantees on the convergence of the estimator) and when it is impossible. The estimator that we will define in next sections is intended to show the possibility of consistent estimation, rather than to provide a practical procedure for recovering the sparsity pattern. Therefore, the estimator will be allowed to depend on different constants appearing in conditions imposed on the regression function f and on some characteristics of the noise.

To make the consistent estimation of the set J realizable, we impose some smoothness and identifiability assumptions on f. In order to describe the smoothness assumption imposed on f, let us introduce the trigonometric Fourier basis: $\varphi_{\mathbf{0}} \equiv 1$ and

$$\varphi_{\boldsymbol{k}}(\boldsymbol{x}) = \begin{cases} \sqrt{2}\cos(2\pi\,\boldsymbol{k}\cdot\boldsymbol{x}), & \boldsymbol{k} \in (\mathbb{Z}^d)_+, \\ \sqrt{2}\sin(2\pi\,\boldsymbol{k}\cdot\boldsymbol{x}), & -\boldsymbol{k} \in (\mathbb{Z}^d)_+, \end{cases} \tag{1}$$

where $(\mathbb{Z}^d)_+$ denotes the set of all $\boldsymbol{k} \in \mathbb{Z}^d \setminus \{0\}$ such that the first nonzero element of $\boldsymbol{k}$ is positive and $\boldsymbol{k}\cdot\boldsymbol{x}$ stands for the usual inner product in $\mathbb{R}^d$. In what follows, we use the notation $\langle\cdot,\cdot\rangle$ for designing the scalar product in $L^2([0,1]^d;\mathbb{R})$, that is $\langle\mathsf{h},\tilde{\mathsf{h}}\rangle = \int_{[0,1]^d} \mathsf{h}(\boldsymbol{x})\tilde{\mathsf{h}}(\boldsymbol{x})\,d\boldsymbol{x}$ for every $\mathsf{h},\tilde{\mathsf{h}} \in L^2([0,1]^d;\mathbb{R})$. Using this orthonormal Fourier basis, we define

$$\Sigma_L = \left\{ f : \sum_{\boldsymbol{k}\in\mathbb{Z}^d} k_j^2 \langle\mathsf{f},\varphi_{\boldsymbol{k}}\rangle^2 \leq L; \quad \forall j \in \{1,\ldots,d\} \right\}.$$

To ease notation, we set $\theta_{\boldsymbol{k}}[\mathsf{f}] = \langle\mathsf{f},\varphi_{\boldsymbol{k}}\rangle$ for all $\boldsymbol{k} \in \mathbb{Z}^d$. In addition to the smoothness, we need also to require that the relevant variables are sufficiently relevant for making their identification possible. This is done by means of the following condition.

[**C1**(κ, L)] The regression function f belongs to Σ_L. Furthermore, for some subset $J \subset \{1,\ldots,d\}$ of cardinality $\leq d^*$, there exists a function $\bar{\mathsf{f}} : \mathbb{R}^{\mathrm{Card}(J)} \to \mathbb{R}$ such that $\mathsf{f}(\boldsymbol{x}) = \bar{\mathsf{f}}(\boldsymbol{x}_J)$, $\forall\boldsymbol{x} \in \mathbb{R}^d$ and it holds that

$$Q_j[\mathsf{f}] = \sum_{\boldsymbol{k}:k_j\neq 0} \theta_{\boldsymbol{k}}[\mathsf{f}]^2 \geq \kappa, \ \forall j \in J. \tag{2}$$

One easily checks that $Q_j[\mathsf{f}] = 0$ for every j that does not lie in the sparsity pattern. This provides a characterization of the sparsity pattern as the set of indices of nonzero coefficients of the vector $\boldsymbol{Q}[\mathsf{f}] = (Q_1[\mathsf{f}], \dots, Q_d[\mathsf{f}])$.

Prior to describing the procedures for estimating J, let us comment Condition [**C1**]. It is important to note that the identifiability assumption (2) can be rewritten as $\int_{[0,1]^d} \left(\mathsf{f}(\boldsymbol{x}) - \int_0^1 \mathsf{f}(\boldsymbol{x}) \, dx_j\right)^2 d\boldsymbol{x} \geq \kappa$ and, therefore, is not intrinsically related to the basis we have chosen. In the case of continuously differentiable and 1-periodic function f, the smoothness assumption $\mathsf{f} \in \Sigma_L$ as well can be rewritten without using the trigonometric basis, since $\sum_{\boldsymbol{k} \in \mathbb{Z}^d} k_j^2 \theta_{\boldsymbol{k}}[\mathsf{f}]^2 = (2\pi)^{-2} \int_{[0,1]^d} [\partial_j \mathsf{f}(\boldsymbol{x})]^2 \, d\boldsymbol{x}$. Thus, condition [**C1**] is essentially a constraint on the function f itself and not on its representation in the specific basis of trigonometric functions.

The results of this work can be extended with minor modifications to other types of smoothness conditions imposed on f, such as Hölder continuity or Besov-regularity. In these cases the trigonometric basis (1) should be replaced by a basis adapted to the smoothness condition (spline, wavelet, etc.). Furthermore, even in the case of Sobolev smoothness, one can replace the set Σ_L corresponding to smoothness order 1 by any Sobolev ellipsoid of smoothness $\beta > 0$, see for instance [10] where the case $\beta = 2$ is explored. Roughly speaking, the role of the smoothness assumption is to reduce the statistical model with infinite-dimensional parameter f to a finite-dimensional model having good approximation properties. Any value of smoothness order $\beta > 0$ leads to this reduction. The value $\beta = 1$ is chosen for simplicity of exposition only.

3. Idealized setup: Gaussian white noise model.   To convey the main ideas without taking care of some technical details, we start by focusing our attention on the Gaussian white noise model, that was proved to be asymptotically equivalent to the model of regression [6, 35], as well as to other nonparametric models [7, 13]. Thus, we assume that the available data consists of Gaussian process $\{\boldsymbol{Y}(\phi) : \phi \in L^2([0,1]^d; \mathbb{R})\}$ such that

$$\mathbb{E}_{\mathsf{f}}[\boldsymbol{Y}(\phi)] = \int_{[0,1]^d} \mathsf{f}(\boldsymbol{x})\phi(\boldsymbol{x}) \, d\boldsymbol{x}, \qquad \mathbf{Cov}_{\mathsf{f}}(\boldsymbol{Y}(\phi), \boldsymbol{Y}(\phi')) = \frac{1}{n} \int_{[0,1]^d} \phi(\boldsymbol{x})\phi'(\boldsymbol{x}) \, d\boldsymbol{x}.$$

It is well-known that these two properties uniquely characterize the probability distribution of a Gaussian process. An alternative representation of $\boldsymbol{Y}$ is

$$d\boldsymbol{Y}(\boldsymbol{x}) = \mathsf{f}(\boldsymbol{x}) \, d\boldsymbol{x} + n^{-1/2} dW(\boldsymbol{x}), \qquad \boldsymbol{x} \in [0,1]^d,$$

where $W(\boldsymbol{x})$ is a d-parameter Brownian sheet. Note that minimax estimation and detection of the function f in this set-up (but without sparsity assumption) was studied by [23].

3.1. *Estimation of J by multiple hypotheses testing.* We intend to tackle the variable selection problem by multiple hypotheses testing; each hypothesis concerns a group of the Fourier coefficients of the observed signal and suggests that all the elements within the group are zero. The rationale behind this approach is the following simple observation: since the trigonometric basis is orthonormal and contains the constant function,

$$j \notin J \iff \theta_{\boldsymbol{k}}[\mathrm{f}] = \langle f, \varphi_{\boldsymbol{k}} \rangle = 0, \ \forall \boldsymbol{k} \ \text{s.t.} \ k_j \neq 0. \tag{3}$$

This observation entails that if the intrinsic dimension $|J|$ is small as compared to d, then the sequence of Fourier coefficients is sparse. Furthermore, as explained below, there is a sort of group sparsity with overlapping groups.

For every $\ell \in \{1, \ldots, d^*\}$, we denote by P_ℓ^d the set of all subsets I of $\{1, \ldots, d\}$ having exactly ℓ elements: $P_\ell^d = \{I \subset \{1, \ldots, d\} \ : \ \mathrm{Card}(I) = \ell\}$. For every multi-index $\boldsymbol{k} \in \mathbb{Z}^d$, we denote by $\mathrm{supp}(\boldsymbol{k})$ the set of indices corresponding to nonzero entries of $\boldsymbol{k}$. To define the blocks of coefficients $\theta_{\boldsymbol{k}}$ that will be tested for significance, we introduce the following notation: for every $I \subset \{1, \ldots, d\}$ and for every $j \in I$, we set

$$V_I^j[\mathrm{f}] = \Big(\theta_{\boldsymbol{k}}[\mathrm{f}] : j \in \mathrm{supp}(\boldsymbol{k}) \subset I \Big).$$

It follows from (3) that the characterization

$$j \notin J \iff \max_I \big\| V_I^j[\mathrm{f}] \big\|_p = 0, \tag{4}$$

holds true for every $p \in [0, +\infty]$. Furthermore, again in view of (3), the maximum over I of the norms $\big\| V_I^j[\mathrm{f}] \big\|_p$ is attained when $I = J$ and is equal to the maximum over all subsets I such that $\mathrm{Card}(I) \leq d^*$. Summarizing these arguments, we can formulate the problem of variable selection as a problem of testing d null hypotheses

$$H_{0j}: \quad \big\| V_I^j[\mathrm{f}] \big\|_p = 0 \quad \forall I \subset \{1, \ldots, d\} \text{ such that } \mathrm{Card}(I) \leq d^*. \tag{5}$$

If the hypothesis H_{0j} is rejected, then the jth covariate is declared as relevant. Note that by virtue of assumption [**C1**], the alternatives can be written as

$$H_{1j}: \quad \big\| V_I^j[\mathrm{f}] \big\|_2^2 \geq \kappa \quad \text{for some } I \subset \{1, \ldots, d\} \text{ such that } \mathrm{Card}(I) \leq d^*. \tag{6}$$

Our estimator is based on this characterization of the sparsity pattern. If we denote by $y_{\boldsymbol{k}}$ the observable random variable $\boldsymbol{Y}(\varphi_{\boldsymbol{k}})$, we have

$$y_{\boldsymbol{k}} = \theta_{\boldsymbol{k}}[\mathrm{f}] + n^{-1/2} \xi_{\boldsymbol{k}}, \qquad \theta_{\boldsymbol{k}} = \langle f, \varphi_{\boldsymbol{k}} \rangle, \quad \boldsymbol{k} \in \mathbb{Z}^d, \tag{7}$$

where $\{\xi_{\boldsymbol{k}}; \boldsymbol{k} \in \mathbb{Z}^d\}$ form a countable family of independent Gaussian random variables with zero mean and variance equal to one. According to this property,

$y_{\mathbf{k}}$ is a good estimate of $\theta_{\mathbf{k}}[\mathsf{f}]$: it is unbiased and with a mean squared error equal to $1/n$. Using the plug-in argument, this suggests to estimate V_I^j by $\widehat{V}_I^j = (y_{\mathbf{k}} : j \in \mathrm{supp}(\mathbf{k}) \subset I)$ and the norm of V_I^j by the norm of $\widehat{V}_I^j$. However, since this amounts to estimating an infinite-dimensional vector, the error of estimation will be infinitely large. To cope with this issue, we restrict the set of indices for which $\theta_{\mathbf{k}}$ is estimated by $y_{\mathbf{k}}$ to a finite set, outside of which $\theta_{\mathbf{k}}$ will be merely estimated by 0. Such a restriction is justified by the fact that f is assumed to be smooth: Fourier coefficients corresponding to very high frequencies are very small.

Let us fix an integer $m > 0$, the cut-off level, and denote, for $j \in I \subset \{1, \ldots, d\}$,

$$S_{m,I}^j = \left\{ \mathbf{k} \in \mathbb{Z}^d : \|\mathbf{k}\|_2 \le m \text{ and } \{j\} \subset \mathrm{supp}(\mathbf{k}) \subset I \right\}.$$

Since the alternatives H_{1j} are concerned with the 2-norm, we build our test statistic on an estimate of the norm $\|V_I^j[\mathsf{f}]\|_2$. To this end, we introduce

$$\widehat{Q}_{m,I}^j = \sum_{\mathbf{k} \in S_{m,I}^j} \left(y_{\mathbf{k}}^2 - \frac{1}{n} \right),$$

which is an unbiased estimator of $Q_{m,I}^j = \sum_{\mathbf{k} \in S_{m,I}^j} \theta_{\mathbf{k}}^2$. Note that when $m \to \infty$, the quantity $Q_{m,I}^j$ approaches $\|V_I^j[\mathsf{f}]\|_2^2$. It is clear that larger values of m lead to a smaller bias while the variance get increased. Moreover, the variance of $\widehat{Q}_{m,I}^j$ is proportional to the cardinality of the set $S_{m,I}^j$. The latter is an increasing function of $\mathrm{Card}(I)$. Therefore, if we aim at getting comparable estimation accuracies when estimating the functionals $\|V_I^j[\mathsf{f}]\|_2^2$ by $\widehat{Q}_{m,I}^j$ for various I's, it is reasonable to make the cut-off level m vary with the cardinality of I.

Thus, we consider a multivariate cut-off $\mathbf{m} = (m_1, \ldots, m_{d^*}) \in \mathbb{N}^{d^*}$. For a subset I of cardinality $\ell \le d^*$, we test significance of the vector $V_I^j[\mathsf{f}]$ by comparing its estimate $\widehat{Q}_{m_\ell,I}^j$ with a prescribed threshold λ_ℓ. This leads us to define an estimator of the set J by

$$\widehat{J}_n(\mathbf{m}, \boldsymbol{\lambda}) = \left\{ j \in \{1, \ldots, d\} : \max_{\ell \le d^*} \lambda_\ell^{-1} \max_{I \in P_\ell^d} \widehat{Q}_{m_\ell,I}^j \ge 1 \right\}.$$

where $\mathbf{m} = (m_1, \ldots, m_{d^*}) \in \mathbb{N}^{d^*}$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{d^*}) \in \mathbb{R}_+^{d^*}$ are two vectors of tuning parameters. As already mentioned, the role of $\mathbf{m}$ is to ensure that the truncated sums $Q_{m,J}^j$ do not deviate too much from the complete sums Q_J^j. Quantitatively speaking, for a given $\tau > 0$, we would like to choose m_ℓ's so that $Q_{m_s,J}^j \ge \kappa \tau / \tau + 1$, where $s = \mathrm{Card}(J)$. This guarantee can be achieved due to the smoothness assumption. Indeed, as proved in (26) (cf. Appendix B), it holds that

$$Q_{m_s,J}^j \ge \kappa - m_s^{-2} Ls, \qquad \forall j \in J.$$

Therefore, choosing $m_\ell = \left(\ell L(1+\tau)/\kappa\right)^{1/2}$, for every $\ell = 1,\ldots,d^*$, entails the inequality $Q_{m_s,J}^j \geq \kappa\tau/\tau + 1$, which indicates that the relevance of variables is not affected too much by the truncation.

Pushing further the analogy with the hypotheses testing, we define Type I error of an estimator $\widehat{J}_n$ of J as the one of having $\widehat{J}_n \not\subset J$, *i.e.*, classifying some irrelevant variables as relevant. The Type II error is then that of having $J \not\subset \widehat{J}$, which amounts to classifying some relevant variables as irrelevant. As in testing problem, handling the Type I error is easier since the distribution of the test statistic is independent of f. In fact, this is the max of a finite family of random variables drawn from translated and scaled χ^2-distributions. Using the Bonferroni adjustment, leads to the following control of the first kind error.

PROPOSITION 1. *Let us denote by $N(\ell,\gamma)$ the cardinality of the set $\{k \in \mathbb{Z}^\ell :$ $\|k\|_2^2 \leq \gamma\ell \ \& \ k_1 \neq 0\}$. If for some $A > 1$ and for every $\ell = 1,\ldots,d^*$,*

$$\lambda_\ell \geq \frac{2\sqrt{AN(\ell,m_\ell^2/\ell)d^* \log(2ed/d^*)} + 2Ad^* \log(2ed/d^*)}{n}, \tag{8}$$

then the Type I error $\mathbf{P}\left(\widehat{J}_n(m,\lambda) \not\subset J\right)$ is upper-bounded by $(2ed/d^*)^{-d^*(A-1)}$, and therefore tends to 0 as $d \to +\infty$.

This proposition shows that the Type I error of a variable selection procedure may be made small by choosing a sufficiently high threshold. By doing this, we run the risk to reject H_{0j} very often and to drastically underestimate the set of relevant variables. The next result establishes a necessary condition, which will be shown to be tight, ensuring that such an underestimation does not occur.

THEOREM 1. *Let condition $[\mathbf{C1}(\kappa,L)]$ be satisfied with some known constants $\kappa > 0$ and $L < \infty$ and let $s = \mathrm{Card}(J)$. For some real numbers $\tau > 0$ and $A > 1$, set $m_\ell = \left(\ell L(1+\tau)/\kappa\right)^{1/2}$, $\ell = 1,\ldots,d^*$, and define λ_ℓ to be equal to the right-hand side of (8). If the condition*

$$4\lambda_s \leq \kappa\tau/(1+\tau) \tag{9}$$

is fulfilled, then $\widehat{J}_n(m,\lambda)$ is consistent and satisfies the inequalities $\mathbf{P}\left(\widehat{J}_n(m,\lambda) \not\supset J\right) \leq 2(2ed/d^*)^{-d^*(A-1)}$ and $\mathbf{P}\left(\widehat{J}_n(m,\lambda) \neq J\right) \leq 3(2ed/d^*)^{-d^*(A-1)}$.

Condition (9) ensuring the consistency of the variable selection procedure $\widehat{J}_n$ admits a very natural interpretation: It is possible to detect relevant variables if the degree of relevance κ is larger than a multiple of the threshold λ_s, the latter being chosen according to the noise level.

A first observation is that this theorem provides interesting insight to the possibility of consistent recovery of the sparsity pattern J in the context of fixed intrinsic dimension. In fact, when d^* remains bounded from above when $n \to \infty$ and $d \to \infty$, then we get that $\mathbf{P}(\widehat{J}_1(m,\lambda) = J) \to_{n,d\to\infty} 1$ provided that

$$\log d \le \text{Const} \cdot n. \tag{10}$$

Although we did not find (exactly) this result in the statistical literature on variable selection, it can be checked that (10) is a necessary and sufficient condition for recovering the sparsity pattern J in linear regression with fixed sparsity d^* and growing dimension d and sample size n. Thus, in the regime of fixed or bounded d^*, the sparsity pattern estimation in nonparametric regression is not more difficult than in the parametric linear regression, as far as only the consistency of estimation is considered and the precise value of the constant in (10) is neglected. Furthermore, there is a simple estimator $\widehat{J}_n^{(1)}$ of J (cf. Eq. (3) in [10]), which is provably consistent under condition (10). This estimator can be seen as a procedure of testing hypotheses H_{0j} of form (5) with $p = \infty$ and, therefore, it does not really exploit the structure of the Fourier coefficients of the regression function. To some extent, this is the reason why in the regime of growing intrinsic dimension $d^* \to \infty$, the estimator $\widehat{J}_n^{(1)}$ proposed by [10] is no longer optimal.

In fact, when $d^* \to \infty$, the term $N(s, m_s^2/s)$ present in (9) tends to infinity as well. Furthermore, as we show in Section 4, this convergence takes place at an exponential rate in d^*. Therefore, in this asymptotic set-up it is crucial to have the right order of $N(s, m_s^2/s)$ in the condition that ensures the consistency. As shown in Section 5, this is the case for condition (9).

REMARK 1. An apparent drawback of the estimator $\widehat{J}_n$ is the large dimensionality of tuning parameters involved in $\widehat{J}_n$. However, Theorem 1 reveals that for achieving good selection power, it is sufficient to select the $2d^*$-dimensional tuning parameter $(\boldsymbol{m}, \boldsymbol{\lambda})$ on a one-dimensional curve parameterized by $\vartheta = L(1 + \tau)/\kappa$. Indeed, once the value of ϑ is given, Thm. 1 advocates for choosing

$$m_\ell = (\ell\vartheta)^{1/2} \quad \text{and} \quad \lambda_\ell = \frac{2\sqrt{AN(\ell,\vartheta)d^*\log(2ed/d^*)} + 2Ad^*\log(2ed/d^*)}{n} \tag{11}$$

for every $\ell = 1,\ldots,d^*$. As discussed in Section 6.1, this property allows to relax the requirement that the values L and κ involved in [**C1**] are known in advance.

REMARK 2. The result of the last theorem is in some sense adaptive w.r.t. the unknown sparsity. Indeed, while the estimator $\widehat{J}_n$ involves d^*, which is merely a known upper bound on the true sparsity $s = \text{Card}(J)$ and may be significantly

larger than s, it is the true sparsity s that appears in condition (9) as a first argument of the quantity $N(\cdot, \vartheta)$. This point is important given the exponential rate of divergence of $N(\cdot, \vartheta)$ when its first argument tends to infinity. On the other hand, if condition (9) is satisfied with $N(d^*, \vartheta)$ instead of $N(\text{Card}(J), \vartheta)$, then the consistent estimation of J can be achieved by a slightly simpler procedure:

$$\widetilde{J}_n(\boldsymbol{m}, \boldsymbol{\lambda}) = \left\{ j \in \{1, \ldots, d\} \; : \quad \max_{I \in P_{d^*}^d} \widehat{Q}_{m_{d^*}, I}^j \geq \lambda_{d^*} \right\}.$$

The proof of this statement is similar to that of Thm. 1 and will be omitted.

4. Counting lattice points in a ball. The aim of the present section is to investigate the properties of the quantity $N(d^*, \gamma)$ that is involved in the conditions ensuring the consistency of the proposed procedures. Quite surprisingly, the asymptotic behavior of $N(d^*, \gamma)$ turns out to be related to the Jacobi θ-function. To show this, let us introduce some notation. For a positive number γ, we set

$$\mathscr{C}_1(d^*, \gamma) = \left\{ \boldsymbol{k} \in \mathbb{Z}^{d^*} : k_1^2 + \ldots + k_{d^*}^2 \leq \gamma d^* \right\}, \quad \mathscr{C}_2(d^*, \gamma) = \left\{ \boldsymbol{k} \in \mathscr{C}_1(d^*, \gamma) : k_1 = 0 \right\}$$

along with $N_1(d^*, \gamma) = \text{Card}\,\mathscr{C}_1(d^*, \gamma)$ and $N_2(d^*, \gamma) = \text{Card}\,\mathscr{C}_2(d^*, \gamma)$. In simple words, $N_1(d^*, \gamma)$ is the number of lattice points lying in the d^*-dimensional ball with radius $(\gamma d^*)^{1/2}$ and centered at the origin, while $N_2(d^*, \gamma)$ is the number of (integer) lattice points lying in the $(d^* - 1)$-dimensional ball with radius $(\gamma d^*)^{1/2}$ and centered at the origin. With this notation, the quantity $N(\ell, \cdot)$ of Theorem 1 can be written as $N_1(\ell, \cdot) - N_2(\ell, \cdot)$. By volumetric arguments, one can check that $V(d^*)(\sqrt{\gamma} - 1)^{d^*}(d^*)^{d^*/2} \leq N_1(d^*, \gamma) \leq V(d^*)(\sqrt{\gamma} + 1)^{d^*}(d^*)^{d^*/2}$, where $V(d^*) = \pi^{d^*/2}/\Gamma(1 + d^*/2)$ is the volume of the unit ball in $\mathbb{R}^{d^*}$. Furthermore, similar bounds hold true for $N_2(d^*, \gamma)$ as well. Unfortunately, when $d^* \to \infty$, these inequalities are not accurate enough to yield non-trivial results in the problem of variable selection we are dealing with. This is especially true for the results on impossibility of consistent estimation stated in Section 5.

In order to determine the asymptotic behavior of $N_1(d^*, \gamma)$ and $N_2(d^*, \gamma)$ when d^* tends to infinity, we will rely on their integral representation through Jacobi's θ-function. Recall that the latter is given by $\mathsf{h}(z) = \sum_{r \in \mathbb{Z}} z^{r^2}$, which is well defined for any complex number z belonging to the unit ball $|z| < 1$. To briefly explain where the relation between $N_i(d^*, \gamma)$ and the θ-function comes from, let us denote by $\{a_r\}$ the sequence of coefficients of the power series of $\mathsf{h}(z)^{d^*}$, that is $\mathsf{h}(z)^{d^*} = \sum_{r \geq 0} a_r z^r$. One easily checks that $\forall r \in \mathbb{N}$, $a_r = \text{Card}\{\boldsymbol{k} \in \mathbb{Z}^{d^*} : k_1^2 + \ldots + k_{d^*}^2 = r\}$. Thus, for every γ such that γd^* is integer, we have $N_1(d^*, \gamma) = \sum_{r=0}^{\gamma d^*} a_r$. As a consequence of Cauchy's theorem, we get :

$$N_1(d^*, \gamma) = \frac{1}{2\pi i} \oint \frac{\mathsf{h}(z)^{d^*}}{z^{\gamma d^*}} \frac{dz}{z(1 - z)}.$$
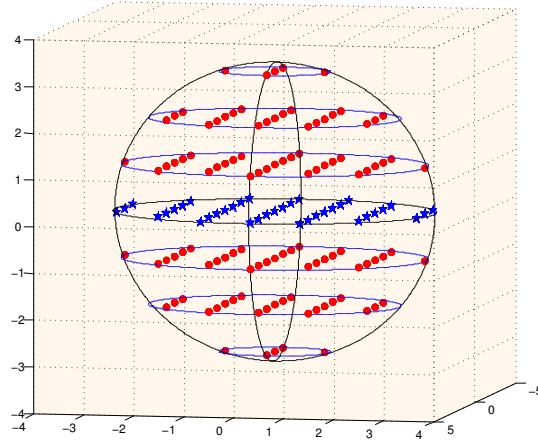
FIG 1. *Lattice points in a ball of radius $R = \gamma d^* = 3.2$ in the three dimensional space ($d^* = 3$). Red points are those of $\mathscr{C}_1(d^*,\gamma) \setminus \mathscr{C}_2(d^*,\gamma)$ while blue stars are those of $\mathscr{C}_2(d^*,\gamma)$. In this example, $N(d^*,\gamma) = N(3,1.07) = 110$.*

where the integral is taken over any circle $|z| = w$ with $0 < w < 1$. Exploiting this representation and applying the saddle-point method thoroughly described in [14], we get the following result.

PROPOSITION 2.   *Let $\gamma > 0$ be an integer and let $\mathsf{l}_\gamma(z) = \log \mathsf{h}(z) - \gamma \log z$.*

1. *There is a unique solution z_γ in $(0,1)$ to the equation $\mathsf{l}'_\gamma(z) = 0$. Furthermore, the function $\gamma \mapsto z_\gamma$ is increasing and $\mathsf{l}''_\gamma(z) > 0$.*
2. *For $i = 1, 2$, the following equivalences hold true:*

$$N_i(d^*,\gamma) = \left(\frac{\mathsf{h}(z_\gamma)}{z_\gamma^\gamma}\right)^{d^*} \frac{1 + o(1)}{\mathsf{h}(z_\gamma)^{i-1} z_\gamma (1 - z_\gamma)(2\mathsf{l}''_\gamma(z_\gamma)\pi d^*)^{1/2}},$$

as d^* tends to infinity.

Hereafter, it will be useful to note that the second part of Prop. 2 yields

$$\log\left(N_1(d^*,\gamma) - N_2(d^*,\gamma)\right) = d^* \mathsf{l}_\gamma(z_\gamma) - \frac{1}{2}\log d^* + c_\gamma + o(1), \quad \text{as } d^* \to \infty, \quad (12)$$

with $c_\gamma = \log\left(\frac{\mathsf{h}(z_\gamma)-1}{\mathsf{h}(z_\gamma)z_\gamma(1-z_\gamma)\sqrt{2\pi\mathsf{l}''_\gamma(z_\gamma)}}\right)$. Furthermore, while the asymptotic equivalences of Prop. 2 are established for integer values of $\gamma > 0$, relation $\log\left(N_1(d^*,\gamma) - N_2(d^*,\gamma)\right) = d^*\mathsf{l}_\gamma(z_\gamma)(1 + o(1))$ holds true for any positive real number γ [30]. In order to get an idea of how the terms z_γ and $\mathsf{l}_\gamma(z_\gamma)$ depend on γ, we depicted in Fig. 2 the plots of these quantities as functions of $\gamma > 0$.

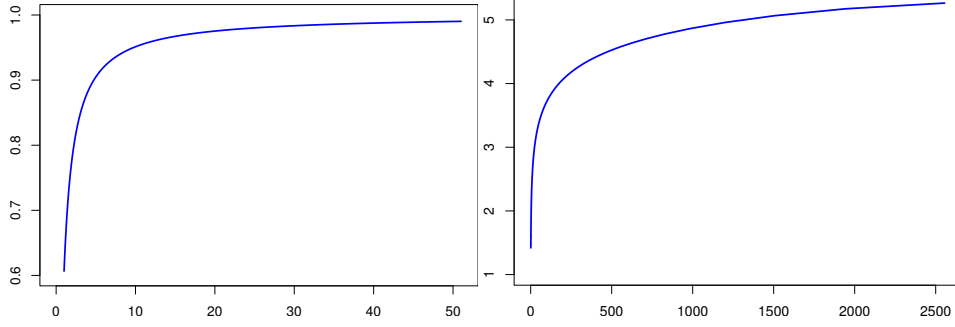Combining relation (12) with Thm. 1, we get the following result.

FIG 2. *The plots of mappings $\gamma \mapsto z_\gamma$ and $\gamma \mapsto l_\gamma(z_\gamma)$. One can observe that both functions are increasing, the first one converges to 1 very rapidly, while the second one seems to diverge very slowly.*

COROLLARY 3. *Let condition $[\mathbf{C1}(\kappa, L)]$ be satisfied with some known constants $\kappa > 0$ and $L < \infty$. Consider the asymptotic set-up in which both $d = d_n$ and $d^* = d_n^*$ tend to infinity as $n \to \infty$. Assume that d grows at a sub-exponential rate in n, that is $\log\log d = o(\log n)$. If*

$$\limsup_{n\to\infty} \frac{d^*}{\log n} < \frac{2}{l_\gamma(z_\gamma)}$$

with $\gamma = L/\kappa$, then consistent estimation of J is possible and can be achieved, for instance, by the estimator $\widehat{J}_n$.

5. Tightness of the assumptions. In this section, we focus our attention on the functional class $\Sigma(\kappa, L)$ of all functions satisfying assumption $[\mathbf{C1}(\kappa, L)]$. For emphasizing that J is the sparsity pattern of the function f, we write J_f instead of J. We assume that $s = \mathrm{Card}(J) = d^*$. The goal is to provide conditions under which the consistent estimation of the sparsity support is impossible, that is there exists a constant $c > 0$ and an integer $n_0 \in \mathbb{N}$ such that, if $n \geq n_0$,

$$\inf_{\widetilde{J}} \sup_{f \in \Sigma(\kappa, L)} \mathbf{P}_f(\widetilde{J} \neq J_f) \geq c,$$

where the inf is over all possible estimators of J_f. To this end, we introduce a set of $M + 1$ probability distributions $\mu_0, \dots, \mu_M$ on $\Sigma(\kappa, L)$ and use the fact that

$$\inf_{\widetilde{J}} \sup_{f \in \widetilde{\Sigma}(\kappa, L)} \mathbf{P}_f(\widetilde{J} \neq J_f) \geq \inf_{\widetilde{J}} \frac{1}{M} \sum_{\ell=1}^{M} \int_{\Sigma(\kappa, L)} \mathbf{P}_f(\widetilde{J} \neq J_f) \mu_\ell(df). \tag{13}$$

These measures μ_ℓ will be chosen in such a way that for each $\ell \geq 1$ there is a set J_ℓ of cardinality d^* such that $\mu_\ell\{J_f = J_\ell\} = 1$ and all the sets $J_1, \dots, J_M$ are distinct. The measure μ_0 is the Dirac measure in 0. Considering these μ_ℓs as "priors" on $\Sigma(\kappa, L)$ and defining the corresponding "posteriors" $\mathbb{P}_0, \mathbb{P}_1, \dots, \mathbb{P}_M$ by

$$\mathbb{P}_\ell(A) = \int_{\Sigma(\kappa, L)} \mathbf{P}_f(A) \mu_\ell(df), \quad \text{for every measurable set } A \subset \mathbb{R}^n,$$

we can write the inequality (13) as

$$\inf_{\widetilde{J}} \sup_{f \in \Sigma(\kappa, L)} \mathbf{P}_f(\widetilde{J} \neq J_f) \geq \inf_{\psi} \frac{1}{M} \sum_{\ell=1}^{M} \mathbb{P}_\ell(\psi \neq \ell), \tag{14}$$

where the inf is taken over all random ψ taking values in $\{0, \ldots, M\}$. The latter inf will be controlled using a suitable version of the Fano lemma. To state it, we denote by $\mathcal{K}(P, Q)$ the Kullback-Leibler divergence between two probability measures P and Q defined on the same probability space.

LEMMA 4 (Corollary 2.6 of [39]).   *Let $M \geq 3$ be an integer, $(\mathcal{X}, \mathcal{A})$ be a measurable space and let $P_0, \ldots, P_M$ be probability measures on $(\mathcal{X}, \mathcal{A})$. Let us set $\bar{p}_{e,M} = \inf_\psi M^{-1} \sum_{\ell=1}^{M} P_\ell(\psi \neq \ell)$, where the inf is taken over all measurable functions $\psi : \mathcal{X} \rightarrow \{1, \ldots, M\}$. If for some $0 < \alpha < 1$, $\frac{1}{M+1} \sum_{\ell=1}^{M} \mathcal{K}(P_\ell, P_0) \leq \alpha \log M$, then $\bar{p}_{e,M} \geq \frac{1}{2} - \alpha$.*

We apply this lemma with $\mathcal{X}$ being the set of all arrays $\boldsymbol{y} = \{y_{\boldsymbol{k}} : \boldsymbol{k} \in \mathbb{Z}^d\}$ such that for some $K > 0$ the entries $y_{\boldsymbol{k}} = 0$ for every $\boldsymbol{k}$ larger than K in ℓ_2-norm. It follows from Fano's lemma that one can deduce a lower bound on $\bar{p}_{e,M}$, the quantity we are interested in, from an upper bound on the average Kullback-Leibler divergence between $\mathbb{P}_\ell$ and $\mathbb{P}_0$. With these tools at hand, we are in a position to state the main result on the impossibility of consistent estimation of the sparsity pattern in the case when the conditions of Thm. 1 are violated.

THEOREM 2.   *Assume that $\vartheta = L/\kappa > 1$ and $\binom{d}{d^*} \geq 3$. Let γ_ϑ be the largest integer satisfying $\gamma\left(1 + \left(\mathsf{h}(z_\gamma) - 1\right)^{-1}\right) \leq \vartheta$, where the Jacobi θ-function h and z_γ are those defined in Section 4.*
  i) If for some $\alpha \in (0, 1/2)$,

$$\frac{N(d^*, \gamma_\vartheta) d^* \log(d/d^*)}{n^2} \geq \frac{\vartheta}{\alpha \gamma_\vartheta} \kappa^2, \tag{15}$$

  then, for d^* large enough, $\inf_{\widetilde{J}} \sup_{f \in \Sigma} \mathbf{P}_f\left(\widetilde{J} \neq J_f\right) \geq \frac{1}{2} - \alpha$.
  ii) If for some $\alpha \in (0, 1/2)$,

$$\frac{d^* \log(d/d^*)}{n} \geq \frac{\kappa}{\alpha}, \tag{16}$$

  then $\inf_{\widetilde{J}} \sup_{f \in \Sigma} \mathbf{P}_f\left(\widetilde{J} \neq J_f\right) \geq \frac{1}{2} - \alpha$.

It is worth stressing here that condition (15) is the converse of condition (9) of Thm. 1 in the case $d^* \rightarrow \infty$, in the sense that condition (9) amounts to requiring that the left-hand side of (15) is smaller than some constant. There is however

one difference between the quantities involved in these conditions: the term $N(d^*, \vartheta(1+\tau))$ of (9) is replaced by $N(d^*, \gamma_\vartheta)$ in condition (15). One can wonder how close γ_ϑ is to ϑ. To give a qualitative answer to this question, we plotted in Figure 3 the curve of the mapping $\vartheta \mapsto \gamma_\vartheta$ along with the bisector $\vartheta \mapsto \vartheta$. We observe that the difference between two curves is small compared to ϑ. As we discuss it later, this property shows that the constants involved in the necessary condition and in the sufficient condition for consistent estimation of J are very close, especially for large values of ϑ.
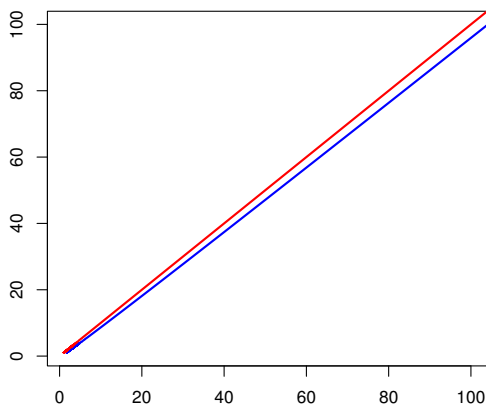

FIG 3. *The curve of the function $L \mapsto \gamma_L$ (blue) and the bisector (red).*

6. Adaptivity and minimax rates of separation.

6.1. *Adaptation with respect to L and κ.* The estimator $\widehat{J}(\boldsymbol{m}, \boldsymbol{\lambda})$ we have introduced in Section 3 is clearly nonadaptive: the tuning parameters $(\boldsymbol{m}, \boldsymbol{\lambda})$ recommended by the developed theory involve the values L and κ, which are generally unknown. Fortunately, we can take advantage of the fact that the choice of $\boldsymbol{m}$ and $\boldsymbol{\lambda}$ is governed by the one-dimensional parameter $\vartheta = L(1+\tau)/\kappa$. Therefore, it is realistic to assume that a finite grid of values $1 < \vartheta_1 \le \ldots \le \vartheta_K < \infty$ is available containing a true value of ϑ. The following result provides an adaptive procedure of variable selection with guaranteed control of the error.

PROPOSITION 5. *Let $1 < \vartheta_1 \le \ldots \le \vartheta_K < \infty$ and $\tau > 0$ be given values and set*[1]

$$i^* = \min\left\{ i : (1+\tau)\frac{\max_{j=1,\ldots,d} \sum_{\boldsymbol{k}\in\mathbb{Z}^d} k_j^2 \theta_{\boldsymbol{k}}^2}{\min_{j\in J} \sum_{\boldsymbol{k}:k_j\neq 0} \theta_{\boldsymbol{k}}^2} \le \vartheta_i \right\} \le K.$$

For every $i, \ell \in \mathbb{N}$, let us denote $\widehat{J}_n(i) = \widehat{J}_n(\boldsymbol{m}(\vartheta_i), \boldsymbol{\lambda}(\vartheta_i))$ with $m_\ell(\vartheta) = (\vartheta\ell)^{1/2}$ and

$$\lambda_\ell(\vartheta) = \frac{2\sqrt{2N(\ell, \vartheta)d^* \log(2ed/d^*)} + 4d^* \log(2ed/d^*)}{n}.$$

---

[1]We use the convention that the minimum over an empty set equals $+\infty$.

If the condition $4\lambda_s(\vartheta_{i^*}) < \kappa\tau/(1+\tau)$ *is fulfilled, then the estimator* $\widehat{J}_n^{\mathrm{ad}} = \bigcup_{i=1}^K \widehat{J}_n(i)$ *satisfies* $\mathbf{P}(\widehat{J}_n^{\mathrm{ad}} \neq J) \leq (K+2)(d^*/2ed)^{d^*}$.

In simple words, if the grid of possible values $\{\vartheta_i\}$ has a cardinality K which is not too large (that is $K(d^*/d)^{d^*} \to 0$), then declaring a variable relevant if at least one of the procedures $\widehat{J}_n(i)$ suggests its relevance provides a consistent and adaptive variable selection strategy. The proof of this statement follows immediately from Prop. 1 and Thm. 1. Indeed, applying Prop. 1 with $A = 2$ yields $\mathbf{P}(\widehat{J}_n^{\mathrm{ad}} \not\subset J) \leq \sum_{i=1}^K \mathbf{P}(\widehat{J}_n(i) \not\subset J) \leq K(d^*/2ed)^{d^*}$, while Thm. 1 ensures that $\mathbf{P}(\widehat{J}_n^{\mathrm{ad}} \not\supset J) \leq \mathbf{P}(\widehat{J}_n(i^*) \not\supset J) \leq 2(d^*/2ed)^{d^*}$.

6.2. *Minimax rates of separation.*   Since the methodology of Section 3 takes its roots in the theory of hypotheses testing, one naturally wonders what are the minimax rates of separation in the problem of variable selection. The results stated in foregoing sections allow us to answer this question in the case of Sobolev smoothness 1 and alternatives separated in L^2-norm. The following result, the proof of which is postponed to the Appendix E provides minimax rates. We assume herein that the true sparsity $s = \mathrm{Card}(J)$ and its known upper estimate d^* are such that d^*/s is bounded from above by some constant.

PROPOSITION 6.   *There is a constant* D^* *depending only on L such that if*

$$\kappa \geq D^*\left\{\left(\frac{\log(d/s)}{n^2}\right)^{-2/(4+s)} \bigvee \frac{s\log(d/s)}{n}\right\},$$

then there exists a consistent estimator of J. Furthermore, the consistency is uniform in $\mathfrak{f} \in \Sigma(\kappa, L)$. *On the other hand, there is a constant* D_* *depending only on L such that if*

$$\kappa \leq D_*\left\{\left(\frac{\log(d/s)}{n^2}\right)^{-2/(4+s)} \bigvee \frac{s\log(d/s)}{n}\right\},$$

then uniformly consistent estimation of J is impossible.

Borrowing the terminology of the theory of hypotheses testing, we say that $\left(\frac{\log(d/s)}{n^2}\right)^{-2/(4+s)} \vee \frac{s\log(d/s)}{n}$ is the minimax rate of separation in the problem of variable selection for Sobolev smoothness one. These results readily extend to Sobolev smoothness of any order $\beta \geq 1$, in which case the rate of separation takes the form $\left(\frac{\log(d/s)}{n^2}\right)^{-2\beta/(4\beta+s)} \vee \frac{s\log(d/s)}{n}$. The first term in this maximum coincides, up to the logarithmic term, with the minimax rate of separation in the problem of detection of an s-dimensional signal [22]. Note, however, that in our case this logarithmic inflation is unavoidable. It is the price to pay for not knowing in advance which s variables are relevant.

7. Nonparametric regression with random design. So far, we have analyzed the situation in which noisy observations of the regression function $f(\cdot)$ are available at all points $\boldsymbol{x} \in [0,1]^d$. Let us turn now to the more realistic model of nonparametric regression, when the observed noisy values of f are sampled at random in the unit hypercube $[0,1]^d$. More precisely, we assume that n independent and identically distributed pairs of input-output variables $(\boldsymbol{X}_i, Y_i)$, $i = 1, \ldots, n$ are observed that obey the regression model

$$Y_i = f(\boldsymbol{X}_i) + \sigma \varepsilon_i, \qquad i = 1, \ldots, n.$$

The input variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are assumed to take values in $\mathbb{R}^d$ while the output variables $Y_1, \ldots, Y_n$ are scalar. As usual, $\varepsilon_1, \ldots, \varepsilon_n$ are such that $\mathbf{E}[\varepsilon_i | \boldsymbol{X}_i] = 0$, $i = 1, \ldots, n$; additional conditions will be imposed later. Without requiring from f to be of a special parametric form, we aim at recovering the set $J \subset \{1, \ldots, d\}$ of its relevant variables. The noise magnitude σ is assumed to be known.

It is clear that the estimation of J cannot be accomplished without imposing some further assumptions on f and on the distribution P_X of the input variables. Roughly speaking, we will assume that f is differentiable with a squared integrable gradient and that P_X admits a density which is bounded from below. More precisely, let g denote the density of P_X w.r.t. the Lebesgue measure.

[C2] $g(\boldsymbol{x}) = 0$ for any $\boldsymbol{x} \notin [0,1]^d$ and that $g(\boldsymbol{x}) \geq g_{\min}$ for any $\boldsymbol{x} \in [0,1]^d$.

The next assumptions imposed to the regression function and to the noise require their boundedness in an appropriate sense. These assumptions are needed in order to prove, by means of a concentration inequality, the closeness of the empirical coefficients to the true ones.

[C3(L_∞, L_2)] The $L^\infty([0,1]^d, \mathbb{R}, P_X)$ and $L^2([0,1]^d, \mathbb{R}, P_X)$ norms of the function f are bounded from above respectively by L_∞ and L_2, i.e., $\mathbf{P}(|f(\boldsymbol{X})| \leq L_\infty) = 1$ and $\mathbf{E}[f(\boldsymbol{X})^2] \leq L_2^2$.

[C4] The noise variables satisfy a.e. $\mathbf{E}[e^{t\varepsilon_i} | \boldsymbol{X}_i] \leq e^{t^2/2}$ for all $t > 0$.

We stress once again that the primary aim of this work is merely to understand when it is possible to consistently estimate the sparsity pattern. The estimator that we will define is intended to show the possibility of consistent estimation, rather than being a practical procedure for recovering the sparsity pattern. Therefore, the estimator will be allowed to depend on the parameters $g_{\min}$, L, κ and L_2 appearing in conditions **[C1-C3]**.

7.1. *An estimator of J and its consistency.* The estimator of the sparsity pattern J that we are going to introduce now is based on the following simple ob-

servation: if $j \notin J$ then $\theta_{\boldsymbol{k}}[\mathsf{f}] = 0$ for every $\boldsymbol{k}$ such that $k_j \neq 0$. In contrast, if $j \in J$ then there exists $\boldsymbol{k} \in \mathbb{Z}^d$ with $k_j \neq 0$ such that $|\theta_{\boldsymbol{k}}[\mathsf{f}]| > 0$. To turn this observation into an estimator of J, we start by estimating the Fourier coefficients $\theta_{\boldsymbol{k}}[\mathsf{f}]$ by their empirical counterparts:

$$\widehat{\theta}_{\boldsymbol{k}} = \frac{1}{n} \sum_{i=1}^n \frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\mathsf{g}(\boldsymbol{X}_i)} Y_i, \qquad \boldsymbol{k} \in \mathbb{Z}^d.$$

Then, for every $\ell \in \mathbb{N}$ and for any $\gamma > 0$, we introduce the notation $S_{m,\ell}^j = \{\boldsymbol{k} \in \mathbb{Z}^d : \|\boldsymbol{k}\|_2 \leq m, \|\boldsymbol{k}\|_0 \leq \ell, k_j \neq 0\}$. The estimator of J is defined by

$$\widehat{J}_n^{(1)}(m, \lambda) = \left\{ j \in \{1, \ldots, d\} : \max_{\boldsymbol{k} \in S_{m,d^*}^j} |\widehat{\theta}_{\boldsymbol{k}}| > \lambda \right\}, \tag{17}$$

where m and λ are some parameters to be defined later. The next result, the proof of which is placed in the supplementary material, provides consistency guarantees for $\widehat{J}_n^{(1)}(m, \lambda)$.

THEOREM 3.   *Let conditions* [**C1-C4**] *be fulfilled with some known values* $g_{\min}$, $\vartheta = 2L/\kappa$ *and* L_2. *Assume furthermore that the design density* g *and an upper estimate on the noise magnitude* σ *are available. Set* $m = (\vartheta d^*)^{1/2}$ *and* $\lambda = 4(\sigma + L_2)(d^* \log(24\sqrt{\vartheta}d/d^*)/n\,g_{\min}^2)^{1/2}$. *If the following conditions are satisfied:*

$$\frac{d^* \log(24\sqrt{\vartheta}d/d^*)}{n} \leq \frac{L_2^2}{L_\infty^2}, \quad \frac{128(\sigma + L_2)^2 d^* N(d^*, \vartheta) \log(24\sqrt{\vartheta}d/d^*)}{n\,g_{\min}^2} < \kappa, \tag{18}$$

then the estimator $\widehat{J}^{(1)}(m, \lambda)$ *satisfies* $\mathbf{P}\big(\widehat{J}^{(1)}(m, \lambda) \neq J\big) \leq (8d/d^*)^{-d^*}$.

If we take a look at the conditions of Theorem 3 ensuring the consistency of $\widehat{J}_n^{(1)}$, it becomes clear that the strongest requirement is the second inequality in (18). Roughly speaking, this condition requires that $d^* N(d^*, \vartheta) \log(d/d^*)/n$ is bounded from above by some constant. According to results stated in Section 4, $N(d^*, \vartheta)$ diverges exponentially fast, making inequality (18) impossible for d^* larger than $\log n$ up to a multiplicative constant.

It is also worth stressing that although we require the P_X-a.e. boundedness of f by some constant L_∞, this constant is not needed for computing the estimator proposed in Thm. 3. Only constants related to some quadratic functionals of the sequence of Fourier coefficients $\theta_{\boldsymbol{k}}[\mathsf{f}]$ are involved in the tuning parameters m and λ. This point might be important for designing practical estimators of J, since the estimation of quadratic functionals is more realistic, see for instance [27, 9], than the estimation of sup-norm.

Theorem 3 can be reformulated to characterize the level of relevance κ for the relevant components of $\boldsymbol{X}$ making their identification possible. In fact, an

alternative way of stating Theorem 3 is the following: under conditions [**C1-C4**] if ϑ is an arbitrary tuning parameter satisfying the first inequality in (18), then the estimator $\widehat{J}_n^{(1)}(m,\lambda)$—with m and λ chosen as in Theorem 3—satisfies $\mathbf{P}(\widehat{J}_n^{(1)}(m,\lambda) \neq J) \leq (8d/d^*)^{-d^*}$ if the smallest level of relevance κ for components X_j of $\boldsymbol{X}$ with $j \in J$ is not smaller than $8\lambda^2 N(d^*, m^2/d^*)$. This statement can be easily deduced from the proof of Theorem 3 (cf. supplementary material).

7.2. *Tightness of the assumptions.* A natural question is now to check that the assumptions of Thm. 3 are tight in the asymptotic regimes of fixed sparsity and increasing ambient dimension, as well as increasing sparsity. We will only establish an analogue of claim ii) of Thm. 2. An attempt to prove a result similar to claim i) of Thm. 2 was done in [11, Theorem 2]. However, the result of [11] involves a stringent assumption on the empirical Gram matrix (cf. condition (6) in [11]) and, unfortunately, we are unable to prove the existence of a sampling scheme for which this assumption is fulfilled.

We assume that the errors ε_i are i.i.d. standard Gaussian and we focus our attention on the functional class $\Sigma(\kappa, L)$. The following simple result shows that the conditions of Thm. 3 are tight in the case of fixed intrinsic dimension.

PROPOSITION 7. *Let the design $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0,1]^d$ be either deterministic or random. If for some positive $\alpha < 1/2$, the inequality*

$$\frac{d^* \log(d/d^*)}{n} \geq \kappa \alpha^{-1}$$

holds true, then there is a constant $c > 0$ such that $\inf_{\widetilde{J}_n} \sup_{f \in \Sigma(\kappa, L)} \mathbf{P}_f(\widetilde{J}_n \neq J_f) \geq c$.

8. Concluding remarks. The results proved in previous sections almost exhaustively answer the questions on the existence of consistent estimators of the sparsity pattern in the model of Gaussian white noise and, to a smaller extent, in nonparametric regression. In fact as far as only rates of convergence are of interest, the result obtained in Thm. 1 is shown in Section 5 to be unimprovable. Thus only the problem of finding sharp constants remains open. To make these statements more precise, let us consider the simplified set-up $\sigma = \kappa = 1$ and define the following two regimes:

- The regime of fixed sparsity, *i.e.*, when the sample size n and the ambient dimension d tend to infinity but the intrinsic dimension d^* remains constant or bounded.
- The regime of increasing sparsity, *i.e.*, when the intrinsic dimension d^* tends to infinity along with the sample size n and the ambient dimension d. For simplicity, we will assume that $d^* = O(d^{1-\epsilon})$ for some $\epsilon > 0$.

In the fixed sparsity regime, in view of Theorems 1 and 3, consistent estimation of the sparsity pattern can be achieved both in Gaussian white noise model and nonparametric regression as soon as $\limsup_{n\to\infty}(d^*\log d)/n < c_\star$, where $c_\star$ is the constant defined by $c_* = 1/8$ for the Gaussian white noise model and

$$c_\star = \min\Big(\frac{L_2^2}{2L_\infty^2}, \frac{g_{\min}^2}{2^8(1+L_2)^2 N(d^*, 2L)}\Big)$$

for the regression model. On the other hand, by Thm. 2 and Prop. 7, consistent estimation of the sparsity pattern is impossible if $\liminf_{n\to\infty}(d^*\log d)/n > c^\star$ with $c^\star = 2$. Thus, up to multiplicative constants $c_\star$ and $c^\star$ (which are clearly not sharp), the results of Theorems 1 and 3 cannot be improved in the regime of fixed sparsity.

In the regime of increasing sparsity, the results we get in the model of Gaussian white noise are much stronger than those for nonparametric regression. In the former model, taking the logarithm of both sides of inequality (9) and using formula (12) for $N(d^*, \cdot) = N_1(d^*, \cdot) - N_2(d^*, \cdot)$, we see that consistent estimation of J is possible when, for some $\tau > 0$ and for all n, the following two conditions are fulfilled:

$$\begin{cases} \mathsf{l}_{L+\tau}(z_{L+\tau})d^* + \frac{1}{2}\log d^* + \log\log(d/d^*) - 2\log n < \underline{c}_1, \\ \log d^* + \log\log(d/d^*) - \log n \leq \underline{c}_1' \end{cases} \tag{19}$$

with some constants $\underline{c}_1 = \underline{c}_1(L, \tau)$ and $\underline{c}_1' = \underline{c}_1'(L, \tau)$. On the other hand, Thm. 2 yields that there are some constants $\bar{c}_1$ and $\bar{c}_1'$ such that it is impossible to consistently estimate J if either one of the conditions

$$\mathsf{l}_{\gamma_L}(z_{\gamma_L})d^* + \frac{1}{2}\log d^* + \log\log(d/d^*) - 2\log n \geq \bar{c}_1, \tag{20}$$

$$\log d^* + \log\log(d/d^*) - \log n \geq \bar{c}_1', \tag{21}$$

is satisfied. First note that the left-hand side of the second condition in (19) is exactly the same as the left-hand side of (21). If we compare now the left-hand side of the first condition in (19) with the left-hand side of (20), we see that only the coefficients of d^* differ. To measure the degree of difference of these two coefficients we draw in Figure 4 the plots of the functions $L \mapsto \mathsf{l}_L(z_L)$ and $L \mapsto \mathsf{l}_{\gamma_L}(z_{\gamma_L})$, with γ_L as is Thm. 2. One can observe that the two curves are very close especially for relatively large values of L. This implies that the conditions (19) are tight. A simple consequence of inequalities (19) and (20) is that the consistent recovery of the sparsity pattern is possible under the condition $d^*/\log n \to 0$ and impossible for $d^*/\log n \to \infty$ as $n \to \infty$, provided that $\log\log(d/d^*) = o(\log n)$.
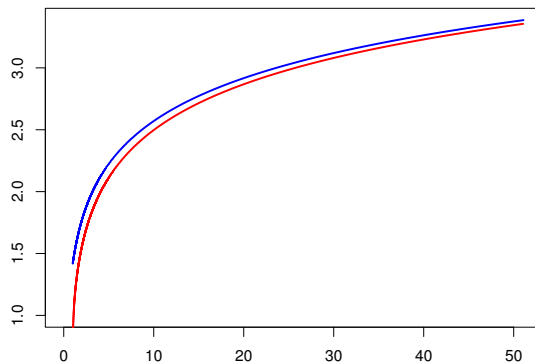
FIG 4. *The curves of functions $L \mapsto I_L(z_L)$ (blue curve) and $L \mapsto I_{\gamma_L}(z_{\gamma_L})$ (red curve).*

Still in the regime of increasing sparsity, but for nonparametric regression, we proved that consistent estimation of the sparsity pattern is possible whenever

$$\begin{cases} I_{L+\tau}(z_{L+\tau})d^* + \frac{1}{2}\log d^* + \log\log(d/d^*) - \log n < \underline{c}_2, \\ \log d^* + \log\log d - \log n < \underline{c}_2' \end{cases} \tag{22}$$

with some constants $\underline{c}_2 = \underline{c}_2(g_{\min}, \sigma, L_2, L)$ and $\underline{c}_2' = 2\log(L_2/L_\infty)$. As we have already mentioned, the second condition in (22) is tight, up to the choice of $\underline{c}_2'$, in view of Proposition 7. It is natural to expect that the first condition is tight as well, since it is in the model of Gaussian white noise, which has the reputation of being simpler than the model of nonparametric regression. However, we do not have a mathematical proof of this statement.

Let us stress now that, all over this work, we have deliberately avoided any discussion on the computational aspects of the variable selection in nonparametric regression. The goal in this paper was to investigate the possibility of consistent recovery without paying attention to the complexity of the selection procedure. This lead to some conditions that could be considered a benchmark for assessing the properties of sparsity pattern estimators. As for the estimators proposed in Section 3, it is worth noting that their computational complexity is not always prohibitively large. A recommended strategy is to compute the coefficients $\widehat{\theta}_{\pmb{k}}$ in a stepwise manner; at each step $K = 1, 2, \ldots, d^*$ only the coefficients $\widehat{\theta}_{\pmb{k}}$ with $\|\pmb{k}\|_0 = K$ need to be computed and compared with the threshold. If some $\widehat{\theta}_{\pmb{k}}$ exceeds the threshold, then all the variables X^j corresponding to nonzero coordinates of $\pmb{k}$ are considered as relevant. We can stop this computation as soon as the number of variables classified as relevant attains d^*. While the worst-case complexity of this procedure is exponential, there are many functions f for which the complexity of the procedure will be polynomial in d. For example, this is the case for additive models in which $f(\pmb{x}) = f_1(x_{i_1}) + \ldots + f_{d^*}(x_{i_{d^*}})$ for some univariate functions $f_1, \ldots, f_{d^*}$.

Note also that in the present study we focused exclusively on the consistency of variable selection without paying any attention to the consistency of regression function estimation. A thorough analysis of the latter problem being left to a future work, let us simply remark that in the case of fixed d^*, under the conditions of Thm. 3, it is straightforward to construct a consistent estimator of the regression function. In fact, it suffices to use a projection estimator with a properly chosen truncation parameter on the set of relevant variables. The situation is much more delicate in the case when the sparsity d^* grows to infinity along with the sample size n. Presumably, condition (19) is no longer sufficient for consistently estimating the regression function. The rationale behind this conjecture is that the minimax rate of convergence for estimating f in our context, if we assume in addition that the set of relevant variables is known, is equal to $n^{-2/(2+d^*)} = \exp(-2\log n/(2+d^*))$. If the left-hand side of (19) is equal to a constant and $\log\log d = o(\log n)$, then the aforementioned minimax rate does not tend to zero, making thus the estimator inconsistent.

Finally, we would like to mention that the selection of relevant variables is a challenging statistical task, which might be useful to perform independently of the task of regression function estimation. Indeed, if we succeed in identifying relevant variables on a data-set having a small sample size, we can continue the data collection process more efficiently by recording only the values of relevant variables. This may considerably reduce the memory costs related to the data storage and the financial costs necessary for collecting new data. Then, the regression function may be estimated more accurately on the base of this new (larger) data-set.

APPENDIX A: PROOF OF PROPOSITION 1

To ease notation, we write $\widehat{J}_n$ instead of $\widehat{J}_n(\boldsymbol{m}, \boldsymbol{\lambda})$. It is clear that $\widehat{J}_n \not\subset J$ if and only if $\exists j \in J^c$ such that $\max_{\ell \leq d^*} \lambda_\ell^{-1} \max_{I \in P_\ell^d} \widehat{Q}_{m,I}^j \geq 1$, where $Q_{m,I}^j = \sum_{\boldsymbol{k} \in S_{m,I}^j} \theta_{\boldsymbol{k}}^2$. For every $j \in \{1, \ldots, d\}$, let us set $R_{m,I}^j = \sum_{\boldsymbol{k} \in S_{m,I}^j} (\xi_{\boldsymbol{k}}^2 - 1)$ and $N_{m,I}^j = (Q_{m,I}^j)^{-1/2} \sum_{\boldsymbol{k} \in S_{m,I}^j} \theta_{\boldsymbol{k}} \xi_{\boldsymbol{k}}$ so that

$$\widehat{Q}_{m,I}^j = \sum_{\boldsymbol{k} \in S_{m,I}^j} \left(y_{\boldsymbol{k}}^2 - \frac{1}{n} \right) = Q_{m,I}^j + \frac{2\sqrt{Q_{m,I}^j}}{\sqrt{n}} N_{m,I}^j + \frac{1}{n} R_{m,I}^j. \tag{23}$$

For $j \in J^c$, the first two terms of the last sum vanish and, therefore, we have

$$\left\{ \widehat{J}_n \not\subset J \right\} = \bigcup_{j \in J^c} \bigcup_{\ell \leq d^*} \bigcup_{I \in P_\ell^d} \left\{ R_{m,I}^j \geq n\lambda_\ell \right\} = \bigcup_{\ell \leq d^*} \bigcup_{I \in P_\ell^d} \bigcup_{j \in J^c \cap I} \left\{ R_{m,I}^j \geq n\lambda_\ell \right\},$$

where the last equality results from the fact that $R_{m,I}^j = 0$ if $j \notin I$. The random

variable $R_{m,I}^{j}$, being a centered sum of squares of independent standard Gaussian random variables, follows a translated χ^2-distribution. The tails of this distribution can be evaluated using the following result.

LEMMA 8 (cf. Lemma 1 in [27]).   *Let $\xi_1,...,\xi_D$ be independent standard Gaussian random variables. For every $x \geq 0$ and for every vector $\boldsymbol{a} = (a_1,...,a_D) \in \mathbb{R}_+^D$, the following inequalities hold true:*

$$\mathbf{P}\Big(\sum_{i=1}^{D} a_i(\xi_i^2 - 1) \geq 2\|\boldsymbol{a}\|_2\sqrt{x} + 2\|\boldsymbol{a}\|_\infty x\Big) \leq \exp(-x),$$
$$\mathbf{P}\Big(\sum_{i=1}^{D} a_i(\xi_i^2 - 1) \leq -2\|\boldsymbol{a}\|_2\sqrt{x}\Big) \leq \exp(-x).$$

We apply this lemma to $R_{m_\ell,I}^{i}$, for which $\|\boldsymbol{a}\|_\infty = 1$ and $\|\boldsymbol{a}\|_2^2 = N(\ell, m_\ell^2/\ell)$. Setting $n\lambda_\ell = 2\sqrt{N(\ell, m_\ell^2/\ell)x} + 2x$ and using the union bound, we get

$$\mathbf{P}\Big(\widehat{J}_n \not\subset J\Big) \leq \mathbf{P}\Big(\bigcup_{\ell=1}^{d^*}\Big\{\max_{I \in P_\ell^d; i \in I} R_{m_\ell,I}^{i} \geq n\lambda_\ell\Big\}\Big)$$
$$\leq \sum_{\ell=1}^{d^*} \ell \operatorname{Card}(P_\ell^d) \max_{I \in P_\ell^d; i \in I} \mathbf{P}\Big(R_{m_\ell,I}^{i} > n\lambda_\ell\Big) \leq e^{-x}\sum_{\ell=1}^{d^*} \ell\binom{d}{\ell}.$$

One checks that $\sum_{\ell=1}^{d^*} \ell\binom{d}{\ell} \leq (2ed/d^*)^{d^*}$ holds true for every pair of integers (d^*,d) such that $1 \leq d^* \leq d$ (cf. supplementary material for a proof). Hence, for $x = Ad^*\log(2ed/d^*)$, we get $\mathbf{P}\big(\widehat{J}_n \not\subset J\big) \leq (2ed/d^*)^{-(A-1)d^*}$.

APPENDIX B: PROOF OF THEOREM 1

We begin with proving a stronger result that implies the claim of Thm. 1.

PROPOSITION 9.   *Let α be a real number from $(0,1)$. If for every $j \in J$ and for $s = \operatorname{Card}(J)$ the inequality*

$$Q_{m_s,J}^{j} \geq \left\{\left[\lambda_s + \frac{2\sqrt{N(s, m_s^2/s)\log(2s/\alpha)} + 1}{n}\right]^{1/2} + \left[\frac{2\log(2s/\alpha)}{n}\right]^{1/2}\right\}^2 \quad (24)$$

holds true, then $\mathbf{P}(J \not\subset \widehat{J}_n) \leq \alpha$.

PROOF. To bound from above the probability of Type II error, we rely on the equivalence: $J \not\subset \widehat{J}_n$ if and only if $\exists j \in J$ such that $\max_{\ell \leq d^*} \lambda_\ell^{-1} \max_{I \in P_\ell^d} \widehat{Q}_{m,I}^{j} \leq 1$.

Recall that $s = \mathrm{Card}(J)$. Using Bonferroni's inequality, we get

$$\mathbf{P}(J \not\subset \widehat{J}_n) \leq \sum_{j \in J} \mathbf{P}\Big(\max_{\ell \leq d^*} \lambda_\ell^{-1} \max_{I \in P_\ell^d} \widehat{Q}_{m,I}^j \leq 1\Big)$$

$$\leq \sum_{j \in J} \mathbf{P}\Big(\widehat{Q}_{m_s,J}^j \leq \lambda_s\Big) \leq s \max_{j \in J} \mathbf{P}\Big(\widehat{Q}_{m_s,J}^j \leq \lambda_s\Big). \qquad (25)$$

By virtue of decomposition (23),

$$\mathbf{P}\Big(\widehat{Q}_{m_s,J}^j \leq \lambda_s\Big) = \mathbf{P}\Bigg(\Big(\sqrt{Q_{m_s,J}^j} + \frac{1}{\sqrt{n}} N_{m_s,J}^j\Big)^2 + \frac{1}{n}\Big(R_{m_s,J}^j - (N_{m_s,J}^j)^2\Big) \leq \lambda_s\Bigg).$$

One checks that $R_{m_s,J}^j - (N_{m_s,J}^j)^2 + N(s, m_s^2/s)$ is a drawn from χ^2-distribution with $N(s, m_s^2/s) - 1$ degrees of freedom. Therefore, using Lemma 8 stated in previous section, we get $\mathbf{P}\big(\frac{1}{n}\big(R_{m_s,J}^j - (N_{m_s,J}^j)^2\big) + \frac{1}{n} \leq -2\sqrt{N(s, m_s^2/s)\log(2s/\alpha)}\big) \leq \frac{\alpha}{2s}$. Therefore, $\mathbf{P}\big(\widehat{Q}_{m_s,J}^j \leq \lambda_s\big)$ is upper-bounded by

$$\frac{\alpha}{2s} + \mathbf{P}\Bigg(\Big(\sqrt{Q_{m_s,J}^j} + \frac{1}{\sqrt{n}} N_{m_s,J}^j\Big)^2 \leq \lambda_s + \frac{2\sqrt{N(s, m_s^2/s)\log(2s/\alpha)} + 1}{n}\Bigg).$$

Using the condition of the proposition, we get $\mathbf{P}\big(\widehat{Q}_{m_s,J}^j \leq \lambda_s\big) \leq \frac{\alpha}{2s} + \mathbf{P}\big(N_{m_s,J}^j \leq -\sqrt{2\log(2s/\alpha)}\big) \leq \frac{\alpha}{s}$. Combining this inequality with (25), we get the result of Proposition 9. $\qquad\square$

To deduce the claim of Thm. 1 from that of Prop. 9, we use the following lower bound:

$$Q_{m_s,J}^j = Q^j - \sum_{j \in \mathrm{supp}(\boldsymbol{k}) \subset J} \theta_{\boldsymbol{k}}^2 \mathbf{1}_{\{\|\boldsymbol{k}\|_2 \geq m_s\}} \geq \kappa - \sum_{j \in \mathrm{supp}(\boldsymbol{k}) \subset J} \theta_{\boldsymbol{k}}^2 \mathbf{1}_{\{\|\boldsymbol{k}\|_2 \geq m_s\}}$$

$$\geq \kappa - m_s^{-2} \sum_{j \in \mathrm{supp}(\boldsymbol{k}) \subset J} \theta_{\boldsymbol{k}}^2 \|\boldsymbol{k}\|_2^2 \geq \kappa - m_s^{-2} L s, \qquad (26)$$

for every $j \in J$. Our choice of m_s, $m_s = \sqrt{sL(1+\tau)/\kappa}$, ensures that $Q_{m_s,J}^j \geq \kappa\tau/(1+\tau)$. Finally, using a very rough bound (which is sufficient for our purposes), the right-hand side in (24) can be upper-bounded by $4\lambda_s$ if α is chosen to be equal to $2(2ed/d^*)^{-(A-1)d^*}$. Therefore, if $\frac{\kappa\tau}{1+\tau} \geq 4\lambda_s$, then (24) holds true with $\alpha = 2(2ed/d^*)^{-(A-1)d^*}$ and, therefore, the type II error has a probability less than or equal to $2(2ed/d^*)^{-(A-1)d^*}$.

APPENDIX C: PROOF OF PROPOSITION 2

Proof of the first assertion. This proof can be found in [30], we repeat here the arguments therein for the sake of keeping the paper self-contained. Recall that

$N_1(d^*, \gamma)$ admits an integral representation with the integrand:

$$\frac{\mathsf{h}(z)^{d^*}}{z^{\gamma d^*}} \frac{1}{z(1-z)} = \frac{1}{z(1-z)} \exp\left[d^* \log\left(\frac{\mathsf{h}(z)}{z^\gamma}\right)\right].$$

For any $y > 0$, we define $\phi(y) = e^{-y} \mathsf{h}'(e^{-y})/\mathsf{h}(e^{-y}) = \sum_{k \in \mathbb{Z}} k^2 e^{-yk^2}/\sum_{k \in \mathbb{Z}} e^{-yk^2}$ in such a way that

$$\phi(y) = \gamma \iff \frac{\mathsf{h}'(e^{-y})}{\mathsf{h}(e^{-y})} = \frac{\gamma}{e^{-y}} \iff \mathsf{l}'_\gamma(e^{-y}) = 0.$$

By virtue of the Cauchy-Schwarz inequality, it holds that $\sum k^4 e^{-yk^2} \sum e^{-yk^2} > (\sum k^2 e^{-yk^2})^2$, $\forall y \in (0, \infty)$, implying that $\phi'(y) < 0$ for all $y \in (0, \infty)$, *i.e.*, ϕ is strictly decreasing. Furthermore, ϕ is obviously continuous with $\lim_{y \to 0} \phi(y) = +\infty$ and $\lim_{y \to \infty} \phi(y) = 0$. These properties imply the existence and the uniqueness of $y_\gamma \in (0, \infty)$ such that $\phi(y_\gamma) = \gamma$. Furthermore, as the inverse of a decreasing function, the function $\gamma \mapsto y_\gamma$ is decreasing as well. We set $z_\gamma = e^{-y_\gamma}$ so that $\gamma \mapsto z_\gamma$ is increasing. We also have

$$\mathsf{l}''_\gamma(z_\gamma) = \frac{\mathsf{h}''\mathsf{h} - (\mathsf{h}')^2}{\mathsf{h}^2}(z_\gamma) + \frac{\gamma}{z_\gamma^2} = z_\gamma^{-2}\left\{\frac{\sum_k (k^4 - k^2) z_\gamma^{k^2}}{\sum_k z_\gamma^{k^2}} - \left(\frac{\sum_k k^2 z_\gamma^{k^2}}{\sum_k z_\gamma^{k^2}}\right)^2 + \gamma\right\}$$

$$= z_\gamma^{-2}\{-\phi'(y_\gamma) - \phi(y_\gamma) + \gamma\} = -z_\gamma^{-2}\phi'(y_\gamma) > 0.$$

Proof of the second assertion.   We apply the saddle-point method to the integral representing N_1 see, *e.g.*, Chapter IX in [14]. It holds that

$$N_1(d^*, \gamma) = \frac{1}{2\pi i} \oint_{|z|=z_\gamma} \frac{\mathsf{h}(z)^{d^*}}{z^{\gamma d^*}} \frac{dz}{z(1-z)} = \frac{1}{2\pi i} \oint_{|z|=z_\gamma} \{z(1-z)\}^{-1} e^{d^* \mathsf{l}_\gamma(z)} dz. \quad (27)$$

The first assertion of the proposition provided us with a real number z_γ such that $\mathsf{l}'_\gamma(z_\gamma) = 0$ and $\mathsf{l}''_\gamma(z_\gamma) > 0$. The tangent to the steepest descent curve at z_γ is vertical. The path we choose for integration is the circle with center 0 and radius z_γ. As this circle and the steepest descent curve have the same tangent at z_γ, applying formula (1.8.1) of [14] (with $\alpha = 0$ since $\mathsf{l}''(z_\gamma)$ is real and positive), we get that

$$\oint_{|z|=z_\gamma} \{z(1-z)\}^{-1} e^{d^* \mathsf{l}_\gamma(z)} dz = \sqrt{\frac{2\pi}{d^* \mathsf{l}''_\gamma(z_\gamma)}} e^{i\pi/2} \{z_\gamma(1-z_\gamma)\}^{-1} e^{d^* \mathsf{l}_\gamma(z_\gamma)}(1 + o(1)),$$

when $d^* \to \infty$, as soon as the condition[2] $\Re[\mathsf{l}_\gamma(z) - \mathsf{l}_\gamma(z_\gamma)] \leq -\mu$ is satisfied for some $\mu > 0$ and for any z belonging to the circle $|z| = |z_\gamma|$ and lying not too close to z_γ. To check that this is indeed the case, we remark that $\Re[\mathsf{l}_\gamma(z)] = \log\left|\frac{\mathsf{h}(z)}{z^\gamma}\right|$.

---

[2]$\Re u$ stands for the real part of the complex number u.

Hence, if $z = z_\gamma e^{i\omega}$ with $\omega \in [\omega_0, 2\pi - \omega_0]$ for some $\omega_0 \in]0, \pi[$, then

$$\left| \frac{\mathsf{h}(z)}{z^\gamma} \right| = \frac{|1 + 2z + 2\sum_{k>1} z^{k^2}|}{z_\gamma^\gamma} \leq \frac{|1 + z| + z_\gamma + 2\sum_{k>1} z_\gamma^{k^2}}{z_\gamma^\gamma}$$

$$\leq \frac{|1 + e^{i\omega_0} z_\gamma| + z_\gamma + 2\sum_{k>1} z_\gamma^{k^2}}{z_\gamma^\gamma}.$$

Therefore $\Re[\mathsf{l}_\gamma(z) - \Re\mathsf{l}_\gamma(z_\gamma)] \leq -\mu$ with $\mu = \log\left(\frac{1 + 2z_\gamma + 2\sum_{k>1} z_\gamma^{k^2}}{|1 + z_\gamma e^{i\omega_0}| + z_\gamma + 2\sum_{k>1} z_\gamma^{k^2}} \right) > 0$. This completes the proof for the term $N_1(d^*, \gamma)$. The term $N_2(d^*, \gamma)$ can be dealt in the same way.

APPENDIX D: PROOF OF THEOREM 2

To prove i) we apply Lemma 4 with $M = \binom{d}{d^*}$ in conjunction with a standard result the proof of which can be found in [11] and in the supplementary material.

LEMMA 10. *Let S be a subset of $\mathbb{Z}^d$ of cardinality $|S|$ and A be a constant. Define μ_S as a discrete measure supported on the finite set of functions $\{\mathsf{f}_{\boldsymbol{\omega}} = \sum_{\boldsymbol{k} \in S} A\omega_{\boldsymbol{k}} \varphi_{\boldsymbol{k}} : \boldsymbol{\omega} \in \{\pm 1\}^S\}$ such that $\mu_S(\mathsf{f} = \mathsf{f}_{\boldsymbol{\omega}}) = 2^{-|S|}$ for every $\boldsymbol{\omega} \in \{\pm 1\}^S$. If we define the probability measure $\mathbb{P}_S$ by $\mathbb{P}_S(A) = \int_{\Sigma(\kappa, L)} \mathbf{P}_{\mathsf{f}}(A) \mu_S(d\mathsf{f})$, for every measurable set $A \subset \mathbb{R}^n$, and $\mathbb{P}_0 = \mathbf{P}_{\mathsf{f}_0}$, then $\mathscr{K}(\mathbb{P}_S, \mathbb{P}_0) \leq |S| A^4 n^2$.*

Without loss of generality, we can assume $\kappa = 1$ (the general case can be reduced to this one by replacing L and n respectively by L/κ and $n\kappa$). Thus, $\vartheta = L$. We denote the set $\Sigma(1, L)$ by Σ_L and choose $\mu_0, \ldots, \mu_M$ as follows: μ_0 is the Dirac measure δ_0, μ_1 is defined as in Lemma 10 with $S = \mathscr{C}_1(d^*, \gamma_L)$ and $A = \left[N(d^*, \gamma_L) \right]^{-1/2}$. The measures $\mu_2, \ldots, \mu_M$ are defined similarly and correspond to the $M - 1$ remaining sparsity patterns of cardinality d^*.

In view of inequality (14) and Lemma 4, it suffices to show that the measures μ_ℓ satisfy $\mu_\ell(\Sigma_L) = 1$ and $\sum_{\ell=0}^M \mathscr{K}(\mathbb{P}_\ell, \mathbb{P}_0) \leq (M+1)\alpha \log M$. Combining Lemma 10 with Card$(S) = N_1(d^*, \gamma_L)$ and inequality (15), we get $\mathscr{K}(\mathbb{P}_\ell, \mathbb{P}_0) \leq \frac{n^2 N_1(d^*, \gamma_L)}{N(d^*, \gamma_L)^2} \leq \frac{n^2 L}{\gamma_L N(d^*, \gamma_L)} \leq \alpha \log M$. Now, let us show that $\mu_1(\Sigma_L) = 1$. By symmetry, this will imply that $\mu_\ell(\Sigma_L) = 1$ for every ℓ. Since μ_1 is supported by the set $\{\mathsf{f}_{\boldsymbol{\omega}} : \boldsymbol{\omega} \in \{\pm 1\}^{\mathscr{C}_1(d^*, \gamma_L)}\}$, it is clear that $\sum_{k_1 \neq 0} \theta_{\boldsymbol{k}}^2[\mathsf{f}_{\boldsymbol{\omega}}] = A^2[N_1(d^*, \gamma_L) - N_2(d^*, \gamma_L)] = 1$ and,

$$\sum_{\boldsymbol{k} \in \mathbb{Z}^d} k_j^2 \theta_{\boldsymbol{k}}^2[\mathsf{f}_{\boldsymbol{\omega}}] = \sum_{\boldsymbol{k} \in \mathscr{C}_1(d^*, \gamma_L)} k_j^2 A^2 = \frac{1}{d^*} \sum_{j=1}^{d^*} \sum_{\boldsymbol{k} \in \mathscr{C}_1(d^*, \gamma_L)} k_j^2 A^2 \leq A^2 \gamma_L N_1(d^*, \gamma_L)$$

$$\leq \gamma_L \frac{N_1(d^*, \gamma_L)}{N(d^*, \gamma_L)}, \qquad j = 1, \ldots, d^*.$$

The results stated in Section 4 imply that $N_1(d^*, \gamma_L)/N(d^*, \gamma_L) \sim_{d^* \to \infty} 1 + (\mathsf{h}(z_\gamma) - 1)^{-1}$. Our choice of γ_L ensures that, for d^* large enough, $\mathsf{f}_{\boldsymbol{\omega}} \in \Sigma_L$. This completes the proof of claim i). To prove ii), we still use Lemma 4 with $\mu_0 = \delta_0$ and $\mu_\ell = \delta_{\mathsf{f}_\ell}$, where for every $\ell \in \{1, \ldots, M\}$, f_ℓ is chosen as follows. Let $I_1, \ldots, I_M$ be all the subsets of $\{1, \ldots, d\}$ containing exactly d^* elements. We define f_ℓ, for $\ell \neq 0$, by its Fourier coefficients $\{\theta_{\boldsymbol{k}}^\ell : \boldsymbol{k} \in \mathbb{Z}^d\}$ as follows:

$$
\theta_{\boldsymbol{k}}^\ell = \begin{cases} 1, & \boldsymbol{k} = (k_1, \ldots, k_d) = (\mathbf{1}_{1 \in I_\ell}, \ldots, \mathbf{1}_{d \in I_\ell}), \\ 0, & \text{otherwise.} \end{cases}
$$

Obviously, all the functions f_ℓ belong to Σ and, moreover, each f_ℓ has I_ℓ as sparsity pattern. One easily checks that our choice of f_ℓ implies $\mathscr{K}(\mathbf{P}_{\mathsf{f}_\ell}, \mathbf{P}_{\mathsf{f}_0}) = n \|\mathsf{f}_\ell - \mathsf{f}_0\|_2^2 = n$. Therefore, if $\alpha \log M = \alpha \log \binom{d}{d^*} \geq n$, the desired inequality is satisfied. To conclude it suffices to note that $\log \binom{d}{d^*} \geq d^* \log(d/d^*)$.

APPENDIX E: PROOF OF PROPOSITION 6

In view of Thm. 1, applied with $A = 2$ and $\tau = 1$, the consistent (uniformly in $\mathsf{f} \in \Sigma(\kappa, L)$) estimation of J is possible if $\frac{8\sqrt{2N(s, 2L/\kappa)d^* \log(d/d^*)} + 16 d^* \log(d/d^*)}{n} \leq \frac{\kappa}{2}$. Since d^*/s is upper-bounded by some constant, there is a constant D_1^* such that the left-hand side of the last display is upper-bounded by

$$
D_1^* \left\{ \frac{\sqrt{N(s, 2L/\kappa)s \log(d/s)}}{n} \bigvee \frac{s \log(d/s)}{n} \right\}.
$$

As proved in Lemma 11 below, $N(s, 2L/\kappa) \leq 0.3 (18\pi e L/\kappa)^{s/2}$. Thus, there is a constant D_2 such that

$$
\left\{ \frac{\sqrt{N(s, 2L/\kappa)s \log(d/s)}}{n} \bigvee \frac{s \log(d/s)}{n} \right\} \leq \frac{D_2^s \kappa^{-s/4} \sqrt{s \log(d/s)}}{n} \bigvee \frac{s \log(d/s)}{n}.
$$

Combining these results, we see that under the conditions $2D_1^* s \log(d/s)/n \leq \kappa$ and

$$
2D_1^* \frac{D_2^s \sqrt{s \log(d/s)}}{n} \leq \kappa^{1 + \frac{s}{4}},
$$

consistent estimation of J is possible. Taking $D^* = 2D_1^*(1 + D_2^4)$, we complete the proof of the first claim of the proposition. To prove the second assertion, we apply Thm. 2. Since it holds that $2\gamma_\vartheta \geq \gamma_\vartheta + 1 \geq \frac{\vartheta}{1 + (\mathsf{h}(z_{\gamma_\vartheta + 1}) - 1)^{-1}} \geq \frac{\vartheta}{1 + (2z_1)^{-1}}$, we deduce from Thm. 2 that there are some constants D_3 and D_4 such that if

$$
D_3 \left\{ \frac{\sqrt{N(s, D_4/\kappa)s \log(d/s)}}{n} \bigvee \frac{s \log(d/s)}{n} \right\} \geq \kappa
$$

then consistent estimation of J is impossible. Since the s-dimensional L_2 ball with radius $\sqrt{s\gamma}$ contains the L_∞ ball of radius $\sqrt{\gamma}$, $N(s, D_4/\kappa) \geq (D_5)^s \kappa^{-s/2}$ for some constant D_5. By rearranging different terms, we get the desired result.

LEMMA 11. *For every $\gamma \geq 1$ and $d^* \in \mathbb{N}$, $N_1(d^*, \gamma) \leq 0.3(9\pi e \gamma)^{d^*/2}$.*

PROOF. One readily checks that if $\|\boldsymbol{k}\|_2^2 \leq d^*\gamma$, then the hypercube centered at $\boldsymbol{k}$ with side of length 1 is included in the ball centered at the origin and having radius $\sqrt{d^*\gamma} + 0.5\sqrt{d^*}$. Therefore, $N_1(d^*, \gamma) \leq (\sqrt{d^*\gamma} + 0.5\sqrt{d^*})^{d^*}\text{Vol}[B_{d^*}(0; 1)]$, where $\text{Vol}[B_{d^*}(0; 1)]$ stands for the volume of the unit ball in $\mathbb{R}^{d^*}$. Using the well-known formula for the latter and the Stirling approximation, for every $d^* \geq 1$, we get $\text{Vol}[B_{d^*}(0; 1)] = \frac{2\pi^{d^*/2}}{d^*\Gamma(d^*/2)} \leq 0.4\frac{(4\pi e/d^*)^{d^*/2}}{\sqrt{2d^*}}$. This implies that $N_1(d^*, \gamma) \leq 0.4\left(\frac{9\gamma d^*}{4}\right)^{d^*/2}\frac{(4\pi e/d^*)^{d^*/2}}{\sqrt{2d^*}} \leq 0.3(9\pi e \gamma)^{d^*/2}$ and the result follows. □

APPENDIX F: SUPPLEMENT TO "TIGHT CONDITIONS FOR CONSISTENCY OF VARIABLE SELECTION IN THE CONTEXT OF HIGH DIMENSIONALITY"

This supplementary material provides the proofs of Theorem 3, Proposition 7, Corollary 3 and Lemma 10 of the article "Tight conditions for consistency of variable selection in the context of high dimensionality".

F.1. Proof of Theorem 3. To ease notation, we write $\widehat{J}$ instead of $\widehat{J}_n^{(1)}$ throughout this proof. The empirical Fourier coefficients can be decomposed as follows:

$$\widehat{\theta}_{\boldsymbol{k}} = \tilde{\theta}_{\boldsymbol{k}} + z_{\boldsymbol{k}}, \quad \text{where} \quad \tilde{\theta}_{\boldsymbol{k}} = \frac{1}{n}\sum_{i=1}^n \frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\text{g}(\boldsymbol{X}_i)}\text{f}(\boldsymbol{X}_i) \text{ and } z_{\boldsymbol{k}} = \frac{\sigma}{n}\sum_{i=1}^n \frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\text{g}(\boldsymbol{X}_i)}\varepsilon_i. \quad (28)$$

If, for a multi index $\boldsymbol{k}$, $\theta_{\boldsymbol{k}} = 0$, then the corresponding empirical Fourier coefficient will be close to zero with high probability. To show this, let us first look at what happens with $z_{\boldsymbol{k}}$'s. We have, for every real number x,

$$\mathbf{P}\big(|z_{\boldsymbol{k}}| > x \,\big|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\big) \leq \exp\left(-\frac{x^2}{2\sigma_{\boldsymbol{k}}^2}\right) \quad \forall \boldsymbol{k} \in S_{m,d^*}$$

with

$$\sigma_{\boldsymbol{k}}^2 = \frac{\sigma^2}{n^2}\sum_{i=1}^n \frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)^2}{\text{g}(\boldsymbol{X}_i)^2} \leq \frac{2\sigma^2}{\text{g}_{\min}^2 n}.$$

Therefore, it holds that $\max_{\boldsymbol{k}\in S_{m,d^*}} \mathbf{P}\big(|z_{\boldsymbol{k}}| > x | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\big) \leq \exp(-n\text{g}_{\min}^2 x^2/4\sigma^2)$. This entails that by setting $\lambda_1 = (8\sigma^2 d^* \log(24\sqrt{\vartheta}d/d^*)/n\text{g}_{\min}^2)^{1/2}$ and by using

the inequalities (cf. Lemma 13 below)

$$\text{Card}(S_{m,d^*}) \leq 0.3 \left(\frac{24\sqrt{\vartheta}d}{d^*} \right)^{d^*},$$

we get

$$\mathbf{P}\left(\max_{\boldsymbol{k} \in S_{m,d^*}} |z_{\boldsymbol{k}}| > \lambda_1 \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \right) \leq \sum_{\boldsymbol{k} \in S_{m,d^*}} \mathbf{P}\left(|z_{\boldsymbol{k}}| > \lambda_1 \,|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \right)$$

$$\leq 2\text{Card}(S_{m,d^*}) e^{-n g_{\min}^2 \lambda_1^2 / 4\sigma^2} \leq 0.6(24\sqrt{\vartheta}d/d^*)^{-d^*}.$$

Next, we use a concentration inequality for controlling large deviations of $\tilde{\theta}_{\boldsymbol{k}}$'s from $\theta_{\boldsymbol{k}}$'s. Recall that in view of the definition $\tilde{\theta}_{\boldsymbol{k}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\mathsf{g}(\boldsymbol{X}_i)} \mathsf{f}(\boldsymbol{X}_i)$, we have $\mathbb{E}(\tilde{\theta}_{\boldsymbol{k}}) = \theta_{\boldsymbol{k}}$. The boundedness of f yields $|\frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\mathsf{g}(\boldsymbol{X}_i)} \mathsf{f}(\boldsymbol{X}_i)| \leq \sqrt{2} L_{\infty}/g_{\min}$. Furthermore, the bound $V = \text{Var}(\frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\mathsf{g}(\boldsymbol{X}_i)} \mathsf{f}(\boldsymbol{X}_i)) \leq \int \mathsf{f}^2(\boldsymbol{x}) \frac{\varphi_{\boldsymbol{k}}^2(\boldsymbol{x})}{\mathsf{g}(\boldsymbol{x})} d\boldsymbol{x} \leq 2L_2^2/g_{\min}^2$ combined with Bernstein's inequality implies that

$$\mathbf{P}(|\tilde{\theta}_{\boldsymbol{k}} - \theta_{\boldsymbol{k}}| > t) \leq 2\exp\left(-\frac{nt^2}{2(V + t\sqrt{2}L_{\infty}/3g_{\min})} \right)$$

$$\leq 2\exp\left(-\frac{g_{\min}^2 nt^2}{4L_2^2 + tL_{\infty}g_{\min}} \right), \qquad \forall t > 0.$$

Let us define $\lambda_2 = 4L_2 \left(\frac{d^* \log(24\sqrt{\vartheta}d/d^*)}{n g_{\min}^2} \right)^{1/2}$. Then,

$$\mathbf{P}(|\tilde{\theta}_{\boldsymbol{k}} - \theta_{\boldsymbol{k}}| > \lambda_2) \leq 2\exp\left(-\frac{4L_2^2 d^* \log(24\sqrt{\vartheta}d/d^*)}{L_2^2 + L_{\infty}L_2 \left(\frac{d^* \log(24\sqrt{\vartheta}d/d^*)}{n} \right)^{1/2}} \right).$$

The first inequality in the main condition of the theorem implies that the denominator in the exponential is not larger than $2L_2^2$. Hence, $\mathbf{P}\left(\max_{\boldsymbol{k} \in S_{m,d^*}} |\tilde{\theta}_{\boldsymbol{k}} - \theta_{\boldsymbol{k}}| > \lambda_2 \right) \leq 0.6/(24\sqrt{\vartheta}d/d^*)^{d^*}$. Let

$$\mathscr{A}_1 = \left\{ \max_{\boldsymbol{k} \in S_{m,d^*}} |z_{\boldsymbol{k}}| \leq \lambda_1 \right\} \quad \text{and} \quad \mathscr{A}_2 = \left\{ \max_{\boldsymbol{k} \in S_{m,d^*}} |\tilde{\theta}_{\boldsymbol{k}}| \leq \lambda_2 \right\}.$$

One easily checks that

$$\mathbf{P}(J^c \not\subset \hat{J}^c) \leq \mathbf{P}(\mathscr{A}_1^c) + \mathbf{P}(\mathscr{A}_2^c) \leq 1.2/(24\sqrt{\vartheta}d/d^*)^{d^*}.$$

Since $d^* \geq 1$ and $\vartheta > 1$, the last inequality implies

$$\mathbf{P}(J^c \not\subset \hat{J}^c) \leq \mathbf{P}(\mathscr{A}_1^c) + \mathbf{P}(\mathscr{A}_2^c) \leq 0.4/(8\sqrt{\vartheta}d/d^*)^{d^*}.$$

As for the converse inclusion, we have

$$\mathbf{P}(J \not\subset \widehat{J}) \le \mathbf{P}\left(\exists j \in J \text{ s.t. } \max_{\boldsymbol{k} \in S_{m,d^*}: k_j \ne 0} |\widehat{\theta}_{\boldsymbol{k}}| \le \lambda\right)$$

$$\le \mathbf{1}\left\{\exists j \in J \text{ s.t. } \max_{\boldsymbol{k} \in S_{m,d^*}: k_j \ne 0} |\theta_{\boldsymbol{k}}| \le 2\lambda\right\} + \mathbf{P}(\mathscr{A}_1^c) + \mathbf{P}(\mathscr{A}_2^c).$$

We show now that the first term in the last line is equal to zero. If this was not the case, then for some value j_0 we would have $Q_{j_0} \ge \kappa$ and $|\theta_{\boldsymbol{k}}| \le 2\lambda$, for all $\boldsymbol{k} \in S_{m,d^*}$ such that $k_{j_0} \ne 0$. This would imply that

$$Q_{j_0,m,d^*} = \sum_{\boldsymbol{k} \in S_{m,d^*}^{j_0}} \theta_{\boldsymbol{k}}^2 \le 4\lambda^2 N(d^*, m^2/d^*).$$

On the other hand,

$$Q_{j_0} - Q_{j_0,m,d^*} \le \sum_{\|\boldsymbol{k}\|_2 \ge m} \theta_{\boldsymbol{k}}^2 \le m^{-2} \sum_{\|\boldsymbol{k}\|_2 \ge m} \sum_{j \in J} k_j^2 \theta_{\boldsymbol{k}}^2 \le \frac{Ld^*}{m^2}.$$

Remark now that the choice of the truncation parameter m proposed in the statement of the proposition implies that $Q_{j_0} - Q_{j_0,m,d^*} \le \kappa/2$. Combining these estimates, we get

$$Q_{j_0} \le \frac{\kappa}{2} + 4\lambda^2 N(d^*, m^2/d^*),$$

which is impossible since $Q_{j_0} \ge \kappa$.

F.2. Proof of Proposition 7. Let $M = \binom{d}{d^*}$ and let $\{f_0, f_1, \ldots, f_M\}$ be a set included in Σ_L. Let $I_1, \ldots, I_M$ be all the subsets of $\{1, \ldots, d\}$ containing exactly d^* elements somehow enumerated. Let us set $f_0 \equiv 0$ and define f_ℓ, for $\ell \ne 0$, by its Fourier coefficients $\{\theta_{\boldsymbol{k}}^\ell : \boldsymbol{k} \in \mathbb{Z}^d\}$ as follows:

$$\theta_{\boldsymbol{k}}^\ell = \begin{cases} 1, & \boldsymbol{k} = (k_1, \ldots, k_d) = (\mathbf{1}_{1 \in I_\ell}, \ldots, \mathbf{1}_{d \in I_\ell}), \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, all the functions f_ℓ belong to Σ and, moreover, each f_ℓ has I_ℓ as sparsity pattern. One easily checks that our choice of f_ℓ implies $\mathscr{K}(\mathbf{P}_{f_\ell}, \mathbf{P}_{f_0}) = n\|f_\ell - f_0\|_2^2 = n$. Therefore, if $\alpha \log M = \alpha \log \binom{d}{d^*} \ge n$, the desired inequality is satisfied. To conclude it suffices to note that $\log \binom{d}{d^*}$ is larger than or equal to $d^* \log(d/d^*) = d^*(\log d - \log d^*)$.

F.3. Proof of Lemma 10. First we specify the notation. Let $\boldsymbol{y} = \{y_{\boldsymbol{k}} : \boldsymbol{k} \in S\}$ and, for $\boldsymbol{\omega} \in \{\pm 1\}^S$, $\boldsymbol{\theta}_{\boldsymbol{\omega}} = \{A\boldsymbol{\omega}_{\boldsymbol{k}} : \boldsymbol{k} \in S\}$. The likelihood ratio between $\mathbf{P}_{\boldsymbol{\omega}} = \mathbf{P}_{f_{\boldsymbol{\omega}}}$ and $\mathbf{P}_0 = \mathbf{P}_{f_0}$ is

$$\frac{d\mathbf{P}_{\boldsymbol{\omega}}}{d\mathbf{P}_0}(\boldsymbol{y}) = \exp\left[-\frac{n\boldsymbol{\theta}_{\boldsymbol{\omega}}^2}{2} + n\boldsymbol{\theta}_{\boldsymbol{\omega}} \cdot \boldsymbol{y} \right],$$

where, for a vector $\boldsymbol{z} \in \mathbb{Z}^d$, $\boldsymbol{z}^2 = \boldsymbol{z} \cdot \boldsymbol{z}$. Then, the likelihood ratio between $\mathbb{P}_S$ and $\mathbb{P}_0$ is

$$\frac{d\mathbb{P}_S}{d\mathbb{P}_0}(\boldsymbol{y}) = \exp\left(-\frac{nA^2|S|}{2} \right) \sum_{\boldsymbol{\omega} \in \{\pm 1\}^S} \frac{1}{2^{|S|}} \exp\left(n\boldsymbol{\theta}_{\boldsymbol{\omega}} \cdot \boldsymbol{y} \right).$$

As a consequence, simple algebra yields

$$\int \left(\frac{d\mathbb{P}_S}{d\mathbb{P}_0}(\boldsymbol{y}) \right)^2 \mathbb{P}_0(d\boldsymbol{y}) = \left(\frac{\exp(nA^2) + \exp(-nA^2)}{2} \right)^{|S|} \leq \exp(|S|n^2 A^4).$$

The last inequality follows from the elementary inequality $\cosh(x) \leq e^{x^2}$, which can be checked by decomposing the functions $\cosh(x)$ and e^{x^2} in Taylor series and comparing the corresponding terms.

F.4. Proof of Corollary 3. Let us set $\gamma = L/\kappa$ and $\gamma_\tau = (1 + \tau)\gamma$. Applying Theorem 1 with $A = 2$, we get that

$$\mathbf{P}\left(\widehat{J}_n \neq J \right) \leq 3(2ed/d^*)^{-d^*}$$

provided that the condition

$$\frac{8\sqrt{2N(d^*, \gamma_\tau)d^* \log(2ed/d^*)}}{n} + \frac{16d^* \log(2ed/d^*)}{n} \leq \frac{\kappa\tau}{1 + \tau}. \qquad (29)$$

is satisfied for some $\tau > 0$. Clearly, when $d \to \infty$, for every $d^* \geq 1$, it holds that

$$(2ed/d^*)^{-d^*} \leq d^{-1/2} \to 0.$$

Therefore, it is sufficient to check that the assumptions

$$\lim_{n \to \infty} \frac{\log\log d}{\log n} = 0, \qquad \limsup_{n \to \infty} \frac{d^*}{\log n} < \frac{2}{l_\gamma(z_\gamma)} \qquad (30)$$

imply that (29) is true for sufficiently large values of n. We will show that the left-hand side of 29 tends to 0 as $n \to \infty$.

First remark that (30) yields

$$\log d \leq n^{1/3}, \qquad d^* \leq n^{1/3}$$

for sufficiently large n. Therefore,

$$\lim_{n\to\infty} \frac{16d^* \log(2ed/d^*)}{n} = 0.$$

Second, by continuity of the mappings $\gamma \mapsto l_\gamma$ and $\gamma \mapsto z_\gamma$, there exists $\tau > 0$ such that

$$\limsup_{n\to\infty} \frac{d^*}{\log n} < \frac{2}{l_{\gamma_\tau}(z_{\gamma_\tau})} \iff \limsup_{n\to\infty} \left(\frac{d^* l_{\gamma_\tau}(z_{\gamma_\tau})}{2\log n} - 1 \right) < 0. \qquad (31)$$

This inequality, combined with the relation $\log N(d^*, \gamma_\tau) = d^* l_{\gamma_\tau}(z_{\gamma_\tau})(1 + o(1))$ (cf. Eq. (12) in the manuscript [12]), implies that

$$\log\left\{ \frac{\sqrt{N(d^*,\gamma_\tau)d^* \log(2ed/d^*)}}{n} \right\}$$

$$\leq \log n \left(\frac{\log N(d^*,\gamma_\tau)}{2\log n} - 1 + \frac{\log d^*}{2\log n} + \frac{\log\log(2ed)}{2\log n} \right)$$

$$\leq \log n \left(\underbrace{\frac{d^* l_{\gamma_\tau}(z_{\gamma_\tau})}{2\log n}(1 + o(1)) - 1}_{\text{tends to a negative number}} + \underbrace{\frac{\log d^*}{d^*} \frac{d^*}{2\log n} + \frac{\log\log d}{\log n}}_{\text{tends to 0}} \right)$$

$$\to -\infty.$$

This entails that the first term in the left-hand side of (29) tends to zero, which completes the proof.

F.5. Some technical lemmas.

LEMMA 12. *For every $\gamma \geq 1$ the numbers $N_1(d^*,\gamma) = \{k \in \mathbb{Z}^{d^*} : \|k\|_2^2 \leq d^*\gamma\}$ admit the following upper bound:*

$$N_1(d^*,\gamma) \leq 0.3\left(9\pi e\gamma\right)^{d^*/2}.$$

PROOF. One readily checks that if $\|k\|_2^2 \leq d^*\gamma$, then the hypercube centered at k with side of length 1 is included in the ball centered at the origin and having radius $m + 0.5\sqrt{d^*}$. Therefore,

$$N_1(d^*,\gamma) \leq (\sqrt{d^*\gamma} + 0.5\sqrt{d^*})^{d^*} \text{Vol}[B_{d^*}(0;1)],$$

where $\text{Vol}[B_{d^*}(0;1)]$ stands for the volume of unit ball in $\mathbb{R}^{d^*}$. Using the well-known formula for the latter and the Stirling approximation, for every $d^* \geq 1$, we get:

$$\text{Vol}[B_{d^*}(0;1)] = \frac{2\pi^{d^*/2}}{d^*\Gamma(d^*/2)} \in \left(0.3\frac{(4\pi e/d^*)^{d^*/2}}{\sqrt{2d^*}}; 0.4\frac{(4\pi e/d^*)^{d^*/2}}{\sqrt{2d^*}} \right).$$

This implies that

$$N_1(d^*,\gamma) \leq \left(\frac{9\gamma d^*}{4}\right)^{d^*/2} \mathrm{Vol}[B_{d^*}(0;1)] \leq 0.4\left(\frac{9\gamma d^*}{4}\right)^{d^*/2} \frac{(4\pi e/d^*)^{d^*/2}}{\sqrt{2d^*}} \leq 0.3(9\pi e\gamma)^{d^*/2}.$$

This proves the result.   □

LEMMA 13.   *Let $S_{m,d^*} = \{k \in \mathbb{Z}^d : \|k\|_2 \leq m, \ \& \ \|k\|_0 \leq d^*\}$. If $m = \sqrt{\gamma d^*}$ with $\gamma \geq 1$, then $\mathrm{Card}(S_{m,d^*}) \leq 0.3(24\sqrt{\gamma}d/d^*)^{d^*}$.*

PROOF. It is clear that

$$\mathrm{Card}(S_{m,d^*}) \leq \binom{d}{d^*} N_1(d^*,\gamma).$$

Combining this with the inequality

$$\binom{d}{d^*} \leq \left(\frac{ed}{d^*}\right)^{d^*},$$

and the previous lemma, we get

$$\mathrm{Card}(S_{m,d^*}) \leq \left(\frac{ed}{d^*}\right)^{d^*} 0.3(9\pi e\gamma)^{d^*/2}.$$

The claim of the lemma follows now from the inequlaity $\sqrt{9\pi e}e \leq 24$.   □

LEMMA 14.   *For every pair of positive integers (d,d^*) such that $d^* \leq d$:*

$$\sum_{\ell=1}^{d^*} \ell\binom{d}{\ell} \leq (2ed/d^*)^{d^*}.$$

PROOF. We will proceed by induction over d^*. If $d^* = 1$, we have

$$\sum_{\ell=1}^{d^*} \ell\binom{d}{\ell} = d \leq ed = (ed/d^*)^{d^*}.$$

Assume that the inequality

$$\sum_{\ell=1}^{d^*} \ell\binom{d}{\ell} \leq (2ed/d^*)^{d^*}$$

is true for some $1 \leq d^* < d$. Let us show that this entails the inequality

$$\sum_{\ell=1}^{d^*+1} \ell\binom{d}{\ell} \leq (2ed/(d^*+1))^{d^*+1}.$$

It holds that

$$\sum_{\ell=1}^{d^*+1} \ell \binom{d}{\ell} \le (2ed/d^*)^{d^*} + d\binom{d-1}{d^*}$$

$$= (2ed/(d^*+1))^{d^*+1} \frac{d^*+1}{2ed}\left(1+\frac{1}{d^*}\right)^{d^*} + d\left(\frac{e(d-1)}{d^*}\right)^{d^*}$$

$$= (2ed/(d^*+1))^{d^*+1} \frac{1}{2e}\left(1+\frac{1}{d^*}\right)^{d^*}\left(\frac{d^*+1}{d} + \frac{d^*+1}{2^{d^*}}(1-1/d)^{d^*}\right).$$

Using that $(1+1/d^*)^{d^*} \le e$, $d^* \ge 1$ and $d \ge d^*+1$, we get

$$\sum_{\ell=1}^{d^*+1} \ell \binom{d}{\ell} \le (2ed/(d^*+1))^{d^*+1} \frac{1}{2}\underbrace{\left(\frac{d^*+1}{d} + \frac{d^*+1}{2^{d^*}}\right)}_{\le 1+1}$$

and the result follows.   □

REFERENCES

[1] Hirotsugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[2] Pierre Alquier. Iterative feature selection in least square regression estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 44(1):47–88, 2008.

[3] Francis Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. Technical report, arXiv:0909.0844, 2009.

[4] Karine Bertin and Guillaume Lecué. Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.*, 2:1224–1241, 2008.

[5] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Hierarchical selection of variables in sparse high-dimensional regression. *Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown. IMS Collections*, 6:56–69, 2010.

[6] Lawrence D. Brown and Mark G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398, 1996.

[7] Lawrence D. Brown, Andrew V. Carter, Mark G. Low, and Cun-Hui Zhang. Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.*, 32(5):2074–2097, 2004.

[8] Florentina Bunea and Adrian Barbu. Dimension reduction and variable selection in case control studies via regularized likelihood optimization. *Electron. J. Stat.*, 3:1257–1287, 2009.

[9] Tony T. Cai and Mark G. Low. Optimal adaptive estimation of a quadratic functional. *Ann. Statist.*, 34(5):2298–2325, 2006.

[10] Laetitia Comminges. Conditions minimales de consistance pour la sélection de variables en grande dimension. *Comptes Rendus Mathematique*, 349(7-8):469–472, feb 2011.

[11] Laëtitia Comminges and Arnak S. Dalalyan. Tight conditions for consistent variable selection in high dimensional nonparametric regression. *J.Mach. Learn. Res. - Proceedings Track*, 19:187–206, 2011.

[12] Laëtitia Comminges and Arnak S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *submitted*, arxiv:math.ST/1106.4293v2, 2012.

[13] Arnak Dalalyan and Markus Reiß. Asymptotic statistical equivalence for scalar ergodic diffusions. *Probab. Theory Related Fields*, 134(2):248–282, 2006.

[14] Jean Dieudonné. *Calcul infinitésimal*. Hermann, Paris, 1968.

[15] David Donoho and Jiashun Jin. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 367(1906):4449–4470, 2009.

[16] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

[17] Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory*, 57(8):5467–5484, 2011.

[18] Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, 10:2013–2038, 2009.

[19] Ghislaine Gayraud and Yuri Ingster. Detection of sparse variable functions. Technical report, arXiv:1011.6369v3, 2011.

[20] Mohamed Hebiri. Sparse conformal predictors. *Statistics and Computing*, 20:253–266, April 2010.

[21] Junzhou Huang and Tong Zhang. The benefit of group sparsity. *Ann. Statist.*, 38(4):1978–2004, 2010.

[22] Yuri Ingster and Natalia Stepanova. Estimation and detection of functions from anisotropic Sobolev classes. *Electron. J. Stat.*, 5:484–506, 2011.

[23] Yuri Ingster and Irina Suslina. Estimation and hypothesis testing for functions from tensor products of spaces. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 351(Veroyatnost i Statistika. 12):180–218, 301–302, 2007.

[24] Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *J.Mach. Learn. Res.*, 12 (Oct):2777–2824, 2011.

[25] Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *Ann. Statist.*, 38(6):3660–3695, 2010.

[26] John Lafferty and Larry Wasserman. Rodeo: sparse, greedy nonparametric regression. *Ann. Statist.*, 36(1):28–63, 2008.

[27] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[28] Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara van de Geer. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39 (4):2164–2204, 2011.

[29] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, Nov. 1973.

[30] James Mazo and Andrew Odlyzko. Lattice points in high-dimensional spheres. *Monatsh. Math.*, 110(1):47–61, 1990.

[31] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B*

Stat. Methodol., 72(4):417–473, 2010.

[32] Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *Ann. Statist.*, 39(1):1–47, 2011.

[33] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. Technical report, arXiv:1008.3654v2, 2011.

[34] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.*, 38(3): 1287–1319, 2010.

[35] Markus Reiß. Asymptotic equivalence for nonparametric regression with multivariate and random design. *Ann. Statist.*, 36(4):1957–1982, 2008.

[36] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[37] James G. Scott and James O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.*, 38(5):2587–2619, 2010.

[38] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[39] Alexandre B. Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009.

[40] Nicolas Verzelen. Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. *Electron. J. Statist.*, 6:38–90, 2012.

[41] Martin J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory*, 55(12):5728–5741, 2009.

[42] Larry Wasserman and Kathryn Roeder. High-dimensional variable selection. *Ann. Statist.*, 37(5A):2178–2201, 2009.

[43] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

[44] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.

[45] Tong Zhang. On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.*, 10:555–568, 2009.

[46] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

[47] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.*, 37(6A):3468–3497, 2009.

LIGM/IMAGINE
77455 MARNE-LA-VALLÉE CEDEX 2 - FRANCE
E-MAIL: laetitia.comminges,dalalyan@imagine.enpc.fr