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

Laëtitia Comminges, Arnak S. Dalalyan

Submitted

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

BY LAËTITIA COMMINGES AND ARNAK S. DALALYAN

Université Paris Est/ ENPC

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 procedures that are provably consistent under these tight conditions are simple: they are based on comparing the empirical Fourier coefficients with an appropriately chosen threshold value.

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 completely 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 $\log n$.

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 (1973), Akaike (1973), Schwarz (1978) who introduced respectively the famous criteria C_p, AIC and BIC, the problem of variable selection has been 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 in the case where 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 the ambient dimension d.

In the important particular case of linear regression, the latter scenario has been the subject of a number of recent studies. Many of them rely on ℓ_1-norm penalization—as for instance in (Tibshirani, 1996, Zhao and Yu, 2006, Meinshausen and Bühlmann, 2010)—and constitute an attractive

AMS 2000 subject classifications: Primary 62G08, ; secondary 62H12
Keywords and phrases: variable selection, nonparametric regression, set estimation, sparsity pattern

alternative to iterative variable selection procedures proposed by Alquier (2008), Zhang (2009), Ting et al. (2010) and to marginal regression or correlation screening explored in (Wasserman and Roeder, 2009, Fan et al., 2009). Promising results for feature selection are also obtained by conformal prediction (Hebiri, 2010), minimax concave penalties in (Zhang, 2010), by Bayesian approach in (Scott and Berger, 2010) and by higher criticism in (Donoho and Jin, 2009). Extensions to other settings including logistic regression, generalized linear model and Ising model have been carried out in (Bunea and Barbu, 2009, Ravikumar et al., 2010, Fan et al., 2009), respectively. Variable selection in the context of groups of variables with disjoint or overlapping groups has been studied by Jenatton et al. (2009), Lounici et al. (2010), Obozinski et al. (2011). Hierarchical procedures for selection of relevant variables have been proposed by Bach (2009), Bickel et al. (2010) and Zhao et al. (2009).

It is now well understood that in the high-dimensional linear regression, if the Gram matrix satisfies some variant of irrepresentable condition, then 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 $n \to \infty$. 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. 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 (Lafferty and Wasserman, 2008) and (Bertin and Lecué, 2008), closely related to the present work, consider the problem of variable selection in nonparametric Gaussian regression model. They prove the consistency of the proposed procedures under some assumptions that—in the light of the present work—turn out to be suboptimal. More precisely, in (Lafferty and Wasserman, 2008), the unknown regression function is assumed to be four times continuously differentiable with bounded derivatives. The algorithm they propose, termed Rodeo, is a greedy procedure performing simultaneously local bandwidth choice and variable selection. Under the assumption that the density of the sampling design is continuously differentiable and strictly positive, 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 (2008) propose a procedure based on the ℓ_1-penalization of local polynomial estimators and prove its consistency when $d^* = O(1)$ but d is allowed to be as large as $\log n$, up to a multiplicative constant. They also have a weaker assumption on the regression function which is merely assumed to belong to the Holder class with smoothness $\beta > 1$.

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 well chosen empirical Fourier coefficients to a prescribed significance level. 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 $(\log d)/n \le c_*$ the estimator proposed in Section 3 is consistent and (b) no estimator of the sparsity pattern may be consistent if $(\log d)/n \ge c^*$.
- 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 $\underline{c}_1 d^* + \log\log d - \log n < \underline{c}_2$ the estimator proposed in Section 3 is consistent and (b) no estimator of the sparsity pattern may be consistent if $\bar{c}_1 d^* + \log\log d - \log n > \bar{c}_2$.
- 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 $d^* \le c_0 \log n$ the estimator proposed in Section 3 is consistent and (b) no estimator of the sparsity pattern may be consistent if $d^* \ge c^0 \log n$.

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, two estimators of the set of relevant variables are introduced and their consistency is established, in the case where the data come from the Gaussian white noise model. The main condition required in the consistency results 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 via the Jacobi θ-function and the saddle point method. Results on impossibility of consistent estimation of the sparsity pattern are derived in Section 5. Then, in Section 6, we show that some of our results can be extended to the model of nonparametric regression, under some additional assumptions, which are quite common in the context of regression. The relation between consistency and inconsistency results are discussed in Section 7. 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 $\mathsf{f} : [0,1]^d \to \mathbb{R}$ 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 $\{1,\dots,d\}$ satisfying $\mathrm{Card}(J) \le d^*$, there is a function $\bar{\mathsf{f}} : \mathbb{R}^{\mathrm{Card}(J)} \to \mathbb{R}$ such that

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

where $\boldsymbol{x}_J$ stands for the subvector of $\boldsymbol{x}$ obtained by removing from $\boldsymbol{x}$ all the coordinates with in-

dices 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 Card(J); d^* is merely a known upper bound on the intrinsic dimension.

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

$$\varphi_{\boldsymbol{k}}(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{k} = 0, \\ \sqrt{2}\cos(2\pi\,\boldsymbol{k}\cdot\boldsymbol{x}), & \boldsymbol{k} \in (\mathbb{Z}^d)_+, \\ \sqrt{2}\sin(2\pi\,\boldsymbol{k}\cdot\boldsymbol{x}), & -\boldsymbol{k} \in (\mathbb{Z}^d)_+, \end{cases} \tag{1}$$

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 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

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

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 $J \subset \{1,\dots,d\}$ of cardinality $\le d^*$, 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

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

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 $\int_{[0,1]^d} \big(\mathsf{f}(\boldsymbol{x}) - \int_0^1 \mathsf{f}(\boldsymbol{x})\,dx_j\big)^2 d\boldsymbol{x} \ge \kappa$ 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 (Comminges, 2011) for some additional details. 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. In order 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 has been proved to be asymptotically equivalent to the model of regression (Brown and Low, 1996, Carter, 2007, Reiß, 2008), as well as to other nonparametric models (Brown et al., 2004, Dalalyan and Reiß, 2006, Golubev et al., 2010). 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

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

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

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

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) has been studied by Ingster and Suslina (2007).

3.1. *Estimation of J by thresholding Fourier coefficients.* A natural approach for variable selection consists in computing the Fourier coefficients of the observed signal and in comparing them to a properly chosen threshold. Since the trigonometric basis is orthonormal and contains the constant function, it follows that

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

It is not difficult to show that the converse is also true. Therefore, one has

$$J = \Big\{ j \in \{1,\dots,d\} : \exists \boldsymbol{k} \ \text{ s.t. } \ k_j \neq 0 \ \text{ and } \ \theta_{\boldsymbol{k}}[\mathsf{f}] \neq 0 \Big\}. \tag{4}$$

Our first 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

$$y_{\boldsymbol{k}} = \theta_{\boldsymbol{k}}[\mathsf{f}] + n^{-1/2}\,\xi_{\boldsymbol{k}}, \qquad \theta_{\boldsymbol{k}} = \langle f, \varphi_{\boldsymbol{k}}\rangle, \quad \boldsymbol{k} \in \mathbb{Z}^d, \tag{5}$$

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, $y_{\boldsymbol{k}}$ is a good estimate (unbiased and with a mean squared error equal to $1/n$) of $\theta_{\boldsymbol{k}}[\mathsf{f}]$ and, therefore, it seems natural to estimate the set J by the set

$$\tilde{J}_\lambda = \Big\{ j \in \{1,\dots,d\} : \sup_{\boldsymbol{k} \in \mathbb{Z}^d : k_j \neq 0} |y_{\boldsymbol{k}}| > \lambda \Big\}, \tag{6}$$

for some positive parameter λ. However, this estimator suffers from two major flows. First, it is impossible to compute, since for every j an infinite number of comparisons should be performed. Second, even if we neglect the computational aspect, the estimator $\tilde{J}_\lambda$ is definitely not consistent since for every λ and for every $j \notin J$, one can find (with probability one) a random variable among the infinite sequence $\{\xi_{\boldsymbol{k}} : k_j \neq 0\}$ that exceeds the threshold λ in absolute value. Therefore, the probability of having $\tilde{J}_\lambda = \{1,\dots,d\}$ is equal to one.

To cope with this issue, we restrict the set of indices for which the comparison with the threshold is performed. We use the standard notation for the vector norms:

$$\|\boldsymbol{x}\|_0 = \sum_{j=1}^{d} \mathbf{1}(x_j \neq 0), \qquad \|\boldsymbol{x}\|_p^p = \sum_{j=1}^{d} |x_j|^p, \ \forall p \in [1,\infty), \qquad \|\boldsymbol{x}\|_\infty = \sup_{j=1,\dots,d} |x_j|,$$

for every $\boldsymbol{x} \in \mathbb{R}^d$. Let us fix an integer $m > 0$ and denote, for $j \in \{1,\dots,d\}$,

$$S^j_{m,d^*} = \Big\{ \boldsymbol{k} \in \mathbb{Z}^d : \|\boldsymbol{k}\|_2 \leq m, \quad \|\boldsymbol{k}\|_0 \leq d^* \ \text{ and } \ \boldsymbol{k}_j \neq 0 \Big\}.$$

Using the sets S^j_{m,d^*}, we define the estimator of J by the formula

$$\widehat{J}_{1,n}(m,\lambda) = \Big\{ j \in \{1,\dots,d\} : \max_{\boldsymbol{k} \in S^j_{m,d^*}} |y_{\boldsymbol{k}}| > \lambda \Big\}. \tag{7}$$

This estimator depends on two parameters, m and λ, having natural physical meaning. The parameter m allows to adapt the estimator to the smoothness of the underlying regression function, while λ is the threshold used for discriminating significant coefficients from the noise. Thus, it is not surprising that in the next theorem, m is a decreasing function of L and λ is proportional to the noise level $n^{-1/2}$.

THEOREM 1. *Let condition* $[\mathbf{C1}(\kappa, L)]$ *be satisfied with some known constants* $\kappa > 0$ *and* $L < \infty$. *For some real number* $A > 1$, *set*

$$m = (2Ld^*/\kappa)^{1/2} \qquad \text{and} \qquad \lambda = \Big(\frac{2Ad^* \log(6md)}{n} \Big)^{1/2}.$$

For any $\gamma > 0$, *let us define* $N(d^*, \gamma)$ *as the cardinality of the set* $\{\boldsymbol{k} \in \mathbb{Z}^{d^*} : \|\boldsymbol{k}\|_2^2 \leq \gamma d^* \,\&\, k_1 \neq 0\}$. *If*

$$\frac{d^* N\big(d^*, 2L/\kappa\big) \log(6md)}{n} \leq \frac{\kappa}{16A} \tag{8}$$

then $\widehat{J}_1(m,\lambda)$ *is a consistent estimator of* J *and the probability of the event* $\widehat{J}_1(m,\lambda) \neq J$ *is upper bounded by* $6d^{-d^*(A-1)}$.

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

$$\log d \leq \mathrm{Const} \cdot n. \tag{9}$$

Although we did not find this result in the statistical literature on variable selection, it can be checked that (9) 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 (9) is neglected. Furthermore, the estimator of J which is proved to be consistent under condition (9) does not really exploit the structure of the Fourier coefficients of the regression function.

3.2. *Estimation of J by group-thresholding.* In the regime of growing intrinsic dimension $d^* \to \infty$, the term $N(d^*, m^2/d^*)$ present in (8) tends to infinity as well. Furthermore, as we show in Section 4, this convergence takes place at an exponential rate in d^*. It is possible then to construct an estimator of J which is consistent under a weaker condition than the one given by (8). Roughly speaking, we are going to demonstrate that if in (8) the quantity $N(d^*, m^2/d^*)$ is replaced by its square root, then consistent estimation of J is still possible and can be done by a group-thresholding procedure.

We now consider an estimator of J that is based on testing subsets of variables. For every $\ell \in \{1,\dots,d^*\}$, we denote by P_ℓ^d the set of all subsets I of $\{1,\dots,d\}$ having exactly ℓ elements:

$$P_\ell^d = \Big\{ I \subset \{1,\dots,d\} \,:\, \mathrm{Card}(I) = \ell \Big\}.$$

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 notation

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

Using this notation, one easily checks that for every $j \in J$, the following relation holds true: $Q_j = \lim_{m\to\infty} \sum_{\boldsymbol{k}\in S_{m,J}^j} \theta_{\boldsymbol{k}}^2 \ge \kappa$. Therefore, for every $j \in J$ and $\tau > 0$, and for m large enough, there exists a set $I \in P_{d^*}^d$ such that $Q_{m,I}^j \triangleq \sum_{\boldsymbol{k}\in S_{m,I}^j} \theta_{\boldsymbol{k}}^2 \ge \kappa\tau/(L+\tau)$. This property, combined with the fact that

$$\widehat{Q}_{m,I}^j \triangleq \sum_{\boldsymbol{k}\in S_{m,I}^j} \Big(y_{\boldsymbol{k}}^2 - \frac{1}{n} \Big)$$

is an unbiased estimate of $Q_{m,I}^j$, leads us to define an estimator of the set J by

$$\widehat{J}_{2,n}(\boldsymbol{m},\boldsymbol{\lambda}) = \Big\{ j \in \{1,\dots,d\} \,:\, \max_{\ell \le d^*} \lambda_\ell^{-1} \max_{I \in P_\ell^d} \widehat{Q}_{m_\ell,I}^j \ge 1 \Big\}.$$

where $\boldsymbol{m} = (m_1,\dots,m_{d^*}) \in \mathbb{N}^{d^*}$ and $\boldsymbol{\lambda} = (\lambda_1,\dots,\lambda_{d^*}) \in \mathbb{R}_+^{d^*}$ are two vectors of tuning parameters.

THEOREM 2. *Let condition* [**C1**(κ, L)] *be satisfied with some known constants $\kappa > 0$ and $L < \infty$. For some real numbers $\tau > 0$ and $A > 1$, set*

$$m_\ell = \big((L+\tau)\ell/\kappa\big)^{1/2} \qquad \text{and} \qquad \lambda_\ell = \frac{2\sqrt{AN(\ell, m_\ell^2/\ell) d^* \log d} + 2Ad^* \log d}{n}.$$

If the condition

$$\frac{8\sqrt{AN(d^*, (L+\tau)/\kappa) d^* \log d} + 10Ad^* \log d}{n} \le \frac{\kappa\tau}{L+\tau} \tag{10}$$

is fulfilled, then the estimator $\widehat{J}_{2,n}(\boldsymbol{m},\boldsymbol{\lambda})$ is consistent and $\mathbf{P}\Big(\widehat{J}_{2,n}(\boldsymbol{m},\boldsymbol{\lambda}) \ne J\Big) \le 6d^{-d^*(A-1)}$.

In order to fully understand the constraints imposed on d, d^* and n by condition (10), we need to describe the behavior of $N(d^*,\gamma)$ when $\gamma>0$ is fixed and $d^*\to\infty$. This will be done in the next section. Let us simply note here that condition (10) is obviously weaker than condition (8), since the latter requires that $N(d^*,\gamma)d^*\log d$ is bounded by n up to a multiplicative constant, while the former requires from the same quantity to be bounded by n^2, still up to a multiplicative constant.

Another important observation concerns the number of tuning parameters involved in each estimation procedure $\widehat{f}_{1,n}$ and $\widehat{f}_{2,n}$. In fact, Theorems 1 and 2 reveal that the genuine tuning parameters are not (m,λ) or $(\boldsymbol{m},\boldsymbol{\lambda})$, but rather $\vartheta=(L+\tau)/\kappa$. Indeed, once the value of $(L+\tau)/\kappa$ is given, the parameters (m,λ) and $(\boldsymbol{m},\boldsymbol{\lambda})$ can be computed in a unique manner. Since in practice neither L nor κ is known, the optimal value of ϑ should be determined from the data. We envisage to tackle this issue in a future work.

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. In order to show this, let us introduce some notation. For a positive number γ, we set

$$\mathscr{C}_1(d^*,\gamma)=\Big\{\boldsymbol{k}\in\mathbb{Z}^{d^*}:k_1^2+\ldots+k_{d^*}^2\le\gamma d^*\Big\},\quad \mathscr{C}_2(d^*,\gamma)=\Big\{\boldsymbol{k}\in\mathbb{Z}^{d^*}:k_2^2+\ldots+k_{d^*}^2\le\gamma d^*\ \&\ k_1=0\Big\}$$

along with $N_1(d^*,\gamma)=\mathrm{Card}\mathscr{C}_1(d^*,\gamma)$ and $N_2(d^*,\gamma)=\mathrm{Card}\mathscr{C}_2(d^*,\gamma)$. In simple words, $N_1(d^*,\gamma)$ is the number of (integer) 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 $N(d^*,\cdot)$ of Theorems 1 and 2 can be written as $N_1(d^*,\cdot)-N_2(d^*,\cdot)$. Using relatively elementary volumetric arguments, one can check that $V(d^*)(\sqrt{\gamma}-1)^{d^*}(d^*)^{d^*/2}\le N_1(d^*,\gamma)\le V(d^*)(\sqrt{\gamma}+1)^{d^*}(d^*)^{d^*/2}$, 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 $\mathsf{h}(z)^{d^*}=\sum_{r\ge0}a_rz^r$. One easily checks that $\forall r\in\mathbb{N}$, $a_r=\mathrm{Card}\{\boldsymbol{k}\in\mathbb{Z}^{d^*}:k_1^2+\ldots+k_{d^*}^2=r\}$. 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 :

$$N_1(d^*,\gamma)=\frac{1}{2\pi i}\oint\frac{\mathsf{h}(z)^{d^*}}{z^{\gamma d^*}}\frac{dz}{z(1-z)}.$$

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 Dieudonné (1968), we get the following result.

PROPOSITION 1. *Let $\gamma>0$ be an integer and let $\mathsf{l}_\gamma(z)=\log\mathsf{h}(z)-\gamma\log z$.*

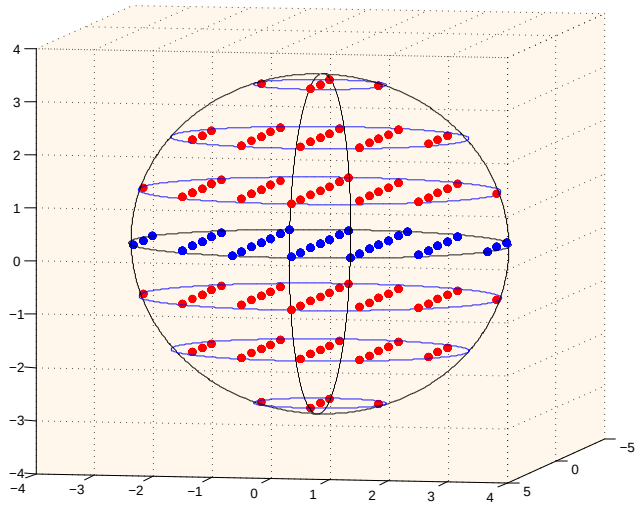

FIG 1. *Lattice points in a ball of radius $R = \gamma d^* = 3.2$ in the three dimensional space ($d^* = 3$). Red points are those of $\mathscr{C}_2(d^*,\gamma)$ while blue points are those of $\mathscr{C}_1(d^*,\gamma)\setminus\mathscr{C}_2(d^*,\gamma)$. In this example, $N(d^*,\gamma) = N(3,1.07) = 110$.*

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. *The following equivalences hold true:*

$$N_1(d^*,\gamma) = \left(\frac{\mathsf{h}(z_\gamma)}{z_\gamma^\gamma}\right)^{d^*} \frac{1+o(1)}{z_\gamma(1-z_\gamma)(2\mathsf{l}''_\gamma(z_\gamma)\pi d^*)^{1/2}},$$

$$N_2(d^*,\gamma) = \left(\frac{\mathsf{h}(z_\gamma)}{z_\gamma^\gamma}\right)^{d^*} \frac{1+o(1)}{\mathsf{h}(z_\gamma)z_\gamma(1-z_\gamma)(2\mathsf{l}''_\gamma(z_\gamma)\pi d^*)^{1/2}},$$

as d^* tends to infinity.

In the sequel, it will be useful to remark that the second part of Proposition 1 yields

$$\log\big(N_1(d^*,\gamma) - N_2(d^*,\gamma)\big) = d^*\mathsf{l}_\gamma(z_\gamma) - \frac{1}{2}\log d^* + c_\gamma + o(1), \qquad \text{as} \quad d^* \to \infty, \tag{11}$$

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 Proposition 1 are established for integer values of $\gamma > 0$, relation (11) holds true for any positive real number γ. In order to get an idea of how the terms z_γ and $\mathsf{l}_\gamma(z_\gamma)$ depend on γ, we depicted in Figure 2 the plots of these quantities as functions of $\gamma > 0$.

Combining relation (11) with Theorem 2, we get the following result.

COROLLARY 2. *Let condition* **[C1**(κ, 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*

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

with $\gamma = L/\kappa$, then consistent estimation of J is possible and can be achieved, for instance, by the estimator $\widehat{J}_{2,n}$.

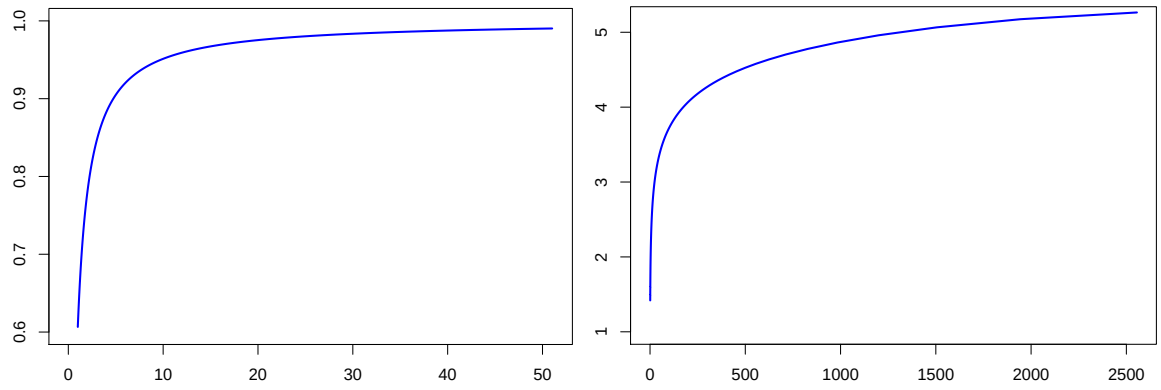

FIG 2. *The plots of mappings* $\gamma \mapsto z_\gamma$ *and* $\gamma \mapsto \mathsf{l}_\gamma(z_\gamma)$. *One can observe that both functions are increasing, the first one converges to* 1 *very rapidly, while the second one seems to diverge very slowly.*

5. Tightness of the assumptions. In this section, we focus our attention on the functional class $\Sigma(\kappa, L)$ of all functions satisfying assumption **[C1**(κ, L)**]**. In order to avoid irrelevant technicalities and to better convey the main results, we assume that $\kappa = 1$ and denote $\Sigma_L = \Sigma(1, L)$. For emphasize that J is the sparsity pattern of the function f that has been used in the data generation process, we write J_f instead of J. The goal in this section is to provide conditions under which the consistent estimation of the sparsity support is impossible, that is there exists a positive constant $c > 0$ and an integer $n_0 \in \mathbb{N}$ such that, if $n \geq n_0$,

$$\inf_{\widetilde{J}} \sup_{\mathsf{f} \in \Sigma_L} \mathbf{P}_\mathsf{f}(\widetilde{J} \neq J_\mathsf{f}) \geq c,$$

where the inf is over all possible estimators of J_f. To lower bound the left hand side of the last inequality, we introduce a set of $M+1$ functions $\mathsf{f}_0, \ldots, \mathsf{f}_M$ such that $\mathsf{f}_0 \equiv 0$ and $\mathsf{f}_\ell \in \Sigma_L$ for every $\ell = 1, \ldots, M$ and use the fact that

$$\inf_{\widetilde{J}} \sup_{\mathsf{f} \in \Sigma_L} \mathbf{P}_\mathsf{f}(\widetilde{J} \neq J_\mathsf{f}) \geq \inf_{\widetilde{J}} \frac{1}{M} \sum_{\ell=1}^{M} \mathbf{P}_{\mathsf{f}_\ell}(\widetilde{J} \neq J_{\mathsf{f}_\ell}). \tag{12}$$

These functions f_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 $J_{\mathsf{f}_\ell} = J_\ell$ and all the sets $J_1, \ldots, J_M$ are distinct. We can write the inequality (12) as

$$\inf_{\widetilde{J}} \sup_{\mathsf{f} \in \Sigma_L} \mathbf{P}_\mathsf{f}(\widetilde{J} \neq J_\mathsf{f}) \geq \inf_{\psi} \frac{1}{M} \sum_{\ell=1}^{M} \mathbf{P}_{\mathsf{f}_\ell}(\psi \neq \ell), \tag{13}$$

where the inf is taken over all random variables ψ taking values in $\{1, \ldots, M\}$. The latter inf will be controlled using a suitable version of the Fano lemma, see Fano (1961). In what follows, we denote by $\mathscr{K}(P, Q)$ the Kullback-Leibler divergence between two probability measures P and Q defined on the same probability space.

LEMMA 3 (Corollary 2.6 of Tsybakov (2009)). *Let $M \geq 3$ be an integer, $(\mathscr{X}, \mathscr{A})$ be a measurable space and let $P_0, \ldots, P_M$ be probability measures on $(\mathscr{X}, \mathscr{A})$. 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 $\psi : \mathscr{X} \to \{1, \ldots, M\}$. If for some $0 < \alpha < 1$*

$$\frac{1}{M+1} \sum_{\ell=1}^{M} \mathscr{K}(P_\ell, P_0) \leq \alpha \log M,$$

then

$$\bar{p}_{e,M} \geq \frac{(M+1)(\log(M+1)-\log 2)}{M\log M} - \frac{1}{M} - \alpha \geq \frac{1}{2} - \alpha.$$

We apply this lemma with the choice $\mathscr{X} = \Sigma_L \cup \{\mathsf{f}_0\}$. It follows from Fano's lemma that one can deduce a lower bound on $\bar{p}_{e,M}$, which is the quantity we are interested in, from an upper bound on the average Kullback-Leibler divergence between the measures $\mathbf{P}_{\mathsf{f}_\ell}$ and $\mathbf{P}_{\mathsf{f}_0}$. This roughly means that the functions f_ℓ should not be very far from f_0 but they should differ one from another (and from f_0) in terms of the sparsity pattern.

5.1. *A simple approach for fixed sparsity.* One can observe that in the case when d^* remains bounded from above when $n \to \infty$, both estimators $\widehat{J}_{1,n}$ and $\widehat{J}_{2,n}$ are consistent as soon as $\log d \leq cn$ for every n, where c is a given constant depending on κ and L. In this subsection, we show that if this condition is not verified, then for every estimator, there is at least one function f—satisfying the assumptions made in the previous section—such that the probability of choosing the true sparsity pattern does not tend to one.

PROPOSITION 4. *If for some $\alpha \in (0,1)$ the inequality $d^*(\log d - \log d^*) \geq \alpha^{-1} n$ holds true for every $n \geq 1$, then there is a constant $c > 0$ such that*

$$\inf_{\widetilde{J}_n} \sup_{f \in \Sigma_L} \mathbf{P}_{\mathsf{f}}(\widetilde{J}_n \neq J_{\mathsf{f}}) \geq \frac{1}{2} - \alpha, \qquad \forall n \geq 1,$$

where the inf *is taken over the collection of all the estimator $\widetilde{J}_n$.*

5.2. *Tightness of the conditions for growing sparsity.* In the case when the intrinsic dimension $d^* = d^*_n \to \infty$, the condition ensuring the validity of Theorem 2 is weaker than the one ensuring the validity of Theorem 1. In particular, it follows from Theorem 2 that if the sequences $(d^* \log d)/n$ and $(N(d^*, L+\tau)d^* \log d)^{1/2}/n$, are uniformly in n bounded by some small constant, then the sparsity pattern estimator $\widehat{J}_{2,n}$ is consistent. We have already established in Proposition 4 that the condition on $(d^* \log d)/n$ is optimal up to constant. The aim is now to show that the condition on $(N(d^*, L+\tau)d^* \log d)^{1/2}/n$ is optimal up to constant as well.

The first step consists in proving that one can restrict the set of all estimators $\widetilde{J}_n$ to the set of estimators that depend only on the absolute values of the observations. In other words, instead of applying Lemma 3 to the measures $\mathbf{P}_{\mathsf{f}}$ one can do the same with the measures $\bar{\mathbf{P}}_{\mathsf{f}}$ defined as the probability distribution of the sequence $\{y^2_{\boldsymbol{k}} : \boldsymbol{k} \in \mathbb{Z}^d \ \& \ \|\boldsymbol{k}\|_\infty \leq n\}$, where $y_{\boldsymbol{k}}$ is as defined in (5). The advantage of considering $\bar{\mathbf{P}}_{\mathsf{f}}$ instead of $\mathbf{P}_{\mathsf{f}}$ is that, in general, the Kullback-Leibler divergence between $\bar{\mathbf{P}}_{\mathsf{f}}$ and $\bar{\mathbf{P}}_{\mathsf{f}'}$ is smaller than the Kullback-Leibler divergence between $\mathbf{P}_{\mathsf{f}}$ and $\mathbf{P}_{\mathsf{f}'}$. To state the next result, we need the additional notation $\Sigma_{L,n} = \{\mathsf{f} \in \Sigma_L : \sup_{\|\boldsymbol{k}\|_\infty > n} |\langle \mathsf{f}, \varphi_{\boldsymbol{k}} \rangle| = 0\}$.

LEMMA 5. *Assume that for some $M \geq 3$ and some $\alpha \in (0,1)$ there exist functions $\mathsf{f}_1, \dots, \mathsf{f}_M$ in $\Sigma_{L,n}$ such that*

$$\frac{1}{M} \sum_{\ell=1}^{M} \mathscr{K}(\bar{\mathbf{P}}_{\mathsf{f}_\ell}, \bar{\mathbf{P}}_{\mathsf{f}_0}) \leq \alpha \log M$$

and the sparsity patterns of functions f_ℓ, for $\ell = 1, \dots, M$, are all distinct. Then, the following inequality holds true:

$$\inf_{\widetilde{J}_n} \sup_{\mathsf{f} \in \Sigma_{L,n}} \mathbf{P}_{\mathsf{f}}\Big(\widetilde{J}_n \neq J_{\mathsf{f}}\Big) \geq \frac{1}{2} - \alpha.$$

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 Theorem 2 are violated.

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

$$\frac{N(d^*,\gamma_L)d^*\log(d/d^*)}{n^2}\geq L\alpha^{-1}, \tag{14}$$

then, for d^* is large enough,

$$\inf_{\widetilde{J}}\sup_{\mathsf{f}\in\Sigma}\mathbf{P}_{\mathsf{f}}\Big(\widetilde{J}\neq J_{\mathsf{f}}\Big)\geq\frac{1}{2}-\alpha.$$

It is worth stressing here that condition (14) is the converse of condition (10) of Theorem 2 in the case $d^*\to\infty$, in the sense that condition (10) amounts to requiring that the left hand side of (14) is smaller than some constant. There is however one difference between the quantities involved in these conditions: the term $N(d^*,L+\tau)$ of (10) is replaced by $N(d^*,\gamma_L)$ in condition (14). A natural question arises: how close γ_L is to L ? To give a qualitative answer to this question, we plotted in Figure 3 the curve of the mapping $L\mapsto\gamma_L$ along with the bissectrice $L\mapsto L$. We observe that the difference between two curves is small compared to L. 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 L.

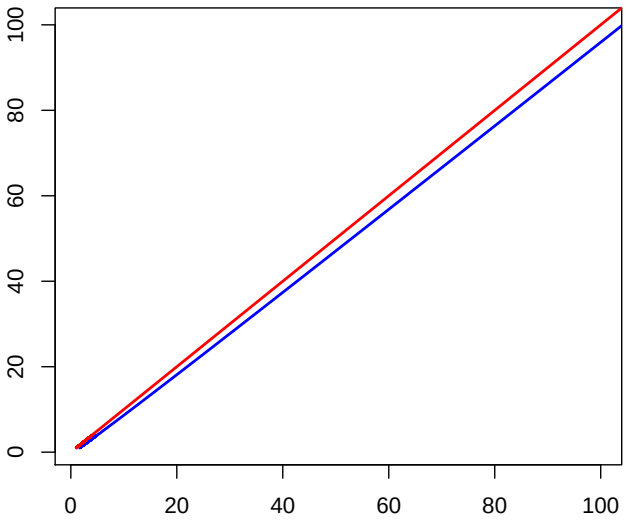

FIG 3. *The curve of the function $L\mapsto\gamma_L$ (blue) and the bissectrice (red).*

6. Nonparametric regression with random design. So far, we have analyzed the situation in which noisy observations of the regression function $\mathsf{f}(\cdot)$ are available at all points $\boldsymbol{x}\in[0,1]^d$. Let us turn now to the more realistic model of nonparametric regression, corresponding to the case where 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 $(\boldsymbol{X}_i, Y_i)$, $i=1,\dots,n$ are observed that obey the regression model

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

The input variables $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ are assumed to take values in $\mathbb{R}^d$ while the output variables $Y_1,\dots,Y_n$ are scalar. As usual, the noise $\varepsilon_1,\dots,\varepsilon_n$ is such that $\mathbf{E}[\varepsilon_i|\boldsymbol{X}_i]=0$, $i=1,\dots,n$; some additional conditions will be imposed later. Without requiring from f to be of a special parametric form, we aim at recovering the set $J\subset\{1,\dots,d\}$ 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 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] We assume that $\mathsf{g}(\boldsymbol{x})=0$ for any $\boldsymbol{x}\notin[0,1]^d$ and that $\mathsf{g}(\boldsymbol{x})\geq g_{\min}$ 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 $L_\infty>0$ and L_2, *i.e.*,

$$P_X\big(\boldsymbol{x}\in[0,1]^d : |\mathsf{f}(\boldsymbol{x})|\leq L_\infty\big)=1 \quad \text{and} \quad \int_{[0,1]^d}\mathsf{f}(\boldsymbol{x})^2\mathsf{g}(\boldsymbol{x})\,d\boldsymbol{x}\leq L_2^2.$$

[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 M appearing in conditions **[C1-C3]**.

6.1. *An estimator of J and its consistency.* The estimator of the sparsity pattern J that we are going to introduce now is the analogue of $\widehat{J}_{1,n}$ studied in Section 3. Recall that it is based on the following simple observation: 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 $S_{m,\ell}=\big\{\boldsymbol{k}\in\mathbb{Z}^d : \|\boldsymbol{k}\|_2\leq m,\ \|\boldsymbol{k}\|_0\leq\ell\big\}$ and $N(d^*,\gamma)=\mathrm{Card}\{\boldsymbol{k}\in\mathbb{Z}^{d^*}:\|\boldsymbol{k}\|_2^2\leq\gamma d^*\,\&\,k_1\neq 0\}$. Finally our estimator is defined by

$$\widehat{J}_n(m,\lambda)=\Big\{j\in\{1,\dots,d\}:\max_{\boldsymbol{k}\in S_{m,d^*}:k_j\neq 0}|\widehat{\theta}_{\boldsymbol{k}}|>\lambda\Big\}, \tag{15}$$

where m and λ are some parameters to be defined later. The notation $a\wedge b$, for two real numbers a and b, stands for $\min(a,b)$.

THEOREM 4. *Let conditions* **[C1-C4]** *be fulfilled with some known constants* $g_{\min}, L, \kappa$ *and* L_2. *Assume furthermore that the design density* g *and an upper estimate on the noise magnitude* σ *are available. Set* $m = (2Ld^*/\kappa)^{1/2}$ *and* $\lambda = 4(\sigma + L_2)\big(d^* \log(6md)/n g_{\min}^2\big)^{1/2}$. *If*

$$\frac{L_\infty^2 d^* \log(6md)}{n} \le L_2^2, \quad \textit{and} \quad \frac{128(\sigma + L_2)^2 d^* N(d^*, 2L/\kappa)\log(6md)}{n g_{\min}^2} \le \kappa, \tag{16}$$

then the estimator $\widehat{J}(m,\lambda)$ *satisfies* $\mathbf{P}\big(\widehat{J}(m,\lambda) \neq J\big) \le 3(6md)^{-d^*}$.

If we take a look at the conditions of Theorem 4 ensuring the consistency of the estimator $\widehat{J}$, it becomes clear that the strongest requirement is the second inequality in (16). To some extent, this condition requires that $(d^* N(d^*, 2L/\kappa)\log d)/n$ is bounded from above by some constant. To further analyze the interplay between d^*, d and n implied by this condition, we need an equivalent to $N(d^*, 2L/\kappa)$ as the intrinsic dimension d^* tends to infinity. As proved in the next section, $N(d^*, 2L/\kappa)$ diverges exponentially fast, making inequality (16) 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 Theorem 4. Only constants related to some quadratic functionals of the sequence of Fourier coefficients $\theta_{\boldsymbol{k}}[\mathrm{f}]$ 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 (Laurent and Massart, 2000, Cai and Low, 2006), than the estimation of sup-norm.

The result stated above can be reformulated to provide also a level of relevance κ for the variables of $\boldsymbol{X}$ making their identification possible. In fact, an alternative way of stating Theorem 4 is the following: if conditions **[C1-C4]** and $L_\infty^2 d^* \log(6md) \le n L_2^2$ are fulfilled, then the estimator $\widehat{J}(m,\lambda)$—with arbitrary tuning parameters m and λ—satisfies $\mathbf{P}(\widehat{J}(m,\lambda) \neq J) \le 3(6md)^{-d^*}$ provided that 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 presented in Appendix C.

6.2. *Tightness of the assumptions.* A natural question is now to check that the assumptions of Theorem 4 are tight in the asymptotic regimes of fixed sparsity and increasing ambient dimension, as well as increasing sparsity. In the present subsection, we will only establish an analogue of Proposition 4, which entails the tightness of assumptions in the case of fixed (or bounded) sparsity. An attempt to prove a result similar to Theorem 3 was done in (Comminges and Dalalyan, 2011, Theorem 2). However, the result of (Comminges and Dalalyan, 2011) involves a stringent assumption on the empirical Gram matrix (cf. condition (6)) 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. Gaussian with zero mean and variance 1 and we focus our attention on the functional class Σ_L of all functions satisfying assumption **[C1**$(1, L)$**]**. The goal is now to provide conditions under which the consistent estimation of the sparsity support is impossible, that is there exists a positive constant $c > 0$ and an integer $n_0 \in \mathbb{N}$ such that, if $n \ge n_0$,

$$\inf_{\widetilde{J}} \sup_{\mathrm{f} \in \Sigma_L} \mathbf{P}_{\mathrm{f}}(\widetilde{J} \neq J_{\mathrm{f}}) \ge c,$$

where the inf is over all possible estimators of J_{f}.

The following simple result shows that the conditions of Theorem 4 are tight in the case of fixed intrinsic dimension.

PROPOSITION 6. *Let the design $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n \in [0,1]^d$ be either deterministic or random. If for some positive $\alpha < (\log 3 - \log 2)/\log 3$, the inequality*

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

holds true, then there is a constant $c>0$ such that $\inf_{\widetilde{J}_n} \sup_{\mathsf{f}\in\Sigma_L} \mathbf{P}_{\mathsf{f}}(\widetilde{J}_n \neq J_{\mathsf{f}}) \geq c$.

7. 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 Theorem 2 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 4, consistent estimation of the sparsity pattern can be achieved both in Gaussian white noise model and nonparametric regression as soon as $(\log d)/n \leq c_\star$, where $c_\star$ is the constant defined by $c_* = \big(2^5 d^* N(d^*, 2L)\big)^{-1}$ for the Gaussian white noise model (we choose $A=2$) and

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

for the model of nonparametric regression. This follows from the fact that the tuning parameter m is fixed and that the probability of the error, bounded by $3(6md)^{d^*}$ tends to zero as $d\to\infty$. On the other hand, by virtue of Propositions 4 and 6, consistent estimation of the sparsity pattern is impossible if $(\log d)/n > c^\star$, where $c^\star = 2\log 3/(d^*\log(3/2))$. Thus, up to multiplicative constants $c_\star$ and $c^\star$ (which are clearly not sharp), the results of Theorems 1 and 4 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 (10) and using formula (11) for $N(d^*, L+\tau) = N_1(d^*, L+\tau) - N_2(d^*, L+\tau)$, we see that consistent estimation of J is possible when, for some $\tau>0$ and for all n, the following two conditions are fulfilled:

$$\begin{cases} \mathsf{l}_{L+\tau}(z_{L+\tau})d^* + \frac{1}{2}\log d^* + \log\log d - 2\log n < \underline{c}_1, \\ \log d^* + \log\log d - \log n \leq \underline{c}_1' \end{cases} \tag{17}$$

with $\underline{c}_1 = \log\tau - \log(2^9(L+\tau)) - c_\gamma$ and $\underline{c}_1' = \log\tau - \log(40(L+\tau))$. On the other hand, Proposition 4 and Theorem 3 yield that there are some constants $\bar{c}_1$ and $\bar{c}_1'$ such that it is impossible to consistently estimate J if either

$$\mathsf{l}_{\gamma_L}(z_{\gamma_L})d^* + \frac{1}{2}\log d^* + \log\log d - 2\log n \geq \bar{c}_1, \tag{18}$$

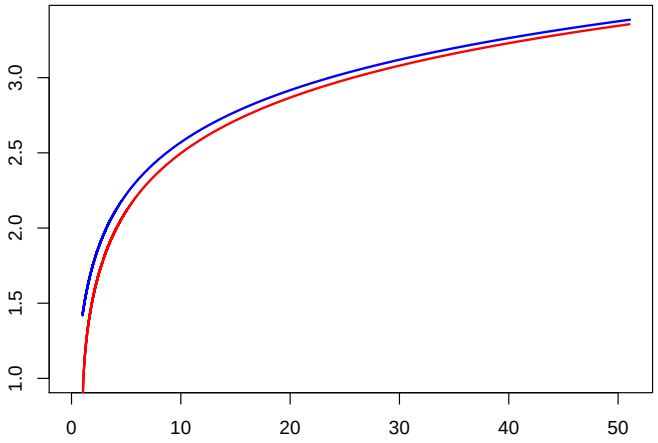

FIG 4. *The curves of functions* $L \mapsto \mathsf{I}_L(z_L)$ *(blue curve) and* $L \mapsto \mathsf{I}_{\gamma_L}(z_{\gamma_L})$ *(red curve).*

or

$$\log d^* + \log\log d - \log n \geq \bar{c}_1'. \tag{19}$$

First note that the left hand side of the second condition in (17) is exactly the same as the left hand side of (19). If we compare now the left hand side of the first condition in (17) with the left hand side of (18), 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 $L \mapsto \mathsf{I}_L(z_L)$ and $L \mapsto \mathsf{I}_{\gamma_L}(z_{\gamma_L})$, with γ_L being defined as is Theorem 3. One can observe that the two curves are very close especially for relatively large values of L. This implies that the conditions (17) are tight. A very simple consequence of inequalities (17) and (18) 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 $\log\log d = o(\log n)$.

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

$$\begin{cases} \mathsf{I}_{L+\tau}(z_{L+\tau})d^* + \frac{1}{2}\log d^* + \log\log d - \log n < \underline{c}_2, \\ \log d^* + \log\log d - \log n < \underline{c}_2' \end{cases} \tag{20}$$

with $\underline{c}_2 = 2\big(\log(g_{\min}) - \log(12(\sigma + L_2))\big) - c_{2L}$ and $\underline{c}_2' = 2\log(L_2/L_\infty)$, where c_{2L} is the constant c_γ of (11) evaluated at $\gamma = 2L$. As we have already mentioned, the second condition in (20) is tight, up to the choice of $\underline{c}_2'$, in view of Proposition 6. 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 $\widehat{\theta}_{\boldsymbol{k}}$ in a stepwise manner; at each step $K=1,2,\ldots,d^*$ only the coefficients $\widehat{\theta}_{\boldsymbol{k}}$ with $\|\boldsymbol{k}\|_0=K$ need to be computed and compared with the threshold. If some $\widehat{\theta}_{\boldsymbol{k}}$ exceeds the threshold, then all the variables X^j corresponding to nonzero coordinates of $\boldsymbol{k}$ 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 $\mathsf{f}(\boldsymbol{x})=\mathsf{f}_1(x_{i_1})+\ldots+\mathsf{f}_{d^*}(x_{i_{d^*}})$ for some univariate functions $\mathsf{f}_1,\ldots,\mathsf{f}_{d^*}$.

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 Theorem 4, 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 (17) 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 $n^{-2/(2+d^*)}=\exp(-2\log n/(2+d^*))$. If the left hand side of (17) 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. This heuristical argument shows that there is still some work to do for getting tight conditions ensuring the consistent estimation of the regression function in the high dimensional set-up.

Finally, we would like to mention that the selection of relevant variables is a challenging statistical task, which is useful to perform independently of the task of regression function estimation. Indeed, in 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 data storage and financial costs necessary for collecting new data. Then, the regression function may be estimated in a more accurate fashion on the base of this new (larger) data-set.

APPENDIX A: PROOF OF THEOREM 1

To ease notation, we write $\widehat{J}_1$ instead of $\widehat{J}_{1,n}(m,\lambda)$. On the one hand, if the event

$$\mathscr{A}_\lambda=\Big\{\max\nolimits_{j,\boldsymbol{k}\in S^j_{m,d^*}}|\xi_{\boldsymbol{k}}|\le n^{1/2}\lambda\Big\}$$

is realized, then $J^c\subset\widehat{J}_1^c$ (or, equivalently, $\widehat{J}_1\subset J$). Thus, the probability of $\widehat{J}_1\subset J$ is at least as large as $\mathbf{P}(\mathscr{A}_\lambda)$. On the other hand, we show below that under condition (8) the event $J\subset\widehat{J}_1$ is at least as $\mathscr{A}_\lambda$ or, equivalently, that the event $J\not\subset\widehat{J}_1$ is at most as large as $\mathscr{A}_\lambda^c$. It is clear that

$$\begin{aligned}\{J\not\subset\widehat{J}_1\}&=\Big\{\exists j\in J\text{ s.t. }\max\nolimits_{\boldsymbol{k}\in S^j_{m,d^*}}|y_{\boldsymbol{k}}|\le\lambda\Big\}\\&\subset\Big\{\exists j\in J\text{ s.t. }|\theta_{\boldsymbol{k}}[\mathsf{f}]|\le\lambda+n^{-1/2}|\xi_{\boldsymbol{k}}|,\ \forall\boldsymbol{k}\in S^j_{m,d^*}\Big\}\\&\subset\Big(\Big\{\exists j\in J\text{ s.t. }|\theta_{\boldsymbol{k}}[\mathsf{f}]|\le\lambda+n^{-1/2}|\xi_{\boldsymbol{k}}|,\ \forall\boldsymbol{k}\in S^j_{m,d^*}\Big\}\cap\mathscr{A}_\lambda\Big)\cup\mathscr{A}_\lambda^c\\&\subset\Big\{\exists j\in J\text{ s.t. }\max\nolimits_{\boldsymbol{k}\in S^j_{m,d^*}}|\theta_{\boldsymbol{k}}[\mathsf{f}]|\le 2\lambda\Big\}\cup\mathscr{A}_\lambda^c.\end{aligned}$$

We show now that the first set in the last line is empty. If this was not the case, then for some value j_0 we would have $Q_{j_0} \geq \kappa$ and $|\theta_{\boldsymbol{k}}[\mathrm{f}]| \leq 2\lambda$, for all $\boldsymbol{k} \in S^{j_0}_{m,d^*}$. This would imply that

$$Q_{j_0,m,d^*} \triangleq \sum_{\boldsymbol{k} \in S^{j_0}_{m,d^*}} \theta_{\boldsymbol{k}}[\mathrm{f}]^2 \leq 4\lambda^2 N(d^*, m^2/d^*).$$

On the other hand,

$$Q_{j_0} - Q_{j_0,m,d^*} \leq \sum_{\|\boldsymbol{k}\|_2 \geq m} \theta_{\boldsymbol{k}}[\mathrm{f}]^2 \leq m^{-2} \sum_{\|\boldsymbol{k}\|_2 \geq m} \sum_{j \in J} k_j^2 \theta_{\boldsymbol{k}}[\mathrm{f}]^2 \leq m^{-2} L d^*.$$

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

$$Q_{j_0} \leq 0.5\kappa + 4\lambda^2 N(d^*, m^2/d^*),$$

which is impossible since $Q_{j_0} \geq \kappa$. Thus, we proved that $\{J \not\subset \widehat{J}_1\} \subset \mathscr{A}_\lambda^c$, which implies that $\mathscr{A}_\lambda \subset \{J \subset \widehat{J}_1\}$. Therefore,

$$\mathscr{A}_\lambda \subset \{J \subset \widehat{J}_1\} \cap \{\widehat{J}_1 \subset J\} = \{\widehat{J}_1 = J\}.$$

The cardinality of this set $S_{m,d^*} = \cup_j S^j_{m,d^*}$ admits the simple upper bound:

$$\mathrm{Card}(S_{m,d^*}) = \sum_{s=0}^{d^*} \binom{d}{s} (2m)^s \leq (2m)^{d^*} \sum_{s=0}^{d^*} \frac{(d)^s}{s!} \leq 3(2md)^{d^*}.$$

This bound is very rough but it will be sufficient for the purposes of the present work. Using the Bonferroni inequality and the well-known inequalities for the tails of standard Gaussian distribution, we get

$$\mathbf{P}(\widehat{J}_1 \neq J) \leq \mathbf{P}(\mathscr{A}_\lambda^c) \leq \sum_{\boldsymbol{k} \in S_{m,d^*}} \mathbf{P}(|\xi_{\boldsymbol{k}}| > n^{1/2}\lambda) \leq 6(2md)^{d^*} e^{-n\lambda^2/2} \leq 6d^{-d^*(A-1)}$$

and the theorem follows.

APPENDIX B: PROOF OF THEOREM 2

Let us recall that for every $j \in \{1, \ldots, d\}$, we use the notation

$$Q^j_{m,I} = \sum_{\boldsymbol{k} \in S^j_{m,I}} \theta^2_{\boldsymbol{k}}, \qquad \widehat{Q}^j_{m,I} = \sum_{\boldsymbol{k} \in S^j_{m,I}} \Big(y^2_{\boldsymbol{k}} - \frac{1}{n}\Big).$$

We will also need the notation

$$R^j_{m,I} \triangleq \sum_{\boldsymbol{k} \in S^j_{m,I}} (\xi^2_{\boldsymbol{k}} - 1), \qquad N^j_{m,I} \triangleq \sum_{\boldsymbol{k} \in S^j_{m,I}} \theta_{\boldsymbol{k}} \xi_{\boldsymbol{k}}.$$

It is clear that $N^j_{m,I}$ is drawn from the centered Gaussian distribution with variance $Q^j_{m,I}$. Furthermore, we have

$$\widehat{Q}^j_{m,I} = Q^j_{m,I} + \frac{2}{\sqrt{n}} N^j_{m,I} + \frac{1}{n} R^j_{m,I}. \tag{21}$$

This implies, in particular, that for every $j \notin J$, $\widehat{Q}_j^{I,m} = n^{-1} R^j_{m,I}$ for every I. Therefore, on the event

$$\mathscr{A} = \bigcap_{\ell=1}^{d^*} \left[\left\{ \max_{I \in P_\ell^d} \max_{i \in \{1,\dots,d\}} |R^i_{m_\ell,I}| \le n\lambda_\ell \right\} \bigcap \left\{ \max_{I \in P_\ell^d} \max_{i \in \{1,\dots,d\}} (N^i_{m_\ell,I})^2 / Q^i_{m_\ell,I} \le 2Ad^* \log d \right\} \right],$$

the inclusion $J^c \subset \widehat{J}_2^c$ is necessarily true (we use the convention $0/0 = 0$). This entails that $\mathscr{A} \cap \{J^c \not\subset \widehat{J}_2^c\} = \varnothing$ and hence

$$\begin{aligned} \mathbf{P}(\widehat{J}_2 \neq J) &= \mathbf{P}(\{J^c \not\subset \widehat{J}_2^c\} \cup \{J \not\subset \widehat{J}_2\}) \\ &\le \mathbf{P}(\mathscr{A}^c) + \mathbf{P}(\mathscr{A} \cap \{J \not\subset \widehat{J}_2\}). \end{aligned}$$

The term $\mathbf{P}(\mathscr{A}^c)$ is bounded using standard concentration inequalities for Gaussian and chi-square random variables. In particular, we will use the following result.

LEMMA 7 (cf. Laurent and Massart (2000)). *Let $(\xi_1, \dots, \xi_D)$ be independent Gaussian random variables with mean zero and variance 1. For every vector $\boldsymbol{a} = (a_1, \dots, a_D) \in \mathbb{R}_+^D$ and for every $x \ge 0$, the following inequality holds true:*

$$\mathbf{P}\left(\left| \sum_{i=1}^D a_i (\xi_i^2 - 1) \right| \ge 2\|\boldsymbol{a}\|_2 \sqrt{x} + 2\|\boldsymbol{a}\|_\infty x \right) \le 2\exp(-x).$$

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

$$\begin{aligned} \mathbf{P}\left(\bigcup_{\ell=1}^{d^*} \left\{ \max_{I \in P_\ell^d} \max_{i \in \{1,\dots,d\}} |R^i_{m_\ell,I}| > n\lambda_\ell \right\} \right) &\le \sum_{\ell=1}^{d^*} \mathbf{P}\left(\max_{I \in P_\ell^d; i \in I} |R^i_{m_\ell,I}| > n\lambda_\ell \right) \\ &\le \sum_{\ell=1}^{d^*} \ell \operatorname{Card}(P_\ell^d) \max_{I \in P_\ell^d; i \in I} \mathbf{P}\left(|R^i_{m_\ell,I}| > n\lambda_\ell \right) \\ &\le 2e^{-x} \sum_{\ell=1}^{d^*} \frac{d \cdot (d-1) \cdot \ldots \cdot (d-\ell+1)}{(\ell-1)!} \le 3e^{-x} d^{d^*}. \end{aligned}$$

For $x = Ad^* \log d$, we get

$$\mathbf{P}\left(\bigcup_{\ell=1}^{d^*} \left\{ \max_{I \in P_\ell^d} \max_{i \in \{1,\dots,d\}} |R^i_{m_\ell,I}| > n\lambda_\ell \right\} \right) \le 3d^{-(A-1)d^*}.$$

Consider now the noise terms $N^i_{m_\ell,I}$. Using the classical bounds on the tails of standard Gaussian random variables, for any $x > 0$, we get

$$\begin{aligned} \mathbf{P}\left(\bigcup_{\ell=1}^{d^*} \left\{ \max_{I \in P_\ell^d} \max_{i \in \{1,\dots,d\}} \frac{(N^i_{m_\ell,I})^2}{Q^i_{m_\ell,I}} > x \right\} \right) &\le \sum_{\ell=1}^{d^*} \sum_{I \in P_\ell^d} \sum_{i \in I} \mathbf{P}\left(\frac{(N^i_{m_\ell,I})^2}{Q^i_{m_\ell,I}} > x \right) \\ &\le 2e^{-x/2} \sum_{\ell=1}^{d^*} \ell \binom{d}{\ell} \le 3e^{-x/2} d^{d^*}. \end{aligned}$$

Setting $x = 2Ad^* \log d$, we bound the last probability by $3d^{-(A-1)d^*}$ and, therefore,

$$\mathbf{P}(\mathscr{A}^c) \le 6d^{-(A-1)d^*}. \tag{22}$$

Let us show now that under the conditions of the theorem, $\mathscr{A} \cap \{J \not\subset \widehat{J}_2\} = \varnothing$. We start by rewriting the event

$$\{J \not\subset \widehat{J}_2\} = \bigcup_{j_0 \in J} \bigcap_{\ell=1}^{d^*} \Big\{ \max_{I \in P_\ell^d} \widehat{Q}^{j_0}_{m_\ell, I} < \lambda_\ell \Big\} \subset \bigcup_{j_0 \in J} \Big\{ \widehat{Q}^{j_0}_{m_{d^*}, J} < \lambda_{d^*} \Big\}.$$

This implies that

$$\mathscr{A} \cap \{J \not\subset \widehat{J}_2\} \subset \bigcup_{j_0 \in J} \Big(\mathscr{A} \cap \Big\{ \widehat{Q}^{j_0}_{m_{d^*}, J} < \lambda_{d^*} \Big\} \Big)$$

In the rest of the proof, m and λ will stand for m_{d^*} and λ_{d^*}. Following the lines of proof[1] of Theorem 1, one easily checks that

$$Q^{j_0}_{m,J} \ge \kappa\tau/(L+\tau) \tag{23}$$

for every $j_0 \in J$. Recall also that

$$\widehat{Q}^{j_0}_{m,J} = Q^{j_0}_{m,J} + \frac{1}{n} R^{j_0}_{m,J} + \frac{2}{\sqrt{n}} N^{j_0}_{m,J}.$$

Thus, on the event $\mathscr{A}$, we have

$$\widehat{Q}^{j_0}_{m,J} - \lambda \ge Q^{j_0}_{m,J} - 2\left(Q^{j_0}_{m,J}\right)^{1/2} \sqrt{\frac{2Ad^* \log d}{n}} - 2\lambda. \tag{24}$$

Elementary computations show that the function $x \mapsto x^2 - 2ax - b$ is positive when $x \ge 4a^2 + 2b$. Applying this property to the right hand side of (24), we get that

$$\widehat{Q}^{j_0}_{m,J} - \lambda \ge 0 \tag{25}$$

provided that

$$Q^{j_0}_{m,J} \ge 4\lambda + \frac{8Ad^* \log d}{n} = \frac{8\sqrt{AN(d^*, m^2_{d^*}/d^*) d^* \log d} + 10Ad^* \log d}{n},$$

which is satisfied by virtue of (23) and (10). This implies that under the conditions of the theorem, on the event $\mathscr{A}$, we have $\hat{Q}^{j_0}_{m,J} \ge \lambda$, $\forall j_0 \in J$. Consequently, the event $\mathscr{A} \cap \{J \not\subset \widehat{J}_2\}$ is impossible. This completes the proof of the theorem.

APPENDIX C: PROOF OF THEOREM 4

The empirical Fourier coefficients can be decomposed as follows:

$$\widehat{\theta}_{\boldsymbol{k}} = \tilde{\theta}_{\boldsymbol{k}} + z_{\boldsymbol{k}}, \qquad \text{where} \qquad \tilde{\theta}_{\boldsymbol{k}} = \frac{1}{n} \sum_{i=1}^n \frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\mathrm{g}(\boldsymbol{X}_i)} \mathrm{f}(\boldsymbol{X}_i) \qquad \text{and} \qquad z_{\boldsymbol{k}} = \frac{\sigma}{n} \sum_{i=1}^n \frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\mathrm{g}(\boldsymbol{X}_i)} \boldsymbol{\varepsilon}_i. \tag{26}$$

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,

$$\mathbf{P}\big(|z_{\boldsymbol{k}}| > x \,\big|\, \boldsymbol{X}_1, \dots, \boldsymbol{X}_n\big) \le \exp\Big(-\frac{x^2}{2\sigma^2_{\boldsymbol{k}}}\Big) \quad \forall \boldsymbol{k} \in S_{m,d^*}$$

[1] In Appendix A, this property is established for $\tau = L$.

with

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

Therefore, for every $\boldsymbol{k} \in S_{m,d^*}$, it holds that $\mathbf{P}\big(|z_{\boldsymbol{k}}| > x|\boldsymbol{X}_1,\dots,\boldsymbol{X}_n\big) \le \exp(-n g_{\min}^2 x^2/4\sigma^2)$. This entails that by setting $\lambda_1 = (8\sigma^2 d^* \log(6md)/n g_{\min}^2)^{1/2}$ and by using the inequalities

$$\begin{aligned}\mathrm{Card}(S_{m,d^*}) &= \sum_{i=0}^{d^*}\binom{d}{i}(2m)^i \le (2m)^{d^*}\sum_{i=0}^{d^*}\frac{d^i}{i!}\\ &\le 3(2md)^{d^*} \le (6md)^{d^*},\end{aligned}$$

we get

$$\begin{aligned}\mathbf{P}\Big(\max_{\boldsymbol{k}\in S_{m,d^*}}|z_{\boldsymbol{k}}| > \lambda_1\,|\boldsymbol{X}_1,\dots,\boldsymbol{X}_n\Big) &\le \sum_{\boldsymbol{k}\in S_{m,d^*}}\mathbf{P}\Big(|z_{\boldsymbol{k}}| > \lambda_1\,|\boldsymbol{X}_1,\dots,\boldsymbol{X}_n\Big)\\ &\le \mathrm{Card}(S_{m,d^*})e^{-n g_{\min}^2\lambda_1^2/4\sigma^2} \le (6md)^{-d^*}.\end{aligned}$$

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 $\tilde\theta_{\boldsymbol{k}} = \frac{1}{n}\sum_{i=1}^n \frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\mathrm{g}(\boldsymbol{X}_i)}\mathrm{f}(\boldsymbol{X}_i)$, we have $\mathbb{E}(\tilde\theta_{\boldsymbol{k}}) = \theta_{\boldsymbol{k}}$. By virtue of the boundedness of f, it holds that $|\frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\mathrm{g}(\boldsymbol{X}_i)}\mathrm{f}(\boldsymbol{X}_i)| \le \sqrt{2}L_\infty/g_{\min}$. Furthermore, the bound $V \triangleq \mathrm{Var}\big(\frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\mathrm{g}(\boldsymbol{X}_i)}\mathrm{f}(\boldsymbol{X}_i)\big) \le \int f^2(\boldsymbol{x})\frac{\varphi_{\boldsymbol{k}}^2(\boldsymbol{x})}{\mathrm{g}(\boldsymbol{x})}d\boldsymbol{x} \le 2L_2^2/g_{\min}^2$ combined with Bernstein's inequality yields

$$\begin{aligned}\mathbf{P}\big(|\tilde\theta_{\boldsymbol{k}} - \theta_{\boldsymbol{k}}| > t\big) &\le 2\exp\Big(-\frac{nt^2}{2(V + t\sqrt{2}L_\infty/3g_{\min})}\Big)\\ &\le 2\exp\Big(-\frac{g_{\min}^2 nt^2}{4L_2^2 + tL_\infty g_{\min}}\Big), \qquad \forall t > 0.\end{aligned}$$

Let us define $\lambda_2 = 4L_2\Big(\frac{d^*\log(6md)}{ng_{\min}^2}\Big)^{1/2}$. Then,

$$\mathbf{P}\big(|\tilde\theta_{\boldsymbol{k}} - \theta_{\boldsymbol{k}}| > \lambda_2\big) \le 2\exp\Big(-\frac{4L_2^2 d^*\log(6md)}{L_2^2 + L_\infty L_2\big(\frac{d^*\log(6md)}{n}\big)^{1/2}}\Big).$$

The first inequality in condition (16) implies that the denominator in the exponential is not larger than $2L_2^2$. Hence,

$$\mathbf{P}\Big(\max_{\boldsymbol{k}\in S_{m,d^*}}|\tilde\theta_{\boldsymbol{k}} - \theta_{\boldsymbol{k}}| > \lambda_2\Big) \le 2/(6md)^{d^*}.$$

Let $\mathscr{A}_1 = \big\{\max_{\boldsymbol{k}\in S_{m,d^*}}|z_{\boldsymbol{k}}| \le \lambda_1\big\}$ and $\mathscr{A}_2 = \big\{\max_{\boldsymbol{k}\in S_{m,d^*}}|\tilde\theta_{\boldsymbol{k}}| \le \lambda_2\big\}$. One easily checks that

$$\mathbf{P}\big(J^c \not\subset \widehat{J}^c\big) \le \mathbf{P}\big(\mathscr{A}_1^c\big) + \mathbf{P}\big(\mathscr{A}_2^c\big) \le 3/(6md)^{d^*}.$$

As for the converse inclusion, we have

$$\begin{aligned}\mathbf{P}(J \not\subset \widehat{J}) &\le \mathbf{P}\Big(\exists j \in J \text{ s.t. } \max_{\boldsymbol{k}\in S_{m,d^*}:\,k_j\neq 0}|\widehat\theta_{\boldsymbol{k}}| \le \lambda\Big)\\ &\le \mathbf{1}\Big\{\exists j \in J \text{ s.t. } \max_{\boldsymbol{k}\in S_{m,d^*}:\,k_j\neq 0}|\theta_{\boldsymbol{k}}| \le 2\lambda\Big\} + \mathbf{P}\big(\mathscr{A}_1^c\big) + \mathbf{P}\big(\mathscr{A}_2^c\big).\end{aligned}$$

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 $Q_{j_0} \geq \kappa$ and $|\theta_{\boldsymbol{k}}| \leq 2\lambda$, for all $\boldsymbol{k} \in S_{m,d^*}$ such that $k_{j_0} \neq 0$. This would imply that

$$Q_{j_0,m,d^*} \triangleq \sum_{\boldsymbol{k} \in S_{m,d^*} : k_{j_0} \neq 0} \theta_{\boldsymbol{k}}^2 \leq 4\lambda^2 N(d^*, m^2/d^*).$$

On the other hand,

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

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

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

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

APPENDIX D: PROOF OF PROPOSITION 1

Proof of the first assertion.. This proof can be found in Mazo and Odlyzko (1990), 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 real number $y > 0$, we define $\phi(y) = e^{-y}\mathsf{h}'(e^{-y})/\mathsf{h}(e^{-y}) = \sum_{k=-\infty}^{k=+\infty} k^2 e^{-yk^2} / \sum_{k=-\infty}^{k=+\infty} e^{-yk^2}$ 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

$$\sum k^4 e^{-yk^2} \sum e^{-yk^2} > \left(\sum k^2 e^{-yk^2} \right)^2, \quad \forall y \in (0, \infty),$$

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

$$\begin{aligned} \mathsf{l}_\gamma''(z_\gamma) &= \frac{\mathsf{h}''\mathsf{h} - (\mathsf{h}')^2}{\mathsf{h}^2}(z_\gamma) + \frac{\gamma}{z_\gamma^2} = z_\gamma^{-2} \left\{ \frac{\sum_k (k^4 - k^2) z_\gamma^{k^2}}{\sum_k z_\gamma^{k^2}} - \left(\frac{\sum_k k^2 z_\gamma^{k^2}}{\sum_k z_\gamma^{k^2}} \right)^2 + \gamma \right\} \\ &= z_\gamma^{-2} \{ -\phi'(y_\gamma) - \phi(y_\gamma) + \gamma \} = -z_\gamma^{-2} \phi'(y_\gamma) > 0. \end{aligned}$$

Proof of the second assertion.. We apply the saddle-point method to the integral representing N_1 see, *e.g.*, Chapter IX in (Dieudonné, 1968). It holds that

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

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 Dieudonné (1968) (with $\alpha=0$ since $\mathsf{l}''(z_\gamma)$ is real and positive), we get that

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

when $d^*\to\infty$, as soon as the condition[2] $\Re[\mathsf{l}_\gamma(z)-\mathsf{l}_\gamma(z_\gamma)]\le-\mu$ 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 $\Re[\mathsf{l}_\gamma(z)]=\log\left|\frac{\mathsf{h}(z)}{z^\gamma}\right|$. Hence, if $z=z_\gamma e^{\mathrm{i}\omega}$ with $\omega\in[\omega_0,2\pi-\omega_0]$ for some $\omega_0\in]0,\pi[$, then

$$\begin{aligned}\left|\frac{\mathsf{h}(z)}{z^\gamma}\right|&=\frac{|1+2z+2\sum_{k>1}z^{k^2}|}{z_\gamma^\gamma}\\&\le\frac{|1+z|+z_\gamma+2\sum_{k>1}z_\gamma^{k^2}}{z_\gamma^\gamma}\\&\le\frac{|1+e^{\mathrm{i}\omega_0}z_\gamma|+z_\gamma+2\sum_{k>1}z_\gamma^{k^2}}{z_\gamma^\gamma}.\end{aligned}$$

Therefore $\Re[\mathsf{l}_\gamma(z)-\Re\mathsf{l}_\gamma(z_\gamma)]\le-\mu$ with $\mu=\log\left(\frac{1+2z_\gamma+\sum_{k\ge1}z_\gamma^{k^2}}{|1+z_\gamma e^{\mathrm{i}\omega_0}|+z_\gamma+\sum_{k\ge1}z_\gamma^{k^2}}\right)>0$. 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 E: PROOF OF PROPOSITION 4

Let $M=\binom{d}{d^*}$, $\mathsf{f}_0\equiv0$ and let $\{\mathsf{f}_1,\dots,\mathsf{f}_M\}$ be a set included in Σ_L. It is clear that

$$\inf_{\widetilde{J}_n}\sup_{\mathsf{f}\in\Sigma}\mathbf{P}_\mathsf{f}(\widetilde{J}_n\neq J_\mathsf{f})\ge\inf_{\widetilde{J}_n}\max_{\mathsf{f}\in\{\mathsf{f}_0,\mathsf{f}_1,\dots,\mathsf{f}_M\}}\mathbf{P}_\mathsf{f}(\hat{J}\neq J_\mathsf{f}). \tag{28}$$

Let $I_1,\dots,I_M$ be all the subsets of $\{1,\dots,d\}$ containing exactly d^* elements somehow enumerated. Let us define f_ℓ, for $\ell\neq0$, by its Fourier coefficients $\{\theta^\ell_{\boldsymbol{k}}:\boldsymbol{k}\in\mathbb{Z}^d\}$ as follows:

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

Obviously, all the functions f_ℓ belong to Σ and, moreover, each f_ℓ has I_ℓ as sparsity pattern. In view of Lemma 3, if the condition

$$\frac{1}{M}\sum_{\ell=1}^M\mathscr{K}(\mathbf{P}_{\mathsf{f}_\ell},\mathbf{P}_{\mathsf{f}_0})\le\alpha\log M \tag{29}$$

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

is satisfied for some $\alpha \in (0,1)$, then the RHS of (28) can be lower bounded as follows:

$$\inf_{\widetilde{J}_n} \max_{\mathsf{f}\in\{\mathsf{f}_0,\mathsf{f}_1,\dots,\mathsf{f}_M\}} \mathbf{P}_\mathsf{f}(\widetilde{J} \neq J_\mathsf{f}) \geq \frac{1}{2} - \alpha.$$

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 $\log \binom{d}{d^*}$ is larger than or equal to $d^* \log(d/d^*) = d^*\big(\log d - \log d^*\big)$.

APPENDIX F: PROOF OF LEMMA 5

Let us use boldface letters to denote the set of all Fourier coefficients: $\boldsymbol{y} = \{y_{\boldsymbol{k}} : \|\boldsymbol{k}\|_\infty \leq n\}$, $\boldsymbol{\theta} = \{\theta_{\boldsymbol{k}} : \|\boldsymbol{k}\|_\infty \leq n\}$ and $\boldsymbol{\xi} = \{\xi_{\boldsymbol{k}} : \|\boldsymbol{k}\|_\infty \leq n\}$. Let us denote by $\mathscr{E}$ the set of all sequences $\boldsymbol{\varepsilon} = \{\varepsilon_{\boldsymbol{k}} : \|\boldsymbol{k}\|_\infty \leq n\}$ such that $\varepsilon_{\boldsymbol{k}} \in \{\pm 1\}$. We use the notation $|\boldsymbol{y}| \triangleq \{|y_{\boldsymbol{k}}| : \|\boldsymbol{k}\|_\infty \leq n\}$ and $\boldsymbol{\varepsilon} \cdot |\boldsymbol{y}| \triangleq \{\varepsilon_{\boldsymbol{k}} \cdot y_{\boldsymbol{k}} : \|\boldsymbol{k}\|_\infty \leq n\}$. Let now π be an arbitrary orthosymmetric probability measure on $\Sigma_{L,n}$, *i.e.*, a probability measure satisfying $\pi(A) = \pi(\{\boldsymbol{\varepsilon} \cdot \boldsymbol{\theta} : \boldsymbol{\theta} \in A\})$ for every measurable set A and every $\boldsymbol{\varepsilon} \in \mathscr{E}$. Denoting by $\mathbb{P}(d\boldsymbol{y}, d\mathsf{f}) = \mathbf{P}_\mathsf{f}(d\boldsymbol{y})\pi(d\mathsf{f})$ the joint probability distribution of $(\boldsymbol{y}, \mathsf{f})$, for every sparsity pattern estimator $\widetilde{J}_n$ we have:

$$\begin{aligned}
\sup_{\mathsf{f}\in\Sigma_{L,n}} \mathbf{P}_\mathsf{f}(\widetilde{J}_n \neq J_\mathsf{f}) &\geq \int_{\Sigma_{L,n}} \mathbf{P}_\mathsf{f}(\widetilde{J}_n \neq J_\mathsf{f})\pi(d\mathsf{f}) = \mathbb{P}(\widetilde{J}_n(\boldsymbol{y}) \neq J_\mathsf{f}) \\
&= \frac{1}{2^{|\mathscr{E}|}} \sum_{\boldsymbol{\varepsilon}\in\mathscr{E}} \mathbb{P}\Big(\widetilde{J}_n(\boldsymbol{y}) \neq J_\mathsf{f} \Big| \operatorname{sign}(\boldsymbol{y}) = \boldsymbol{\varepsilon}\Big) \cdot \mathbb{P}\Big(\operatorname{sign}(\boldsymbol{y}) = \boldsymbol{\varepsilon}\Big) \\
&= \frac{1}{2^{|\mathscr{E}|}} \sum_{\boldsymbol{\varepsilon}\in\mathscr{E}} \mathbb{P}\Big(\widetilde{J}_n(\boldsymbol{\varepsilon} \cdot |\boldsymbol{y}|) \neq J_\mathsf{f} \Big| \operatorname{sign}(\boldsymbol{y}) = \boldsymbol{\varepsilon}\Big) \cdot \mathbb{P}\Big(\operatorname{sign}(\boldsymbol{y}) = \boldsymbol{\varepsilon}\Big).
\end{aligned}$$

Using the facts that $\boldsymbol{y} = \boldsymbol{\theta} + n^{-1/2}\boldsymbol{\xi}$, the distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are orthosymmetric and $\boldsymbol{\theta}$ is independent of $\boldsymbol{\xi}$ under $\mathbb{P}$, one easily checks that the random vectors $\operatorname{sign}(\boldsymbol{y})$ and $(|\boldsymbol{y}|, |\boldsymbol{\theta}|)$ are independent. Hence, under $\mathbb{P}$, the event $\operatorname{sign}(\boldsymbol{y}) = \boldsymbol{\varepsilon}$ is independent of $\widetilde{J}_n(\boldsymbol{\varepsilon} \cdot |\boldsymbol{y}|) \neq J_\mathsf{f}$ and therefore

$$\begin{aligned}
\sup_{f\in\Sigma_{L,n}} \mathbf{P}_\mathsf{f}(\widetilde{J}_n \neq J_\mathsf{f}) &\geq \frac{1}{2^{|\mathscr{E}|}} \sum_{\boldsymbol{\varepsilon}\in\mathscr{E}} \mathbb{P}\Big(\widetilde{J}_n(\boldsymbol{\varepsilon} \cdot |\boldsymbol{y}|) \neq J_\mathsf{f}\Big) \cdot \mathbb{P}\Big(\operatorname{sign}(\boldsymbol{y}) = \boldsymbol{\varepsilon}\Big) \\
&\geq \min_{\boldsymbol{\varepsilon}\in\mathscr{E}} \mathbb{P}\Big(\widetilde{J}_n(\boldsymbol{\varepsilon} \cdot |\boldsymbol{y}|) \neq J_\mathsf{f}\Big) \geq \inf_{\widetilde{J}\in\mathscr{J}_{\mathrm{sym}}} \mathbb{P}\Big(\widetilde{J}_n(\boldsymbol{y}^2) \neq J_\mathsf{f}\Big),
\end{aligned}$$

where the last inf is taken over the set $\mathscr{J}_{\mathrm{sym}}$ of all sparsity pattern estimators depending only on $|\boldsymbol{y}|$, or equivalently on $\boldsymbol{y}^2 \triangleq \boldsymbol{y} \cdot \boldsymbol{y}$. Thus, at this stage we have proven that the inequality

$$\inf_{\widetilde{J}_n} \sup_{\mathsf{f}\in\Sigma_{L,n}} \mathbf{P}_\mathsf{f}(\widetilde{J}_n \neq J_\mathsf{f}) \geq \inf_{\widetilde{J}\in\mathscr{J}_{\mathrm{sym}}} \int_{\Sigma_{L,n}} \mathbf{P}_\mathsf{f}\Big(\widetilde{J}_n(\boldsymbol{y}^2) \neq J_\mathsf{f}\Big)\, \pi(d\mathsf{f}) \tag{30}$$

holds for every orthosymmetric probability measure π. Now, let us define the so called "least favorable prior", that is the probability measure π for which the RHS of the last inequality is strictly positive. Based on the Dirac measures $\mu_\ell = \delta_{\mathsf{f}_\ell}$, $\ell = 1, \dots, M$, we define the prior π as the orthosymmetrized version of the probability measure

$$\pi_0(d\mathsf{f}) = \frac{1}{M} \sum_{\ell=1}^{M} \mu_\ell(d\mathsf{f}).$$

Strictly speaking, for every measurable subset A of $\Sigma_{L,n}$,

$$\pi(A) = \frac{1}{2^{|\mathscr{E}|}} \sum_{\boldsymbol{\varepsilon} \in \mathscr{E}} \pi_0\Big(\Big\{ \sum_{\boldsymbol{k}} \varepsilon_{\boldsymbol{k}} \langle \mathsf{f}, \varphi_{\boldsymbol{k}} \rangle \varphi_{\boldsymbol{k}} : \mathsf{f} \in A \Big\} \Big).$$

Injecting this prior in (30) and using the notation $\mathsf{f}_{\boldsymbol{\varepsilon}} = \sum_{\boldsymbol{k}} \varepsilon_{\boldsymbol{k}} \theta_{\boldsymbol{k}} [\mathsf{f}] \phi_{\boldsymbol{k}}$, we get

$$\begin{aligned}
\inf_{\widetilde{J}_n} \sup_{\mathsf{f} \in \Sigma_{L,n}} \mathbf{P}_{\mathsf{f}}(\widetilde{J}_n \neq J_{\mathsf{f}}) &\geq \inf_{\widetilde{J} \in \mathscr{J}_{\mathrm{sym}}} \frac{1}{2^{|\mathscr{E}|}} \sum_{\boldsymbol{\varepsilon} \in \mathscr{E}} \int_{\Sigma_{L,n}} \mathbf{P}_{\mathsf{f}_{\boldsymbol{\varepsilon}}} \Big(\widetilde{J}_n(\boldsymbol{y}^2) \neq J_{\mathsf{f}_{\boldsymbol{\varepsilon}}} \Big) \pi_0(d\mathsf{f}) \\
&= \inf_{\widetilde{J} \in \mathscr{J}_{\mathrm{sym}}} \int_{\Sigma_{L,n}} \mathbf{P}_{\mathsf{f}} \Big(\widetilde{J}_n(\boldsymbol{y}^2) \neq J_{\mathsf{f}} \Big) \pi_0(d\mathsf{f}) \\
&= \frac{1}{M} \inf_{\widetilde{J} \in \mathscr{J}_{\mathrm{sym}}} \frac{1}{M} \sum_{\ell=1}^{M} \bar{\mathbf{P}}_{\mathsf{f}_\ell} \Big(\widetilde{J}_n(\boldsymbol{z}) \neq J_{\mathsf{f}_\ell} \Big).
\end{aligned}$$

The desired inequality follows now by a simple application of Lemma 3.

APPENDIX G: PROOF OF LEMMA 3

Let $M = \binom{d}{d^*}$, $\mathsf{f}_0 \equiv 0$ and let $\{\mathsf{f}_1, \dots, \mathsf{f}_M\}$ be a set included in Σ_L. Let $J_1, \dots, J_M$ be all the subsets of $\{1, \dots, d\}$ containing exactly d^* elements somehow enumerated. We apply Lemma 5 to the set of functions $\mathsf{f}_0, \dots, \mathsf{f}_M$. Let us define now the Fourier coefficients of f_1 if, for instance, f_1 corresponds to $J_1 = \{1, ..., d^*\}$. We set

$$\theta_{\boldsymbol{k}} = \begin{cases} N(d^*, \gamma)^{-1/2} & \text{if } \boldsymbol{k} \in \mathscr{C}_1(d^*, \gamma) \\ 0 & \text{otherwise} \end{cases}$$

Clearly, the vector $\theta = (\theta_{\boldsymbol{k}})$ is symmetric in the variables $\{1, ..., d^*\}$. Furthermore, it satisfies

$$\sum_{k_1 \neq 0} \theta_{\boldsymbol{k}}^2 = \sum_{\boldsymbol{k} \in \mathscr{C}_1(\gamma) \setminus \mathscr{C}_2(\gamma)} \theta_{\boldsymbol{k}}^2 = 1$$

and

$$\begin{aligned}
\sum_{\boldsymbol{k} \in \mathbb{Z}^{d^*}} k_1^2 \theta_{\boldsymbol{k}}^2 &= N(d^*, \gamma)^{-1} \sum_{\boldsymbol{k} \in \mathscr{C}_1(d^*, \gamma)} k_1^2 \\
&= N(d^*, \gamma)^{-1} \frac{1}{d^*} \sum_{j=1}^{d^*} \sum_{\boldsymbol{k} \in \mathscr{C}_1(d^*, \gamma)} k_j^2 \\
&= N(d^*, \gamma)^{-1} \frac{1}{d^*} \sum_{\boldsymbol{k} \in \mathscr{C}_1(d^*, \gamma)} \|\boldsymbol{k}\|_2^2 \\
&\leq \gamma N_1(d^*, \gamma) / N(d^*, \gamma).
\end{aligned}$$

The results stated in Section 4 imply that $N_1(d^*, \gamma)/N(d^*, \gamma) \sim_{d^* \to \infty} 1 + (\mathsf{h}(z_\gamma) - 1)^{-1}$. According to the choice of γ, for d^* large enough, $\mathsf{f}_1 \in \Sigma_L$. Let us define the functions $\mathsf{f}_2, \mathsf{f}_3, \dots, \mathsf{f}_M$ in a similar way by only changing the sparsity pattern; the function f_ℓ has the sparsity pattern J_ℓ.

To apply Lemma 3, we need to upper-bound $\mathscr{K}(\bar{\mathbf{P}}_{\mathsf{f}_\ell}, \bar{\mathbf{P}}_{\mathsf{f}_0})$ for every $\ell \in \{1, \dots, M\}$. To ease notation, we write $\bar{\mathbf{P}}_\ell$ instead of $\bar{\mathbf{P}}_{\mathsf{f}_\ell}$. Using the symmetry and the well-known inequalities relating the Kullback-Leibler divergence to the χ^2-distance between probability measures, we have $\mathscr{K}(\bar{\mathbf{P}}_j, \bar{\mathbf{P}}_0) =$

$\mathscr{K}(\bar{\mathbf{P}}_1,\bar{\mathbf{P}}_0)\leq\log\big(1+\chi_2(\bar{\mathbf{P}}_1,\bar{\mathbf{P}}_0)\big)$, where the χ_2 distance between two probabilities P and Q such that $P\ll Q$ is defined by :

$$\chi_2(P,Q)=\int\left(\frac{dP}{dQ}-1\right)^2 dQ=\int\left(\frac{dP}{dQ}\right)^2 dQ-1.$$

In view of (Efromovich and Low, 1996, lemme 5.1) , we obtain

$$1+\chi_2(\bar{\mathbf{P}}_1,\bar{\mathbf{P}}_0)\leq\exp\Big(n^2\sum_{\boldsymbol{k}\in\mathbb{Z}^{d^*}}\theta_{\boldsymbol{k}}^4\Big)=\exp\Big(\frac{n^2N_1(d^*,\gamma)}{N(d^*,\gamma)^2}\Big)\leq\exp\Big(\frac{n^2L}{N(d^*,\gamma)}\Big).$$

The assumption made in the theorem entails that $\frac{1}{M}\sum_{j=1}^M K(\bar{\mathbf{P}}_j,\bar{\mathbf{P}}_0)\leq\alpha\log M$ and we conclude thanks to Lemmas 3 and 5.

APPENDIX H: PROOF OF PROPOSITION 6

Let $M=\binom{d}{d^*}$ and let $\{\mathsf{f}_0,\mathsf{f}_1,\ldots,\mathsf{f}_M\}$ 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 $\mathsf{f}_0\equiv 0$ and 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 $\log\binom{d}{d^*}$ is larger than or equal to $d^*\log(d/d^*)=d^*\big(\log d-\log d^*\big)$.

ACKNOWLEDGMENTS

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under the grant PARCIMONIE. The initial version of this paper has been reviewed and accepted by the conference COLT 2011. The authors would like to thank the reviewers for very useful remarks.

REFERENCES

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.

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

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

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

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.

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

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.

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.

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

Andrew V. Carter. Asymptotic approximation of nonparametric regression experiments with unknown variances. *Ann. Statist.*, 35(4):1644–1673, 2007.

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

Laëtitia Comminges and Arnak S. Dalalyan. Tight conditions for consistent variable selection in high dimensional nonparametric regression. In *COLT*, arXiv:1102.3616v1 [math.ST], 2011.

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

Jean Dieudonné. *Calcul infinitésimal*. Hermann, Paris, 1968.

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. With electronic supplementary materials available online.

Sam Efromovich and Mark Low. On optimal adaptive estimation of a quadratic functional. *Ann. Statist.*, 24(3):1106–1125, 1996.

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

Robert M. Fano. *Transmission of information: A statistical theory of communications*. The M.I.T. Press, Cambridge, Mass., 1961.

Georgi K. Golubev, Michael Nussbaum, and Harrison H. Zhou. Asymptotic equivalence of spectral density estimation and Gaussian white noise. *Ann. Statist.*, 38(1):181–214, 2010.

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

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.

Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. Technical report, arXiv:0904.3523, 2009.

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

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

Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara van de Geer. Oracle inequalities and optimal inference under group sparsity. Technical report, arXiv:1007.1771, 2010.

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

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

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(4):417–473, 2010.

Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. High-dimensional union sup-

port recovery in multivariate. *The Annals of Statistics*, to appear, 2011.

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.

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

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

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.

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

Jo-Anne Ting, Aaron D'Souza, Sethu Vijayakumar, and Stefan Schaal. Efficient learning and feature selection in high-dimensional regression. *Neural Comput.*, 22(4):831–886, 2010.

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

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

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

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

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

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
ÉCOLE DES PONTS PARISTECH
6, AV BLAISE PASCAL - CITÉ DESCARTES
CHAMPS-SUR-MARNE
77455 MARNE-LA-VALLÉE CEDEX 2 - FRANCE
E-MAIL: laetitia.comminges,dalalyan@imagine.enpc.fr