Tight conditions for consistent variable selection in high dimensional nonparametric regression

Laëtitia Comminges
Université Paris Est/ ENPC
LIGM/IMAGINE
laetitia.comminges@imagine.enpc.fr

Arnak S. Dalalyan
Université Paris Est/ ENPC
LIGM/IMAGINE
dalalyan@imagine.enpc.fr

Abstract

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

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 ambient dimension of covariates, denoted by d, 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 covariates, 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 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 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 covariates 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 covariates 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 covariates d^* tends to infinity with $n \to \infty$, then there exist real numbers $\underline{c}_i$ and $\bar{c}_i$, $i = 1, \dots, 4$ such that $\underline{c}_i > 0$, $\bar{c}_i > 0$ for $i = 1, 2, 3$ and (a) if $\underline{c}_1 d^* + \underline{c}_2 \log d^* + \underline{c}_3 \log\log d - \log n < \underline{c}_4$ the estimator proposed in Section 3 is consistent and (b) no estimator of the sparsity pattern may be consistent if $\bar{c}_1 d^* + \bar{c}_2 \log d^* + \bar{c}_3 \log\log d - \log n > \bar{c}_4$.

- 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 $d^* \ge c^0 \log n$.

Very surprisingly, the derivation of these results required from us to apply some tools from complex analysis, such as the Jacobi θ-function and the saddle point method, in order to evaluate the number of lattice points lying in a ball of an Euclidean space with increasing dimension.

The rest of the paper is organized as follows. The notation and assumptions necessary for stating our main results are presented in Section 2. In Section 3, an estimator of the set of relevant covariates is introduced and its consistency is established. The principal condition required in the consistency

result involves the number of lattice points in a ball of a high-dimensional Euclidean space. An asymptotic equivalent for this number is obtained 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, while the relation between consistency and inconsistency results are discussed in Section 6. The technical parts of the proofs are postponed to the Appendix.

2 Notation and assumptions

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 covariates.

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.

[C1] 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 $\boldsymbol{x} \in [0,1]^d$.

To describe the smoothness assumption imposed on f, let us introduce the 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

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

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 covariates are sufficiently relevant for making their identification possible. This is done by means of the following condition.

[C2(κ, L)**]** The regression function f belongs to Σ_L. Furthermore, for some subset $J \subset \{1,\dots,d\}$ of cardinality $\leq d^*$, there exists a function $\bar{\mathsf{f}} : \mathbb{R}^{|J|} \to \mathbb{R}$ such that $\mathsf{f}(\boldsymbol{x}) = \bar{\mathsf{f}}(\boldsymbol{x}_J)$, $\forall \boldsymbol{x} \in \mathbb{R}^d$ and it holds that

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

Hereafter, we will refer to J as the sparsity pattern of f.

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 $\boldsymbol{Q}[\mathsf{f}] = (Q_1[\mathsf{f}],\dots,Q_d[\mathsf{f}])$.

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

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

Remark 1 *The primary aim of this work is to understand when it is possible to estimate the sparsity pattern (with theoretical guarantees on the convergence of the estimator) and when it is impossible. The estimator that we will define in the next section 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]**.

3 Consistent estimation of the set of relevant variables

The estimator of the sparsity pattern J that we are going to introduce now is based on the following simple observation: if $j \notin J$ then $\theta_{\boldsymbol{k}}[\mathrm{f}] = 0$ 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 $|\theta_{\boldsymbol{k}}[\mathrm{f}]| > 0$. To turn this observation into an estimator of J, we start by estimating the Fourier coefficients $\theta_{\boldsymbol{k}}[\mathrm{f}]$ by their empirical counterparts:

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

Then, for every $\ell \in \mathbb{N}$ and for any $\gamma > 0$, we introduce the notation $S_{m,\ell} = \{\boldsymbol{k} \in \mathbb{Z}^d : \|\boldsymbol{k}\|_2 \le m,\ \|\boldsymbol{k}\|_0 \le \ell\}$ and $N(d^*, \gamma) = \mathrm{Card}\{\boldsymbol{k} \in \mathbb{Z}^{d^*} : \|\boldsymbol{k}\|_2^2 \le \gamma d^* \,\&\, k_1 \neq 0\}$. Finally our estimator is defined by

$$\widehat{J}_n(m, \lambda) = \Big\{ j \in \{1, \dots, d\} : \max_{\boldsymbol{k} \in S_{m,d^*} : k_j \neq 0} |\widehat{\theta}_{\boldsymbol{k}}| > \lambda \Big\}, \tag{3}$$

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 1 *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*

$$\frac{L_\infty^2 d^* \log(6md)}{n} \le L_2^2, \quad \textit{and} \quad \frac{128(\sigma + L_2)^2 d^* N(d^*, 2L/\kappa) \log(6md)}{n g_{\min}^2} \le \kappa, \tag{4}$$

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 1 ensuring the consistency of the estimator $\widehat{J}$, it becomes clear that the strongest requirement is the second inequality in (4). 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 (4) 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 1. 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), than the estimation of sup-norm.

The result stated above can be reformulated to provide also a level of relevance κ for the covariates of $\boldsymbol{X}$ making their identification possible. In fact, an alternative way of stating Theorem 1 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 A.

4 Counting lattice points in a ball

The aim of the present section is to investigate the properties of the quantity $N(d^*, m^2/d^*)$ that is involved in the conditions ensuring the consistency of the proposed procedure. Quite surprisingly, the

asymptotic behavior of $N(d^*, m^2/d^*)$ 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

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

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 with the first coordinate equal to zero and lying in the d^*-dimensional ball with radius $(\gamma d^*)^{1/2}$ and centered at the origin. With this notation, the quantity $N(d^*,2L/\kappa)$ of Theorem 1 can be written as $N_1(d^*,2L/\kappa)-N_2(d^*,2L/\kappa)$.

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 $N_i(\gamma)$ and the θ-function comes from, let us denote by $\{a_r\}$ the sequence of coefficients of the power series of $\mathsf{h}(z)^{d^*}$, that is $\mathsf{h}(z)^{d^*}=\sum_{r\geq 0} a_r z^r$. One easily checks that $\forall r\in\mathbb{N}$, $a_r=\mathrm{Card}\{\boldsymbol{k}\in\mathbb{Z}^{d^*}: k_1^2+\ldots+k_{d^*}^2=r\}$. 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 such that γd^* is an integer and let $\mathsf{l}_\gamma(z)=\log\mathsf{h}(z)-\gamma\log z$.*

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

Furthermore, the convergence to zero of the terms replaced by $o(1)$ in the previous formulae is uniform in γ on any compact set $[\underline{\gamma},\overline{\gamma}]\subset(0,\infty)$.

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

$$\log\big(N_1(d^*,\gamma)-N_2(d^*,\gamma)\big)=d^*\mathsf{l}_\gamma(z_\gamma)-\frac{1}{2}\log d^*-\log\left\{\frac{\mathsf{h}(z_\gamma)z_\gamma(1-z_\gamma)(2\mathsf{l}''_\gamma(z_\gamma)\pi)^{1/2}}{\mathsf{h}(z_\gamma)-1}\right\}+o(1). \tag{5}$$

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 1 the plots of these quantities as functions of $\gamma>0$.

5 Tightness of the assumptions

In this section, 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 $\widetilde{\Sigma}(\kappa,L)$ of all functions satisfying assumption [**C2**(κ,L)]. In order to avoid irrelevant technicalities and to better convey the main results, we assume that $\kappa=1$ and denote $\widetilde{\Sigma}_L=\widetilde{\Sigma}(1,L)$. Furthermore, we will assume that the design $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ is fixed and satisfies

$$\frac{1}{n}\sum_{i=1}^n \varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)\varphi_{\boldsymbol{k}'}(\boldsymbol{X}_i)\leq\frac{n}{N_1(d^*,L)^2} \tag{6}$$

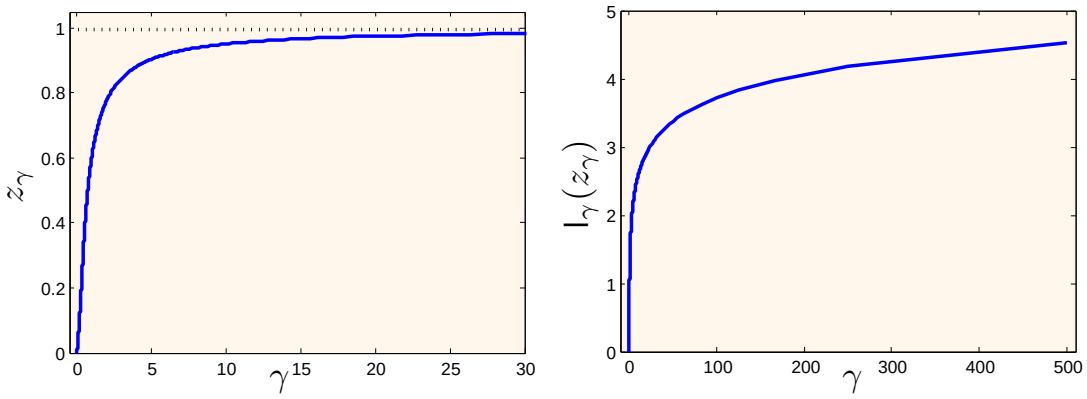

Figure 1: The plots of mappings $\gamma \mapsto z_\gamma$ and $\gamma \mapsto \mathsf{I}_\gamma(z_\gamma)$.

for all distinct $\boldsymbol{k}, \boldsymbol{k}' \in S_{(d^*L)^{1/2},d^*} \subset \mathbb{Z}^d$. 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 $n \geq n_0$,

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

where the inf is over all possible estimators of J_f. To lower bound the LHS of the last inequality, we introduce a set of $M+1$ probability distributions $\mu_0, \ldots, \mu_M$ on $\tilde{\Sigma}_L$ and use the fact that

$$\inf_{\widetilde{J}} \sup_{\mathsf{f} \in \tilde{\Sigma}_L} \mathbf{P}_\mathsf{f}(\widetilde{J} \neq J_\mathsf{f}) \geq \inf_{\widetilde{J}} \frac{1}{M+1} \sum_{\ell=0}^{M} \int_{\tilde{\Sigma}_L} \mathbf{P}_\mathsf{f}(\widetilde{J} \neq J_\mathsf{f}) \mu_\ell(d\mathsf{f}). \tag{7}$$

These measures μ_ℓ will be chosen in such a way that for each $\ell \geq 1$ there is a set J_ℓ of cardinality d^* such that $\mu_\ell\{J_\mathsf{f} = J_\ell\} = 1$ and all the sets $J_1, \ldots, J_M$ are distinct. The measure μ_0 is the Dirac measure in 0. Considering these μ_ℓs as "prior" probability measures on $\tilde{\Sigma}_L$ and defining the corresponding "posterior" probability measures $\mathbb{P}_0, \mathbb{P}_1, \ldots, \mathbb{P}_M$ by

$$\mathbb{P}_\ell(A) = \int_{\tilde{\Sigma}_L} \mathbf{P}_\mathsf{f}(A) \mu_\ell(d\mathsf{f}), \quad \text{for every measurable set } A \subset \mathbb{R}^n,$$

we can write the inequality (7) as

$$\inf_{\widetilde{J}} \sup_{\mathsf{f} \in \tilde{\Sigma}_L} \mathbf{P}_\mathsf{f}(\widetilde{J} \neq J_\mathsf{f}) \geq \inf_{\psi} \frac{1}{M+1} \sum_{\ell=0}^{M} \mathbb{P}_\ell(\psi \neq \ell), \tag{8}$$

where the inf is taken over all random variables ψ taking values in $\{0, \ldots, M\}$. 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 2 (Corollary 2.6 of Tsybakov (2009)) *Let $(\mathscr{X}, \mathscr{A})$ be a measurable space and let $P_0, \ldots, P_M$ be probability measures on $(\mathscr{X}, \mathscr{A})$. Let us set $\bar{p}_{e,M} = \inf_\psi (M+1)^{-1} \sum_{\ell=0}^{M} P_\ell(\psi \neq \ell)$ where the* inf *is taken over all measurable functions $\psi : \mathscr{X} \to \{0, \ldots, M\}$. If for some $0 < \alpha < 1$*

$$\frac{1}{M+1} \sum_{\ell=0}^{M} \mathscr{K}(P_\ell, P_0) \leq \alpha \log M,$$

then

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

It follows from this 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 $\mathbb{P}_\ell$ and $\mathbb{P}_0$. This roughly means that the measures μ_ℓ should not be very far from μ_0 but the probability measures μ_ℓ should be very different one from another in terms of the sparsity pattern of a function f randomly drawn according to μ_ℓ. This property is ensured by the following result.

Lemma 3 *Suppose $\mu_0 = \delta_0$, the Dirac measure at $0 \in \Sigma_L$. Let S be a subset of $\mathbb{Z}^d$ of cardinality $|S|$ and A be a constant. Define μ_S as a discrete measure supported on the finite set of functions $\{\mathsf{f}_{\boldsymbol{\omega}} = \sum_{\boldsymbol{k}\in S} A\omega_{\boldsymbol{k}}\varphi_{\boldsymbol{k}} : \boldsymbol{\omega} \in \{\pm 1\}^S\}$ such that $\mu_S(\mathsf{f} = \mathsf{f}_{\boldsymbol{\omega}}) = 2^{-|S|}$ for every $\boldsymbol{\omega} \in \{\pm 1\}^S$, i.e., the $\omega_{\boldsymbol{k}}$'s are i.i.d. Rademacher random variables under μ_S. If, for some $\epsilon \geq 0$, the condition*

$$\frac{1}{n}\sum_{i=1}^{n} \varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)\varphi_{\boldsymbol{k}'}(\boldsymbol{X}_i) \leq \epsilon \qquad \forall \boldsymbol{k}, \boldsymbol{k}' \in S$$

is fulfilled, then

$$\mathscr{K}(\mathbb{P}_1, \mathbb{P}_0) \leq \log\left[\int \Big(\frac{d\mathbb{P}_1}{d\mathbb{P}_0}(\boldsymbol{y})\Big)^2 \mathbb{P}_0(d\boldsymbol{y})\right] \leq 4|S|A^4 n^2\Big\{1 + \frac{|S|\epsilon}{4nA^2}\Big\}.$$

These evaluations lead to the following theorem, that tells us that the conditions to which we have resorted for proving the consistency in Section 3 are nearly optimal.

Theorem 4 *Let the design $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in [0,1]^d$ be deterministic and satisfy (6). Let γ^* the largest real number such that $d^*\gamma^*$ is integer and $L \geq \gamma^*(1 + 1/2z_{\gamma^*})$. If for some positive number $\alpha < (\log 3 - \log 2)/\log 3$*

$$\frac{(N_1(d^*,\gamma^*) - N_2(d^*,\gamma^*))^2 \log\binom{d}{d^*}}{n^2 N_1(d^*,\gamma^*)} \geq \frac{\alpha}{5}, \tag{9}$$

then there exists a positive constant $c > 0$ and a $d_0 \in \mathbb{N}$ such that, if $d^* \geq d_0$,

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

Proof: We apply the Fano lemma with $M = \binom{d}{d^*}$. We choose $\mu_0, \ldots, \mu_M$ as follows. μ_0 is the Dirac measure δ_0, μ_1 is defined as in Lemma 3 with $S = \mathscr{C}_1(d^*,\gamma^*)$ and $A = \big[N_1(d^*,\gamma^*) - N_2(d^*,\gamma^*)\big]^{-1/2}$. The measures $\mu_2, \ldots, \mu_M$ are defined similarly and correspond to the $M-1$ remaining sparsity patterns of cardinality d^*.

In view of inequality (8) and Lemma 2, it suffices to show that the measures μ_ℓ satisfy $\mu_\ell(\widetilde{\Sigma}_L) = 1$ and $\sum_{\ell=0}^{M} \mathscr{K}(\mathbb{P}_\ell, \mathbb{P}_0) \leq (M+1)\alpha \log M$. Combining Lemma 3 with $\mathrm{Card}(S) = N_1(d^*,\gamma^*)$ and condition (6), one easily checks that equation (9) implies the desired bound on $\sum_{\ell=0}^{M} \mathscr{K}(\mathbb{P}_\ell, \mathbb{P}_0)$.

Let us show now that $\mu_1(\widetilde{\Sigma}_L) = 1$. By symmetry, this will imply that $\mu_\ell(\widetilde{\Sigma}_L) = 1$ for every ℓ. Since μ_1 is supported by the set $\{\mathsf{f}_{\boldsymbol{\omega}} : \boldsymbol{\omega} \in \{\pm 1\}^{\mathscr{C}_1(d^*,\gamma^*)}\}$, it is clear that

$$\sum_{k_1 \neq 0} \theta_{\boldsymbol{k}}^2[\mathsf{f}_{\boldsymbol{\omega}}] = A^2[N_1(d^*,\gamma^*) - N_2(d^*,\gamma^*)] = 1$$

and, for every $j = 1, \ldots, d^*$,

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

By virtue of Proposition 1, as d^* tends to infinity, $N_1(d^*,\gamma^*)/N_2(d^*,\gamma^*)$ is asymptotically equivalent to $\mathsf{h}(z_{\gamma^*}) > 1 + 2z_{\gamma^*}$. Hence, for d^* large enough,

$$A^2 N_1(d^*,\gamma^*) = \frac{N_1(d^*,\gamma^*)}{N_1(d^*,\gamma^*) - N_2(d^*,\gamma^*)} < \frac{1}{2z_{\gamma^*}} + 1.$$

As a consequence, for every $j = 1, \ldots, d^*$,

$$\sum_{\boldsymbol{k}\in\mathbb{Z}^d} k_j^2 \theta_{\boldsymbol{k}}^2[\mathsf{f}_{\boldsymbol{\omega}}] \leq \gamma^*\Big(\frac{1}{2z_{\gamma^*}} + 1\Big) \leq L,$$

where the last inequality follows from the definition of γ^*. ■

Note that Theorem 4 is concerned by the case where the intrinsic dimension is not too small, which is the most interesting case in the present context. However, a much simpler result can be established showing that the conditions of Theorem 1 are tight in the case of fixed intrinsic dimension as well.

Proposition 2 *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 $\inf_{\widetilde{J}_n} \sup_{f\in\widetilde{\Sigma}_L} \mathbf{P}_f(\widetilde{J}_n \neq J_f) \geq c$.

6 Discussion

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

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

In the fixed sparsity regime, in view of Theorem 1, consistent estimation of the sparsity pattern can be achieved using the estimator $\widehat{J}$ as soon as $(\log d)/n \leq c_\star$, where $c_\star$ is the constant defined by

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

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 Proposition 2, 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 result of Theorem 1 cannot be improved.

In the regime of increasing sparsity, the second inequality in (4) is the most stringent one. Taking the logarithm of both sides and using formula (5) for $N(d^*,2L) = N_1(d^*,2L) - N_2(d^*,2L)$, we see that consistent estimation of J is possible when

$$\underline{c}_1 d^* + \frac{1}{2}\log d^* + \log\log d - \log n < \underline{c}_2, \tag{10}$$

with $\underline{c}_1 = \mathsf{I}_{2L}(z_{2L})$ and $\underline{c}_2 = 2(\log(g_{\min}) - \log(17(\sigma+L_2)) + \log\Big\{\frac{\mathsf{h}(z_{2L})z_{2L}(1-z_{2L})(2\mathsf{I}''_{2L}(z_{2L})\pi)^{1/2}}{\mathsf{h}(z_{2L})-1}\Big\}$. On the other hand, by virtue of (5), $\log\Big\{\frac{[N_1(d^*,\gamma)-N_2(d^*,\gamma)]^2}{N_1(d^*,\gamma)}\Big\} = d^*\mathsf{I}_\gamma(z_\gamma) - \frac{1}{2}\log d^* - \log\Big\{\frac{\mathsf{h}(z_\gamma)^2 z_\gamma(1-z_\gamma)(2\mathsf{I}''_\gamma(z_\gamma)\pi)^{1/2}}{(\mathsf{h}(z_\gamma)-1)^2}\Big\} + o(1)$. Therefore, Theorem 4 yields that it is impossible to consistently estimate J if

$$\bar{c}_1 d^* + \frac{1}{2}\log d^* + \log\log d - 2\log n > \bar{c}_2, \tag{11}$$

where $\bar{c}_1 = \mathsf{I}_{\gamma^*}(z_{\gamma^*})$ and $\bar{c}_2 = \log\Big\{\frac{\mathsf{h}(z_{\gamma^*})^2 z_{\gamma^*}(1-z_{\gamma^*})(2\mathsf{I}''_{\gamma^*}(z_{\gamma^*})\pi)^{1/2}}{(\mathsf{h}(z_{\gamma^*})-1)^2}\Big\} + \log\log(3/2) - \log 5 - \log\log 3$. A very simple consequence of inequalities (10) and (11) 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 $\log\log d = o(\log n)$.

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 estimator proposed in Section 3, it is worth noting that its 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 covariates X^j corresponding to nonzero coordinates of $\boldsymbol{k}$ are considered as relevant. We can stop this computation as soon as the number of covariates 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 1, 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 (10) 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 (10) 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.

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.

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.

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.

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.

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 support 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.

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.

Appendix

A Proof of Theorem 1

The empirical Fourier coefficients can be decomposed as follows:

$$\widehat{\theta}_{\boldsymbol{k}} = \tilde{\theta}_{\boldsymbol{k}} + z_{\boldsymbol{k}}, \qquad \text{where} \qquad \tilde{\theta}_{\boldsymbol{k}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\mathrm{g}(\boldsymbol{X}_i)}\mathrm{f}(\boldsymbol{X}_i) \qquad \text{and} \qquad z_{\boldsymbol{k}} = \frac{\sigma}{n}\sum_{i=1}^{n} \frac{\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)}{\mathrm{g}(\boldsymbol{X}_i)}\boldsymbol{\varepsilon}_i. \tag{12}$$

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

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

with

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

Therefore, for every $\boldsymbol{k} \in S_{m,d^*}$, it holds that $\mathbf{P}\big(|z_{\boldsymbol{k}}| > x | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\big) \le \exp(-n g_{\min}^2 x^2/4\sigma^2)$. 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

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

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 $\lambda_2 = 4L_2\Big(\frac{d^*\log(6md)}{n g_{\min}^2}\Big)^{1/2}$. 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 (4) 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 $Q_{j_0} \geq \kappa$.

B 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 $\phi(y) = e^{-y}\mathsf{h}'(e^{-y})/\mathsf{h}(e^{-y}) = \sum_{k=-\infty}^{k=+\infty} k^2 e^{-y k^2} / \sum_{k=-\infty}^{k=+\infty} e^{-y k^2}$ in such a way that

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

By virtue of the Cauchy-Schwarz inequality, it holds that

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

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

$$N_1(d^*, \gamma) = \frac{1}{2\pi i} \oint_{|z| = z_\gamma} \frac{\mathsf{h}(z)^{d^*}}{z^{\gamma d^*}} \frac{dz}{z(1-z)} = \frac{1}{2\pi i} \oint_{|z| = z_\gamma} \{z(1-z)\}^{-1} e^{d^* \mathsf{l}_\gamma(z)} dz. \tag{13}$$

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[1] $\Re[\mathsf{l}_\gamma(z) - \mathsf{l}_\gamma(z_\gamma)] \leq -\mu$ is satisfied for some $\mu > 0$ and for any z belonging to the circle $|z| = |z_\gamma|$ and lying not too close to z_γ. To check that this is indeed the case, we

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

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

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

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

C Proof of Lemma 3

Let $\phi(\cdot)$ be the density of $\mathcal{N}(0,1)$ and let

$$p_{\mathsf{f}}(\boldsymbol{y}) \triangleq \prod_{i=1}^n \phi\Big(y_i - \mathsf{f}(\boldsymbol{X}_i)\Big), \qquad \forall \boldsymbol{y} \in \mathbb{R}^n.$$

Since the errors $\boldsymbol{\varepsilon}_i$ are Gaussian, the posterior probabilities $\mathbb{P}_0$ and $\mathbb{P}_1$ are absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^n$ and admit the densities

$$p_0(\boldsymbol{y}) = \prod_{i=1}^n \phi(y_i), \qquad \text{and} \qquad p_1(\boldsymbol{y}) = \mathbf{E}_{\mathsf{f}\sim\mu_S} p_{\mathsf{f}}(\boldsymbol{y}), \qquad \forall \boldsymbol{y} \in \mathbb{R}^n.$$

Simple algebra yields:

$$p_{\mathsf{f}}(\boldsymbol{y}) = C_{\mathsf{f}}\, p_0(\boldsymbol{y}) \prod_{i=1}^n \exp\Big\{y_i \mathsf{f}(\boldsymbol{X}_i)\Big\}, \quad \forall \boldsymbol{y} \in \mathbb{R}^n,$$

where $C_{\mathsf{f}} = \prod_{i=1}^n \exp\big\{-\mathsf{f}(\boldsymbol{X}_i)^2/2\big\}$. Thus,

$$\frac{p_1}{p_0}(\boldsymbol{y}) = \mathbf{E}_{\mathsf{f}\sim\mu_S}\Big[C_{\mathsf{f}} \prod_{i=1}^n \exp\Big\{y_i \mathsf{f}(\boldsymbol{X}_i)\Big\}\Big].$$

Therefore,

$$\begin{aligned}
\int_{\mathbb{R}^n} \Big(\frac{p_1}{p_0}(\boldsymbol{y})\Big)^2 p_0(\boldsymbol{y}) d\boldsymbol{y} &= \mathbf{E}_{(\mathsf{f},\mathsf{f}')\sim\mu_S\otimes\mu_S}\Big[C_{\mathsf{f}} C_{\mathsf{f}'} \int_{\mathbb{R}^n} \prod_{i=1}^n \Big(\exp\Big\{y_i(\mathsf{f}+\mathsf{f}')(\boldsymbol{X}_i)\Big\}\phi(y_i)\Big) d\boldsymbol{y}\Big] \\
&= \mathbf{E}_{(\mathsf{f},\mathsf{f}')\sim\mu_S\otimes\mu_S}\Big[C_{\mathsf{f}} C_{\mathsf{f}'} \prod_{i=1}^n \exp\Big(\frac{1}{2}(\mathsf{f}+\mathsf{f}')^2(\boldsymbol{X}_i)\Big)\Big] \\
&= \mathbf{E}_{(\mathsf{f},\mathsf{f}')\sim\mu_S\otimes\mu_S}\Big[\exp\Big(\sum_{i=1}^n \mathsf{f}(\boldsymbol{X}_i)\mathsf{f}'(\boldsymbol{X}_i)\Big)\Big] \\
&= \frac{1}{2^{2|S|}} \sum_{\boldsymbol{\omega},\boldsymbol{\omega}'\in\{\pm 1\}^S} \prod_{\boldsymbol{k},\boldsymbol{k}'\in S} \exp\Big(\omega_{\boldsymbol{k}}\omega'_{\boldsymbol{k}'} b_{\boldsymbol{k}\boldsymbol{k}'}\Big),
\end{aligned}$$

where $b_{\boldsymbol{k}\boldsymbol{k}'} = A^2 \sum_{i=1}^n \varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)\varphi_{\boldsymbol{k}'}(\boldsymbol{X}_i)$, for all $\boldsymbol{k}, \boldsymbol{k}' \in S$. Note that $0 \leq b_{\boldsymbol{k}\boldsymbol{k}} \leq 2A^2 n$ and $|b_{\boldsymbol{k}\boldsymbol{k}'}| \leq A^2 n\epsilon$, for all $\boldsymbol{k}, \boldsymbol{k}' \in S$ such that $\boldsymbol{k}' \neq \boldsymbol{k}$. Now, on the one hand, for a fixed pair $(\boldsymbol{\omega}, \boldsymbol{\omega}')$, we have

$$\prod_{\boldsymbol{k}\neq\boldsymbol{k}'} \exp\Big(\omega_{\boldsymbol{k}}\omega'_{\boldsymbol{k}'} b_{\boldsymbol{k}\boldsymbol{k}'}\Big) \leq \exp\big(|S|^2 A^2 n\epsilon\big).$$

On the other hand, if we are given a sequence of numbers $(b_{\boldsymbol{k}\boldsymbol{k}})$ indexed by S, we have

$$\frac{1}{2^{2|S|}} \sum_{\boldsymbol{\omega},\boldsymbol{\omega}'} \prod_{\boldsymbol{k}\in S} e^{\omega_{\boldsymbol{k}}\omega'_{\boldsymbol{k}} b_{\boldsymbol{k}\boldsymbol{k}}} = \prod_{\boldsymbol{k}\in S} \frac{e^{b_{\boldsymbol{k}\boldsymbol{k}}} + e^{-b_{\boldsymbol{k}\boldsymbol{k}}}}{2} \leq \prod_{\boldsymbol{k}\in S} e^{b_{\boldsymbol{k}\boldsymbol{k}}^2} \leq \exp\Big(4|S|A^4 n^2\Big).$$

From these remarks it results that

$$\int_{\mathbb{R}}^d \Big(\frac{p_1}{p_0}(\boldsymbol{y})\Big)^2 p_0(\boldsymbol{y}) d\boldsymbol{y} \leq \exp\Big(4|S|A^4 n^2\Big\{1 + \frac{|S|\epsilon}{4nA^2}\Big\}\Big),$$

and the claim of the lemma follows.

D Proof of Proposition 2

Let $M = \binom{d}{d^*}$ and let $\{\mathsf{f}_0, \mathsf{f}_1, \ldots, \mathsf{f}_M\}$ be a set included in $\widetilde{\Sigma}_L$. 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 $\{\theta^\ell_{\boldsymbol{k}} : \boldsymbol{k} \in \mathbb{Z}^d\}$ as follows:

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

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 $d^* \log(d/d^*) = d^* \big(\log d - \log d^*\big)$.