Slope heuristics and V-Fold model selection in heteroscedastic regression using strongly localized bases

Fabien Navarro, Adrien Saumard

Slope heuristics and V-Fold model selection in heteroscedastic regression using strongly localized bases

Fabien Navarro and Adrien Saumard

CREST-ENSAI, BRUZ, FRANCE

Abstract

We investigate the optimality for model selection of the so-called slope heuristics, V-fold cross-validation and V-fold penalization in a heteroscedatic with random design regression context. We consider a new class of linear models that we call strongly localized bases and that generalize histograms, piecewise polynomials and compactly supported wavelets. We derive sharp oracle inequalities that prove the asymptotic optimality of the slope heuristics—when the optimal penalty shape is known—and V-fold penalization. Furthermore, V-fold cross-validation seems to be suboptimal for a fixed value of V since it recovers asymptotically the oracle learned from a sample size equal to $1 - V^{-1}$ of the original amount of data. Our results are based on genuine concentration inequalities for the true and empirical excess risks that are of independent interest. We show in our experiments the good behavior of the slope heuristics for the selection of linear wavelet models. Furthermore, V-fold cross-validation and V-fold penalization have comparable efficiency.

AMS 2000 subject classifications— 62G08, 62G09
Key words and phrases— Nonparametric regression, heteroscedastic noise, random design, model selection, cross-validation, wavelets

Contents

Introduction

The main goal of this paper is to substantially extend the study, in a heteroscedastic regression with random design context, of the *optimality* of two general model selection devices: the so-called slope heuristics and V-fold resampling strategies. More precisely, we consider projection estimators on some general linear models and investigate from a theoretical perspective the possibility to derive optimal oracle inequalities for the considered model selection procedures. We also experiment and compare the procedures for the selection of linear wavelet models.

The slope heuristics [12] is a recent calibration method of penalization procedures in model selection: from the knowledge of a (good) penalty shape it allows to calibrate a penalty that performs an accurate model selection. It is based on the existence of a minimal penalty, around which there is a drastic change in the behavior of the model selection procedure. Moreover, the optimal penalty is simply linked to the minimal one by a factor two. The slope heuristics is thus a general method for the selection of M-estimators [9] and it has been successfully applied in various methodological studies surveyed in [10].

However, there is a gap between the wide range of applicability of the slope heuristics and its theoretical justification. Indeed, there are only a few studies, in quite restrictive frameworks, that theoretically describe the optimality of this penalty calibration procedure. First, Birgé and Massart [12] have shown the validity of the slope heuristics in a generalized linear Gaussian model setting, including the case of homoscedastic regression with fixed design. Then, Arlot and Massart [9] validated the slope heuristics in a heteroscedastic with random design regression framework, for the selection of linear models of histograms. These result has been extended to the case of piecewise polynomial functions in [37]. Lerasle [28, 29] has shown the optimality of the slope heuristics in least-squares density estimation for the selection of some linear models for both independent and dependent data. It has also been shown in [35]—refining previous partial results of [17]—that the slope heuristics is valid for the selection of histograms in maximum likelihood density estimation. On the negative side, Arlot and Bach [5] proved that the constant two between the minimal penalty and the optimal one is not always valid for the selection of linear estimators in least-squares regression with fixed design. For instance, kernel ridge regression leads to a ratio between the optimal penalty and the minimal one that takes values between 1 and 2. The existence of a minimal penalty—that can be estimated in practice—seems to be general however, even for the selection of linear estimators.

If the noise is homoscedastic, then the shape of the ideal penalty is known and is linear in the dimension of the models as in the case of Mallows' C_p. However, if the noise is heteroscedastic, then Arlot [4] showed that the ideal penalty is not in general a function of the linear dimension of the models. Hence, it is likely that finding a good penalty shape in order to use the slope heuristics will be hard and another approach would be needed. Probably, the most commonly used method to select an hyperparameter—such as the linear dimension of the models in our problem—in practice is the V-fold cross-validation (VFCV) procedure [22], with V classically taken to be equal to 5 or 10.

Despite its wide success in practice, there is still quite few theoretical results concerning VFCV, that are surveyed in [6]. Some asymptotic results are described in [23]. Some papers more specifically address the efficiency of VFCV as a model selection tool by deriving oracle inequalities. But most results, such as in [27] in a general learning context or in [40] for least-squares regression, do not allow to tackle the question of the *optimality* of the procedure as a model selection tool, since they prove oracle inequalities with unknown or suboptimal leading constant. A notable exception is [3], which proves that VFCV for a fixed V is indeed asymptotically suboptimal for the selection of regressograms. This is simply explained by the fact that VFCV gives a biased estimation of the risk, as emphasized earlier by Burman [13], who proposed to remove this bias.

Building on ideas of [13], Arlot [3] defined the so-called V-fold penalization and proved its asymptotic optimality, even for fixed V, for the selection of histograms. In particular, the procedure adapts to the het-

eroscedasticity of the noise, a property of V-fold techniques also putted on emphasis in [7] in the context of change-point detection. The idea is that V-fold penalization gives an unbiased estimate of the risk by adding to the empirical risk a cross-validated estimate of the ideal penalty. However, in practice V-fold penalization and cross-validation roughly give the same accuracy, since the over-penalization performed by cross-validation can actually be an advantage when the sample size is small to moderate. Concerning the choice of V in either VFCV or penalization, Arlot and Lerasle [8] recently justified in a least-squares density estimation context that the choice of $V = 5$ or 10 is a reasonable choice.

The theoretical investigation of optimality of either the slope heuristics or V-fold strategies will be based, among other things, on sharp results that describe the *concentration* of the true and the empirical excess risks when the model is fixed—but with dimension allowed to depend on the sample size. Since the excess risk of an empirical risk minimizer is a central object of the theory of statistical learning, such concentration result and subsequent optimal upper and lower bounds for the excess risk of least-squares estimators are of independent interest. Moreover, excess risk's concentration around a single deterministic point is an exciting new direction of research that refines more classical excess risk bounds. It recently gained interest after the work of Chatterjee [18], proving concentration inequalities for excess risk in least-squares regression under convex constraint and deducing universal admissibility of least-squares estimation in this context.

It is worth noting that one of the main arguments developed in [18] and leading to excess risk's concentration is a formula expressing the excess risk as the maximizer of a functional related to local suprema of a Gaussian process. In fact, such a *representation* of the excess risk of a general M-estimator in terms of an empirical process appeared earlier in Saumard [36]—see Remark 1 of Section 3 therein—and was also used to prove concentration inequalities for the excess risk of a projection estimator in least-squares regression. Building on [18], Muro and van de Geer [33] recently proved concentration inequalities for the excess risk in regularized least-squares regression and van de Geer and Wainwright [39] proposed a generic framework of regularized M-estimation allowing to derive excess risk's concentration. These studies are also both based on excess risk's representation in terms of either a Gaussian or an empirical process.

Let us now detail our contributions:

- We propose a new analytical property, allowing to deal with a lot of functional bases, that we call *strongly localized basis*. We show that it is a refinement on the classical concept of localized basis [11], that encompasses the cases of histograms, piecewise polynomials and compactly supported wavelets. We prove better results for strongly localized bases than for localized bases, while all known examples of localized bases are in fact strongly localized. Therefore, the concept of strongly localized basis is a way to describe some functional bases that is of independent interest and that could be used in many other nonparametric settings.

- We substantially extend the theoretical analysis of the slope heuristics, generalizing the results of [9, 37] to the case of strongly localized bases.

- We prove sharp oracle inequalities for the V-fold cross validation with fixed V, showing that it asymptotically recovers an oracle model learned with a fraction equal to $1 - V^{-1}$ of the original amount of data. Then we improve on these bounds by considering V-fold penalization, which satisfies optimal oracle inequalities. By proving such a result, we generalize a previous study of Arlot [3], from the case of histograms to the case of strongly localized bases.

- We prove concentration bounds for the excess risk of projection estimators, that are of independent interest. These results are based on previous work [36] and on a new approach to sup-norm consistency. We indeed generalize previous representation formulas in terms of empirical process for the excess risk of a (regularized) M-estimator obtained in [36, 39] to *any functional* of a M-estimator and use it to obtain bounds in sup-norm for projection estimators on strongly localized bases. These new representation formulas are also of independent interest, since they are totally general in M-estimation.

- We show in our experiments the good behavior of the slope heuristics for the selection of linear wavelet models. Indeed, it often compares favorably to VFCV and penalization. In addition, Mallows' C_p seems to be also efficient. We also recover in our more general framework some previous observations of Arlot [3]: even if the V-fold penalization has better theoretical guarantees than the V-fold cross validation, it has only comparable efficiency in practice.

The paper is organized as follows. In Section 1, we describe the statistical framework. The concept of strongly localized basis is presented in Section 2. The slope heuristics is validated in Section 3, and V-fold strategies are considered in Section 4. Then we expose our results for a fixed model, that are of independent interest, in Section 5. Numerical experiments are detailed in Section 6. The proofs are postponed to Section 7.

1 Statistical framework

We consider n independent observations $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ with common distribution P, as well as a generic random variable $\xi = (X, Y)$, independent of the sample $(\xi_1, \dots, \xi_n)$, following the same distribution P. The *feature space* $\mathcal{X}$ is a subset of $\mathbb{R}^d$, $d \geq 1$. The marginal distribution of X_i is denoted P^X. We assume that the following relation holds,

$$Y = s_* (X) + \sigma (X) \varepsilon ,$$

where $s_* \in L_2 \left(P^X\right)$ is the *regression function* of Y with respect to X to be estimated. Conditionally to X, the residual ε is normalized, i.e. it has mean zero and variance one. The function $\sigma : \mathcal{X} \to \mathbb{R}_+$ is the unknown *heteroscedastic noise level*.

To estimate s_*, we consider a finite *collection of models* $\mathcal{M}_n$, with cardinality depending on the sample size n. Each model $m \in \mathcal{M}_n$ will be a finite-dimensional vector space of *linear dimension* D_m. The models that we consider in this paper are more precisely defined in Section 2 below.

We write $\|s\|_2 = \left(\int_{\mathcal{X}} s^2 dP^X\right)^{1/2}$ the quadratic norm in $L_2 \left(P^X\right)$ and s_m the *orthogonal projection* of s_* onto m in the Hilbert space $\left(L^2 \left(P^X\right), \|\cdot\|_2\right)$. For a function $f \in L_1 (P)$, we write $P(f) = Pf = \mathbb{E}[f(\xi)]$. By setting $\gamma : L_2 \left(P^X\right) \to L_1 (P)$ the *least-squares contrast*, defined by

$$\gamma (s) : (x, y) \mapsto (y - s (x))^2 \quad , \qquad s \in L_2 \left(P^X\right) ,$$

the regression function s_* is characterized by the following relation,

$$s_* = \arg \min_{s \in L_2 (P^X)} P (\gamma (s)) .$$

The projections s_m also satisfy,

$$s_m = \arg \min_{s \in m} P (\gamma (s)) .$$

For each model $m \in \mathcal{M}_n$, we consider a *least-squares estimator* $\widehat{s}_m$ (possibly non unique), satisfying

$$\begin{aligned} \widehat{s}_m &\in \arg \min_{s \in m} \{P_n (\gamma (s))\} \\ &= \arg \min_{s \in m} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - s (X_i))^2 \right\} , \end{aligned}$$

where $P_n = n^{-1} \sum_{i=1}^n \delta_{\xi_i}$ is the *empirical measure* built from the data.

The performance of the least-squares estimators is tackled through their *excess loss*,

$$\ell (s_*, \widehat{s}_m) := P (\gamma (\widehat{s}_m) - \gamma (s_*)) = \|\widehat{s}_m - s_*\|_2^2 .$$

We split the excess risk into a sum of two terms,

$$\ell (s_*, \widehat{s}_m) = \ell (s_*, s_m) + \ell (s_m, \widehat{s}_m) ,$$

where

$$\ell (s_*, s_m) := P (\gamma (s_m) - \gamma (s_*)) = \|s_m - s_*\|_2^2 \quad \text{ and } \quad \ell (s_m, \widehat{s}_m) := P (\gamma (\widehat{s}_m) - \gamma (s_m)) \geq 0.$$

The quantity $\ell (s_*, s_m)$ is a deterministic term called the *bias* of the model m, while $\ell (s_m, \widehat{s}_m)$ is a random variable that we call the *excess risk* of the least-squares estimator $\widehat{s}_m$ *on the model* m. Notice that by the Pythagorean theorem, it holds

$$\ell (s_m, \widehat{s}_m) = \|\widehat{s}_m - s_m\|_2^2 .$$

Having at hand the collection of models $\mathcal{M}_n$, we want to construct an estimator whose excess risk is as close as possible to the excess risk of an *oracle model* m_*,

$$m_* \in \arg \min_{m \in \mathcal{M}_n} \{\ell (s_*, \widehat{s}_m)\} . \tag{1}$$

We propose to perform this task *via* a penalization procedure: given some penalty pen, that is a function from $\mathcal{M}_n$ to $\mathbb{R}^+$, we consider the following *selected model*,

$$\widehat{m} \in \arg \min_{m \in \mathcal{M}_n} \{P_n (\gamma (\widehat{s}_m)) + \text{pen} (m)\} \ . \tag{2}$$

The goal is then to find a good penalty, such that the selected model $\widehat{m}$ satisfies an *oracle inequality* of the form

$$\ell (s_*, \widehat{s}_{\widehat{m}}) \leq C \times \inf_{m \in \mathcal{M}_n} \ell (s_*, \widehat{s}_m) , \tag{3}$$

with probability close to one and with some constant $C \geq 1$, as close to one as possible.

2 Strongly localized bases

We define here the analytic constraints that we need to put on the models in order to derive our model selection results. We also provide various examples of such models.

2.1 Definition

Let us take a finite-dimensional model m with linear dimension D_m and an orthonormal basis $(\varphi_k)_{k=1}^{D_m}$. The family $(\varphi_k)_{k=1}^{D_m}$ is called a *strongly localized basis* (with respect to the probability measure P^X) if the following assumption is satisfied:

(Aslb) there exist $r_m > 0$, $b_m \in \mathbb{N}_*$, a partition $(\Pi_i)_{i=1}^{b_m}$ of $\{1,\dots,D_m\}$, positive constants $(A_i)_{i=1}^{b_m}$ and an orthonormal basis $(\varphi_k)_{k=1}^{D_m}$ of $(m, \|\cdot\|_2)$ such that $1 \le A_1 \le A_2 \le \ldots \le A_{b_m} < +\infty$,

$$\sum_{i=1}^{b_m} \sqrt{A_i} \le r_m \sqrt{D_m}, \tag{4}$$

and

$$\text{for all } i \in \{1,\dots,b_m\}, \text{ for all } k \in \Pi_i,\ \|\varphi_k\|_\infty \le r_m \sqrt{A_i}. \tag{5}$$

Moreover, for every $(i,j) \in \{1,\dots,b_m\}^2$ and $k \in \Pi_i$, we set

$$\Pi_{j|k} = \left\{ l \in \Pi_j ; \operatorname{supp}(\varphi_k) \bigcap \operatorname{supp}(\varphi_l) \neq \emptyset \right\}$$

and we assume that there exists a positive constant A_c such that for all $j \in \{1,\dots,b_m\}$,

$$\max_{k \in \Pi_i} \operatorname{Card}\left(\Pi_{j|k}\right) \le A_c \left(A_j A_i^{-1} \vee 1 \right). \tag{6}$$

Up to our knowledge, the concept of strongly localized basis is new. In (5), we ask for a control in sup-norm of each element of the considered basis. We also require in (6) a control of the number of intersections between the supports of the elements of the considered orthonormal basis.

As shown in Section 2.2 below, the property of strongly localized basis allows to unify the treatment of some models of histograms, piecewise polynomials and compactly supported wavelets. From this point of view, we may interpret the parameter b_m as the number "scales" in the basis, which in particular equals one for histograms and piecewise polynomials. It is also equal to the number of resolutions in the multi-resolution analysis associated to wavelet models. See Section 2.2 below for details about these examples.

The classical concept of localized basis (Birgé and Massart [11]) also covers the previous examples. More precisely, recall that an orthonormal basis $(\varphi_k)_{k=1}^{D_m}$ of $(m, \|\cdot\|_2)$ is a *localized basis* if there exists $r_\varphi > 0$ such that

$$\text{for all } \beta = (\beta_k)_{k=1}^{D_m} \in \mathbb{R}^{D_m},\ \left\| \sum_{k=1}^{D_m} \beta_k \varphi_k \right\|_\infty \le r_\varphi \sqrt{D_m} \max_{k \in \{1,\dots,D_m\}} |\beta_k|.$$

In fact, we show in the next proposition that strongly localized bases are localized in the classical sense. The interest of strongly localized bases over localized bases then comes from the fact that it allows to derive concentration bounds for the excess risks for models with dimension much larger than what we can prove with localized bases (from $D_m \ll n^{1/3}$ for localized bases to $D_m \ll n$ for strongly localized ones). This point is detailed in Section 5.1.

Proposition 2.1 *If an orthonormal basis $(\varphi_k)_{k=1}^{D_m}$ is strongly localized, then it is localized. More precisely, if $(\varphi_k)_{k=1}^{D_m}$ satisfies (**Aslb**), then for every $\beta = (\beta_k)_{k=1}^{D_m} \in \mathbb{R}^{D_m}$,*

$$\begin{aligned} \left\| \sum_{k=1}^{D_m} \beta_k \varphi_k \right\|_\infty &\le A_c r_m \sum_{i=1}^{b_m} \sqrt{A_i} \max_{l \in \Pi_i} |\beta_l| \\ &\le A_c r_m^2 \sqrt{D_m} \max_{k \in \{1,\dots,D_m\}} |\beta_k|. \end{aligned}$$

Reciprocally, if $(\varphi_k)_{k=1}^{D_m}$ is a localized basis as in (2.1), then it achieves (4) and (5) above with $b = 1$, $A_1 = D_m$ and $r_m = \max\{r_\varphi, 1\}$.

Proposition 2.1 shows that the parameter r_m appearing in the definition of a strongly localized basis is closely related to the parameter r_φ defining a localized basis.
The proof of Proposition 2.1 can be found in Section 7.1.

2.2 Examples

We investigate here the scope of the concept of strongly localized basis by providing some examples of linear models achieving this condition.

2.2.1 Histograms and piecewise polynomials

It is proved in [36] that linear models of histograms and more general piecewise polynomials with bounded degree are localized bases in $L_2\left(P^X\right)$ if the underlying partition $\mathcal{P}$ of $\mathcal{X}$ is lower-regular in the sense that there exists a constant c_m such that

$$0 < c_m < \sqrt{|\mathcal{P}| \inf_{I\in\mathcal{P}} P^X(I)}\,.$$

More precisely, if $r \in \mathbb{N}$ is the maximal degree of the piecewise polynomials—$r = 0$ in the case of histograms—then any orthonormal basis $\{\varphi_{I,j},\ I \in \mathcal{P},\ j \in \{0,\dots,r\}\}$ of $(m, \|\cdot\|_2)$ such that for all $j \in \{0,\dots,r\}$, $\varphi_{I,j}$ is supported by the element I of $\mathcal{P}$, is localized. Hence, by Proposition 2.1, it achieves Inequalities (4) and (5) of the definition of strongly localized basis with $b = 1$ and $A_1 = D_m = (r+1)\mathrm{Card}(\mathcal{P})$. It is also immediately seen that such basis achieves in this case (6) with $A_c = r + 1$. Furthermore, $r_m = c_m^{-1}$ for histograms (Lemma 4, [36]) and it has a more complicated expression for piecewise polynomials (Lemma 7, [36]). As a result, models of histograms and piecewise polynomials with bounded degree and underlying lower-regular partition $\mathcal{P}$ of $\mathcal{X}$ are endowed with a strongly localized structure.

2.2.2 Compactly supported wavelet expansions

We assume here that $\mathcal{X} = [0,1]$ and take $b_m \in \mathbb{N}^\star$. For details about wavelets and interactions with Statistics, we refer to [25]. Set ϕ_0 the father wavelet and ψ_0 the mother wavelet. For every integers $j \geq 0$, $1 \leq k \leq 2^j$, define

$$\psi_{j,k} : x \mapsto 2^{j/2}\psi_0\left(2^j x - k + 1\right).$$

As explained in [19], there are many ways to consider wavelets on the interval. We will consider here one of the most classical solution, that consists of using "periodized" wavelets. To this aim, we associate to a function ψ on $\mathbb{R}$, the 1-periodic function

$$\psi^{\mathrm{per}}(x) = \sum_{p\in\mathbb{Z}} \psi(x+p)\,.$$

Notice that if ψ has a compact support, then the sum at the right-hand side of the latter inequality is finite for any x.

We set for every integers $i, j, l \geq 0$, satisfying $i \leq j$ and $1 \leq l \leq 2^i$,

$$\begin{aligned}\Lambda(j) &= \left\{(j,k)\ ;\ 1 \leq k \leq 2^j\right\},\\ \Lambda(j,i,l) &= \left\{(j,k)\ ;\ 2^{j-i}(l-1)+1 \leq k \leq 2^{j-i}l\right\}.\end{aligned}$$

Moreover, we set $\psi_{-1,k}(x) = \phi_0(x-k+1)$,

$$\Lambda(-1) = \{(-1,k)\,;\mathrm{supp}\,(\psi_{-1,k}) \cap [0,1] \neq \emptyset\} \quad \text{and} \quad \Lambda_{b_m} = \bigcup_{j=-1}^{b_m} \Lambda(j)\ .$$

Notice that for every integers $i, j \geq 0$ such that $i \leq j$, $\left\{\Lambda(j,i,l)\ ;\ 1 \leq l \leq 2^i\right\}$ is a partition of $\Lambda(j)$, which means that

$$\Lambda(j) = \bigcup_{l=1}^{2^i} \Lambda(j,i,l) \text{ and for all } 1 \leq l,h \leq 2^i,\ \Lambda(j,i,l) \bigcap \Lambda(j,i,h) = \emptyset.$$

We consider the model

$$m = \mathrm{Span}\left\{\psi_\lambda^{\mathrm{per}}\ ;\ \lambda \in \Lambda_{b_m}\right\}. \tag{7}$$

Notice that the linear dimension D_m of m satisfies $D_m = 2^{b_m+1}$.

Proposition 2.2 *With the notations above, if ϕ_0 and ψ_0 are compactly supported, then $\{\psi_\lambda^{per}\ ;\ \lambda \in \Lambda_{b_m}\}$ is a strongly localized basis on $([0,1], \mathrm{Leb})$, with parameters b_m as defined above, $A_j = 2^j$ for $j \geq 0$ and $A_{-1} = 1$ (an explicit value of r_m is also given in the proof, but is more complicated).*

The proof of Proposition 2.2 can be found in Section 7.1. Proposition 2.2 proves that periodized compactly supported wavelets on the unit interval form a localized basis for the Lebesgue measure.

Considering the Haar basis, we can avoid the use of periodization and consider more general measures than the Lebesgue one.

Proposition 2.3 *Let us take* $\phi_0 = \mathbf{1}_{[0,1]}$ *and* $\psi_0 = \mathbf{1}_{[0,1/2]} - \mathbf{1}_{(1/2,1]}$ *and consider the model* m *given in (7). Set for every integers* $j \geq 0$, $1 \leq k \leq 2^j$,

$$p_{j,k,-} = P^X\left(\left[2^{-j}(k-1), 2^{-j}\left(k-\frac{1}{2}\right)\right]\right) \ , \ p_{j,k,+} = P^X\left(\left(2^{-j}\left(k-\frac{1}{2}\right), 2^{-j}k\right]\right)$$

$$\psi_{j,k} : x \in [0,1] \mapsto \frac{1}{\sqrt{p_{j,k,+}^2 p_{j,k,-} + p_{j,k,-}^2 p_{j,k,+}}}\left(p_{j,k,+}\mathbf{1}_{[2^{-j}(k-1),2^{-j}(k-1/2)]} - p_{j,k,-}\mathbf{1}_{\left(2^{-j}\left(k-\frac{1}{2}\right),2^{-j}k\right]}\right).$$

Moreover we set $\psi_{-1} = \phi_0$. *Assume that* P^X *has a density* f *with respect to* Leb *on* $[0,1]$ *and that there exists* $c_{\min} > 0$ *such that for all* $x \in [0,1]$,

$$f(x) \geq c_{\min} > 0.$$

Then $\{\psi_\lambda \ ; \ \lambda \in \Lambda_{b_m}\}$ *is a strongly localized orthonormal basis of* $(m, \|\cdot\|_2)$. *Indeed, by setting* $A_{-1} = 1$ *and* $A_j = 2^j$, $j \geq 0$, *we have for every integers* $j \geq 0$, $1 \leq k \leq 2^j$,

$$\|\psi_{j,k}\|_\infty \leq \sqrt{\frac{2}{c_{\min}} A_j}$$

and

$$\sum_{j=-1}^{b_m} \sqrt{A_j} \leq \left(\sqrt{2}+1\right)\sqrt{D_m} \ .$$

Finally, if $\Lambda_{j|\mu} = \{\lambda \in \Lambda_j ; \mathrm{supp}(\varphi_\mu) \bigcap \mathrm{supp}(\varphi_\lambda) \neq \emptyset\}$ *for* $\mu \in \Lambda_{b_m}$ *and* $j \in \{-1, 0, 1, ..., b_m\}$,

$$\max_{\mu \in \Lambda_i} \mathrm{Card}\left(\Lambda_{j|\mu}\right) \leq A_j A_i^{-1} \vee 1.$$

Proposition 2.3, which proof is straightforward and left to the reader, ensures that if P^X has a density which is uniformly bounded away from zero on $\mathcal{X}$, then the Haar basis is a strongly localized orthonormal basis for the $L_2\left(P^X\right)$-norm. More precisely, with notations of (**Aslb**), $r_m = \max\left\{\sqrt{2}+1, \sqrt{2c_{\min}^{-1}}\right\}$ and $A_c = 1$ are convenient.

3 The slope heuristics

3.1 Principles

The slope heuristics is a conjunction of general facts about penalization techniques in model selection, that lead in practice to an efficient penalty calibration procedure. Let us briefly recall the main ideas underlying the slope heuristics.

Consider the model selection problem described in (2). First, there exists a *minimal penalty*, denoted $\mathrm{pen}_{\min}$, such that if $\mathrm{pen}(m_1) < \mathrm{pen}_{\min}(m_1)$ where m_1 is one of the largest models in $\mathcal{M}_n$, then the procedure defined in (2) totally misbehaves in the sense that the dimension of the selected model is one of the largest of the collection, $D_{\widehat{m}} \gtrsim D_{m_1}$, and the excess risk of the selected model explodes compared to the excess risk of the oracle.

Furthermore, if $\mathrm{pen} > \mathrm{pen}_{\min}$ uniformly over the collection of models, then the selected model is of reasonable dimension and achieves an oracle inequality as in (3).

Arlot and Massart [9] conjectured the validity in a large M-estimation context of the following candidate for the minimal penalty,

$$\mathrm{pen}_{\min}(m) = \mathbb{E}\left[\ell_{\mathrm{emp}}\left(\widehat{s}_m, s_m\right)\right], \tag{8}$$

where $\ell_{\mathrm{emp}}\left(\widehat{s}_m, s_m\right)$ is the empirical excess risk on the model $m \in \mathcal{M}_n$, defined to be

$$\ell_{\mathrm{emp}}\left(\widehat{s}_m, s_m\right) = P_n\left(\gamma\left(s_m\right) - \gamma\left(\widehat{s}_m\right)\right) \geq 0. \tag{9}$$

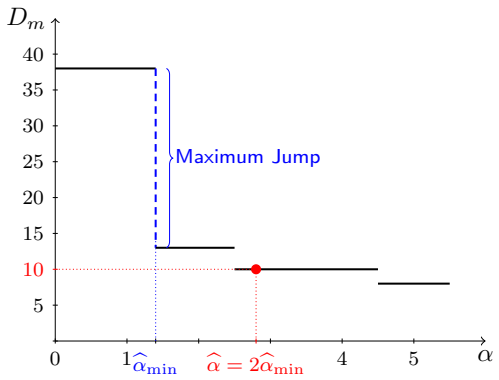


Figure 1: If $\text{pen}_{\min} = \alpha_{\min} \cdot \text{pen}_{\text{shape}}$, for a known penalty shape $\text{pen}_{\text{shape}}$, then one can estimate $\alpha_{\min}$ by using the dimension jump. This gives $\widehat{\alpha}_{\min}$ and $\text{pen}_{\text{opt}} = 2\widehat{\alpha}_{\min} \cdot \text{pen}_{\text{shape}}$ is then an optimal penalty according to the slope heuristics.

Finally, if the penalty satisfies $\text{pen} = 2 \times \text{pen}_{\min}$ then it is optimal in the sense that the excess risk of the selected model converges to the excess risk of the oracle when the amount of data tends to infinity,

$$\frac{\ell\left(s_*, \widehat{s}_{\widehat{m}}\right)}{\inf_{m \in \mathcal{M}_n} \left\{\ell\left(s_*, \widehat{s}_m\right)\right\}} \longrightarrow_{n \to +\infty} 1.$$

From the previous facts, two algorithms have been built in order to optimally calibrate a penalty shape. Both are based on the estimation of the minimal penalty. One takes advantage of the dimension jump of the selected model occurring around the minimal penalty (see Figure 1) and the other is based on formula (8), performing a robust regression of the empirical risk with respect to the penalty shape. We refer to the survey paper [10] for further details about the algorithmic and theoretical works existing on the slope heuristics.

3.2 Assumptions and comments

Set of assumptions : (SA)

(P1) Polynomial complexity of $\mathcal{M}_n$: there exist some constants $c_{\mathcal{M}}$, $\alpha_{\mathcal{M}} > 0$ such that $\text{Card}\left(\mathcal{M}_n\right) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$.

(Auslb) Existence of strongly localized bases: there exist $r_{\mathcal{M}}, A_c > 0$ such that for every $m \in \mathcal{M}_n$, there exist $b_m \in \mathbb{N}_*$, a partition $(\Pi_i)_{i=1}^{b_m}$ of $\{1, \ldots, D_m\}$, positive constants $(A_i)_{i=1}^{b_m}$ and an orthonormal basis $(\varphi_k)_{k=1}^{D_m}$ of $(m, \|\cdot\|_2)$ such that $0 < A_1 \leq A_2 \leq \ldots \leq A_{b_m} < +\infty$,

$$\sum_{i=1}^{b_m} \sqrt{A_i} \leq r_{\mathcal{M}} \sqrt{D_m},$$

and

$$\text{for all } i \in \{1, \ldots, b_m\}, \text{ for all } k \in \Pi_i, \ \|\varphi_k\|_\infty \leq r_{\mathcal{M}} \sqrt{A_i}.$$

Moreover, for every $(i, j) \in \{1, \ldots, b_m\}^2$ and $k \in \Pi_i$, we set

$$\Pi_{j|k} = \left\{ l \in \Pi_j \ ; \text{supp}\left(\varphi_k\right) \bigcap \text{supp}\left(\varphi_l\right) \neq \emptyset \right\}$$

and we assume that for all $j \in \{1, \ldots, b_m\}$,

$$\max_{k \in \Pi_i} \text{Card}\left(\Pi_{j|k}\right) \leq A_c \left(A_j A_i^{-1} \vee 1\right).$$

(P2) Upper bound on dimensions of models in $\mathcal{M}_n$: there exists a positive constant $A_{\mathcal{M},+}$ such that for every $m \in \mathcal{M}_n$, $1 \leq D_m \leq \max\left\{D_m, b_m^2 A_{b_m}\right\} \leq A_{\mathcal{M},+} n \left(\ln n\right)^{-2}$.

(P3) Richness of $\mathcal{M}_n$: there exist $m_0, m_1 \in \mathcal{M}_n$ and some constants c_{rich}, $A_{\text{rich}} > 0$ such that $D_{m_0} \in [\sqrt{n}, c_{\text{rich}} \sqrt{n}]$ and $D_{m_1} \geq A_{\text{rich}} n \left(\ln n\right)^{-2}$.

(Ab) A positive constant A exists, that bounds the data and the projections s_m of the target s_* over the models m of the collection $\mathcal{M}_n$: $|Y_i| \leq A < \infty$, $\|s_m\|_\infty \leq A < \infty$ for all $m \in \mathcal{M}_n$.

(An) Uniform lower-bound on the noise level: $\sigma(X_i) \geq \sigma_{\min} > 0$ *a.s.*

($\mathbf{Ap}_u$) The bias decreases as a power of D_m: there exist $\beta_+ > 0$ and $C_+ > 0$ such that

$$\ell(s_*, s_m) \leq C_+ D_m^{-\beta_+} .$$

The set of Assumptions (**SA**) is very similar—and actually extends—the set of assumptions used in [9] and [37] to prove the validity of the slope heuristics in heteroscedastic least-squares regression, respectively for models of histograms and piecewise polynomials.

The main features in this set of Assumptions (**SA**) are as follows. Assumption (**P1**) amounts to say that we select a model among a "small" collection, as opposed to large collection of models whose cardinal is exponential with respect to the amount of data n. Roughly speaking, this assumption allows to neglect the deviations of the excess risks on each model around their mean, since concentration inequalities shown in Section 5 below for these quantities are exponential.

Then Assumptions (**Auslb**), (**P2**), (**Ab**) and (**An**) enable to apply the desired concentration inequalities for the excess risks established in Section 5. As shown in Section 2.2.2, Assumption (**Auslb**) allows in particular to encompass the case of compactly supported wavelet expansions on the interval.

For further and more detailed comments on the above assumptions, we refer to [9] and [37].

3.3 Statement of the theorems

Let us now state our results validating the slope heuristics for the selection of uniformly strongly localized bases. The first theorem exhibits the empirical excess risk defined in (9) as a (majorant of the) minimal penalty, as conjectured by Arlot and Massart [9].

Theorem 3.1 *Take a positive penalty: for all $m \in \mathcal{M}_n$, $\mathrm{pen}(m) \geq 0$. Suppose that the assumptions (**SA**) of Section 3.2 hold, and furthermore suppose that for $A_{\mathrm{pen}} \in [0,1)$ and $A_p > 0$ the model m_1 of assumption (**P3**) satisfies*

$$0 \leq \mathrm{pen}(m_1) \leq A_{\mathrm{pen}} \mathbb{E}\left[\ell_{\mathrm{emp}}\left(\widehat{s}_{m_1}, s_{m_1}\right)\right] ,$$

*with probability at least $1 - A_p n^{-2}$. Then there exist a constant $L_1 > 0$ only depending on constants in (**SA**), as well as an integer n_0 and a positive constant L_2 only depending on A_{pen} and on constants in (**SA**) such that, for all $n \geq n_0$, it holds with probability at least $1 - L_1 n^{-2}$,*

$$D_{\widehat{m}} \geq L_2 n \ln(n)^{-2}$$

and

$$\ell(s_*, \widehat{s}_{\widehat{m}}) \geq \frac{n^{\beta_+/(1+\beta_+)}}{(\ln n)^3} \inf_{m \in \mathcal{M}_n} \left\{\ell(s_*, \widehat{s}_m)\right\} ,$$

*where $\beta_+ > 0$ is defined in assumption ($\mathbf{Ap}_u$) of (**SA**).*

In order to theoretically validate the slope heuristics described in Section 3.1 above, it remains, in addition to Theorem 3.1, to show that taking a penalty greater than the empirical excess risk ensures an oracle inequality and that taking two times the empirical excess risk yields asymptotic optimality of the procedure. That's what we present now.

Theorem 3.2 *Suppose that the assumptions (**SA**) of Section 3.2 hold, and furthermore suppose that for some $\delta \in [0,1)$ and $A_p, A_r > 0$, there exists an event of probability at least $1 - A_p n^{-2}$ on which, for every model $m \in \mathcal{M}_n$ such that $D_m \geq A_{\mathcal{M},+} (\ln n)^3$, it holds*

$$\left|\mathrm{pen}(m) - 2\mathbb{E}\left[\ell_{\mathrm{emp}}\left(\widehat{s}_m, s_m\right)\right]\right| \leq \delta\left(\ell(s_*, s_m) + \mathbb{E}\left[\ell_{\mathrm{emp}}\left(\widehat{s}_m, s_m\right)\right]\right)$$

together with

$$\left|\mathrm{pen}(m)\right| \leq A_r \left(\frac{\ell(s_*, s_m)}{(\ln n)^2} + \frac{(\ln n)^3}{n}\right) . \tag{10}$$

*Then, for any $\eta \in (0, \beta_+/(1+\beta_+))$, there exist an integer n_0 only depending on η, δ and β_+ and on constants in (**SA**), a positive constant L_3 only depending on $c_{\mathcal{M}}$ given in (**SA**) and on A_p, two positive constants L_4 and L_5 only depending on constants in (**SA**) and on A_r and a sequence*

$$\theta_n \leq \frac{L_4}{(\ln n)^{1/4}}$$

such that it holds for all $n \geq n_0$, with probability at least $1 - L_3 n^{-2}$,

$$D_{\widehat{m}} \leq n^{\eta + 1/(1+\beta_+)}$$

and

$$\ell\left(s_*, \widehat{s}_{\widehat{m}}\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{(1-\delta)^2}\right) \inf_{m \in \mathcal{M}_n} \left\{\ell\left(s_*, \widehat{s}_m\right)\right\} + L_5 \frac{(\ln n)^3}{n}.$$

Assume that in addition, the following assumption holds,

(Ap) *The bias decreases like a power of D_m: there exist $\beta_- \geq \beta_+ > 0$ and $C_+, C_- > 0$ such that*

$$C_- D_m^{-\beta_-} \leq \ell\left(s_*, s_m\right) \leq C_+ D_m^{-\beta_+} \ .$$

Then it holds for all $n \geq n_0\left((\boldsymbol{SA}), C_-, \beta_-, \beta_+, \eta, \delta\right)$, with probability at least $1 - L_3 n^{-2}$,

$$A_{\mathcal{M},+} (\ln n)^3 \leq D_{\widehat{m}} \leq n^{\eta + 1/(1+\beta_+)}$$

and

$$\ell\left(s_*, \widehat{s}_{\widehat{m}}\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{(1-\delta)^2}\right) \inf_{m \in \mathcal{M}_n} \left\{\ell\left(s_*, \widehat{s}_m\right)\right\}. \tag{11}$$

Notice that taking $\delta = 0$ in Theorem 3.2 gives an oracle inequality (11) with leading constant equal to $1 + 5\theta_n$ and thus converging to one when the amount of data tends to infinity. This shows the optimality of the penalty equal to two times the minimal one, thus validating the slope heuristics for the selection of models endowed with a strongly localized basis structure.
The proofs of Theorems 3.1 and 3.2 simply derive from [37] and Theorem 5.1 above (see Section 7.2 for more details).

4 V-fold model selection

We need some further notations and we follow here the notations of Arlot [3]. In order to highlight the dependence in the training set, we will denote $\widehat{s}_m(P_n)$ for the least-squares estimator learned from the empirical distribution $P_n = 1/n \sum_{i=1}^n \delta_{(X_i, Y_i)}$. In V-fold sampling, we choose some partition $(B_j)_{1 \leq j \leq V}$ of the index set $\{1, \ldots, n\}$ and define

$$P_n^{(j)} = \frac{1}{\mathrm{Card}\,(B_j)} \sum_{i \in B_j} \delta_{(X_i, Y_i)} \quad \text{and} \quad P_n^{(-j)} = \frac{1}{n - \mathrm{Card}\,(B_j)} \sum_{i \notin B_j} \delta_{(X_i, Y_i)}$$

together with the estimators,

$$\widehat{s}_m^{(-j)} = \widehat{s}_m\left(P_n^{(-j)}\right).$$

4.1 Classical V-fold cross-validation

In the VFCV procedure, the selected model $\widehat{m}_{\mathrm{VFCV}}$ optimizes the classical V-fold criterion,

$$\widehat{m}_{\mathrm{VFCV}} \in \arg\min_{m \in \mathcal{M}_n} \left\{\mathrm{crit}_{\mathrm{VFCV}}(m)\right\}, \tag{12}$$

where

$$\mathrm{crit}_{\mathrm{VFCV}}(m) = \frac{1}{V} \sum_{j=1}^{V} P_n^{(j)} \gamma\left(\widehat{s}_m^{(-j)}\right). \tag{13}$$

We assume that the partition is regular in the sense that for all $j \in \{1, \ldots, V\}$, $\mathrm{Card}\,(B_j) = n/V$ and in practice we can always ensure that for all j, $|\mathrm{Card}\,(B_j) - n/V| < 1$.

Theorem 4.1 *Assume that (**SA**) holds. Let $r \in (2, +\infty)$ and $V \in \{2, \dots, n-1\}$ satisfying $1 < V \leq r$. Define the VFCV procedure as the model selection procedure given by (12). Then, there exists a constant $L_{(\boldsymbol{SA}),r} > 0$ such that for all $n \geq n_0((\boldsymbol{SA}), r)$, with probability at least $1 - L_{(\boldsymbol{SA}),r} n^{-2}$,*

$$\ell(s_*, \widehat{s}_{\widehat{m}_{\mathrm{VFCV}}}) \leq \left(1 + \frac{L_{(\boldsymbol{SA}),r}}{\sqrt{\ln n}}\right) \inf_{m \in \mathcal{M}_n} \left\{\ell\left(s_*, \widehat{s}_m^{(-1)}\right)\right\} + L_{(\boldsymbol{SA}),r} \frac{(\ln n)^3}{n}.$$

The proof of Theorem 4.1 can be found in Section 7.3.

In Theorem 4.1, we show an oracle inequality with leading constant converging to one when the amount of data tends to infinity, that compares the excess risk of the model selected *via* VFCV to the excess risk of the best estimator learned with a fraction of the amount of data equal to $1 - V^{-1}$. Thus, VFCV allows to asymptotically recover the oracle learned with a fraction of the amount of data equal to $1 - V^{-1}$. This is natural since the V-fold criterion given in (13) is an unbiased estimate of the risk of estimators learned with a fraction $1 - V^{-1}$ of the data.

Consequently, it seems from Theorem 4.1 that there is some room to improve the performances of VFCV for fixed V, since the oracle learned with all the data has better performances (smaller excess risk) than the oracle learned with only part of the initial data. Furthermore, using the concentration inequalities derived in Theorem 5.2, we roughly have, for any $m \in \mathcal{M}_n$,

$$\begin{aligned}
\mathbb{E}\left[\ell\left(s_*, \widehat{s}_m^{(-1)}\right)\right] &= \ell(s_*, s_m) + \mathbb{E}\left[\ell\left(s_m, \widehat{s}_m^{(-1)}\right)\right] \\
&\sim \ell(s_*, s_m) + \frac{1}{4} \frac{\mathcal{C}_m}{(1 - V^{-1})n} \\
&\sim \ell(s_*, s_m) + \frac{V}{V-1} \mathbb{E}\left[\ell(s_m, \widehat{s}_m)\right] \\
&\leq \frac{V}{V-1} \mathbb{E}\left[\ell(s_*, \widehat{s}_m)\right].
\end{aligned}$$

The natural idea to overcome this issue is to try to select a model using an unbiased estimate of the risk of the estimators $\widehat{s}_m$ (rather than $\widehat{s}_m^{(-1)}$ for VFCV). This is what we propose in the following section.

4.2 V-fold penalization

Let us consider the following penalization procedure, proposed by Arlot [3] and called V-fold penalization,

$$\widehat{m}_{\mathrm{penVF}} \in \arg\min_{m \in \mathcal{M}_n} \left\{\mathrm{crit}_{\mathrm{penVF}}(m)\right\},$$

where

$$\mathrm{crit}_{\mathrm{penVF}}(m) = P_n\left(\gamma\left(\widehat{s}_m\right)\right) + \mathrm{pen}_{\mathrm{VF}}(m), \tag{14}$$

with

$$\mathrm{pen}_{\mathrm{VF}}(m) = \frac{V-1}{V} \sum_{j=1}^{V} \left[P_n \gamma\left(\widehat{s}_m^{(-j)}\right) - P_n^{(-j)} \gamma\left(\widehat{s}_m^{(-j)}\right)\right]. \tag{15}$$

The idea behind V-fold penalization is to use the V-fold penalty $\mathrm{pen}_{\mathrm{VF}}$ as an unbiased estimate of the *ideal penalty* $\mathrm{pen}_{\mathrm{id}}$, the latter allowing to recover exactly the oracle m_* defined in 1. Indeed, we can write

$$m_* \in \arg\min_{m \in \mathcal{M}_n} \left\{P_n\left(\gamma\left(\widehat{s}_m\right)\right) + \mathrm{pen}_{\mathrm{id}}(m)\right\},$$

where

$$\mathrm{pen}_{\mathrm{id}}(m) = P\left(\gamma\left(\widehat{s}_m\right)\right) - P_n\left(\gamma\left(\widehat{s}_m\right)\right). \tag{16}$$

Comparing (15) and (16), it is now clear that the V-fold penalty is a *resampling estimate* of the ideal penalty where for each $j \in \{1, \dots, V\}$ the role of P is played by P_n and the role of P_n is played by $P_n^{(-j)}$.

Now, the benefit compared to VFCV is that V-fold penalization is asymptotically optimal, as stated in the following theorem.

Theorem 4.2 *Assume that (**SA**) holds. Let $r \in (2, +\infty)$ and $V \in \{2, \dots, n-1\}$ satisfying $1 < V \leq r$. Define the V-fold penalization procedure as the model selection procedure given in (14). Then, there exists a constant $L_{(\boldsymbol{SA}),r} > 0$ such that for all $n \geq n_0((\boldsymbol{SA}), r)$, with probability at least $1 - L_{(\boldsymbol{SA}),r} n^{-2}$,*

$$\ell\left(s_*, \widehat{s}_{\widehat{m}_{\mathrm{penVF}}}\right) \leq \left(1 + \frac{L_{(\boldsymbol{SA}),r}}{\sqrt{\ln n}}\right) \inf_{m \in \mathcal{M}_n} \left\{\ell\left(s_*, \widehat{s}_m\right)\right\} + L_{(\boldsymbol{SA}),r} \frac{(\ln n)^3}{n}.$$

The proof of Theorem 4.2 can be found in Section 7.3.2.
Theorem 4.2 exhibits an oracle inequality with leading constant converging to one, comparing the risk of the model selected by V-fold penalization to the risk of an oracle model. This shows asymptotic optimality of the procedure and extends to the case of the selection of linear models endowed with a strongly localized basis structure, previous optimality results obtained by Arlot [3] for the selection of histograms, also in heteroscedatic regression with random design.

5 Excess risks' concentration

We formulate in this section optimal upper and lower bounds that describe the concentration of the excess risks for a fixed parametric model, but with dimension depending on the sample size. In the case of the existence of a strongly localized basis, we prove optimal bounds for models with dimension roughly smaller than n (up to logarithmic factors).

The proofs, which involve sophisticated arguments from empirical process theory, are partly based on earlier work by Saumard [36]. Furthermore, we use some representation formulas for functionals of M-estimators, which generalize previous excess risks representations exposed by Saumard [36], Chatterjee [18], Muro and van de Geer [33] and van de Geer and Wainwright [39]. We give these formulas in Section 5.2.

5.1 Strongly localized bases case

The following result of consistency in sup-norm for the least-squares estimator is a preliminary result that will be needed in the proof of our optimal concentration bounds.

Theorem 5.1 *Let $\alpha > 0$. Assume that m is a linear vector space of finite dimension D_m satisfying (**Aslb**) and use notations of (**Aslb**). Assume moreover that the following assumption holds:*

(Ab(m)) *A positive constant A exists, that bounds the data and the projection s_m of the target s_* on the model m: $|Y_i| \leq A < \infty$, $\|s_m\|_\infty \leq A < \infty$.*

If there exists $A_+ > 0$ such that

$$\max\left\{D_m, b_m^2 A_{b_m}\right\} \leq A_+ \frac{n}{(\ln n)^2},$$

then there exists a positive constant $L_{A,r_m,\alpha}$ such that, for all $n \geq n_0\left(A_+, A_c, r_m, \alpha\right)$,

$$\mathbb{P}\left(\|\widehat{s}_m - s_m\|_\infty \geq L_{A,r_m,\alpha}\sqrt{\frac{D_m \ln n}{n}}\right) \leq n^{-\alpha}. \tag{17}$$

Theorem 5.1 extends to the case of strongly localized bases previous results obtained in [36] for the consistency in sup-norm of least-squares estimators on linear models of histograms and piecewise polynomials. Note that minimax rates of convergence in sup-norm—and more general L_q norms, $1 \leq q \leq \infty$—for random design regression have been obtained by Stone [38].

Theorem 5.1 is based on new formulas for functionals of M-estimators that are described in Section 5.2 below.

Remark 5.1 *The main results of our paper are proved for models endowed with a strongly localized basis. In fact, we can also prove some results for the slightly weaker and more classical assumption of localized basis, defined in (2.1). The main difference is that with models having a strongly localized basis we can describe the optimality of model selection procedures for the selection of models with dimension up to $n/\left(\ln n\right)^2$, whereas for the localized basis case, we describe optimal results for models with dimension smaller than $n^{1/3}/\left(\ln n\right)^2$. This is an issue for instance in the slope heuristics, where the two algorithms of detection of the minimal penalty are based on the behavior of the largest models in the collection at hand. At a technical level, the essential gap is that for models with localized bases, we are able to prove Inequality* (17) *in Theorem 5.1 for models with linear dimension $D_m \ll n^{1/3}$ (see Remark 7.1).*

Let us now detail our concentration bounds for the excess risks. Theorem 5.2 below is a corollary of Theorem 2 of [36] and Theorem 5.1 above.

Theorem 5.2 *Let $A_+, A_-, \alpha > 0$. Assume that m is a linear vector space of finite dimension D_m satisfying (**Aslb**) and use notations of (**Aslb**). Assume moreover that Assumption (**Ab**(m)) defined in Theorem 5.1 holds. If we have*

$$A_- (\ln n)^2 \leq D_m \leq \max\left\{D_m, b_m^2 A_{b_m}\right\} \leq A_+ \frac{n}{(\ln n)^2} ,$$

then a positive constant L_0 exists, only depending on α, A_- and on the constants $A, \sigma_{\min}$ and r_m such that by setting

$$\varepsilon_n = L_0 \max\left\{ \left(\frac{\ln n}{D_m}\right)^{1/4} , \left(\frac{D_m \ln n}{n}\right)^{1/4} \right\} , \tag{18}$$

we have for all $n \geq n_0 (A_-, A_+, A, r_m, \sigma_{\min}, \alpha)$,

$$\mathbb{P}\left[(1-\varepsilon_n) \frac{\mathcal{C}_m}{n} \leq \ell(s_m, \widehat{s}_m) \leq (1+\varepsilon_n) \frac{\mathcal{C}_m}{n} \right] \geq 1 - 10 n^{-\alpha} , \tag{19}$$

$$\mathbb{P}\left[\left(1-\varepsilon_n^2\right) \frac{\mathcal{C}_m}{n} \leq \ell_{\text{emp}}(\widehat{s}_m, s_m) \leq \left(1+\varepsilon_n^2\right) \frac{\mathcal{C}_m}{n} \right] \geq 1 - 5 n^{-\alpha} , \tag{20}$$

where $\mathcal{C}_m = \sum_{k=1}^{D_m} \mathrm{Var}\left((Y - s_m(X)) \cdot \varphi_k(X)\right)$.

Theorem 5.2 exhibits the concentration of the excess risk and the empirical excess risk around the same value equal to $n^{-1}\mathcal{C}_m$. Furthermore, it is easy to check that the term $\mathcal{C}_m$ is of the order of the linear dimension D_m. More precisely, it satisfies

$$0 < \frac{\sigma_{\min} D_m}{2} \leq \mathcal{C}_m \leq \frac{3 A D_m}{2}.$$

See [36], Section 4.3 for the details, noticing that with the notations of [36], it holds $\mathcal{C}_m = D_m \mathcal{K}_{1,m}^2/4$. It is also worth noticing that the empirical excess risk concentrates better than the true excess risk, the rate of concentration for the empirical excess risk—given by the term ε_n^2—being the square of the concentration rate ε_n of the excess risk. This will be explained at a heuristic level in Section 5.2 using representation formulas for the excess risks in terms of empirical process.

Compared to other concentration results established in [18], [33] and [39] for the excess risk of least-squares or more general M-estimators, Inequalities (19) and (20) share the strong feature of computing the exact concentration point, which is equal to $n^{-1}\mathcal{C}_m$. On contrary, the methodology built by Chatterjee [18] and extended in [33] and [39], gives the concentration of the excess risk around a point, but says nothing on the value of this point. We explain further this important aspect in Section 5.2 below.

5.2 Representation formulas for functionals of M-estimators

In this section only, we assume that the contrast γ defining the estimator $\widehat{s}_m$ is general, so that $\widehat{s}_m$ is a general M-estimator—assumed to exist—on a model m,

$$\begin{aligned} \widehat{s}_m &\in \arg\min_{s \in m} \{P_n(\gamma(s))\} \\ &= \arg\min_{s \in m} \left\{ \frac{1}{n} \sum_{i=1}^n \gamma(s)(Z_i) \right\}, \end{aligned}$$

where $(Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ is a sample of random variables living in some general measurable space $\mathcal{Z}$.

Define $\mathcal{F}$ a nonnegative functional from m to $\mathbb{R}_+$: $\forall s \in m$, $\mathcal{F}(s) \geq 0$. Then the following *representation* of $\mathcal{F}(\widehat{s}_m)$ in terms of local extrema of the empirical process of interest holds.

Proposition 5.3 *With the notations above, let us also write m_C (resp. d_C), $C \geq 0$, the subset of the model m such that the values of the functional $\mathcal{F}$ on this subset are bounded above by (resp. equal to) C:*

$$m_C = \{s \in m \ ; \ \mathcal{F}(s) \leq C\} \ \text{ and } \ d_C = \{s \in m \ ; \ \mathcal{F}(s) = C\} .$$

Then,

$$\mathcal{F}(\widehat{s}_m) \in \arg\min_{C \geq 0} \left\{ \inf_{s \in d_C} P_n(\gamma(s)) \right\} \tag{21}$$

and

$$\mathcal{F}(\widehat{s}_m) \in \arg\min_{C \geq 0} \left\{ \inf_{s \in m_C} P_n(\gamma(s)) \right\} . \tag{22}$$

Proposition 5.3, whose proof is simple and written in Section 7.4.2, casts the problem of bounding *any* functional of a M-estimator into an empirical process question, consisting of comparing local extrema of the empirical measure P_n taken on the contrasted functions of the model. Up to our knowledge, such a result is new.

Considering the particular case of the sup-norm, formula (21) is our starting point to prove Theorem 5.1. More precisely, we use the fact that taking

$$\mathcal{F}\left(\widehat{s}_m\right) = \left\|\widehat{s}_m - s_m\right\|_\infty \ ,$$

formula (21) directly implies that for any $C \geq 0$,

$$\begin{aligned} & \mathbb{P}\left(\left\|\widehat{s}_m - s_m\right\|_\infty \geq C\right) \\ \leq \ & \mathbb{P}\left(\inf_{s \in m \setminus m_C} P_n\left(\gamma\left(s\right)\right) \leq \inf_{s \in m_C} P_n\left(\gamma\left(s\right)\right)\right). \end{aligned}$$

See Section 7 for the complete proofs.

Another interesting application of Proposition 5.3 would be to derive bounds for the L_p, $p \geq 1$, moments—or more general Orlicz norms—of a M-estimator. We postpone this question for future work.

Remark 5.4 *Nonnegativity of $\mathcal{F}$ is not essential (but suitable to our needs) and considering functionals with negative values is also possible, with straightforward adaptations of formulas of Proposition 5.3.*

Taking $\mathcal{F}$ to be the true or the empirical excess risk on m, we get the following results, refining the representation formulas previously obtained by [36]—see Remark 1 of Section 3 therein.

Proposition 5.5 *With the notations above, let also $\mathcal{G}$ be a nonnegative functional on m and $R_0 \in \mathbb{R}_+ \cup \{+\infty\}$. If the following event holds $\{\mathcal{G}\left(\widehat{s}_m\right) \leq R_0\}$ (the case $R_0 = +\infty$ corresponds to the trivial total event), then by setting*

$$\widetilde{m}_C = \left\{s \in m \ ; \ \mathcal{F}\left(s\right) \leq C \ \ \& \ \ \mathcal{G}\left(s\right) \leq R_0\right\} \ \text{ and } \ \widetilde{d}_C = \left\{s \in m \ ; \ \mathcal{F}\left(s\right) = C \ \ \& \ \ \mathcal{G}\left(s\right) \leq R_0\right\},$$

it holds

$$\ell\left(s_m, \widehat{s}_m\right) \in \arg\max_{C \geq 0} \left\{\sup_{s \in \widetilde{d}_C} \left\{\left(P_n - P\right)\left(\gamma\left(s_m\right) - \gamma\left(s\right)\right)\right\} - C\right\}, \tag{23}$$

$$\ell\left(s_m, \widehat{s}_m\right) \in \arg\max_{C \geq 0} \left\{\sup_{s \in \widetilde{m}_C} \left\{\left(P_n - P\right)\left(\gamma\left(s_m\right) - \gamma\left(s\right)\right)\right\} - C\right\}, \tag{24}$$

$$\ell_{\mathrm{emp}}\left(s_m, \widehat{s}_m\right) = \max_{C \geq 0} \left\{\sup_{s \in \widetilde{d}_C} \left\{\left(P_n - P\right)\left(\gamma\left(s_m\right) - \gamma\left(s\right)\right)\right\} - C\right\}, \tag{25}$$

and

$$\ell_{\mathrm{emp}}\left(s_m, \widehat{s}_m\right) = \max_{C \geq 0} \left\{\sup_{s \in \widetilde{m}_C} \left\{\left(P_n - P\right)\left(\gamma\left(s_m\right) - \gamma\left(s\right)\right)\right\} - C\right\} \tag{26}$$

The same type of excess risks representation as the one obtained in (24) are at the core of the approach to excess risk's concentration recently developed by Chatterjee [18], Muro and van de Geer [33] and van de Geer and Wainwright [39]. The main difference with our approach is that these authors rather use the parametrization $t = \sqrt{C}$ and take into advantage an argument of concavity with respect to t of the supremum of the empirical process on "balls" of excess risk smaller than t^2. We refer to van de Geer and Wainwright [39] for more details about this concavity argument (called "second order margin condition" by these authors). But with this concavity argument, nothing can be said *a priori* about the point around which the excess risk concentrates. To obtain optimal bounds on this point, as in Theorem 5.2 above, we rather apply a technology developed in [36] and based on the least-squares *contrast expansion* around the projection s_m of the target. We refer to Section 3 of [36] for a detailed presentation of the latter approach.

Proposition 5.5 also allows to make it transparent the fact the empirical excess risk has better concentration rates—given by the term ε_n^2 in Theorem 5.2—than the excess risk—which concentrates at the rate ε_n. Indeed, if we set

$$\Gamma_n\left(C\right) := \sup_{s \in \widetilde{d}_C} \left\{\left(P_n - P\right)\left(\gamma\left(s_m\right) - \gamma\left(s\right)\right)\right\} - C,$$

with $\left\{\mathcal{G}\left(\widehat{s}_m\right) \leq R_0\right\} = \left\{\left\|\widehat{s}_m - s_m\right\|_\infty \leq L\sqrt{\frac{D_m \ln n}{n}}\right\}$, the proof of Theorem 5.2 shows that $\Gamma_n\left(C\right)$ concentrates around the quantity $2\sqrt{n^{-1}\mathcal{C}_m C} - C$, which is parabolic around its maximum. The conclusion can now be directly read in Figure 2.

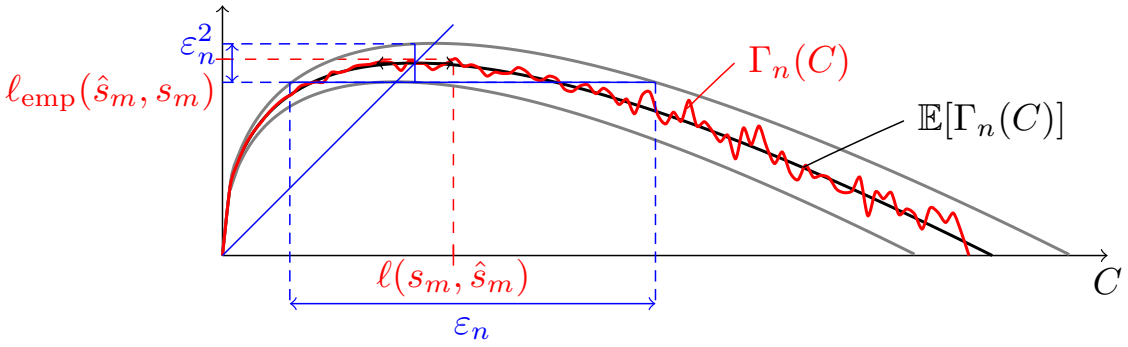


Figure 2: The true and empirical excess risks are given respectively as the maximizer and the maximum of the same function Γ_n. If Γ_n is regular around its maximum this explains why concentration rate for the empirical excess risk—given by ε_n^2—is better than for true excess risk—given by ε_n.

6 Numerical experiments

A simulation study was conducted in order to compare the numerical performances of the model selection procedures we have discussed. We consider wavelet models as non trivial illustrative examples of the theory developed above for the selection of linear estimators using the slope heuristics and V-fold model selection. However, it is a rather different question than designing the best possible estimators using wavelet expansions, since these estimators are likely to be nonlinear as for the thresholding strategies (see e.g., [1] for a comparative simulation study of wavelet based estimators). Although a linear wavelet estimator is not as flexible, or potentially as powerful, as a nonlinear one, it still preserves the computational benefits of wavelet methods. See e.g., [2] which is a key reference for linear wavelet methods in nonparametric regression. All simulations have been conducted with Matlab and the wavelet toolbox Wavelab850 [20] that is freely available from `http://statweb.stanford.edu/~wavelab/`. In order to reproduce all the experiments, the codes used to generate the numerical results presented in this paper will be available online at `https://github.com/fabnavarro`.

6.1 Computational aspects

For sample sizes $n = 256, 1024, 4096$, data were generated according to

$$Y_i = s_*(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \dots, n$$

where X_i's are uniformly distributed on $[0, 1]$, ε_i's are independent $\mathcal{N}(0, 1)$ variables and independent of X_i's. In the case of fixed design, thanks to Mallat's pyramid algorithm (see [30]), the computation of wavelet-based estimators is straightforward and fast. In the case where the function s_* is observed on a random grid, the implementation requires some extra precautions and several strategies have been proposed in the literature (see e.g. [15, 24]). In the context of random uniform design regression estimation, [16] have examined convergence rates when the unknown function is in a Hölder class. They showed that the standard equispaced wavelet method with universal thresholding can be directly applied to the nonequispaced data (without a loss in the rate of convergence). In this simulations study, we have adopted this approach, since it preserves the computational simplicity and efficiency of the equispaced algorithm. The same choice was made in the context of wavelet regression in random design with heteroscedastic dependent errors by [26]. Thus, in this case, the collection of models is computed by a simple application of Mallat's algorithm using the ordered Y_i's as input variables.

6.2 Examples

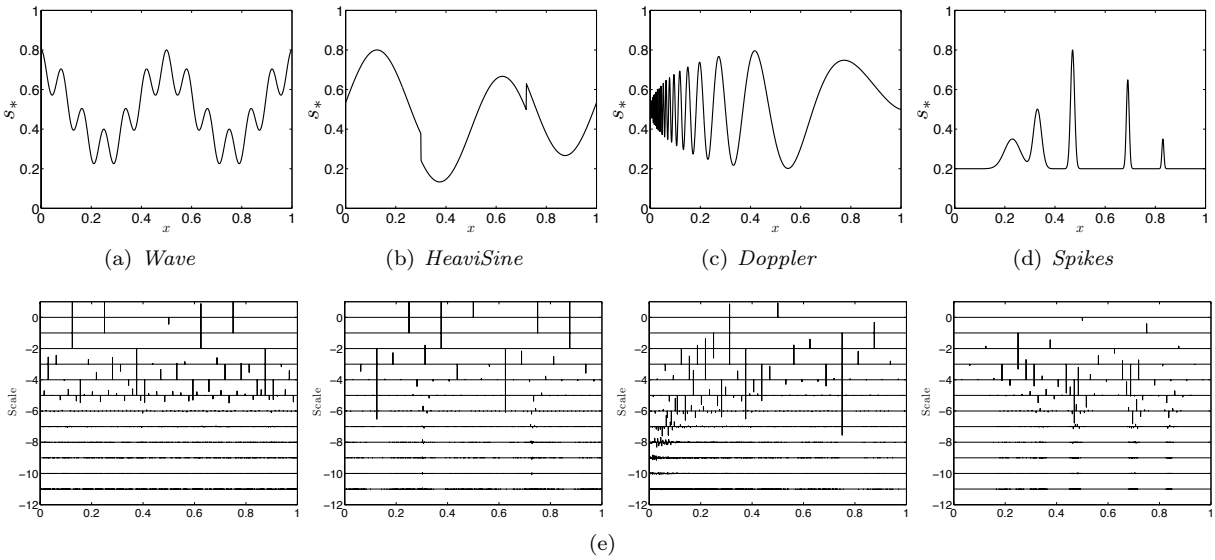
Four standard regression functions representing different level of spatial variability (*Wave, HeaviSine, Doppler* and *Spikes*, see [21, 31, 14]) and the following four $\sigma(\cdot)$ scenarios were considered:

(a) *Low Homoscedastic Noise*: $\sigma_{l1}(x) = 0.01$;

(b) *Low Heteroscedastic Noise*: $\sigma_{l2}(x) = 0.02x$;

(c) *High Homoscedastic Noise*: $\sigma_{h1}(x) = 0.05$;

(d) *High Heteroscedastic Noise*: $\sigma_{h2}(x) = 0.1x$.



(a) *Wave* (b) *HeaviSine* (c) *Doppler* (d) *Spikes*

(e)

Figure 3: (a)–(d): The four test functions used in the simulation study sampled at 4096 points. (e): Wavelet coefficients of the test functions.

The test functions are plotted in Figure 3 and a visual idea of the four noise levels is given in Figures 4(a)–(d). Several different wavelets were used. In the following, we only report in detail the results for Daubechies' compactly supported wavelet with 8 vanishing moments.

6.3 Four model selection procedures

The performance of the following four model selection methods were compared:

- The slope heuritics (SH):

$$\widehat{m}_{\mathrm{SH}} \in \arg\min_{m\in\mathcal{M}_n} \left\{\mathrm{crit}_{\mathrm{SH}}(m)\right\},$$

with

$$\mathrm{crit}_{\mathrm{SH}}(m) = P_n\left(\gamma\left(\widehat{s}_m\right)\right) + \mathrm{pen}_{\mathrm{SH}}(m),$$

and

$$\mathrm{pen}_{\mathrm{SH}}(m) = 2\frac{\widehat{\alpha}_{\min} D_m}{n}.$$

where $\widehat{\alpha}_{\min}$ is obtained from the dimension jump method (see Figure 1). Practical issues about SH are addressed in [10] and our implementation is based on the Matlab package CAPUSHE.

- Mallow's C_p (Cp):

$$\widehat{m}_{\mathrm{Cp}} \in \arg\min_{m\in\mathcal{M}_n} \left\{\mathrm{crit}_{\mathrm{Cp}}(m)\right\},$$

with

$$\mathrm{crit}_{\mathrm{Cp}}(m) = P_n\left(\gamma\left(\widehat{s}_m\right)\right) + \mathrm{pen}_{\mathrm{Cp}}(m),$$

and

$$\mathrm{pen}_{\mathrm{Cp}}(m) = 2\frac{\widehat{\sigma}^2 D_m}{n},$$

where $\widehat{\sigma}^2$ is globally estimated by the classical variance estimator defined as

$$\widehat{\sigma}^2 = \frac{d^2(Y_{1\dots n}, m_{n/2})}{n - n/2},$$

where $Y_{1\dots n} = (Y_i)_{1\le i\le n} \in \mathbb{R}^n$, $m_{n/2}$ is the largest model of dimension $n/2$, and d is the Euclidean distance on $\mathbb{R}^n$.

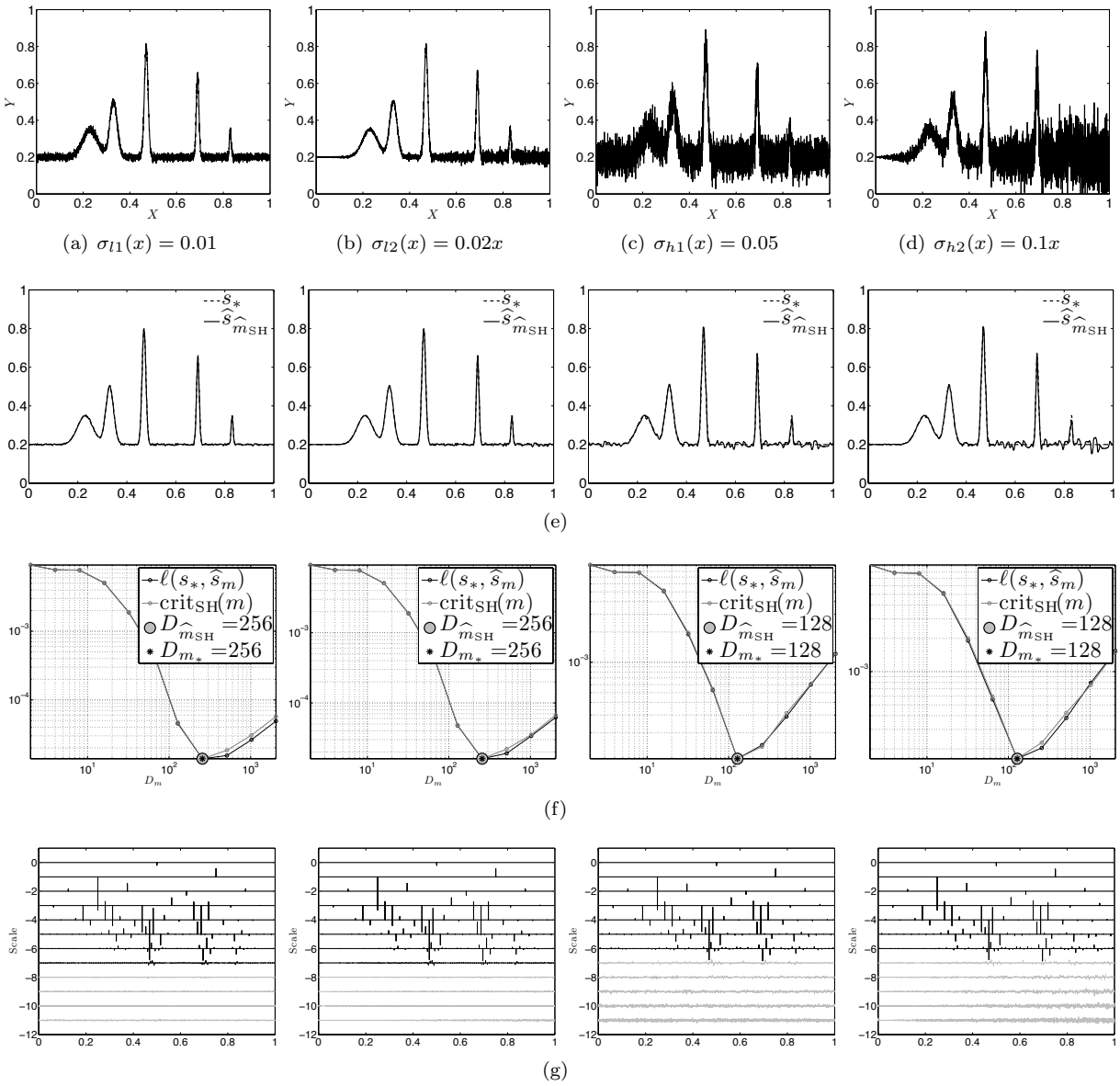


Figure 4: (a)-(d): Noisy version of *Spikes* for each $\sigma(\cdot)$ scenarios. (e): Typical reconstructions from a single simulation with $n = 4096$. The dotted line is the true signal and the solid one depicts the estimates $\widehat{s}_{\widehat{m}_{\mathrm{SH}}}$. (f): Graph of the excess risk $\ell(s_*, \widehat{s}_m)$ against the dimension D_m and (shifted) $\mathrm{crit}_{\mathrm{SH}}(m)$ (in a log-log scale). The gray circle represents the global minimizer $\widehat{m}$ of $\mathrm{crit}_{\mathrm{SH}}(m)$ and the black star the oracle model m_*. (g): Noisy and selected (black) wavelet coefficients (see Figure 3(e) for a visual comparison with the original wavelet coefficients).

- Nason's 2-fold cross-validation (2FCV). Nason adjusted the usual 2FCV method—which cannot be applied directly to wavelet estimation—for choosing the threshold parameter in wavelet shrinkage [34]. Adapting his strategy to our context, we test, for every model of the collection, an interpolated wavelet estimator learned from the (ordered) even-indexed data against the odd-indexed data and vice versa. More precisely, considering the data X_i are ordered, the selected model $\widehat{m}_{\mathrm{2FCV}}$ is obtained by minimizing (13) with $V = 2$, $B_1 = \{2, 4, \ldots, n\}$ and $B_2 = \{1, 3, \ldots, n-1\}$.

- A penalized version of Nason's 2-fold cross-validation (pen2F). As for the 2FCV, we compute $\widehat{m}_{\mathrm{pen2F}}$ by minimizing (14) with $V = 2$, $B_1 = \{2, 4, \ldots, n\}$ and $B_2 = \{1, 3, \ldots, n-1\}$.

For each method, the model collection described in Section 2.2.2 is constructed by adding successively whole resolution levels of wavelet coefficients. Thus, the considered dimensions are $\{D_m, m \in \mathcal{M}_n\} = \{2^j, j = 1, \ldots, \log_2(n)-1\}$. Note that unlike the local behaviours of the nonlinear models (e.g. thresholding), these linear

models operate in a global fashion since entire scale levels of coefficients are suppressed (see Figures 5(g),4(g) for an illustration).

Typical estimations from a single simulation with $n = 4096$ are depicted in 4(e) for the *Spikes* function. Figure 4(f) also contains a plot of the excess risk $\ell(s_*, \widehat{s}_m)$ against the dimension D_m and a vertical shift of the curve $\mathrm{crit}_{\mathrm{SH}}(m)$ is also overlayed for visualization purposes. It can be observed that $\mathrm{crit}_{\mathrm{SH}}(m)$ gives a very reliable estimate for the risk $\ell(s_*, \widehat{s}_m)$, and in turn, also a high-quality estimate of the optimal model. Indeed, for all cases, SH consistently selects the best model.

6.4 Model selection performances

We compared the procedures on $N = 1000$ independent data sets of size n ranging from 256 to 4096. As in Arlot [3], we estimate the quality of the model-selection strategies through the following constant

$$C_{\mathrm{or}} = \mathbb{E}\left[\frac{\|\widehat{s}_{\widehat{m}} - s_*\|_2^2}{\inf_{m \in \mathcal{M}_n} \|\widehat{s}_{m_*} - s_*\|_2^2}\right]$$

which represents the constant that would appear in front of an oracle inequality. This ratio, which is greater than 1, represents the accuracy of the model selection procedure. The average C_{or} over 1000 replications are given in Tables 1 and 2.

6.5 Results and discussion

It can be seen from Tables 1 and 2 that none of the methods clearly outperforms the others in all cases. However, in our experiments, Mallows' C_p seems to perform slightly better in many situations, both in the low and high noise regimes and for either homoscedastic and heteroscedastic noise. Also, the slope heuristics has roughly comparable results with Mallows' C_p, except for the small sample size case $n = 256$, where Mallows' C_p performs better, especially in the low noise regime. The quite bad behavior of the slope heuristics in the latter case (low noise, small sample size) can be explained by the fact that in such situation, the oracle model is the greatest model, that the slope heuristics tries to avoid through the use of the dimension jump.

In the low noise regime (Table 1), 2-fold penalization is slightly better than 2-fold cross-validation, especially when the sample size is small ($n = 256$). Moreover, 2-fold penalization is competitive with Mallows' C_p in the low noise regime. When the noise is high (Table 2), 2FCV and pen2F give roughly equivalent results.

Finally, it seems surprising that Mallows' C_p and the slope heuristics, that are based on linear penalties, outperform cross-validation methods in the heteroscedastic noise case. Indeed linear penalties are proved to be asymptotically suboptimal in such case, see Arlot [4], while we proved in Theorem 4.2 that V-fold penalization for a fixed V is asymptotically optimal. However, in order to be able to use Mallat's algorithm for the discrete wavelet transform, we restricted ourselves to the 2-fold and this could be the reason for the rather mild performances of the cross-validation techniques compared to Mallows' C_p. Indeed, it is well-known that in general, it is better to take $V = 5$ or 10 instead of 2 (see for instance [6]), because it reduces the variance of the cross-validation criterion. Also, Nason's cross-validation for wavelet models allows to use Mallat's algorithm, but at the price of an approximation of the original cross-validation criterion. These two aspects might be at the origin of the superiority of Mallows' C_p over the cross-validation techniques, at least in the heteroscedastic case.

s_*	$\sigma.$	n	**SH**	**Cp**	**2FCV**	**pen2F**
Wave	$l1$	256	1.980 ± 0.011	1.106 ± 0.008	1.406 ± 0.019	1.034 ± 0.005
		1024	1.051 ± 0.002	1.031 ± 0.002	1.062 ± 0.002	1.056 ± 0.004
		4096	1.021 ± 0.001	1.021 ± 0.001	1.055 ± 0.002	1.021 ± 0.001
	$l2$	256	1.799 ± 0.009	1.140 ± 0.008	1.341 ± 0.015	1.042 ± 0.005
		1024	1.021 ± 0.002	1.027 ± 0.002	1.029 ± 0.003	1.084 ± 0.006
		4096	1.033 ± 0.002	1.032 ± 0.002	1.015 ± 0.001	1.039 ± 0.002
HeaviSine	$l1$	256	1.482 ± 0.014	1.157 ± 0.005	1.437 ± 0.016	1.084 ± 0.006
		1024	1.065 ± 0.003	1.023 ± 0.002	1.155 ± 0.006	1.062 ± 0.004
		4096	1.011 ± 0.001	1.008 ± 0.001	1.101 ± 0.004	1.010 ± 0.001
	$l2$	256	1.357 ± 0.012	1.122 ± 0.005	1.357 ± 0.013	1.063 ± 0.004
		1024	1.048 ± 0.003	1.032 ± 0.002	1.133 ± 0.006	1.093 ± 0.006
		4096	1.016 ± 0.001	1.013 ± 0.001	1.064 ± 0.003	1.020 ± 0.001
Doppler	$l1$	256	2.890 ± 0.039	1.106 ± 0.008	1.852 ± 0.038	1.072 ± 0.008
		1024	2.091 ± 0.015	1.064 ± 0.006	1.486 ± 0.022	1.013 ± 0.003
		4096	1.010 ± 0.001	1.000 ± 0.000	1.141 ± 0.007	1.025 ± 0.003
	$l2$	256	2.820 ± 0.040	1.127 ± 0.009	1.784 ± 0.036	1.059 ± 0.006
		1024	1.874 ± 0.013	1.078 ± 0.006	1.419 ± 0.016	1.009 ± 0.002
		4096	1.024 ± 0.002	1.002 ± 0.000	1.187 ± 0.006	1.019 ± 0.003
Spikes	$l1$	256	3.541 ± 0.071	1.092 ± 0.007	2.075 ± 0.062	1.062 ± 0.010
		1024	1.077 ± 0.006	1.021 ± 0.002	1.198 ± 0.012	1.045 ± 0.003
		4096	1.008 ± 0.001	1.008 ± 0.001	1.029 ± 0.002	1.014 ± 0.001
	$l2$	256	3.236 ± 0.058	1.087 ± 0.007	2.008 ± 0.055	1.071 ± 0.011
		1024	1.054 ± 0.004	1.013 ± 0.001	1.187 ± 0.012	1.069 ± 0.004
		4096	1.007 ± 0.001	1.007 ± 0.001	1.009 ± 0.001	1.019 ± 0.002

Table 1: Comparison of mean performance C_{or} for each procedure over $N = 1000$ realizations of the low noise level setting with corresponding empirical standard deviation divided by $\sqrt{N}$.

s_*	$\sigma.$	n	**SH**	**Cp**	**2FCV**	**pen2F**
Wave	$h1$	256	1.029 ± 0.004	1.016 ± 0.003	1.236 ± 0.011	1.158 ± 0.009
		1024	1.003 ± 0.001	1.002 ± 0.001	1.002 ± 0.001	1.033 ± 0.005
		4096	1.011 ± 0.002	1.008 ± 0.002	1.000 ± 0.000	1.040 ± 0.004
	$h2$	256	1.076 ± 0.006	1.052 ± 0.006	1.252 ± 0.010	1.244 ± 0.012
		1024	1.022 ± 0.005	1.014 ± 0.004	1.004 ± 0.002	1.072 ± 0.008
		4096	1.020 ± 0.004	1.019 ± 0.004	1.006 ± 0.002	1.067 ± 0.007
HeaviSine	$h1$	256	1.096 ± 0.005	1.090 ± 0.005	1.115 ± 0.006	1.185 ± 0.013
		1024	1.057 ± 0.003	1.054 ± 0.003	1.123 ± 0.006	1.075 ± 0.004
		4096	1.029 ± 0.002	1.028 ± 0.002	1.081 ± 0.004	1.041 ± 0.003
	$h2$	256	1.155 ± 0.009	1.153 ± 0.011	1.125 ± 0.008	1.300 ± 0.020
		1024	1.101 ± 0.006	1.091 ± 0.006	1.133 ± 0.007	1.159 ± 0.010
		4096	1.047 ± 0.003	1.046 ± 0.003	1.122 ± 0.006	1.083 ± 0.005
Doppler	$h1$	256	1.330 ± 0.011	1.107 ± 0.005	1.347 ± 0.013	1.043 ± 0.003
		1024	1.054 ± 0.003	1.025 ± 0.002	1.108 ± 0.005	1.067 ± 0.005
		4096	1.013 ± 0.001	1.014 ± 0.001	1.029 ± 0.002	1.021 ± 0.001
	$h2$	256	1.224 ± 0.010	1.076 ± 0.004	1.291 ± 0.011	1.053 ± 0.003
		1024	1.035 ± 0.002	1.031 ± 0.002	1.079 ± 0.004	1.098 ± 0.007
		4096	1.010 ± 0.001	1.009 ± 0.001	1.022 ± 0.003	1.023 ± 0.002
Spikes	$h1$	256	1.156 ± 0.009	1.047 ± 0.003	1.282 ± 0.014	1.076 ± 0.005
		1024	1.006 ± 0.001	1.005 ± 0.001	1.094 ± 0.007	1.029 ± 0.004
		4096	1.012 ± 0.002	1.010 ± 0.001	1.009 ± 0.002	1.021 ± 0.002
	$h2$	256	1.119 ± 0.008	1.052 ± 0.004	1.284 ± 0.014	1.126 ± 0.006
		1024	1.015 ± 0.002	1.014 ± 0.002	1.137 ± 0.008	1.059 ± 0.008
		4096	1.015 ± 0.002	1.011 ± 0.002	1.014 ± 0.003	1.030 ± 0.004

Table 2: Comparison of mean performance C_{or} for each procedure over $N = 1000$ realizations of the high noise level setting with corresponding empirical standard deviation divided by $\sqrt{N}$.

7 Proofs

7.1 Proofs related to Section 2

Proof of Proposition 2.1. The proof simply follows from the following computations. For every $\beta = (\beta_k)_{k=1}^{D_m} \in \mathbb{R}^{D_m}$,

$$\begin{aligned}\left\|\sum_{k=1}^{D_m}\beta_k\varphi_k\right\|_\infty &\leq \sum_{i=1}^{b_m}\left\|\sum_{l\in\Pi_i}\beta_l\varphi_l\right\|_\infty \\ &\leq \sum_{i=1}^{b_m} A_c \max_{l\in\Pi_i}\|\varphi_l\|_\infty \times \max_{l\in\Pi_i}|\beta_l| \\ &\leq A_c r_m \sum_{i=1}^{b_m}\sqrt{A_i}\max_{l\in\Pi_i}|\beta_l| \\ &\leq A_c r_m^2\sqrt{D_m}\max_{k\in\{1,\dots,D_m\}}|\beta_k|\,.\end{aligned}$$

■

Proof of Proposition 2.2. The fact that $\{\psi_\lambda^{\rm per}\ ;\ \lambda\in\Lambda_{b_m}\}$ is an orthonormal family - and thus an orthonormal basis of m - is a classical fact of wavelet theory (see for instance [19]). Take $m>0$ such that

$$\mathrm{supp}\,(\psi_0)\bigcup\mathrm{supp}\,(\phi_0)\subset[0,m]\,.$$

For $j\geq 0$ and $1\leq k\leq 2^j$, we have

$$\left\|\psi_{j,k}^{\rm per}\right\|_\infty \leq ([m]+2)\,\|\psi_{j,k}\|_\infty \leq ([m]+2)\,2^{j/2}\,\|\psi_0\|_\infty\,,$$

where $[m]$ is the integer part of m. We thus take $A_j=2^j$ for $j\geq 0$ and $A_{-1}=1$, which gives

$$\sum_{i=-1}^{b_m}\sqrt{A_i}\leq\left(1+\sqrt{2}\right)\sqrt{D_m},$$

since $D_m=2^{b_m+1}$. By taking $r_m=\max\left\{([m]+2)\,\|\psi_0\|_\infty, 1+\sqrt{2}\right\}$, we thus get, for any $j\geq -1$ and $k\in\left\{1,\dots,2^j\right\}$,

$$\left\|\psi_{j,k}^{\rm per}\right\|_\infty\leq r_m\sqrt{A_j}\quad\text{and}\quad\sum_{i=-1}^{b_m}\sqrt{A_i}\leq r_m\sqrt{D_m}\,.$$

It remains to prove that there exists $A_c>0$ such that, by denoting for $\mu\in\Lambda_{b_m}$ and $j\in\{-1,0,1,\dots,m\}$,

$$\Lambda_{j|\mu}=\left\{\lambda\in\Lambda\,(j)\ ;\mathrm{supp}\,(\psi_\mu)\bigcap\mathrm{supp}\,(\psi_\lambda)\neq\emptyset\right\},$$

one has

$$\max_{\mu\in\Lambda(i)}\mathrm{Card}\left(\Lambda_{j|\mu}\right)\leq A_c\left(A_jA_i^{-1}\vee 1\right). \tag{27}$$

Take $j_0=\max\left\{[\log_2(m)]+1,0\right\}$. Then for all $j\geq j_0$ and $k\in\left\{1,\dots,2^{j-j_0}\right\}$, $\mathrm{supp}\,(\psi_{j,k})\subset[0,1)$. Furthermore, for every $k\in\left\{1,\dots,2^{j-j_0}\right\}$ set $\Gamma(k)=\left\{2^{j-j_0}l+k; l\in\left\{0,\dots,2^{j_0}-1\right\}\right\}$. Then $\left\{\Gamma(k)\,; k\in\left\{1,\dots,2^{j-j_0}\right\}\right\}$ form a partition of $\left\{1,\dots,2^j\right\}$ and for $k,k'\in\left\{1,\dots,2^{j-j_0}\right\}$, $k\neq k'$,

$$\mathrm{supp}\,(\psi_{j,k})\bigcap\mathrm{supp}\,(\psi_{j,k'})=\emptyset.$$

It is then easy to see that taking $A_c=2^{j_0}$ gives (27). ■

7.2 Proofs related to the slope heuristics

We first notice that, from [37], Section 5, Theorems 3.1, 3.2 are valid under the following general set of assumptions (i.e. by replacing (**SA**) by **(GSA)** in the statement of the theorems):

General set of assumptions: (GSA)

Assume (**P1**), (**P2**), (**P3**), (**Ab**), (**An**) and (**Ap**$_u$) of (**SA**). Furthermore suppose that,

(Alb) there exists a constant $r_{\mathcal{M}}$ such that for each $m \in \mathcal{M}_n$ one can find an orthonormal basis $(\varphi_k)_{k=1}^{D_m}$ satisfying, for all $(\beta_k)_{k=1}^{D_m} \in \mathbb{R}^{D_m}$,

$$\left\| \sum_{k=1}^{D_m} \beta_k \varphi_k \right\|_\infty \leq r_{\mathcal{M}} \sqrt{D_m} \, |\beta|_\infty \ ,$$

where $|\beta|_\infty = \max \{|\beta_k| \, ; k \in \{1, \ldots, D_m\}\}$.

($\mathbf{Ac}_\infty$) a positive integer n_1 exists such that, for all $n \geq n_1$, there exist a positive constant A_{cons} and an event Ω_∞ of probability at least $1 - n^{-2-\alpha_{\mathcal{M}}}$, on which for all $m \in \mathcal{M}_n$,

$$\|\widehat{s}_m - s_m\|_\infty \leq A_{cons} \sqrt{\frac{D_m \ln n}{n}} \ .$$

Now the proofs of Theorems 3.1 and 3.2 simply rely on the fact that assumptions (**Alb**) and ($\mathbf{Ac}_\infty$) in (**GSA**) are ensured under (**SA**). Indeed, assumption (**Alb**) in (**GSA**) is satisfied under assumption (**Auslb**) in the set of assumptions (**SA**), see Proposition 2.1. Furthermore, Theorem 5.1 shows that assumption ($\mathbf{Ac}_\infty$) in (**GSA**) is also satisfied under assumption (**Auslb**).

7.3 Proofs related to V-fold procedures

7.3.1 Proofs related to V-fold cross-validation

Theorem 4.1 is a straightforward consequence of the following result, that will be proved below. Recall that the set of assumptions (**GSA**) is defined in Section 7.2 above.

Theorem 7.1 *Assume that (**GSA**) holds. Let $r \in (2, +\infty)$ and $V \in \{2, \ldots, n-1\}$ satisfying $1 < V \leq r$. Define the VFCV procedure as the model selection procedure given by (12) and (13). Then, for all $n \geq n_0 ((\mathbf{GSA}), r)$, with probability at least $1 - L_{(\mathbf{GSA}),r} n^{-2}$,*

$$\ell\left(s_*, \widehat{s}_{\widehat{m}}^{(-1)}\right) \leq \left(1 + \frac{L_{(\mathbf{GSA}),r}}{\sqrt{\ln n}}\right) \inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s_*, \widehat{s}_m^{(-1)}\right) \right\} + L_{(\mathbf{GSA}),r} \frac{(\ln n)^3}{n}.$$

Proof of Theorem 7.1. All along the proof, the value of the constant $L_{(\mathbf{GSA}),r}$ may vary from line to line. We set

$$\mathrm{crit}^0_{\mathrm{VFCV}}(m) = \mathrm{crit}_{\mathrm{VFCV}}(m) - \frac{1}{V} \sum_{j=1}^{V} P_n^{(j)} (\gamma(s_*)) \, .$$

It is worth noting that the difference between $\mathrm{crit}^0_{\mathrm{VFCV}}(m)$ and $\mathrm{crit}_{\mathrm{VFCV}}(m)$ is a quantity independent of m, when m varies in $\mathcal{M}_n$. Hence, the procedure defined by $\mathrm{crit}^0_{\mathrm{VFCV}}$ gives the same result as the VFCV procedure defined by $\mathrm{crit}_{\mathrm{VFCV}}$. It will be convenient for our analysis to consider $\mathrm{crit}^0_{\mathrm{VFCV}}$ instead of $\mathrm{crit}_{\mathrm{VFCV}}$.

We get for all $m \in \mathcal{M}_n$,

$$\begin{aligned}
\mathrm{crit}^0_{\mathrm{VFCV}}(m) &= \frac{1}{V} \sum_{j=1}^{V} P_n^{(j)} \left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma(s_*) \right) \\
&= \frac{1}{V} \sum_{j=1}^{V} \Big[P_n^{(j)} \left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma(s_m) \right) \\
&\quad + \left(P_n^{(j)} - P \right) (\gamma(s_m) - \gamma(s_*)) + P(\gamma(s_m) - \gamma(s_*)) \Big] \\
&= \ell\left(s_*, \widehat{s}_m^{(-1)}\right) + \Delta_V(m) + \bar{\delta}(m) \qquad (28)
\end{aligned}$$

where

$$\Delta_V(m) = \frac{1}{V} \sum_{j=1}^{V} P_n^{(j)} \left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma(s_m) \right) - P\left(\gamma\left(\widehat{s}_m^{(-1)}\right) - \gamma(s_m) \right),$$

and $\bar{\delta}(m)$ has been defined in Lemma 7.5. Furthermore denote

$$\mathrm{p}_1^{(-1)}(m) = P\left(\gamma\left(\widehat{s}_m^{(-1)}\right) - \gamma(s_m) \right) \quad \text{and} \quad \mathrm{p}_2^{(-1)}(m) = P_n^{(-1)}\left(\gamma(s_m) - \gamma\left(\widehat{s}_m^{(-1)}\right) \right).$$

Let Ω_n be the event on which:

- For all models $m \in \mathcal{M}_n$ of dimension D_m such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_m$, it holds

$$\left| \mathrm{p}_1^{(-1)}(m) - \mathbb{E}\left[\mathrm{p}_2^{(-1)}(m)\right] \right| \leq L_{(\mathbf{GSA}),r} \varepsilon_n(m) \, \mathbb{E}\left[\mathrm{p}_2^{(-1)}(m)\right] \tag{29}$$
$$\left| \mathrm{p}_2^{(-1)}(m) - \mathbb{E}\left[\mathrm{p}_2^{(-1)}(m)\right] \right| \leq L_{(\mathbf{GSA}),r} \varepsilon_n^2(m) \, \mathbb{E}\left[\mathrm{p}_2^{(-1)}(m)\right]$$

  together with

$$|\Delta_V(m)| \leq L_{(\mathbf{GSA}),r} \varepsilon_n(m) \, \mathbb{E}\left[\mathrm{p}_2^{(-1)}(m)\right] \tag{30}$$
$$\left|\bar{\delta}(m)\right| \leq \frac{\ell(s_*, s_m)}{\sqrt{D_m}} + L_{(\mathbf{GSA}),r} \frac{\ln n}{\sqrt{D_m}} \mathbb{E}\left[\mathrm{p}_2^{(-1)}(m)\right] \tag{31}$$

- For all models $m \in \mathcal{M}_n$ of dimension D_m such that $D_m \leq A_{\mathcal{M},+} (\ln n)^3$, it holds

$$|\Delta_V(m)| \leq L_{(\mathbf{GSA}),r} \frac{(\ln n)^2}{n} \tag{32}$$
$$\left|\bar{\delta}(m)\right| \leq L_{(\mathbf{GSA}),r} \left(\sqrt{\frac{\ell(s_*, s_m) \ln n}{n}} + \frac{\ln n}{n} \right)$$
$$\mathrm{p}_2^{(-1)}(m) \leq L_{(\mathbf{GSA}),r} \frac{D_m \vee \ln n}{n} \leq L_{(\mathbf{GSA}),r} \frac{(\ln n)^3}{n}$$
$$\mathrm{p}_1^{(-1)}(m) \leq L_{(\mathbf{GSA}),r} \frac{D_m \vee \ln n}{n} \leq L_{(\mathbf{GSA}),r} \frac{(\ln n)^3}{n} \tag{33}$$

By Theorem 2 of [36] and Lemma 4 of [37] applied with $\alpha = 2 + \alpha_{\mathcal{M}}$ and sample size $n_V = n(V-1)/V$, Corollary 7.4 and Lemma 7.5 applied with $\alpha = 2 + \alpha_{\mathcal{M}}$, we get for all $n \geq n_0((\mathbf{GSA}), r)$,

$$\mathbb{P}(\Omega_n) \geq 1 - L_{(\mathbf{GSA}),r} \sum_{m \in \mathcal{M}_n} n^{-2-\alpha_{\mathcal{M}}} \geq 1 - L_{(\mathbf{GSA}),r} n^{-2} \, .$$

Control on the criterion $\mathrm{crit}^0_{\mathrm{VFCV}}$ for models of dimension not too small:

We consider models $m \in \mathcal{M}_n$ such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_m$.

$$\begin{aligned}
\mathrm{crit}^0_{\mathrm{VFCV}}(m) &= \frac{1}{V} \sum_{j=1}^{V} P_n^{(j)} \left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma(s_*) \right) \\
&= \frac{1}{V} \sum_{j=1}^{V} \Big[P_n^{(j)} \left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma(s_m) \right) \\
&\quad + \left(P_n^{(j)} - P \right) (\gamma(s_m) - \gamma(s_*)) + P(\gamma(s_m) - \gamma(s_*)) \Big] \\
&= \ell\left(s_*, \widehat{s}_m^{(-1)}\right) + \Delta_V(m) + \bar{\delta}(m)
\end{aligned}$$

By (29), (30) and (31) we have on Ω_n,

$$\begin{aligned}
\max\left\{ |\Delta_V(m)|, \left|\bar{\delta}(m)\right| \right\} &\leq L_{(\mathbf{GSA}),r} \varepsilon_n(m) \left(\ell(s_*, s_m) + \mathbb{E}\left[\mathrm{p}_2^{(-1)}(m)\right] \right) \\
&\leq L_{(\mathbf{GSA}),r} \varepsilon_n(m) \, \ell\left(s_*, \widehat{s}_m^{(-1)}\right) .
\end{aligned}$$

Hence, identity (28) gives

$$\left| \mathrm{crit}^0_{\mathrm{VFCV}}(m) - \ell\left(s_*, \widehat{s}_m^{(-1)}\right) \right| \leq L_{(\mathbf{GSA}),r} \varepsilon_n(m) \, \ell\left(s_*, \widehat{s}_m^{(-1)}\right) . \tag{34}$$

Control on the criterion $\mathrm{crit}^0_{\mathrm{VFCV}}$ for models of small dimension:

We consider models $m \in \mathcal{M}_n$ such that $D_m \leq A_{\mathcal{M},+} (\ln n)^3$. By (32), (59) and (33), it holds on Ω_n, for any $\tau > 0$ and for all $m \in \mathcal{M}_n$ such that $D_m \leq A_{\mathcal{M},+} (\ln n)^3$,

$$\begin{aligned}
&\left|\mathrm{crit}^0_{\mathrm{VFCV}}(m) - \ell\left(s_*, \widehat{s}_m^{(-1)}\right)\right| \\
&\leq L_{(\mathbf{GSA}),r}\frac{(\ln n)^2}{n} + L_{(\mathbf{GSA}),r}\left(\sqrt{\frac{\ell(s_*, s_m)\ln n}{n}} + \frac{\ln n}{n}\right) \\
&\leq L_{(\mathbf{GSA}),r}\frac{(\ln n)^2}{n} + L_{(\mathbf{GSA}),r}\tau\ell(s_*, s_m) + \left(\tau^{-1}+1\right)L_{(\mathbf{GSA})}\frac{\ln n}{n}.
\end{aligned}$$

Hence, by taking $\tau = (\ln n)^{-2}$ in the last display we get,

$$\left|\mathrm{crit}^0_{\mathrm{VFCV}}(m) - \ell\left(s_*, \widehat{s}_m^{(-1)}\right)\right| \leq L_{(\mathbf{GSA}),r}\left(\frac{\ell\left(s_*, \widehat{s}_m^{(-1)}\right)}{(\ln n)^2} + \frac{(\ln n)^3}{n}\right). \tag{35}$$

Oracle inequalities:

We exploit the following inequality, that defines the selected model $\widehat{m}$,

$$\mathrm{crit}^0_{\mathrm{VFCV}}(\widehat{m}) \leq \inf_{m\in\mathcal{M}_n}\left\{\mathrm{crit}^0_{\mathrm{VFCV}}(m)\right\}. \tag{36}$$

Indeed, using (34) and (35), we get that on Ω_n it holds,

$$\begin{aligned}
&\mathrm{crit}^0_{\mathrm{VFCV}}(\widehat{m}) \\
\geq\ &\left(1 - L_{(\mathbf{GSA}),r}\left[\frac{1}{(\ln n)^2} - \sup_{m:D_m\geq A_{\mathcal{M},+}(\ln n)^3}\varepsilon_n(m)\right]\right)\ell\left(s_*, \widehat{s}_{\widehat{m}}^{(-1)}\right) - L_{(\mathbf{GSA}),r}\frac{(\ln n)^3}{n} \\
\geq\ &\left(1 - \frac{L_{(\mathbf{GSA}),r}}{\sqrt{\ln n}}\right)\ell\left(s_*, \widehat{s}_{\widehat{m}}^{(-1)}\right) - L_{(\mathbf{GSA}),r}\frac{(\ln n)^3}{n}.
\end{aligned} \tag{37}$$

Furthermore, using again (34) and (35), we get

$$\begin{aligned}
&\inf_{m\in\mathcal{M}_n}\left\{\mathrm{crit}^0_{\mathrm{VFCV}}(m)\right\} \\
\leq\ &\left(1 + \frac{L_{(\mathbf{GSA}),r}}{\sqrt{\ln n}}\right)\inf_{m\in\mathcal{M}_n}\left\{\ell\left(s_*, \widehat{s}_m^{(-1)}\right)\right\} + L_{(\mathbf{GSA}),r}\frac{(\ln n)^3}{n}.
\end{aligned} \tag{38}$$

Putting (37) and (38) in (36), we get that for all $n \geq n_0((\mathbf{GSA}),r)$,

$$\begin{aligned}
\ell\left(s_*, \widehat{s}_{\widehat{m}}^{(-1)}\right) &\leq \left(1 - \frac{L_{(\mathbf{GSA}),r}}{\sqrt{\ln n}}\right)^{-1}\left[\left(1 + \frac{L_{(\mathbf{GSA}),r}}{\sqrt{\ln n}}\right)\inf_{m\in\mathcal{M}_n}\left\{\ell\left(s_*, \widehat{s}_m^{(-1)}\right)\right\} + L_{(\mathbf{GSA}),r}\frac{(\ln n)^3}{n}\right] \\
&\leq \left(1 + \frac{L_{(\mathbf{GSA}),r}}{\sqrt{\ln n}}\right)\inf_{m\in\mathcal{M}_n}\left\{\ell\left(s_*, \widehat{s}_m^{(-1)}\right)\right\} + L_{(\mathbf{GSA}),r}\frac{(\ln n)^3}{n}.
\end{aligned}$$

This concludes the proof of Theorem 7.1.

7.3.2 Proofs related to V-fold penalization

Recall that the set of assumptions (**GSA**) is defined in Section 7.2 above. The proof of Theorem 4.2 will be based on the following theorem, proved in [37] - see Theorem 2 and its proof under (**GSA**) therein.

Theorem 7.2 *Suppose that the assumptions (**GSA**) of Section 3.2 hold, and furthermore suppose that for some $\delta \in [0,1)$ and $A_p, A_r > 0$, there exists an event of probability at least $1 - A_p n^{-2}$ on which, for every model $m \in \mathcal{M}_n$ such that $D_m \geq A_{\mathcal{M},+}(\ln n)^3$, it holds*

$$|\mathrm{pen}(m) - 2\mathbb{E}[P_n(\gamma(s_m) - \gamma(\widehat{s}_m))]| \leq \delta\left(\ell(s_*, s_m) + \mathbb{E}[P_n(\gamma(s_m) - \gamma(\widehat{s}_m))]\right)$$

together with

$$|\mathrm{pen}(m)| \leq A_r\left(\frac{\ell(s_*, s_m)}{(\ln n)^2} + \frac{(\ln n)^3}{n}\right).$$

*Then there exist an integer n_0 only depending on δ and β_+ and on constants in (**GSA**), a positive constant L_3 only depending on $c_{\mathcal{M}}$ given in (**GSA**) and on A_p, two positive constants L_4 and L_5 only depending on constants in (**GSA**) and on A_r and a sequence*

$$\theta_n \leq \frac{L_4}{(\ln n)^{1/4}}$$

such that it holds for all $n \geq n_0$, with probability at least $1 - L_3 n^{-2}$,

$$\ell\left(s_*, \widehat{s}_{\widehat{m}}\right) \leq \left(\frac{1+\delta}{1-\delta} + \frac{5\theta_n}{(1-\delta)^2}\right) \inf_{m \in \mathcal{M}_n} \left\{\ell\left(s_*, \widehat{s}_m\right)\right\} + L_5 \frac{(\ln n)^3}{n}.$$

We now prove Theorem 4.2.

Proof of Theorem 4.2. We set

$$\mathrm{pen}_0(m) = \mathrm{pen}_{\mathrm{VF}}(m) - \frac{V-1}{V} \sum_{j=1}^{V} \left(P_n\left(\gamma\left(s_*\right)\right) - P_n^{(-j)}\left(\gamma\left(s_*\right)\right)\right) .$$

It is worth noting that the penalization procedure defined by pen_0 gives the same result as the procedure defined by $\mathrm{pen}_{\mathrm{VF}}$. It will be convenient for our analysis to consider pen_0 instead of $\mathrm{pen}_{\mathrm{VF}}$. Our strategy is to derive Theorem 4.2 as a corollary of Theorem 7.2 applied with $\mathrm{pen} \equiv \mathrm{pen}_0$.

As $P_n = (1 - V^{-1}) P_n^{(-j)} + V^{-1} P_n^{(j)}$, we get for all $m \in \mathcal{M}_n$,

$$\begin{aligned}
\mathrm{pen}_0(m) &= \frac{V-1}{V} \sum_{j=1}^{V} \left(P_n\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma\left(s_*\right)\right) - P_n^{(-j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma\left(s_*\right)\right)\right) \\
&= \frac{V-1}{V^2} \sum_{j=1}^{V} \left(P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma\left(s_*\right)\right) - P_n^{(-j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma\left(s_*\right)\right)\right) \\
&= \frac{V-1}{V^2} \sum_{j=1}^{V} \left(P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma\left(s_m\right)\right) - P_n^{(-j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma\left(s_m\right)\right)\right) \\
&\quad + \frac{V-1}{V^2} \sum_{j=1}^{V} \left(\left(P_n^{(j)} - P\right)\left(\gamma\left(s_m\right) - \gamma\left(s_*\right)\right) - \left(P_n^{(-j)} - P\right)\left(\gamma\left(s_m\right) - \gamma\left(s_*\right)\right)\right) \\
&= \frac{V-1}{V} \left(\bar{\mathrm{p}}_1(m) + \bar{\mathrm{p}}_2(m) + \bar{\delta}(m) - \bar{\delta}'(m)\right)
\end{aligned}$$

where

$$\bar{\mathrm{p}}_1(m) = \frac{1}{V} \sum_{j=1}^{V} P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma\left(s_m\right)\right) \ , \ \bar{\mathrm{p}}_2(m) = \frac{1}{V} \sum_{j=1}^{V} P_n^{(-j)}\left(\gamma\left(s_m\right) - \gamma\left(\widehat{s}_m^{(-j)}\right)\right),$$

and $\bar{\delta}(m)$ and $\bar{\delta}'(m)$ have been defined in Lemma 7.5. We also set

$$\mathrm{p}_1(m) = P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m\right)\right) \quad \text{and} \quad \mathrm{p}_2(m) = P_n\left(\gamma\left(s_m\right) - \gamma\left(\widehat{s}_m\right)\right).$$

Let Ω_n be the event on which:

- For all models $m \in \mathcal{M}_n$ of dimension D_m such that $A_{\mathcal{M},+} (\ln n)^3 \leq D_m$, it holds

$$\begin{aligned}
\left|\mathrm{p}_1(m) - \mathbb{E}\left[\mathrm{p}_2(m)\right]\right| &\leq L_{(\mathbf{GSA})} \varepsilon_n(m) \, \mathbb{E}\left[\mathrm{p}_2(m)\right] \\
\left|\mathrm{p}_2(m) - \mathbb{E}\left[\mathrm{p}_2(m)\right]\right| &\leq L_{(\mathbf{GSA})} \varepsilon_n^2(m) \, \mathbb{E}\left[\mathrm{p}_2(m)\right]
\end{aligned}$$

  where $\varepsilon_n(m)$ is defined in Theorem 5.2, together with

$$\begin{aligned}
\left|\bar{\mathrm{p}}_1(m) - \frac{V}{V-1}\mathbb{E}\left[\mathrm{p}_2(m)\right]\right| &\leq L_{(\mathbf{GSA}),r} \varepsilon_n(m) \, \mathbb{E}\left[\mathrm{p}_2(m)\right] \\
\left|\bar{\mathrm{p}}_2(m) - \frac{V}{V-1}\mathbb{E}\left[\mathrm{p}_2(m)\right]\right| &\leq L_{(\mathbf{GSA}),r} \varepsilon_n^2(m) \, \mathbb{E}\left[\mathrm{p}_2(m)\right] \\
\max\left\{\left|\bar{\delta}(m)\right|, \left|\bar{\delta}'(m)\right|\right\} &\leq \frac{\ell\left(s_*, s_m\right)}{\sqrt{D_m}} + L_{(\mathbf{GSA}),r} \frac{\ln n}{\sqrt{D_m}} \mathbb{E}\left[\mathrm{p}_2(m)\right] \qquad (39)
\end{aligned}$$

- For all models $m \in \mathcal{M}_n$ of dimension D_m such that $D_m \leq A_{\mathcal{M},+} (\ln n)^3$, it holds

$$\max\left\{\left|\bar{\delta}(m)\right|, \left|\bar{\delta}'(m)\right|\right\} \leq L_{(\mathbf{GSA}),r}\left(\sqrt{\frac{\ell(s_*, s_m)\ln n}{n}} + \frac{\ln n}{n}\right) \tag{40}$$

$$\bar{\mathrm{p}}_2(m) \leq L_{(\mathbf{GSA}),r}\frac{D_m \vee \ln n}{n} \leq L_{(\mathbf{GSA}),r}\frac{(\ln n)^3}{n} \tag{41}$$

$$\bar{\mathrm{p}}_1(m) \leq L_{(\mathbf{GSA}),r}\left(\frac{(\ln n)^2}{n} + \frac{D_m \vee \ln n}{n}\right) \leq L_{(\mathbf{GSA}),r}\frac{(\ln n)^3}{n} \tag{42}$$

By Theorem 2 of [36] and Lemma 4 of [37] applied with $\alpha = 2 + \alpha_{\mathcal{M}}$ and sample size $n_V = n(V-1)/V$, Corollary 7.4 and Lemma 7.5 applied with $\alpha = 2 + \alpha_{\mathcal{M}}$, we get for all $n \geq n_0((\mathbf{GSA}), r)$,

$$\mathbb{P}(\Omega_n) \geq 1 - L\sum_{m \in \mathcal{M}_n} n^{-2-\alpha_{\mathcal{M}}} \geq 1 - L_{c_{\mathcal{M}}} n^{-2}\,.$$

We consider models $m \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^3 \leq D_m$. Notice that (39) implies by (18) that, for all $m \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^3 \leq D_m$,

$$\begin{aligned}\max\left\{\left|\bar{\delta}(m)\right|, \left|\bar{\delta}'(m)\right|\right\} &\leq L_{(\mathbf{GSA}),r}\left(\frac{(\ln n)^3}{D_m}\cdot\frac{\ln n}{D_m}\right)^{1/4} \times (\ell(s_*, s_m) + \mathbb{E}[p_2(m)]) \\ &\leq L_{(\mathbf{GSA}),r}\varepsilon_n(m)(\ell(s_*, s_m) + \mathbb{E}[p_2(m)])\,.\end{aligned}$$

We deduce that on Ω_n we have, for all models $m \in \mathcal{M}_n$ such that $A_{\mathcal{M},+}(\ln n)^3 \leq D_m$ and for all $n \geq n_0((\mathbf{GSA}),r)$,

$$\begin{aligned}&|\mathrm{pen}_0(m) - 2\mathbb{E}[\mathrm{p}_2(m)]| \\ \leq\ & \frac{V-1}{V}\left(\left|\bar{\mathrm{p}}_1(m) - \frac{V}{V-1}\mathbb{E}[\mathrm{p}_2(m)]\right| + \left|\bar{\mathrm{p}}_2(m) - \frac{V}{V-1}\mathbb{E}[\mathrm{p}_2(m)]\right|\right) \\ &+ \max\left\{\left|\bar{\delta}(m)\right|, \left|\bar{\delta}'(m)\right|\right\} \\ \leq\ & L_{(\mathbf{GSA}),r}\varepsilon_n(m)(\ell(s_*, s_m) + \mathbb{E}[\mathrm{p}_2(m)])\end{aligned} \tag{43}$$

Let us now consider models $m \in \mathcal{M}_n$ such that $D_m \leq A_{\mathcal{M},+}(\ln n)^3$. By (40), (41) and (42), we have on Ω_n,

$$\begin{aligned}|\mathrm{pen}_0(m)| &= \frac{V-1}{V}\left|\bar{\mathrm{p}}_1(m) + \bar{\mathrm{p}}_2(m) + \bar{\delta}(m) - \bar{\delta}'(m)\right| \\ &\leq L_{(\mathbf{GSA}),r}\left(\sqrt{\frac{\ell(s_*, s_m)\ln n}{n}} + \frac{(\ln n)^3}{n}\right) \\ &\leq L_{(\mathbf{GSA}),r}\left(\frac{\ell(s_*, s_m)}{(\ln n)^2} + \frac{(\ln n)^3}{n}\right)\end{aligned} \tag{44}$$

Inequality (44) implies that inequality (10) of Theorem 3.2 is satisfied with $A_r = L_{(\mathbf{GSA}),r}$. From (43) and (44), we thus apply Theorem 7.2 with $A_p = L_{A_p, c_{\mathcal{M}}}$, and this gives Theorem 4.2 with

$$\theta_n = L_{(\mathbf{GSA}),r}\left((\ln n)^{-2} + \sup_{m \in \mathcal{M}_n}\left\{\varepsilon_n(m);\ A_{\mathcal{M},+}(\ln n)^3 \leq D_m \leq n^{\eta + 1/(1+\beta_+)}\right\}\right).$$

7.4 Proofs related to Section 5

7.4.1 Proofs for strongly localized bases

Proof of Theorem 5.1. Let $C > 0$. Set

$$\mathcal{F}_C^\infty := \{s \in m\,;\, \|s - s_m\|_\infty \leq C\}$$

and

$$\mathcal{F}_{>C}^\infty := \{s \in m\,;\, \|s - s_m\|_\infty > C\} = m \backslash \mathcal{F}_C^\infty.$$

Take an orthonormal basis $(\varphi_k)_{k=1}^{D_m}$ of $(m, \|\cdot\|_2)$ satisfying (**Aslb**). By Lemma 7.7, we get that there exists $L^{(1)}_{A,r_m,\alpha} > 0$ such that, by setting

$$\Omega_1 = \left\{ \max_{k\in\{1,\dots,D_m\}} |(P_n - P)(\psi_m \cdot \varphi_k)| \leq L^{(1)}_{A,r_m,\alpha} \sqrt{\frac{\ln n}{n}} \right\},$$

we have for all $n \geq n_0(A_+)$, $\mathbb{P}(\Omega_1) \geq 1 - n^{-\alpha}$. Moreover, we set

$$\Omega_2 = \left\{ \max_{(k,l)\in\{1,\dots,D_m\}^2} |(P_n - P)(\varphi_k \cdot \varphi_l)| \leq L^{(2)}_{\alpha,r_m} \min\left\{\|\varphi_k\|_\infty ; \|\varphi_l\|_\infty\right\} \sqrt{\frac{\ln n}{n}} \right\},$$

where $L^{(2)}_{\alpha,r_m}$ is defined in Lemma 7.6. By Lemma 7.6, we have that for all $n \geq n_0(A_+)$, $\mathbb{P}(\Omega_2) \geq 1 - n^{-\alpha}$ and so, for all $n \geq n_0(A_+)$,

$$\mathbb{P}\left(\Omega_1 \bigcap \Omega_2\right) \geq 1 - 2n^{-\alpha}.$$

We thus have for all $n \geq n_0(A_+)$,

$$\begin{aligned}
& \mathbb{P}\left(\|\widehat{s}_m - s_m\|_\infty > C\right) \\
\leq\ & \mathbb{P}\left(\inf_{s\in\mathcal{F}^\infty_{>C}} P_n(\gamma(s) - \gamma(s_m)) \leq \inf_{s\in\mathcal{F}^\infty_{C}} P_n(\gamma(s) - \gamma(s_m)) \right) \\
=\ & \mathbb{P}\left(\sup_{s\in\mathcal{F}^\infty_{>C}} P_n(\gamma(s_m) - \gamma(s)) \geq \sup_{s\in\mathcal{F}^\infty_{C}} P_n(\gamma(s_m) - \gamma(s)) \right) \\
\leq\ & \mathbb{P}\left(\left\{ \sup_{s\in\mathcal{F}^\infty_{>C}} P_n(\gamma(s_m) - \gamma(s)) \geq \sup_{s\in\mathcal{F}^\infty_{C/2}} P_n(\gamma(s_m) - \gamma(s)) \right\} \bigcap \Omega_1 \bigcap \Omega_2 \right) + 2n^{-\alpha}. \qquad (45)
\end{aligned}$$

Now, for any $s \in m$ such that

$$s - s_m = \sum_{k=1}^{D_m} \beta_k \varphi_k,\ \beta = (\beta_k)_{k=1}^{D_m} \in \mathbb{R}^{D_m},$$

we have

$$\begin{aligned}
& P_n(\gamma(s_m) - \gamma(s)) \\
=\ & (P_n - P)(\psi_m \cdot (s_m - s)) - (P_n - P)\left((s - s_m)^2\right) - P(\gamma(s) - \gamma(s_m)) \\
=\ & \sum_{k=1}^{D_m} \beta_k (P_n - P)(\psi_m \cdot \varphi_k) - \sum_{k,l=1}^{D_m} \beta_k \beta_l (P_n - P)(\varphi_k \cdot \varphi_l) - \sum_{k=1}^{D_m} \beta_k^2.
\end{aligned}$$

We set for any $(k,l) \in \{1,\dots,D_m\}^2$,

$$R^{(1)}_{n,k} = (P_n - P)(\psi_m \cdot \varphi_k) \quad \text{and} \quad R^{(2)}_{n,k,l} = (P_n - P)(\varphi_k \cdot \varphi_l).$$

Moreover, we set a function h_n, defined as follows,

$$h_n : \beta = (\beta_k)_{k=1}^{D_m} \longmapsto \sum_{k=1}^{D_m} \beta_k R^{(1)}_{n,k} - \sum_{k,l=1}^{D_m} \beta_k \beta_l R^{(2)}_{n,k,l} - \sum_{k=1}^{D_m} \beta_k^2.$$

We thus have for any $s \in m$ such that $s - s_m = \sum_{k=1}^{D_m} \beta_k \varphi_k$, $\beta = (\beta_k)_{k=1}^{D_m} \in \mathbb{R}^{D_m}$,

$$P_n(\gamma(s_m) - \gamma(s)) = h_n(\beta). \qquad (46)$$

In addition we set for any $\beta = (\beta_k)_{k=1}^{D_m} \in \mathbb{R}^{D_m}$,

$$|\beta|_{m,\infty} = r_m \sum_{i=1}^{b_m} \sqrt{A_i} \max_{k\in\Pi_i} |\beta_k|.$$

It is straightforward to see that $|\cdot|_{m,\infty}$ is a norm on $\mathbb{R}^{D_m}$. We also set for a real $D_m \times D_m$ matrix B, its operator norm $\|B\|_m$ associated to the norm $|\cdot|_{m,\infty}$ on the D_m-dimensional vectors. More explicitly, we set for any $B \in \mathbb{R}^{D_m \times D_m}$,

$$\|B\|_m := \sup_{\beta \in \mathbb{R}^D,\ \beta \neq 0} \frac{|B\beta|_{m,\infty}}{|\beta|_{m,\infty}}.$$

We have, for any $B = (B_{k,l})_{k,l=1,\dots D_m} \in \mathbb{R}^{D_m \times D_m}$,

$$\begin{aligned}
\|B\|_m &= \sup_{\beta \in \mathbb{R}^{D_m},\ |\beta|_{m,\infty}=1} \left\{ r_m \sum_{i=1}^{b_m} \sqrt{A_i} \max_{k \in \Pi_i} \left| \sum_{l=1}^{D_m} B_{k,l}\beta_l \right| \right\} \\
&= \sup_{\beta \in \mathbb{R}^{D_m},\ |\beta|_{m,\infty}=1} \left\{ r_m \sum_{i=1}^{b_m} \sqrt{A_i} \max_{k \in \Pi_i} \left| \sum_{j=1}^{b_m} \sum_{l \in \Pi_j} B_{k,l}\beta_l \right| \right\} \\
&= \sup_{\beta \in \mathbb{R}^{D_m},\ |\beta|_{m,\infty}=1} \left\{ \sum_{i=1}^{b_m} \sqrt{A_i} \max_{k \in \Pi_i} \left\{ r_m \sum_{j=1}^{b_m} \sqrt{A_j} \max_{l \in \Pi_j} |\beta_l| \left(\sqrt{A_j^{-1}} \sum_{l \in \Pi_j} |B_{k,l}| \right) \right\} \right\} \\
&= \sum_{i=1}^{b_m} \sqrt{A_i} \max_{k \in \Pi_i} \left\{ \max_{j \in \{1,\dots,b_m\}} \left\{ \sqrt{A_j^{-1}} \sum_{l \in \Pi_j} |B_{k,l}| \right\} \right\}.
\end{aligned}$$

Notice that by Inequality (5) of (**Aslb**), it holds

$$\mathcal{F}^{\infty}_{>C} \subset \left\{ s \in m \ ;\ s - s_m = \sum_{k=1}^{D_m} \beta_k \varphi_k \ \ \& \ \ |\beta|_{m,\infty} \geq C \right\} \tag{47}$$

and

$$\mathcal{F}^{\infty}_{C/2} \supset \left\{ s \in m \ ;\ s - s_m = \sum_{k=1}^{D_m} \beta_k \varphi_k \ \ \& \ \ |\beta|_{m,\infty} \leq C/2 \right\}. \tag{48}$$

Hence, from (45), (46) (47) and (48) we deduce that if we find on $\Omega_1 \bigcap \Omega_2$ a value of C such that

$$\sup_{\beta \in \mathbb{R}^{D_m},\ |\beta|_{m,\infty} \geq C} h_n(\beta) < \sup_{\beta \in \mathbb{R}^{D_m},\ |\beta|_{m,\infty} \leq C/2} h_n(\beta),$$

then Inequality (17) follows and Theorem 5.1 is proved. Taking the partial derivatives of h_n with respect to the coordinates of its arguments, it then holds for any $(k,l) \in \{1,\dots,D_m\}^2$ and $\beta = (\beta_i)_{i=1}^{D_m} \in \mathbb{R}^{D_m}$,

$$\frac{\partial h_n}{\partial \beta_k}(\beta) = R^{(1)}_{n,k} - 2\sum_{i=1}^{D_m} \beta_i R^{(2)}_{n,k,i} - 2\beta_k \tag{49}$$

We look now at the set of solutions β of the following system,

$$\frac{\partial h_n}{\partial \beta_k}(\beta) = 0 \ , \forall k \in \{1,\dots,D_m\}. \tag{50}$$

We define the $D_m \times D_m$ matrix $R^{(2)}_n$ to be

$$R^{(2)}_n := \left(R^{(2)}_{n,k,l} \right)_{k,l=1,\dots,D_m}$$

and by (49), the system given in (50) can be written

$$2\left(I_{D_m} + R^{(2)}_n \right) \beta = R^{(1)}_n, \tag{**S**}$$

where $R^{(1)}_n$ is a D-dimensional vector defined by

$$R^{(1)}_n = \left(R^{(1)}_{n,k} \right)_{k=1,\dots,D_m}.$$

Let us give an upper bound of the norm $\left\|R_n^{(2)}\right\|_m$, in order to show that the matrix $I_{D_m}+R_n^{(2)}$ is nonsingular. On Ω_2 we have, using (6),

$$\begin{aligned}
\left\|R_n^{(2)}\right\|_m &= \sum_{i=1}^{b_m}\sqrt{A_i}\max_{k\in\Pi_i}\left\{\max_{j\in\{1,\dots,b_m\}}\left\{\sqrt{A_j^{-1}}\sum_{l\in\Pi_j}\left|R_{n,k,l}^{(2)}\right|\right\}\right\}\\
&= \sum_{i=1}^{b_m}\sqrt{A_i}\max_{k\in\Pi_i}\left\{\max_{j\in\{1,\dots,b_m\}}\left\{\sqrt{A_j^{-1}}\sum_{l\in\Pi_{j|k}}\left|R_{n,k,l}^{(2)}\right|\right\}\right\}\\
&\leq \sum_{i=1}^{b_m}\sqrt{A_i}\max_{k\in\Pi_i}\left\{\max_{j\in\{1,\dots,b_m\}}\left\{\sqrt{A_j^{-1}}\mathrm{Card}\left(\Pi_{j|k}\right)\max_{l\in\Pi_j}\left|\left(P_n-P\right)\left(\varphi_k\cdot\varphi_l\right)\right|\right\}\right\}\\
&\leq A_c L_{\alpha,r_m}^{(2)}\sqrt{\frac{\ln n}{n}}\sum_{i=1}^{b_m}\max_{j\in\{1,\dots,b_m\}}\left\{\sqrt{\frac{A_i}{A_j}}\left(\frac{A_j}{A_i}\vee 1\right)\sqrt{\min\left\{A_i;A_j\right\}}\right\} \qquad (51)
\end{aligned}$$

We deduce from (4) and (51) that on Ω_2,

$$\left\|R_n^{(2)}\right\|_m \leq L_{A_c,\alpha,r_m}\cdot b_m\sqrt{\frac{A_{b_m}\ln n}{n}}. \qquad (52)$$

Hence, from (52) and the fact that $b_m^2 A_{b_m}\leq A_+\frac{n}{(\ln n)^2}$, we get that for all $n\geq n_0\left(A_+,A_c,r_m,\alpha\right)$, it holds on Ω_2,

$$\left\|R_n^{(2)}\right\|_m \leq \frac{1}{2}$$

and the matrix $\left(I_{D_m}+R_n^{(2)}\right)$ is nonsingular, of inverse $\left(I_{D_m}+R_n^{(2)}\right)^{-1}=\sum_{u=0}^{+\infty}\left(-R_n^{(2)}\right)^u$. Hence, the system (**S**) admits a unique solution $\beta^{(n)}$, given by

$$\beta^{(n)}=\frac{1}{2}\left(I_{D_m}+R_n^{(2)}\right)^{-1}R_n^{(1)}.$$

Now, on Ω_1 we have by (4),

$$\left|R_n^{(1)}\right|_{m,\infty}\leq r_m\left(\sum_{i=1}^{b_m}\sqrt{A_i}\right)\max_{k\in\{1,\dots,D_m\}}\left|\left(P_n-P\right)\left(\psi_m\cdot\varphi_k\right)\right|\leq r_m L_{A_m,r_m,\alpha}^{(1)}\sqrt{\frac{D_m\ln n}{n}}$$

and we deduce that for all $n_0\left(A_+,A_c,r_m,\alpha\right)$, it holds on $\Omega_2\bigcap\Omega_1$,

$$\left|\beta^{(n)}\right|_{m,\infty}\leq\frac{1}{2}\left\|\left(I_d+R_n^{(2)}\right)^{-1}\right\|_m\left|R_n^{(1)}\right|_{m,\infty}\leq r_m L_{A,r_m,\alpha}^{(1)}\sqrt{\frac{D_m\ln n}{n}}. \qquad (53)$$

Moreover, by the formula (46) we have

$$h_n\left(\beta\right)=P_n\left(\gamma\left(s_m\right)\right)-P_n\left(Y-\sum_{k=1}^{D_m}\beta_k\varphi_k\right)^2$$

and we thus see that h_n is concave. Hence, for all $n_0\left(A_+,A_c,r_m,\alpha\right)$, we get that on Ω_2, $\beta^{(n)}$ is the unique maximum of h_n and on $\Omega_2\bigcap\Omega_1$, by (53), concavity of h_n and uniqueness of $\beta^{(n)}$, we get

$$h_n\left(\beta^{(n)}\right)=\sup_{\beta\in\mathbb{R}^{D_m},\ |\beta|_{m,\infty}\leq C/2}h_n\left(\beta\right)>\sup_{\beta\in\mathbb{R}^{D_m},\ |\beta|_{m,\infty}\geq C}h_n\left(\beta\right),$$

with $C=2r_m L_{A,r_m,\alpha}^{(1)}\sqrt{\frac{D_m\ln n}{n}}$, which concludes the proof. ■

Remark 7.1 *The proof of Theorem 5.1 can be adapted for models endowed with a localized basis structure. Indeed, if we set for any* $B\in\mathbb{R}^{D_m\times D_m}$,

$$\|B\|_m:=\sup_{\beta\in\mathbb{R}^{D_m},\ \beta\neq 0}\frac{|B\beta|_{m,\infty}}{|\beta|_{m,\infty}}=\sup_{\beta\in\mathbb{R}^{D_m},\ \beta\neq 0}\frac{|B\beta|_\infty}{|\beta|_\infty},$$

then we have, the following classical formula

$$\|B\|_m = \max_{k\in\{1,\dots,D_m\}} \left\{ \left\{ \sum_{l\in\{1,\dots,D_m\}} |B_{k,l}| \right\} \right\}.$$

Now, it holds,

$$\begin{aligned}
\left\| R_n^{(2)} \right\|_m &= \max_{k\in\{1,\dots,D_m\}} \left\{ \left\{ \sum_{l\in\{1,\dots,D_m\}} |(P_n - P)(\varphi_k \cdot \varphi_l)| \right\} \right\} \\
&\leq L^{(2)}_{\alpha,r_m} \max_{k\in\{1,\dots,D_m\}} \left\{ \left\{ \sum_{l\in\{1,\dots,D_m\}} \min\left\{ \|\varphi_k\|_\infty ; \|\varphi_l\|_\infty \right\} \sqrt{\frac{\ln n}{n}} \right\} \right\} \\
&\leq r_m L^{(2)}_{\alpha,r_m} \sqrt{\frac{D_m^3 \ln n}{n}}
\end{aligned}$$

The previous bound tends to zero if $D_m \leq n^{1/3}/ln^2(n)$ and this is the essential reason why results for localized bases are restricted to models with dimension lower that $n^{1/3}/ln^2(n)$ while for strongly localized bases we can go as far as $D_m \leq n/ln^2(n)$ (see also Remark 5.1).

7.4.2 Proofs related to excess risks' representations

Proof of Proposition 5.3. Let us write $C_* := \mathcal{F}(\widehat{s}_m)$. It holds

$$\begin{aligned}
\inf_{s\in d_{C_*}} P_n(\gamma(s)) &= \inf_{s\in m} P_n(\gamma(s)) \\
&\leq \min_{C\geq 0} \left\{ \inf_{s\in d_C} P_n(\gamma(s)) \right\},
\end{aligned}$$

which readily proves Formula (21). Formula (22) is a direct consequence of (21), since $m_C = \bigcup_{R\leq C} d_C$. ■

Proof of Proposition 5.5. We will only prove the case where $R_0 = +\infty$. Then the situation where $R_0 \in \mathbb{R}_+$ can be deduced easily by noticing that the subset $\{s \in m \ ; \ \mathcal{G}(s) \leq R_0\}$ of m actually plays the role of m in this latter case.

When $R_0 = +\infty$, we have with the notations of Proposition 5.3 and by taking $\mathcal{F} = P(\gamma(\cdot) - \gamma(s_m))$, $\widetilde{d}_C = d_C$ and $\widetilde{m}_C = m_C$. From formula (21) we thus get

$$\begin{aligned}
P(\gamma(\widehat{s}_m) - \gamma(s_m)) &\in \arg\min_{C\geq 0} \left\{ \inf_{s\in\widetilde{d}_C} P_n(\gamma(s)) \right\} \\
&= \arg\max_{C\geq 0} \left\{ \sup_{s\in\widetilde{d}_C} P_n(\gamma(s_m) - \gamma(s)) \right\} \\
&= \arg\max_{C\geq 0} \left\{ \sup_{s\in\widetilde{d}_C} (P_n - P)(\gamma(s_m) - \gamma(s)) - C \right\}.
\end{aligned}$$

Hence, Formula (23) is proved. Now, for (24), take any $C \geq 0$ and notice that there exists a random variable $C_1 \in [0, C]$ such that

$$\begin{aligned}
\sup_{s\in\widetilde{m}_C} \{(P_n - P)(\gamma(s_m) - \gamma(s))\} - C &= \sup_{s\in\widetilde{d}_{C_1}} (P_n - P)(\gamma(s_m) - \gamma(s)) - C \\
&\leq \sup_{s\in\widetilde{d}_{C_1}} (P_n - P)(\gamma(s_m) - \gamma(s)) - C_1 \\
&\leq \sup_{s\in\widetilde{d}_{C_*}} (P_n - P)(\gamma(s_m) - \gamma(s)) - C_* \ , \qquad (54)
\end{aligned}$$

where $C_* := P(\gamma(\widehat{s}_m) - \gamma(s_m))$. Taking $C = C_*$, we get

$$\sup_{s\in\widetilde{m}_{C_*}} \{(P_n - P)(\gamma(s_m) - \gamma(s))\} - C_* \leq \sup_{s\in\widetilde{d}_{C_*}} (P_n - P)(\gamma(s_m) - \gamma(s)) - C_*$$

and since $\widetilde{d}_{C_*} \subset \widetilde{m}_{C_*}$, this implies

$$\sup_{s \in \widetilde{m}_{C_*}} \{(P_n - P)(\gamma(s_m) - \gamma(s))\} - C_* = \sup_{s \in \widetilde{d}_{C_*}} (P_n - P)(\gamma(s_m) - \gamma(s)) - C_*.$$

Together with (54), the latter equality gives that for any $C \geq 0$,

$$\sup_{s \in \widetilde{m}_C} \{(P_n - P)(\gamma(s_m) - \gamma(s))\} - C \leq \sup_{s \in \widetilde{m}_{C_*}} \{(P_n - P)(\gamma(s_m) - \gamma(s))\} - C_*,$$

which is another way to write (24).

Now, considering the case of the empirical excess risk, we could again apply Proposition 5.3, but we will follow a more direct proof. We have, by definition of $\widehat{s}_m$,

$$\begin{aligned} \ell_{\mathrm{emp}}(s_m, \widehat{s}_m) &= P_n(\gamma(s_m) - \gamma(\widehat{s}_m)) \\ &= \max_{s \in m} \{P_n(\gamma(s_m) - \gamma(s))\}. \end{aligned}$$

Now, as $\{\mathcal{G}(\widehat{s}_m) \leq R_0\} = \bigcup_{C \geq 0} \widetilde{d}_C = \bigcup_{C \geq 0} \widetilde{m}_C$, we get

$$\begin{aligned} \ell_{\mathrm{emp}}(s_m, \widehat{s}_m) &= \max_{s \in m} \{P_n(\gamma(s_m) - \gamma(s))\} \\ &= \max_{C \geq 0} \sup_{s \in \widetilde{d}_C} \{P_n(\gamma(s_m) - \gamma(s))\} \\ &= \max_{C \geq 0} \left\{ \sup_{s \in \widetilde{d}_C} \{(P_n - P)(\gamma(s_m) - \gamma(s))\} - C \right\}, \end{aligned}$$

that is (25). Now formula (26) follows from the kind of arguments that allow to prove (24) based on (25). ∎

Acknowledgements

The authors are very grateful to the Associate Editor and the anonymous referees for their careful reading and valuable comments that have led to several improvements and new developments of the paper.

References

[1] A. Antoniadis, J. Bigot, and T. Sapatinas. Wavelet estimators in nonparametric regression: a comparative simulation study. *Journal of Statistical Software*, 6:1–83, 2001.

[2] A. Antoniadis, G. Gregoire, and I. McKeague. Wavelet methods for curve estimation. *J. Amer. Statist. Assoc.*, 89(428):1340–1353, 1994.

[3] S. Arlot. V-fold cross-validation improved: V-fold penalization, Feb. 2008. arXiv:0802.0566v2.

[4] S. Arlot. Choosing a penalty for model selection in heteroscedastic regression, June 2010. arXiv:0812.3141.

[5] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.

[6] S. Arlot and A. Célisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

[7] S. Arlot and A. Célisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Stat. Comput.*, 21(4):613–632, 2011.

[8] S. Arlot and M. Lerasle. Choice of V for V-fold cross-validation in least-squares density estimation. *J. Mach. Learn. Res.*, 17(208):1–50, 2016.

[9] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[10] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470, 2012.

[11] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[12] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[13] P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.

[14] T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.*, 27(3):898–924, 1999.

[15] T. Cai and L. Brown. Wavelet shrinkage for nonequispaced samples. *Ann. Statist.*, 26:1783–1799, 1998.

[16] T. Cai and L. Brown. Wavelet estimation for samples with random uniform design. *Statist. Probab. Lett.*, 42(3):313–321, 1999.

[17] G. Castellan. Modified Akaike's criterion for histogram density estimation. *Technical report ♯99.61, Université Paris-Sud*, 1999.

[18] S. Chatterjee. A new perspective on least squares under convex constraint. *Ann. Statist.*, 42(6):2340–2381, 12 2014.

[19] A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1):54–81, 1993.

[20] A. Donoho, D. Maleki and M. Shahram. Wavelab 850. 2006.

[21] D. Donoho and I. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[22] S. Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

[23] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[24] P. Hall and B. Turlach. Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *Ann. Statist.*, 25(5):1912–1925, 1997.

[25] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.

[26] R. Kulik and M. Raimondo. Wavelet regression in random design with heteroscedastic dependent errors. *Ann. Statist.*, 37(6A):3396–3430, 2009.

[27] G. Lecué and C. Mitchell. Oracle inequalities for cross-validation type procedures. *Electron. J. Stat.*, 6:1803–1837, 2012.

[28] M. Lerasle. Optimal model selection for density estimation of stationary data under various mixing conditions. *Ann. Statist.*, 39(4):1852–1877, 2011.

[29] M. Lerasle. Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(3):884–908, 2012.

[30] S. Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

[31] J. Marron, S. Adak, I. Johnstone, M. Neumann, and P. Patil. Exact risk analysis of wavelet regression. *J. Comput. Graph. Statist.*, 7(3):278–309, 1998.

[32] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[33] A. Muro and S. van de Geer. Concentration behavior of the penalized least squares estimator. *arXiv preprint arXiv:1511.08698*, 2015.

[34] G. Nason. Wavelet shrinkage using cross-validation. *J. R. Stat. Soc. Ser. B*, pages 463–479, 1996.

[35] A. Saumard. Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models, Aug. 2010. hal-00512310.

[36] A. Saumard. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. *Electron. J. Statist.*, 6(1-2):579–655, 2012.

[37] A. Saumard. Optimal model selection in heteroscedastic regression using piecewise polynomial functions. *Electron. J. Statist.*, 7:1184–1223, 2013.

[38] C. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982.

[39] S. van de Geer and M. Wainwright. On concentration for (regularized) empirical risk minimization. *arXiv preprint arXiv:1512.00677*, 2016.

[40] M. Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1):252–273, 2003.

Appendix

7.5 Some lemmas instrumental in the proofs

We gather here the lemmas that are used in the proofs of Section 7.

In the next lemma, we apply Lemma 9 of [37], with $n_2 = n/V$, $n_1 = n - n_2 = n\,(1 - 1/V)$, $\tau = 2$ and we set $r = c^{-1} \in (1, +\infty)$. Furthermore, the notations P_{n_2} and $s_{n_1}(m)$ used in [37] correspond respectively to the quantities $P_n^{(j)}$ and $\widehat{s}_m^{(-j)}$.

Lemma 7.3 *Assume that (**GSA**) holds. Let $r \in (2, +\infty)$ and $V \in \{2, \dots, n-1\}$ satisfying $1 < V \leq r$. Then there exists $L = L_{(\mathbf{GSA}),r} > 0$ such that for all $m \in \mathcal{M}_n$ satisfying $D_m \geq A_{\mathcal{M},+} (\ln n)^3$, by setting*

$$\varepsilon_n^{(1)}(m) = L\sqrt{\frac{\ln n}{D_m}} \leq \frac{L}{\ln n},$$

it holds for all $n \geq n_0((\mathbf{GSA}), r)$ and for all $j \in \{1, \dots, V\}$,

$$\mathbb{P}\left(\left|P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma(s_m)\right) - P\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma(s_m)\right)\right| \geq \varepsilon_n^{(1)}(m)\,\mathbb{E}\left[\mathrm{p}_2^{(-1)}(m)\right]\right) \leq 12n^{-2-\alpha_{\mathcal{M}}},$$

where $\mathrm{p}_2^{(-1)}(m) = P_n^{(-1)}\left(\gamma(s_m) - \gamma\left(\widehat{s}_m^{(-1)}\right)\right)$. If $D_m \leq A_{\mathcal{M},+}(\ln n)^3$, then for all $n \geq n_0((\mathbf{GSA}), r)$,

$$\mathbb{P}\left(\left|P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma(s_m)\right) - P\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma(s_m)\right)\right| \geq L\frac{(\ln n)^2}{n}\right) \leq 12n^{-2-\alpha_{\mathcal{M}}}.$$

Taking into account the averaging between the blocks of the V-fold, we get from Lemma 7.3 the following corollary.

Corollary 7.4 *Assume that (**GSA**) holds. Let $r \in (0, 1)$ and $V \in \{2, \dots, n-1\}$ satisfying $1 < V \leq r$. Then there exists $L = L_{(\mathbf{GSA}),r} > 0$ such that for all $m \in \mathcal{M}_n$ satisfying $D_m \geq A_{\mathcal{M},+}(\ln n)^3$, it holds for all $n \geq n_0((\mathbf{GSA}), r)$,*

$$\mathbb{P}\left(\left|P\left(\gamma\left(\widehat{s}_m^{(-1)}\right) - \gamma(s_m)\right) - \frac{1}{V}\sum_{j=1}^{V} P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma(s_m)\right)\right| \geq L\varepsilon_n(m)\,\mathbb{E}\left[\mathrm{p}_2^{(-1)}(m)\right]\right) \leq 22rn^{-2-\alpha_{\mathcal{M}}}, \tag{55}$$

where $\mathrm{p}_2^{(-1)}(m) = P_n^{(-1)}\left(\gamma(s_m) - \gamma\left(\widehat{s}_m^{(-1)}\right)\right)$ and $\varepsilon_n(m)$ is defined in Theorem 5.2. If $D_m \leq A_{\mathcal{M},+}(\ln n)^3$, then for all $n \geq n_0((\mathbf{GSA}), r)$,

$$\mathbb{P}\left(\left|P\left(\gamma\left(\widehat{s}_m^{(-1)}\right) - \gamma(s_m)\right) - \frac{1}{V}\sum_{j=1}^{V} P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right) - \gamma(s_m)\right)\right| \geq L\frac{(\ln n)^2}{n}\right) \leq 22rn^{-2-\alpha_{\mathcal{M}}}. \tag{56}$$

Proof. First we prove the following inequality,

$$\mathbb{P}\left(\left|\tfrac{1}{V}\sum_{j=1}^{V} P\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)-\tfrac{1}{V}\sum_{j=1}^{V} P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)\right|\geq \varepsilon_n^{(1)}\left(m\right)\mathbb{E}\left[\mathrm{p}_2^{(-1)}\left(m\right)\right]\right)\leq 12Vn^{-2-\alpha_{\mathcal{M}}}. \tag{57}$$

Indeed, it easily derives from Lemma 7.3 together with a union bound along the V blocks, taking advantage of the following formula

$$\begin{aligned}&\left|\frac{1}{V}\sum_{j=1}^{V} P\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)-\frac{1}{V}\sum_{j=1}^{V} P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)\right|\\ \leq\ &\max_{j\in\{1,\dots,V\}}\left|P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)-P\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)\right|.\end{aligned}$$

Then, we show that the quantity $\frac{1}{V}\sum_{j=1}^{V} P\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)$ is close enough to $P\left(\gamma\left(\widehat{s}_m^{(-1)}\right)-\gamma\left(s_m\right)\right)$ with probability close to one. Indeed, it holds for any $C\geq 0$,

$$\begin{aligned}&\mathbb{P}\left(\left|P\left(\gamma\left(\widehat{s}_m^{(-1)}\right)-\gamma\left(s_m\right)\right)-\frac{1}{V}\sum_{j=1}^{V} P\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)\right|\geq C\right)\\ =\ &\mathbb{P}\left(\left|\frac{1}{V}\sum_{j=2}^{V}\left[P\left(\gamma\left(\widehat{s}_m^{(-1)}\right)-\gamma\left(s_m\right)\right)-P\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)\right]\right|\geq C\right)\\ \leq\ &\mathbb{P}\left(\max_{j\in\{2,\dots,V\}}\left|P\left(\gamma\left(\widehat{s}_m^{(-1)}\right)-\gamma\left(s_m\right)\right)-P\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)\right|\geq C\right)\\ \leq\ &\sum_{j=2}^{V}\mathbb{P}\left(\left|P\left(\gamma\left(\widehat{s}_m^{(-1)}\right)-\gamma\left(s_m\right)\right)-P\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)\right|\geq C\right)\\ =\ &\sum_{j=2}^{V}\mathbb{P}\left(\left|\left[P\left(\gamma\left(\widehat{s}_m^{(-1)}\right)-\gamma\left(s_m\right)\right)-\frac{\mathcal{C}_m}{n}\right]-\left[P\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)-\frac{\mathcal{C}_m}{n}\right]\right|\geq C\right)\\ \leq\ &2V\mathbb{P}\left(\left|P\left(\gamma\left(\widehat{s}_m^{(-1)}\right)-\gamma\left(s_m\right)\right)-\frac{\mathcal{C}_m}{n}\right|\geq\frac{C}{2}\right).\end{aligned}$$

Hence, from Theorem 2 of [36] applied with $\alpha=2+\alpha_{\mathcal{M}}$ and sample size equal to $n_V=nV/\left(V-1\right)$, we get that by taking

$$C=2\varepsilon_{n_V}\left(m\right)\frac{\mathcal{C}_m}{n_V}\leq L_{(\mathbf{GSA}),r}\varepsilon_n\left(m\right)\mathbb{E}\left[\mathrm{p}_2^{(-1)}\left(m\right)\right],$$

it holds

$$\mathbb{P}\left(\left|P\left(\gamma\left(\widehat{s}_m^{(-1)}\right)-\gamma\left(s_m\right)\right)-\frac{1}{V}\sum_{j=1}^{V} P\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma\left(s_m\right)\right)\right|\geq C\right)\leq 10Vn^{-2-\alpha_{\mathcal{M}}}. \tag{58}$$

Inequality (55) now follows from combining (57) with (58) and noticing that $\varepsilon_n^{(1)}\left(m\right)\leq L_{(\mathbf{GSA}),r}\varepsilon_n\left(m\right)$. Inequality (56) also derives from Lemma 7.3 with the same type of reasoning and further details are left to the reader. ■

Lemma 7.5 *Let $\alpha>0$. Assume that (**GSA**) is satisfied and that $1<V\leq r$. Then by setting*

$$\bar{\delta}\left(m\right)=\frac{1}{V}\sum_{j=1}^{V}\left(P_n^{(j)}-P\right)\left(\gamma\left(s_m\right)-\gamma\left(s_*\right)\right)\quad\text{and}\quad\bar{\delta}'\left(m\right)=\frac{1}{V}\sum_{j=1}^{V}\left(P_n^{(-j)}-P\right)\left(\gamma\left(s_m\right)-\gamma\left(s_*\right)\right),$$

we have for all $m\in\mathcal{M}_n$,

$$\mathbb{P}\left(\max\left\{\left|\bar{\delta}\left(m\right)\right|,\left|\bar{\delta}'\left(m\right)\right|\right\}\geq L_{(\mathbf{GSA}),r}\left(\sqrt{\frac{\ell\left(s_*,s_m\right)\ln n}{n}}+\frac{\ln n}{n}\right)\right)\leq 2rn^{-\alpha}\ . \tag{59}$$

Furthermore, for all $m\in\mathcal{M}_n$ such that $A_{\mathcal{M},+}\left(\ln n\right)^2\leq D_m$ and for all $n\geq n_0\left((\mathbf{GSA}),\alpha\right)$, we have

$$\mathbb{P}\left(\max\left\{\left|\bar{\delta}\left(m\right)\right|,\left|\bar{\delta}'\left(m\right)\right|\right\}\geq\frac{\ell\left(s_*,s_m\right)}{\sqrt{D_m}}+L_{(\mathbf{GSA}),r}\frac{\ln n}{\sqrt{D_m}}\mathbb{E}\left[\mathrm{p}_2^{(-1)}\left(m\right)\right]\right)\leq 2rn^{-\alpha},$$

where $\mathrm{p}_2^{(-1)}(m) := P_n^{(-1)}\left(\gamma(s_m) - \gamma\left(\widehat{s}_m^{(-1)}\right)\right) \geq 0$.

Proof. Notice that for any $C \geq 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\bar{\delta}(m)\right| \geq C\right) &\leq \mathbb{P}\left(\max_{j\in\{1,\dots,V\}}\left|\left(P_n^{(j)} - P\right)(\gamma(s_m) - \gamma(s_*))\right| \geq C\right) \\ &\leq \sum_{j=1}^{V}\mathbb{P}\left(\left|\left(P_n^{(j)} - P\right)(\gamma(s_m) - \gamma(s_*))\right| \geq C\right).\end{aligned}$$

Then use j times Lemma 5 of [37] with a sample size equal to n/V in order to control the summands at the right-hand side of the inequality in the last display. The same reasoning holds for $\left|\bar{\delta}'(m)\right|$. Further details are left to the reader. ■

Lemma 7.6 *Let $\alpha > 0$. Consider a finite-dimensional linear model m of linear dimension D and assume that $(\varphi_k)_{k=1}^{D_m}$ is a localized orthonormal basis of $(m, \|\cdot\|_2)$ with index of localization $r_m > 0$. More explicitly, we thus assume that for all $\beta = (\beta_k)_{k=1}^{D_m} \in \mathbb{R}^{D_m}$,*

$$\left\|\sum_{k=1}^{D_m}\beta_k\varphi_k\right\|_\infty \leq r_m\sqrt{D_m}\,|\beta|_\infty.$$

*If (**Ab**(m)) given in Theorem 5.1 holds and if for some positive constant A_+,*

$$D_m \leq A_+\frac{n}{(\ln n)^2},$$

then there exists a positive constant $L^{(2)}_{\alpha,r_m}$ such that for all $n \geq n_0(A_+)$, we have

$$\mathbb{P}\left(\max_{(k,l)\in\{1,\dots,D_m\}^2}|(P_n - P)(\varphi_k\cdot\varphi_l)| \geq L^{(2)}_{\alpha,r_m}\min\left\{\|\varphi_k\|_\infty;\|\varphi_l\|_\infty\right\}\sqrt{\frac{\ln n}{n}}\right) \leq n^{-\alpha}. \tag{60}$$

Proof. For any $(k,l) \in \{1,\dots,D_m\}^2$, we have

$$\mathbb{E}\left[(\varphi_k\cdot\varphi_l)^2(X)\right] \leq \min\left\{\|\varphi_k\|_\infty^2;\|\varphi_l\|_\infty^2\right\}$$

and

$$\begin{aligned}\|\varphi_k\cdot\varphi_l\|_\infty &\leq \min\left\{\|\varphi_k\|_\infty;\|\varphi_l\|_\infty\right\}\times\max\left\{\|\varphi_k\|_\infty;\|\varphi_l\|_\infty\right\} \\ &\leq \min\left\{\|\varphi_k\|_\infty;\|\varphi_l\|_\infty\right\}\times r_m\sqrt{D_m}.\end{aligned}$$

Hence, we apply Bernstein's inequality (see Proposition 2.9 in [32]) and we get, for all $\gamma > 0$,

$$\mathbb{P}\left(|(P_n - P)(\varphi_k\cdot\varphi_l)| \geq \min\left\{\|\varphi_k\|_\infty;\|\varphi_l\|_\infty\right\}\left(\sqrt{\frac{2\gamma\ln n}{n}} + \frac{r_m\sqrt{D_m}\gamma\ln n}{3n}\right)\right) \leq 2n^{-\gamma}. \tag{61}$$

Since, for all $n \geq n_0(A_+)$,

$$\frac{r_m\sqrt{D_m}\ln n}{n} \leq \frac{r_m\sqrt{A_+}}{\sqrt{\ln n}}\cdot\sqrt{\frac{\ln n}{n}} \leq r_m\sqrt{\frac{\ln n}{n}},$$

we get from (61) that for all $n \geq n_0(A_+)$,

$$\begin{aligned}&\mathbb{P}\left(\max_{(k,l)\in\{1,\dots,D_m\}^2}|(P_n - P)(\varphi_k\cdot\varphi_l)| \geq \left(\sqrt{2\gamma} + \frac{\gamma r_m}{3}\right)\min\left\{\|\varphi_k\|_\infty;\|\varphi_l\|_\infty\right\}\sqrt{\frac{\ln n}{n}}\right) \\ \leq &\sum_{(k,l)\in\{1,\dots,D_m\}^2}\mathbb{P}\left(|(P_n - P)(\varphi_k\cdot\varphi_l)| \geq \left(\sqrt{2\gamma} + \frac{\gamma r_m}{3}\right)\min\left\{\|\varphi_k\|_\infty;\|\varphi_l\|_\infty\right\}\sqrt{\frac{\ln n}{n}}\right) \\ \leq &\sum_{(k,l)\in\{1,\dots,D_m\}^2}\mathbb{P}\left(|(P_n - P)(\varphi_k\cdot\varphi_l)| \geq \min\left\{\|\varphi_k\|_\infty;\|\varphi_l\|_\infty\right\}\sqrt{\frac{2\gamma\ln n}{n}} + \frac{r_m\sqrt{D_m}\gamma\ln n}{3n}\right) \\ \leq &\ 2D^2n^{-\gamma} \leq n^{-\gamma+2}.\end{aligned} \tag{62}$$

We deduce from (62) that (60) holds with $L^{(2)}_{\alpha,r_m} = \sqrt{2\alpha+4} + (\alpha+2)\,r_m/3 > 0$. ■

Lemma 7.7 *Under the assumptions of Lemma 7.6 there exists a positive constant* $L^{(1)}_{A,r_m,\alpha}$ *such that for all* $n \geq n_0(A_+)$*, we have*

$$\mathbb{P}\left(\max_{k\in\{1,\dots,D_m\}} |(P_n - P)(\psi_m \cdot \varphi_k)| \geq L^{(1)}_{A,r_m,\alpha}\sqrt{\frac{\ln n}{n}}\right) \leq n^{-\alpha},$$

where $\psi_m(x,y) = -2(y - s_m(x))$.

Proof. Let $\beta > 0$. Notice that by $(\mathbf{Ab}(m))$,

$$|\psi_m(X,Y)| \leq 4A \quad a.s.$$

Then by Bernstein's inequality, we get by straightforward computations (in the spirit of the proof of Lemma 7.6) that there exists $L(1)_{A,r_m,\beta} > 0$ such that, for all $k \in \{1,\dots,D_m\}$,

$$\mathbb{P}\left(|(P_n - P)(\psi_m \cdot \varphi_k)| \geq L^{(1)}_{A,r_m,\beta}\sqrt{\frac{\ln n}{n}}\right) \leq n^{-\beta}.$$

Now the result follows from a simple union bound with $\beta = \alpha + 1$. ■

7.6 Additional simulation results

This section provides additional simulation results to those in Section 6. Figure 5 is an analogy to Figure 4 which illustrates the difference between the test functions *Spikes* and *Wave* for a smaller sample size (i.e. $n = 1024$).

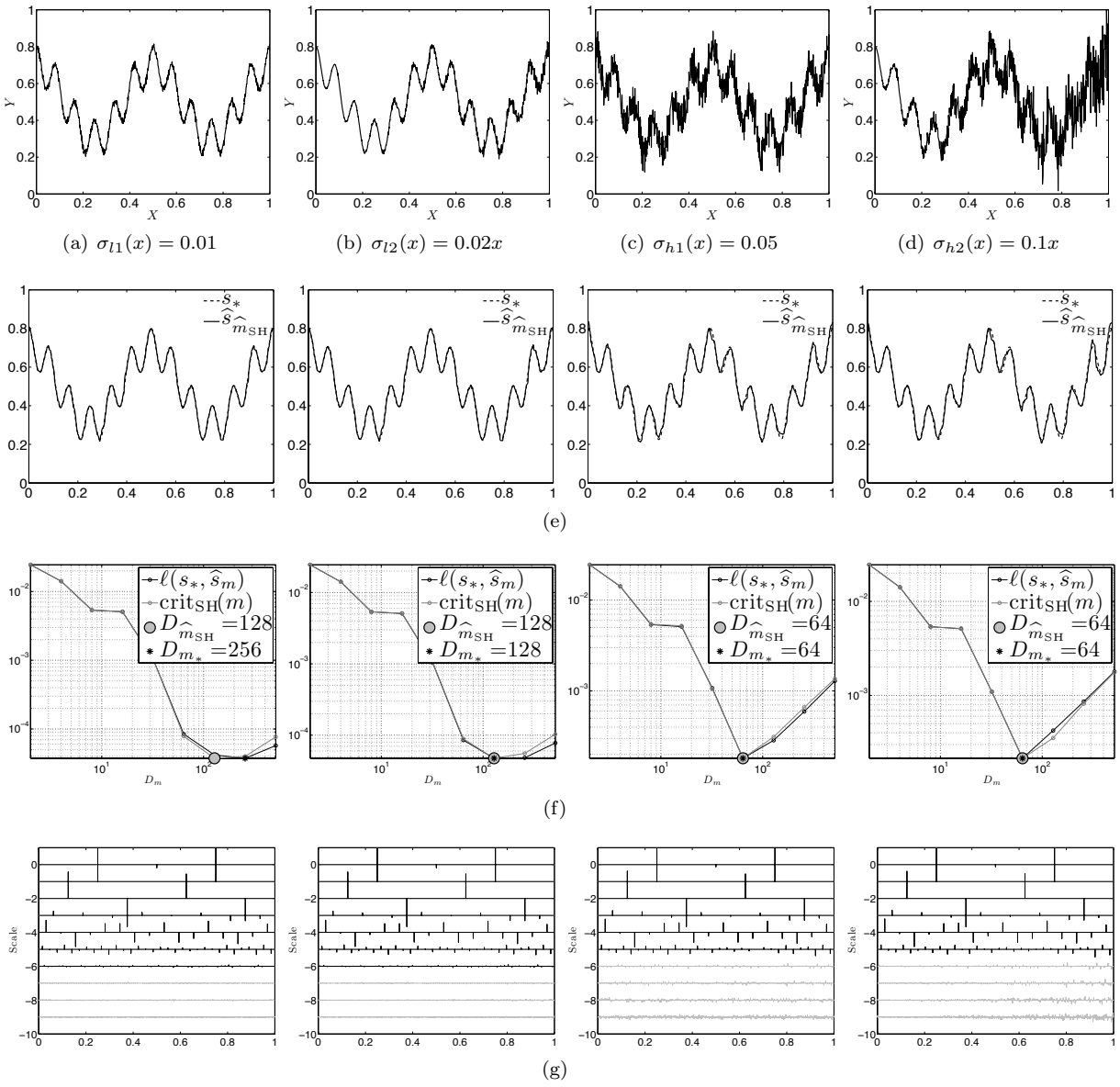


Figure 5: (a)-(d): Noisy version of *Wave* for each $\sigma(\cdot)$ scenarios. (e): Typical reconstructions from a single simulation with $n = 1024$. The dotted line is the true signal and the solid one depicts the estimates $\widehat{s}_{\widehat{m}_{\mathrm{SH}}}$. (f): Graph of the excess risk $\ell(s_*, \widehat{s}_m)$ against the dimension D_m and (shifted) $\mathrm{crit}_{\mathrm{SH}}(m)$ (in a log-log scale). The gray circle represents the global minimizer $\widehat{m}$ of $\mathrm{crit}_{\mathrm{SH}}(m)$ and the black star the oracle model m_*. (g): Noisy and selected (black) wavelet coefficients (see Figure 3(e) for a visual comparison with the original wavelet coefficients).