HAL
open science

Efficiency of the V-fold model selection for localized bases

Fabien Navarro, Adrien Saumard

To cite this version:

Fabien Navarro, Adrien Saumard. Efficiency of the V-fold model selection for localized bases. Conference of the International Society for Non-Parametric Statistics, Jun 2016, Avignon, France. 10.1007/978-3-319-96941-1_4 . hal-01505514v2

HAL Id: hal-01505514

https://hal.science/hal-01505514v2

Submitted on 16 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Efficiency of the V-fold model selection for localized bases

Fabien Navarro and Adrien Saumard

Abstract Many interesting functional bases, such as piecewise polynomials or wavelets, are examples of localized bases. We investigate the optimality of V-fold cross-validation and a variant called V-fold penalization in the context of the selection of linear models generated by localized bases in a heteroscedastic framework. It appears that while V-fold cross-validation is not asymptotically optimal when V is fixed, the V-fold penalization procedure is optimal. Simulation studies are also presented.

Key words: V-fold cross-validation, V-fold penalization, model selection, non-parametric regression, heteroscedastic noise, random design, wavelets

1 Introduction

V-fold cross-validation type procedures are extremely used in Statistics and machine learning, with however a rather small set of theoretical results on it ([3]). This paper aims at investigating from the theoretical point of view and on simulations, the efficiency of two V-fold strategies for model selection in a heteroscedastic regression setting, with random design. From the one hand, we investigate the behaviour of the classical V-fold cross-validation to select, among other examples, linear models of wavelets. As pointed out in the case of histogram selection in [2], this procedure is not asymptotically optimal when V is fixed, as it is the case in practice where V is usually taken to be equal to 5 or 10. On the other hand, we study the V-fold penalization proposed by Arlot [2] and show its efficiency in our general context.

Fabien Navarro
CREST-ENSAI-UBL, BRUZ, FRANCE e-mail: fabien.navarro@ensai.fr

Adrien Saumard
CREST-ENSAI-UBL, BRUZ, FRANCE e-mail: adrien.saumard@ensai.fr

More precisely, the present paper is devoted to an extension of some results obtained in [16] related to efficiency of cross-validation type procedures. Indeed, as remarked in [16] (see Remark 5.1 therein) our results obtained for the selection of linear models endowed with a strongly localized basis (see Definition (**Aslb**), Section 2.1 of [16]) can be extended to more general and more classical localized bases, at the price of considering only models with sufficiently small dimensions. Rigorous proofs are given here and further simulation studies are explored.

The paper is organized as follows. In Section 2, we describe our model selection setting. Then V-fold cross-validation is considered in Section 3, while the efficiency of V-fold penalization is tackled in Section 4. A simulation study is reported in Section 5. The proofs are exposed in Section 6.

2 Model selection setting

Assume that we observe n independent pairs of random variables $\xi_i = (X_i, Y_i) \in \mathscr{X} \times \mathbb{R}$ with common distribution P. For convenience, we also denote by $\xi = (X, Y)$ a random pair, independent of the sample $(\xi_1, ..., \xi_n)$, following the same distribution P. The set $\mathscr{X}$ is called the feature space and we assume $\mathscr{X} \subset \mathbb{R}^d$, $d \geq 1$. We denote P^X the marginal distribution of the design X. We assume that the following regression relation is valid,

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

with $s_* \in L_2\left(P^X\right)$ the regression function that we aim at estimating. Conditionally to X, the residual ε is normalized, i.e. it has mean zero and variance one. The function $\sigma : \mathscr{X} \to \mathbb{R}_+$ is a heteroscedastic noise level, assumed to be unknown.

To produce an estimator of s_*, we are given a finite collection of models $\mathscr{M}_n$, with cardinality depending on the amount n of data. Each model $m \in \mathscr{M}_n$ is taken to be a finite-dimensional vector space, of linear dimension D_m. We will further detail in a few lines the analytical structure of the models.

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

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

Using these notations, the regression function s_* is the unique minimizer of the risk,

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

The projections s_m are also characterized by

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

To each model $m \in \mathscr{M}_n$, we associate a least squares estimator $\hat{s}_m$, defined by

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

where $P_n = n^{-1} \sum_{i=1}^{n} \delta_{\xi_i}$ is the empirical measure associated to the sample.

The accuracy of estimation is tackled through the excess loss of the estimators,

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

The following "bias-variance" decomposition holds,

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

where

$$\begin{aligned}\ell(s_*, s_m) &:= P(\gamma(s_m) - \gamma(s_*)) = \|s_m - s_*\|_2^2 \\ \ell(s_m, \hat{s}_m) &:= P(\gamma(\hat{s}_m) - \gamma(s_m)) \geq 0 \ .\end{aligned}$$

The deterministic quantity $\ell(s_*, s_m)$ is called the bias of the model m, while the random variable $\ell(s_m, \hat{s}_m)$ is called the excess loss of the least squares estimator $\hat{s}_m$ on the model m. By the Pythagorean Theorem, we have

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

From the collection of models $\mathscr{M}_n$, we aim at proposing an estimator that is as close as possible in terms of excess loss to an oracle model m_*, defined by

$$m_* \in \arg\min_{m \in \mathscr{M}_n} \{\ell(s_*, \hat{s}_m)\} \ .$$

We choose to select an estimator from the collection $\{\hat{s}_m \ ; \ m \in \mathscr{M}_n\}$. Hence, the selected model is denoted $\hat{m}$. The goal is to ensure that the selected estimator achieves an oracle inequality of the form

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

for a constant $C \geq 1$ as close as possible to one and on an event of probability close to one.

3 V-fold cross-validation

For convenience, let us denote in the following $\widehat{s}_m(P_n)$ the least squares estimator built from the empirical distribution $P_n = 1/n\sum_{i=1}^n \delta_{(X_i,Y_i)}$. To perform the V-fold cross-validation (VFCV) procedure, we consider a partition $(B_j)_{1\leq j\leq V}$ of the index set $\{1,...,n\}$ and set,

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

We assume that the partition $(B_j)_{1\leq j\leq V}$ is regular: for all $j\in\{1,...,V\}$, $\mathrm{Card}(B_j) = n/V$. It is worth noting that we can always define our partition such $\sup_j \left|\mathrm{Card}(B_j) - n/V\right| < 1$ so that the assumption of regular partition is only a slight approximation of the general case. Let us write $\widehat{s}_m^{(-j)} = \widehat{s}_m\left(P_n^{(-j)}\right)$ the estimators built from the data in the block B_j. Now, the selected model $\widehat{m}_{\mathrm{VFCV}}$ is taken equal to any model optimizing the V-fold criterion,

$$\widehat{m}_{\mathrm{VFCV}} \in \arg\min_{n\in\mathscr{M}_n} \left\{\mathrm{crit}_{\mathrm{VFCV}}(m)\right\} \ , \tag{1}$$

where

$$\mathrm{crit}_{\mathrm{VFCV}}(m) = \frac{1}{V}\sum_{j=1}^{V} P_n^{(j)} \gamma\left(\widehat{s}_m^{(-j)}\right) \ . \tag{2}$$

Let us now detail the set of assumptions under which we will investigate the accuracy of VFCV.

Set of assumptions: (SA)

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

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

$$\left\|\sum_{k=1}^{D_m} \beta_k \varphi_k\right\|_\infty \leq r_{\mathscr{M}} \sqrt{D_m}\, |\beta|_\infty \ , \tag{3}$$

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

(**P2**) Upper bound on dimensions of models in $\mathscr{M}_n$: there exists a positive constant $A_{\mathscr{M},+}$ such that for every $m\in\mathscr{M}_n$, $D_m \leq A_{\mathscr{M},+} n^{1/3} (\ln n)^{-2}$.

(**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 $\mathscr{M}_n$: $|Y_i| \leq A < \infty$, $\|s_m\|_\infty \leq A < \infty$ for all $m\in\mathscr{M}_n$.

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

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

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

Assumption (**Alb**) refers to the classical concept of localized basis (Birgé and Massart [6]). It is proved in [5], Section 3.2.1, that linear models of piecewise polynomials with bounded degree on a regular partition of a bounded domain of $\mathbb{R}^d$ are endowed with a localized basis. It is also proved that compactly supported wavelet expansions are also fulfilled with a localized basis on $\mathbb{R}^d$. However, the Fourier basis is not a localized basis. For some sharp concentration results related to the excess loss of least squares estimators built from the Fourier basis, we refer to [19].

The assumption (**Alb**) is more general than the assumption of strongly localized basis used in [16], but the price to pay for such generality is that, according to (**P2**) we can only consider models with dimensions $D_m << n^{1/3}$.

Assumption (**P1**) states that the collection as a polynomial cardinality with respect to the sample size, allowing in particular to consider a collection of models built from a basis expansion.

Then Assumption (**Ab**) is related to boundedness of the data and enables in particular to use Talagrand's type concentration inequalities for the empirical process. Going beyond the bounded setting would in particular bring much more technicalities that might darken our work. For an example of results in an unbounded setting, see for instance [4], dealing with optimal selection of regressograms (histograms being a very particular case of our general framework). Assumption (**An**) is essentially a technical assumption that allows to obtain sharp lower bounds for the excess losses of the estimators. Condition (Ap_u) is a very classical assumption in the model selection literature, specifying a rate of decay for the biases of the models. This assumption is classically satisfied for piecewise polynomials when the regression function belongs to a Sobolew space and for wavelet models whenever the target belongs to some Besov space (see for instance [5] for more details). The specific value of β_+ parameter will only affect the value of the constants in the derived oracle inequalities.

Theorem 1. *Assume that (**SA**) holds. Let $r \in (2, +\infty)$ and $V \in \{2, ..., n-1\}$ satisfying $1 < V \leq r$. Define the V-fold cross-validation procedure as the model selection procedure given by (1). Then, for all $n \geq n_0\left((\boldsymbol{SA}), r\right)$, with probability at least $1 - L_{(\boldsymbol{SA}),r} n^{-2}$,*

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

In Theorem 1, we prove an oracle inequality with principal constant tending to one when the sample size goes to infinity. This inequality bounds from above the excess loss of the selected estimator by the excess loss of the oracle learned with a fraction $1 - V^{-1}$ of the original data. Ideally, one would, however, expect from an optimal procedure to recover the oracle built from the entire data. The next section is devoted to this task.

Parameter V (or r) is considered in Theorem 1 as a constant, essentially for ease of presentation. Actually, the value of V may be allowed to depend on n but also on

the dimensions D_m, meaning that we may take different values of V according to the different models of the collection. More precisely, it can be seen from the arguments in the proofs (especially from Theorem 8 in [18]) that for each model $m \in \mathscr{M}_n$, it suffices to have $V \leq \max\{D_m(\ln n)^{-\tau}; 2\}$ where τ is any number in $(1,3)$ to ensure an oracle inequality with leading constant tending to one when the amount of data tends to infinity. In this case, r can not be considered as a parameter independent from the sample size anymore, but it can be checked that for the latter constraints on V, the constants $n_0((\mathbf{SA}), r)$ and $L_{(\mathbf{SA}),r}$ do not explode but are still uniformly bounded with respect to n and thus can be still considered as independent from n.

4 V-fold penalization

Now we investigate the behaviour of a penalization procedure proposed by Arlot [2] and called V-fold penalization,

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

where

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

with

$$\mathrm{pen}_{\mathrm{VF}}(m) = \frac{V-1}{V} \sum_{j=1}^{V} \left[P_n \gamma\left(\widehat{s}_m^{(-j)}\right) - P_n^{(-j)} \gamma\left(\widehat{s}_m^{(-j)}\right) \right] . \tag{4}$$

The property underlying the V-fold penalization is that the V-fold penalty $\mathrm{pen}_{\mathrm{VF}}$ is an unbiased estimate of the ideal penalty $\mathrm{pen}_{\mathrm{id}}$, the latter allowing to identify the oracle m_*,

$$\begin{aligned} m_* &\in \arg\min_{m \in \mathscr{M}_n} \{P(\gamma(\widehat{s}_m))\} \\ &= \arg\min_{m \in \mathscr{M}_n} \{P_n(\gamma(\widehat{s}_m)) + \mathrm{pen}_{\mathrm{id}}(m)\} , \end{aligned}$$

where

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

The following theorem states the asymptotic optimality of the V-fold penalization procedure for a fixed V.

Theorem 2. *Assume that (**SA**) holds. Let $r \in (2, +\infty)$ and $V \in \{2, \ldots, n-1\}$ satisfying $1 < V \leq r$. Define the V-fold cross-validation procedure as the model selection procedure given by,*

$$\widehat{m}_{penVF} \in \arg\min_{n \in \mathscr{M}_n} \{P_n(\gamma(\widehat{s}_m)) + \mathrm{pen}_{VF}(m)\} .$$

Then, for all $n \geq n_0((\boldsymbol{SA}), r)$, with probability at least $1 - L_{(\boldsymbol{SA}),r} n^{-2}$,

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

As for Theorem 1 above, parameter V (or r) is considered in Theorem 2 as a constant but in fact, the value of V may be allowed to depend on n and even on the dimensions D_m, this case corresponding to possibly different choices V according to the models of the collection. As for Theorem 1, it is allowed to have $V \leq \max\{D_m(\ln n)^{-\tau}; 2\}$ where τ is any number in $(1,3)$ to ensure an oracle inequality with leading constant tending to one when the amount of data tends to infinity.

5 Simulation study

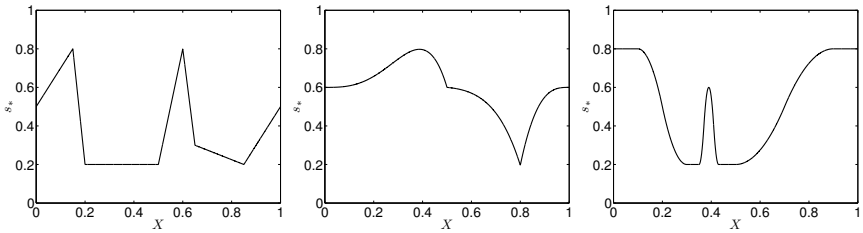
In order to assess the numerical performances of the model selection procedures we have discussed, a short simulation study was conducted. Particularly, to illustrate the theory developed above for the selection of linear estimators using the V-fold cross-validation and V-fold penalization, linear wavelet models were considered.

Despite the fact that a linear wavelet estimator is not as flexible, or potentially as powerful, as a nonlinear one, it still preserves the computational efficiency of wavelet methods and can provide comparative results to thresholding estimator, particularly when the unknown function is sufficiently smooth (see [1]).

The simulations were carried out using Matlab and the wavelet toolbox Wavelab850 [10]. The codes used to replicate the numerical results presented here will be available at https://github.com/fabnavarro. For more details on the numerical simulations and comparisons with other model selection procedures, we refer the reader to [19].

The simulated data were generated according to $Y_i = s_*(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \ldots, n$, where $n = 4096$, X_i's are uniformly distributed on $[0,1]$, ε_i's are independent $\mathscr{N}(0,1)$ variables and independent of X_i's. The heteroscedastic noise level $\sigma(x) = |\cos(10x)|/10$. Daubechies' compactly-supported wavelet with 8 vanishing moments were used. Three standard regression functions with different degrees of smoothness (*Angle, Corner* and *Parabolas*, see [14, 7]) were considered. They are plotted in Figure 1 and a visual idea of the noise level is given in Figures 2(b).

The computation of wavelet-based estimators is straightforward and fast in the fixed design case, thanks to Mallat's pyramidal algorithm ([13]). In the case of random design, the implementation requires some changes and several strategies have been developed in the literature (see e.g. [8, 11]). In the regression with uniform design [9] 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). We have followed this approach since it preserves the computational simplicity and efficiency of the equispaced algorithm. In the context of wavelet regression in random design with heteroscedastic dependent errors [12]



(a) *Angle* (b) *Corner* (c) *Parabolas*

Fig. 1 (a)–(c): The three test functions used in the simulation study.

have also adopted this approach. Thus, the wavelet coefficients of the collection of models is computed by a simple application of Mallat's algorithm using the ordered Y_i's as input variables. The collection is then constructed by successively adding whole resolution levels of wavelet coefficients. Thus, the considered dimensions are $\{D_m, m \in \mathscr{M}_n\} = \{2^j, j = 1, \ldots, J-1\}$, where $J = \log 2(n)$ (the finest resolution level). Finally, the selected model are obtained by minimizing (2) and (4) over the set $m \in \mathscr{M}_n$. Note that these linear models operate in a global fashion since whole levels of coefficients are suppressed as opposed to thresholding methods.

For choosing the threshold parameter in wavelet shrinkage Nason [15] adjusted the usual 2FCV method—which cannot be applied directly to wavelet estimation. In order to implement its strategy in a linear 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 $\widehat{m}_{\mathrm{2FCV}}$ (resp. $\widehat{m}_{\mathrm{pen2F}}$) is obtained by minimizing (2) (resp. (4)) with $V = 2$, $B_1 = \{2, 4, \ldots, n\}$ and $B_2 = \{1, 3, \ldots, n-1\}$.

For one Monte Carlo simulation with a sample size $n = 4096$, we display the estimation results in Figure 2(b). Plots of the excess risk $\ell(s_*, \widehat{s}_m)$ against the dimension D_m are plotted in Figure 2(c). The curve $\mathrm{crit}_{\mathrm{2FCV}}(m)$ and $\mathrm{crit}_{\mathrm{pen2F}}(m)$ are also displayed in Figure 2(c). It can be observed that $\mathrm{crit}_{\mathrm{2FCV}}(m)$ and $\mathrm{crit}_{\mathrm{pen2F}}(m)$ give very reliable estimate for the risk $\ell(s_*, \widehat{s}_m)$, and in turn, also a high-quality estimate of the optimal model. Indeed, in this case, both methods consistently select the oracle model m_*.

6 Proofs

As a preliminary result, let us first prove the consistency in sup-norm of our least squares estimators. This is in fact the main change compared to the strongly localized case treated in [16].

Theorem 3. *Let $\alpha > 0$. Assume that m is a linear vector space satisfying Assumption **(Alb)** and use the notations given in the statement of **(Alb)**. Assume also that Assumption **(Ab)** holds. If there exists $A_+ > 0$ such that*

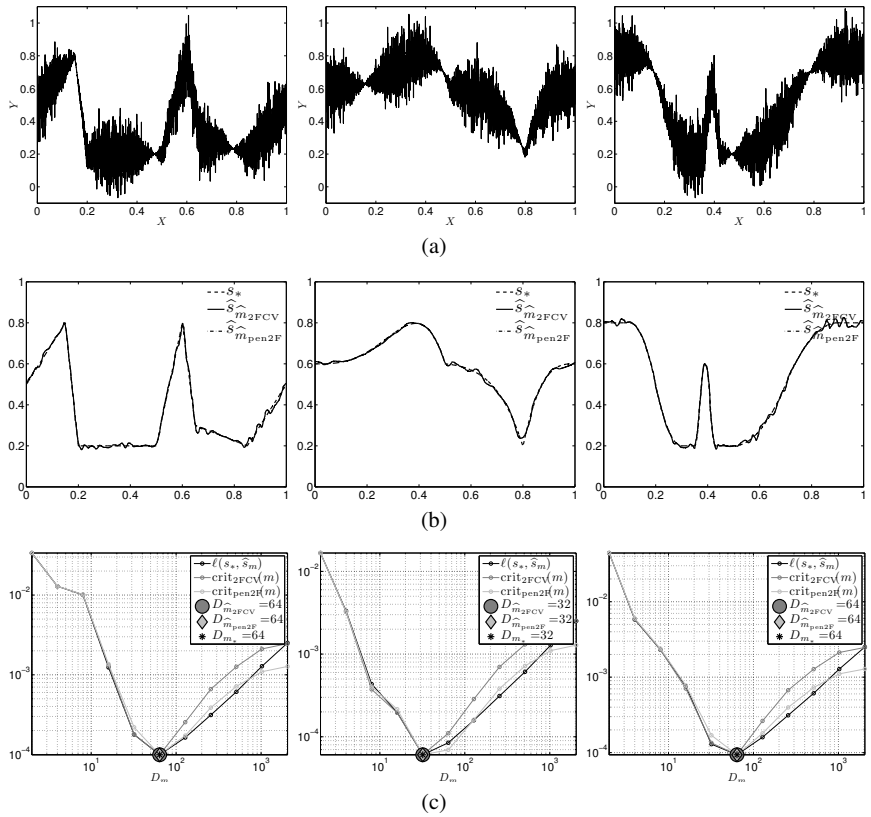


Fig. 2 (a): Noisy observations. (b):Typical reconstructions from a single simulation with $n = 4096$. Dashed line indicates the true function s_*, solid line corresponds to the estimates $\widehat{s}_{\widehat{m}_{\mathrm{2FCV}}}$ and dashed-dotted line to $\widehat{s}_{\widehat{m}_{\mathrm{pen2F}}}$. (c): Graph of the excess risk $\ell(s_*, \widehat{s}_m)$ (black) against the dimension D_m and (rescaled) $\mathrm{crit}_{\mathrm{2FCV}}(m)$ (gray) and $\mathrm{crit}_{\mathrm{pen2F}}(m)$ (light-gray) (in a log-log scale). The gray circle represents the global minimizer $\widehat{m}_{\mathrm{2FCV}}$ of $\mathrm{crit}_{\mathrm{2FCV}}(m)$, the light-gray diamond corresponds to the global minimizer $\widehat{m}_{\mathrm{2FCV}}$ of $\mathrm{crit}_{\mathrm{pen2F}}(m)$ and the black star the oracle model m_*.

$$D_m \leq A_+ \frac{n^{1/3}}{(\ln n)^2} \ ,$$

then there exists a positive constant $L_{A,r_{\mathscr{M}},\alpha}$ such that, for all $n \geq n_0(r_{\mathscr{M}}, \alpha)$,

$$\mathbb{P}\left(\|\hat{s}_m - s_m\|_\infty \geq L_{A,r_{\mathscr{M}},\alpha} \sqrt{\frac{D_m \ln n}{n}} \right) \leq n^{-\alpha} \ .$$

Proof (Proof of Theorem 3). Let $C > 0$. Set

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

and

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

Take an orthonormal basis $(\varphi_k)_{k=1}^{D_m}$ of $(m, \|\cdot\|_2)$ satisfying (**Alb**). By Lemma 19 of [16], we get that there exists $L_{A,r_m,\alpha}^{(1)} > 0$ such that, by setting

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

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

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

where $L_{\alpha,r_m}^{(2)}$ is defined in Lemma 18 of [16]. By Lemma 18 of [16], 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 $n \geq n_0(A_+)$,

$$\begin{aligned}
&\mathbb{P}(\|s_n - s_m\|_\infty > C) \\
\leq &\mathbb{P}\left(\inf_{s\in\mathscr{F}_{>C}^{\infty}} P_n(\gamma(s) - \gamma(s_m)) \leq \inf_{s\in\mathscr{F}_C^{\infty}} P_n(\gamma(s) - \gamma(s_m)) \right) \\
= &\mathbb{P}\left(\sup_{s\in\mathscr{F}_{>C}^{\infty}} P_n(\gamma(s_m) - \gamma(s)) \geq \sup_{s\in\mathscr{F}_C^{\infty}} P_n(\gamma(s_m) - \gamma(s)) \right) \\
\leq &\mathbb{P}\left(\left\{ \sup_{s\in\mathscr{F}_{>C}^{\infty}} P_n(\gamma(s_m) - \gamma(s)) \geq \sup_{s\in\mathscr{F}_{C/2}^{\infty}} P_n(\gamma(s_m) - \gamma(s)) \right\} \bigcap \Omega_1 \bigcap \Omega_2 \right) + 2n^{-\alpha} \,.
\end{aligned} \tag{5}$$

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 $(k,l) \in \{1,...,D_m\}^2$,

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

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

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

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

$$P_n(\gamma(s_m) - \gamma(s)) = h_n(\beta) \ . \tag{6}$$

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

$$|\beta|_{m,\infty} = r_m \sqrt{D_m} |\beta|_\infty \ .$$

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

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

We have, for any $B = \left(B_{k,l}\right)_{k,l=1,...,D_m} \in \mathbb{R}^{D_m \times D_m}$, the following classical formula

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

Notice that by inequality (3) of (**Alb**), it holds

$$\mathscr{F}_{>C}^{\infty} \subset \left\{ s \in m \ ; \ s - s_m = \sum_{k=1}^{D_m} \beta_k \varphi_k \ \ \& \ \ |\beta|_{m,\infty} \geq C \right\} \tag{7}$$

and

$$\mathscr{F}_{C/2}^{\infty} \supset \left\{ s \in m \ ; \ s - s_m = \sum_{k=1}^{D_m} \beta_k \varphi_k \ \ \& \ \ |\beta|_{m,\infty} \leq C/2 \right\} \ . \tag{8}$$

Hence, from (5), (6) (8) and (7) 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 we will get

$$\mathbb{P}\left(\|\widehat{s}_m - s_m\|_\infty > C\right) \leq 2n^{-\alpha} \ .$$

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

$$\frac{\partial h_n}{\partial \beta_k}(\beta) = R_{n,k}^{(1)} - 2\sum_{i=1}^{D_m} \beta_i R_{n,k,i}^{(2)} - 2\beta_k \tag{9}$$

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

$$\frac{\partial h_n}{\partial \beta_k}(\beta) = 0 \ , \forall k \in \{1,...,D_m\} \ . \tag{10}$$

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

$$R_n^{(2)} := \left(R_{n,k,l}^{(2)}\right)_{k,l=1,...,D_m}$$

and by (9), the system given in (10) can be written

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

where $R_n^{(1)}$ is a D_m-dimensional vector defined by

$$R_n^{(1)} = \left(R_{n,k}^{(1)}\right)_{k=1,...,D_m} \ .$$

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,

$$\begin{aligned}
\left\|R_n^{(2)}\right\|_m &= \max_{k\in\{1,...,D_m\}} \left\{ \left\{ \sum_{l\in\{1,...,D_m\}} |(P_n - P)(\varphi_k \cdot \varphi_l)| \right\} \right\} \\
&\leq L_{\alpha,r_m}^{(2)} \max_{k\in\{1,...,D_m\}} \left\{ \left\{ \sum_{l\in\{1,...,D_m\}} \min\{\|\varphi_k\|_\infty ; \|\varphi_l\|_\infty\} \sqrt{\frac{\ln n}{n}} \right\} \right\} \\
&\leq r_m L_{\alpha,r_m}^{(2)} \sqrt{\frac{D_m^3 \ln n}{n}}
\end{aligned} \tag{11}$$

Hence, from (11) and the fact that $D_m \leq A_+ \frac{n^{1/3}}{(\ln n)^2}$, we get that for all $n \geq n_0(r_m, \alpha)$, it holds on Ω_2,

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

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

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

Now, on Ω_1 we have,

$$\left|R_n^{(1)}\right|_{m,\infty} \leq r_m\sqrt{D_m} \max_{k\in\{1,\ldots,D_m\}} \left|\left(P_n - P\right)\left(\psi_m \cdot \varphi_k\right)\right| \leq r_m L_{A,r_m,\alpha}^{(1)} \sqrt{\frac{D_m \ln n}{n}}$$

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

$$\left|\beta^{(n)}\right|_{m,\infty} \leq \frac{1}{2}\left\|\left(I_d + R_n^{(2)}\right)^{-1}\right\|_m \left|R_n^{(1)}\right|_{m,\infty} \leq r_m L_{A,r_m,\alpha}^{(1)} \sqrt{\frac{D_m \ln n}{n}} . \qquad (12)$$

Moreover, by the formula (6) we have

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

and we thus see that h_n is concave. Hence, for all $n_0\left(r_m,\alpha\right)$, we get that on Ω_2, $\beta^{(n)}$ is the unique maximum of h_n and on $\Omega_2 \bigcap \Omega_1$, by (12), 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.

From Theorem 2 of [17] and Theorem 3 above, we deduce the following excess risks bounds.

Theorem 4. *Let $A_+, A_-, \alpha > 0$. Assume that m is a linear vector space of finite dimension D_m satisfying (**Alb(m)**) and use notations of (**Alb(m)**). Assume, moreover, that the following assumption holds:*

(Ab(m)) There exists a constant $A > 0$, such that $\|s_m\|_\infty \leq A$ and $|Y| \leq A$ a.s.

If it holds

$$A_- \left(\ln n\right)^2 \leq D_m \leq A_+ \frac{n^{1/3}}{\left(\ln n\right)^2} ,$$

then a positive constant A_0 exists, only depending on α, A_- and on the constants $A, \sigma_{\min}$ and r_m such that by setting

$$\varepsilon_n = A_0 \max\left\{\left(\frac{\ln n}{D_m}\right)^{1/4}, \left(\frac{D_m \ln n}{n}\right)^{1/4}\right\} ,$$

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

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

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

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

Having at hand Theorem 4, the proofs of Theorems 1 and 4 follow from the exact same lines as the proofs of Theorems 6 and 7 of [16]. To give a more precise view of the ideas involved, let us detail the essential arguments of the proof of Theorem 1.

We set

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

The difference between $\mathrm{crit}^0_{\mathrm{VFCV}}\left(m\right)$ and $\mathrm{crit}_{\mathrm{VFCV}}\left(m\right)$ being a quantity independent of $m\in\mathscr{M}_n$, the procedure defined by $\mathrm{crit}^0_{\mathrm{VFCV}}$ gives the same result as the VFCV procedure defined by $\mathrm{crit}_{\mathrm{VFCV}}$.

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

$$\begin{aligned}
\mathrm{crit}^0_{\mathrm{VFCV}}\left(m\right) &= \frac{1}{V}\sum_{j=1}^{V} P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma(s_*)\right)\\
&= \frac{1}{V}\sum_{j=1}^{V}\Big[P_n^{(j)}\left(\gamma\left(\widehat{s}_m^{(-j)}\right)-\gamma(s_m)\right)\\
&\quad + \left(P_n^{(j)}-P\right)\left(\gamma(s_m)-\gamma(s_*)\right)+P\left(\gamma(s_m)-\gamma(s_*)\right)\Big]\\
&= \ell\left(s_*,\widehat{s}_m^{(-1)}\right)+\Delta_V\left(m\right)+\bar{\delta}\left(m\right) \qquad (13)
\end{aligned}$$

where

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

and

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

Now, we have to show that $\Delta_V\left(m\right)$ and $\bar{\delta}\left(m\right)$ are negligeable in front of $\ell\left(s_*,\widehat{s}_m^{(-1)}\right)$. For $\bar{\delta}\left(m\right)$, this is done by using Bernstein's concentration inequality (see Lemma 7.5 of [16]). To control $\Delta_V\left(m\right)$, we also make use of Bernstein's concentration inequality, but by conditioning successively on the data used to learn the estimators $\widehat{s}_m^{(-j)}$, $j=1,\ldots,V$ (see Lemma 7.3 and Corollary 7.4 of [16]).

References

1. Antoniadis, A., Gregoire, G., McKeague, I.: Wavelet methods for curve estimation. J. Amer. Statist. Assoc. **89**(428), 1340–1353 (1994)
2. Arlot, S.: V-fold cross-validation improved: V-fold penalization (2008). URL http://hal.archives-ouvertes.fr/hal-00239182/en/. ArXiv:0802.0566v2
3. Arlot, S., Célisse, A.: A survey of cross-validation procedures for model selection. Stat. Surv. **4**, 40–79 (2010)
4. Arlot, S., Massart, P.: Data-driven calibration of penalties for least-squares regression. J. Mach. Learn. Res. **10**, 245–279 (electronic) (2009)
5. Barron, A., Birgé, L., Massart, P.: Risk bounds for model selection via penalization. Probab. Theory Related Fields **113**(3), 301–413 (1999)
6. Birgé, L., Massart, P.: Minimum contrast estimators on sieves: exponential bounds and rates of convergence. Bernoulli **4**(3), 329–375 (1998)
7. Cai, T.: Adaptive wavelet estimation: a block thresholding and oracle inequality approach. Ann. Statist. **27**(3), 898–924 (1999)
8. Cai, T., Brown, L.: Wavelet shrinkage for nonequispaced samples. Ann. Statist. **26**, 1783–1799 (1998)
9. Cai, T., Brown, L.: Wavelet estimation for samples with random uniform design. Statist. Probab. Lett. **42**(3), 313–321 (1999)
10. Donoho D. Maleki, A., Shahram, M.: Wavelab 850 (2006). URL http://statweb.stanford.edu/ wavelab/
11. Hall, P., Turlach, B.: Interpolation methods for nonlinear wavelet regression with irregularly spaced design. Ann. Statist. **25**(5), 1912–1925 (1997)
12. Kulik, R., Raimondo, M.: Wavelet regression in random design with heteroscedastic dependent errors. Ann. Statist. **37**(6A), 3396–3430 (2009)
13. Mallat, S.: A wavelet tour of signal processing: the sparse way. Academic press (2008)
14. Marron, J., Adak, S., Johnstone, I., Neumann, M., Patil, P.: Exact risk analysis of wavelet regression. J. Comput. Graph. Statist. **7**(3), 278–309 (1998)
15. Nason, G.: Wavelet shrinkage using cross-validation. J. R. Stat. Soc. Ser. B pp. 463–479 (1996)
16. Navarro, F., Saumard, A.: Slope heuristics and v-fold model selection in heteroscedastic regression using strongly localized bases. ESAIM: Probability and Statistics (in press)
17. Saumard, A.: 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)
18. Saumard, A.: Optimal model selection in heteroscedastic regression using piecewise polynomial functions. Electron. J. Statist. **7**, 1184–1223 (2013)
19. Saumard, A.: On optimality of empirical risk minimization in linear aggregation. Bernoulli (2017). To appear, arXiv:1605.03433