Regression function estimation on non compact support in an heteroskedastik model

Fabienne Comte, Valentine Genon-Catalot

REGRESSION FUNCTION ESTIMATION ON NON COMPACT SUPPORT IN AN HETEROSKESDASTIC MODEL

F. COMTE AND V. GENON-CATALOT

Abstract. We study the problem of non parametric regression function estimation on non necessarily compact support in a heteroskedastic model with unbounded variance. A collection of least squares projection estimators on m-dimensional functional linear spaces is built. We prove new risk bounds for the estimator with fixed m and propose a new selection procedure relying on inverse problems methods leading to an adaptive estimator. Contrary to more standard cases, the data-driven dimension is chosen within a random set and the penalty is random. Examples and numerical simulations results show that the procedure is easy to implement and provides satisfactory estimators. December 31, 2018

Keywords: Heteroskedatic regression model. Least squares estimation. Model selection. Projection estimator

1. Introduction

This paper is concerned with nonparametric regression function estimation under some non standard assumptions. Consider observations $(X_i, Y_i)_{i=1,\dots,n}$ such that

$$Y_i = b(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where the random variables (X_i) are real-valued, independent and identically distributed (*i.i.d.*), with density f, the noise variables (ε_i) are *i.i.d.* with $\mathbb{E}(\varepsilon_1) = 0$, $\mathrm{Var}(\varepsilon_1) = 1$ and the sequences (X_i), (ε_i) are independent. The function $b(.) : \mathbb{R} \to \mathbb{R}$ is unknown and we aim at estimating b from the sample $(X_i, Y_i)_{i=1,\dots,n}$. This problem has been the subject of a huge number of contributions and various methods have been developped (see *e.g.* Tsybakov, 2009, for a reference book). Here, we are concerned with nonparametric projection estimation of b where estimators are obtained by minimization of a least squares contrast

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}[t^2(X_i) - 2Y_i t(X_i)] = \frac{1}{n}\sum_{i=1}^{n}[Y_i - t(X_i)]^2 - \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \tag{2}$$

over a collection of finite dimensional subspaces of $\mathbb{L}^2(A, dx)$ for $A \subset \mathbb{R}$. This approach provides directly an estimator of b without requiring the estimation of the unknown density f of the X_is. A model selection procedure allows to determine the best choice of the dimension m. On the other hand, the estimation of b is restricted to the set A (see *e.g.* Birgé and Massart (1998), Barron *et al.* (1999), Baraud (2002) among others).
Papers concerned with model selection in a heteroskedastic regression model are not so numerous. Comte and Rozenholc (2002), estimate the pair (b, σ) by a two-step procedure

Université Paris Descartes, Laboratoire MAP5 UMR CNRS 8145
fabienne.comte@parisdescartes.fr, valentine.genon-catalot@parisdescartes.fr.

(one for b and one for σ) and give risk bounds separately for each function in $\mathbb{L}^2$-norm. Under the assumption of bounded data, in Arlot (2007, chapter 6), estimation of the regression function is studied. For polynomial collection of models, the author provides oracle inequalities for the quadratic risk relying on resampling penalties. Galtchouk and Pergamenshchikov (2009), propose an adaptive nonparametric estimation procedure leading to an oracle inequality for the quadratic risk under some regularity assumptions. Gendre (2008) deals with estimation using a Kullback risk. In a recent paper, Jin *et al.* (2015) study the problem of estimation in heteroskedastic regression (see also references therein), but the point of view is rather different: they consider asymptotic properties in pointwise setting, while we are interested in global risk from nonasymptotic point of view. Moreover, we do not study here the estimation of the volatility function, but its specific impact on regression function estimation.
In all of the above references and in most references, the estimation set A is assumed to be compact, and this plays a crucial role in the assumptions and bounds. In addition, it is fixed in the theory but in practice, adjusted on the data which contradicts the theoretical assumption. This is why we intend to overcome this drawback and investigate the case where A can be the whole real line or $\mathbb{R}^+$, and in any case a non compact set. This raises specific difficulties which were investigated in a recent paper (Comte and Genon-Catalot, 2018) for the homoskedatic regression, *i.e.* $\sigma(x) \equiv 1$. Now, we study the heteroskedastic case with the additional difficulty that we do not assume σ bounded. In particular, we provide a new formula for the variance term of the procedure, which makes the influence of the function $\sigma(.)$ clear, and leads to a more precise evaluation of it in the model selection procedure. This involves real difficulties, as we both estimate matrices defining the penalty function, and the collection of models, which is also random and data driven.

Projection estimators of b on a fixed space are studied in Section 2. For the risk bound defined as the expectation of an empirical norm and as the expectation of the $\mathbb{L}^2(A, f(x)dx)$-norm, we obtain a variance term which is new. Moreover the risk bounds require a constraint for the possible dimensions of the projection spaces, called "stability condition" following the terminology introduced in Cohen *et al.* (2013). Section 3 concerns the model selection procedure where a data driven choice of the projection space dimension is proposed. The non compacity of the estimation set and the unboundedness of σ induce a specific treatment. First, the data driven dimension is chosen in a random set and the penalty too is random. For the selection procedure, the function σ is supposed to be known. Section 4 contains examples of bases with non compact support and numerical simulation results on various models. To estimate the regression function on $\mathbb{R}$, we propose to use the Hermite basis and for estimation on $\mathbb{R}^+$, the Laguerre basis is convenient. These two bases have been recently used for density estimation on non compact support (see *e.g.* Comte and Genon-Catalot, 2018a) and for regression in a homoskedastic model in Comte and Genon-Catalot (2018b). In practice, σ is unknown and we show on simulations how to estimate this function in order to make the procedure implementable. Section 5 gives some concluding remarks.

2. Projection estimator on a fixed space

Let $A \subset \mathbb{R}$ and consider S_m a finite-dimensional subspace of $\mathbb{L}^2(A, dx)$ spanned by an orthonormal basis of A-supported functions $(\varphi_0, \ldots, \varphi_{m-1})$. We assume that for all j,

$\int \varphi_j^2(x)f(x)dx < +\infty$. With γ_n defined in (2), we set

$$\hat{b}_m = \arg\min_{t\in S_m} \gamma_n(t).$$

The computation of $\hat{b}_m$ using the basis $(\varphi_0, \dots, \varphi_{m-1})$ is very classical. To recall it, let us introduce some notations. For functions s, t and for u the vector $(u_1, \dots, u_n)'$ (u' denotes the transpose of u), we set

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^n t^2(X_i), \quad \langle s,t\rangle_n := \frac{1}{n}\sum_{i=1}^n s(X_i)t(X_i), \quad \langle u,t\rangle_n := \frac{1}{n}\sum_{i=1}^n u_i t(X_i).$$

We introduce the matrices

$$\widehat{\Phi}_m = (\varphi_j(X_i))_{1\le i\le n, 0\le j\le m-1}, \quad \widehat{\Psi}_m = (\langle \varphi_j, \varphi_k\rangle_n)_{0\le j,k\le m-1} = \frac{1}{n}\widehat{\Phi}_m'\widehat{\Phi}_m,$$

and

$$\Psi_m = \left(\int \varphi_j(x)\varphi_k(x)f(x)dx\right)_{0\le j,k\le m-1} = \mathbb{E}(\widehat{\Psi}_m). \tag{3}$$

Set $\vec{Y} = (Y_1, \dots, Y_n)'$. Then, the m-dimensional vector $\vec{\hat{a}}^{(m)} = (\hat{a}_0^{(m)}, \dots, \hat{a}_{m-1}^{(m)})'$ such that $\hat{b}_m = \sum_{j=0}^{m-1} \hat{a}_j^{(m)}\varphi_j$ is given, if $\widehat{\Psi}_m$ is invertible almost surely (a.s.), by

$$\vec{\hat{a}}^{(m)} = (\widehat{\Phi}_m'\widehat{\Phi}_m)^{-1}\widehat{\Phi}_m'\vec{Y} = \frac{1}{n}\widehat{\Psi}_m^{-1}\widehat{\Phi}_m'\vec{Y}. \tag{4}$$

2.1. **Risk bound in empirical norm.** We now evaluate the risk of the estimator $\hat{b}_m$. The following notations are used in the sequel. For h a function, $h_A := h\mathbf{1}_A$, $\|h\|$ is the $\mathbb{L}^2(A, dx)$ norm, $\|h\|_f$ is the $\mathbb{L}^2(A, f(x)dx)$-norm, $\|h\|_\infty$ is the sup-norm on A. For M a matrix, we denote by $\|M\|_{\text{op}}$ the operator norm defined as the square root of the largest eigenvalue of MM'. If M is symmetric, $\|M\|_{\text{op}} = \sup\{|\lambda_i|\}$ where λ_i are the eigenvalues of M. If M, N are two matrices with compatible product MN, $\|MN\|_{\text{op}} \le \|M\|_{\text{op}}\|N\|_{\text{op}}$. The trace of a matrix M is denoted $\text{Tr}(M)$.
Assuming that $\mathbb{E}[\sigma^2(X_1)] < +\infty$, we define the matrices

$$\widehat{\Psi}_{m,\sigma^2} = \left(\frac{1}{n}\sum_{i=1}^n \sigma^2(X_i)\varphi_j(X_i)\varphi_k(X_i)\right)_{0\le,j,k\le m-1} \quad \text{and} \quad \Psi_{m,\sigma^2} = \mathbb{E}[\widehat{\Psi}_{m,\sigma^2}]. \tag{5}$$

In other words,

$$\Psi_{m,\sigma^2} := \left(\int \varphi_j(x)\varphi_k(x)\sigma^2(x)f(x)dx\right)_{0\le j,k\le m-1}. \tag{6}$$

Proposition 2.1. *Let $(X_i, Y_i)_{1\le i\le n}$ be observations drawn from model (1). Assume that $\widehat{\Psi}_m$ is invertible a.s. and consider the least squares estimator $\hat{b}_m$ of $b_A = b\mathbf{1}_A$. Then*

$$\mathbb{E}\big[\|\hat{b}_m - b_A\|_n^2\big] \le \inf_{t\in S_m}\|b_A - t\|_f^2 + \frac{\mathbb{E}[\text{Tr}(\widehat{\Psi}_m^{-1}\widehat{\Psi}_{m,\sigma^2})]}{n}.$$

If in addition $\sigma_A = \sigma\mathbf{1}_A$ is bounded, then $\mathbb{E}[\text{Tr}(\widehat{\Psi}_m^{-1}\widehat{\Psi}_{m,\sigma^2})] \le m\|\sigma_A\|_\infty^2$.

If $\sigma^2(x)=\sigma^2$ is constant, $\widehat{\Psi}_{m,\sigma^2}=\sigma^2\widehat{\Psi}_m$ and the variance term is simply $\sigma^2 m/n$. This exactly coincides with the homoskedastic results. If $\sigma^2(x)$ is not constant, the variance term becomes $\mathbb{E}[\mathrm{Tr}(\widehat{\Psi}_m^{-1}\widehat{\Psi}_{m,\sigma^2})]/n$ which is new. Now, to have a better understanding of its rate, assume that

$$L(m):=\sup_{x\in A}\sum_{j=0}^{m-1}\varphi_j^2(x)<+\infty. \tag{7}$$

The quantity $L(m)$ was introduced in Comte and Genon-Catalot (2018a). It is independent of the choice of the $\mathbb{L}^2(dx)$-orthonormal basis of S_m. Moreover, for nested spaces (i.e. $m\leq m'\Rightarrow S_m\subset S_{m'}$), the map $m\mapsto L(m)$ is increasing. We show below on examples that condition (7) is not stringent and that $L(m)$ is on classical examples of order m (see Section 4).

Proposition 2.2. *Let $(X_i,Y_i)_{1\leq i\leq n}$ be observations drawn from model (1). Assume that $\widehat{\Psi}_m$ is invertible a.s. and that $\mathbb{E}(\sigma^4(X_1))<+\infty, \mathbb{E}(b^4(X_1))<+\infty$. Let m be such that*

$$L(m)(\|\Psi_m^{-1}\|_{\mathrm{op}}\vee 1)\leq \frac{\mathfrak{c}}{2}\frac{n}{\log(n)},\quad \mathfrak{c}=\frac{1-\log(2)}{5}. \tag{8}$$

Then, the least squares estimator $\hat{b}_m$ of b_A satisfies

$$\mathbb{E}\big[\|\hat{b}_m-b\|_n^2\big]\leq \inf_{t\in S_m}\|b_A-t\|_f^2+\frac{2}{n}\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right]+\frac{c}{n},$$

where c is a constant depending on $\mathbb{E}(\varepsilon_1^4)$ and $\int b_A^4(x)f(x)dx$.

We mention that the value of $\mathfrak{c}$ is chosen in order that Lemma 6.1 holds.
Condition (8) is to be interpreted as a stability condition. If m is too large, then the least-squares estimator may be very unstable. Such a condition was introduced in Cohen *et al.* (2013) and used also in Comte and Genon-Catalot (2018b). On the other hand, if A is compact and f is lower bounded by f_0 on A, then $\|\Psi_m^{-1}\|_{\mathrm{op}}\leq 1/f_0$ (see Proposition 4.1 in Comte and Genon-Catalot (2018b)). This means that condition (8) can be written $L(m)\leq \mathfrak{c}((f_0\wedge 1)/2)n/\log(n)$: this constraint is very weak, especially when $L(m)$ is of order m, see (12).

2.2. **Risk bound in integral norm.** To bound the risk in $\mathbb{L}^2(f(x)dx)$-norm, we introduce a cutoff and define

$$\tilde{b}_m:=\hat{b}_m\mathbf{1}_{L(m)(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}\vee 1)\leq \mathfrak{c}n/\log(n)}, \tag{9}$$

where $L(m)$ is defined by (7) and $\mathfrak{c}$ in (8). For nested spaces, it is proved in Comte and Genon-Catalot (2018a) that the maps $m\mapsto\|\widehat{\Psi}_m^{-1}\|$ and $m\mapsto\|\Psi_m^{-1}\|$ are nondecreasing (see Proposition 2.2).

Proposition 2.3. *Let $(X_i,Y_i)_{1\leq i\leq n}$ be observations drawn from model (1). Assume that $\widehat{\Psi}_m$ is invertible a.s., and that $\mathbb{E}(\sigma^4(X_1))<+\infty, \mathbb{E}(b^4(X_1))<+\infty$. Consider the estimator $\tilde{b}_m$ of b_A. Let m satisfy condition (8), then*

$$\mathbb{E}\big[\|\tilde{b}_m-b_A\|_f^2\big]\leq\left(1+\frac{8\mathfrak{c}}{\log(n)}\right)\inf_{t\in S_m}\|b_A-t\|_f^2+\frac{8}{n}\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right]+\frac{c'}{n}, \tag{10}$$

where c' is a constant depending on $\mathbb{E}(\varepsilon_1^4)$, $\int b_A^4(x)f(x)dx$, $\int\sigma_A^4(x)f(x)dx$.

If $\sigma^2(x) \equiv \sigma^2$, $\Psi_{m,\sigma^2} = \sigma^2\Psi_m$ and $\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] = \sigma^2 m$. In all cases, the variance term $\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right]$ has the following properties (see Proposition 3.2 in Comte and Genon-Catalot (2019)).

Proposition 2.4. *Let m be an integer. Assume that Ψ_m is invertible and $\mathbb{E}\sigma^2(X_1) < +\infty$.*

(1) *If the spaces S_m are nested, then $m \mapsto \mathrm{Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}]$ is non-decreasing.*
(2) *If σ is bounded on A, then $\mathrm{Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}] \le \|\sigma_A\|_\infty^2 m$.*
(3) *Under condition (7), $\mathrm{Tr}[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}] \le \mathbb{E}[\sigma_A^2(X_1)]L(m)\|\Psi_m^{-1}\|_{\mathrm{op}}$.*

When σ^2 is unknown, the upper bound *(3)* is interesting as $\mathbb{E}[\sigma_A^2(X_1)]$ is easy to estimate and this quantity has been used as penalty term in the context of drift estimation in stochastic differential equations studied in Comte and Genon-Catalot (2019). However, this upper bound is not sharp. On simulations, we observe that the left-hand side seems proportional to m while the right-hand side increases very fast with m.

2.3. **Discussion on rates and lower bound.** The evaluation of risk rates in this problem is delicate. Indeed, we do not know explicitly the variance rate in the risk bound (10) and the assessment of the bias term requires to introduce specific regularity spaces linked with f. In Comte and Genon-Catalot (2018b), the following spaces were introduced:

$$W_f^s(A,R) = \left\{h \in \mathbb{L}^2(A, f(x)dx), \forall \ell \ge 1, \|h - h_\ell^f\|_f^2 \le R\ell^{-s}\right\} \tag{11}$$

where h_ℓ^f is the $\mathbb{L}^2(A, f(x)dx)$-orthogonal projection of h on S_ℓ. Hence, for $f \in W_f^s(A,R)$, $\inf_{t\in S_m}\|b_A - t\|_f^2 \le Rm^{-s}$. If the function σ is upper bounded, by Proposition 2.4, the variance term is upper bounded by a term of order m/n. Therefore, if $m_{\mathrm{opt}} := n^{1/(s+1)}$ satisfies (8), then, the risk of $\tilde{b}_{m_{\mathrm{opt}}}$ is upper bounded by $Cn^{-s/(s+1)}$.
If in addition σ is lower bounded ($\sigma(x) \ge \sigma_0 > 0$), the proof of Theorem 3.1 in Comte and Genon-Catalot (2018b) can be extended without difficulty and leads to the following lower bound: for $\varepsilon_1 \sim \mathcal{N}(0, \sigma_\varepsilon^2)$,

$$\liminf_{n\to+\infty} \inf_{T_n} \sup_{b_A \in W_f^s(A,R)} \mathbb{E}_{b_A}[n^{s/(s+1)}\|T_n - b_A\|_f^2] \ge c$$

where $\inf_{T_n}$ denotes the infimum over all estimators and the constant $c > 0$ depends on s and R.

3. Adaptive procedure

We present now a model selection procedure and associated risk bounds where the following assumptions are used.

(A1) We consider a nested collection of spaces $(S_m, m \in \mathcal{M}_n)$ (that is $S_m \subset S_{m'}$ for $m \le m'$) such that, for each m, the basis $(\varphi_0, \ldots, \varphi_{m-1})$ of S_m satisfies

$$\|\sum_{j=0}^{m-1}\varphi_j^2\|_\infty \le c_\varphi^2 m \quad \text{for} \quad c_\varphi^2 > 0 \quad \text{a constant.} \tag{12}$$

(A2) $\|f\|_\infty < +\infty$.
(A3) For $\alpha \ge 1$ and $m \ge 1$, $\|\Psi_m^{-1}\|_{\mathrm{op}}^2 \ge c^\star m^\alpha$, where $c^\star$ is a positive constant.

(A4) $\sum_{m\geq 1} m\|\Psi_m^{-1}\|_{\mathrm{op}} e^{-m/12} \leq \Sigma < +\infty.$

Assumption **(A1)** is satisfied on most examples of compactly supported bases and for the non compactly supported Laguerre and Hermite bases (see below Section 4). Assumption **(A2)** is standard and rather weak. Assumption **(A3)** holds for the Laguerre and Hermite bases (see Proposition 3.4 in Comte and Genon-Catalot (2018)) with $\alpha = 1$ (see Section 4). Assumption **(A4)** holds for instance when $\|\Psi_m^{-1}\|_{\mathrm{op}}$ has polynomial order which is the case if the unknown density f is $\mathbb{R}^+$-supported, satisfies $f(x) \geq c/(1+x)^k$ for all $x \geq 0$ and the Laguerre basis is used or if f satisfies $f(x) \geq c/(1+x^2)^k$ for all x and the Hermite basis is used (see Comte and Genon-Catalot (2018), Proposition 3.5). Let us mention that if $\|\Psi_m^{-1}\|_{\mathrm{op}}^2$ is bounded by a constant $1/f_0$, Assumption **(A4)** is automatically fulfilled, but Assumption **(A3)** is not. However, Assumption **(A3)** is useful to deal with the random penalty involving several matrices. In fact, in the case of bounded $\|\Psi_m^{-1}\|_{\mathrm{op}}^2$, the penalty may be taken proportional to $c_\varphi^2 \mathbb{E}(\sigma^2(X_1))m/(f_0 n)$ (see Proposition 2.4, (3)) and studied as in the homoskedastic case (see Comte and Genon-Catalot (2018a)).

Under **(A2)**, we define[1]

(13)
$$\mathcal{M}_n = \left\{ m \in \mathbb{N}, c_\varphi^2 m\, (\|\Psi_m^{-1}\|_{\mathrm{op}}^2 \vee 1) \leq \frac{\mathfrak{d}}{4}\frac{n}{\log(n)} \right\},\ \mathfrak{d} = \min\{\frac{1/192}{(\|f\|_\infty \vee 1) + 1/3}, \frac{3}{8}\sqrt{c^*}\}.$$

To select the most relevant space S_m, we set

$$\hat{m} = \arg \min_{m \in \widehat{\mathcal{M}}_n} \left\{ -\|\hat{b}_m\|_n^2 + \widehat{\mathrm{pen}}(m) \right\}, \tag{14}$$

$$\widehat{\mathrm{pen}}(m) = 2\kappa \frac{m}{n} \widehat{V}(m), \qquad \widehat{V}(m) = \|\widehat{\Psi}_m^{-1/2} \widehat{\Psi}_{m,\sigma^2} \widehat{\Psi}_m^{-1/2}\|_{\mathrm{op}} + 1, \tag{15}$$

where κ is a numerical constant and $\widehat{\mathcal{M}}_n$ is a collection of models defined by

$$\widehat{\mathcal{M}}_n = \left\{ m \in \mathbb{N}, c_\varphi^2 m (\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}}^2 \vee 1) \leq \mathfrak{d}\frac{n}{\log(n)} \right\}, \tag{16}$$

The set $\widehat{\mathcal{M}}_n$ where $\hat{m}$ is chosen is random which is not usual in such procedures. It is the empirical counterpart of $\mathcal{M}_n$ defined by (13) (with a change of constant). Similarly, the theoretical counterpart of $\widehat{\mathrm{pen}}(m)$ is

$$\mathrm{pen}(m) = \kappa \frac{m}{n} V(m), \qquad V(m) = \|\Psi_m^{-1/2} \Psi_{m,\sigma^2} \Psi_m^{-1/2}\|_{\mathrm{op}} + 1. \tag{17}$$

Note that if $\sigma^2(x) \equiv \sigma^2$ is a constant, then $\widehat{V}(m) = V(m) = \sigma^2 + 1$ and the penalty $\widehat{\mathrm{pen}}(m)$ is nonrandom. Note that both $m \mapsto \widehat{V}(m)$ and $m \mapsto V(m)$ are nondecreasing as can be easily checked.

Theorem 3.1. *Let $(X_i, Y_i)_{1\leq i\leq n}$ be observations from model (1). Assume that* **(A1)**-**(A4)** *hold, that $\mathbb{E}(\varepsilon_1^{10}) < +\infty$, $\mathbb{E}\big[b^4(X_1)\big] < +\infty$, and $\mathbb{E}\big[\sigma_A^{4+56/\alpha}(X_1)\big] < +\infty$. Then,*

[1]The constant $\mathfrak{d}$ is calibrated in order that (33) and (42) both hold.

there exists a numerical constant κ_0 such that for $\kappa \geq \kappa_0$, we have

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_n^2\big] \leq C \inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|b_A - t\|_f^2 + \mathrm{pen}(m)\right) + \frac{C'}{n}, \tag{18}$$

and

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_f^2\big] \leq C_1 \inf_{m\in\mathcal{M}_n}\left(\inf_{t\in S_m}\|b_A - t\|_f^2 + \mathrm{pen}(m)\right) + \frac{C_1'}{n} \tag{19}$$

where C, C_1 are a numerical constants and C', C_1' are constants depending on f, b, σ.

Note that we do not use the exact value of variance in the empirical and theoretical penalty, but an upper bound on it, as for a $m \times m$ positive matrix, $\mathrm{Tr}[M] \leq m\|M\|_{\mathrm{op}}$. As usual, the data-driven choice (14) of the dimension m is dictated by the squared-bias-variance compromise. Two difficulties are to be stressed. First, $\hat{m}$ is chosen in a random set (this was already the case in the homoscedastic model treated in Comte and Genon-Catalot (2018)). Second we have to deal here with a random penalty which was not the case in the homoscedastic regression model. Handling this in the proof is rather difficult. For practical implementation, the constant κ is calibrated by preliminary simulations instead of applying the theoretical value κ_0 provided by the proof which is not optimal. The penalty (15) depends on the vector $(\sigma^2(X_i), i = 1, \ldots, n)$. In Section 4, we propose to replace these values by $((Y_i - \hat{b}_m(X_i))^2)$ for a m taken close to the maximal possible value and show that this works well on simulations.

4. Examples and numerical simulations

For implementation, we consider either the Laguerre basis ($A = \mathbb{R}^+$) or the Hermite basis ($A = \mathbb{R}$). Both are easy to handle in practice.

- Laguerre basis, $A = \mathbb{R}^+$. The Laguerre polynomial L_j and the Laguerre function ℓ_j of order j are given by

$$L_j(x) = \sum_{k=0}^{j}(-1)^k\binom{j}{k}\frac{x^k}{k!}, \qquad \ell_j(x) = \sqrt{2}L_j(2x)e^{-x}\mathbf{1}_{x\geq 0}, \quad j \geq 0. \tag{20}$$

The collection $(\ell_j)_{j\geq 0}$ constitutes a complete orthonormal system on $\mathbb{L}^2(\mathbb{R}^+)$ satisfying (see Abramowitz and Stegun (1964)): $\forall j \geq 0,\ \forall x \in \mathbb{R}^+,\ |\ell_j(x)| \leq \sqrt{2}$. The collection of models $(S_m = \mathrm{span}\{\ell_0, \ldots, \ell_{m-1}\})$ is nested and obviously (12) holds with $c_\varphi^2 = 2$.

- Hermite basis, $A = \mathbb{R}$. The Hermite polynomial H_j and the Hermite function of order j are given, for $j \geq 0$, by:

$$H_j(x) = (-1)^j e^{x^2}\frac{d^j}{dx^j}(e^{-x^2}), \quad h_j(x) = c_j H_j(x)e^{-x^2/2}, \quad c_j = \left(2^j j!\sqrt{\pi}\right)^{-1/2} \tag{21}$$

The sequence $(h_j, j \geq 0)$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R}, dx)$. Moreover (see Abramowitz and Stegun (1964), Szegö (1959) p.242), $\|h_j\|_\infty \leq \Phi_0, \Phi_0 \simeq 1,086435/\pi^{1/4} \simeq 0.8160$, so that (12) holds with $c_\varphi^2 = \Phi_0^2$. The collection of models $(S_m = \mathrm{span}\{h_0, \ldots, h_{m-1}\})$ is obviously nested.

Laguerre polynomials were computed using formula (20) and Hermite polynomials with $H_0(x) \equiv 1$, $H_1(x) = x$ and the recursion $H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$.

$X\sim$	$\sigma_1(x)\equiv\sigma$				$\sigma_2(x)=\sigma\sqrt{\lvert x\rvert}$				$\sigma_3(x)=\sigma\sqrt{1+x^2}$			
	$\mathcal{U}([-1,1])$		$\mathcal{N}(0,1/3)$		$\mathcal{U}([-1,1])$		$\mathcal{N}(0,1/3)$		$\mathcal{U}([-1,1])$		$\mathcal{N}(0,1/3)$	
	$\hat\sigma$	σ	$\hat\sigma$	σ	$\hat\sigma$	σ	$\hat\sigma$	σ	$\hat\sigma$	σ	$\hat\sigma$	σ
$b_1(x)$												
Emp.	0.14	0.14	0.25	0.25	0.09	0.08	0.25	0.23	0.16	0.17	0.78	0.69
	(0.1)	(0.1)	(0.5)	(0.5)	(0.06)	(0.06)	(0.2)	(0.17)	(0.10)	(0.12)	(0.47)	(0.46)
L^2	0.13	0.13	0.22	0.22	0.07	0.07	0.18	0.17	0.14	0.16	0.52	0.46
	(0.1)	(0.1)	(0.5)	(0.5)	(0.05)	(0.05)	(0.16)	(0.13)	(0.09)	(0.11)	(0.40)	(0.35)
dim	4.3	4.3	6.4	6.7	4.2	4.1	6.2	6.0	4.0	4.1	5.0	5.2
	(0.7)	(0.8)	(1.1)	(1.5)	(0.5)	(0.3)	(0.2)	(0.8)	(0.2)	(0.5)	(1.3)	(1.2)
$b_2(x)$												
Emp	0.17	0.17	0.24	0.24	0.12	0.13	0.30	0.34	0.21	0.22	0.81	0.59
	(0.08)	(0.09)	(0.12)	(0.13)	(0.06)	(0.05)	(0.70)	(0.70)	(0.10)	(0.11)	(0.79)	(0.35)
L^2	0.15	0.15	0.21	0.22	0.10	0.11	0.23	0.25	0.19	0.20	0.46	0.40
	(0.07)	(0.08)	(0.14)	(0.15)	(0.05)	(0.04)	(0.68)	(0.68)	(0.10)	(0.11)	(0.32)	(0.27)
dim	3.8	4.0	6.5	6.9	3.9	3.7	6.4	5.8	3.1	3.3	5.1	5.2
	(1.1)	(1.2)	(1.17)	(1.17)	(1.2)	(1.0)	(1.5)	(1.2)	(0.5)	(0.7)	(1.1)	(0.6)
$b_3(x)$												
Emp	0.14	0.15	0.21	0.22	0.09	0.09	0.34	0.23	0.24	0.21	0.60	0.57
	(0.08)	(0.09)	(0.11)	(0.12)	(0.06)	(0.06)	(2.19)	(0.15)	(0.22)	(0.16)	(0.72)	(0.29)
L^2	0.13	0.13	0.18	0.19	0.08	0.08	0.27	0.16	0.21	0.19	0.44	0.40
	(0.08)	(0.08)	(0.13)	(0.13)	(0.05)	(0.05)	(2.1)	(0.12)	(0.18)	(0.14)	(0.64)	(0.28)
dim	5.1	5.2	6.5	6.7	5.1	5.1	6.3	6.0	5.0	5.1	5.5	5.4
	(0.3)	(0.7)	(1.7)	(1.8)	(0.4)	(0.3)	(1.5)	(1.2)	(0.3)	(0.4)	(0.7)	(0.6)
$b_4(x)$												
Emp.	0.92	0.91	5.02	2.42	0.81	0.81	5.65	2.47	1.25	1.03	9.38	3.07
	(0.18)	(0.11)	(4.65)	(0.67)	(0.08)	(0.08)	(4.96)	(0.72)	(0.81)	(0.17)	(7.12)	(0.81)
L^2	0.89	0.88	4.47	2.19	0.78	0.78	4.93	2.16	1.19	0.97	8.04	2.50
	(0.19)	(0.12)	(3.98)	(0.65)	(0.08)	(0.08)	(4.35)	(0.69)	(0.79)	(0.17)	(6.23)	(0.85)
dim	11	11	14.5	16.0	11	11	14.3	16.1	10.8	11	13.0	16.1
	(0.1)	(0.0)	(2.0)	(1.0)	(0.0)	(0.0)	(2.1)	(1.1)	(0.6)	(0.0)	(2.2)	(1.1)

TABLE 1. Emp.= empirical risk $\times$ 100 (with std $\times$ 100), L^2 = standard $\mathbb{L}^2$-risk $\times$ 100 (with std $\times$ 100), dim= mean of the selected dimensions (with std). 400 samples with size $n = 1000$.

We consider several models. In all cases, we generate the ε_i's as i.i.d. standard Gaussian random variables. For the distribution of X, we choose $X\sim\mathcal{U}([-1,1])$ (compact support case), or $X\sim\mathcal{N}(0,1)/\sqrt{3}$ (non compact support case), both have the same variance equal to $1/3$. Graphs are also given for X with Gamma distribution $\gamma(3,1/4)$ for implementation using the Laguerre basis. The uniform case allows to check that the general procedure works also in the compact setting. For the function $b(x)$ we experimented

$$b_1(x)=-2x+1,\ b_2(x)=1-x^2,\ b_3(x)=\sin(\pi x+\pi/3),\ b_4(x)=2(x+2e^{-16x^2})$$

jointly with different functions $\sigma(x)$, with $\sigma = 0.5$ in all cases:

$$\sigma_1(x) = \sigma, \quad \sigma_2(x) = \sigma\sqrt{|x|}, \quad \sigma_3(x) = \sqrt{1+x^2}.$$

The maximal dimension $\mathfrak{d}n/\log(n)$ is replaced by $D_{\max} = \lfloor n/\log^2(1+n)\rfloor - 1$, which from theoretical and practical point of view avoids to look for a specific value of $\mathfrak{d}$. The cutoff to fix the random collection of models associated with each path is taken much larger than recommended by the definition of $\widehat{\mathcal{M}}_n$ in (16), namely we consider all dimensions $m \le D_{\max}$ such that $2\|\widehat{\Psi}_m^{-1}\|_{\text{op}}^{1/4} \le D_{\max}$, this defines a maximal value of m, $\widehat{M}_n$. The choice of $\widehat{M}_n$ is really delicate and implementing the true stability constraint $m\|\widehat{\Psi}_m^{-1}\|_{\text{op}} \le D_{\max}$ is probably possible but further numerical investigations would be required to avoid rare explosion events. The penalty constant is roughly calibrated to the value $\kappa = 3$ in the two cases where the true function $\sigma(x)$ is used for computing $\|\widehat{\Psi}_{m,\sigma^2}\|_{\text{op}}$ and where the terms $\sigma^2(X_i)$ in the matrix are replaced by $((Y_i - \hat{b}_{m^\star}(X_i))^2)$, with $m^\star = \widehat{M}_n - 2$.

Simulations results are given for sample sizes $n = 400$. For Table 1, $K = 400$ repetitions are done to compute the risks. Column "$\hat{\sigma}$" means that the coefficients of the matrix $\widehat{\Psi}_{m,\sigma^2}$ are estimated as specified previously and column "σ" means that function σ is assumed to be known (in the computation of the matrix $\widehat{\Psi}_{m,\sigma^2}$ in the penalty).

Table 1 gives the empirical mean squared error (multiplied by 100) (Emp) computed as

$$\frac{1}{K}\sum_{j=1}^{K}\frac{1}{n}\sum_{i=1}^{n}\left[\hat{b}_{\hat{m}^{(j)}}(X_i^{(j)}) - b(X_i^{(j)})\right]^2,$$

where $X_i^{(j)}$ is the ith observation of the jth simulated path, and the $\mathbb{L}^2$-risk (multiplied by 100) (L^2) computed as

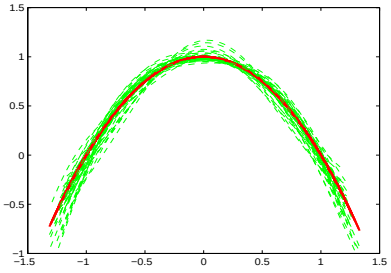
$$\frac{1}{K}\sum_{j=1}^{K}\frac{1}{100}\sum_{\ell=1}^{100}\left[\hat{b}_{\hat{m}^{(j)}}(x_\ell^{(j)}) - b(x_l^{(j)})\right]^2,$$

where $(x_\ell^{(j)})_{1\le j\le 100}$ is a set of 100 equispaced points on $[a^{(j)}, b^{(j)}]$, $a^{(j)}$ is the 2% quantile of the sample $(X_i^{(j)})_{1\le i\le n}$ and $b^{(j)}$ the 98% quantile. Standard deviations multiplied by 100 are given below in parenthesis. We also give the mean of the selected dimensions along the 400 repetitions, "dim", together with its standard deviation in parenthesis.

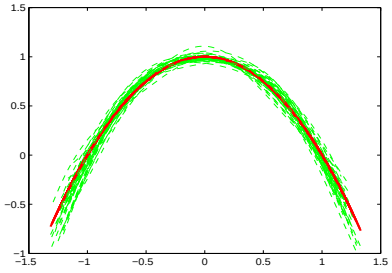
The results in Table 1 show that the procedure works quite well. We can see that the MSEs are smaller for uniform than for Gaussian X_i's. Empirical MSEs are generally larger than standard $\mathbb{L}^2$ error, but this may be due to the fact that the partition excludes points near the borders. Globally, estimating $\sigma(x)$ for the penalty seems to work well, except for b_4 and Gaussian X, where the error is much larger for estimated $\sigma(x)$ than for known $\sigma(x)$. For uniform X, the risk is smaller for $\sigma_2(x)$ that for constant function σ_1. This may be due to the variance reduction associated with the factor $\sqrt{|x|}$ when $|x| \le 1$, and this is coherent with the observed increase for $\sigma_3(x)$.

Figure 1 presents beams of estimated functions for 40 different paths. The results are rather typical of the method for such sample size, when using Hermite (on $\mathbb{R}$) or Laguerre

Hermite basis, $b_2(x) = 1 - x^2$ $\sigma(x) = \sigma\sqrt{1+x^2}$, $X \sim \mathcal{N}(0, 1/3)$

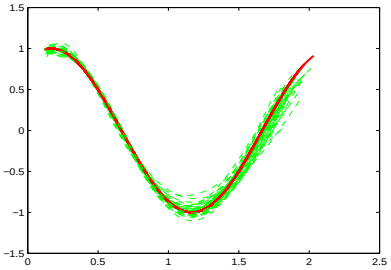


$\bar{\hat{m}} = 4.9$ (1.1), Emp. $= 0.9$ (0.8)

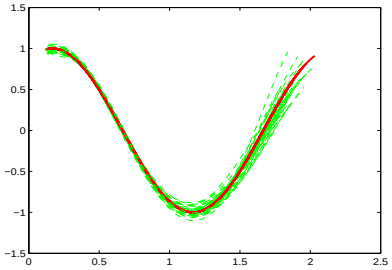


$\bar{\hat{m}} = 5.2$ (0.7), Emp. $= 0.6$ (0.3)

Laguerre basis, $b_3(x) = \sin(\pi x + \pi/3)$, $\sigma(x) = \sigma\sqrt{|x|}$, $X \sim \gamma(3, 1/4)$

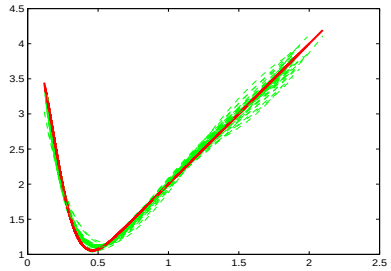


$\bar{\hat{m}} = 6.5$ (1.4), Emp. $= 0.4$ (0.4)

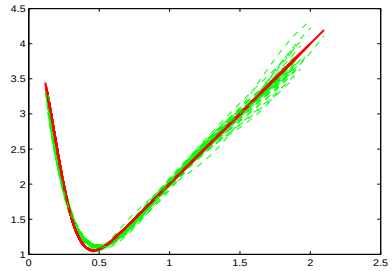


$\bar{\hat{m}} = 6.8$ (1.4), Emp. $= 0.3$ (0.2)

Laguerre basis, $b_4(x) = 2(x + 2e^{-16x^2})$, $\sigma(x) = \sigma\sqrt{1+x^2}$, $X \sim \gamma(3, 1/4)$



$\bar{\hat{m}} = 5.4$ (1.5), Emp. $= 0.8$ (0.5)



$\bar{\hat{m}} = 5.7$ (1.6), Emp. $= 0.7$ (0.2)

FIGURE 1. The true function b in bold (red-black) and 40 estimated curves (green-grey) in Hermite basis (top) or Laguerre basis (middle and bottom), the true in bold (red), $n = 1000$, $\bar{\hat{m}}$: mean selected dimension (std), Emp. $= 100\times$ empirical risk (100 std).

basis (on $\mathbb{R}^+$). The empirical risks given below graphs are computed as for Table 1 but for

the 40 paths of the illustration. The method is stable as shown by the variability bands and the values of risks.

5. Concluding remarks

In this paper, we consider the nonparametric regression function estimation by projection method allowing the estimation set to be non compact and the variance to be unbounded. This paper completes and extends the paper Comte and Genon-Catalot (2018b) where the homoskedastic regression model is studied with the analogous method. Introducing an unbounded variance term changes a lot the theoretical study. First, the upper bound of the estimators risk shows a new variance term whose explicit rate is not easy to determine. This makes the determination of the optimal rate difficult. The problem of finding it and proving a corresponding lower bound is open and worth of interest.

In both the homoskedastic and heteroskedastic models, the data-driven dimension is chosen in a random set which is not standard. In the heteroskedastic case, the model selection procedure relies on a random penalty which is not the standard case too. The resulting estimator is adaptive in the sense that its risk automatically achieves the square-bias-variance compromise. As illustrated by our simulations, the method is easy to implement and works well.

6. Proofs

6.1. **Proof of Proposition 2.1.** Let us denote by Π_m the orthogonal projection (for the scalar product of $\mathbb{R}^n$) on the sub-space $\big\{\big(t(X_1),\dots,t(X_n)\big)', t\in S_m\big\}$ of $\mathbb{R}^n$ and by $\Pi_m b$ the projection of the vector $(b(X_1),\dots,b(X_n))'$. The following equality holds,

$$\|\hat b_m - b_A\|_n^2 = \|\Pi_m b - b_A\|_n^2 + \|\hat b_m - \Pi_m b\|_n^2 = \inf_{t\in S_m} \|t - b_A\|_n^2 + \|\hat b_m - \Pi_m b\|_n^2 \tag{22}$$

By taking expectation, we obtain

$$\mathbb{E}\big[\|\hat b_m - b_A\|_n^2\big] \le \inf_{t\in S_m} \|t - b_A\|_f^2 + \mathbb{E}\left[\|\hat b_m - \Pi_m b\|_n^2\right].$$

Denote by $b(X) = (b(X_1),\dots,b(X_n))'$ and $b_A(X) = (b_A(X_1),\dots,b_A(X_n))'$. We can write

$$\hat b_m(X) = (\hat b_m(X_1),\dots,\hat b_m(X_n))' = \widehat\Phi_m \vec{\hat a}^{(m)},$$

where $\vec{\hat a}^{(m)}$ is given by (4), and

$$\Pi_m b = \widehat\Phi_m \vec a^{(m)}, \quad \vec a^{(m)} = (\widehat\Phi_m'\widehat\Phi_m)^{-1}\widehat\Phi_m' b(X).$$

Now, denoting by $\mathbf{P}(X) := \widehat\Phi_m(\widehat\Phi_m'\widehat\Phi_m)^{-1}\widehat\Phi_m'$, and by $\overrightarrow{\sigma_A\varepsilon}$ the $n\times 1$-vector with coordinates $\sigma_A(X_i)\varepsilon_i$, $i=1,\dots,n$, we get, as the φ_j are A-supported,

$$\|\hat b_m - \Pi_m b\|_n^2 = \|\mathbf{P}(X)\overrightarrow{\sigma_A\varepsilon}\|_n^2 = \frac{1}{n}\|\mathbf{P}(X)\overrightarrow{\sigma\varepsilon}\|_{2,n}^2 = \frac{1}{n}(\overrightarrow{\sigma\varepsilon})'\mathbf{P}(X)(\overrightarrow{\sigma_A\varepsilon}),$$

as $\mathbf{P}(X)'\mathbf{P}(X) = \mathbf{P}(X)$ and $\mathbf{P}(X)$ is the $n\times n$-matrix of the euclidean orthogonal projection on the subspace of $\mathbb{R}^n$ generated by the vectors $\varphi_0(X),\dots,\varphi_{m-1}(X)$, where $\varphi_j(X) = (\varphi_j(X_1),\dots,\varphi_j(X_n))'$. Note that

$$\mathbb{E}(\|\mathbf{P}(X)\overrightarrow{\sigma_A\varepsilon}\|_{2,n}^2) \le \mathbb{E}(\|\overrightarrow{\sigma_A\varepsilon}\|_{2,n}^2) \le \|\sigma_A\|_\infty \mathbb{E}(\|\vec\varepsilon\|_{2,n}^2) < +\infty.$$

Next, using that $\mathbf{P}(X)$ has coefficients depending on the X_i's only,

$$\begin{aligned}\mathbb{E}\big[(\overrightarrow{\sigma\varepsilon})'\mathbf{P}(X)(\overrightarrow{\sigma\varepsilon})\big] &= \sum_{i,\ell}\mathbb{E}\big[\varepsilon_i\varepsilon_\ell\sigma(X_i)\sigma(X_\ell)[\mathbf{P}(X)]_{i,\ell}\big]=\sum_{i=1}^n\mathbb{E}\big[\sigma^2(X_i)[\mathbf{P}(X)]_{i,i}\big]\\ &= \frac{1}{n}\sum_{i=1}^n\sum_{0\le j,k\le m-1}\mathbb{E}\big[\sigma^2(X_i)\varphi_j(X_i)\varphi_k(X_i)[\widehat{\Psi}_m^{-1}]_{j,k}\big]\\ &= \mathbb{E}\left\{\sum_{j,k}[\widehat{\Psi}_{m,\sigma^2}]_{j,k}[\widehat{\Psi}_m^{-1}]_{j,k}\right\}=\mathbb{E}\left[\mathrm{Tr}[\widehat{\Psi}_{m,\sigma^2}\widehat{\Psi}_m^{-1}]\right].\end{aligned}$$

If σ is bounded on A, we have $\mathbb{E}\big[(\overrightarrow{\sigma\varepsilon})'\mathbf{P}(X)(\overrightarrow{\sigma\varepsilon})\big]\le\|\sigma_A\|_\infty^2\mathbb{E}\big[\mathrm{Tr}(\mathbf{P}(X))\big]=m\|\sigma_A\|_\infty^2$, as $\mathrm{Tr}(\mathbf{P}(X))=\mathrm{Tr}[(\widehat{\Phi}_m'\widehat{\Phi}_m)^{-1}\widehat{\Phi}_m'\widehat{\Phi}_m]=m$. We obtain the result of Proposition 2.1. □

6.2. **Proof of Proposition 2.2.** We define the set,

$$\Omega_m=\left\{\left|\frac{\|t\|_n^2}{\|t\|_f^2}-1\right|\le\frac{1}{2},\forall t\in S_m\right\}. \tag{23}$$

It is easy to see that (see Proposition 2.3 in Comte and Genon-Catalot (2018)):

$$\Omega_m=\left\{\|\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}-\mathrm{Id}_m\|_{\mathrm{op}}\le\frac{1}{2}\right\}. \tag{24}$$

This implies that on Ω_m, all the eigenvalues of $\Psi_m^{-1/2}\widehat{\Psi}_m\Psi_m^{-1/2}$ belong to $[1/2,3/2]$. The following lemma is proved in Comte and Genon-Catalot (2018) (see Proposition 2.3 and Lemma 6.3) and determines the value of $\mathfrak{c}$.

Lemma 6.1. *Under the assumptions of Proposition 2.2, for m satisfying condition (8), we have $\mathbb{P}(\Omega_m^c)\le c/n^4$, where c is a positive constant.*

Now, we write

$$\|\hat{b}_m-b_A\|_n^2 = \|\hat{b}_m-b_A\|_n^2\mathbf{1}_{\Omega_m}+\|\hat{b}_m-b_A\|_n^2\mathbf{1}_{\Omega_m^c} \tag{25}$$

For the last term, we have

$$\|b_A-\hat{b}_m\|_n^2=\|b_A-\Pi_m b_A\|_n^2+\|\Pi_m\sigma\varepsilon\|_n^2\le\|b\|_n^2+n^{-1}\sum_{k=1}^n\sigma^2(X_k)\varepsilon_k^2. \tag{26}$$

Thus

$$\begin{aligned}\mathbb{E}\big[\|b_A-\hat{b}_m\|_n^2\mathbf{1}_{\Omega_m^c}\big] &\le \mathbb{E}\big[\|b\|_n^2\mathbf{1}_{\Omega_m^c}\big]+\frac{1}{n}\sum_{k=1}^n\mathbb{E}\big[\sigma^2(X_1)\varepsilon_k^2\mathbf{1}_{\Omega_m^c}\big]\\ &\le (\mathbb{E}^{1/2}\big[b^4(X_1)\big]+\mathbb{E}^{1/2}\big[\sigma^4(X_1)\big]\mathbb{E}^{1/2}\big[\varepsilon_1^4\big])\mathbb{P}^{1/2}(\Omega_m^c)\le\frac{C}{n^2}. \qquad (27)\end{aligned}$$

Next,

$$\begin{aligned}\mathbb{E}[\|b_A-\hat{b}_m\|_n^2\mathbf{1}_{\Omega_m}] &= \mathbf{E}[\|b_A-\Pi_m b_A\|_n^2\mathbf{1}_{\Omega_m}]+\mathbb{E}[\|\Pi_m\sigma\varepsilon\|_n^2\mathbf{1}_{\Omega_m}]\\ &\le \inf_{t\in S_m}\|b_A-t\|_f^2+\mathbb{E}[\|\Pi_m\sigma\varepsilon\|_n^2\mathbf{1}_{\Omega_m}] \qquad (28)\end{aligned}$$

From the proof of Proposition 2.1, and using that Ω_m only depends on $X_1, \dots, X_n$, we have

$$\mathbb{E}[\|\Pi_m \sigma\varepsilon\|_n^2 \mathbf{1}_{\Omega_m}] = \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}(\widehat{\Psi}_m^{-1}\widehat{\Psi}_{m,\sigma^2})\mathbf{1}_{\Omega_m}\right].$$

We note that

$$\mathrm{Tr}(\widehat{\Psi}_m^{-1}\widehat{\Psi}_{m,\sigma^2}) = \mathrm{Tr}(\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2}\Psi_m^{-1/2}\widehat{\Psi}_{m,\sigma^2}\Psi_m^{-1/2}).$$

We know that the eigenvalues $(\lambda_j)_{1\le j\le m}$ of $\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2}$ belong to $[2/3, 2]$ on Ω_m. Write $\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2} = P'DP$, with $D = diag(\lambda_i)$ and $PP' = P'P = \mathrm{Id}_m$ and denote by $M = P\Psi_m^{-1/2}\widehat{\Psi}_{m,\sigma^2}^{-1}\Psi_m^{-1/2}P$. The matrix M is symmetric nonnegative so that $[M]_{j,j} \ge 0$ for all j. Thus

$$\mathrm{Tr}(\Psi_m^{1/2}\widehat{\Psi}_m^{-1}\Psi_m^{1/2}\Psi_m^{-1/2}\widehat{\Psi}_{m,\sigma^2}\Psi_m^{-1/2}) = \mathrm{Tr}[DM] = \sum_{j=1}^m \lambda_j [M]_{j,j} \le 2\mathrm{Tr}(M).$$

Now as $\mathrm{Tr}(M) = \mathrm{Tr}(\Psi_m^{-1/2}\widehat{\Psi}_{m,\sigma^2}\Psi_m^{-1/2})$, we get

$$\mathbb{E}[\|\Pi_m \sigma\varepsilon\|_n^2 \mathbf{1}_{\Omega_m}] = \frac{1}{n}\mathbb{E}\left[\mathrm{Tr}(\widehat{\Psi}_m^{-1}\widehat{\Psi}_{m,\sigma^2})\mathbf{1}_{\Omega_m}\right] \le \frac{2}{n}\left[\mathrm{Tr}(\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2})\right]. \tag{29}$$

Now, gathering (29), (28) and (27) and (25) gives the result of Proposition 2.2. □.

6.3. **Proof of Proposition 2.3.** Let (see (9))

$$\Lambda_m = \left\{ L(m)(\|\widehat{\Psi}_m^{-1}\|_{\mathrm{op}} \vee 1) \le \mathfrak{c}\frac{n}{\log(n)} \right\},$$

Lemma 6.2. *Under the assumptions of Proposition 2.2, for m satisfying condition (8), we have $\mathbb{P}(\Lambda_m^c) \le c/n^4$, where c is a positive constant.*

Now, we write

$$\|\tilde{b}_m - b_A\|_f^2 = \|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Omega_m\cap\Lambda_m} + \|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Omega_m^c\cap\Lambda_m} + \|b_A\|_f^2 \mathbf{1}_{\Lambda_m^c} \tag{30}$$

The last term is obviously negligible as $\mathbb{E}[\|b_A\|_f^2 \mathbf{1}_{\Lambda_m^c}] \le \|b_A\|_f^2 \mathbb{P}(\Lambda_m^c) \le c\mathbb{E}[b^2(X_1)]/n$.

For the middle term, we have from (42) in Comte and Genon-Catalot (2018) (proof of Proposition 3.1), that

$$\mathbb{E}(\|\hat{b}_m - b_A\|_f^2 \mathbf{1}_{\Omega_m^c\cap\Lambda_m}) \le \frac{c}{n}.$$

So we turn to the main term and denote by $b_m^{(n)}$ the orthogonal projection of b on S_m w.r.t. the norm $\|.\|_n$. Set also (see Cohen *et al.* (2013)) $g = b - b_m^{(f)}$ where $b_m^{(f)}$ is the orthogonal projection of b on S_m w.r.t. the norm $\|.\|_f$. Note that $\|g\|_f^2 = \inf_{t\in S_m}\|b_A - t\|_f^2$ and $g_m^{(n)} = b_m^{(n)} - b_m^{(f)}$.

$$\begin{aligned}\|b_A - \hat{b}_m\|_f^2 &= \|g - g_m^{(n)} - (\hat{b}_m - b_m^{(n)})\|_f^2 = \|g\|_f^2 + \|g_m^{(n)} - (\hat{b}_m - b_m^{(n)})\|_f^2 \\ &\le \|g\|_f^2 + 2\|g_m^{(n)}\|_f^2 + 2\|(\hat{b}_m - b_m^{(n)})\|_f^2\end{aligned}$$

It follows from Theorem 2 in Cohen *et al.* (2013) that

$$\mathbb{E}(\|g_m^{(n)}\|_f^2 \mathbf{1}_{\Omega_m\cap\Lambda_m}) \le 4\frac{\mathfrak{c}}{\log(n)}\|g\|_f^2.$$

Moreover

$$\mathbb{E}(\|(\hat{b}_m - b_m^{(n)})\|_f^2 \mathbf{1}_{\Omega_m \cap \Lambda_m}) \leq 2\mathbb{E}(\|(\hat{b}_m - b_m^{(n)})\|_n^2 \mathbf{1}_{\Omega_m \cap \Lambda_m}) = 2\mathbb{E}(\|\Pi_m \sigma\varepsilon\|_n^2 \mathbf{1}_{\Omega_m}),$$

and we just proved in (29) that this term is less than $(4/n)\mathrm{Tr}(\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2})$. Joining all terms gives the result. □

6.4. **Proof of Theorem 3.1.** We denote by $\widehat{M}_n$ the maximal element of $\widehat{\mathcal{M}}_n$ (see (16)) and by M_n the maximal element of $\mathcal{M}_n$ (see (13)). We need also:

$$\mathcal{M}_n^+ = \left\{ m \in \mathbb{N}, \quad c_\varphi^2 m \, (\|\Psi_m^{-1}\|_{\mathrm{op}}^2 \vee 1) \leq 4\mathfrak{d} \frac{n}{\log(n)} \right\}, \tag{31}$$

with $\mathfrak{d}$ give in (16). Let M_n^+ denote the maximal element of $\mathcal{M}_n^+$. Heuristically, with large probability, considering the constants associated with the sets, we should have $M_n \leq \widehat{M}_n \leq M_n^+$ or equivalently $\mathcal{M}_n \subset \widehat{\mathcal{M}}_n \subset \mathcal{M}_n^+$, and on this set, we really bound the risk; otherwise, we bound the probability of the complement. More precisely, we denote by

$$\Xi_n := \left\{ \mathcal{M}_n \subset \widehat{\mathcal{M}}_n \subset \mathcal{M}_n^+ \right\}, \tag{32}$$

and use that Lemma 6.6 in Comte and Genon-Catalot (2018) states that: for c a positive constant,

$$\mathbb{P}(\Xi_n^c) = \mathbb{P}\left(\left\{ \mathcal{M}_n \not\subset \widehat{\mathcal{M}}_n \text{ or } \widehat{\mathcal{M}}_n \not\subset \mathcal{M}_n^+ \right\} \right) \leq \frac{c}{n^2}. \tag{33}$$

Then we write the decomposition:

$$\widehat{b}_{\hat{m}} - b_A = (\widehat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_n} + (\widehat{b}_{\hat{m}} - b_A)\mathbf{1}_{\Xi_n^c}. \tag{34}$$

Proceeding as in (27), we get that

$$\mathbb{E}\big[\|b - \hat{b}_{\hat{m}}\|_n^2 \mathbf{1}_{\Xi_n^c}\big] \leq \frac{c'}{n}.$$

The following Lemma allows to obtain the first Inequality of Theorem 3.1.

Lemma 6.3. *Under the assumptions of Theorem 3.1, there exists κ_0 such that for $\kappa \geq \kappa_0$, we have (see (17))*

$$\mathbb{E}\big[\|\hat{b}_{\hat{m}} - b_A\|_n^2 \mathbf{1}_{\Xi_n}\big] \leq C \inf_{m \in \mathcal{M}_n} \left(\inf_{t \in S_m} \|t - b_A\|_f^2 \; + \kappa c_\varphi^2 \frac{V(m)}{n} \right) + \frac{C'}{n}$$

where C is a numerical constant and C' is a constant depending on f, b, σ_ε.

The second inequality of Theorem 3.1 can be deduced from the first one as in Comte and Genon-Catalot (2018), Section 6.3.4. □

6.5. **Proof of Lemma 6.3.** To begin with, we note that $\gamma_n(\hat{b}_m) = -\|\hat{b}_m\|_n^2$. Indeed, using formula (4) and $\widehat{\Phi}_m'\widehat{\Phi}_m = n\widehat{\Psi}_m$, we have

$$\gamma_n\big(\hat{b}_m\big) = \big\|\widehat{\Phi}_m \vec{\hat{a}}^{(m)}\big\|_n^2 - 2\big(\vec{\hat{a}}^{(m)}\big)'\widehat{\Phi}_m' \vec{Y} = -\big(\vec{\hat{a}}^{(m)}\big)'\widehat{\Phi}_m' \vec{Y} = -\big\|\widehat{\Phi}_m \vec{\hat{a}}^{(m)}\big\|_n^2.$$

Consequently, we can write

$$\hat{m} = \arg \min_{m \in \widehat{\mathcal{M}}_n} \{\gamma_n(\hat{b}_m) + \widehat{\mathrm{pen}}(m)\},$$

where $\widehat{\text{pen}}(m)$ is defined by (15). Thus, using the definition of the contrast, we have, for any $m \in \widehat{\mathcal{M}}_n$, and any $b_m \in S_m$,

$$\gamma_n(\hat{b}_{\hat{m}}) + \widehat{\text{pen}}(\hat{m}) \leq \gamma_n(b_m) + \widehat{\text{pen}}(m). \tag{35}$$

On the set $\Xi_n = \left\{\mathcal{M}_n \subset \widehat{\mathcal{M}}_n \subset \mathcal{M}_n^+\right\}$, we have $\hat{m} \leq \widehat{M}_n \leq M_n^+$ and either $M_n \leq \hat{m} \leq \widehat{M}_n \leq M_n^+$ or $\hat{m} < M_n \leq \widehat{M}_n \leq M_n^+$. In the first case, $\hat{m}$ is upper and lower bounded by deterministic bounds, and in the second,

$$\hat{m} = \arg\min_{m \in \mathcal{M}_n} \{\gamma_n(\hat{b}_m) + \widehat{\text{pen}}(m)\}.$$

Thus, on Ξ_n, (35) holds for any $m \in \mathcal{M}_n$ and any $b_m \in S_m$. The decomposition $\gamma_n(t) - \gamma_n(s) = \|t-b\|_n^2 - \|s-b\|_n^2 + 2\nu_n(t-s)$, where $\nu_n(t) = \langle \overrightarrow{\sigma\varepsilon}, t\rangle_n$, yields, for any $m \in \mathcal{M}_n$ and any $b_m \in S_m$,

$$\|\hat{b}_{\hat{m}} - b\|_n^2 \leq \|b_m - b\|_n^2 + 2\nu_n(\hat{b}_{\hat{m}} - b_m) + \widehat{\text{pen}}(m) - \widehat{\text{pen}}(\hat{m}).$$

We introduce, for $\|t\|_f^2 = \int t^2(u) f(u) du$, the unit ball

$$B^f_{m,m'}(0,1) = \{t \in S_m + S_{m'}, \|t\|_f = 1\}$$

and the set

$$\Omega_n = \left\{ \left| \frac{\|t\|_n^2}{\|t\|_f^2} - 1 \right| \leq \frac{1}{2}, \ \forall t \in \bigcup_{m,m' \in \mathcal{M}_n^+} (S_m + S_{m'}) \setminus \{0\} \right\}. \tag{36}$$

We start by studying the expectation on Ω_n. On this set, the following inequality holds: $\|t\|_f^2 \leq 2\|t\|_n^2$. We get, on $\Xi_n \cap \Omega_n$,

$$\begin{aligned}
\|\hat{b}_{\hat{m}} - b\|_n^2 \leq & \|b_m - b\|_n^2 + \frac{1}{8}\|\hat{b}_{\hat{m}} - b_m\|_f^2 + (8 \sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_n^2(t) + \widehat{\text{pen}}(m) - \widehat{\text{pen}}(\hat{m})) \\
\leq & \Big(1 + \frac{1}{2}\Big)\|b_m - b\|_n^2 + \frac{1}{2}\|\hat{b}_{\hat{m}} - b\|_n^2 + 8\Big(\sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_n^2(t) - p(m,\hat{m})\Big)_+ \\
& + \widehat{\text{pen}}(m) + 8p(m,\hat{m}) - \widehat{\text{pen}}(\hat{m}).
\end{aligned} \tag{37}$$

Here we state the following Lemma:

Lemma 6.4. *Assume that* **(A1)**, **(A3)** *and* **(A4)** *hold, and that* $\mathbb{E}(\varepsilon_1^{10}) < +\infty$, $\mathbb{E}(\sigma_A^{10}(X_1)) < +\infty$. *Then* $\nu_n(t) = \langle \overrightarrow{\sigma\varepsilon}, t\rangle_n$ *satisfies*

$$\mathbb{E}\left[\Big(\sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_n^2(t) - p(m,\hat{m})\Big)_+ \mathbf{1}_{\Xi_n \cap \Omega_n} \right] \leq \frac{C}{n}$$

where $p(m,m') = \sup(p(m), p(m'))$ *with* $p(m) = 8mV(m)/n$ *(see (17)).*

For $\kappa \geq 64$, $8p(m,m') \leq \text{pen}(m) + \text{pen}(m')$. Therefore, plugging the result of Lemma 6.4 in (37) and taking expectation yield that

$$\begin{aligned}
\frac{1}{2}\mathbb{E}(\|\hat{b}_{\hat{m}} - b\|_n^2 \mathbf{1}_{\Xi_n \cap \Omega_n}) \leq & \frac{3}{2}\|b_m - b\|_n^2 + \text{pen}(m) + \frac{C}{n} \\
& + \mathbb{E}(\widehat{\text{pen}}(m)\mathbf{1}_{\Xi_n \cap \Omega_n}) + \mathbb{E}[(\text{pen}(\hat{m}) - \widehat{\text{pen}}(\hat{m}))_+ \mathbf{1}_{\Xi_n \cap \Omega_n}).
\end{aligned}$$

Lemma 6.5. *Under the assumptions of Theorem 3.1, there exist constants $c_1, c_2 > 0$ such that for $m \in \mathcal{M}_n$ and $\hat{m} \in \widehat{\mathcal{M}}_n$,*

$$\mathbb{E}(\widehat{\mathrm{pen}}(m)\mathbf{1}_{\Xi_n \cap \Omega_n}) \leq c_1 \mathrm{pen}(m) + \frac{c_2}{n} \tag{38}$$

$$\mathbb{E}[(\mathrm{pen}(\hat{m}) - \widehat{\mathrm{pen}}(\hat{m}))_+ \mathbf{1}_{\Xi_n \cap \Omega_n}) \leq \frac{c_2}{n}. \tag{39}$$

Lemma 6.5 concludes the study of the expectation of the empirical risk on $\Xi_n \cap \Omega_n$. □

6.6. **Proof of Lemma 6.4.** In order to apply the Talagrand Inequality, we make the following decompositions, where k_n, ℓ_n are to be chosen later:

$$\varepsilon_i = \eta_i + \xi_i,\ \eta_i = \varepsilon_i \mathbf{1}_{|\varepsilon_i| \leq k_n} - \mathbb{E}\big[\varepsilon_i \mathbf{1}_{|\varepsilon_i| \leq k_n}\big],$$

and set

$$\tau(x) = \sigma_A(x)\mathbf{1}_{\sigma_A^2(x) \leq \ell_n},\ \text{and } \sigma_A(x) = \tau(x) + \theta(x),$$

Then we have $\nu_n(t) = \nu_{n,1}(t) + \nu_{n,2}(t) + \nu_{n,3}(t)$, where

$$\nu_{n,1}(t) = \langle \tau\eta, t\rangle_n, \quad \nu_{n,2}(t) = \langle \theta\eta, t\rangle_n, \quad \nu_{n,3}(t) = \langle \sigma\xi, t\rangle_n.$$

We write

$$\begin{aligned}\Big(\sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_n^2(t) - p(m,\hat{m})\Big)_+ \ \leq\ & \Big(\sup_{t \in B^f_{\hat{m},m}(0,1)} 2\nu_{n,1}^2(t) - p(m,\hat{m})\Big)_+ \\ & +4 \sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_{n,2}^2(t) + 4 \sup_{t \in B^f_{\hat{m},m}(0,1)} \nu_{n,3}^2(t).\end{aligned} \tag{40}$$

We successively bound the three terms. To bound the first term, we use the Talagrand inequality applied to the process $\nu_{n,1}$.
Let $t = \sum_{j=0}^{m-1} a_j \varphi_j$ where $\vec{a} = \Psi_m^{-1/2}\vec{u}$ ($a_j = \sum_{k=0}^{m-1} [\Psi_m^{-1/2}]_{j,k} u_k$) and $\|\vec{u}\|_{2,m} = 1$. Then, $\|t\|_f^2 = \vec{a}'\Psi_m\vec{a} = 1$ and

$$t = \sum_{k=0}^{m-1} u_k \left(\sum_{j=0}^{m-1} [\Psi_m^{-1/2}]_{j,k}\varphi_j\right) \quad \text{and} \quad \langle \tau\eta, t\rangle_n^2 \leq \sum_{k=0}^{m-1} \langle \tau\eta, \sum_{j=0}^{m-1} [\Psi_m^{-1/2}]_{j,k}\varphi_j\rangle_n^2.$$

Therefore,

$$\mathbb{E}\left(\sup_{t \in S_m, \|\Psi_m^{1/2}\vec{a}\|_{2,m}=1} \langle \tau\eta, t\rangle_n^2\right) \leq \sum_{k=0}^{m-1} \mathbb{E}\left(\langle \tau\eta, \sum_{j=0}^{m-1} [\Psi_m^{-1/2}]_{j,k}\varphi_j\rangle_n^2\right).$$

Then using that the (η_i, X_i) are independent and the terms are centered, we get

$$\begin{aligned}
\mathbb{E}\left(\langle \tau\eta, \sum_{j=0}^{m-1}[\Psi_m^{-1/2}]_{j,k}\varphi_j\rangle_n^2\right) &= \frac{1}{n}\mathbb{E}\left[\tau^2(X_1)\left(\sum_{j=0}^{m-1}[\Psi_m^{-1/2}]_{j,k}\varphi_j(X_1)\right)^2\right] \\
&\leq \frac{1}{n}\mathbb{E}\left[\sigma^2(X_1)\left(\sum_{j=0}^{m-1}[\Psi_m^{-1/2}]_{j,k}\varphi_j(X_1)\right)^2\right] \\
&= \frac{1}{n}\sum_{0\leq j,\ell\leq m-1}[\Psi_m^{-1/2}]_{j,k}[\Psi_m^{-1/2}]_{\ell,k}[\Psi_{m,\sigma^2}]_{j,\ell} \\
&= \frac{1}{n}\mathrm{Tr}\left[\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\right] \leq \frac{mV(m)}{n}.
\end{aligned}$$

Therefore

$$\mathbb{E}\Big[\sup_{t\in B^f_{m',m}(0,1)}\nu_{n,1}^2(t)\Big] \leq \frac{mV(m)\ \vee\ m'V(m')}{n} := H^2.$$

Next

$$\begin{aligned}
\sup_{t\in B^f_{m',m}(0,1)}\mathrm{Var}(\eta_1\tau(X_1)t(X_1)) &\leq \mathbb{E}\big[\eta_1^2\big]\sup_{t\in B^f_{m',m}(0,1)}\mathbb{E}\big[\tau^2(X_1)t^2(X_1)\big] \\
&\leq \sup_{t\in B^f_{m',m}(0,1)}\mathbb{E}\big[\sigma^2(X_1)t^2(X_1)\big]
\end{aligned}$$

Now, to deal with this last term, we remark

$$\begin{aligned}
\sup_{t\in B^f_m(0,1)}\int t^2\sigma^2 f &= \sup_{\|\vec{u}\|^2_{2,m}=1}\int\left(\sum_{k=0}^{m-1}u_k\left(\sum_{j=0}^{m-1}[\Psi_m^{-1/2}]_{j,k}\varphi_j\right)\right)^2\sigma^2 f \\
&= \sup_{\|\vec{u}\|^2_{2,m}=1}\int\left(\sum_{j=0}^{m-1}\left(\sum_{k=0}^{m-1}u_k[\Psi_m^{-1/2}]_{j,k}\right)\varphi_j\right)^2\sigma^2 f \\
&= \sup_{\|\vec{u}\|^2_{2,m}=1}\int\sum_{j,\ell=0}^{m-1}\left(\sum_{k=0}^{m-1}u_k[\Psi_m^{-1/2}]_{j,k}\right)\left(\sum_{k=0}^{m-1}u_k[\Psi_m^{-1/2}]_{\ell,k}\right)\varphi_j\varphi_\ell\sigma^2 f \\
&= \sup_{\|\vec{u}\|^2_{2,m}=1}\sum_{j,\ell=0}^{m-1}\left(\sum_{k=0}^{m-1}u_k[\Psi_m^{-1/2}]_{j,k}\right)\left(\sum_{k=0}^{m-1}u_k[\Psi_m^{-1/2}]_{\ell,k}\right)[\Psi_{m,\sigma^2}]_{j,\ell} \\
&= \sup_{\|\vec{u}\|^2_{2,m}=1}\vec{u}'\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\vec{u} = \|\Psi_m^{-1/2}\Psi_{m,\sigma^2}\Psi_m^{-1/2}\|_{\mathrm{op}} \leq V(m).
\end{aligned}$$

As a consequence, we have

$$\sup_{t\in B^f_{m',m}(0,1)}\mathrm{Var}(\eta_1\tau(X_1)t(X_1)) \leq V(m)\vee V(m') := v$$

Lastly, setting $m^* := \max(m, m')$, for $t = \sum_{j=0}^{m^*-1} a_j \varphi_j \in B^f_{m',m}(0,1)$, $\vec{a}'\vec{a} = \vec{u}'\Psi^{-1}_{m^*}\vec{u}$ with $\vec{u}'\vec{u} = 1$. Therefore,

$$\sup_{t\in B^f_{m',m}(0,1)} \sup_{(u,x)} \left(|u|\mathbf{1}_{|u|\le k_n}|\tau(x)||t(x)|\right) \le k_n\sqrt{\ell_n} \sup_{t\in B^f_{m',m}(0,1)} \sup_x |t(x)|$$
$$\le c_\varphi k_n\sqrt{\ell_n}\sqrt{m^*\|\Psi^{-1}_{m^*}\|_{\rm op}} := M_1.$$

Consequently, the Talagrand Inequality implies, for $p(m,m')$ defined in Lemma 6.4, i.e. $\frac{1}{2}p(m,m') = 2(1+2\epsilon)H^2$ with $\epsilon = 1/2$,

$$\mathbb{E}\Big(\sup_{t\in B^f_{m,m'}(0,1)} [\nu_{n,1}]^2(t) - \frac{1}{2}p(m,m')\Big)_+ \quad \le \quad \frac{C_1}{n}\left(T_1 + T_2\right),$$

with $\quad T_1 = V(m^*)\, e^{-m^*/12}, \quad T_2 = \frac{k_n^2\ell_n\sqrt{m^*}}{\sqrt{n\log(n)}} e^{-C_3\frac{n^{1/2}}{k_n\sqrt{\ell_n}}\frac{V^{1/2}(m^*)}{\|\Psi^{-1}_{m^*}\|^{1/2}_{\rm op}}}$.
We have $V(m^*) \ge 1$, and we choose k_n, ℓ_n such that $k_n^2\ell_n = n^{1/2}$. As $\|\Psi^{-1}_{m^*}\|_{\rm op} \le \sqrt{\mathfrak{d}n/(m\log n)}$, we get, provided that $\log n \ge 1$, $T_2 \le \sqrt{m^*}\exp(-C_5(m^*)^{1/4})$.
As $\|\Psi^{-1/2}_m \Psi_{m,\sigma^2}\Psi^{-1/2}_m\|_{\rm op} \le \|\Psi^{-1}_m\|_{\rm op}\|\Psi_{m,\sigma^2}\|_{\rm op}$ and

$$\|\Psi_{m,\sigma^2}\|_{\rm op} = \sup_{\|\vec{x}\|_{2,m}=1} \vec{x}'\Psi_{m,\sigma^2}\vec{x} = \sup_{\|\vec{x}\|_{2,m}=1} \int (\sum_{j=0}^{m-1} x_j\varphi_j(u))^2\sigma^2(u)f(u)du \le c_\varphi^2 m\mathbb{E}[\sigma^2(X_1)]$$

we get $V(m) \le c\varphi^2 m\|\Psi^{-1}_m\|_{\rm op}\mathbb{E}[\sigma^2(X_1)] + 1$. So, from Assumption **(A4)**, we have

$$\sum_{m'\in\mathcal{M}_n^+} V(m^*)\, e^{-m^*/12} \le c_\varphi^2\Sigma < +\infty.$$

This yields, by summing up all terms over $m' \in \mathcal{M}_n^+$,

$$\mathbb{E}\Big(\sup_{t\in B^f_{\hat{m},m}(0,1)} [\nu_{n,1}]^2(t) - \frac{1}{2}p(m,\hat{m})\Big)_+ \quad \le \quad \sum_{m'} \mathbb{E}\Big(\sup_{t\in B^f_{m',m}(0,1)} [\nu_{n,1}]^2(t) - \frac{1}{2}p(m,m')\Big)_+$$
$$\le \quad \frac{C}{n}. \tag{41}$$

Now, we study the second term in (40). Recall that $M_n^+ \leq n$ is the dimension of the largest space of the collection $\mathcal{M}_n^+$. Then we have

$$\begin{aligned}
\mathbb{E}\Big[\Big(\sup_{t\in B^f_{\hat{m},m}(0,1)} \nu^2_{n,2}(t)\mathbf{1}_{\Xi_n}\Big)_+\Big] &\leq \|\Psi^{-1}_{M_n^+}\|_{\mathrm{op}} \sum_{j=0}^{M_n^+-1} \mathbb{E}\big[\langle \eta\theta, \varphi_j\rangle_n^2\big] \\
&= \|\Psi^{-1}_{M_n^+}\|_{\mathrm{op}} \sum_{j=0}^{M_n^+-1} \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^n \eta_i\theta(X_i)\varphi_j(X_i)\Big) \\
&\leq \frac{c_\varphi^2 M_n^+ \|\Psi^{-1}_{M_n^+}\|_{\mathrm{op}}}{n}\mathbb{E}\big[\eta_1^2\big]\mathbb{E}\big[\theta_A(X_1)^2\big] \\
&\leq c_\varphi^2 \frac{M_n^+ \|\Psi^{-1}_{M_n^+}\|_{\mathrm{op}}}{n}\mathbb{E}\big[\sigma_A^2(X_1)\mathbf{1}_{\sigma_A^2(X_1)>\ell_n}\big] \\
&\leq C\frac{\mathbb{E}\big[|\sigma_A(X_1)|^{2+q}\big]}{\log(n)\ell_n^{q/2}} = C\frac{\mathbb{E}\big[\sigma_A^{10}(X_1)\big]}{n}
\end{aligned}$$

by taking $\ell_n = n^{1/4}$ and $q = 8$.
Let us now study the third term in (40) with $k_n = n^{1/8}$.We have

$$\begin{aligned}
\mathbb{E}\Big[\Big(\sup_{t\in B^f_{\hat{m},m}(0,1)} \nu^2_{n,3}(t)\Big)_+\Big] &\leq \|\Psi^{-1}_{M_n^+}\|_{\mathrm{op}} \sum_{j=0}^{M_n^+-1} \mathbb{E}\big[\langle \xi, \sigma_A\varphi_j\rangle_n^2\big] \\
&= \|\Psi^{-1}_{M_n^+}\|_{\mathrm{op}} \sum_{j=0}^{M_n^+-1} \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^n \xi_i\sigma_A(X_i)\varphi_j(X_i)\Big) \\
&\leq \frac{c_\varphi^2 M_n^+ \|\Psi^{-1}_{M_n^+}\|_{\mathrm{op}}}{n}\mathbb{E}[\sigma^2(X_1)]\mathbb{E}\big[\xi_1^2\big] \\
&\leq c_\varphi^2\mathbb{E}[\sigma^2(X_1)]\frac{M_n^+ \|\Psi^{-1}_{M_n^+}\|_{\mathrm{op}}}{n}\mathbb{E}\big[\varepsilon_1^2\mathbf{1}_{|\varepsilon_1|>k_n}\big] \leq C\frac{\mathbb{E}\big[\varepsilon_1^{10}\big]}{n}.
\end{aligned}$$

This bound together with (41) plugged in (40) gives the result of Lemma 6.4. □

6.7. **Proof of Lemma 6.5.** Take pen(m) as in (17) and set $\widehat{\mathrm{pen}}(m) = \kappa' m\widehat{V}(m)/n$ to determine κ'.
On Ω_m (see (23)), $|\|t\|_n^2/\|t\|_f^2 - 1| \leq 1/2$, which implies $2/3 \leq \|t\|_f^2/\|t\|_n^2 \leq 2$, so that

$$\begin{aligned}
\|\widehat{\Psi}_m^{-1/2}\widehat{\Psi}_{m,\sigma^2}\widehat{\Psi}_m^{-1/2}\|_{\mathrm{op}} &= \sup_{t\in S_m, \|t\|_n^2=1} \|t\sigma\|_n^2 = \sup_{t\in S_m}\frac{\|t\sigma\|_n^2}{\|t\|_n^2} = \sup_{t\in S_m}\frac{\|t\sigma\|_n^2}{\|t\|_f^2}\frac{\|t\|_f^2}{\|t\|_n^2} \\
&\leq 2\sup_{t\in S_m}\frac{\|t\sigma\|_n^2}{\|t\|_f^2} = 2\|\Psi_m^{-1/2}\widehat{\Psi}_{m,\sigma^2}\Psi_m^{-1/2}\|_{\mathrm{op}}.
\end{aligned}$$

Thus

$$\begin{aligned}
\widehat{V}(m)\mathbf{1}_{\Omega_m} &\leq (2\|\Psi_m^{-1/2}\widehat{\Psi}_{m,\sigma^2}\Psi_m^{-1/2}\|_{\mathrm{op}} + 1)\mathbf{1}_{\Omega_m} \\
&\leq 2V(m)\mathbf{1}_{\Omega_m} + 2\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma^2} - \Psi_{m,\sigma^2})\Psi_m^{-1/2}\|_{\mathrm{op}}\mathbf{1}_{\Omega_m}
\end{aligned}$$

Now we set

$$\sigma^2(x) = \sigma^2_m(x) + s^2_m(x), \text{ where } \sigma^2_m(x) := \sigma^2(x)\mathbf{1}_{\sigma^2(x)\le c_m}.$$

We decompose accordingly

$$\widehat{\Psi}_{m,\sigma^2} - \Psi_{m,\sigma^2} = [\widehat{\Psi}_{m,\sigma^2_m} - \Psi_{m,\sigma^2_m}] + [\widehat{\Psi}_{m,s^2_m} - \Psi_{m,s^2_m}].$$

Let

$$\Omega_{m,\sigma^2_m} := \left\{ \|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma^2_m} - \Psi_{m,\sigma^2_m})\Psi_m^{-1/2}\|_{\rm op} \le \frac{1}{2} \right\}.$$

We can prove (see section 6.8) that, for $c_m \ge 1$,

$$\mathbb{P}(\Omega^c_{m,\sigma^2_m}) \le \frac{c}{n^4} \text{ if } c_\varphi^2 m c_m^2 \|\Psi_m^{-1}\|_{\rm op} \le \frac{3}{8}\frac{n}{\log(n)}. \tag{42}$$

Under our constraint (13), the above holds for $c_m = m^{\alpha/4}$ as under **(A3)**, $\|\Psi_m^{-1}\|_{\rm op} \ge \sqrt{c^\star} m^{\alpha/2}$.
We write

$$\begin{aligned}\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma^2} - \Psi_{m,\sigma^2})\Psi_m^{-1/2}\|_{\rm op} &\le \|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma^2_m} - \Psi_{m,\sigma^2_m})\Psi_m^{-1/2}\|_{\rm op}\mathbf{1}_{\Omega_{m,\sigma^2_m}} \\ &+\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma^2_m} - \Psi_{m,\sigma^2_m})\Psi_m^{-1/2}\|_{\rm op}\mathbf{1}_{\Omega^c_{m,\sigma^2_m}} + \|\Psi_m^{-1/2}(\widehat{\Psi}_{m,s^2_m} - \Psi_{m,s^2_m})\Psi_m^{-1/2}\|_{\rm op}.\end{aligned}$$

and we study the three terms.

First,

$$T_1 = \|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma^2} - \Psi_{m,\sigma^2})\Psi_m^{-1/2}\|_{\rm op}\mathbf{1}_{\Omega_{m,\sigma^2_m}} \le \frac{1}{2} \le \frac{V(m)}{2}.$$

Next,

$$\begin{aligned} T_2 &= \mathbb{E}(\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma^2_m} - \Psi_{m,\sigma^2_m})\Psi_m^{-1/2}\|_{\rm op}\mathbf{1}_{\Omega^c_{m,\sigma^2_m}}) \\ &\le \mathbb{E}^{1/2}(\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma^2_m} - \Psi_{m,\sigma^2_m})\Psi_m^{-1/2}\|^2_{\rm op})\mathbb{P}^{1/2}(\Omega^c_{m,\sigma^2_m}). \end{aligned} \tag{43}$$

So, we must bound $\mathbb{E}(\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma^2_m} - \Psi_{m,\sigma^2_m})\Psi_m^{-1/2}\|^2_{\rm op})$, and the term

$$T_3 = \mathbb{E}(\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,s^2_m} - \Psi_{m,s^2_m})\Psi_m^{-1/2}\|_{\rm op}).$$

We have

$$\begin{aligned} T_3^2 &\le \mathbb{E}\left(\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,s^2_m} - \Psi_{m,s^2_m})\Psi_m^{-1/2}\|^2_{\rm op}\right) \le \|\Psi_m^{-1}\|^2_{\rm op}\mathbb{E}\left(\sup_{\|\vec{x}\|_{2,m}=1}\|(\widehat{\Psi}_{m,s^2_m} - \Psi_{m,s^2_m})\vec{x}\|^2_{2,m}\right) \\ &\le \|\Psi_m^{-1}\|^2_{\rm op}\mathbb{E}\left\{\sup_{\|x\|_{2,m}=1}\sum_{j=0}^{m-1}\left(\sum_{k=0}^{m-1}\left[\widehat{\Psi}_{m,s^2_m} - \Psi_{m,s^2_m}\right]_{j,k}x_k\right)^2\right\} \\ &\le \|\Psi_m^{-1}\|^2_{\rm op}\mathbb{E}\left\{\sum_{j=0}^{m-1}\sum_{k=0}^{m-1}\left[\widehat{\Psi}_{m,s^2_m} - \Psi_{m,s^2_m}\right]^2_{j,k}\right\} = \frac{\|\Psi_m^{-1}\|^2_{\rm op}}{n}\sum_{j=0}^{m-1}\sum_{k=0}^{m-1}{\rm Var}\left(\varphi_j(X_1)\varphi_k(X_1)s^2_m(X_1)\right) \\ &\le \frac{c_\varphi^4 m^2\|\Psi_m^{-1}\|^2_{\rm op}}{n}\mathbb{E}(\sigma^4(X_1)\mathbf{1}_{\sigma^2(X_1)\ge c_m}) \le \frac{c_\varphi^4 m\|\Psi_m^{-1}\|^2_{\rm op}}{n}\frac{m\mathbb{E}(\sigma^{4+2q}(X_1))}{c_m^q} \le \frac{c}{m^6}, \end{aligned} \tag{44}$$

for $\alpha q/4 = 7$ as $c_m^q = m^{\alpha q/4}$. Thus $4+2q = 4+56/\alpha$ and we require $\mathbb{E}(\sigma^{4+56/\alpha}(X_1)) < +\infty$. Using (44) and the Schwarz Inequality yields

$$T_3 = \mathbb{E}(\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,s_m^2} - \Psi_{m,s_m^2})\Psi_m^{-1/2}\|_{\mathrm{op}}) \leq \frac{c}{m^3}.$$

Similarly,

$$\begin{aligned}\mathbb{E}\left(\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma_m^2} - \Psi_{m,\sigma_m^2})\Psi_m^{-1/2}\|_{\mathrm{op}}^2\right) &\leq \frac{c_\varphi^4 m^2 \|\Psi_m^{-1}\|_{\mathrm{op}}^2}{n}\mathbb{E}(\sigma^4(X_1)\mathbf{1}_{\sigma^2(X_1)\leq c_m}) \\ &\leq c\mathbb{E}[\sigma^4(X_1)]\frac{n}{\log(n)}. \qquad (45)\end{aligned}$$

This yields, plugging (45) and (42) in (43),

$$T_2 = \mathbb{E}(\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma_m^2} - \Psi_{m,\sigma_m^2})\Psi_m^{-1/2}\|_{\mathrm{op}}\mathbf{1}_{\Omega^c_{m,\sigma_m^2}}) \leq c/n^{3/2},$$

Thus for $m \geq 1$, and using that under our assumptions $m \in \mathcal{M}_n^+$ satisfies $m \leq \sqrt{n}$ as under **(A3)**, $\|\Psi_m^{-1}\|_{\mathrm{op}}^2 \geq c^\star m$,

$$\frac{m}{n}T_2 = \frac{m}{n}\mathbb{E}(\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,\sigma_m^2} - \Psi_{m,\sigma_m^2})\Psi_m^{-1/2}\|_{\mathrm{op}}\mathbf{1}_{\Omega^c_{m,\sigma_m^2}}) \leq c/n^2$$

and

$$\frac{m}{n}T_3 = \frac{m}{n}\mathbb{E}(\|\Psi_m^{-1/2}(\widehat{\Psi}_{m,s_m^2} - \Psi_{m,s_m^2})\Psi_m^{-1/2}\|_{\mathrm{op}}) \leq c/n.$$

This implies that

$$\mathbb{E}(\widehat{\mathrm{pen}}(m)\mathbf{1}_{\Xi_n\cap\Omega_n}) = \mathbb{E}(\kappa'\frac{m}{n}\widehat{V}(m)\mathbf{1}_{\Xi_n\cap\Omega_n}) \leq 2\kappa'\frac{m}{n}V(m) + \frac{c}{n}$$

and concludes the proof of (38).

In the same way, on Ω_m, for all m,

$$\begin{aligned}&(\mathrm{pen}(m) - \widehat{\mathrm{pen}}(m))_+ = \frac{m}{n}\left(\kappa V(m) - \kappa'\widehat{V}(m)\right)_+ \\ \leq\ &\frac{m}{n}\left(\kappa\|\Psi_m^{-1/2}(\Psi_{m,\sigma^2} - \widehat{\Psi}_{m,\sigma^2})\Psi_m^{-1/2}\|_{\mathrm{op}} + \kappa\|\Psi_m^{-1/2}\widehat{\Psi}_{m,\sigma^2}\Psi_m^{-1/2}\|_{\mathrm{op}} + \kappa - \kappa'\widehat{V}(m)\right)_+ \\ \leq\ &\frac{m}{n}\left(\kappa\|\Psi_m^{-1/2}(\Psi_{m,\sigma^2} - \widehat{\Psi}_{m,\sigma^2})\Psi_m^{-1/2}\|_{\mathrm{op}} - \frac{\kappa}{2} + (\frac{3}{2}\kappa - \kappa')\widehat{V}(m)\right)_+\end{aligned}$$

using that on Ω_m,

$$\|\Psi_m^{-1/2}\widehat{\Psi}_{m,\sigma^2}\Psi_m^{-1/2}\|_{\mathrm{op}} + \frac{3}{2} \leq \frac{3}{2}\|\widehat{\Psi}_m^{-1/2}\widehat{\Psi}_{m,\sigma^2}\widehat{\Psi}_m^{-1/2}\|_{\mathrm{op}} + \frac{3}{2} = \frac{3}{2}\widehat{V}(m).$$

Then, using the same decomposition on Ω_{m,σ^2} and its complement as above,

$$\begin{aligned}&(\mathrm{pen}(m) - \widehat{\mathrm{pen}}(m))_+ \\ \leq\ &\frac{m}{n}\Big(\kappa\|\Psi_m^{-1/2}(\Psi_{m,\sigma_m^2} - \widehat{\Psi}_{m,\sigma_m^2})\Psi_m^{-1/2}\|_{\mathrm{op}}\mathbf{1}_{\Omega^c_{m,\sigma_m^2}} + \kappa\|\Psi_m^{-1/2}(\Psi_{m,s_m^2} - \widehat{\Psi}_{m,s_m^2})\Psi_m^{-1/2}\|_{\mathrm{op}} \\ &\qquad +(\frac{3}{2}\kappa - \kappa')\widehat{V}(m)\Big)_+\end{aligned}$$

Choose $(\frac{3}{2}\kappa - \kappa') \leq 0$, hence $\kappa' = 2\kappa$ is convenient, to get

$$\begin{aligned}(\mathrm{pen}(\hat{m}) - \widehat{\mathrm{pen}}(\hat{m}))_+ &\leq \frac{\hat{m}}{n}\kappa \|\Psi_{\hat{m}}^{-1/2}(\Psi_{\hat{m},\sigma^2} - \widehat{\Psi}_{\hat{m},\sigma^2})\Psi_{\hat{m}}^{-1/2}\|_{\mathrm{op}} \mathbf{1}_{\Omega^c_{\hat{m},\sigma^2_{\hat{m}}}} \\ &\quad + \kappa \frac{\hat{m}}{n}\|\Psi_{\hat{m}}^{-1/2}(\Psi_{\hat{m},s^2_{\hat{m}}} - \widehat{\Psi}_{\hat{m},s^2_{\hat{m}}})\Psi_{\hat{m}}^{-1/2}\|_{\mathrm{op}}\end{aligned}$$

Then

$$\begin{aligned}&\mathbb{E}\left(\mathrm{pen}(\hat{m}) - \widehat{\mathrm{pen}}(\hat{m})\right)_+ \mathbf{1}_{\Omega_n \cap \Xi_n}) \\ \leq &\sum_{m' \in \mathcal{M}_n^+} \mathbb{E}\left[\frac{m'}{n}\kappa\left(\|\Psi_{m'}^{-1/2}(\Psi_{m',\sigma^2_{m'}} - \widehat{\Psi}_{m',\sigma^2_{m'}})\Psi_{m'}^{-1/2}\|_{\mathrm{op}} \mathbf{1}_{\Omega^c_{m',\sigma^2_{m'}}}\right.\right. \\ &\qquad \left.\left. + \|\Psi_{m'}^{-1/2}(\Psi_{m',s^2_{m'}} - \widehat{\Psi}_{m',s^2_{m'}})\Psi_{m'}^{-1/2}\|_{\mathrm{op}}\right)\right] \leq c/n\end{aligned}$$

as the cardinality of $\mathcal{M}_n^+$ is smaller than n and $m \leq n$ and $(1/n)\sum_{m'}(1/(m')^2) \leq c/n$. □

6.8. **Proof of Inequality (42).** To get the announced result, we apply a Bernstein matrix inequality (see Theorem 7.2). Thus we write $\Psi_m^{-1/2}\widehat{\Psi}_{m,\sigma^2_m}\Psi_m^{-1/2}$ as a sum of a sequence of independent matrices

$$\Psi_m^{-1/2}\widehat{\Psi}_{m,\sigma^2_m}\Psi_m^{-1/2} = \frac{1}{n}\sum_{i=1}^n \mathbf{K}_m(X_i),$$

with

$$\mathbf{K}_m(X_i) = \Psi_m^{-1/2}\Sigma_m(X_i)\Psi_m^{-1/2}, \quad \Sigma_m(X_i) = (\varphi_j(X_i)\varphi_k(X_i)\sigma_m^2(X_i))_{0\leq j,k\leq m-1}.$$

Considering $\mathbf{S}_m = \frac{1}{n}\sum_{i=1}^n \mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right]$, we compute $\mathbf{L}$ and a bound on $\nu(\mathbf{S}_m)$ to apply Theorem 7.2.

• Bound on $\|\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]\|_{\mathrm{op}}/n$.

First we can write that

$$\|\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]\|_{\mathrm{op}} \leq \|\mathbf{K}_m(X_1)\|_{\mathrm{op}} + \|\mathbb{E}\left[\mathbf{K}_m(X_1)\right]\|_{\mathrm{op}},$$

and we bound the first term, the other one being similar. As $\mathbf{K}_m(X_1)$ is symmetric and nonnegative a.s., we have a.s.

$$\begin{aligned}\|\mathbf{K}_m(X_1)\|_{\mathrm{op}} &= \sup_{\|\vec{x}\|_{2,m}=1} \sum_{0\leq j,k\leq m-1} [\Psi_m^{-1/2}\vec{x}]_j [\Psi_m^{-1/2}\vec{x}]_k \varphi_j(X_1)\varphi_k(X_1)\sigma_m^2(X_1) \\ &\leq \|\Psi_m^{-1}\|_{\mathrm{op}} \sup_{\|\vec{y}\|_{2,m}\leq 1} \sum_{0\leq j,k\leq m-1} y_j y_k \varphi_j(X_1)\varphi_k(X_1)\sigma_m^2(X_1) \\ &= \|\Psi_m^{-1}\|_{\mathrm{op}} \sup_{\|\vec{x}\|_{2,m}\leq 1} \left[\left(\sum_{j=0}^{m-1} y_j \varphi_j(X_1)\sigma_m^2(X_1)\right)^2\right] \leq c_\varphi^2 m c_m \|\Psi_m^{-1}\|_{\mathrm{op}}.\end{aligned}$$

So we get that, a.s.

$$\frac{1}{n}\|\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]\|_{\mathrm{op}} \leq \frac{2c_\varphi^2 m c_m \|\Psi_m^{-1}\|_{\mathrm{op}}}{n} := \mathbf{L}. \tag{46}$$

• Bound on $\nu(\mathbf{S}_m) = \| \sum_{i=1}^n \mathbb{E}\left[(\mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right])' (\mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right])\right] \|_{\mathrm{op}}/n^2$. By definition of the operator norm we have

$$\begin{aligned}
\nu(\mathbf{S}_m) &= \frac{1}{n^2} \sup_{\|\vec{x}\|_{2,m}=1} \vec{x}' \sum_{i=1}^n \mathbb{E}\left[(\mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right])' (\mathbf{K}_m(X_i) - \mathbb{E}\left[\mathbf{K}_m(X_i)\right])\right] \vec{x} \\
&= \frac{1}{n} \sup_{\|\vec{x}\|_{2,m}=1} \vec{x}' \, \mathbb{E}\left[(\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right])' (\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right])\right] \vec{x} \\
&= \frac{1}{n} \sup_{\|\vec{x}\|_{2,m}=1} \mathbb{E}\, \|(\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]) \vec{x}\|_{2,m}^2
\end{aligned}$$

It yields that, for $\vec{x}' = (x_0, \ldots, x_{m-1})$,

$$\begin{aligned}
\mathbb{E}_1 := \mathbb{E}\, \|(\mathbf{K}_m(X_1) - \mathbb{E}\left[\mathbf{K}_m(X_1)\right]) \vec{x}\|_{2,m}^2 &= \sum_{j=0}^{m-1} \mathrm{Var}\left[\sum_{k=0}^{m-1} [\Psi_m^{-1/2} \Sigma_m(X_1)]_{j,k} [\Psi_m^{-1/2} \vec{x}]_k\right] \\
&\leq \sum_{j=0}^{m-1} \mathbb{E}\left(\sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} [\Psi_m^{-1/2}]_{j,\ell} [\Sigma_m(X_1)]_{\ell,k} [\Psi_m^{-1/2}\vec{x}]_k\right)^2 \\
&= \sum_{j=0}^{m-1} \mathbb{E}\left[\left(\sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} [\Psi_m^{-1/2}]_{j,\ell} \varphi_\ell(X_1) \varphi_k(X_1) \sigma_m^2(X_1) [\Psi_m^{-1/2}\vec{x}]_k\right)^2\right] \\
&= \sum_{j=0}^{m-1} \mathbb{E}\left[\left(\sigma_m^2(X_1) \sum_{\ell=0}^{m-1} [\Psi_m^{-1/2}]_{j,\ell} \varphi_\ell(X_1) \sum_{k=0}^{m-1} \varphi_k(X_1) [\Psi_m^{-1/2}\vec{x}]_k\right)^2\right]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_1 &\leq \sum_{j=0}^{m-1} \mathbb{E}\left[\left(\sigma_m^2(X_1) [\Psi_m^{-1/2} \overrightarrow{\varphi(X_1)}]_j \, \overrightarrow{\varphi(X_1)}' \Psi_m^{-1/2} \vec{x}\right)^2\right] \\
&= \mathbb{E}\left[\sigma_m^4(X_1) \|\Psi_m^{-1/2} \overrightarrow{\varphi(X_1)}\|_{2,m}^2 \left(\overrightarrow{\varphi(X_1)}' \Psi_m^{-1/2} \vec{x}\right)^2\right] \\
&\leq c_m^2 \sup_{x \in A} \|\Psi_m^{-1/2} \overrightarrow{\varphi(x)}\|_{2,m}^2 \mathbb{E}\left[\vec{x}' \Psi_m^{-1/2} \overrightarrow{\varphi(X_1)} \overrightarrow{\varphi(X_1)}' \Psi_m^{-1/2} \vec{x}\right] \\
&\leq c_m^2 \|\Psi_m^{-1}\|_{\mathrm{op}} c_\varphi^2 m \|\vec{x}\|_{2,m}^2
\end{aligned}$$

Then we get that $\nu(\mathbf{S}_m) \leq \dfrac{c_\varphi^2 m c_m^2 \|\Psi_m^{-1}\|_{\mathrm{op}}}{n}$. Applying Theorem 7.2 yields that for all $u > 0$ and $c_m \geq 1$,

$$\mathbb{P}\left[\|\Psi_m^{-1/2} (\widehat{\Psi}_{m,\sigma_m^2} - \Psi_{m,\sigma_m^2}) \Psi_m^{-1/2} \geq u\right] \leq 2m \exp\left(-\frac{n}{c_\varphi^2 m c_m^2 \|\Psi_m^{-1}\|_{\mathrm{op}}} \frac{u^2/2}{1 + 2u/3)}\right).$$

Then choosing $u = 1/2$ and

$$(32/3) c_\varphi^2 m c_m^2 \|\Psi_m^{-1}\|_{\mathrm{op}} \leq 4n/\log(n)$$

ensures that (42) holds. □

7. Theoretical tools

We recall the Talagrand concentration inequality given in Klein and Rio (2005).

Theorem 7.1. *Consider $n \in \mathbb{N}^*$, $\mathcal{F}$ a class at most countable of measurable functions, and $(X_i)_{i\in\{1,\dots,n\}}$ a family of real independent random variables. Define, for $f \in \mathcal{F}$, $\nu_n(f) = (1/n)\sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)])$, and assume that there are three positive constants M_1, H and v such that $\sup_{f\in\mathcal{F}} \|f\|_\infty \le M_1$, $\mathbb{E}[\sup_{f\in\mathcal{F}} |\nu_n(f)|] \le H$, and $\sup_{f\in\mathcal{F}} (1/n)\sum_{i=1}^n \mathrm{Var}(f(X_i)) \le v$. Then for all $\epsilon > 0$,*

$$\mathbb{E}\left[\left(\sup_{f\in\mathcal{F}} |\nu_n(f)|^2 - 2(1+2\epsilon)H^2\right)_+\right] \le \frac{4}{b}\left(\frac{v}{n}e^{-b\epsilon\frac{nH^2}{v}} + \frac{49M_1^2}{bC^2(\epsilon)n^2}e^{-\frac{\sqrt{2}bC(\epsilon)\sqrt{\epsilon}}{7}\frac{nH}{M_1}}\right)$$

with $C(\epsilon) = (\sqrt{1+\epsilon} - 1) \wedge 1$, and $b = \frac{1}{6}$.

By density arguments, this result can be extended to the case where $\mathcal{F}$ is a unit ball of a linear normed space, after checking that $f \to \nu_n(f)$ is continuous and $\mathcal{F}$ contains a countable dense family.

Theorem 7.2 (Bernstein Matrix inequality)**.** *Consider a finite sequence $\{\mathbf{S}_k\}$ of independent, random matrices with common dimension $d_1 \times d_2$. Assume that*

$$\mathbb{E}\mathbf{S}_k = 0 \quad \text{and} \quad \|\mathbf{S}_k\|_{\mathrm{op}} \le L \quad \text{for each index } k.$$

Introduce the random matrix $\mathbf{Z} = \sum_k \mathbf{S}_k$. Let $\nu(\mathbf{Z})$ be the the variance statistic of the sum: $\nu(\mathbf{Z}) = \max\{\lambda_{\max}(\mathbb{E}[\mathbf{Z}'\mathbf{Z}]), \lambda_{\max}(\mathbb{E}[\mathbf{Z}\mathbf{Z}'])\}$. Then, for all $t \ge 0$

$$\mathbb{P}\left[\|\mathbf{Z}\|_{\mathrm{op}} \ge t\right] \le (d_1 + d_2)\exp\left(-\frac{t^2/2}{\nu(\mathbf{Z}) + Lt/3}\right).$$

A proof can be found in Tropp (2012) or Tropp (2015).

8. Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[Abramowitz and Stegun, 1964] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition.

[Arlot, 2007] Arlot, S. (2007). Rééchantillonnage et sélection de modèles. Ph.D. thesis, Université Paris Sud.

[Baraud, 2002] Baraud, Y. (2002) Model selection for regression on a random design. *ESAIM Probab. Statist.* **6**, 127-146.

[Barron et al., 1999] Barron, A., Birgé, L. and Massart, P. (1999) Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113**, 301-413.

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

[Cohen et al., 2013] Cohen, A., Davenport, M.A. and Leviatan, D. (2013). On the stability and acuracy of least squares approximations. *Found. Comput. math.* **13**, 819-834.

[Comte and Genon-Catalot, 2018a] Comte, F. and Genon-Catalot, V. (2018a). Laguerre and Hermite bases for inverse problems. *Journal of the Korean Statistical Society*, **47**, 273-296.

[Comte and Genon-Catalot, 2018b] Comte, F. and Genon-Catalot, V. (2018b). Regression function estimation as a partly inverse problem. Preprint Hal

[Comte and Genon-Catalot, 2019] Comte, F. and Genon-Catalot, V. (2019). Drift estimation on non compact support for diffusion models. Preprint Hal

[Comte and Rozenholc, 2002] Comte, F. and Rozenholc, Y. (2002) Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Process. Appl.* **97**, 111-145.

[Galtchouk and Pergamenshchikov, 2009] Galtchouk, L. and Pergamenshchikov, S. (2009) Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression. *J. Korean Statist. Soc.* **38**, 305-322.

[Gendre, 2008] Gendre, X. (2008) Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression. *Electron. J. Stat.* **2**, 1345-1372.

[Jin *et al.* (2015)] Jin, S., Su, L. and Xiao, Z. (2015) Adaptive nonparametric regression with conditional heteroskedasticity. *Econometric Theory* **31**, 1153-1191.

[Klein and Rio, 2005] Klein, T. and Rio, E. (2005) Concentration around the mean for maxima of empirical processes. *Ann. Probab.* **33**, no. 3, 1060-1077.

[Szegö, 1975] Szegö, G. (1975) *Orthogonal polynomials.* Fourth edition. American Mathematical Society, Colloquium Publications, Vol. XXIII. American mathematical Society, Providence, R.I.

[Tropp, 2012] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434.

[Tropp, 2015] Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230.

[Tsybakov, 2009] Tsybakov, A. B. (2009) Introduction to nonparametric estimation. Springer Series in Statistics. Springer, New York.